UNIVERSIDADE FEDERAL DE JUIZ DE FORA
PROGRAMA DE PÓS-GRADUAÇÃO EM MODELAGEM COMPUTACIONAL

João Paulo Scoralick de Oliveira

Cenários de aplicação de técnicas de aprendizado de máquina para a predição de estágios da doença renal crônica em uma base de dados do sistema público de saúde do Brasil

Juiz de Fora

2024

João Paulo Scoralick de Oliveira

Cenários de aplicação de técnicas de aprendizado de máquina para a predição de estágios da doença renal crônica em uma base de dados do sistema público de saúde do Brasil

Tese apresentada ao Programa de Pós-graduação em Modelagem Computacional da Universidade Federal de Juiz de Fora como requisito à obtenção parcial do título de Doutor em Modelagem Computacional. Área de concentração: Modelagem Computacional.

Orientadora: Prof[a] D.Sc. Priscila Vanessa Zabala Capriles Goliatt

Coorientador: Prof. D.Sc. Leonardo Goliatt da Fonseca

Juiz de Fora

2024

Ficha catalográfica elaborada através do programa de geração
automática da Biblioteca Universitária da UFJF,
com os dados fornecidos pelo(a) autor(a)

de Oliveira, Joao Paulo Scoralick .
Cenários de aplicação de técnicas de aprendizado de máquina para a predição de estágios da doença renal crônica em uma base de dados do sistema público de saúde do Brasil / Joao Paulo Scoralick de Oliveira. -- 2024.
168 p. : il.

Orientadora: Priscila Vanessa Zabala Capriles Goliatt
Coorientador: Leonardo Goliatt da Fonseca
Tese (doutorado) - Universidade Federal de Juiz de Fora, ICE/Engenharia. Programa de Pós-Graduação em Modelagem Computacional, 2024.

1. Aprendizado de Máquina. 2. Algoritmo de Classificação. 3. Doença Renal Crônica. 4. Taxa de Filtração Glomerular. 5. Sistema Único de Saúde. I. Goliatt, Priscila Vanessa Zabala Capriles, orient. II. da Fonseca, Leonardo Goliatt , coorient. III. Título.

João Paulo Scoralick de Oliveira

Cenários de aplicação de técnicas de aprendizado de máquina para a predição de estágios da doença renal crônica em uma base de dados do sistema público de saúde do Brasil

Tese apresentada ao Programa de Pós-Graduação em Modelagem Computacional da Universidade Federal de Juiz de Fora como requisito parcial à obtenção do título de Doutor em Modelagem Computacional. Área de concentração: Modelagem Computacional.

Aprovada em 12 de setembro de 2024.

BANCA EXAMINADORA

Profª.Drª. Priscila Vanessa Zabala Capriles Goliatt - Orientadora
Universidade Federal de Juiz de Fora

Prof.Dr. Leonardo Goliatt da Fonseca - Coorientador
Universidade Federal de Juiz de Fora

Prof.Dr. Heder Soares Bernardino
Universidade Federal de Juiz de Fora

Prof.Dr. Fernando Antonio Basile Colugnati
Universidade Federal de Juiz de Fora

Prof.Dr. Douglas Adriano Augusto

Fundação Oswaldo Cruz

Prof.Dr. Eduardo Krempser da Silva

Fundação Oswaldo Cruz

Juiz de Fora, 01/11/2024.

Documento assinado eletronicamente por **Leonardo Goliatt da Fonseca**, **Professor(a)**, em 01/11/2024, às 15:25, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Heder Soares Bernardino**, **Professor(a)**, em 03/11/2024, às 01:00, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Fernando Antonio Basile Colugnati**, **Servidor(a)**, em 03/11/2024, às 20:44, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Priscila Vanessa Zabala Capriles Goliatt**, **Professor(a)**, em 04/11/2024, às 05:43, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Douglas Adriano Augusto**, **Usuário Externo**, em 04/11/2024, às 10:09, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Eduardo Krempser da Silva**, **Usuário Externo**, em 06/11/2024, às 15:13, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

A autenticidade deste documento pode ser conferida no Portal do SEI-Ufjf (www2.ufjf.br/SEI) através do ícone Conferência de Documentos, informando o código verificador **2074559** e o código CRC **FD97F8BD**.

Dedicado à minha família.

AGRADECIMENTOS

Agradeço a Deus por ter guiado os meus passos desde sempre.

Agradeço à minha esposa, Carolina, pelo amor, companheirismo, amizade, respeito, incentivo, paciência e compreensão imensuráveis desde a adolescência, que só aumentam ao longo dos anos. Não consigo descrever em palavras tudo o que você fez e faz por mim. É o mais puro clichê, mas sem você não teria conseguido chegar até a esta etapa. Você sabe muito bem disso! E, além de tudo, ainda se tornou doutora antes de mim.

Agradeço aos meus pais, Paulo e Rose, por serem meus grandes amigos e ídolos. Simplesmente, sem o amor incondicional deles, absolutamente nada na minha vida teria sido possível, nem teria feito sentido. Espero que um dia eu consiga retribuir pelo menos um pouco de tudo o que vocês fizeram por mim.

Aos meus irmãos, Carlos Eduardo e Pedro Henrique, pela amizade, lealdade, parceria e apoio constantes. Tenho muito orgulho de ter vocês como meus irmãos, não só por serem grandes botafoguenses, mas, principalmente por serem pessoas diferenciadas e especiais.

Aos meus queridos avós, Bernardo, Mauro, Mabel e Rachel (*in memoriam*), pelo amor, incentivo, carinho e presença constantes ao longo de todas as etapas da minha vida. Aos meus tios e tias, especialmente às minhas madrinhas, Rosane e Neuza, e ao meu padrinho, Kiko, pelo incentivo e apoio desde sempre.

Agradeço também aos meus sogros, Conceição e José Mauro (*in memoriam*), por terem me recebido e acolhido em sua família desde o início, sempre me respeitando, me apoiando e me tratando como se eu fosse mais um filho.

Agradeço profundamente aos meus orientadores, profª Priscila Capriles e profº Leonardo Goliatt, pelo apoio, acolhimento, incentivo, ensinamentos, compreensão e amizade desde o período de preparação do pré-projeto para ingresso no curso de doutorado, e durante todos os momentos e etapas desta jornada. Vocês foram mais do que fundamentais e necessários para que este trabalho pudesse ser realizado. Antes de serem excelentes profissionais, são pessoas da melhor qualidade. O meu agradecimento a vocês é eterno.

À colega Gabriele Iwashima, pela longa parceria e pelas contribuições fundamentais proporcionadas desde o início do desenvolvimento da pesquisa apresentada neste trabalho. E também à colega Larissa de Lima e Silva, pelo inestimável auxílio prestado na elaboração de conceitos e testes desenvolvidos.

Ao prof. Fernando Colugnati, pela colaboração constante e de primeira hora, que foi fundamental para a minha compreensão de todo o contexto que envolve a doença renal crônica. E também pela disponibilidade e pelo pronto atendimento em participar da banca desta tese. Aos professores Heder Bernardino e Eduardo Krempser, pela atenção e

disponibilidade dedicadas à leitura e à avaliação deste trabalho desde a etapa de qualificação. E também ao professor Douglas Augusto, pela disponibilidade e pelo interesse em participar da banca de defesa desta tese. Muito obrigado!

Aos professores Bernardo Martins e Flávia Bastos, e à Maíra Macário de Assis, que me auxiliaram em diversos momentos do curso, oferecendo suporte em questões essenciais para a continuidade do meu trabalho.

Uma parte fundamental desta etapa de conclusão do doutorado passa pelo meu período de trabalho no Itaú Unibanco. Agradeço profundamente à Cristina Schöwe e ao Eduardo Yokoyama por terem acreditado no meu potencial, me proporcionando a oportunidade que mudou a minha vida. À futura doutora Désirée Fadel, pela amizade, parceria, ensinamentos e incentivos constantes desde o início da jornada. Aos amigos César Seiji, Barbara Moraes, Ingrid Toro, Karoliny dos Anjos, Vinícius Gava e Leonardo José Azevedo, pela grande parceria e amizade ao longo dos anos. Não poderia deixar de agradecer a três pessoas que foram fundamentais para a conclusão desta etapa: Raphael Pastore, Bruno Ito Vargas e Francis Pontes, pela valorização do meu trabalho no doutorado, pelos esforços empregados em minha defesa e pela oportunidade que foi imprescindível para que eu tivesse a tranquilidade e o apoio necessários para a finalização do meu doutorado.

Por fim, agradeço ao Programa de Pós-Graduação em Modelagem Computacional da Universidade Federal de Juiz de Fora, em nome de seus professores e funcionários, pela oportunidade de cursar meu doutorado e de ampliar a minha formação acadêmica e o meu desenvolvimento pessoal.

RESUMO

A doença renal crônica (DRC) é um grave problema de saúde pública tanto no Brasil quanto no mundo. Caracterizada pela perda progressiva e irreversível da função renal, a DRC apresenta uma prevalência média projetada de 9,5% na população adulta mundial. No Brasil, milhões de indivíduos possuem o diagnóstico da doença, tendo esse número atingido um total estimado de 17 milhões em 2017. Intervenções precoces podem retardar sua progressão e reduzir a necessidade de terapias renais substitutivas. A antecipação da detecção da DRC, em cada um de seus seis estágios de gravidade, é essencial para o adequado manejo clínico dos pacientes, representando o método mais eficaz para a redução dos custos associados ao tratamento. O objetivo desta tese foi desenvolver cenários de aplicação de algoritmos e técnicas de aprendizado de máquina (AM) para a predição dos estágios da DRC, utilizando uma base de dados de saúde pública do Brasil, composta por mais de sete mil registros que incluem informações pessoais, socioeconômicas, clínicas e laboratoriais dos pacientes. A proposta foi explorar diferentes combinações de variáveis preditoras que pudessem servir de base para a aplicação de métodos de AM capazes de predizer a progressão da doença, reduzindo a dependência de marcadores tradicionais, como a creatinina sérica, amplamente utilizada nos trabalhos da literatura relacionada. A partir dessas combinações de variáveis relacionadas a exames clínicos e dados pessoais, esta tese também teve como objetivo identificar o menor conjunto viável de variáveis preditoras dos seis estágios da DRC. Para o desenvolvimento dos cenários, foram exploradas três diferentes configurações da base de dados, com o objetivo de extrair distintas perspectivas sobre a representação do conteúdo. Métodos variados foram aplicados para a inferência de dados faltantes, juntamente com técnicas voltadas ao balanceamento dos dados, seleção de variáveis relevantes e divisão dos dados para fins de classificação. Adicionalmente, uma série de algoritmos supervisionados de AM, com diferentes fundamentações e objetivos, foram empregados para permitir uma análise comparativa dos resultados. Em cada cenário proposto, os resultados obtidos variaram. No primeiro, com classificações realizadas pelo algoritmo de floresta aleatória (RF, do inglês *random forest*), todas as abordagens apresentaram valores elevados de acurácia, exceto o agrupamento que não incluiu a creatinina sérica como variável preditora. No segundo cenário, composto por 25 variáveis e sem a inclusão da creatinina, o algoritmo *extreme gradient boosting* (XGBoost) apresentou alta acurácia, comparável aos valores reportados na literatura, em estudos que fazem uso da creatinina em suas análises. No terceiro cenário, o desbalanceamento da base de dados foi tratado com diferentes métodos e a classificação considerou apenas três variáveis preditoras. Embora os resultados gerais tenham ficado aquém do esperado, alguns se revelaram promissores para a detecção dos estágios iniciais da DRC. No quarto cenário, a inferência de dados faltantes foi abordada por meio do conceito de cópulas, mas os resultados foram insatisfatórios. Por fim, o quinto

cenário foi o mais completo em termos de organização, tratamento, seleção e classificação dos dados. Entretanto, as novas abordagens não resultaram em melhorias significativas nos resultados. Em conclusão, uma parte dos cenários desenvolvidos foi bem sucedida em corresponder aos objetivos delineados nesta tese, sobretudo por não prescindir do uso de marcadores tradicionais da doença. Os resultados promissores possivelmente poderiam ser avaliados para o uso na prática clínica diária e no auxílio ao diagnóstico precoce da doença renal crônica.

Palavras-chave: Aprendizado de Máquina. Algoritmo de Classificação. Doença Renal Crônica. Taxa de Filtração Glomerular. Sistema Único de Saúde.

ABSTRACT

Chronic kidney disease (CKD) represents a significant public health concern in Brazil and globally. Characterized by the progressive and irreversible loss of kidney function, CKD has an estimated average prevalence of 9.5% among the global adult population. In Brazil, millions have been diagnosed with the disease, with the total reaching an estimated 17 million in 2017. Early interventions can slow disease progression and reduce the need for renal replacement therapies. Early detection of CKD across its six clinical stages is critical for appropriate clinical management and is the most effective approach to reducing treatment-associated costs.

This thesis aimed to develop application scenarios for algorithms and machine learning (ML) techniques to predict CKD stages using a Brazilian public health database comprising over seven thousand records containing personal, socioeconomic, clinical, and laboratory information from patients. The proposal explored various combinations of predictor variables to serve as a basis for ML methods capable of predicting disease progression, thereby reducing reliance on traditional markers such as serum creatinine, commonly used in related literature. Based on these variable combinations, connected to clinical exams and personal data, this work also sought to identify the smallest viable set of predictor variables for the six stages of CKD.

In developing the scenarios, three different dataset configurations were explored to derive distinct perspectives on content representation. Various methods were applied to infer missing data, along with techniques aimed at balancing the data, selecting relevant variables, and partitioning the data for classification purposes. Additionally, supervised ML algorithms with diverse theoretical foundations and objectives were employed to facilitate a comparative analysis of the results.

The outcomes varied across the proposed scenarios. In the first scenario, classifications were performed using the random forest (RF) algorithm, with all approaches achieving high accuracy, except for the dataset excluding serum creatinine as a predictor variable. In the second scenario, which included 25 variables but excluded creatinine, the extreme gradient boosting (XGBoost) algorithm demonstrated high accuracy comparable to values reported in the literature, despite the latter's inclusion of creatinine. In the third scenario, dataset imbalance was addressed using different methods, and classification was performed based on only three predictor variables. Although the overall results did not meet expectations, some findings were promising for detecting early CKD stages. In the fourth scenario, missing data inference was handled using the copula-based approach, but results were unsatisfactory. Lastly, the fifth scenario was the most comprehensive in terms of data organization, processing, selection, and classification; however, the new approaches did not lead to significant improvements in results.

In conclusion, some of the developed scenarios successfully met the objectives outlined in this thesis, especially as they retained the use of traditional disease markers. The promising results may have potential applications in daily clinical practice and could assist in the early diagnosis of chronic kidney disease.

Keywords: Machine Learning. Classification Algorithm. Chronic Kidney Disease. Glomerular Filtration Rate. Public Health System.

LISTA DE ILUSTRAÇÕES

LISTA DE ABREVIATURAS E SIGLAS

AdaBoost	*Adaptive Boosting*
ADASYN	*Adaptive Synthetic Sampling*
AE	Algoritmos de *ensemble*
AM	Aprendizado de Máquina
ANN	*Artificial Neural Network*
ANS	Aprendizado Não Supervisionado
AR	Aprendizado por Reforço
AS	Aprendizado Supervisionado
ASS	Aprendizado Semissupervisionado
BERT	*Bidirectional Encoder Representations from Transformers*
BS	*Borderline Synthetic Minority Over-sampling Technique*
CART	*Classification and Regression Tree*
CatBoost	*Categorical Boosting*
CBD	Censo Brasileiro de Diálise
CDC	*Centers for Disease Control and Prevention*
CEAE	Centro Estadual de Assistência Especializada
CG	*Cockcroft-Gault*
CH	Centro Hiperdia de Juiz de Fora
CHDM	Centro Hiperdia Minas
CHIRP	*Composite Hypercube on Iterated Random Projection*
CKD-EPI	*Chronic Kidney Disease Epidemiology Collaboration*
CPU	*Central Processing Unit*
DCNT	Doença Crônica Não Transmissível
DM	Diabetes *Mellitus*
DRC	Doença Renal Crônica
DT	*Decision Tree*
EL	*Ensemble Learning*
ELM	*Extreme Learning Machine*
ENN	*Edited Nearest Neighbor*
ETC	*Extra Trees Classifier*
EUA	Estados Unidos da América
FN	*False Negative*
FNN	*Feedforward Neuronal Network*
FP	*False Positive*
FT	*Functional Trees*
GAN	*Generative Adversarial Networks*
GB	*Gradient Boosting*
GPT	*Generative Pre-trained Transformer*
HAS	Hipertensão Arterial Sistêmica
IBGE	Instituto Brasileiro de Geografia e Estatística

IA	Inteligência Artificial
IMEPEN	Fundação Instituto Mineiro de Estudos e Pesquisas em Nefrologia
IMC	Índice de Massa Corporal
IRC	Insuficiência Renal Crônica
KDIGO	*Kidney Disease: Improving Global Outcome*
KDOQI	*Kidney Disease Outcome Quality Initiative*
KNN	*K-Nearest Neighbors*
LDA	*Linear Discriminant Analysis*
LightGBM	*Light Gradient-Boosting Machine*
LIME	*Local Interpretable Model-agnostic Explanations*
LogR	*Logistic Regression*
LR	*Linear Regression*
MAR	*Missing at Random*
MCAR	*Missing Completely at Random*
MDRD	*Modification of Diet in Renal Disease*
MICE	*Multiple Imputation by Chained Equations*
MLP	*Multilayer Perceptron*
MLR	*Multiple Linear Regression*
MMQ	Método dos Mínimos Quadrados
MNAR	*Missing Not At Random*
MSE	*Mean Squared Error*
NB	*Naive Bayes*
NBTree	*Naive Bayes Tree*
NFK	*National Kidney Foundation*
NLP	*Natural Language Processing*
OMS	Organização Mundial da Saúde
PaLM	*Pathways Language Model*
PCA	*Principal Component Analysis*
PD	Pré-diálise
pmp	Pacientes por Milhão da População
PNN	*Probabilistic Neural Network*
ppm	Partes por Milhão
PR	*Polynomial Regression*
PRCS	Primeiro Registro de Creatinina Sérica
PRE	Primeiro Registro de Estágio
PRISMA	*Preferred Reporting Items for Systematic Reviews and Meta-Analyses*
RAM	*Random Access Memory*
RBF	*Radial Basis Function*
RF	*Random Forest*
RFE	*Recursive Feature Elimination*
ROC	*Receiver Operating Characteristic*
ROC AUC	*Area Under the Receiver Operating Characteristic Curve*

SBN	Sociedade Brasileira de Nefrologia
SE	*Synthetic Minority Over-sampling Technique with Edited Nearest Neighbor*
SES/MG	Secretaria de Estado de Saúde de Minas Gerais
SHAP	*SHapley Additive exPlanations*
SLANH	Sociedade Latino Americana de Nefrologia e Hipertensão
SLFN	*Single Hidden Layer Feedforward Neural Network*
SLR	*Simple Linear Regression*
SMOTE	*Synthetic Minority Over-sampling Technique*
SMOTE-TEK	*Synthetic Minority Over-sampling Technique for Time Series with Efficient Kernels*
ST	*Synthetic Minority Over-sampling Technique with Edited Nearest Neighbor with Tomek Links*
SUS	Sistema Único de Saúde
SVM	*Support Vector Machine*
TD	Terapia Dialítica
TFG	Taxa de Filtração Glomerular
TGP	Transaminase Glutâmico Pirúvica
TN	*True Negative*
TP	*True Positive*
TR	Transplante Renal
TRS	Terapia Renal Substitutiva
TSH	*Thyroid Stimulating Hormone*
UCI	*University of California, Irvine*
UFJF	Universidade Federal de Juiz de Fora
UFS	*Univariate Feature Selection*
URE	Último Registro de Estágio
US	*Univariate Selection*
XAI	*Explainable Artificial Intelligence*
XGBoost	*Extreme Gradient Boosting*

SUMÁRIO

1 INTRODUÇÃO

As doenças crônicas não transmissíveis (DCNTs) são as principais causadoras de morte em todo o mundo (155) (167) (267). No Brasil, esse cenário não é diferente, com as DCNTs sendo responsáveis por aproximadamente três quartos dos óbitos em 2017 (263). Embora o cenário atual indique uma tendência de redução no número de mortes (156), com uma previsão de diminuição contínua até 2025 (155), as DCNTs permanecem sendo um desafio significativo para a saúde pública, especialmente em países onde uma parcela considerável da população pertence a grupos em situação de vulnerabilidade socioeconômica, como aqueles de baixa escolaridade e de baixa renda (155) (156) (263).

1.1 Caracterização do problema

A doença renal é um tipo de DCNT de elevada prevalência a nível global, superando em incidência outras doenças crônicas não transmissíveis priorizadas pela Organização Mundial da Saúde (OMS), como as doenças cardiovasculares, o câncer, a diabetes *mellitus* (DM) e a hipertensão arterial (HAS) (137) (263). Em particular, a DM e a HAS desempenham um papel significativo na deterioração da função renal (56), ao promoverem lesões nos néfrons, as unidades funcionais dos rins responsáveis pela filtração renal. Tais lesões podem comprometer, a longo prazo, tanto a anatomia quanto a função renal, levando ao desenvolvimento da doença renal crônica (DRC), que atualmente afeta cerca de 10% da população mundial, número que pode ser ainda maior devido aos casos de subdiagnóstico (268).

A DRC é uma doença silenciosa, não raramente assintomática em seus estágios iniciais, de relevância e incidência globais, mas que ainda não recebe o financiamento e a abordagem adequados para o seu efetivo combate (85). Apenas 48% dos governos nacionais em todo o planeta consideram a DRC e suas medidas de tratamento e prevenção como uma prioridade nas políticas de saúde (137). Intervenções eficazes podem ser implementadas no nível da atenção primária dos sistemas de saúde público, possibilitando a prevenção da progressão da doença para estágios mais avançados e reduzindo a necessidade de terapias renais substitutivas (TRS), como a diálise e o transplante renal (TR) (264). No contexto brasileiro, a implantação de programas de atendimento pré-dialítico demonstra o potencial de gerar uma economia superior a 33 mil reais no custo médio para cada ano de tratamento dialítico que seja evitado (165).

A estratificação de risco de pacientes doentes renais é realizada por meio de seis estágios classificados através de dados pessoais e clínicos — com destaque para a creatinina sérica — inclusos em equações que calculam a taxa de filtração glomerular (TFG) (264). É fundamental a compreensão das razões clínicas e individuais responsáveis por encaminhar os pacientes a cada um desses estágios e como isso acontece. A detecção precoce da

DRC também é crucial para identificar os estágios de gravidade da doença, auxiliando na diminuição de custos e possibilitando a aplicação de outros métodos de tratamento mais eficientes. Durante o seu acompanhamento clínico, cada paciente tem seu valor de TFG calculado de acordo com a demanda, uma vez que pode ter seu estágio da DRC modificado ao longo do tempo, indicando o grau de evolução do seu tratamento (265).

A utilização de métodos de inteligência artificial (IA), com ênfase nos algoritmos de aprendizado de máquina (AM), tem se expandido substancialmente nos últimos anos. Esta expansão é resultado do contínuo aprimoramento das técnicas de AM e de sua capacidade de elevar tanto a eficácia quanto a personalização no cuidado à saúde, contribuindo para avanços significativos no diagnóstico e na predição de desfechos clínicos em diversas doenças (82).

Diversos trabalhos (9), (10), (90), (105), (126), (136), (180), (191), (199), (201), (204) e (243) utilizaram a base de dados de doentes renais crônicos do repositório *online* da *University of California, Irvine* (UCI) (210) para a predição do risco da DRC. Essa base é formada por dados de 400 indivíduos de origem indiana e possui 25 atributos, entre dados pessoais e clínicos, utilizados em um problema binário de predição do risco da DRC. Um parcela expressiva dos estudos encontrados na literatura de AM e DRC faz uso desses dados por meio da implementação de diferentes métodos de inferência de dados faltantes e de algoritmos de classificação, como árvores funcionais (FT, do inglês *functional trees*), floresta aleatória (RF, do inglês *random forest*), máquina de vetores de suporte (SVM, do inglês *support vector machines*) e *extreme gradient boosting* (XGBoost). Muitos desses trabalhos chegaram a resultados com valores médios de acurácia de 97% na predição do risco da DRC (259). No entanto, a base de dados da UCI apresenta vieses importantes, os quais foram destacados por Wang *et al.* (2021) (259). Esses dados representam uma população com perfil étnico homogêneo, característica que pode ser limitadora na generalização dos resultados para outras populações (259).

O estudo de Debal e Sitote (2022) (65) utilizou os algoritmos árvores de decisão (DT, do inglês *decision tree*), RF, SVM e XGBoost na predição dos estágios da DRC em uma base de dados de pacientes etíopes. Já Ghosh e Khandoker (2024) (96), fizeram uso de dados provenientes de um hospital dos Emirados Árabes Unidos. Silveira *et al.* (2021) (228) utilizaram dados clínicos e pessoais de pacientes atendidos em um hospital público de Maceió, no Brasil. Embora este estudo tenha sido o único, entre os mencionados, a considerar um cenário com potencial para variabilidade étnica, a base de dados não incluiu variáveis relacionadas a fatores raciais. Além disso, as análises foram realizadas com uma amostra limitada a apenas 60 pacientes.

Conforme detalhado na revisão sistemática conduzida por Sanmarchi *et al.* (2023), os 68 estudos selecionados, que empregam técnicas de aprendizado de máquina aplicadas à doença renal crônica, utilizaram bases de dados provenientes da Ásia, Europa, América do

Norte e África. A América do Sul é citada em apenas um estudo, mas de forma equivocada, dado que Guo *et al.* (2020) (103) utilizaram dados clínicos de hospitais localizados na China, com pacientes de perfil étnico chinês, segundo os próprios autores. De forma similar, a revisão sistemática realizada por Khalid *et al.* (2024) (135) também destacou exclusivamente estudos dos primeiros quatro continentes mencionados. Nesta tese, os dados analisados são provenientes de um centro interdisciplinar de saúde pública no Brasil, compreendendo mais de sete mil registros pessoais, socioeconômicos, étnicos, clínicos e laboratoriais de pacientes atendidos pelo Sistema Único de Saúde (SUS), ao longo de um período de acompanhamento aproximado de quatro anos.

Por ser um marcador fundamental da função renal devido à sua relação direta com o cálculo da TFG, a creatinina sérica é uma variável amplamente utilizada em modelos de classificação da progressão da DRC. Todos os estudos revisados nesta tese a utilizaram como variável preditora nos processos de classificação desenvolvidos. E Sanmarchi *et al.* (2023) também a destacaram entre as variáveis mais utilizadas nos estudos selecionados pelos autores.

Diante do panorama apresentado, torna-se imperativa a realização de pesquisas que proponham novas abordagens para a predição dos estágios da DRC. Os estudos devem focar não apenas em bases de dados que possuam maior variabilidade étnica e socioeconômica (3) (148) (208), como também em conjuntos de dados com volume de registros de pacientes superiores aos observados na maioria das pesquisas disponíveis na literatura. Além disso, é fundamental fomentar investigações que permitam o acompanhamento da progressão da DRC em diferentes estágios, utilizando exames laboratoriais simples e amplamente acessíveis, evitando a dependência de exames convencionais, como a creatinina sérica, que pode ser limitada em certos contextos. A exploração de métodos alternativos pode contribuir para a melhoria da predição e do manejo clínico da DRC em populações diversas.

1.2 Motivação

Com base nos fatos apresentados, a principal motivação desta tese é contribuir para o desenvolvimento de cenários diversos de aplicação de algoritmos e técnicas de aprendizado de máquina voltados à predição dos estágios clínicos da doença renal crônica. Nesse contexto, a inovação central reside na identificação empírica do menor conjunto possível de variáveis, composto por dados pessoais e exames clínicos, que permita a predição dos estágios da DRC sem a necessidade de utilização de marcadores laboratoriais convencionais, como a creatinina sérica.

O diferencial desta proposta reside na identificação de modelos preditivos capazes de utilizar de forma eficaz variáveis alternativas aos padrões clínicos convencionais. Essa abordagem visa não apenas simplificar o processo de predição dos estágios da DRC, mas também torná-lo mais acessível em contextos nos quais exames laboratoriais complexos ou

de alto custo representam uma limitação. Além disso, devido à estruturação e à coleta dos dados na base utilizada, o problema de classificação dos estágios da DRC apresenta-se como uma tarefa de natureza multiclasse, diferindo da maioria dos estudos mencionados.

Com a proposição de diferentes cenários, espera-se que os resultados obtidos possam contribuir para aprimorar o tratamento e o acompanhamento da evolução de pacientes doentes renais crônicos em diferentes estágios de gravidade, promovendo maior eficácia e redução de custos. Como demonstrado por Moraes *et al.* (2021) (165), a otimização do tratamento pré-dialítico, em detrimento dos diferentes tipos de TRS, constitui o método mais eficaz para a redução dos custos associados à DRC.

Para atingir os objetivos propostos, foram desenvolvidos cinco cenários que exploram diferentes formas de organização dos dados e de aplicação de algoritmos e técnicas de AM para a predição dos estágios da DRC na base de dados considerada. Os dois primeiros cenários adotaram abordagens mais objetivas, focadas na classificação e na organização das variáveis preditoras em agrupamentos variados, com o intuito de realizar uma análise exploratória e inicial dos dados e compará-los com os estudos previamente encontrados na literatura. Já os três cenários subsequentes, seguiram abordagens mais detalhadas, englobando diversas etapas e técnicas de tratamento, organização, classificação e avaliação dos dados. O objetivo central desses últimos cenários foi identificar o menor conjunto possível de exames capaz de predizer os estágios da DRC com a maior eficácia possível. Os resultados obtidos apresentaram variações em termos de eficácia: alguns mostraram-se promissores e alinhados com a literatura existente, ao passo que outros introduziram abordagens inovadoras. Houve também cenários que não atingiram os resultados esperados, mas que, ainda assim, geraram discussões relevantes, capazes de fundamentar futuras análises e orientar o desenvolvimento de novos estudos.

1.3 Objetivos

1.3.1 Objetivo Geral

O objetivo deste trabalho foi desenvolver cenários de aplicação de técnicas de aprendizado de máquina voltadas à predição dos estágios da doença renal crônica, por meio da identificação de variáveis que pudessem se distanciar dos padrões tradicionais empregados na prática clínica e laboratorial.

1.3.2 Objetivos Específicos

A partir da definição do objetivo principal, podem ser destacados os seguintes objetivos específicos.

- Desenvolver cenários que considerem vários agrupamentos distintos de dados pessoais e exames clínicos para a predição dos estágios da DRC, de forma a se obter um

entendimento analítico dos dados considerados;

- Selecionar e aplicar diferentes algoritmos e técnicas de aprendizado de máquina em cada um dos cenários desenvolvidos;
- Identificar o menor conjunto factível de variáveis preditoras dos estágios da doença renal crônica, que seja viável, clinicamente adequado e que se diferencie dos padrões clínicos e laboratoriais convencionais;
- Avaliar as principais contribuições de cada um dos cenários propostos;
- Possibilitar que as estratégias desenvolvidas nos cenários possam contribuir para a detecção precoce dos estágios da DRC, criando condições alternativas e favoráveis à postergação ou até mesmo à prevenção da necessidade de terapias renais substitutivas. Esse avanço poderia resultar em uma significativa redução de custos e melhorias substanciais na qualidade de vida dos pacientes.

2 A DOENÇA RENAL CRÔNICA

A doença renal crônica (DRC) é definida como a perda gradativa e sem possibilidade de reversão da estrutura ou da função renal (206) por um período superior a três meses (17) (73) (174). Em sua fase terminal, quando os rins se tornam incapazes de suprir as necessidades básicas de um indivíduo, a DRC é definida como insuficiência renal crônica (IRC) (206).

Pacientes em estágios mais avançados da DRC, que geralmente apresentam quadro de IRC e passam por algum tipo de terapia renal substitutiva, podem sofrer graves limitações que influenciam o seu dia-a-dia, suas relações sociais, suas aptidões físicas e psicológicas, além de diversas outras áreas (89).

Assim como ocorre para as doenças crônicas não transmissíveis (DCNTs), os principais fatores que influenciam no desenvolvimento da DRC são a diabetes *mellitus* (DM) e a hipertensão arterial sistêmica (HAS), sendo esta presente em mais de 75% dos pacientes de todas as idades (15). Ademais, diversos fatores também podem influenciar o surgimento e a progressão da DRC: idade, doenças cardiovasculares, doenças respiratórias, doenças autoimunes, doenças hepáticas, câncer, relação de parentesco com doentes renais crônicos, entre outros (15) (165) (166).

2.1 Taxa de Filtração Glomerular

A taxa de filtração glomerular (TFG) é um marcador da função renal que realiza a medição da depuração (*clearance*) da filtração do plasma sanguíneo pelos glomérulos renais em uma determinada unidade de tempo (34) (234). Desse modo, é utilizada tanto para a detecção, quanto para o tratamento da DRC.

Seguindo as orientações da *Kidney Disease: Improving Global Outcomes* (KDIGO) — organização global sem fins lucrativos que desenvolve e implementa diretrizes de prática clínica baseadas em evidências das doenças renais — a TFG deve ser estimada a partir da creatinina, substância presente na corrente sanguínea e que é um produto do catabolismo da creatina, composto de aminoácidos encontrado sobretudo no tecido muscular na forma de fosfocreatina (242).

A filtração da creatinina ocorre principalmente nos glomérulos renais (234) e, em uma pequena porção, nos túbulos renais (16). Os rins também são os únicos órgãos responsáveis pela secreção da creatinina, uma vez que ela não é reabsorvida nem reaproveitada pelo organismo, razão também pela qual a concentração de creatinina no plasma sanguíneo influencia diretamente os valores de TFG. Quanto maior o valor de creatinina, há uma menor filtragem do plasma sanguíneo, o que indica que o funcionamento dos rins está deficiente (242).

Na prática clínica, a depuração da creatinina é comumente utilizada para a avaliação da TFG, por meio da coleta de urina durante um período de 24 horas, no qual a proporção entre as quantidades de creatinina no sangue e na urina é avaliada. Essa abordagem apresenta algumas limitações, como a necessidade da coleta intermitente de urina durante as 24 horas — que pode ser problemática para certos grupos de pacientes — e o fato da creatinina não ser secretada exclusivamente pelos glomérulos, de forma que a concentração urinária não representa a totalidade da concentração de creatinina. Portanto, o uso da depuração de creatinina é indicado apenas em casos específicos, geralmente em pacientes que apresentam valores de TFG superiores a 60 mL/min/1,73 m^2 (valor da superfície corporal média) e certas patologias associadas (16).

Com o intuito de minimizar as limitações sobre o uso da creatinina e de sua depuração na avaliação da função renal, diversas equações para o cálculo do valor de TFG tem sido elaboradas (234). As três mais utilizadas e frequentemente encontradas na literatura são: *Cockcroft-Gault* (CG) (54), *Modification of Diet in Renal Disease* (MDRD) (142) (143) (244) e *Chronic Kidney Disease Epidemiology Collaboration* (CKD-EPI) (145).

A Equação CG (2.1), inicialmente publicada em 1973 (54), foi pioneira em obter ampla aceitação na comunidade científica para a estimativa da depuração da creatinina sérica (34). Seu desenvolvimento baseou-se em um estudo conduzido com 249 homens de etnia caucasiana, hospitalizados, sem lesões renais e com função renal normal, apresentando uma idade média de 57 anos, variando entre 18 e 92 anos.

$$CG(ml/min) = \frac{(140 - Idade) \times Peso}{72 \times Creatinina} \times 0,85 \text{ (se for mulher)} \tag{2.1}$$

sendo a idade expressa em anos, o peso em quilogramas e a creatinina em mg/dL. Na equação, foi desconsiderada a superfície corporal média dos indivíduos; portanto, para as mulheres, deve-se adicionar um fator de correção de 0,85 (16) (34).

Um dos grandes limitadores do uso da Equação CG é a superestimação dos seus resultados para a obtenção da TFG, uma vez que a porção da secreção de creatinina realizada pelos túbulos renais não é considerada. Ademais, o valor do peso corporal de um paciente nem sempre está disponível para o cálculo de sua TFG e está sempre sujeito a variações frequentes (34).

A Equação MDRD foi inicialmente publicada em 1992 (244) e englobou apenas pacientes doentes renais crônicos. Ao contrário da depuração de creatinina utilizada na Equação CG, a elaboração da MDRD foi baseada no *clearance* de iotalamato-I^{125}, composto orgânico de iodo cuja depuração ocorre via filtração glomerular sem secreção ou reabsorção nos túbulos renais, podendo ser utilizado como ferramenta diagnóstica para a determinação da TFG (16) (34).

Do grupo de indivíduos considerados no estudo de proposição da Equação MDRD (244), 88% eram brancos, não transplantados, parte deles apresentando quadro de nefropatia

diabética, com média de idade de 51 anos e valores de TFG iguais a 40 mL/min/1,73 m^2. Esta unidade de medida de TFG leva em consideração a taxa de depuração, dada em mililitros por minuto, para uma superfície corporal média de 1,73 m^2, valor referente aos indivíduos que participaram do estudo (16).

A Equação MDRD passou por pequenas alterações ao longo dos anos, principalmente em 1999 (142) e 2006 (144), sendo a modificação realizada neste último ano a mais importante, principalmente por apresentar a mesma precisão da equação original e apenas quatro variáveis: valor de creatinina sérica, idade, raça e gênero, como evidencia a Equação 2.2 (expressa em mL/min/1,73 m^2).

$$\begin{aligned} MDRD = 186 \times Creatinina^{-1,154} \times Idade^{-0,203} \times \\ \times 0,742 \text{ (se for mulher)} \times 1,21 \text{ (se for negro)} \end{aligned} \tag{2.2}$$

sendo a idade expressa em anos e a creatinina em mg/dL.

Para mulheres, à Equação 2.2 deve ser adicionado um multiplicador no valor de 0,742 devido à diferença de massa muscular média em relação aos homens. E, por apresentarem maior prevalência para diabetes e para hipertensão arterial (139), no caso de indivíduos negros, a Equação 2.2 deve ser multiplicada por 1,21.

Especialmente para pacientes idosos, a Equação MDRD demonstra precisão elevada na avaliação da função renal (17) (153) (193). Ao contrário do que ocorre com a equação Cockcroft-Gault, a MDRD é menos suscetível a vieses com relação ao peso, uma vez que considera apenas a superfície corporal média de um indivíduo.

Como o estudo MDRD não englobou pessoas saudáveis, a acurácia de sua equação é prejudicada quando aplicada em indivíduos que apresentam função renal dentro dos padrões da normalidade. Para esses casos, o valor de TFG tende a ser superestimado (194). Já para valores de TFG inferiores a 60 mL/min/1,73 m^2, a Equação MDRD apresenta maior acurácia na obtenção dos resultados.

Outro importante limitador da Equação MDRD é o cálculo da TFG para diferentes grupos étnicos (3) (148) (208). A inclusão do multiplicador para pessoas negras na Equação 2.2 é mais eficaz em populações com maior distinção étnica, como nos Estados Unidos da América (EUA), por exemplo, ao contrário da população brasileira em que há uma maior miscigenação racial (34).

Em 2009, foi desenvolvida a Equação CKD-EPI, batizada com o nome do grupo de pesquisa responsável pela sua elaboração. Ao contrário do estudo que gerou a Equação MDRD, o estudo principal que elaborou a CKD-EPI (145) considerou uma coorte composta não somente por indivíduos doentes renais crônicos, mas também por indivíduos que não apresentavam o diagnóstico de DRC. Ao todo, para a elaboração da equação, foram realizados dez estudos com um total de 8.254 participantes. Outros 16 estudos, totalizando 3.896 participantes, foram desenvolvidos para a validação dos resultados anteriores, sendo

que 72% deste último total foram compostos por pacientes oriundos de populações com características de elevado risco para a DRC (16) (34).

Como mostra a Equação 2.3 (145), a CKD-EPI também apresenta as mesmas quatro variáveis da Equação 2.2 e coeficientes específicos para diferentes grupos populacionais. Para mulheres e negros, multiplicadores no valor de 1,018 e no valor de 1,159 devem ser adicionados à Equação 2.3, respectivamente.

$$\begin{aligned} CKD/EPI = 141 \times min(Creat/\kappa, 1)^{\alpha} \times max(Creat/\kappa, 1)^{-1,209} \\ \times 0,993^{Idade} \times 1,018 \text{ (se for mulher)} \times 1,159 \text{ (se for negro)} \end{aligned} \tag{2.3}$$

sendo $\kappa = 0,7$ para mulheres e 0,9 para homens; $\alpha = -0,329$ para mulheres e -0,411 para homens; e sendo o resultado expresso expresso em mL/min/1,73 m², a idade em anos e a creatinina (*Creat*) em mg/dL.

Ao ser comparada com a MDRD, a Equação CKD-EPI possui menor viés na obtenção dos dados e maior desempenho e eficácia na predição de riscos, sobretudo para valores de TFG superiores a 60 mL/min/1,73 m². Já para pacientes que estejam consideravelmente abaixo do peso ideal, a CKD-EPI pode superestimar a TFG. Por outro lado, para pacientes obesos, a equação pode subestimar essa taxa (34).

A CKD-EPI vem sendo considerada, pelos especialistas em nefrologia, a equação mais recomendada para a avaliação da função renal. Todavia, o mais indicado é sempre aplicar a equação cujas características se adequem melhor a cada grupo de pacientes, seguindo as melhores práticas e experiências clínicas (16) (34).

2.2 Diagnóstico

A organização de saúde voluntária estadunidense conhecida como *National Kidney Foundation* (NFK) publicou em 2002, através do seu guia prático de diretrizes, o *Kidney Disease Outcome Quality Initiative* (KDOQI), novas orientações conceituais para o diagnóstico da doença renal crônica. Três foram os componentes principais definidos (17):

- anatômico ou estrutural;
- funcional, baseado na TFG;
- e um componente temporal com pelo menos três meses de duração.

A partir desses componentes, um indivíduo poderia ser diagnosticado como doente renal crônico se apresentasse valor de TFG abaixo de 60 mL/min/1,73 m². Também poderiam ser diagnosticados com DRC os indivíduos com TFG igual ou acima deste valor e em associação a algum dano renal presente há mais de 3 meses, como, por exemplo, a albuminúria persistente com valor superior a 30 mg por 24 horas (17). Outros marcadores também podem ser levados em consideração, como: anormalidades no sedimento urinário,

distúrbios eletrolíticos e de outros tipos devido a lesões tubulares; anormalidades estruturais detectadas por exame de imagem; e histórico de TR (118) (119) (138) (165) (166). Todas essas definições foram referendadas pela Sociedade Brasileira de Nefrologia (SBN) (15).

Já em 2013, a KDIGO ampliou as diretrizes para o diagnóstico da DRC (138): alterações na estrutura e na anatomia dos rins não são, necessariamente, indicativos irrefutáveis de lesão renal. Assim, a KDIGO determinou que com relação aos componentes de alterações anatômicas e estruturais houvesse uma ponderação no sentido de que elas devem, de fato, acarretar implicações para a saúde do indivíduo, sempre com o componente temporal atrelado (138).

Com a combinação dos dois principais marcadores, TFG e albuminúria, a estratificação de risco da DRC deve ocorrer por meio dos seguintes seis estágios: G1, G2, G3a, G3b, G4 e G5, como mostra a Figura 1. Portanto, quanto menor for o valor de TFG e quanto maior for o valor de albuminúria, mais avançado será o estágio clínico de um indivíduo.

Figura 1 – Relação entre os valores de TFG e os níveis de albuminúria para a estratificação dos estágios da DRC. As cores refletem os riscos de morte, morbidade e progressão, do melhor para o pior: verde (baixo risco para DRC), amarelo (risco moderadamente elevado), laranja (alto risco), vermelho (risco muito alto) e vermelho escuro (máximo risco).

				CATEGORIAS DE ALBUMINÚRIA		
				A1	A2	A3
				Normal a ligeiramente aumentado	Moderadamente aumentado	Severamente aumentado
ESTÁGIOS DA DRC DE ACORDO COM O VALOR DE TFG				< 30 mg/g < 3 mg/mmol	30-299 mg/g 3-29 mg/mmol	≥ 300 mg/g ≥ 30 mg/mmol
Estágios da TFG (ml/min/1.73 m ²)	G1	Normal ou elevado	≥ 90			
	G2	Levemente diminuída	60-90			
	G3a	Levemente a demoradamente diminuída	45-59			
	G3b	Moderadamente a severamente diminuída	30-44			
	G4	Severamente diminuída	15-29			
	G5	Insuficiência renal	< 15			

Fonte: Adaptado de (175).

As cores também podem servir de referência para o acompanhamento clínico anual de um paciente. Se a DRC estiver presente, a cor verde pode indicar estabilidade da doença. No caso da cor amarela, é indicado o acompanhamento anual do indivíduo associado a cuidados complementares. Pacientes identificados em estágios com a cor laranja requerem aproximadamente duas medições de TFG por ano. Já a cor vermelha indica pelo menos

um acompanhamento quadrimestral do paciente. Por fim, a cor vermelha escura sugere um quadro clínico grave do paciente e, pelo menos, acompanhamento trimestral (166) (174).

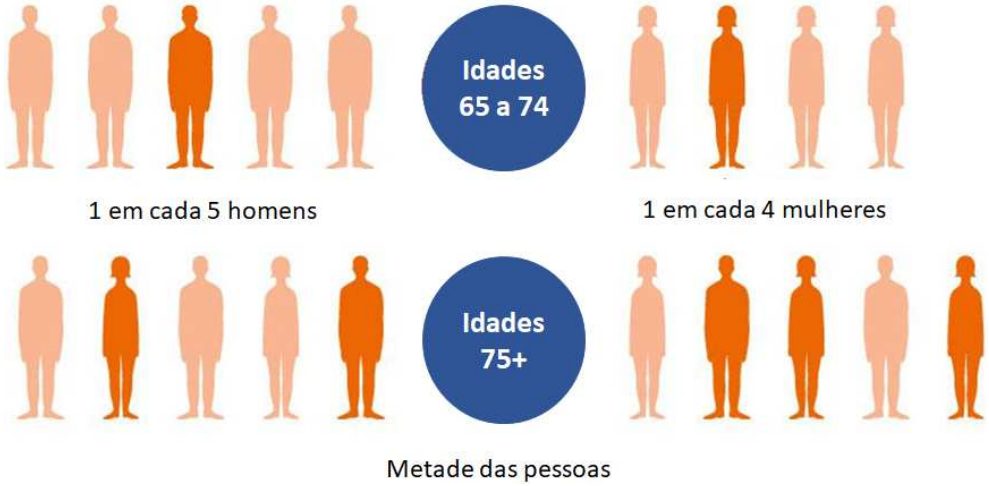
No quinto e último estágio da DRC, quando a TFG apresenta valor inferior a 15 mL/min/1,73 m² e o paciente apresenta um quadro de IRC, é recomendada a aplicação de alguma modalidade de TRS, como a hemodiálise, a diálise peritoneal ou o transplante renal (16) (118) (119).

2.3 Epidemiologia

Em todo o mundo, no ano de 2017 foram registrados 695,5 milhões de casos de DRC e 1,2 milhão de mortes causadas pela doença, colocando-a na décima segunda posição entre as principais causas de óbito em todo o mundo (29) (39). O Brasil ficou no grupo de países com mais de 10 milhões de casos contabilizados, totalizando quase 17 milhões de indivíduos com DRC (29). Entre 1990 e 2017, a mortalidade global por DRC aumentou 41,5% (39). E, de acordo com o estudo publicado por LV e colaboradores (151) em 2019, a prevalência da doença renal crônica na população mundial foi de 13,4%.

Como evidencia a Figura 2, cerca de 1 a cada 5 homens e 1 a cada 4 mulheres, entre as idades de 65 a 74 anos, possui doença renal crônica no mundo todo. Já na faixa etária a partir de 75 anos, cerca de metade das pessoas sofrem de DRC.

Figura 2 – Prevalência global da DRC entre adultos a partir de 65 anos.



Fonte: Adaptado de (231).

O *Centers for Disease Control and Prevention* (CDC) dos EUA mostrou que, no ano de 2021, mais de 15% da população adulta deste país, aproximadamente 37 milhões de pessoas, foram considerados como possíveis doentes renais crônicos. Desse total, 9 em cada 10 estadunidenses não possui conhecimento do diagnóstico. Ademais, 2 em cada 5 adultos não sabem que possuem quadros severos da DRC. Nessa mesma população, a

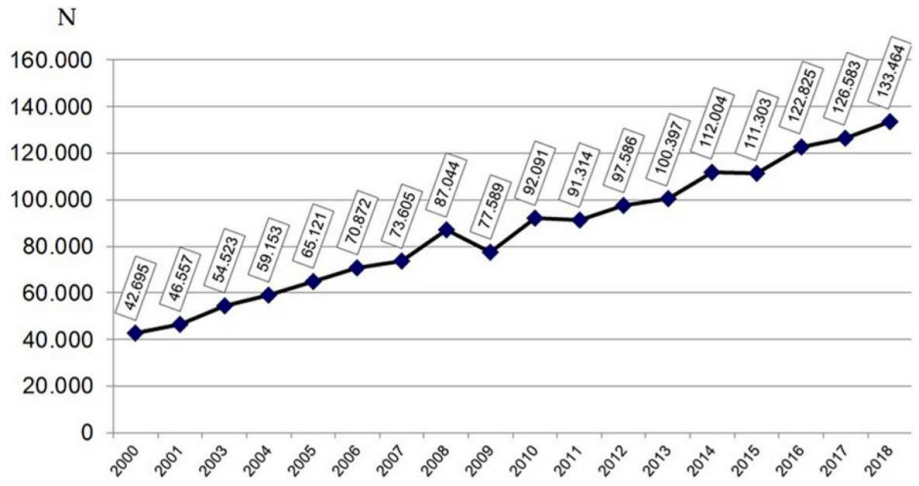
maioria dos doentes, 38%, possui mais de 65 anos de idade. A prevalência é maior entre as mulheres e entre os adultos negros não-hispânicos (41).

No Brasil e em todo o mundo, há uma tendência de crescimento do número de pacientes em diálise e das taxas de prevalência (45) (254). No ano de 2010, o número aproximado de pacientes em diálise no planeta era de 2 milhões e esse total deve ser duplicado até 2030 (45). Dados de 2019 (151) indicam que pacientes com IRC em estágios finais da doença e submetidos a algum tipo de TRS são estimados entre 4,902 milhões a 7,083 milhões em todo o mundo.

Segundo a Sociedade Latino Americana de Nefrologia e Hipertensão (SLANH) (230), em 2018 a taxa média de prevalência de pacientes em TRS era de 805 partes por milhão (ppm) de pessoas na América Latina. Os países com as maiores taxas foram Porto Rico (2,129 ppm), Chile (1,541 ppm) e México (1,405 ppm). Já nos EUA, em 2017 a taxa de prevalência era de 2,203 ppm (254). No Brasil, a taxa estimada de pacientes em TRS é de 876 (178).

A SBN realiza anualmente o Censo Brasileiro de Diálise (CBD). A pesquisa nacional *online* tem por objetivo avaliar as dimensões epidemiológicas e clínicas de pacientes em diálise crônica. No ano de 2014, cerca de 112 mil brasileiros passaram por algum tipo de tratamento dialítico (216). Segundo os dados do CBD da década de 2009 a 2018, estruturado por Neves e colaboradores (178) e disponibilizados pela SBN, houve um crescimento considerável do total de centros clínicos ativos que disponibilizam o tratamento de diálise crônica. Em 2009, eram 594 centros, ao passo que eram 786 em 2018. Neste mesmo período, também houve um aumento médio de 5.587 pacientes por ano que passaram a se submeter à diálise crônica, como mostra a Figura 3.

Figura 3 – Estimativa anual de pacientes em diálise crônica.



Fonte: Extraído de (178).

A última edição divulgada do CBD corresponde ao ano de 2022, sendo que 28%

($n = 243$) dos centros de diálise cadastrados na SBN responderam à pesquisa. No mês de julho, 153.831 pacientes eram estimados para a realização da diálise. A prevalência e a incidência de pacientes por milhão de habitantes foram de 758 e 214, nesta ordem. Levando em consideração essa prevalência, 95,3% dos pacientes realizaram hemodiálise, 4,6% hemodiafiltração e 4,7% diálise peritoneal. A pesquisa concluiu que o número e as taxas de prevalência de pacientes submetidos à diálise crônica estão aumentando, enquanto a taxa de mortalidade diminuiu. Essa redução foi justificada pelo fim da pandemia de COVID-19. Vale ressaltar que a pesquisa é voluntária, e a baixa adesão pode levar a limitações metodológicas quanto às estimativas apresentadas. (177).

A fim de mapear o perfil epidemiológico da insuficiência renal do Brasil, entre os anos de 2012 e 2021, realizou-se um estudo baseado nos dados fornecidos pelo departamento de informação de Saúde do SUS (Sistema Único de Saúde). Entre os resultados obtidos, é possível averiguar um total de 1.185.600 internações, das quais 152.053 sucederam a óbito. Com relação à amostra, a parcela mais acometida pela doença foi formada por homens (56,9%), da raça branca (35,2%), entre 59 e 69 anos de idade (41,3%). A região do país com maior número internações, óbitos e gastos hospitalares foi o Sudeste. A prevalência elevada nesta região pode ser justificada pelo maior acesso aos serviços de saúde, o que facilita a disponibilidade de tratamento e, como consequência, aumenta o número de notificações da doença (75).

Ainda no contexto, o Atlas Global de Saúde Renal da Sociedade Internacional de Nefrologia tem como objetivo mapear a disponibilidade de tratamento para essa doença em todo o mundo (149). Dados sobre a prevalência da DRC estão presentes em 73,9% dos países, apresentando uma prevalência média global de 9,5%. Nesse contexto, destacam-se as prevalências em diferentes regiões, com a Europa Central e Oriental apresentando a taxa mais elevada (12,8%) e a África, a mais baixa (4,2%). Os países com as maiores prevalências são Japão (20,2%), Estônia (16,8%) e Porto Rico (16,8%), ao passo que Chade (3,2%), Somália (3,0%) e Uganda (3,0%) apresentam os menores índices. A incidência da doença tende a ser mais alta em países com maior rendimento. A distribuição apresenta-se da seguinte forma: países de rendimento baixo (3,6%), países de rendimento médio-baixo (7,5%), países de rendimento médio-alto (10,7%) e países de alta renda (11,1%) (22).

2.4 Atendimento especializado em Juiz de Fora e região

Para coordenar a estruturação da rede de atenção à saúde e o monitoramento de DCNTs como a hipertensão arterial sistêmica, a diabetes *mellitus*, doenças cardiovasculares e a doença renal crônica, a Secretaria de Estado de Saúde de Minas Gerais (SES/MG) criou em 2010 o Programa Hiperdia Minas (223). Com a criação de um sistema regionalizado e integrado de ações em saúde, o Hiperdia Minas determinou a implementação dos chamados Centros Hiperdia Minas (CHDM), para que pacientes com condições crônicas de maior

complexidade pudessem ter atenção especializada e interdisciplinar no nível secundário de atenção à saúde (245).

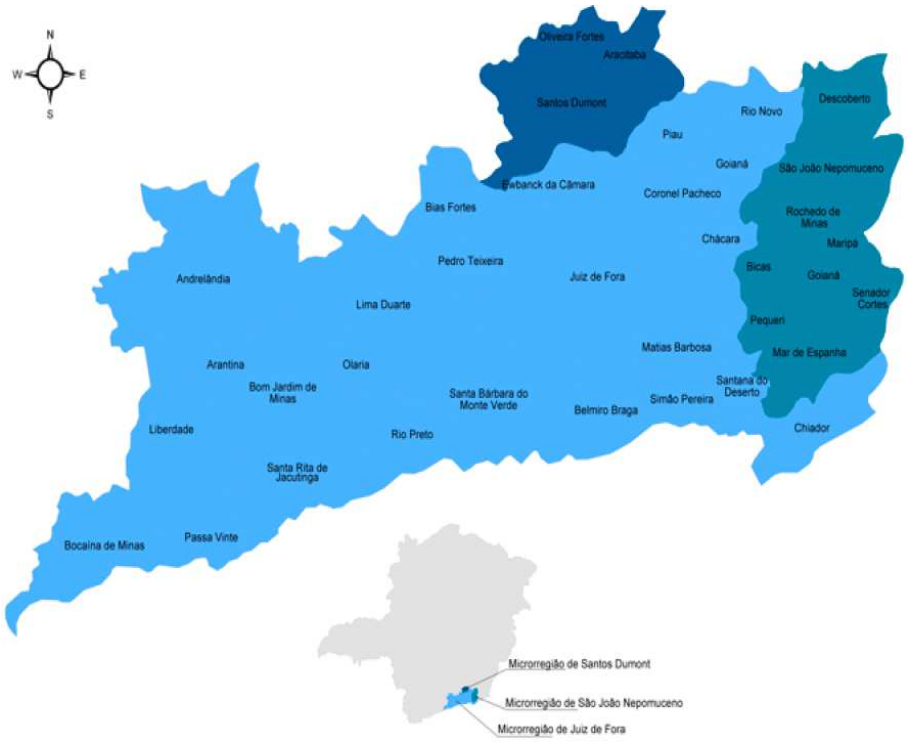
Fundada em 1986 por professores da Faculdade de Medicina da Universidade Federal de Juiz de Fora (UFJF), a Fundação Instituto Mineiro de Estudos e Pesquisas em Nefrologia (IMEPEN) foi a responsável, em 2010, pela implementação do Centro Hiperdia da cidade de Juiz de Fora (CH), na região da Zona da Mata do estado de Minas Gerais. A criação do CH diversificou e ampliou a atuação do Programa Multiprofissional de Prevenção das Doenças Renais, o PREVENRIM, criado em 2002 (118).

A atuação do CH abrangeu três microrregiões determinadas pelo Instituto Brasileiro de Geografia e Estatística (IBGE):

- Juiz de Fora, com 25 municípios;
- Santos Dumont, com 3 municípios;
- e São João Nepomuceno, com 9 municípios.

No total, a população atendida nesses 37 municípios, mostrados pela Figura 4, chegou a mais de 837 mil pessoas, total que representa pouco mais de 4% da população do estado (118) (119).

Figura 4 – Mapa com os municípios atendidos pelo Centro Hiperdia de Juiz de Fora.



Fonte: Extraído de (118).

Seguindo a proposta original de implementação dos CHDM, o CH oferece atenção ambulatorial compartilhada e interdisciplinar com foco não somente em atendimento médico, mas também em atendimento psicológico, nutricional, sociológico, fisioterápico, entre outros. O atendimento inicial a um paciente deve ser realizado em sua cidade de origem, onde profissionais de saúde da atenção primária decidem se é necessário o encaminhamento do paciente ao CH, seguindo os critérios estabelecidos pelas Linhas Guias de Atenção à Saúde do Adulto da SES/MG com relação a informações clínicas relativas à DM, à HAS e à DRC (118) (119).

E para o registro completo do atendimento a um usuário em todas as suas fases, incluindo o registro dos prontuários médicos, foi criado um sistema de registro eletrônico de pacientes para o CH. A validação dos dados do registro eletrônico e a caracterização do perfil dos pacientes a partir de indicadores clínicos foram realizados por Huaira em 2017 (118) (119). Todos os dados coletados foram autorizados pelo Comitê de Ética da UFJF e foram aprovados sob o protocolo de número 36345514.1.0000.5139.

Sob nova gestão estadual, em 2015 a SES/MG realizou modificações no Programa Hiperdia Minas, que passou a ser denominado Centro Estadual de Assistência Especializada (CEAE) (245).

3 TRABALHOS RELACIONADOS

A revisão dos trabalhos relacionados ao tema desta tese foi dividida em dois tópicos principais: a utilização da base de dados de pacientes atendidos pelo CH e a aplicação de métodos de inteligência artificial para a predição dos estágios da DRC a partir de conjuntos de dados pessoais e clínicos de um grupo de indivíduos. Esses dois tópicos constituem a base para a construção, implementação e obtenção dos cenários que serão propostos.

3.1 Base de dados do Centro Hiperdia de Juiz de Fora

Os três trabalhos descritos a seguir são oriundos do Programa de Pós-graduação em Saúde da UFJF e utilizaram os registros de pacientes atendidos majoritariamente entre os anos de 2011 e 2014 no CH. Cada estudo adotou uma abordagem específica aos dados, ampliando a compreensão sobre os mesmos e viabilizando o desenvolvimento de pesquisas futuras complementares.

A dissertação de mestrado de Huaira (2017) (118) e a publicação baseada em seus resultados (2018) (119) objetivaram a validação dos registros eletrônicos criados para o acompanhamento do histórico dos pacientes atendidos pelo CH. A validação foi essencial para que os dados armazenados pudessem ser utilizados na qualificação da gestão do CH como um todo e para possibilitar o seu uso em outros trabalhos de pesquisa, uma vez que a consistência das informações presentes em uma base de dados é fundamental.

Os outros objetivos foram construir o perfil demográfico dos usuários atendidos e utilizar indicadores clínicos de qualidade para caracterizar os pacientes em relação à HAS, à DM e à DRC. Os mais de 63 mil registros de atendimentos, que vieram de 7.266 pacientes presentes na base de dados, foram coletados entre agosto de 2010 e dezembro de 2014. A eles foi aplicada uma série de filtros para a seleção dos dados considerados clinicamente relevantes, excluindo, assim, erros no ato de registro, informações incompletas, pacientes que foram encaminhados apenas uma vez ao CH, pacientes com menos de 18 anos de idade e pacientes sem consulta ao ambulatório de DRC. Assim, para a caracterização do perfil dos pacientes doentes renais crônicos atendidos pelo CH, foram considerados os registros de 1.997 usuários.

Os resultados apontaram que 51,4% dos pacientes eram homens, com idade acima dos 64 anos e 81,6% estavam acima do peso. O tempo médio de acompanhamento dos pacientes foi de 21 meses e 33,7% apresentaram queda nos valores de TFG ao longo do tempo. Em geral, o grupo de pacientes analisados é idoso, obeso, usuário de diferentes tipos de medicamentos, com baixa renda e escolaridade, sendo, portanto, uma população que carece de atenção médica e social. Resultados relacionados a medicamentos e a outros dados clínicos também foram apresentados. A autora concluiu que a validação dos registros e a elaboração do perfil do paciente com DRC são resultados que contribuem para

melhorias na gestão do CH, principalmente devido à vulnerabilidade social e econômica da maioria dos usuários, apontada pelos resultados obtidos. Por fim, os estudos contribuíram, principalmente, na estruturação dos dados de forma que pudessem ser utilizados em outros trabalhos de pesquisa.

Dentro do contexto de pacientes diagnosticados com HAS, DM ou DRC, a tese de doutorado de Tirapani (2018) (245) e a respectiva publicação que a originou (2015) (246) preconizaram a avaliação da prevalência de fatores de riscos socioeconômicos nessa população, através de informações como renda, cor e educação, e a sua correlação com indicadores clínicos de controle. Com o uso da base de dados do CH, a autora estabeleceu dois critérios de seleção de dados: usuários com mais de 18 anos de idade e com registro de pelo menos duas consultas no CH. E para cada um dos ambulatórios - HAS, DM e DRC - foram utilizados critérios clínicos específicos para a seleção de pacientes, além da avaliação de variáveis demográficas, clínicas e relacionadas a medicamentos. Assim, a amostra resultante utilizada na tese foi composta pelos dados de 6.369 usuários.

Após a etapa de análise de dados, as conclusões obtidas com o estudo foram de que informações sobre a cor, a renda e a educação dos pacientes demonstraram pequena associação na progressão da HAS, da DM e da DRC. A única exceção foi o impacto gerado pela renda na progressão da diabetes *mellitus*, fato que pode ser explicado pelo acesso limitado apenas às medicações disponibilizadas pelo SUS, das quais a população de menor renda pode usufruir. Por fim, Tirapani (2018) (245) apontou que a universalidade e a interdisciplinaridade do SUS como um todo podem ter contribuído de maneira eficaz para mitigar a atuação dos fatores de riscos socioeconômicos sobre a saúde da população considerada.

Uma abordagem econômica dos custos para o orçamento público do tratamento em pré-diálise (PD) e da terapia dialítica (TD) foi proposta pela tese de doutorado de Moraes Júnior (2019) (166), que gerou o artigo (165), publicado em 2021. A hipótese apresentada pelo autor considerou que o monitoramento pré-diálise atua como redutor de custos se comparado à TD ambulatorial. Consequentemente, foi proposto um estudo financeiro junto a prestadores de serviço de PD do SUS. O estudo foi focado no atendimento a pacientes no CH e propôs duas abordagens de custos, uma *top-down* e uma *bottom-up*. Ambas foram utilizadas para o apontamento da evolução dos custos relativos aos seis estágios da DRC — já que os valores gastos com tratamento tendem a aumentar com a progressão da doença — e para o aumento dos custos com o encaminhamento dos pacientes à TD.

A partir de um total de registros de 5.689 pacientes atendidos pelos ambulatórios de HAS, de DM e DRC, foram aplicados critérios para a seleção da amostra populacional a ser estudada. Foram excluídos todos os pacientes atendidos antes de 2010 e depois de 2014, e doentes renais crônicos nos estágios G1 a G4 sem registro na base de dados entre 2011 e 2014. O autor assumiu que foram encaminhados para a terapia dialítica todos os pacientes

no estágio G5 que não mais seguiram tratamento no CH, uma vez que informações sobre os desfechos clínicos dos pacientes não estavam disponíveis na base de dados. A amostra resultante foi composta pelo registro de 537 pacientes, sobre os quais foram considerados e estudados, além de outros fatores, dados sociodemográficos e informações sobre os estágios da DRC em que se encontrava a população estudada e as probabilidades de transição entre tais estágios. Os resultados apontaram que tratamentos de PD podem reduzir os custos do orçamento público em R$ 33.023,12 (± R$ 1.676,80), em média, por paciente para cada ano evitado em terapia dialítica. Ademais, os resultados indicam que ações preventivas de detecção precoce e de cuidado a pacientes doentes renais crônicos podem reduzir o montante de gastos futuros com o atendimento em terapia dialítica.

3.2 Aplicação de métodos de inteligência artificial

A utilização de métodos de inteligência artificial (IA) para o auxílio ao diagnóstico de diferentes doenças tem sido cada vez mais constante em trabalhos da literatura relacionada. Especificamente, algoritmos de AM estão sendo progressivamente utilizados em trabalhos interdisciplinares para análises de dados biomédicos e clínicos de alta dimensionalidade (82). Assim, processos automatizados são desenvolvidos para os mais diversos fins, como a detecção precoce de doenças, predição de desfecho clínico, avaliação do uso de medicamentos, reconhecimento de padrões clínicos, entre outros.

3.2.1 Base de dados da UCI

A literatura relacionada à aplicação de métodos de IA para a predição dos estágios da DRC fundamenta-se, em grande parte, em dados provenientes do repositório *online* da *University of California, Irvine* (UCI) denominado *Chronic Kidney Disease Data Set* (210). Essa base de dados contém um total de 400 pacientes, sendo 250 doentes renais crônicos e 150 sem o diagnóstico de DRC. São, ao todo, 25 variáveis, das quais 24 são atributos multivariados, entre numéricos e categóricos, como creatinina sérica, idade, pressão sanguínea, gravidade específica, albumina, potássio, hemoglobina, sódio e nível de açúcar, entre outros. E o atributo alvo é `class`, que identifica se um paciente possui ou não o diagnóstico de DRC sendo, portanto, um problema de classificação binária. Os dados foram doados em 2015 e são provenientes da rede de hospitais indiana *Apollo Hospitals*, da cidade de Karaikudi, localizada no estado de Tâmil Nadu, no sul da Índia.

Trabalhos como Polat *et al.* (2017) (191), Tekale *et al.* (2018) (243), Almasoud e Ward (2019) (10), Almansour *et al.* (2019) (9), Rady e Anwar (2019) (201), Khan *et al.* (2020) (136), Ogunleye e Wang (2020) (180), Reshma *et al.* (2020) (204), Qin *et al.* (2020) (199), Ganie *et al.* (2023) (90), Islam *et al.* (2023) (126) e Halder *et al.* (2024) (105) são alguns exemplos de pesquisas que utilizaram, com sucesso, a base de dados da UCI (210) para aplicação de diferentes algoritmos de classificação na predição do risco da

DRC. Críticas importantes sobre os dados do repositório da UCI (210) foram feitas por Wang *et al.* (2021) (259) e estão descritas no fim desta seção.

Polat *et al.* (2017) (191) utilizaram todos os 24 atributos do conjunto de dados da UCI (210). Os dados faltantes foram inferidos através da média dos dados presentes. Para reduzir a dimensionalidade do conjunto de dados e melhorar a eficácia da classificação, foram aplicados métodos de seleção de atributos do tipo empacotador (*wrapper*) e do tipo filtro (*filter*) através do software WEKA (versão 3.6.13). Para o primeiro, foram utilizados os métodos *ClassifierSubsetEval*, com busca passo-a-passo gulosa (*greedy stepwise search*), e *WrapperSubsetEval*, com busca pelo melhor primeiro (*best first*). Para o segundo, foram utilizados os métodos *CfsSubsetEval*, com busca passo-a-passo gulosa e o *FilterSubsetEval* com busca pelo melhor primeiro. A validação cruzada *k-fold*, com 10 subconjuntos, foi escolhida para a separação dos conjuntos de teste e treinamento, e como algoritmo de classificação foi implementado o SVM. Os resultados do estudo mostraram que a combinação do SVM com métodos de seleção de características melhorou significativamente a acurácia do diagnóstico da DRC. A aplicação do método *FilterSubsetEval*, com busca pelo melhor primeiro, reduziu a dimensão do conjunto de dados para 13 atributos e alcançou a maior taxa de acurácia, 98,5%. Outros métodos como *ClassifierSubsetEval*, com busca passo-a-passo gulosa, e *WrapperSubsetEval*, com busca pelo melhor primeiro, também mostraram uma melhoria na acurácia em comparação com a abordagem sem seleção de atributos, com taxas de acurácia de 98% e 98,25%, respectivamente. Sem a seleção de características, o SVM teve uma taxa de acurácia de 97,75%. Portanto, a redução da dimensionalidade dos dados e a seleção de atributos possuem a capacidade de proporcionar melhorias significativas no desempenho dos algoritmos de classificação, reduzindo a complexidade computacional e minimizando o risco de sobreajuste (*overfitting*).

No estudo de Tekale *et al.* (2018) (243), os autores utilizaram um total de 14 dos 25 atributos disponíveis, e a média foi utilizada como método de inferência de dados faltantes. Para a predição, os autores implementaram os algoritmos DT e SVM, com os quais obtiveram acurácias de 91,75% e 96,75%, respectivamente.

O trabalho de Almasoud e Ward (2019) (10) aplicou à base de dados da UCI (210), quatro algoritmos de AM: regressão logística (LogR, do inglês *logistic regression*), *gradient boosting* (GB), SVM e RF, que foram treinados e testados por meio de uma validação cruzada *k-fold* com 10 subconjuntos. O tratamento dos dados faltantes foi realizado a partir de múltiplas inferências baseadas nos algoritmos de regressão linear (LR, do inglês *linear regression*). Como resultados, foram gerados cinco diferentes conjuntos de dados. Para a remoção de redundância entre os atributos dos dados utilizados e a diminuição de dimensionalidade, alguns testes estatísticos foram aplicados, como o teste ANOVA, a correlação de Pearson e o teste V-quadrado de Cramer. O conjunto de dados resultante foi composto por apenas três atributos: albumina, hemoglobina e gravidade específica. A maior acurácia entre os resultados foi de 99%, o maior valor de F1-*score* foi de 99,1% e a

maior precisão foi de 99,5%, todos valores resultantes da aplicação do classificador GB. Os resultados também evidenciaram que, além da creatinina, a hemoglobina exerceu um papel fundamental na classificação do risco da DRC para os algoritmos GB e RF.

Com a aplicação comparativa dos algoritmos de SVM e de redes neuronais artificiais (ANN, do inglês *artificial neural networks*) na base da UCI (210), Almansour *et al.* (2019) (9) consideraram todos os 24 atributos disponíveis. A média foi o método escolhido para a inferência dos dados faltantes. Para a separação dos dados, foi utilizada a validação cruzada *k-fold* com 10 subconjuntos. Toda as três definições anteriores também foram realizadas por Polat *et al.* (2017) (191). Foram aplicadas etapas de otimização de parâmetros, tanto para o SVM (tipo de *kernel* e custo) quanto para redes (número de camadas ocultas e taxa de aprendizado). A acurácia dos resultados foi de 99,75% para as ANNs (maior tempo de execução) e de 97,75% para SVM (menor tempo de execução), e para ambos os algoritmos, a curva ROC AUC (*area under the receiver operating characteristic curve*) ficou muito próxima de 1. Os autores concluíram que, embora ambas as técnicas de AM sejam eficazes para a predição de DRC, as ANNs foram capazes de oferecer acurácia ligeiramente superior em comparação com o SVM. No entanto, a escolha entre os dois modelos pode depender das necessidades específicas de tempo de execução e complexidade dos dados.

Algoritmos de mineração de dados foram utilizados por Rady e Anwar (2019) (201) para a predição dos estágios da DRC: *probabilistic neuronal networks* (PNN), *radial basis function* (RBF), o *perceptron* multicamadas (MLP, do inglês *multilayer perceptron*) e o SVM. Foram utilizadas 361 das 400 amostras disponíveis no repositório da UCI (210) e todas as 25 variáveis. O valor estimado da TFG foi obtido para cada paciente por meio de uma versão alternativa da Equação 2.2. E os dados faltantes foram preenchidos com a média dos valores. Diferentemente de outros trabalhos que objetivaram unicamente a predição do risco da DRC, Rady e Anwar (2019) (201) expandiram essa abordagem para cinco dos seis estágios da DRC, já que G3a e G3b foram considerados como sendo estágio G3. Os quatro algoritmos foram implementados para pacientes de cada um dos estágios e os resultados foram comparados. O algoritmo PNN apresentou o melhor desempenho na predição das severidades de todos os cinco estágios, com valores de acurácia variando de 96,9 a 99,7%; de precisão, variando de 89 a 100%; e de F1-*score*, variando de 93,6% a 99,37%. Os resultados de Rady e Anwar (2019) (201) mostraram que as variáveis preditoras mais importantes na construção do modelo de classificação foram, respectivamente: creatinina sérica, ureia, albumina, idade, hemoglobina e hipertensão.

Já Khan *et al.* (2020) (136) utilizaram todas as 25 variáveis da base e também trataram os dados ausentes com os valores de média. Foram aplicados sete algoritmos de classificação, entre eles *composite hypercube on iterated random projection* (CHIRP), *naive bayes tree* (NBTree), SVM, MLP, LR, DT e J48. Os resultados de acurácia variaram de 95,75% a 99,75%.

Seis algoritmos de classificação foram implementados e comparados por Ogunleye e Wang (2020) (180): *linear discriminant analysis* (LDA), *k*-vizinhos mais próximos (KNN, do inglês *k-nearest neighbors*), *classification and regression tree* (CART), LR, SVM e XGBoost. A aplicação desses métodos na base de dados da UCI (210) apresentou os melhores resultados para o XGBoost, com valores de acurácia, F1-*score*, sensibilidade e ROC AUC iguais a 0,987±0,016, 0,97±0,06, 0,98±0,08 e 1,00±0,00, respectivamente. Em seguida, o modelo de aplicação do XGBoost foi refinado por meio de um processo de otimização de parâmetros realizado a partir de três técnicas: eliminação recursiva de atributos (RFE, do inglês *recursive feature elimination*), *extra trees classifier* (ETC) e *univariate selection* (US). Os resultados apontaram valores de acurácia, sensibilidade e especificidade próximos a 1,0. Outro modelo foi criado a partir da redução de cerca da metade dos parâmetros da base dados e os resultados obtidos foram os mesmos do modelo aplicado na base completa. Os autores concluíram que a utilização desses modelos, aliada à experiência de especialistas em nefrologia, pode ser eficaz na redução de tempo e de custo do diagnóstico da DRC em um paciente. Ademais, os autores indicaram que a extração de características a partir de imagens geradas por alguns testes de detecção da DRC e a implementação de algoritmos de AM devem ser considerados como propostas de trabalhos futuros.

O valor aproximado de 96% foi obtido pela aplicação do método SVM associado a algoritmos de colônia de formigas no trabalho de Reshma *et al.* (2020) (204). Foram utilizados apenas 12 atributos da base de dados e a média novamente foi utilizada para inferir dados faltantes.

Os algoritmos de classificação implementados por Qin *et al.* (2020) (199) foram: *feedforward neuronal network* (FNN), KNN, LR, *naive bayes* (NB), RF e SVM para a predição do risco da DRC. Em cada implementação foi utilizada uma quantidade específica de atributos da base de dados, que variou de 8 a 15 variáveis. Diferentes implementações do algoritmo KNN foram aplicadas para o tratamento dos dados ausentes, com k sendo igual a 3, 5, 7, 9 e 11. Também foram utilizadas a média e a moda dos valores. Dependendo da combinação entre todos os fatores citados, os resultados obtidos apresentaram acurácia variando de 88% a 99%.

O estudo proposto por Ganie *et al.* (2023) (90) aplicou cinco algoritmos de *boosting*: *adaptive boosting* (AdaBoost), *categorical boosting* (CatBoost), *gradient-boosting machine* (LightGBM), GB e XGBoost para a detecção precoce da DRC. Os valores ausentes foram inferidos com a média e a mediana para as variáveis numéricas e categóricas, respectivamente, e os dados foram normalizados para um intervalo de [0,1]. Para o desbalanceamento dos dados, foi utilizado o método *synthetic minority over-sampling technique* (SMOTE) e por meio da seleção de atributos, a hemoglobina, a creatinina e a glicose no sangue foram identificadas como as variáveis de maior relevância. Os melhores resultados foram obtidos com o AdaBoost, 98,47% de acurácia, 98,50% de precisão e de

revocação, e 98,60% para a ROC AUC.

Por meio da análise de componentes principais (PCA, do inglês *principal component analysis*) e da RFE, Islam e colaboradores (2023) (126) reduziram os 25 atributos da base de dados da UCI (210) para 30% do total. Para a inferência de dados, as variáveis numéricas ausentes foram preenchidas pela média dos valores observados; as variáveis categóricas ausentes foram preenchidas pela moda; e as variáveis faltantes, relativas a dados fisiológicos, foram preenchidas por meio do algoritmo KNN, já que indivíduos com condições físicas semelhantes tendem a possuir características fisiológicas comparáveis. Para garantir que os modelos criados não estivessem sobreajustados aos dados de treinamento, foi utilizada a validação cruzada *k-fold* com 10 subconjuntos. Doze algoritmos de AM foram implementados, entre eles podem ser citados AdaBoost, DT, GB, KNN, RF e XGBoost. Este último apresentou maior acurácia, 0,98. O PCA ajudou a identificar as variáveis mais importantes, ao passo que a RFE refinou ainda mais o conjunto de variáveis, removendo as menos relevantes e contribuindo significativamente para a eficiência dos modelos. No entanto, destaca-se a importância de dados de alta qualidade e técnicas de pré-processamento robustas. Os autores sugeriram que futuras pesquisas devem focar na integração de dados clínicos adicionais e na validação dos modelos em cenários clínicos reais para aumentar a aplicabilidade prática das soluções propostas.

O último dos trabalhos supracitados, que fizeram uso do conjunto de dados da UCI (210), foi o de Halder *et al.* (2024) (105), que desenvolveu uma aplicação *web* baseada em AM para facilitar e popularizar o diagnóstico precoce da DRC. Na etapa de preparação dos dados, os autores realizaram a inferência de dados faltantes por meio da média, aplicaram o método de normalização Min-Max, além de diversas técnicas de seleção de atributos. Sete classificadores foram utilizados: AdaBoost, DT, GB, RF, SVM, NB e XGBoost, sendo todos avaliados pela acurácia, pela matriz de confusão e pela ROC AUC. A validação cruzada *k-fold* com 10 subconjuntos também foi utilizada. Os resultados mostraram que AdaBoost e RF alcançaram taxas de acurácia próximas a 100% e AUC perfeita, enquanto NB se destacou pela eficiência em tempo de execução. A aplicação *web* desenvolvida operacionalizou o modelo, melhorando a acessibilidade para profissionais de saúde. Apesar das forças do estudo, como a robustez do modelo e a inovação da aplicação *web*, limitações como a ausência de análise de complexidade temporal e escalabilidade foram destacadas. Os autores concluíram que a combinação de técnicas avançadas e classificadores robustos pode melhorar a predição de DRC, contribuindo para a melhoria do cuidado ao paciente e para a redução de custos associados ao tratamento tardio.

3.2.2 Demais trabalhos

O estudo de Chan *et al.* (2020) (42) realizou a pesquisa dos termos "*machine learning*" e "*kidney*" ("rim", em inglês) na plataforma para pesquisa de publicações

científicas na área de saúde, PubMed (253), restringindo a busca apenas a estudos com humanos. Foram encontrados 207 resultados, sendo que metade deles foi publicada entre 2018 e 2019. Uma grande parcela dessas publicações é limitada por generalizações em suas abordagens e por falta de perspectiva na validação dos dados resultantes com especialistas da área de saúde, o que dificulta a manutenção de processos de melhoria contínua e que gerem maior confiabilidade em seus resultados. Segundo os autores, os principais temas abordados pelas publicações encontradas foram a predição do diagnóstico de doenças renais, o reconhecimento de padrões por meio de imagens e biópsias, o processamento de linguagem natural e a identificação de subtipos em doenças complexas. Os autores concluíram que há uma necessidade urgente para o desenvolvimento de parcerias e investimento financeiro para fomentar estudos que foquem na melhoria do atendimento, do diagnóstico e do tratamento dos pacientes doentes renais crônicos.

Na contramão de todos os trabalhos citados na subseção anterior, Wang *et al.* (2021) (259) apresentaram críticas à disseminação do uso da base de dados da UCI (210). A primeira delas disserta que, embora a acurácia média dos trabalhos que utilizam essa base seja de 97%, os dados apresentam um viés com relação à proporção entre doentes renais crônicos e pacientes sem o diagnóstico de DRC, que é de 250 para 150, respectivamente. Essa razão é diferente da proporção comumente encontrada na realidade clínica, em que há uma elevada predominância de pacientes que não possuem DRC. A segunda crítica fundamenta que, entre os pacientes com DRC, há 140 (56% do total) cujos valores de creatinina sérica excedem 10 mL/min, valor que é considerado bastante elevado e que, por isso, pode ser inadequado para a predição do risco da DRC. Outra crítica levantada pelos autores foi direcionada não somente ao repositório da UCI (210), mas também a todos os trabalhos que consideram a creatinina como variável preditora, já que seu valor está diretamente relacionado aos valores de TFG (como evidenciam as equações 2.1, 2.2 e 2.3) e, consequentemente, ao estágio da DRC em que se encontra o paciente. Wang *et al.* (2021) (259) fizeram o uso de uma base com 1 milhão de amostras coletadas no ano de 2017, proveniente do *National Health Insurance Data Sharing Service*, da Coreia do Sul, e composta por dados de indivíduos com idades entre 25 e 90 anos. No total, a base de dados possui 24 atributos, incluindo a creatinina sérica. Mais três atributos foram adicionados pelos autores: o valor de TFG, o estágio da DRC e uma variável binária para a classe do paciente (risco da DRC), indicando se ele possui ou não DRC (foram considerados doentes renais crônicos todos os pacientes classificados em estágios superiores ao G2). A metodologia proposta foi dividida em duas abordagens. Na primeira, foi desenvolvido um modelo de regressão que não considerou a creatinina como variável preditora, mas sim como o alvo das 23 variáveis restantes. Para a aplicação da segunda abordagem, como os dados de creatinina são extremamente desbalanceados, os autores aplicaram um método de *undersampling* e propuseram uma nova função de erro quadrático médio sensível ao custo, dividindo os dados em alguns subconjuntos. Para potencializar os resultados da

predição da creatinina, foram utilizados dois métodos de AM, o RF e o XGBoost, e uma rede neuronal residual para a regressão linear. Dessa forma, foi construído um método de *ensemble learning* (EL) a partir da elaboração de oito preditores distintos. Por fim, para a finalização da segunda abordagem, os valores preditos de creatinina foram combinados aos 23 atributos originais para a predição das classes dos pacientes. A partir da seleção da melhor estratégia de *undersampling*, o melhor resultado do EL utilizando o coeficiente de determinação R^2 foi de 0,5590. Com a utilização do valor predito de creatinina o modelo obteve uma ROC AUC no valor de 0,76 na predição das classes de DRC. O algoritmo RF foi utilizado para avaliar a importância de cada variável na classificação. Os seis atributos que mais influenciaram o valor de creatinina foram: sexo, idade, hemoglobina, níveis de proteína na urina, circunferência abdominal e o tabagismo. Por fim, os autores concluíram que o impacto da DRC na saúde pública pode ser reduzido com a detecção precoce da doença mesmo em situações em que o valor de creatinina dos pacientes não está disponível.

O estudo brasileiro de Silveira *et al.* (2021) (228) utilizou registros médicos oriundos de um hospital alagoano para a avaliação do risco da DRC. Os dados são relacionados a 60 indivíduos brasileiros, com ou sem diagnóstico da doença, incluindo variáveis como hipertensão, diabetes *mellitus*, creatinina, ureia, albuminúria, idade, sexo e TFG. A base de dados possui quatro classes de acordo com risco da DRC: baixo, médio, alto e muito alto. Duas abordagens foram aplicadas para o aumento da base: aumento manual dos dados, validado por um nefrologista, e aumento automático usando o SMOTE e duas de suas variações. Modelos de classificação foram implementados com algoritmos como DT, RF e AdaBoost multiclasse. Técnicas como a validação cruzada estratificada (*stratified cross-validation*) e a validação cruzada aninhada (*nested cross-validation*) foram empregadas. Os melhores resultados foram obtidos com os modelos de DT e AdaBoost multiclasse quando aplicados o SMOTE e as validações, com acurácias de 98,99% e 98%, nesta ordem para os algoritmos.

Debal e Sitote (2022) (65) publicaram um estudo baseado na ideia de inteligência artificial explicável (XAI, do inglês *explainable artificial intelligence*), o que permite aos humanos compreenderem e confiarem nas decisões ou previsões feitas por sistemas de inteligência artificial. O foco da XAI é criar modelos de IA que não apenas tenham bom desempenho, mas também sejam transparentes em suas operações e decisões, o que é particularmente importante em aplicações críticas, como saúde, na qual decisões de IA podem ter consequências significativas (2). Os autores aplicaram os algoritmos DT, RF, SVM e XGBoost na predição da DRC em diferentes estágios, utilizando dados de 1.718 pacientes do Hospital St. Paulo, em Adis Abeba, Etiópia. Entre as 19 variáveis consideradas há idade, gênero, pressão arterial, creatinina sérica, nitrogênio ureico, hemoglobina, entre outras. A etapa de preparação dos dados envolveu a remoção de valores atípicos (*outliers*), por meio da aplicação do intervalo interquartil (Q1 - 1,5 e Q3 + 1,5), e a suavização de dados ruidosos. Ademais, a inferência de dados faltantes foi realizada pela média, e a

técnica z-*score*, de normalização, foi aplicada para garantir que todas as características numéricas estivessem na mesma escala, transformando os valores com base na média e no desvio padrão. Variáveis categóricas foram convertidas para valores numéricos binários. Dois métodos de seleção de atributos foram aplicados: RFE com validação cruzada e seleção de recursos univariados (UFS, do inglês *univariate feature selection*). Os quatro algoritmos de AM foram treinados e testados utilizando validação cruzada *k-fold* com 10 subconjuntos. O estudo também utilizou técnicas de *oversampling* para corrigir a distribuição desequilibrada das classes. Os resultados indicaram que a combinação do modelo RF com a RFE com validação cruzada apresentou o melhor desempenho para a classificação binária, com uma acurácia de 99,8%, ao passo que o XGBoost teve a melhor performance para a classificação multiclasses com uma acurácia de 82,56%. Os métodos de seleção de características mostraram que creatinina sérica, nitrogênio ureico, hemoglobina e gravidade específica foram as variáveis com maior influência na predição.

O trabalho publicado por Ghosh e Khandoker (2024) (96) utilizou uma base de dados dos Emirados Árabes Unidos com dados clínicos, bioquímicos e demográficos de 491 pacientes, dos quais 56 com DRC e 435 sem a doença. Para a inferência dos dados ausentes, foi utilizada a média dos valores presentes. Cinco algoritmos de AM foram empregados para a predição da DRC: DT, LogR, NB, RF e XGBoost, sendo este último o de melhor desempenho com ROC AUC de 0,9689 e acurácia de 93,29%. Para avaliar a influência das variáveis nos modelos, os métodos SHAP (*shapley additive explanations*) e *local interpretable model-agnostic explanations* (LIME) foram utilizados. As análises com o SHAP e o LIME forneceram uma visualização detalhada da influência de cada variável nas predições dos modelos, aumentando a transparência e a confiabilidade dos modelos obtidos. Como resultado, ambos os métodos indicaram que a creatinina, a hemoglobina glicada e a idade foram as três variáveis mais influentes na predição do risco da DRC pelo modelo XGBoost.

Delrue e colaboradores (2024) (68) realizaram um balanço acerca do uso de algoritmos de AM na avaliação da DRC. Entre as limitações citadas pelos autores estão o sobreajuste (*overfitting*), a necessidade de dados de alta qualidade e a complexidade dos modelos que podem dificultar a interpretação pelos clínicos. A falta de padronização e a necessidade de validação externa são obstáculos importantes a serem superados. Além da discussão sobre os algoritmos e as métricas de avaliação mais utilizados em estudos recentes, os autores apresentaram perspectivas futuras de pesquisas sobre o tema. O futuro da união entre AM e DRC envolve a expansão e a integração de fontes de dados heterogêneas, incluindo registros eletrônicos de saúde, dados genômicos e de imagem, possibilitando uma visão mais completa da DRC, além de diagnósticos e tratamentos mais personalizados e precisos. Modelos mais sofisticados, como redes neurais profundas, têm o potencial de identificar padrões complexos nos dados, automatizando a quantificação de lesões renais e melhorando a eficiência dos diagnósticos patológicos. Ademais, a personali-

zação do tratamento é um dos principais objetivos da aplicação de AM em DRC. Modelos preditivos que consideram a variabilidade individual nos dados clínicos e genômicos podem desenvolver estratégias terapêuticas personalizadas, incluindo a otimização de doses de medicamentos e intervenções preventivas específicas para cada paciente. A implementação de sistemas de apoio à decisão clínica baseados em AM pode auxiliar os médicos na tomada de decisões mais informadas e precisas, fornecendo previsões em tempo real sobre a progressão da DRC e sugerindo intervenções terapêuticas. No entanto, a validação e a padronização dos modelos, além da gestão ética dos dados e a transparência nos algoritmos, são essenciais para garantir a robustez e a aceitação dessas tecnologias na prática clínica. A integração eficaz do AM depende de uma colaboração contínua entre clínicos, pesquisadores e profissionais que trabalham com dados — como também sugeriram Schena *et al.* (2022) (217) — além da implementação de políticas robustas de compartilhamento de dados. A validação contínua e o aprimoramento dos modelos são essenciais para garantir sua eficácia e aceitação na prática médica.

3.2.3 Revisões sistemáticas e análise bibliométrica

A revisão sistemática realizada por Sanmarchi e colaboradores (2023) (215) abordou a aplicação de métodos de AM na prevenção, no diagnóstico e no tratamento da DRC por meio da abordagem PRISMA (*preferred reporting items for systematic reviews and meta-analyses*) (197) e da inclusão de publicações em língua inglesa disponíveis no PubMed (253) até 20 de outubro de 2021. Um resumo dos principais itens avaliados pelos autores está especificado na Tabela 1.

Tabela 1 – Resumo das principais características avaliadas pela revisão sistemática proposta por Sanmarchi e colaboradores (2023) (215).

Item	**Valor mais comum**	**Total**
Objetivo	Predição	41%
Origem	Ásia	48,5%
População	Indivíduos com DRC e indivíduos saudáveis	26
Base de dados	Registros hospitalares	48,5%
Variáveis (geral)	PA, idade, hemoglobina, creatinina sérica e sexo	25,34%
Variáveis (prognóstico)	Idade, hemoglobina e proteinúria	-
Algoritmos de AM	DT, RF e XGBoost	33,53%
Avaliação	Acurácia e ROC AUC	34,10%

Para cada estudo, foram extraídas 16 variáveis, entre elas: objetivo principal, população estudada, fonte de dados, tamanho da amostra, tipo de problema, preditores usados e métricas de desempenho. Após etapas de análises prévias, 68 artigos foram selecionados, dos quais 48,5% são relativos a estudos oriundos da Ásia, 25% da Europa, 17,6% da América do Norte, 7,35% da África e 1,47% da América do Sul. É importante ressaltar que o único estudo sul-americano mencionado, Guo *et al.* (2020) (103), na

realidade, se refere a uma pesquisa conduzida utilizando uma base de dados de três hospitais situados na China, composta por pacientes de perfil étnico chinês, de acordo com os próprios autores. A maioria dos estudos teve foco na predição do prognóstico da DRC (progressão e mortalidade). Em geral, a população mais comum entre todos os trabalhos era composta por 26 indivíduos, divididos em doentes renais crônicos e em pacientes saudáveis. Os dados foram obtidos majoritariamente por meio de registros hospitalares contendo dados socioeconômicos, clínicos e laboratoriais de indivíduos. Com relação à aplicação de algoritmos de AM, o total variou de 1 a 10 modelos por trabalho, sendo os baseados em árvores os mais frequentes — DT, RF e XGBoost — seguidos por redes neuronais artificiais — com destaque para o MLP — e pelo SVM e a LogR. Entre as métricas de avaliação dos resultados, a mais utilizada foi a acurácia, seguida pela ROC AUC, pela sensibilidade, especificidade, precisão e F1-*score* para a classificação. Nos estudos de prognóstico da DRC, os valores da ROC AUC ficaram entre 0,69 e 0,96; os valores de acurácia entre 0,54 a 0,99; sensibilidade, de 0,54 a 0,93; e a especificidade, de 0,78 a 0,99. De todos os trabalhos considerados, 83,8% envolveram classificação e 16,2% aplicaram técnicas de regressão. O total de variáveis consideradas nos estudos ficou entre o mínimo de 4 e o máximo de 6.624, sendo 813 variáveis distintas entre si ao todo. As mais comuns entre todos os estudos foram pressão arterial (PA), idade, hemoglobina, creatinina sérica e sexo, nesta ordem. E as variáveis mais frequentes (em ordem decrescente) de acordo com o objetivo de cada trabalho foram:

- Prognóstico da DRC: PA, idade, colesterol sérico, sexo, potássio sérico, hemoglobina, sódio sérico, cálcio sérico, índice de massa corporal (IMC) e albumina sérica;
- Diagnóstico da DRC: PA, hemoglobina, infecção urinária, glicose sérica, idade, creatinina sérica, glóbulos vermelhos, diabetes *mellitus*, gravidade específica da urina e histórico de doença cardiovascular;
- Risco de desenvolver DRC: PA, idade, sexo, TFG, histórico de doença cardiovascular, glóbulos vermelhos, colesterol sérico, diabetes *mellitus*, creatinina sérica e albumina sérica;
- Tratamento da DRC: ferro sérico, hemoglobina, glóbulos brancos, volume globular médio, medicamentos utilizados, proteína C reativa, altura, creatinina sérica, cálcio sérico e albumina sérica.

A generalização dos algoritmos e os testes em populações diversas raramente foram considerados. Questões de viés e equidade dos modelos também não foram abordadas com frequência, levando à obtenção de resultados potencialmente injustos. Situações como essa podem não ser adequadas na representação de populações diversas, exacerbando a disparidades em certos grupos demográficos e impactando negativamente populações sub-representadas. Somente 12 dos 68 estudos incluíram variável referente a raça ou a etnias,

dos quais somente 10 também incluíram o gênero. E quando a população considerada envolveu pessoas de fenótipo não branco e não hispânico, elas formavam a maioria dos indivíduos alvos dos estudos. Outro problema identificado foi a falta de transparência na seleção de variáveis e na explicabilidade dos modelos. A revisão também destacou a necessidade de se abordar questões de interpretabilidade, generalização e equidade dos modelos para garantir a aplicação segura dessas tecnologias na prática clínica. A maioria dos estudos não considerou a implementação clínica dos modelos, e poucos avaliaram a integração desses na prática médica diária. Por fim, Sanmarchi e colaboradores (2023) (215) apontaram que o AM possui importante potencial para melhorar o diagnóstico, prognóstico e tratamento da DRC, embora sejam necessárias mais pesquisas para abordar a interpretabilidade, a generalização e a equidade dos modelos antes que possam ser aplicados de forma segura e eficaz na prática médica diária. A validação clínica rigorosa e a adoção de diretrizes de boas práticas de elaboração de relatórios são essenciais para o avanço na implementação dessas tecnologias. Portanto, futuras investigações devem focar na integração desses modelos na prática clínica, considerando as implicações éticas e sociais.

Também por meio da metodologia PRISMA, Khalid e colaboradores (2024) (135) propuseram uma revisão sistemática através da pesquisa por trabalhos disponíveis nas bases de dados Embase (78), PubMed (253), Scopus (220) e Web of Science (55), utilizando termos relacionados ao AM, à IA e à DRC. Foram encontrados 137 estudos, dos quais 13 atenderam aos critérios de inclusão após a triagem de títulos, de resumos e do texto completo. Os critérios de inclusão foram artigos revisados por pares, estudos observacionais e ensaios clínicos publicados nos últimos cinco anos, que aplicaram AM na predição da progressão da DRC. Os estudos selecionados foram publicados entre 2019 e 2023 e são oriundos de seis países diferentes: China, EUA, Egito, Israel, Japão e Países Baixos. As populações de pacientes abordadas variaram de 500 a 550.000 indivíduos, incluindo tanto doentes renais crônicos quanto indivíduos saudáveis, e dados representando informações demográficas, medições clínicas, resultados laboratoriais e registros longitudinais. Entre os modelos de AM mais implementados estão o RF (utilizado devido à sua robustez e capacidade de lidar com grandes conjuntos de dados), o SVM (devido à eficácia em classificações binárias), a LogR (utilizada em vários estudos por sua simplicidade e capacidade de interpretação) e as ANNs (promissoras na modelagem de relações complexas e não lineares). Alguns dos outros algoritmos de AM utilizados foram DT, GB, KNN, NB e XGBoost. Vários dos modelos analisados apresentaram desempenho promissor na predição da progressão da DRC, com vários estudos relatando elevados valores de acurácia, de sensibilidade, de especificidade e da curva ROC AUC. Diversos estudos ressaltaram a relevância da incorporação de dados longitudinais, características de base e biomarcadores específicos ou atributos clínicos para aprimorar a acurácia das previsões. Como resultado da revisão sistemática, os autores apontaram que os algoritmos de AM possuem a habilidade

de identificar padrões e relações complexas de alta dimensionalidade, que podem não ser detectáveis por métodos estatísticos convencionais. Essa característica é especialmente útil no contexto da DRC, uma vez que a doença sofre a influência de diversos fatores, como idade, comorbidades, predisposição genética e exposições ambientais. Os modelos tradicionais de predição de risco frequentemente enfrentam dificuldades para considerar essas interações complexas, resultando em um desempenho subótimo. Variáveis clínicas e pessoais como idade, sexo, comorbidades, TFG basal, proteinúria, albumina sérica, níveis de hemoglobina e marcadores inflamatórios são frequentemente apontadas como preditores significativos da progressão da DRC. Esse fator possibilita que os médicos concentrem seus esforços no monitoramento e gerenciamento desses fatores críticos, potencialmente retardando a progressão da doença e adiando a necessidade de terapia renal substitutiva. Por fim, os autores também apontaram ser fundamental o reconhecimento das limitações e dos desafios na aplicação de técnicas de AM em ambientes clínicos. Outros fatores importantes são a questão acerca da qualidade dos dados, de possíveis vieses e de questões éticas relacionadas à privacidade. Além disso, a transparência dos dados precisa ser cuidadosamente abordada. São necessários estudos de validação externa e prospectivos para avaliar a generalização e o desempenho desses modelos preditivos no mundo real, em diferentes populações de pacientes e sistemas de saúde.

A análise bibliométrica desenvolvida por Wu e colaboradores (2024) (270) objetivou a compreensão das tendências de pesquisas em IA aplicadas à doença renal. Foram coletados artigos publicados na língua inglesa entre 2010 e 2023 a partir da base de dados Web of Science (55). A estratégia de busca foi construída com base em termos encontrados em artigos previamente publicados e em consultas com especialistas em estudos bibliométricos além de serem direcionados à doença renal e ao uso de modelos de IA. A análise incluiu a identificação de tendências de publicação, produtividade dos países, instituições, autores, e a análise de ocorrência de palavras-chave para determinar os temas de pesquisa predominantes. Ao final da aplicação dos critérios de seleção, restaram 631 trabalhos — sendo 172 do ano de 2022 e 135 de 2023 — publicados em 324 diferentes revistas científicas. Os resultados indicaram um crescimento exponencial nas publicações anuais sobre a aplicação de IA na doença renal, com um aumento significativo após 2015. As principais contribuições, 60,54%, vieram de três países: EUA, em primeiro lugar, seguidos pela China e pela Índia. Com relação aos temas de pesquisa mais frequentes, os principais foram “aprendizado de máquina”, “aprendizado profundo”, “inteligência artificial”, “florestas aleatórias” e “redes neuronais artificiais”. O trabalho com maior número de citações é o de Almansour *et al.* (2019) (9) — já comentado neste Capítulo — com um total de 104, seguido pelo estudo de Polat *et al.* (2017) (191), também já descrito, com 88 citações. A análise de Wu e colaboradores (2024) (270) trouxe uma visão detalhada do panorama atual das pesquisas sobre IA e a doença renal, evidenciando as principais tendências, possíveis contribuições e temas emergentes na área, além de permitir que

pesquisadores e formuladores de políticas públicas identifiquem áreas de grande impacto e inovação. No tocante ao futuro das pesquisas, podem ser citados sete temas principais da contribuição da IA para a doença renal e suas variantes:

- Detecção e diagnóstico precoces: reconhecer padrões e sinais precoces de doença renal e identificar biomarcadores sutis e fatores de risco que podem não ser imediatamente perceptíveis para os clínicos humanos;
- Tratamentos personalizados: assim como proposto por Sanmarchi *et al.* (2023) (215), é fundamental buscar a elaboração de planos de tratamentos personalizados com base em dados individuais do paciente, como genética, estilo de vida e resposta ao tratamento, antecipando a progressão da doença e ajustando as intervenções conforme o necessário;
- Otimização da gestão de medicamentos: aumentar a adesão à medicação através de lembretes, monitoramento de efeitos colaterais e ajuste de dosagens com base em dados em tempo real do paciente, identificando aqueles com maior risco de não seguirem o tratamento e permitindo intervenções proativas;
- Monitoramento remoto de pacientes: realizar o monitoramento contínuo dos sinais vitais e outros parâmetros de saúde importantes, possibilitando a supervisão remota dos pacientes;
- Análise preditiva para complicações: antecipar complicações relacionadas à doença renal, como lesão renal aguda, possibilitando intervenções precoces e ações preventivas;
- Alocação eficiente de recursos: aprimorar a alocação de recursos ao prever taxas de admissão de pacientes, identificar populações de alto risco e distribuir os recursos conforme necessário.

4 REFERENCIAL TEÓRICO

4.1 Inteligência Artificial

A inteligência artificial é um ramo da ciência da computação e da engenharia que propõe, desenvolve e implementa métodos computacionais destinados à simulação de sistemas caracterizados por sua complexidade e dinamismo, como a inteligência humana. Esses métodos baseiam-se em processos naturais, biológicos e sociais, distinguindo-se pela capacidade de aprendizado, adaptação e robustez. A IA pode ser especialmente eficaz em contextos nos quais os métodos computacionais tradicionais se mostram inadequados devido à interdisciplinaridade, à complexidade, à incerteza ou à variabilidade inerentes ao problema considerado (80).

As origens da IA remontam ao trabalho seminal de McCulloch e Pitts (161) publicado em 1943, cujo pioneirismo, segundo Piccinini *et al.* (2004) (189), se destacou pela utilização da lógica matemática e da computação na compreensão da atividade neuronal e, por consequência, da atividade mental. Entre as principais contribuições do trabalho, podem ser destacadas a criação de um formalismo, cujo aprimoramento e generalização resultaram na noção de autômatos finitos, conceito crucial na teoria da computabilidade; o desenvolvimento de uma técnica que inspirou a noção de design lógico, componente essencial do projeto de computadores modernos; a aplicação pioneira da computação para abordar o problema mente-corpo; e a formulação da primeira teoria computacional moderna da mente e do cérebro (189).

Já na década de 1950, Alan Turing (249) estabeleceu o "Teste de Turing" como preceito fundamental para a avaliação da inteligência de máquinas computacionais. Desde então, as pesquisas sobre IA avançaram e fomentaram a criação de novas áreas de estudo e a formalização de áreas já existentes à época, como a inteligência artificial (IA). Em 1958 e tendo como inspiração o trabalho de McCulloch e Pitts, Frank Rosenblatt (207) formalizou a ideia de *perceptron*, uma unidade básica de processamento que poderia reconhecer padrões e aprender com base em exemplos. Assim, o autor estabeleceu as bases fundamentais do conceito de redes neuronais artificiais para simular a capacidade do cérebro de aprender, reconhecer padrões e tomar decisões.

No ano seguinte, Arthur Samuel (214) sugeriu o termo aprendizado de máquina (*machine learning*) ao propor um programa para o jogo de damas, que possivelmente está na história como o primeiro programa de autoaprendizagem do mundo e um dos principais trabalhos já produzidos na área de IA (266). Citando a publicação de Samuel (214): *dois procedimentos de aprendizado de máquina foram investigados com algum detalhe usando o jogo de damas. Trabalho suficiente foi feito para verificar o fato de que um computador pode ser programado de modo que aprenda a jogar damas melhor do que aquele que pode ser jogado pela pessoa que escreveu o programa. Além disso, ele pode aprender a fazer*

isso em um período de tempo notavelmente curto (8 ou 10 horas de jogo de máquina) quando lhe são dadas apenas as regras do jogo, um senso de direção e uma lista redundante e incompleta de parâmetros que são pensados para ter algo a ver com o jogo, mas cujos sinais corretos e pesos relativos são desconhecidos e não especificados. Os princípios do aprendizado de máquina verificados por esses experimentos são, obviamente, aplicáveis a muitas outras situações. Os resultados obtidos por Samuel foram fundamentais para o desenvolvimento de novos tipos de algoritmos de AM, principalmente a partir da década de 1980 com o surgimento das redes neuronais artificiais multicamadas e dos algoritmos de retropropagação.

Em 1965, Lotfi Zadeh (271) introduziu o conceito de lógica *fuzzy*, que trouxe importante inovação no tratamento da incerteza e da imprecisão, possibilitando a tomada de decisão computacional em ambientes ambíguos. Inspirado na teoria da evolução darwiniana, John Holland (115) propôs, ainda na década 1960, os conceitos inicias que levaram ao desenvolvimento dos algoritmos genéticos e evolutivos. Ao longo das décadas seguintes, Holland solidificou e ampliou essas novas abordagens em outros diversos trabalhos que se tornaram referências (116).

As décadas de 1970 e de 1980 trouxeram o surgimento de novos conceitos como as árvores de decisão, métodos de aprendizado não supervisionado e redes bayesianas, bem como o aprimoramento de algoritmos como o KNN. Em 1986, Rumelhart *et al.* (212) propuseram uma análise experimental do algoritmo de retropropagação (*backpropagation*), possibilitando o treinamento de ANNs multicamadas. Nos anos 1990, em trabalhos como (28) e (57), Vapnik e colaboradores introduziram o conceito de máquinas de vetores de suporte, um dos principais métodos utilizados para classificação e regressão. Houve um importante fomento nas pesquisas sobre a lógica *fuzzy* e sobre agentes inteligentes, como o trabalho de Wooldrige e Jennins (269).

Desde então e com maior destaque sobretudo a partir dos anos 2000, a IA tem passado por um crescimento significativo e sem precedentes. Esse cenário pode ser explicado sobretudo pelo progressivo e vertiginoso aumento da capacidade computacional, aliado à crescente disponibilização e digitalização de grandes volumes de dados (*big data*) de diversos tipos e fontes, e à popularização mundial da *internet*.

Em 2001, Leo Breiman (33) propôs o método de floresta aleatória, um dos mais utilizados em problemas de classificação e regressão. Anos depois, Hinton e Salakhutdinov publicaram um trabalho (114) que fomentou a intensificação das pesquisas a respeito das redes neuronais artificiais de aprendizado profundo (*deep learning*) (140). Por serem multicamadas, essas redes permitem o treinamento de cada camada separadamente e, dessa forma, são capazes de trabalhar com grandes volumes de dados (44). O estudo de Goodfellow *et al.* (2014) (99) abordou as redes adversárias generativas (GANs, do inglês *generative adversarial networks*). Elas são formadas por duas camadas de ANNs,

uma geradora e uma discriminadora, que são processadas em competição. A primeira é responsável pela sintetização de dados a partir da alteração de uma parte do conjunto de dados considerado inicialmente. Já a segunda, tem a função de verificar o grau de veracidade dos dados gerados, diferenciados das informações reais. As GANs são amplamente utilizadas na geração de imagens e vídeos (6).

Em anos mais recentes, tem-se destacado pesquisas no campo de processamento de linguagem natural (NLP, do inglês *natural language processing*), sobretudo com foco na arquitetura *Transformer*, proposta por Vaswani *et al.* (2017) (258). A grande contribuição deste trabalho foi fundamentar a utilização do mecanismo de atenção (*attention mechanism*) no processamento de sequência de dados, possibilitando eficiência, profundidade e paralelismo nas etapas de processamento. A arquitetura *Transformer* fomentou o surgimento dos modelos de linguagem de larga escala (LLMs, do inglês *large language models*), capazes de capturar padrões complexos e nuances de linguagem em grandes volumes de texto. Neste contexto, foram propostos o BERT (*bidirectional encoder representations from transformers*) (69), e o transformador generativo pré-treinado (GPT, do inglês *generative pre-trained transformer*) (200), modelo que utiliza o mecanismo de atenção para potencializar o processamento de *big data*, focando em partes distintas de um texto de entrada, capturando as dependências de longo alcance em contextos complexos de NLP. Os GPTs tem revolucionado a área de inteligência artificial generativa em diversos segmentos como pesquisas e diagnósticos em saúde, automação de processos industriais, ensino e aprendizado em diferentes níveis educacionais, geração de conteúdo multiplataforma para jornalistas, artistas em geral, profissionais liberais, cientistas , entre outros. Desde o surgimento do primeiro GPT em 2018 (200), novos modelos foram desenvolvidos por diferentes empresas e institutos de pesquisa, acarretando em um avanço vertiginoso das LLMs e das IAs generativas, como um todo e para os mais variados propósitos. Atualmente, em 2024, a versão mais recente do GPT da OpenIA é a de número 5 (181) e outras ferramentas de IA generativa podem ser destacadas como a PaLM 2 (*pathways language model*) (12) e a Gemini (93), do Google, a Megatron-Turing NLG (232), da NVIDIA e Microsoft, entre outras.

4.2 Aprendizado de Máquina

O aprendizado de máquina pode ser conceituado como uma das principais áreas da inteligência artificial e suas pesquisas remontam a trabalhos como os supracitados de Turing (249), Rosenblatt (207), Samuel (214), entre outros.

O objetivo principal das investigações, que levaram ao surgimento da AM, emergiu do anseio em desenvolver sistemas capazes de adquirir conhecimento a partir de conjuntos de dados e aprimorar seu desempenho ao longo do tempo, sem a necessidade de programação explícita para cada tarefa específica. Dessa forma, o AM pode ser utilizado em uma

ampla variedade de aplicações como a visão computacional, *e.g.* segmentação de imagens, detecção de objetos e reconhecimento facial; processamento de linguagem natural; sistemas de recomendação, *e.g.* conteúdos de redes sociais e plataformas multimídias; saúde, *e.g.* em análises de imagens médicas, predição e auxílio a diagnósticos variados; análises financeiras, *e.g.* através da construção de modelos preditivos (163).

Um problema de AM pode ser formalmente definido pela divisão dos dados em dois subconjuntos: X, denominado preditor, que compreende as variáveis de entrada; e Y, conhecido como resposta, que engloba as variáveis de saída vinculadas às entradas em X (132). Seja f uma função que mapeia os dados do conjunto X ao Y. E seja ϵ o erro de aproximação de um modelo desenvolvido. Os diferentes algoritmos de AM procuram relacionar os subconjuntos X e Y por meio da Equação 4.1, para a qual a função f pode ser aproximada por meio da geração de inúmeras funções (132).

$$Y = f(X) + \epsilon \tag{4.1}$$

Um caso de exemplo pode ser a aplicação de AM em um problema relacionado à saúde pública. O subconjunto X pode representar os dados clínicos e pessoais de um indivíduo, ao passo que Y pode ser composto pelo desfecho clínico do paciente (72).

4.2.1 Tipos de Aprendizado de Máquina

Os algoritmos de AM podem ser treinados por meio de quatro abordagens distintas que são diretamente baseadas no tipo e na estruturação do conjunto de dados a ser considerado (11).

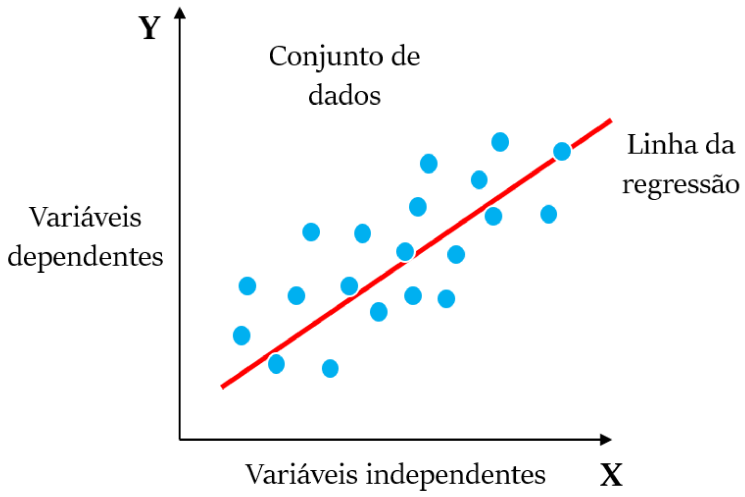
4.2.1.1 Aprendizado Supervisionado

Na abordagem do aprendizado supervisionado (AS), os dados de entrada estão rotulados de acordo com as saídas previamente conhecidas e associadas a eles. Logo, o treinamento aplicado ao algoritmo de AM é vinculado ao rótulo dos dados e o objetivo do processo é obter uma função capaz de mapear as entradas com a saídas desejadas ($X \rightarrow Y$). Dessa forma, o modelo a ser construído aprenderá com os resultados do treinamento e, consequentemente, será capaz de realizar previsões potencialmente corretas para novos dados (140).

Em termos matemáticos, o AS pode ser descrito como o problema de minimizar uma função de custo, que quantifica a discrepância entre as previsões do modelo e os valores reais. A minimização desta função de custo é realizada através de técnicas de otimização, como o gradiente descendente (132). Em termos de categorias de aplicação, o AS pode ser divido em duas principais: regressão e classificação.

Regressão é um conceito estatístico que se refere a um tipo de análise cujo objetivo é identificar e quantificar a relação entre uma variável dependente (também conhecida como variável resposta) e uma ou mais variáveis independentes (denominadas preditores) (11). O exemplo da Figura 5 evidencia a linha de regressão traçada em vermelho, a qual procura representar a maioria dos pontos do conjunto de dados (em azul). Este, por sua vez, representa a relação entre as variáveis X e Y.

Figura 5 – Exemplo do traçado da linha de uma regressão que representa a maioria dos pontos do conjunto de dados em azul.



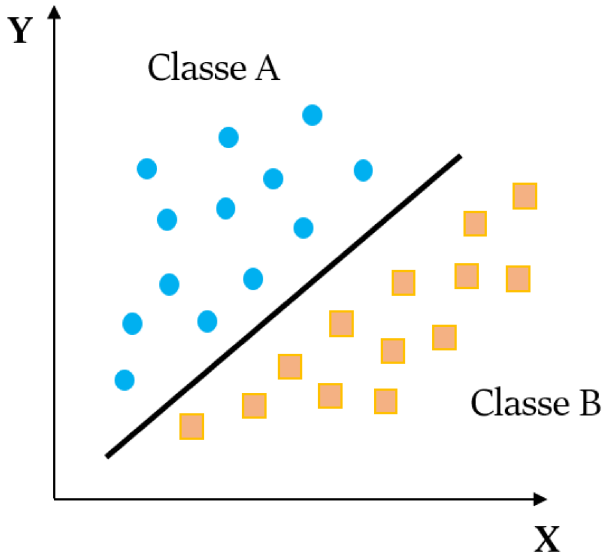
Fonte: Elaborada pelo autor.

A regressão é amplamente utilizada em diversos campos, como a análise de tendências de mercado, a previsão do tempo, a precificação de imóveis e a estimação de demandas futuras com base em dados históricos. A forma mais frequentemente empregada de regressão é a linear (11), que busca modelar a relação entre as variáveis através de uma equação linear, além da regressão logística e das árvores de regressão.

Já a classificação, é um método de AS cujo objetivo é predizer uma categoria ou rótulo discreto para uma dada entrada X, organizando os resultados obtidos em grupos predefinidos: os alvos (classes). A Figura 6 exemplifica a separação dos dados em duas classes distintas: A e B, que poderiam representar, por exemplo, pacientes portadores de uma doença e pacientes saudáveis (11).

O processo de classificação é iterativo e pode ser aplicado tanto em dados estruturados quanto em não estruturados. O objetivo central é determinar um padrão desconhecido a uma classe conhecida. Para isso, o modelo é treinado em um conjunto de dados rotulados, sendo cada instância de dados associada a uma classe específica. A partir desses dados, o modelo aprende a identificar padrões e a distinguir entre diferentes classes, permitindo que ele faça previsões potencialmente precisas para novas instâncias (11).

Figura 6 – Exemplo de classificação que separa o conjunto de dados em duas classes distintas: A e B.



Fonte: Elaborada pelo autor.

4.2.1.2 Aprendizado Não Supervisionado

Ao contrário do AS, o aprendizado não supervisionado (ANS) é desprovido da classificação dos dados de entrada. Uma vez sem os rótulos, cabe exclusivamente ao modelo desenvolvido quais são os dados que devem ou não ser considerados em um problema proposto. Para tal, grandes quantidades de dados são necessárias (172).

O principal objetivo do ANS é encontrar padrões desconhecidos no conjunto de dados. Essa vantagem é especialmente útil para problemas nos quais a estruturação dos dados não é previamente conhecida. Dessa forma, o próprio algoritmo pode ser capaz de identificar padrões inesperados entre os dados considerados (11).

Entre os algoritmos mais utilizados para o ANS estão o agrupamento (*clustering*), as regras de associação (*association analysis*), a detecção de anomalias (*anomaly detection*), a análise de componentes principais (*principal component analysis*, PCA) e os codificadores automáticos (*autoencoders*) (11) (172).

4.2.1.3 Aprendizado Semissupervisionado

O aprendizado semissupervisionado (ASS) é uma abordagem intermediária que combina os dois conceitos anteriores com o objetivo de sobrepujar as desvantagens de ambos (184). A sua utilização é adequada para problemas em que a proporção de dados não rotulados é sobremaneira superior ao total de dados rotulados, principalmente quando a obtenção de dados rotulados é complexa devido a limitações de recursos ou à própria dificuldade do processo de coleta dos dados (11).

Entre as principais aplicações do aprendizado semissupervisionado estão a análise de discursos (*speech analysis*), classificação de páginas web (*web content classification*), classificação de sequências de proteínas (*protein sequence classification*) e classificação de documentos (*document classification*) (11).

4.2.1.4 Aprendizado por Reforço

O aprendizado por reforço (AR) é um método de AM baseado na abordagem de tentativa e erro (129), que possibilita que agentes aprendam a tomada de decisões sequenciais em um ambiente dinâmico que objetiva a maximização da recompensa cumulativa e o descarte de respostas indesejadas. Esse tipo de aprendizado não requer pares de entrada/saída rotulados, uma vez que o aprendizado ocorre através da interação contínua do agente com o ambiente, sem a supervisão explícita de terceiros (11).

Embora o AR seja, em algumas situações, definido como uma subcategoria do ASS, sua originalidade é comumente destacada devido à sua abrangência e aplicabilidade (11). Entre as suas aplicações, podem ser citadas a teoria de jogos, pesquisa operacional, sistemas multiagentes, algoritmos genéticos e em jogos como o Go (227).

4.2.2 Algoritmos

Nesta Subseção, serão apresentados e definidos os conceitos fundamentais de diversos algoritmos de aprendizado de máquina.

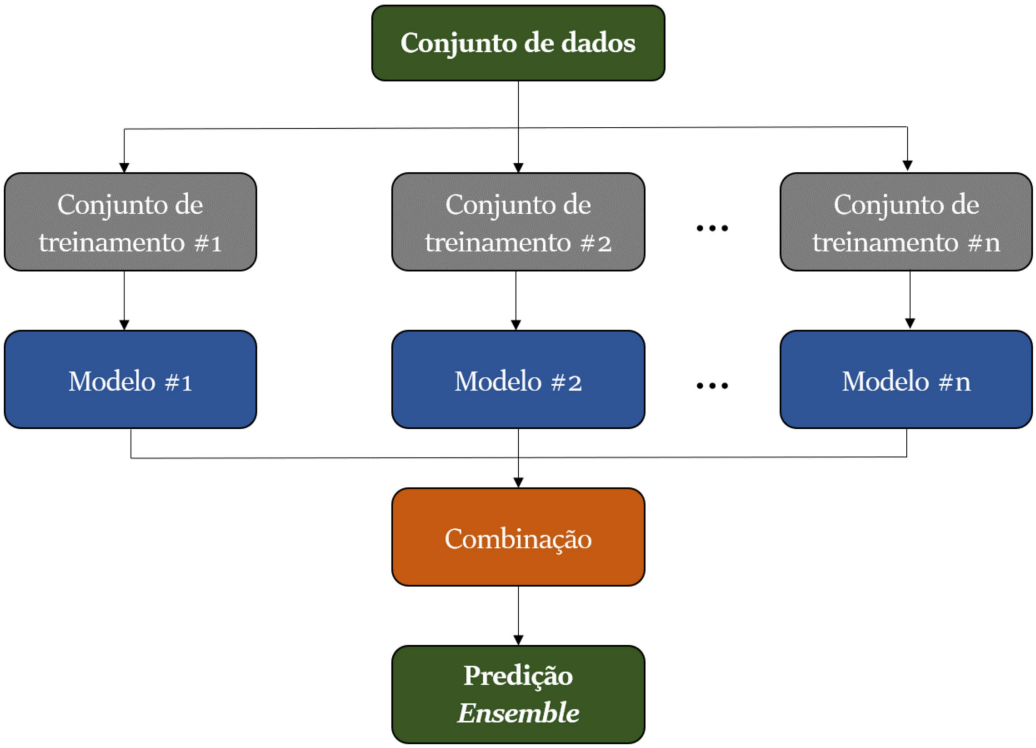
4.2.2.1 Algoritmos de *Ensemble*

A definição de *ensemble* constitui uma classe de métodos de aprendizado de máquina que combinam múltiplos modelos (definidos como modelos "base" ou "fracos") com o intuito de potencializar os seus resultados preditivos. Essa melhoria pode ser alcançada por meio da redução da variância das predições e, consequentemente, a melhoria da eficácia dos resultados. Logo, a combinação dos modelos tende a alcançar resultados preditivos mais robustos do que a aplicação isolada de cada um deles (192). A Figura 7 esquematiza a estrutura de um algoritmo de *ensemble* (AE) por meio da construção de um modelo preditivo "forte" de *ensemble* a partir do treinamento dos modelos fracos.

Uma das principais vantagens dos algoritmos de *ensemble* é que eles são menos suscetíveis ao sobreajuste (*overfitting*), que se manifesta nos casos em que um modelo apresenta previsões muito satisfatórias para os dados de treinamento, porém falha ao generalizar essas previsões para novos dados (164).

Dois dos principais tipos de AE são a agregação (*bagging*) e o *boosting*. O tipo de treinamento dos modelos preditores base é diferente para cada uma das duas abordagens: é simultâneo no *bagging* e é sequencial no *boosting* (182). Ambos os métodos podem ser utilizados para classificação e para a regressão.

Figura 7 – Esquematização da combinação de n modelos base para a formação do modelo preditivo de *ensemble*.



Fonte: Adaptado de (11) e (182).

No *bagging*, diversos subconjuntos da amostra total são gerados com o intuito de desenvolver modelos preditivos com maior variação, ajustando uma redistribuição estocástica dos conjuntos de treinamento (164). Já a abordagem sequencial do *boosting*, permite a disposição adaptativa de vários modelos preditivos fracos, *i.e.*, para cada novo modelo gerado há a tentativa de correção dos erros do modelo anterior. Logo, a cada nova iteração, a influência dos dados classificados incorretamente na etapa anterior é evidenciada, auxiliando o algoritmo nos ajustes necessários para a redução dos danos para as próximas iterações (182).

4.2.2.2 Árvores de Decisão

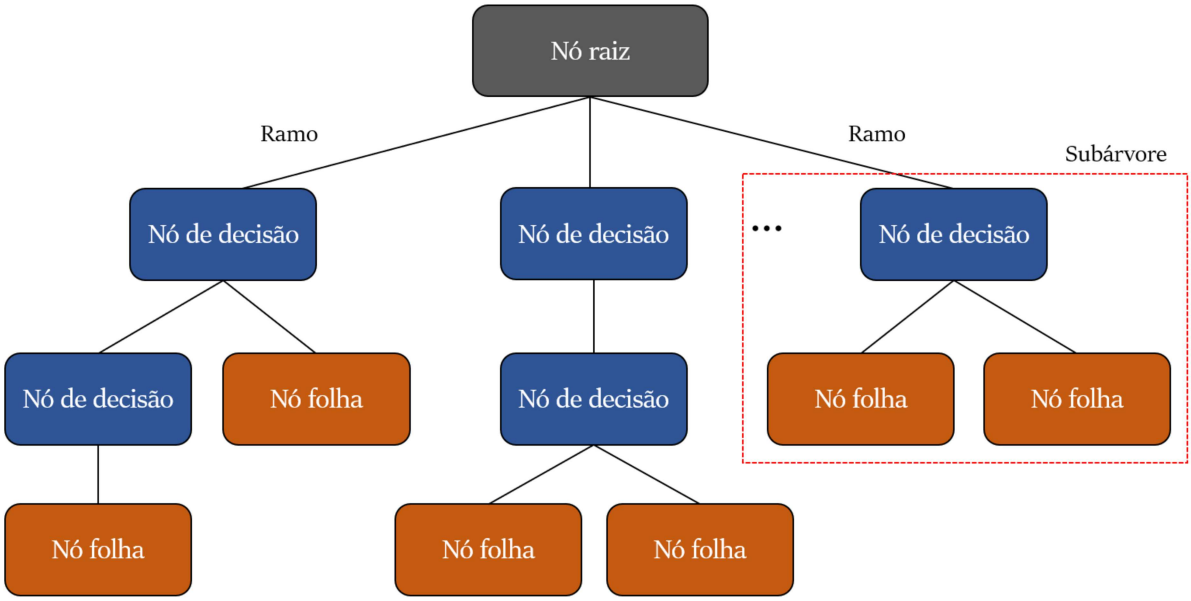
As árvores de decisão (DT, do inglês *decision tree*) são um algoritmo de AS cujo funcionamento é baseado na estrutura natural de uma simples árvore, com suas raízes, ramos, galhos e folhas. Do mesmo modo, a construção de uma DT tem início com o seu nó raiz, a partir do qual os demais são dispostos comumente na direção de cima para baixo e da esquerda para a direita. Os chamados "nós de decisão" são relacionados a condições que devem ser consideradas pelos algoritmos. Já os "folhas" — que são conectados por segmentos chamados de "ramos" — são os nós em que há o término de cada ramificação

(subárvore) (8).

A Figura 8 mostra um exemplo da estrutura de uma DT. A construção da árvore é fundamentada na divisão do conjunto de dados em partes menores determinadas pelas características que melhor o segmentem. Um nó representa uma característica específica do conjunto de dados considerado no algoritmo de DT, ao passo que os ramos correspondem a intervalos de valores, que funcionam como pontos de divisão para o conjunto de valores de uma característica particular. O processo é aplicado recursivamente a cada subconjunto particionado dos itens de dados e é concluído quando todos os itens de dados no subconjunto atual pertencem à mesma classe (8).

O nó raiz representa o melhor preditor para o conjunto de dados considerado. Os nós de decisão são constituídos pelos pontos de divisão nos quais o conjunto de dados é particionado com base em uma determinada característica e seguindo o método de divisão e conquista (205). Já os ramos se conectam aos nós e representam os resultados dos testes aplicados nos nós de decisão. Por fim, os nós folha representam a saída final do algoritmo, que pode ser uma classificação ou uma decisão final, a depender do método de AS considerado (8).

Figura 8 – Representação de uma árvore de decisão com seus diferentes tipos de nós, ramos e folhas.



Fonte: Adaptado de (11).

Entre as principais vantagens das DT estão a simplicidade e a interpretabilidade, uma vez que a construção de uma árvore é intuitiva, de fácil interpretação e pode ser aplicada a um vasto número de problemas. As DTs também são capaz de modelar relações

não lineares entre as variáveis e, por serem um algoritmo não paramétrico, não requerem muitas condições para serem implementados e baixo custo computacional (11). Ademais, as DTs podem trabalhar tanto com dados nominais quanto numéricos, além de poderem lidar com bases de dados com erros, dados faltantes (205) e valores atípicos (*outliers*) (235). Já entre as desvantagens, há o fato das DTs serem propensas à ocorrência de sobreajuste dos dados de treinamento. Nesses casos, o algoritmo apresenta dificuldade na generalização para novos dados e, assim, alta variância. Outra desvantagem ocorre quando a construção de um problema pode se tornar complexa ao ser estruturada como uma DT, sobretudo em casos que envolvam grande volume de dados. As DTs também podem apresentar instabilidade quando pequenas variações nos dados ocasionam a construção de árvores radicalmente distintas (11) (205).

4.2.2.3 Florestas Aleatórias

No trabalho supracitado publicado em 2001 (33), Breiman apresentou uma extensão do conceito de *bagging* para árvores de decisão. Surgiram assim as florestas aleatórias (RF, do inglês *random forest*), algoritmo de aprendizado supervisionado e *ensemble* que constrói com os dados de treinamento múltiplas DTs — "floresta" — e realiza a seleção estocástica de subconjuntos de atributos para a construção individual de cada árvore. Os principais objetivos do RF são o aprimoramento da eficiência e da robustez dos modelos preditivos, além da redução do risco de sobreajuste (33).

Um modelo preditivo de RF pode ser expresso pela Equação 4.2 (108), na qual y é a predição final, T é o total de árvores e $h_t(x)$ é a predição da árvore t para a entrada x.

$$y = \frac{1}{T}\sum_{t=1}^{T} h_t(x) \tag{4.2}$$

Durante a etapa de amostragem de dados, é selecionado, com reposição, um subconjunto aleatório de dados para cada árvore utilizando a técnica de *bagging*. Em seguida, para cada divisão de nó na árvore, é escolhido um subconjunto aleatório de características. Essa abordagem ajuda a garantir que as árvores não apresentem correlação, o que é essencial para o desempenho do modelo. Finalmente, na etapa de predição, cada árvore faz uma predição para um novo exemplo, e a estimativa final do modelo de RF é obtida através da votação (no caso de classificação) ou pela média (no caso de regressão) das previsões de todas as árvores (11) (108). Um exemplo do funcionamento das etapas do algoritmo de RF está descrito nos pseudocódigos 1 e 2.

Entre as principais vantagens das florestas aleatórias estão a sua capacidade de lidar com grandes volumes de dados, com valores atípicos e com ruídos, além de serem um algoritmo eficaz na redução da variância e do sobreajuste (11). Em um modelo RF, a proximidade par-a-par entre as amostras pode ser mensurada pelo conjunto de dados de

treinamento (196).

Algoritmo 1 Pseudocódigo do algoritmo RF. Adaptado de (102).

Requer: Conjunto de dados $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{n}$, número de árvores B, número de atributos m a serem considerados em cada nó
Garante: Floresta de árvores de decisão
1: **Para** $b = 1$ até B **faça**
2: $\mathcal{D}_b \leftarrow$ Amostragem com reposição de $\mathcal{D}$
3: $T_b \leftarrow$ ConstruirÁrvore($\mathcal{D}_b$, m)
4: **Fim Para**
5: **Retorne** $\{T_b\}_{b=1}^{B}$

Algoritmo 2 ConstruirÁrvore

Requer: Subconjunto de dados $\mathcal{D}_b$, número de atributos m
Garante: Árvore de decisão T
1: $T \leftarrow$ Nó raiz
2: **Enquanto** Critério de parada não for satisfeito **faça**
3: $\mathcal{F}_b \leftarrow$ Selecionar m atributos aleatórios de x
4: $A_{\text{best}} \leftarrow$ MelhorAtributo($\mathcal{D}_b$, $\mathcal{F}_b$)
5: Dividir nó usando A_{best}
6: **Fim Enquanto**
7: **Retorne** T

Com relação às desvantagens, o RF costuma necessitar de um tempo maior para o treinamento dos dados quando comparado a outros algoritmos. Ademais, o RF pode exigir um elevado custo computacional quando modelos mais complexos são construídos (11). E nos casos de conjunto de dados em que há variáveis categóricas com níveis distintos, o RF tende a apresentar viés em função de tais variáveis (196).

4.2.2.4 Gradient Boosting

Por meio de uma generalização do conceito de *boosting*, Jerome Friedman propôs em 1999 (88) um novo algoritmo de *ensemble* que utiliza gradientes para minimizar funções de perda. O *gradient boosting* (GB) é baseado na construção sequencial de modelos preditivos com o objetivo de que um novo modelo seja treinado de forma que os erros dos anteriores sejam minimizados.

Para atingir seu objetivo, o algoritmo desenvolve uma aproximação adaptativa da função $F(x)$, dada pela Equação 4.3, que relaciona cada variável de entrada, x, à sua respectiva saída resultante, y, por meio de uma soma ponderada de funções. A variável λ_m é o peso da m-ésima função, $h_m(x)$ (23).

$$F_m(x) = F_{m-1}(x) + \lambda_m h_m(x) \tag{4.3}$$

O processo iterativo do GB tem início com a definição de um modelo, $F_0(x)$, que é inicializado como uma aproximação constante, como mostra a Equação 4.4, na qual L é a função de perda, y_i são os valores reais que estão sendo considerados, N é o total de amostras do conjunto de treinamento e c é uma constante (23) (88).

$$F_0(x) = \arg\min_c \sum_{i=1}^{N} L(y_i, c) \tag{4.4}$$

As diferenças entre os valores reais e os preditos dos modelos a cada iteração são chamadas de resíduos. Estes são utilizados no ajuste do próximo modelo do processo sequencial e iterativo de aprendizado do GB, de forma que os erros residuais tenham propagação mínima. Na m-ésima iteração, os resíduos r_{im} são calculados por meio da Equação 4.5, na qual é calculado o gradiente da função de perda $L(y, F(x))$ em relação à predição atual $F_{m-1}(x)$. As demais variáveis são y_i, que é o valor do i-ésimo exemplo; $F(x_i)$, a predição do modelo para o i-ésimo exemplo; e $F_{m-1}(x_i)$, que é a predição após $m-1$ iterações (88) (173).

$$r_{im} = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right]_{F(x)=Fm-1(x)} \tag{4.5}$$

Em seguida, o novo modelo, $F_m(x)$, é ajustado pelos resíduos, como mostra a Equação 4.6, e a m-ésima predição realizada pelo GB é atualizada pela Equação 4.3 (23) (88).

$$F_m(x) = \arg\min_F \sum_{i=1}^{N} \left[r_{im} - F(x_i)\right]^2 \tag{4.6}$$

Por fim, como mostra a Equação 4.7, a predição final do algoritmo GB, dada por F_M, é constituída pela soma individual dos M modelos obtidos ao longo de todo o processo iterativo e sequencial. A variável $F(x)$ representa o m-ésimo modelo e λ é a taxa de aprendizado acumulada ao longo de todas as iterações (23) (88).

$$F_M(x) = \sum_{m=1}^{M} \lambda F_m(x) \tag{4.7}$$

Uma visão resumida de todas as etapas do algoritmo GB está descrita em pseudo-código no Algoritmo 3.

Como vantagens da utilização do *gradient boosting* podem ser citadas a sua capacidade de capturar dependências complexas e não lineares entre variáveis do conjunto de dados, permitindo que o algoritmo consiga resultados de acurácia superiores a outros métodos. Ademais, o GB é flexível e adaptável para diferentes requisitos práticos (173).

Já a principal desvantagem do algoritmo é o sobreajuste, sobretudo nos casos em que processo iterativo não é devidamente regularizado. Portanto, o GB pode não ser

capaz de ser generalizado para uma ampla gama de novos casos de entrada (23). Outro inconveniente é o elevado custo de memória, comumente ocasionado por aplicações que exigem um grande número de iterações do algoritmo (173).

Algoritmo 3 Pseudocódigo do algoritmo *gradient boosting*. Adaptado de (88).

1: **Entradas:** $\{(x_i, y_i)\}_{i=1}^{N}$: conjunto de treinamento; M: total de iterações, λ: taxa de aprendizado
2: **Inicialização:** $F_0(x) = \arg\min_c \sum_{i=1}^{N} L(y_i, c)$
3: **Para** $m = 1$ até M **faça**
4: Calcule os resíduos: $r_{im} = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right]_{F(x)=F_{m-1}(x)}$
5: Ajuste do modelo $F_m(x)$ com os resíduos $\{r_{im}\}_{i=1}^{N}$
6: Atualização do modelo: $F_m(x) = F_{m-1}(x) + \lambda_m f_m(x)$
7: **Fim Para**
8: **Saída:** modelo final $F_M(x)$

4.2.2.5 Adaptive Boosting

Introduzido por Freund e Schapire em 1997, o *adaptive boosting* (AdaBoost) é um algoritmo de *ensemble* amplamente utilizado em problemas de classificação. O princípio básico do AdaBoost é treinar uma série de classificadores fracos, denotados por $h_t(x)$, sendo t o índice do classificador, de modo que cada novo classificador se concentre em corrigir os erros dos anteriores.

Inicialmente, cada exemplo de treinamento recebe um peso igual. A cada iteração, o peso dos exemplos mal classificados aumenta, e os exemplos corretamente classificados têm seus pesos reduzidos, fazendo com que os classificadores subsequentes deem mais ênfase aos erros anteriores (21) (50).

A função de perda exponencial é dada pela Equação 4.8, na qual a cada iteração, o AdaBoost realiza o ajuste dos pesos dos exemplos e a combinação com classificadores fracos para construir um modelo forte, ponderando-os com base na eficiência de cada classificador fraco, Portanto, o AdaBoost realiza a minimização da função durante o treinamento (21) (50) (87).

$$L(w) = \sum_{i=1}^{n} e^{-y_i f(x_i)} \tag{4.8}$$

Na Equação 4.8, y_i é a classe verdadeira do exemplo i, $f(x)$ é a predição do modelo e w representa o peso dos exemplos. A construção do modelo final, $F(x)$, é dada pela combinação ponderada dos classificadores fracos, como mostra a Equação 4.9. Nesta, $h_t(x)$ é o classificador fraco na t-ésima iteração; α_t é o peso atribuído ao classificador h_t, calculado com base na sua taxa de erro; e T é o número total de classificadores fracos (21).

$$F(x) = \sum_{t=1}^{T} \alpha_t h_t(x) \tag{4.9}$$

O peso α_t é calculado pela Equação 4.10, na qual ϵ_t é a taxa de erro do classificador h_t (21).

$$\alpha_t = \frac{1}{2} \ln\left(\frac{1-\epsilon_t}{\epsilon_t}\right) \tag{4.10}$$

O Algoritmo 4 exibe o pseudocódigo para o processo de iteração do AdaBoost, que começa com a inicialização, na qual cada exemplo de treinamento recebe um peso igual a $w_i = \frac{1}{n}$, sendo n o total de exemplos. Em seguida, o algoritmo realiza o treinamento de um classificador fraco utilizando os pesos atribuídos aos exemplos. Após o treinamento, é calculado o erro do classificador, ϵ_t, como a soma dos pesos dos exemplos que foram mal classificados pelo classificador fraco (50) (87).

Algoritmo 4 Iteração do AdaBoost

1: **Entrada:** Dados de treinamento $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$
2: **Inicialize:** Pesos dos exemplos $w_i = \frac{1}{n}$, $\forall i = 1, \ldots, n$
3: **Para** $t = 1$ até T **faça**
4: Treinar classificador fraco $h_t(x)$ usando os pesos w_i
5: Calcular o erro $\epsilon_t = \sum_{i=1}^{n} w_i \cdot \mathbb{1}\{h_t(x_i) \neq y_i\}$
6: Calcular o peso do classificador $\alpha_t = \frac{1}{2} \ln\left(\frac{1-\epsilon_t}{\epsilon_t}\right)$
7: Atualizar os pesos dos exemplos: $w_i \leftarrow w_i \cdot \exp\left(-\alpha_t \cdot y_i \cdot h_t(x_i)\right)$, normalizar os pesos
8: **Fim Para**
9: **Saída:** Classificador final $F(x) = \sum_{t=1}^{T} \alpha_t h_t(x)$

A próxima etapa envolve a atualização dos pesos dos exemplos, com um aumento para os exemplos mal classificados, para que tenham mais importância nas próximas iterações. Finalmente, após T iterações, o algoritmo combina os classificadores fracos em uma combinação final ponderada, gerando um classificador forte, mais preciso e capaz de generalizar melhor os dados (50) (87).

O AdaBoost é simples de implementar e eficiente em melhorar classificadores fracos, oferecendo boa generalização e flexibilidade em diversos problemas de classificação. Contudo, é sensível a ruídos e a valores atípicos, o que pode levar a sobreajuste, além de depender da qualidade dos classificadores fracos e necessitar de muitas iterações em alguns casos (134).

4.2.2.6 Extreme Gradient Boosting

O *extreme gradiente boosting* (XGBoost) é um algoritmo supervisionado de *ensemble*, altamente escalável e versátil, no qual os modelos preditores fracos são árvores de decisão

regularizadas (23). Em 2016, Chen e Guestrin (48) introduziram o método XGBoost como uma extensão do algoritmo GB, expandindo a definição da sua função de perda (23).

Uma das principais abordagens do algoritmo é o aprendizado regularizado, cujo objetivo é controlar a complexidade do modelo a ser criado evitando ou reduzindo o impacto de possíveis problemas como o sobreajuste (4).

Seja $\hat{y}_i$ a i-ésima predição realizada pelo XGBoost em um conjunto de dados $D = \{(x_i, y_i)\}$, definido como $|D| = n, x_i \in \mathbb{R}^m$ e $y_i \in \mathbb{R}$, sendo m variáveis de entrada e n variáveis de saída. A predição para $\hat{y}_i$ é dada pela Equação 4.11, na qual $\mathcal{F} = \{f(x) = w_{q(x)}\}(q : \mathbb{R}^m \to T, w \in \mathbb{R}^T)$ representa o CART, *i.e.*, o espaço das árvores de regressão (48).

$$\hat{y}_i = \phi(x_i) = \sum_{k=1}^{K} f_k(x_i), f_k \in \mathcal{F} \tag{4.11}$$

A variável K corresponde ao total de funções considerado; q é a definição da estrutura de cada árvore no modelo CART que mapeia um dado para o índice da folha associada; o total de folhas na árvore é dado por T; cada função f_k se refere a um valor de q e a um peso w, ambos estruturas independentes da árvore (48).

Nas tradicionais árvores de decisão, cada folha contém uma categoria específica. Já no XGBoost, toda árvore apresenta uma pontuação contínua em suas folhas, denotada por w_i, para a i-ésima folha (48).

As q regras de decisão nas árvores são aplicadas na classificação de um dado qualquer nas folhas. E pela soma das pontuações nas respectivas w folhas é obtida a predição final. E para a otimização do conjunto de funções empregadas no modelo, a minimização da função de perda regularizada dada pela Equação 4.12 é necessária (23) (48).

$$\begin{gathered} \mathcal{L}(\phi) = \sum_i l(\hat{y}_i, y_i) + \sum_k \Omega(f_k) \\ \text{onde} \quad \Omega(f) = \gamma T + \frac{1}{2}\lambda ||w||^2 \end{gathered} \tag{4.12}$$

As variáveis n e l representam a volumetria de amostras e o erro quadrático médio (MSE, do inglês *mean squared error*), respectivamente. Logo, l denota a diferença entre a predição $\hat{y}_i$ e a variável alvo y_i Com relação às folhas, o total é dado por T e os pesos por w. Já o termo Ω, expressa a regularização, que é representada nos pesos $\mathcal{L}_1$ e $\mathcal{L}_2$, respectivamente, por γ e λ. A inclusão de Ω objetiva a suavização dos pesos finais aprendidos de forma que seja evitada ocorrência de sobreajuste (23) (48).

Métodos tradicionais de otimização do espaço euclidiano não podem ser utilizados para otimizar o modelo *ensemble* da Equação 4.12, um vez que alguns dos parâmetros são funções. Para contornar essa questão, $\hat{y}^(t)$ deve representar predição da i-ésima instância na iteração dada por t, de forma que o modelo seja treinado de forma gulosa com a inclusão de f_t para a minimização da função objetivo da Equação 4.13 (48).

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t) \tag{4.13}$$

À Equação 4.13, $f(t)$ é adicionado o valor ótimo de f_t para a melhoria do modelo. Consequentemente, a aproximação de segunda ordem pode ser utilizada para a otimização, como expressa a Equação 4.14, na qual g_i e h_i são respectivamente, na função de perda, as estatísticas de gradiente de primeira e segunda ordem (48).

$$\begin{gathered}\mathcal{L}^{(t)} \approx \sum_{i=1}^{n}[l(y_i, \hat{y}_i^{(t-1)}) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)] + \Omega(f_t) \\ \text{onde} \quad g_i = \partial_{\hat{y}^{(t-1)}} l(y_i, \hat{y}^{(t-1)}) \quad \text{e} \quad h_i = \partial^2_{\hat{y}^{(t-1)}} l(y_i, \hat{y}^{(t-1)})\end{gathered} \tag{4.14}$$

Por meio da remoção dos seus termos constantes, a Equação 4.14 pode ser simplificada, resultando na Equação 4.15 (48).

$$\tilde{\mathcal{L}^{(t)}} = \sum_{i=1}^{n}[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)] + \Omega(f_t) \tag{4.15}$$

Como mostra a Equação 4.16, Ω pode ser expandido na Equação 4.15 e o termo Seja $I_j = i|q(xi_i) = j$ pode ser inserido como o conjunto de instâncias da folha dada por j (48).

$$\begin{aligned}\tilde{\mathcal{L}^{(t)}} &= \sum_{i=1}^{n}[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)] + \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \\ &= \sum_{j=1}^{T}[(\sum_{i \in I_j} g_i) w_j + \frac{1}{2}(\sum_{i \in I_j} h_i) + \lambda) w_j^2] + \gamma T\end{aligned} \tag{4.16}$$

O valor ótimo do j-ésimo peso de uma folha j, dado por w_j^*, pode ser calculado para a estrutura fixa denotada por $q(x)$, como evidencia a Equação 4.17. Nesta, $\tilde{\mathcal{L}^{(t)}}$ representa o valor ótimo correspondente (48).

$$w_j^* = -\frac{\sum\limits_{i \in I_j} g_i}{\sum\limits_{i \in I_j} h_i + \lambda}$$
$$\tilde{\mathcal{L}^{(t)}}(q) = -\frac{1}{2}\sum_{j=1}^{T} \frac{(\sum\limits_{i \in I_j} g_i)^2}{\sum\limits_{i \in I_j} h_i + \lambda} + \gamma T \tag{4.17}$$

A qualidade de uma estrutura q da árvore pode ser calculada a partir da utilização da Equação 4.17 como uma função de pontuação (*score*), semelhante à pontuação de impureza utilizada na avaliação das árvores de decisão. É impraticável listar todas as possíveis estruturas de árvore q. Como alternativa, pode ser utilizado um algoritmo guloso que inicia com uma única folha e, de forma iterativa, adiciona ramos à árvore. Sejam I_L e I_R conjuntos de instâncias dos nós esquerdo e direito após a separação, respectivamente. Assumindo que $I = I_L \cup I_R$, a Equação 4.18 expressa a redução da perda após a divisão (48).

$$\mathcal{L}_{split} = \frac{1}{2}\left[\frac{(\sum\limits_{i \in I_L} g_i)^2}{\sum\limits_{i \in I_L} h_i + \lambda} + \frac{(\sum\limits_{i \in I_R} g_i)^2}{\sum\limits_{i \in I_R} h_i + \lambda} - \frac{(\sum\limits_{i \in I} g_i)^2}{\sum\limits_{i \in I} h_i + \lambda}\right] - \gamma \tag{4.18}$$

Em aplicações práticas, o termo $\mathcal{L}_{split}$ da Equação 4.18 é utilizado para a avaliação dos candidatos resultantes da divisão. Um dos desafios fundamentais na abordagem com aprendizado de árvores consiste justamente na identificação da divisão ótima $\mathcal{L}_{split}$.

Neste sentido, alguns dos algoritmos e métodos que podem ser empregados para abordar a divisão foram propostos no artigo seminal de Chen e Guestrin (2016) (48): *basic exact greedy*, que enumera todas as possíveis divisões em todas as variáveis de um conjunto de dados a partir de uma abordagem gulosa; *approximate algorithm*, que ao propor um algoritmo de aproximação, evita o alto custo computacional exigido pelo método anterior; *weighted quantile sketch*, método baseado em percentis cuja, ideia geral é dar suporte a operações de mesclagem e poda das folhas com certo nível de eficiência, fazendo com que apenas um subconjunto de divisões candidatas seja testado, possibilitando que grandes conjuntos de dados possam ter seus candidatos divididos com maior eficiência (23), situação que é o principal problema na abordagem por aproximação; e, por fim, *sparsity-aware split finding*, que aborda casos em que o conjunto de dados é esparso.

O XGBoost tem ganhado destaque nos últimos anos no desenvolvimento de modelos preditivos devido aos seus resultados positivos de acurácia em classificações e regressões, eficiência e adaptabilidade. O algoritmo apresenta vantagens como a capacidade de lidar com dados faltantes, fácil utilização em aplicações que envolvam processamento paralelo e aptidão para trabalhar com grandes e complexos conjuntos de dados. Além disso, o

XGBoost inclui diversos parâmetros que podem ser ajustados para melhorar ainda mais sua eficácia (240).

As principais etapas do XGBoost estão detalhadas em pseudocódigo no Algoritmo 5.

Algoritmo 5 Pseudocódigo com as principais etapas do algoritmo *extreme gradient boosting*. Baseado em (48).

Requer: $D = \{(x_i, y_i)\}_{i=1}^{n}$: conjunto de treinamento; η: taxa de aprendizado; T: total de iterações;λ e γ: parâmetros da regularização; e função de perda: $\mathcal{L}$
Garante: Modelo: $f(x) = \sum_{t=1}^{T} f_t(x)$
Inicializa o modelo com $f_0(x) = 0$
Para $t = 1$ até T **faça**
    Calcule os gradientes $g_i = \partial_{\hat{y}^{(t-1)}} l(y_i, \hat{y}^{(t-1)})$ e $h_i = \partial^2_{\hat{y}^{(t-1)}} l(y_i, \hat{y}^{(t-1)})$ para todo i
    Construa uma nova árvore $f_t(x)$ por meio da minimização da regularização
    Para cada folha j na árvore **faça**
        Calcula o peso da folha $w_j^* = -\frac{\sum_{i \in I_j} g_i}{\sum_{i \in I_j} h_i + \lambda}$
        Aplica a penalização da regularização na folha
    Fim Para
    Atualiza o modelo: $\hat{y}_i \leftarrow \hat{y}_i + \eta f_t(x_i)$
Fim Para
Retorne $f(x) = \sum_{t=1}^{T} f_t(x)$

4.2.2.7 Máquina de Vetores de Suporte

No trabalho supracitado publicado em 1992 por Boser, Guyon e Vapnik (28), os autores propuseram um algoritmo supervisionado para classificação e para regressão que viria a ser tornar o algoritmo conhecido hoje como máquina de vetores de suporte (SVM, do inglês *support vector machine*) (62). Trabalhos como o de Vapnik (257) e o também supracitado de Cortes e Vapnik (57) fundamentaram a conceituação do SVM.

A destacada capacidade de generalização do SVM, juntamente ao seu poder discriminativo, tem atraído a atenção da comunidade de aprendizado de máquina em tempos recentes. Devido aos seus sólidos fundamentos teóricos e excelente capacidade de generalização, o SVM se tornou um dos métodos de classificação mais amplamente utilizados, sendo inclusive superiores a outros métodos em diversos cenários (43). Entre as principais inovações proporcionadas por esse algoritmo à área de AM estão o uso explícito de otimização convexa, teoria de aprendizado estatístico e funções *kernel* (62).

Seja um conjunto de dados expresso pela Equação 4.19, na qual x_i é um vetor n-dimensional com rótulos denotados por y_i, e $[-1, 1]$ é o intervalo de valores possíveis para a classificação desses rótulos em classes. O objetivo é desenvolver um processo de generalização no qual deve ser determinada uma função $f(x) = y : \mathbb{R}^n \rightarrow [-1, 1]$ que

obtenha bom desempenho tanto na classificação dos dados de treinamento quanto para os dados cujos padrões não foram previamente definidos (36) (62).

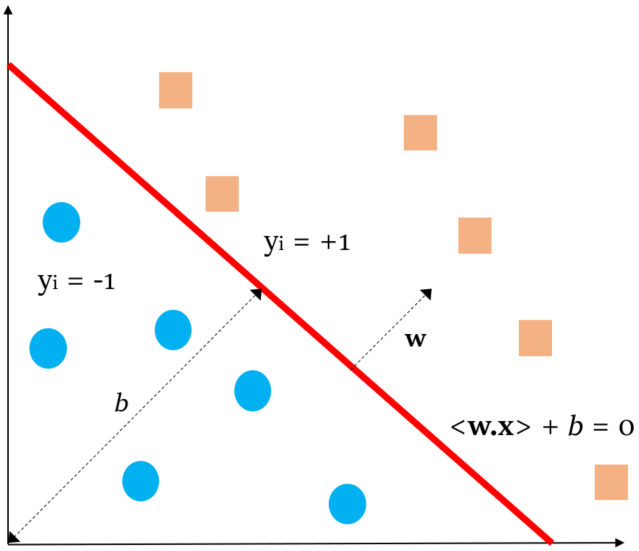
$$(x_1, y_1), \ldots, (x_l, y_l) \in \mathbb{R}^n \times [+1, -1] \tag{4.19}$$

As máquinas de vetores de suporte são baseadas em hiperplanos, cuja definição consta na Equação 4.20, em que $w \in \mathbb{R}^n$ e $b \in \mathbb{R}^n$, sendo w um vetor perpendicular ao hiperplano (43). Esta foi a proposta do trabalho de Vapnik (257), segundo o qual a restrição das classes de funções que um modelo pode aprender deve ser reduzida de forma que a obtenção da função correspondente seja factível (36).

$$\langle w \cdot x \rangle + b = 0 \tag{4.20}$$

A Equação 4.20 divide o espaço de entrada em uma parte, que contém os vetores da classe -1, e outra, com os vetores da classe +1, como mostra a Figura 9.

Figura 9 – Hiperplano (w, b) separando um conjunto de treinamento bidimensional.



Fonte: Adaptado de (36).

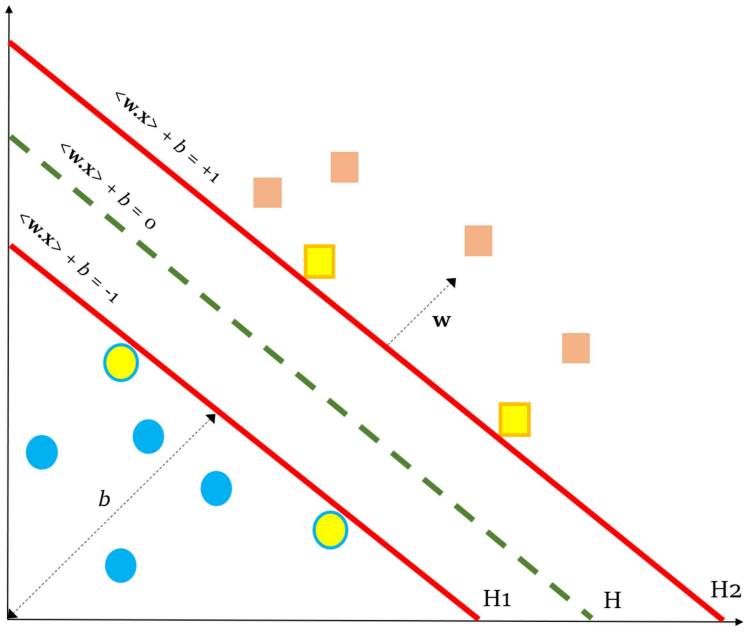
Um conjunto de dados será definido como separável linearmente nos casos em que o hiperplano exemplificado na Figura 9 exista. E para a determinação da classe de um vetor x, pode ser utilizada a Equação 4.21 (36) (62) (81).

$$f(x) = \pm(\langle w \cdot x \rangle + b) \tag{4.21}$$

É possível que haja mais de um hiperplano que classifique adequadamente o conjunto de treinamento. E dentre todos os possíveis, o definido como de margem máxima é o hiperplano ótimo do processo de generalização (61). A margem é a distância entre o hiperplano e o conjunto de treinamento mais próximo a ele (43).

Para construir o hiperplano ideal, é necessário resolver um problema de otimização convexo, que minimiza uma função quadrática sob restrições de desigualdade linear. Este tipo de problema não apresenta mínimos locais, garantindo que qualquer solução encontrada seja global. Quando a distância entre os dois hiperplanos paralelos H_1 e H_2 da Figura 10 é maximizada, alguns pontos do conjunto de dados podem estar dispostos sobre eles, na margem (43). Esses pontos são conhecidos como vetores de suporte — destacados em amarelo na Figura 10 — e são cruciais e suficientes para o problema de classificação, uma vez que possuem todas as informações necessárias (36) (43).

Figura 10 – Classificador ótimo: hiperplano de margem máxima com seus vetores de suporte destacados em amarelo.



Fonte: Adaptado de (36) e de (43).

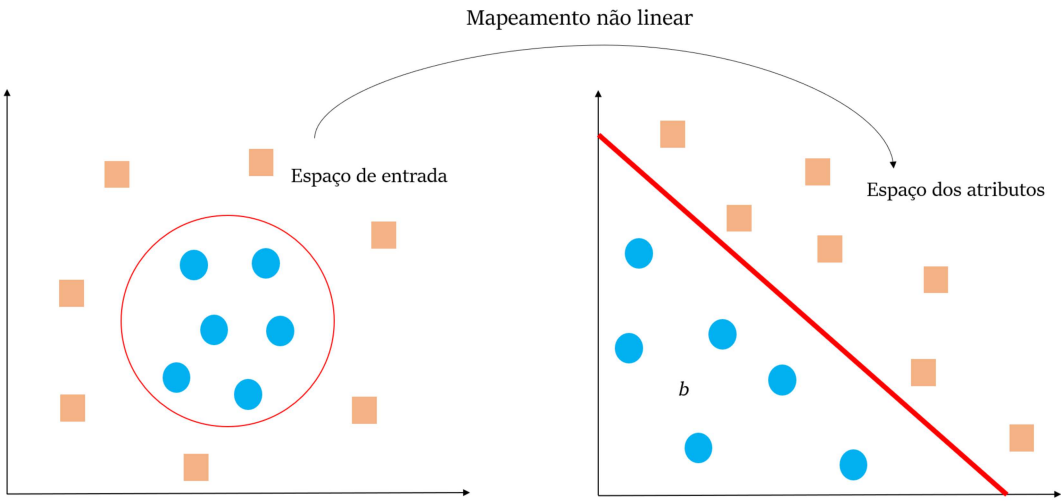
Uma propriedade adicional e importante dos classificadores de hiperplanos é que tanto o problema de otimização, usado para encontrar o hiperplano ótimo, quanto a função de decisão, utilizada para a classificação de novos vetores, podem ser expressos na forma dual, como mostra a Equação 4.22. Esta forma dual depende exclusivamente de produtos escalares entre os vetores (36) (62).

$$f(x) = \pm \left(\sum_{i=1}^{l} y_i \alpha_i \langle x \cdot x_i \rangle + b \right) \tag{4.22}$$

Na Equação 4.22, a variável $\alpha \in \mathbb{R}$ representa o quanto de informação x_i possui. Logo, para vetores que não são de suporte, este valor será zero. E para superar as limitações do SVM em representar dados complexos e realistas, pode ser empregada uma técnica de mapeamento dos valores de entrada para um espaço de características mais sofisticado, geralmente de alta dimensão, no qual os vetores se tornam linearmente separáveis. Este mapeamento é realizado por meio de uma transformação não linear, representada por por ϕ (36) (111).

Um classificador do hiperplano ótimo pode ser utilizado no espaço das variáveis de treinamento para a definição de um hiperplano de separação, como mostra a Figura 11. Assim, como o SVM preconiza, é obtida uma superfície de decisão não linear no espaço de entrada (36) (111).

Figura 11 – Mapeamento não linear do espaço de entrada para o espaço de atributos para a simplificação da classificação.



Fonte: Adaptado de (36) e de (111).

O classificador de hiperplano ótimo usa apenas produtos escalares entre vetores no espaço de entrada, consequentemente, no espaço de atributos haverá: $\langle \phi(x) \cdot \phi(y) \rangle$, que é computacionalmente custo sobretudo em espaços de alta dimensão. Para suplantar essa questão, Boser, Guyon e Vapnik (28) propuseram o conceito de função *kernel*, em que a partir de dois vetores do espaço de entrada, $k(x, y)$, é retornado o produto escalar de suas imagens no espaço de atributos, como mostra a Equação 4.23. E, para cada aplicação específica, pode ser definido um tipo de *kernel* (36) (111).

$$k(x, y) = \langle \phi(x) \cdot \phi(y) \rangle \tag{4.23}$$

Em resumo, pode ser definido que o SVM representa máquinas de aprendizado linear duais capazes de mapear, com o uso de funções *kernel*, os seus vetores de entrada

em um espaço de atributos, no qual será calculado o hiperplano ótimo. Aplicando-se à Equação 4.22 o mapeamento dado por ϕ, é obtida a Equação 4.24 (35) (36) (111).

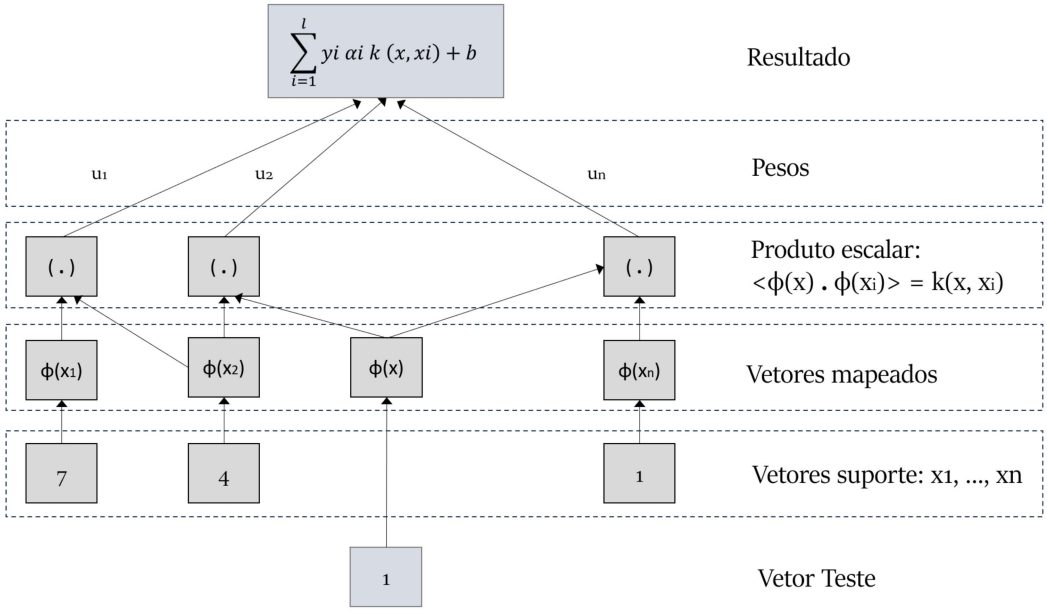
$$f(x) = \pm \left(\sum_{i=1}^{l} y_i \alpha_i \langle \phi(x) \cdot \phi(x_i) \rangle + b \right) \tag{4.24}$$

E como já é sabido que o mapeamento explícito é computacionalmente custoso, funções *kernel* podem ser utilizadas para se obter a função de decisão não linear dada pela 4.25 (35) (36) (111).

$$f(x) = \pm \left(\sum_{i=1}^{l} y_i \alpha_i k(x, x_i) + b \right) \tag{4.25}$$

Uma visão objetiva de algumas das etapas envolvidas na aplicação do algoritmo SVM pode ser vista na Figura 12.

Figura 12 – Etapas da aplicação do algoritmo máquina de vetores de suporte.



Fonte: Adaptado de (111).

4.2.2.8 Regressão Linear

Conforme já definido na Subseção *4.2.1.1*, a regressão é uma técnica de base estatística que, entre outras aplicações, pode servir de fundamentação para o desenvolvimento de um algoritmo supervisionado de AM. O objetivo é modelar a relação entre um conjunto de variáveis dependentes, Y, com um conjunto de variáveis independentes, X, como mostra a Figura 5. Por meio da modelagem, cada predição de um valor do conjunto Y é comparada

ao valor real correspondente e as diferenças obtidas formam o conjunto residual de valores (46).

A depender do tipo de relação entre as variáveis, a regressão pode ser do tipo linear simples, linear multivariada ou polinomial. Na regressão linear simples (SLR, do inglês *simple linear regression*), apenas uma variável independente está relacionada a uma única variável independente, como mostra a Equação 4.26, na qual β_0 é o intercepto do modelo, $\beta_1 x$ o coeficiente da variável independente e ϵ é o resíduo. A SLR objetiva a diferenciação da influência das variáveis independentes sobre a variável dependente, considerando-as isoladamente e sem interferências mútuas (1) (158).

$$x = \beta_0 + \beta_1 x + \epsilon \tag{4.26}$$

Já na regressão linear múltipla (MLR, do inglês *multiple linear regression*), a relação é definida entre um conjunto de variáveis independentes, $x = x_1, x2, \ldots, x_m$, e uma variável dependente, alvo da regressão, como mostra a Equação 4.27 (7) (158). O objetivo é modelar, de forma síncrona, a influência das variáveis independentes sobre a variável dependente. (255).

$$y = \beta_0 + \beta_1 x + \beta_2 x + \cdots + \beta_n x_m + \epsilon \tag{4.27}$$

A Equação 4.27 pode ser reescrita em sua forma matricial para a MLR, resultando na Equação 4.28 (1) (183).

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

$$\text{onde} \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_m \end{bmatrix}, \quad X = \begin{bmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1m} \\ 1 & x_{21} & x_{22} & \ldots & x_{2m} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{nm} \end{bmatrix}, \quad Y = \begin{bmatrix} Y_0 \\ Y_1 \\ \vdots \\ Y_m \end{bmatrix} \tag{4.28}$$

A regressão polinomial (PR, do inglês *polynomial regression*) é uma extensão da MLR na qual o modelo é ajustado para capturar relações não lineares entre a variável dependente e as variáveis independentes, incluindo n termos polinomiais das variáveis independentes. A PR é particularmente útil quando os dados apresentam tendências não lineares, permitindo que o modelo se ajuste de maneira mais adequada às variações observadas, gerando uma modelagem mais complexa e precisa das interações entre as variáveis (183). A definição de um modelo PR está exemplificada na Equação 4.29, na qual h é o grau polinomial (158).

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_h x^h + \epsilon \tag{4.29}$$

Para que sejam determinadas as estimativas dos parâmetros dados por b_0 e b_1, pode ser utilizado o método dos mínimos quadrados (MMQ). O objetivo do MMQ é determinar a curva que melhor se ajusta a um conjunto de pontos de dados, reduzindo a soma dos quadrados das diferenças (parte residual) dos pontos da curva, *i.e.*, entre os atributos (variáveis independentes) e os valores preditos (variáveis dependentes). Para encontrar as predições b_0 e b_1, o MMQ pode ser aplicado para a determinação da mínima distância entre soma dos quadrados das diferenças dos valores reais da resposta de y_i à $\hat{y} = \beta_0 + \beta_1 x_i$, Equação 4.30. Assim, tal resposta deve se aproximar do valor total mínimo entre todos os coeficientes β_0 e β_1 (158).

$$(b_0, b_1) = \min_{(\beta_0, \beta_1)} \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_i)]^2 \tag{4.30}$$

A resolução do sistema dado pela Equação 4.31 fornece o resultado do MMQ para a LR).

$$\begin{aligned} \frac{\partial}{\delta \beta_0} &= \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_i)]^2 = 0 \\ \frac{\partial}{\delta \beta_1} &= \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_i)]^2 = 0 \end{aligned} \tag{4.31}$$

As variáveis b_0 e b_1 podem ser consideradas como as soluções do sistema da equação anterior 4.31. Logo, a linha de regressão representada por $\hat{(y)} = b_0 + b_1 x$ descreve a relação entre as variáveis x e y. E por meio de um modelo linear central, denotado por $y_i = \beta_0^* + \beta_1 (x_i - \overline{x}) + \epsilon_i$, sendo $\beta_0 = \beta_0^* - \beta_1 \overline{x}$, podem ser obtidas as soluções desejadas. Por fim, por meio de uma série de manipulações algébricas possíveis, as soluções b_0 e b_1 estão expressas na Equação 4.32 (158).

$$\begin{aligned} b_0 &= b_0^* - b_1 \overline{x} = \overline{y} - b_1 \overline{x} \\ b_1 &= \frac{\sum_{i=1}^{n} (y_i - \overline{y})(x_i - \overline{x})}{\sum_{i=1}^{n} (x_i - \overline{x})^2} \end{aligned} \tag{4.32}$$

Portanto, são determinadas estimativas de parâmetros tomando a linha "mais próxima" de todos os pontos de dados (x_i, y_i) como resultado final da aplicação do MMQ. A LR residual pode ser determinada para as medições y_i e os valores ajustados $\hat{y}_i$.

Podem ser citadas como vantagens da utilização de algoritmos de LR a facilidade de implementação e a menor complexidade em comparação a outros algoritmos de AM.

Apesar da possibilidade de modelos de LR apresentarem sobreajuste, podem ser empregadas técnicas como redução de dimensionalidade e de regularização (124) (190). Ademais, a LR pode ser utilizada para encontrar a relação natural entre as variáveis consideradas (124). Já o desempenho da LR pode ser afetado por valores atípicos (190) e pelo fato de boa parte dos fenômenos e problemas naturais serem não lineares e, portanto, mais complexos do que a abordagem do algoritmo (124).

4.2.2.9 Regressão Logística

A regressão logística (LogR, do inglês *logistic regression*) é uma técnica estatística amplamente utilizada nas áreas de análise e de mineração de dados, e em aprendizado de máquina. A LogR é particularmente eficaz para a análise e a classificação de conjuntos de dados que possuem respostas binárias, sobretudo devido à sua capacidade inerente de fornecer probabilidades, o que a torna adequada em diversas aplicações práticas. Além disso, a LogR pode ser adaptada para resolver problemas de classificação multiclasse (152) (169) (275). Outra importante vantagem da LogR é a semelhança dos métodos utilizados em sua análise com os princípios empregados na LR, permitindo uma transição mais suave e uma aplicação mais eficiente dos modelos de LogR em diversos contextos analíticos (152).

Entre as diversas áreas de aplicação da LogR, pode se destacar a medicina, em pesquisas envolvendo a classificação de movimentos dos dedos para o controle de próteses de membros superiores; na detecção precoce de doenças cardíacas e da hipertensão (5) (160) (213); e em estudos de análise sobre a pandemia de Covid-19 (157) (170); em engenharias, como na predição da probabilidade de falha de empresas e no tratamento de dados incertos (169) (213); e em pesquisas envolvendo análise de sentimento (213).

Seja um vetor x, com $x \in \mathbb{R}^n$, um vetor com as variáveis referentes aos atributos de um problema, e $y \in [-1, 1]$, a classe binária com rótulos associados aos valores de saída. O modelo de regressão logística é definido pela Equação 4.33, na qual $Pr(y/x)$ denota a probabilidade condicional da variável y dado $x \in \mathbb{R}^n$ (160, 169).

$$Pr(y/x) = \frac{1}{1 + \exp(-y(\beta^T x + \alpha))} = \frac{\exp(y(\beta^T x + \alpha))}{1 + \exp(y(\beta^T x + \alpha))} \tag{4.33}$$

O termo $\beta^T x + \alpha = 0$ define um hiperplano no espaço de atributos, no qual $P(y/x) = 0.5$. Logo, $\alpha \in \mathbb{R}$ é o termo de intercepção e $\beta \in \mathbb{R}$ o vetor de pesos. Se o valor de $\beta^T x + \alpha$ for maior do que 0.5, o termo terá o mesmo sinal que y. Caso contrário, os sinais serão opostos (169).

Em outros termos, por meio da utilização de uma função sigmoide, a regressão logística realiza o mapeamento de uma combinação linear ponderada de atributos em valores reais no intervalo de 0 a 1, *i.e.*, probabilidades de um valor pertencer a uma classe qualquer. A função sigmoide pode ser definida pela Equação 4.34 (213), na qual z é a

combinação linear dos atributos (variáveis independentes) x: $z = \beta_0 + \beta x_1 + \cdots + \beta x_n = \beta^T x$

$$\sigma(z) = \frac{1}{1 + exp - z} \tag{4.34}$$

Para exemplificar a aplicação da Equação 4.33, pode ser utilizado o exemplo proposto por Musa (169), no qual dado um conjunto com m dados observados ou treinados $\{x_i, y_i\}_{i=1}^m$, com $x \in \mathbb{R}^n$ representando a i-ésima amostra e $y \in [-1, 1]$ representando a classe com o conjunto de rótulos correspondentes. Logo, o vetor de probabilidades condicionais das m amostras independentes dado pelo modelo de regressão logística pode ser representado pela Equação 4.35.

$$Pr(\alpha, \beta) = Pr(y_i/x_i) = \frac{\exp y_i(\beta^T x_i + \alpha_i)}{1 + \exp y_i(\beta^T x_i + \alpha_i)}, \qquad i = 1, \ldots, m \tag{4.35}$$

Seja $\Pi_{i=1}^m Pr(\alpha, \beta)_i$ a função de verossimilhança associada às m amostras. Logo, a função de verossimilhança logarítmica associada é dada pela Equação 4.36, na qual $a_i = x_i y_i \in \mathbb{R}^n$ e $f(z) = \log(1 + \exp(-z)$ é a função de perda logística.

$$\sum_{i=1}^{m} \log Pr(\alpha, \beta)_i = -\sum_{i=1}^{m} f(\beta^T \alpha_i + \alpha y_i) \tag{4.36}$$

Aplicando f na Equação 4.36, obtém-se a Equação 4.37.

$$\sum_{i=1}^{m} \log Pr(\alpha, \beta)_i = -\sum_{i=1}^{m} \log(1 + \exp -(\beta^T \alpha_i + \alpha y_i)) \tag{4.37}$$

Dividindo a função 4.37 por m, é possível obter a perda logística média, denotada pela Equação 4.38, na qual a determinação dos parâmetros α e β pode ser realizada por meio do método de estimativa de máxima verossimilhança, a partir dos exemplos observados e da resolução de um problema de otimização convexa dado pela minimização de $l_{avg}(\alpha, \beta)_i$. Este é conhecido como problema de regressão logística, que pode ser resolvido, entre outros métodos, pelo gradiente descendente.

$$l_{avg}(\alpha, \beta)_i = \frac{1}{m} \sum_{i=1}^{m} \log(1 + \exp -(\beta^T \alpha_i + \alpha y_i)) \tag{4.38}$$

Podem ser citadas como vantagens da utilização da regressão logística a baixa variância, a facilidade de aplicação, baixo tempo de treinamento e a capacidade de fornecer probabilidades para as saídas (5) (248). Algumas das desvantagens são a dificuldade de aplicação do algoritmo em conjuntos de dados com atributos altamente correlacionados, além da complexidade da determinação da abordagem mais adequada para o tratamento de variáveis contínuas e da relação entre as variáveis preditoras (202). Por fim, o fato da classificação multiclasse só poder ser realizada com ajustes finos no algoritmo (5, 248).

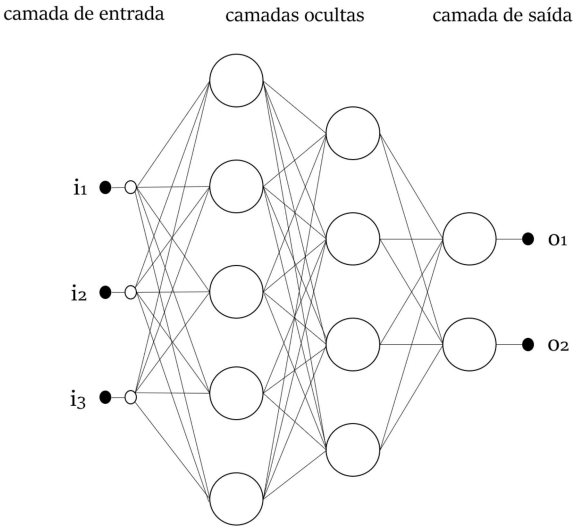
4.2.2.10 Perceptron Multicamadas

O *perceptron* multicamadas (MLP, do inglês *multilayer perceptron*) é um tipo de ANN cujas origens remontam ao trabalho supracitado de Rosenblatt (207), de 1958, que propôs o *perceptron* simples, unidade incipiente voltada ao processamento e reconhecimento de padrões, capaz de resolver apenas problemas linearmente separáveis (168). Décadas mais tarde, em 1986, Rumelhart, Hinton e Williams (212) introduziram o algoritmo de retropropagação, também já citado, que foi responsável por popularizar as pesquisas acerca do MLP.

O MLP é uma extensão do *perceptron* simples, uma vez que com a sua arquitetura de múltiplas camadas, o método possui a capacidade de aprender funções complexas e resolver problemas não lineares. E o avanço no treinamento de redes neurais, através da retropropagação, possibilitou a aplicação do MLP em diversas áreas, como reconhecimento de padrões, classificação e regressão.

Como ilustra a Figura 13, a estrutura do MLP é formada pelo mapeamento não linear entre um vetor de entrada na rede ($x = [x_1, x_2, x_3]$) e um vetor de saída ($y = [y_1, y_2]$), conectados por sistema de neurônios interconectados, os nós. Estes, por sua vez, são interconectados por pesos e sinais de saída, que são uma função da soma das entradas para o nó modificadas por ativação ou alimentação direta (*feedforward*), que é uma função de transferência não linear simples (66) (92). A capacidade do MLP de aproximar funções altamente não lineares deve-se à superposição de múltiplas funções de transferência não lineares simples. Uma função de transferência frequentemente utilizada é a função logística sigmóide, devido à facilidade do cálculo de sua derivada (92).

Figura 13 – Representação de um *perceptron* multicamadas com duas camadas ocultas. O vetor de entrada é representado por x e o vetor de saída por y.



Fonte: Adaptado de (92).

O treinamento do MLP ocorre em duas fases: na primeira, conhecida como ativação (*forward*), o vetor de entrada passa pela estrutura de cálculo da rede com um certo conjunto de pesos e sua saída é comparada com os valores observados, a fim de calcular alguma função de perda. Na segunda fase, conhecida como retrocesso (*backward*), um algoritmo de treinamento é aplicado para minimizar a função de perda ajustando os pesos, de forma que o erro geral do MLP seja reduzido. E uma vez treinado com dados de treinamento adequadamente representativos, o MLP pode generalizar para novos dados de entrada não vistos (66) (92).

No treinamento, ao se variar os pesos por todos os valores possíveis e traçar os erros no espaço tridimensional, pode ser obtida a superfície de erro. O processo de minimização pode ser realizado por meio do gradiente descendente utilizado pelo método de retropropagação, a fim de ser encontrado o mínimo absoluto ou global da superfície de erro. Os pesos na rede, então, são inicialmente definidos para pequenos valores aleatórios. Logo, o algoritmo de retropropagação calcula o gradiente local da superfície de erro e altera os pesos na direção do gradiente. Para uma superfície de erro razoavelmente suave, é esperado que os pesos convirjam para o mínimo global da superfície (92).

Segundo Gardner e Dorling (92), o algoritmo do MLP pode ser resumido nas etapas a seguir.

1. Inicialização dos pesos da rede.
2. A partir dos dados de treinamento, o vetor de entrada é introduzido na rede.
3. Propagação do vetor de entrada pela rede e obtenção do vetor de saída correspondente.
4. O sinal de erro é calculado a partir da comparação da saída real com a saída alvo.
5. Propagação do sinal de erro de volta pela rede.
6. Minimização do erro geral por meio do ajustes dos pesos.
7. Para o próximo vetor de entrada são repetidas as etapas de 2 a 7 até que o erro geral seja razoável e adequado às pretensões desejadas.

Com relação à estrutura do MLP, os n componentes do vetor de entrada da rede, $x = [x_1, x_2, \ldots, x_n])$, são conhecidos como covariáveis e são denotados por x_i. Cada um deles recebe a influência de um peso dado por w_{ij}, que conecta um neurônio i da camada de entrada a um dos k neurônios j da camada oculta, sendo $j = 1, \ldots, k$ (66).

A Equação 4.39 exprime, portanto, que para cada neurônio oculto j existe um somatório de sinais de entrada ponderados pelos pesos sinápticos do neurônio com a adição do termo de viés, dado por b_j (110).

$$v_j = \sum_{i=1}^{n} x_i w_{ij} + b_j \tag{4.39}$$

Uma função de ativação supracitada é aplicada na saída denotada por v_j e conforme a Equação 4.40 (66).

$$\varphi(v_j) = \frac{1}{1 + e^{-v_j}} \tag{4.40}$$

Finalmente, cada neurônio j da camada oculta é ponderado por um peso w_{jh}, sendo h correspondente às camadas de saída. Logo, tais pesos são adicionados com um termo de viés, como evidencia a Equação 4.41. O resultado, então, passa por uma função de ativação (66).

$$\hat{\mu}_h = \sum_{j=1}^{k} w_{jh}\varphi(v_j) + b_h \tag{4.41}$$

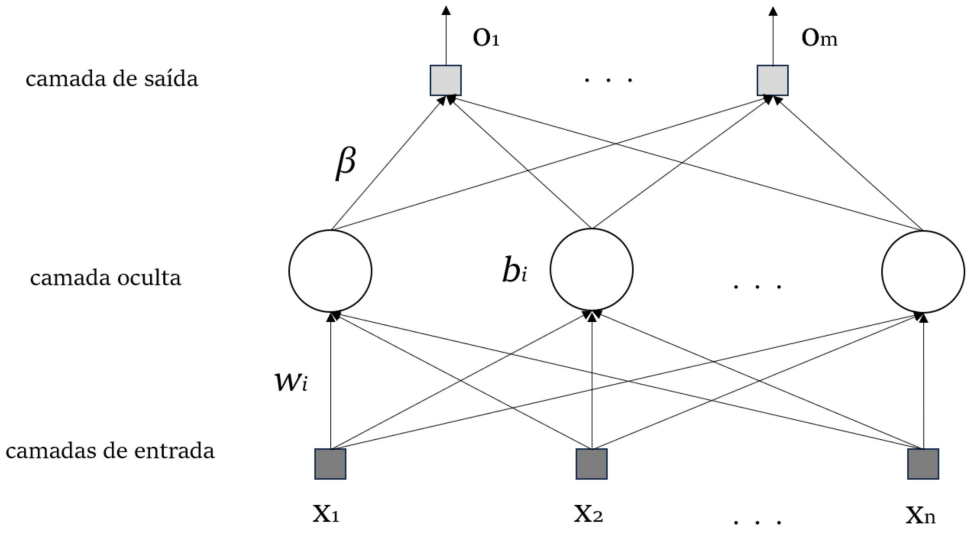
A utilização do *perceptron* multicamadas possui algumas vantagens, como a aplicação para problemas não lineares, o bom desempenho tanto com grandes volumes de dados quanto com pequenos conjuntos e a rapidez na obtenção das predições após a etapa de treinamento. Por outro lado, o MLP também apresenta desvantagens como o desempenho do modelo criado depender sobremaneira da etapa de treinamento. E o fato de não ser claramente determinado o grau de influência de cada variável independente sobre a variável dependente. A realização dos cálculos necessários é complexa e demanda tempo considerável (5).

4.2.2.11 Máquina de Aprendizado Extremo

O algoritmo supervisionado máquina de aprendizado extremo (ELM, do inglês *extreme learning machine*) foi introduzido por Guang-Bin e colaboradores (100) como uma solução eficiente para treinar redes neuronais de alimentação direta *feedforward* com uma única camada oculta (SLFNs, do inglês *single hidden layer feedforward neural networks*). Diferentemente dos métodos tradicionais de treinamento de redes neuronais, como a retropropagação (*backpropagation*), o ELM propõe a determinação aleatória dos pesos entre a camada de entrada e a camada oculta, ao passo que os pesos da camada oculta para a de saída são calculados através de uma solução analítica (101) (260).

Semelhantemente à definição estrutural do MLP e como demonstra a Figura 14, uma SLFN é composta por três camadas: entrada, oculta e saída. Entre as variáveis ilustradas, x_i representa o vetor de entrada da i-ésima amostra; b_i denota o viés do i-ésimo nó; w_i é o vetor de pesos oriundo da camada de entrada para o i-ésimo nó oculto; e β_i denota o vetor de pesos do i-ésimo nó oculto para a camada de saída (260).

Figura 14 – Configuração da estrutura de uma SLFN.



Fonte: Adaptado de (260).

O treinamento de uma rede SLFN consiste na definição dos parâmetros que conduzem à solução ideal. Logo, seja S um conjunto de treinamento expresso por $S = \{(x_i, t_i)|x_i = (x_{i1}, x_{i2}, \ldots, x_{in})^T \in \mathbb{R}^n, t_i = (t_{i1}, t_{i2}, \ldots, t_{in})^T \in \mathbb{R}^m\}$, no qual t_i representa o alvo. Uma saída de um ELM com $\hat{N}$ neurônios ocultos denotada por o pode ser descrita pela Equação 4.42, sendo $g(x)$ a função de ativação na camada oculta (101) (260).

$$\sum_{i=1}^{\hat{N}} \beta_i g(w_i x_j + b_i) = o_j, \quad j = 1, \ldots, N \tag{4.42}$$

No ELM, as funções de ativação adotam uma natureza não linear com o propósito de assegurar um mapeamento não linear para o sistema. Entre os tipos de função que podem ser utilizadas, estão a sigmoide, a cosseno, a de base radial, as quadráticas, entre outras. O treinamento da rede visa, principalmente, à minimização do erro entre o valor alvo e a saída do ELM, sendo que a função objetivo mais frequentemente aplicada para essa finalidade é o erro quadrático médio: $\sum_{i=1}^{N} = (t_{ij} - o_{ij})^2, \quad j = 1, \ldots, m$, sendo i e j os índices para as N amostras de treinamento para o nó da camada de saída. Quanto maior o valor de N, maior a capacidade de aproximação da rede, conforme denota a Equação 4.43, conhecida como capacidade universal de aproximação (101) (260).

$$\sum_{i=1}^{N} \|o_j - t_j\| = 0 \tag{4.43}$$

Substituindo o por t na Equação 4.42, é possível obter a equação matricial $H\beta = T$, na qual cada fator está expresso em detalhes na Equação 4.44 e para a qual a determinação dos valores ótimos para β_i, b_i e w_i é o resultado do treinamento de uma SLFN (101) (260).

$$
\begin{gathered}
H(w_1, \ldots, w_{\hat{N}}, b_1, \ldots, b_{\hat{N}}, x_1, \ldots, x_N) = \\
\begin{bmatrix}
g(w_1 x_1 + b_1 & \ldots & g(w_{\hat{N}} x_1 + b_{\hat{N}}) \\
\vdots & \ddots & \vdots \\
g(w_1 x_N + b_1 & \ldots & g(w_{\hat{N}} x_N + b_{\hat{N}})
\end{bmatrix}
\end{gathered}
\tag{4.44}
$$

$$
\beta = \begin{bmatrix} \beta_1^T \\ \vdots \\ \beta_{\hat{N}}^T \end{bmatrix}_{\hat{N} \times m} \qquad \text{e} \qquad T = \begin{bmatrix} t_1^T \\ \vdots \\ t_N^T \end{bmatrix}_{N \times m}
$$

A síntese do treinamento do ELM pode ser descrita pelas etapas a seguir (260).

1. Definição do conjunto de treinamento $S = [(x_i, y_i) | x_i \in \mathbb{R}^n, y_i \in \mathbb{R}^m, i = 1, \ldots, N]$.
2. Inicialização: atribuição de valores aleatórios para o peso oculto w_i e para o viés b_i, e cálculo da matriz de saída da camada oculta H a partir do conjunto de treinamento.
3. Solução analítica: obtenção de β a partir de $H\beta = T$ pela inversa de Moore-Penrose: $\beta = H^{\dagger}T$, sendo $H^{\dagger}$ a generalização inversa de Moore-Penrose da matriz H.

O ELM possui diversas vantagens em relação a outros sistemas de aprendizado superficial, como o MLP e o SVM. Entre as principais características podem ser citadas a sua taxa de aprendizado superior, sendo adequado tanto para problemas de aprendizado supervisionado quanto não supervisionado (147). Além disso, o ELM é reconhecido por sua rápida e eficiente velocidade de aprendizado, rápida convergência, boa capacidade de generalização e facilidade de implementação (121). Embora o ELM exiba alta complexidade em termos de espaço e tempo de processamento, estratégias de otimização vem sendo implementadas para a mitigação dessas limitações (49). O ELM também foi ampliado para uma vasta gama de tarefas de aprendizado, incluindo agrupamento, seleção de atributos e aprendizado representacional (121). Ademais, sua aplicação se estende a diversos domínios, como engenharia biomédica, visão computacional, identificação e controle de sistemas, além de robótica (121).

Entre as limitações do ELM, está o fato de que os pesos dos nós da camada oculta são definidos aleatoriamente ou de forma artificial e não exigem atualização após a definição inicial. Além disso, o ELM tradicional enfrenta dificuldades para treinar grandes volumes de dados de maneira rápida e eficiente (49). No entanto, estratégias de otimização podem ser implementadas para diminuir essas limitações no ELM tradicional (49). Outro ponto a ser considerado é que o ELM possui uma capacidade limitada de robustez contra valores atípicos, o que pode comprometer sua eficácia em certos contextos (147).

4.3 Métricas de avaliação do desempenho

No contexto de AM e análise de classificadores, a avaliação do desempenho de um modelo é uma etapa crucial. Para isso, utilizam-se diversas métricas que fornecem informações complementares sobre como o modelo está se comportando em diferentes aspectos. As métricas mais utilizadas incluem a acurácia (*accuracy*), revocação (*recall*), precisão (*precision*), F1-*score* e ROC AUC. Cada uma dessas métricas possui suas características e interpretações, sendo mais adequada em determinados cenários e para certos tipos de problemas (113) (195).

Para a representação dos conceitos a seguir, devem ser considerados as seguintes definições (113):

- Falsos positivos: FP, do inglês *false positive*;
- Verdadeiros positivos: TP, do inglês *true positive*;
- Falsos negativos: FN, do inglês *false negative*;
- Verdadeiros negativos: TN, do inglês *true negative*.

4.3.1 Acurácia

Uma das métricas mais simples e frequentemente utilizada para avaliar classificadores é a acurácia. Ela é definida como a proporção de predições corretas em relação ao total de predições realizadas. Sua fórmula está representada pela Equação 4.45 (42) (113) (195).

$$\text{Acurácia} = \frac{TP + TN}{TP + FP + TN + FN} \tag{4.45}$$

A acurácia é eficaz em problemas balanceados, quando o número de exemplos de cada classe é aproximadamente o mesmo. No entanto, em problemas de classificação desbalanceados, essa métrica pode ser enganosa. Por exemplo, em um conjunto de dados onde 95% dos exemplos pertencem à classe negativa, um classificador que sempre prediz a classe negativa teria uma acurácia de 95%, mesmo que não detecte nenhum exemplo da classe positiva (195).

4.3.2 Precisão

A precisão mede a proporção de exemplos corretamente classificados como positivos em relação ao total de exemplos classificados como positivos (incluindo os falsos positivos), como evidencia a Equação 4.46 (42) (113) (195).

$$\text{Precisão} = \frac{TP}{TP + FP} \tag{4.46}$$

Essa métrica é útil quando o foco está em minimizar o número de falsos positivos, como em aplicações de detecção de fraudes ou diagnósticos médicos, situações em que predições erradas de positivo podem ter consequências sérias (195).

4.3.3 Revocação

A revocação, também conhecida como sensibilidade ou taxa de verdadeiros positivos, mede a capacidade do classificador de identificar todos os exemplos positivos reais. A Equação 4.47 fornece a definição de revocação (42) (113) (195).

$$\text{Revocação} = \frac{TP}{TP + FN} \tag{4.47}$$

A revocação é importante em cenários em que a identificação correta de todos os exemplos positivos é crítica. Em problemas como diagnóstico de doenças, é preferível um modelo que identifique todos os casos da doença, mesmo que ocasionalmente classifique alguns indivíduos saudáveis como doentes (195).

4.3.4 F1-*score*

A Equação 4.48 define o conceito de F1-*score*, que é a média harmônica entre precisão e revocação, proporcionando um balanço entre essas duas métricas (42) (113) (195).

$$\text{F1-}score = 2 \times \frac{\text{Precisão} \times \text{Revocação}}{\text{Precisão} + \text{Revocação}} \tag{4.48}$$

O F1-*score* é útil em situações em que tanto falsos positivos quanto falsos negativos devem ser minimizados, e quando há um desbalanceamento entre as classes (195).

4.3.5 ROC AUC

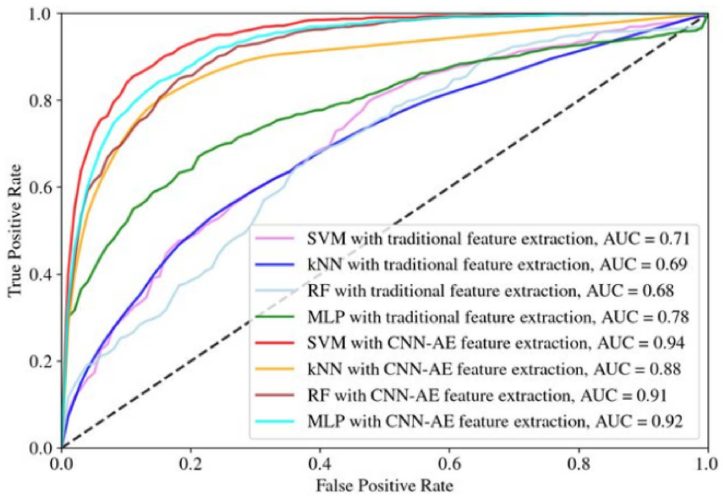
A curva *receiver operating characteristic* (ROC) é uma ferramenta gráfica utilizada para avaliar o desempenho de modelos de classificação. Ela traça a taxa de verdadeiros positivos em função da taxa de falsos positivos para diferentes limiares de decisão. A curva ROC é útil porque permite comparar o desempenho de diferentes modelos ao longo de todos os limiares possíveis, em vez de focar em um único ponto de corte (31) (83).

A área sob a curva ROC, denotada por AUC e exemplificada na Figura 15, é uma métrica que resume a performance de um classificador. A AUC varia de 0 a 1, sendo 1 o valor que indica um modelo perfeito, ao passo que 0,5 indica um modelo equivalente à aleatoriedade. Valores mais próximos de 1 indicam melhor discriminação entre as classes.

Em termos práticos, a AUC representa a probabilidade de que um classificador atribua uma pontuação mais alta para uma amostra positiva do que para uma amostra negativa escolhida aleatoriamente (31) (83).

Matematicamente, a AUC pode ser interpretada como a integral da curva ROC e pode ser calculada a partir do uso de diferentes técnicas, como a aproximação trapezoidal. Além disso, a ROC AUC é robusta contra problemas de desbalanceamento de classes, pois avalia a performance global do modelo em vez de se concentrar em uma única métrica de acurácia.

Figura 15 – Exemplo da ROC AUC com os resultados de oito algoritmos diferentes.



Fonte: Extraído de (125).

4.4 Tratamento dos dados faltantes

Um dos pontos de atenção centrais na análise de dados é o tratamento dos dados faltantes. Este é um passo fundamental para garantir a efetividade e a confiabilidade das análises e dos modelos de inteligência artificial que podem ser desenvolvidos a partir dos dados em questão (273).

Uma parcela considerável das bases utilizadas em indústrias, comércios e em pesquisas epidemiológicas não possui o registro de todas as informações previamente julgadas como necessárias para a sua efetiva utilização. Situações como essa podem ocorrer devido a diversos fatores, como falhas nos processos de inserção manual de dados, erros em equipamentos e medições imprecisas (130) (187).

Um valor faltante pode ser representado em uma base de dados por símbolos como “?”, “-”, “*”, “.”, entre outros; pela indicação “NaN” (*Not a Number*, “não é um número”, em tradução livre); por “Nulo”; simplesmente por uma célula ou campo vazio; entre outros formatos (130).

Uma das abordagens possivelmente problemáticas utilizadas em pesquisas com dados faltantes é a simples remoção dos registros com informações ausentes considerando, portanto, somente as ocorrências com todas as variáveis completas. Uma vez sem todos os registros, as informações restantes podem apresentar uma visão enviesada dos dados como um todo, sobretudo quando a ausência de registros é frequente para um considerável número de campos. Dessa forma, há um cenário de perda da representatividade da amostra original, que também pode ocorrer no caso de abordagens que considerem apenas a substituição irrefletida dos dados faltantes por valores quaisquer. Logo, em ambas abordagens há perda substancial de eficácia (236).

Outras questões concernentes a dados faltantes são a perda de eficiência e o aumento da complexidade do processo de análise. A primeira, se refere ao dispêndio de tempo necessário para o próprio tratamento dos dados em si, o que, via de regra, pode ser custoso de diferentes maneiras. A segunda questão está intrinsecamente vinculada à primeira, uma vez que muitos algoritmos e métodos de inteligência artificial predominantemente não permitem a presença de dados faltantes para que sejam devidamente aplicados (150).

Além da ausência, a volumetria de dados faltantes pode ser indesejadamente elevada em diversos casos, fomentando a necessidade de que os dados presentes sejam analisados adequadamente. Dessa forma, lacunas podem ser preenchidas de maneira significativa e representativa com relação à base de dados integral. E, de acordo com a configuração e o relacionamento entre os dados, cada conjunto pode ser classificado em três categorias, descritas a seguir (187).

4.4.1 Missing at Random

O conceito de *missing at random* (MAR) foi proposto em 1976, por Donald Rubin (209), e pode ser definido como uma condição específica na qual a ausência de dados não é completamente aleatória, mas pode ser explicada por outras variáveis observadas no conjunto de dados. Logo, os dados estão faltando de forma diretamente relacionada e dependente a outras informações presentes na base de dados, mas não do próprio valor faltante. Portanto, um dado cuja ausência se deve somente a razões estruturais pode ser categorizado como MAR (79) (130) (187).

Um exemplo de MAR pode ser uma base com dados clínicos em que os valores de pressão arterial estejam faltando. Se a probabilidade de a pressão arterial estar faltando depende da idade dos pacientes (uma variável observada), mas não dos próprios valores da pressão arterial, então os dados são considerados MAR.

4.4.2 Missing Completely at Random

O termo *missing completely at random* (MCAR) consiste da definição de que a ausência de dados não está relacionada tanto aos dados observados quanto aos não

observados, isto é, a ausência de dados em um conjunto ocorre de maneira totalmente aleatória e independente. Logo, a probabilidade de que qualquer dado esteja faltando é a mesma para todos os casos e não está relacionada a qualquer característica do dado ou do conjunto de dados (146) (187). Um exemplo de conjunto de dados categorizado como MCAR é um caso em que as amostras de uma coleta de campo sejam perdidas ou danificadas de maneira acidental e independentemente das características dos dados em si (130).

Para formalizar o conceito, pode ser utilizado o exemplo proposto por Glas (2010) (97). Considerando k variáveis relativas a dados socioeconômicos de um grupo de indivíduos (com valores de 1 a M), e um grupo de i indivíduos distintos (1 a N). Sendo y_{ik} uma observação aleatória do conjunto de dados, pode ser definida a Equação 4.49:

$$d_{ik} = \begin{cases} 0 & \text{se } y_{ik} \text{ não foi observado} \\ 1 & \text{se } y_{ik} \text{ foi observado} \end{cases} \tag{4.49}$$

Portanto, para que o conjunto de dados possa ser definido como MCAR, d_{ik} e y_{ik} devem ser independentes.

4.4.3 Missing Not at Random

Conjuntos de dados podem ser classificados como *missing not at random* (MNAR) quanto a ausência de dados está diretamente relacionada aos próprios dados não observados, que não estão inclusos no conjunto de dados. Esse tipo de dados faltantes apresenta um padrão específico que não pode ser explicado pelas variáveis observadas (79) (187) (209).

Formalmente, se a probabilidade de uma variável estar ausente depende dos seus próprios valores ou de outras variáveis não observadas no conjunto de dados, este pode ser considerado do tipo MNAR. Isso significa que a ausência dos dados está diretamente relacionada às informações que não estão presentes (130).

Como um exemplo simples, em um estudo sobre o consumo de álcool, os participantes com altos níveis de consumo podem ser menos propensos a relatar seu consumo verdadeiro, resultando em dados faltantes. Nesse caso, a probabilidade de faltar dados sobre consumo de álcool depende dos próprios níveis de consumo (valores não observados). Outro exemplo de caso pode ser uma pesquisa científica acerca do consumo de drogas ilícitas, os entrevistados que possuem dependência química tem menor pretensão de serem sinceros quanto ao seu real consumo de substâncias. Logo, a probabilidade de não se ter todos os dados sobre o consumo de drogas ilícitas depende dos próprios níveis de consumo, os quais não são valores passíveis de observação (130).

4.4.4 Métodos de inferência simples

Algumas das estratégias mais simples e objetivas de inferência de dados faltantes são as substituições por zeros, pela média e pela mediana. No primeiro caso, os registros ausentes são todos essencialmente preenchidos pelo valor zero, independentemente de quais sejam as características e o relacionamento entre as variáveis consideradas (109). Já as substituições pela média, pela mediana e pela moda são alguns dos métodos mais comuns para substituição de dados ausentes (128), os quais são preenchidos pelo valor da média, da mediana e da moda, respectivamente, dos valores presentes no conjunto de dados (273).

Embora sejam métodos simples e de fácil implementação, há desvantagens importantes em suas aplicações. As inferências por zero, pela média e pela mediana utilizam apenas valores fixos para substituir valores ausentes, gerando resultados que podem não representar adequadamente a correlação intrínseca entre as variáveis distintas de um conjunto de dados, servindo apenas como uma estimativa geral, imprecisa e com possíveis vieses (79). Se o número de valores ausentes for elevado, a substituição pelo mesmo valor pode alterar a forma da distribuição dos dados, tornando o desvio padrão menor quando comparado à configuração inicial dos dados. Quanto maior o número de valores ausentes, maior será a redução no desvio padrão (128). No entanto, a média, a mediana e a moda podem ser eficazes em aplicações em que não há muitos valores atípicos, sendo capazes de representar de maneira eficaz as informações mais comuns dos dados (273).

4.4.5 Inferência Múltipla por Equações Encadeadas

A inferência múltipla por equações encadeadas (MICE, do inglês *multiple imputation by chained equations*) é um método para tratamento de dados faltantes que permite preservar a variabilidade dos dados originais, ao passo que estima valores plausíveis para as observações incompletas (14) (133) (256).

O MICE é baseado no princípio de que cada variável com dados faltantes passa por inserção de dados condicionalmente às outras variáveis no conjunto de dados. Ao invés de todos os dados serem inseridos concomitantemente, o processo é realizado de maneira iterativa em etapas ou "equações encadeadas" (*chained equations*), sendo que cada variável sequencialmente recebe a inserção dos dados, ao passo que as demais variáveis são mantidas fixas (256). Assim, cada variável pode ser modelada de acordo com a sua distribuição (14).

Segundo Azur *et al.* (2011) (14), o processo de equações encadeadas pode ser dividido em seis etapas fundamentais:

1. Um conjunto de "marcadores de posição" é construído a partir da inferência básica, *e.g.* pela média, realizada para todo dado faltante da base considerada.

2. Os marcadores de posição indicam que, para uma variável p, as inferências devem ser redefinidas como "ausentes".

3. Para a regressão das demais variáveis do modelo de inferência são utilizados os valores de p da etapa anterior, i.e., p é a variável dependente. Podem ser utilizadas todas as variáveis do conjunto de dados ou somente uma parte delas, as quais formam a parcela independente nos modelos de regressão.

4. Inferências geradas pelo modelo de regressão substituem os valores ausentes. Uma vez que a variável p seja utilizada como independente nos modelos de regressão para outras variáveis, serão utilizados tanto os valores observados como os inferidos.

5. As etapas 2 a 4 são então repetidas para cada variável que possui dados faltantes. Ao final de cada iteração, predições de regressões substituirão os valores ausentes, evidenciando, assim, a relação entre os dados observados.

6. Por um total definido de interações e com inferência de valores em cada um deles, deve ocorrer a repetição das etapas 2 a 4.

O total de ciclos pode ser determinado para cada aplicação específica, mas, em geral, dez são realizados (14) (203). A cada final de ciclo, o conjunto de dados vai recebendo os dados que devem ser inseridos. Ademais, a expectativa é que ocorra a estabilização da distribuição dos parâmetros que determinam tais dados (como, por exemplo, os coeficientes dos modelos de regressão), evitando que a ordem das variáveis inferidas gere algum tipo de dependência (14).

Algumas vantagens do MICE são a preservação da variabilidade dos dados originais, refletindo a incerteza na inferência; a flexibilidade, devido à sua capacidade de lidar com diferentes tipos de variáveis e padrões complexos de dados faltantes; além de ser um método recomendado em muitos campos de pesquisa e análise de dados (106).

4.4.6 K-vizinhos Mais Próximos

As bases conceituais do termo "vizinhos mais próximos", utilizadas no desenvolvimento do algoritmo k-vizinhos mais próximos (KNN, do inglês *k-nearest neighbors*), foram inicialmente propostas em 1967 por Cover *et al.* (58).

O KNN é um método supervisionado de AM fundamentado na proposta de determinar e rotular pontos vizinhos com características similares. Possui vasta aplicação em problemas de regressão, classificação e inferência de dados faltantes, além de ser um método simples, fácil de ser implementado e eficiente para diversos problemas práticos (239).

O algoritmo clássico do KNN define uma variável k que representa o total de vizinhos mais próximos de um ponto. A partir de cada ponto do conjunto pré-definido,

é calculada a sua distância aos demais pontos considerados, de forma que os k vizinhos mais próximos sejam determinados e rotulados (252).

Três das métricas de distância mais utilizadas para o cálculo dos vizinhos mais próximos são a distância Euclidiana, dada pela Equação 4.50; a distância de Manhattan, dada pela Equação 4.51; e uma equação que é a generalização das duas anteriores, a distância de Minkowski (63), Equação 4.52. Em todas as equações, $x = (x_1, x_2, ..., x_n)$ e $y = (y_1, y_2, ..., y_n)$ são dois pontos quaisquer. E a equação da distância de Minkowski é uma generalização da distância de Manhattan para $q = 1$ e da distância Euclidiana para $q = 2$ (252).

$$d(x, y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \tag{4.50}$$

$$d(x, y) = \sum_{i=1}^{n} |(x_i - y_i)| \tag{4.51}$$

$$d(x, y) = \left(\sum_{i=1}^{n} |(x_i - y_i)|^q\right)^{\frac{1}{q}} \tag{4.52}$$

Após o cálculo das distâncias, o algoritmo determina os vizinhos mais próximos ao ponto considerado. Em seguida, diferentes métricas podem ser aplicadas para determinar a classificação do ponto e a sua rotulação. A mais comum e direta é a votação majoritária, na qual a maior quantidade de vizinhos de uma mesma classe vai determinar a classe do ponto em questão (252).

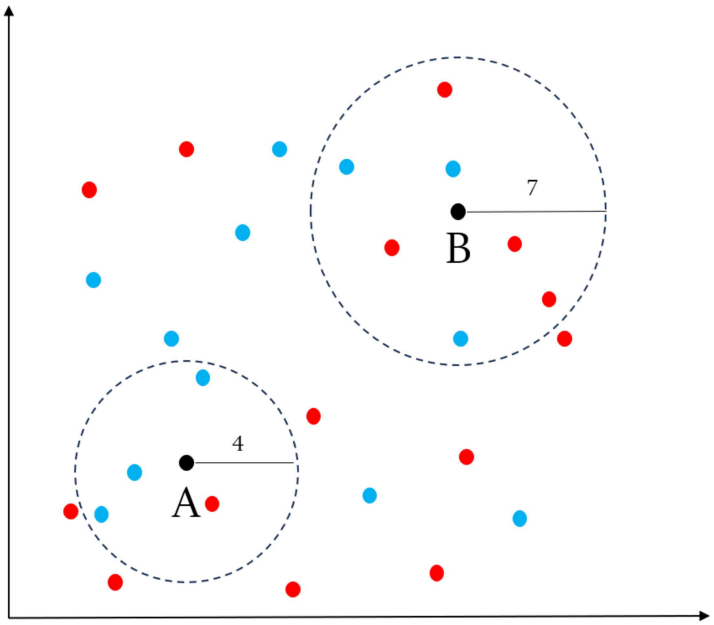
A Figura 16 exemplifica o processo descrito para duas classes representadas pelos pontos azuis e vermelhos. Se o total de vizinhos mais próximos do ponto A, denotado por k, for definido como 4, os três pontos azuis e o único ponto vermelho serão selecionados. Logo, o ponto A será rotulado como sendo da classe "azul". No caso do ponto B, para $k = 7$, há a predominância de quatro vizinhos vermelhos contra três azuis. Logo, B será classificado como "vermelho" (252).

Outras métricas de rotulação, a partir dos vizinhos mais próximos, podem ser utilizadas. Uma delas é a atribuição de pesos maiores aos vizinhos localmente mais próximos, independentemente do total de k amostras da mesma classe. Essa abordagem implementa a votação ponderada pela distância, na qual os votos dos vizinhos têm peso inversamente proporcional às suas distâncias até cada ponto considerado. Uma outra abordagem encontrada na literatura não considera o inverso da distância, mas sim, o seu valor em uma função exponencial (63).

Algumas variantes do KNN clássico são a *adaptive KNN*, cujo foco reside na seleção do valor ótimo para k em cada ponto do conjunto considerado; a *locally adaptive KNN with discrimination class*, a qual determina que a quantidade e a distribuição dos vizinhos

das classes majoritárias principal e secundária devem ser utilizadas para a determinação de um valor ideal para k na vizinhança k de um dado ponto; a *weight adjusted KNN*, que determina pesos para cada um dos pontos do conjunto de dados considerado; entre outras (241) (252).

Figura 16 – Exemplo do funcionamento do algoritmo KNN ao encontrar os vizinhos mais próximos dos pontos A e B a partir dos raios determinados: 4 e 7. O ponto A é classificado como "azul" e o ponto B como "vermelho".



Fonte: Adaptado de (252).

O algoritmo KNN pode ser utilizado também no tratamento de bases de dados com valores faltantes (18), (79), (133), (171). Nesse contexto, a similaridade entre os dados completos pode ser avaliada e utilizada para a geração de dados sintéticos (24). O mecanismo de operação do KNN mantém os mesmos princípios de sua aplicação clássica, com a diferença de que o conjunto de pontos considerados abrange apenas os valores disponíveis na base de dados. A inferência de uma nova amostra é realizada identificando os vizinhos mais próximos no conjunto de treinamento. A imputação dos valores faltantes pode ser realizada de diversas maneiras, sendo que a média (para variáveis contínuas) e a moda (para variáveis categóricas) estão entre as abordagens mais comumente aplicadas (18).

O Algoritmo 6 exemplifica em pseudocódigo o passo-a-passo da aplicação do KNN no tratamento dos dados ausentes, tendo a distância Euclidiana, Equação 4.50, como a métrica de cálculo dos k vizinhos mais próximos.

Algoritmo 6 Pseudocódigo do algoritmo KNN utilizando como métrica para o cálculo dos vizinhos a distância Euclidiana.

Requer: D (conjunto de dados de treinamento), k (número de vizinhos), $\mathbf{x}$ (novo ponto)
Garante: Classe ou valor predito para $\mathbf{x}$

Função KNN($D, k, \mathbf{x}$)
    $distancias \leftarrow []$
    Para cada $(\mathbf{x_i}, y_i) \in D$ **faça**
        $d \leftarrow \sqrt{\sum_{j=1}^{n}(x_j - x_{ij})^2}$
        $distancias$.append((d, y_i))
    Fim Para
    $distancias \leftarrow$ ordena($distancias$)
    $vizinhos \leftarrow$ distancias$[1 : k]$
    $valores \leftarrow \{y : (d, y) \in vizinhos\}$
    Retorne média(valores)
Fim Função

4.4.7 Cópulas

Proposto em 1959 por Sklar (229), o conceito de cópulas desempenha um papel essencial na estatística multivariável por abordar a modelagem de dependência entre dados univariados. Uma vez que as distribuições marginais univariadas sejam conhecidas, as cópulas se relacionam diretamente à abordagem de distribuições multivaridadas (226).

Dado um vetor aleatório n-dimensional $X = (X_1, ..., X_n)$, o Teorema de Sklar (229) define que há uma função de distribuição conjunta, H, com funções marginais $F_{x_1}, ..., F_{x_n}$. Logo, existe uma cópula n-dimensional dada por $C : [0, 1]^d \rightarrow [0, 1]$, que une a função de distribuição conjunta às suas marginais, como evidencia a Equação 4.53 (13):

$$H(x_1, x_2, ..., x_n) = C(F_{x_1}(x_1), F2(x_2), ..., F_{x_n}(x_n)) \tag{4.53}$$

A partir do seu teorema, Sklar demonstrou que para qualquer distribuição conjunta F com marginais $F_{x_1}, ..., F_{x_n}$, exite uma cópula C que satisfaz esta equação. E, se F for contínua, C é única (176).

As cópulas podem ser utilizadas para a geração de dados sintéticos levando em consideração não apenas as distribuições marginais das variáveis, mas também como essas variáveis estão relacionadas umas com as outras. Isso ajuda a preservar a estrutura multivariada dos dados, o que é crucial para manter a coerência nos conjuntos de dados inferidos (176).

Diferentemente de métodos considerados mais tradicionais, as cópulas possibilitam que a inferência de dados faltantes envolva estruturas de dependências complexas e não lineares entre as variáveis. E, por serem invariantes sob transformações marginais, as cópulas oferecem robustez na inferência, mesmo nos casos em que há variação de escala entre as variáveis (176).

A aplicação do fundamento das cópulas em um conjunto de dados fornece como resultado a análise do perfil de distribuição das variáveis envolvidas. E, a depender do tipo de distribuição envolvido, podem ser aplicados diferentes métodos para a inferência dos dados, como a cópula Gaussiana, fundamentada na distribuição normal multivariada, que é eficaz para capturar dependências lineares entre as variáveis e particularmente apropriada para modelagem de dependências simétricas (176); a cópula de Clayton, cuja utilidade está na modelagem de dependências com caudas fortes, com destaque para a cauda inferior, *i.e.*, em situações em que valores inferiores de uma variável estão fortemente associados a valores inferiores de outra variável. É aplicada especialmente em finanças para capturar a dependência em eventos extremos (51); a cópula de Gumbel, utilizada para descrever dependências em caudas, sobretudo na cauda superior, além de ser útil em casos de ocorrência de extremos simultâneos (94). A cópula de Frank, a cópula de Joe e a cópula de Plackett são outros tipos, cujas aplicações podem ser facilmente encontradas na literatura.

4.5 Desbalanceamento de dados

Não é incomum encontrar bases cujos dados se encontram sem qualquer balanceamento em sua distribuição. Esse é um problema recorrente na elaboração de modelos de inteligência artificial e computacional, e sua existência pode ocorrer por diversos fatores, como: coleta imprópria dos dados, erros no registro das informações ou até mesmo pode acontecer devido a características particulares da própria fonte de dados. Por exemplo, uma base de dados com pessoas com a doença de Alzheimer tende a conter majoritariamente indivíduos idosos, deixando a proporção desses indivíduos muito superior aos demais (131).

Um conjunto de dados está desbalanceado quando a proporção de uma ou mais classes é superior às demais. Neste caso, o uso de algoritmos de AM nos dados de treinamento é prejudicado, uma vez que as classes majoritárias podem enviesar a criação do modelo (32).

Para contornar esses problemas, duas das técnicas mais utilizadas são o *oversampling* e o *undersampling*. A primeira, gera dados sintéticos para que a amostra da classe minoritária seja aumentada. Já a segunda, atua no sentido oposto, diminuindo a volumetria da classe majoritária. Em ambos os casos o objetivo central é proporcionar o equilíbrio ao total de dados em cada uma das classes do conjunto (131).

Diferentes abordagens podem ser encontradas na literatura para o tratamento de dados desbalanceados por meio do *undersampling* e do *oversampling* (131). Alguns dos métodos mais utilizados estão descritos a seguir.

4.5.1 SMOTE

O algoritmo *synthetic minority over-sampling technique* (SMOTE) foi proposto em 2002 por Chawla *et al.* (2002) (47) e é um dos métodos mais aplicados para o tratamento de dados desbalanceados. O SMOTE aplica o *oversampling* para gerar amostras sintéticas e adicioná-las à classe minoritária por meio do KNN (76).

O passo-a-passo do método está detalhado em pseudocódigo no Algoritmo 7. Inicialmente, são determinados o conjunto de amostras da classe minoritária ($X_{minoritaria}$), o total de amostras sintéticas a serem geradas, N, e o total de k vizinhos mais próximos. O método começa com a seleção de cada amostra aleatória x_i da classe minoritária. A partir do algoritmo KNN são identificados os vizinhos mais próximos. E, para cada amostra, é gerado um exemplo sintético aleatório por meio da Equação 4.54, na qual x_i é a nova instância sintetizada, x_i é a amostra aleatória original, δ é um valor aleatório entre 0 e 1 e x_ij é a instância vizinha da de x_i.

$$x_{novo} = x_i + \delta \times (x_{ij} - x_i) \tag{4.54}$$

Algoritmo 7 Pseudocódigo do algoritmo SMOTE.

Requer: $X_{\text{minoritária}}$: Conjunto de amostras da classe minoritária
Requer: N: Total de amostras sintéticas a serem geradas
Requer: k: Total de vizinhos mais próximos
Para cada $x_i \in X_{\text{classe minoritária}}$ **faça**
  $S \leftarrow \text{KNN}(x_i, k)$ ▷ Encontra os k vizinhos mais próximos
  Para $n = 1$ até N **faça**
    $x_{ij} \leftarrow \text{AmostraAleatoria}(S)$ ▷ Seleciona um vizinho aleatoriamente
    $\delta \leftarrow \text{Aleatorio}(0, 1)$ ▷ Valor aleatório entre 0 e 1
    $x_{\text{novo}} \leftarrow x_i + \delta \times (x_{ij} - x_i)$ ▷ Gera exemplo sintético
    $X_{\text{sintetico}} \leftarrow X_{\text{sintetico}} \cup \{x_{\text{novo}}\}$
  Fim Para
Fim Para
Garante: $X_{\text{sintetico}}$: Conjunto de amostras sintéticas

Um dos principais problemas do SMOTE é não se basear em fundamentos matemáticos sólidos (76) (77), gerando amostras de forma arbitrária e sem seguir necessariamente a densidade original da classe minoritária (77). Logo, os valores limítrofes entre as classes minoritária e majoritária podem se apresentar de maneira bastante distinta da base original devido às amostras sintetizadas (32). Outra desvantagem do SMOTE é a sintetização de amostras possivelmente ruidosas, isto é, indesejadas, impactando assim na qualidade de análises posteriores (77).

O SMOTE possui variações que foram desenvolvidas a partir de modificações da ideia original do algoritmo. Uma delas é o SMOTE-Tomek, que utiliza o conceito de *tomek links* para identificar e remover, na classe majoritária, as amostras que estão mais

próximas da classe minoritária (19). Outra variação é a *synthetic minority over-sampling technique for time series with efficient kernels* (SMOTE-TEK), que amplia o conceito do SMOTE para a utilização em séries temporais. O *Borderline*-SMOTE tem como objetivo a criação de amostras nas regiões fronteiriças da classe minoritária em relação à majoritária. O SMOTE com *edited nearest neighbor* (ENN), uma variação do algoritmo KNN para identificar os K vizinhos mais próximos de cada observação e verificar se a classe majoritária dentre esses K vizinhos corresponde à classe da observação em questão (91). Há também, entre diversas outras variações, o *safe-level SMOTE*, que preconiza a sintetização a partir das amostras da classe minoritárias que são consideradas "seguras" em algum nível, sendo assim, recomendado principalmente para conjunto de dados em que há uma predominância evidente da classe majoritária (77).

4.5.2 ADASYN

O *adaptive synthetic sampling* (ADASYN) pode ser considerado uma extensão do SMOTE, uma vez que o seu funcionamento é direcionado à sintetização de valores adaptativos, com foco, sobretudo, nas áreas nas quais as instâncias da classe minoritária são mais escassas e mais propensas de serem classificadas de maneira insatisfatória (112).

Para potencializar a eficiência da análise dos dados sobre a classe minoritária, o ADASYN utiliza uma distribuição ponderada para diferentes exemplos dessas classes de acordo com seu nível de dificuldade de aprendizado. Portanto, nos exemplos em que o aprendizado é mais complexo, o algoritmo gera mais dados sintéticos (112).

O Algoritmo 8 exemplifica todas as etapas do funcionamento do algoritmo ADASYN, que é basicamente o mesmo do algoritmo SMOTE. A diferença principal consiste nas etapas iniciais, nas quais para cada valor da classe minoritária, é calculada a proporção, r_i, de valores da classe majoritária entre os k vizinhos mais próximos. Assim, pode ser calculada a distribuição de dificuldade D_i. A seguir, é obtido o total de exemplos sintéticos, G, que devem ser gerados. Por fim, o algoritmo segue os mesmos passos do SMOTE (112), descritos no Algoritmo 7.

Colocando em comparação, SMOTE e ADASYN podem ser utilizados em conjunto ou separadamente, a depender do propósito de suas aplicações e das características principais da distribuição de um conjunto de dados. Diferentes estudos já implementaram ambos os métodos de forma comparativa, sobretudo com relação à qualidade e à eficiência dos resultados obtidos (32). Em geral, a escolha entre SMOTE e ADASYN reside na complexidade do problema e nos ajustes necessários para a sintetização das amostras, de forma a potencializar o desempenho de um modelo em relação à classe minoritária.

Algoritmo 8 Pseudocódigo do algoritmo ADASYN.

Requer: $X_{\text{minoritário}}$: Conjunto de amostras da classe minoritária
Requer: $X_{\text{majoritário}}$: Conjunto de amostras da classe majoritária
Requer: k: Total de vizinhos mais próximos
Requer: β: Nível de balanceamento desejado
$N_{\text{min}} \leftarrow |X_{\text{minoritário}}|$
$N_{\text{maj}} \leftarrow |X_{\text{majoritário}}|$
Para cada $x_i \in X_{\text{minoritário}}$ **faça**
    $S \leftarrow \text{KNN}(x_i, X_{\text{majoritário}}, k)$ ▷ Encontra os k vizinhos majoritários mais próximos
    $r_i \leftarrow \frac{|S|}{k}$ ▷ Proporção da classe majoritária nos vizinhos mais próximos
Fim Para
$D \leftarrow \frac{r_i}{\sum_{i=1}^{N_{\text{min}}} r_i}$ ▷ Normalização da distribuição de dificuldade
$G \leftarrow \beta \times (N_{\text{maj}} - N_{\text{min}})$ ▷ Número total de exemplos sintéticos a serem gerados
Para cada $x_i \in X_{\text{minoritário}}$ **faça**
    $G_i \leftarrow D_i \times G$ ▷ Número de exemplos sintéticos para x_i
    Para $j = 1$ até G_i **faça**
        $x_{ij} \leftarrow \text{AmostraAleatória}(S)$ ▷ Seleciona aleatoriamente um vizinho S
        $\delta \leftarrow \text{Aleatório}(0, 1)$ ▷ Valor aleatório entre 0 e 1
        $x_{\text{novo}} \leftarrow x_i + \delta \times (x_{ij} - x_i)$ ▷ Gera exemplo sintético
        $X_{\text{sintético}} \leftarrow X_{\text{sintético}} \cup \{x_{\text{novo}}\}$
    Fim Para
Fim Para
Garante: $X_{\text{sintético}}$: Conjunto de exemplos sintéticos

5 MATERIAL E MÉTODOS

Neste Capítulo, serão apresentadas e detalhadas a base de dados do Centro Hiperdia de Juiz de Fora e as abordagens desenvolvidas para a aplicação de algoritmos de AM voltados à predição tanto do primeiro quanto do último registro de estágio (PRE e URE) de um paciente. As abordagens foram segmentadas em cenários de aplicação, de modo a possibilitar a comparação entre os métodos, os testes e os respectivos resultados.

5.1 A base de dados

Os dados utilizados nesta tese são oriundos de uma base constituída por registros de pacientes atendidos no CH, do IMEPEN (223), através do SUS, entre agosto de 2010 e dezembro de 2014. A utilização da base de dados para fins de pesquisa foi aprovada pelo Comitê de Ética da UFJF, conforme descrito na Seção 2.4.

A base é formada por 255 campos categóricos ou numéricos, com informações cadastrais (número de identificação, por exemplo), pessoais (idade, sexo, raça, entre outras) e socioeconômicas (renda familiar, nível de escolaridade, entre outras), além de medicamentos, exames clínicos e laboratoriais (creatinina, hemoglobina glicada, glicose de jejum, entre outros) relativos a 7.266 pacientes doentes renais crônicos ou com alguma doença que afeta a função renal, como a diabetes *mellitus* e hipertensão arterial (118) (119). Uma parcela dos campos que compõem a base de dados está exibida na Tabela 2.

Os exames clínicos e laboratoriais possuem majoritariamente dois valores: um relativo ao primeiro registro do paciente na base de dados (denominado, portanto, "inicial") e um relativo ao último registro ("final"). E cada um deles possui um campo referente à data de sua realização. Exclusivamente para a creatinina, há 8 campos de registro, um por semestre entre os anos de 2011 a 2014, como exemplificado na Tabela 3.

5.1.1 Pré-processamento

Conforme descrito na Subseção 2.1, a creatinina possui papel fundamental na determinação da gravidade e da evolução da DRC. Portanto, o primeiro tratamento aplicado aos dados foi a remoção dos pacientes sem nenhum dos oito registros possíveis de creatinina, restando um total igual a 5.689, número também obtido por Moraes Junior (2019) (165) (166). Cada um dos pacientes representa uma linha da base de dados.

O segundo tratamento aplicado foi o cálculo do valor da TFG a partir dos dados disponíveis para cada paciente. Embora a base possua originalmente oito campos para o valor da TFG e o estágio correspondente (1 por semestre de 2011 a 2014), o cálculo dos valores não necessariamente considerou ponderadores importantes que compõem a Equação MDRD (144), descrita na Subseção 2.1: a raça (pacientes declarados como "pretos") e o sexo do indivíduo (feminino). Portanto, a equação foi utilizada para o cálculo dos novos

Tabela 2 – Uma parcela dos campos disponíveis na base de dados correspondente a informações cadastrais, pessoais, socioeconômicas, ambulatoriais e clínicas, além de dados relacionados à DRC. Para todos os exames clínicos, existem campos correspondentes ao valor inicial registrado e ao valor final, ambos acompanhados de campos referentes à data de realização de cada exame.

Tipo de informação	**Campos**
Cadastro	Id
Pessoais	Alt, classe_imc, Codsexo, DN, etilismo, idade, imc, peso, Raça, sedentario e tabagismo
Socioeconômicas	CodCid, CodUBS, instruc, RendaSM, RendaFamiliarSM e TamFamilia
Acompanhamento	datainicial, datafinal, desfecho, retorno e tempoAcompanha
Medicamentos	AAS8, BETABLOQ3, BIGUADINA6, BRAT2, DIUR4, estatina9, FIBRATO13, IECA1, insulina e SULFONIURA7
Ambulatoriais	consultasDM, consultasDRC, consultasHAS, DRC_1_2011, DRC_2_2011, DRC_1_2012, DRC_2_2012, DRC_1_2013, DRC_2_2013, DRC_1_2014, DRC_2_2014, HAS_1_2011, HAS_2_2011, HAS_1_2012, HAS_2_2012, HAS_1_2013, HAS_2_2013, HAS_1_2014, HAS_2_2014, DM_1_2011, DM_2_2011, DM_1_2012, DM_2_2012, DM_1_2013, DM_2_2013, DM_1_2014 e DM_2_2014
Clínicas	AcidoFolico, AcidoUrico, Albumina, AntiHBs, AntiHCV, Bilirrubinatotal, CalcioTotal, CK, Estagio, ColesterolHDL, ColesterolLDL, ColesterolTotal, ECOAE, ECOAO, ECOFE, ECOPP, ECOSIV, Ferritina, FerroSerico, FosfataseAlcalina, Fosforo, GamaGlutamil, GlicemiadeJejum, HBsAG, Hematuria, Hemoglobina, HemoglobinaGlicada, IndicedeSaturacaodaTransferencia, Microalbuminuria, PAD, TSH, PAS, Potassio, Proteinuria24hs, PTHintacto, Rel.AlbuminaCreatininaUAUC, SodioSerico, SodioUrinario, PAS, TGP, Triglicerides, Ureia, Ureia24hs, VITAMINAD e VitaminaB12
DRC	Creatinina_1_2011, Creatinina_1_2012, Creatinina_1_2013, Creatinina_1_2014, Creatinina_2_2011, Creatinina_2_2012, Creatinina_2_2013, Creatinina_2_2014, Estágio_1_2011, Estágio_1_2012, Estágio_1_2013, Estágio_1_2014, Estágio_2_2011, Estágio_2_2012, Estágio_2_2013, Estágio_2_2014, TFG_1_2011, TFG_1_2012, TFG_1_2014, TFG_2_2011, TFG_2_2012, TFG_2_2013, TFG_2_2014, CreatininaI, CreatininaF, EstágioI e EstágioF

Tabela 3 – Trecho da base de dados exemplificando a disposição dos campos e do conteúdo.

ID	**Idade**	**Sexo**	**...**	**Cr. 2011-1**	**...**	**Cr. 2014-2**	**...**	**EstagioF**	**...**
27	74	Masc.	...	1,40	...	Nulo	...	2	...
35	50	Fem.	...	Nulo	...	Nulo	...	3a	...
40	78	Fem.	...	2,50	...	1,68	...	3b	...
42	102	Masc.	...	1,32	...	Nulo	...	3a	...
44	83	Masc.	...	2,50	...	Nulo	...	4	...
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
9835	60	Fem.	...	Nulo	...	1,1	...	2	...

valores de TFG e, consequentemente, para a obtenção das novas informações de estágio para cada paciente. A base de dados resultante dos tratamentos descritos constituiu a configuração inicial adotada em todos os cenários propostos e detalhados nesta tese. Todas as propostas de reorganização, que serão apresentadas neste Capítulo, são derivações da base de dados original, obtida após o processo de pré-processamento. A distribuição de pacientes com valores de TFG registrados por semestre no quadriênio 2011-2014 está descrita na Tabela 4, na qual também constam os totais de pacientes por estágio da DRC.

Tabela 4 – Total de pacientes por estágio com valores de TFG em cada um dos semestres de 2011 a 2014. Os valores circulados representam o maior total de registros de pacientes por semestre.

TFG / **Estágio**	**2011/1**	**2011/2**	**2012/1**	**2012/2**	**2013/1**	**2013/2**	**2014/1**	**2014/2**
1	148	⓪472	196	183	208	216	436	531
2	⓪368	406	⓪553	⓪530	⓪592	⓪672	⓪744	⓪638
3a	341	295	401	430	482	500	496	323
3b	222	199	309	351	410	425	376	315
4	161	178	182	193	209	231	212	190
5	43	54	53	37	59	49	46	41
Total	**1283**	**1604**	**1694**	**1724**	**1960**	**2093**	**2310**	**2038**

Finalmente, uma vez que a aplicação dos métodos de AM nesta análise visa à classificação do URE dos pacientes considerados, em todos os cenários de testes realizados as variáveis categóricas foram convertidas em valores numéricos. Do total de 255 campos, 114 são do tipo categórico, sendo a maior parcela formada por informações de data referentes à realização dos exames. Visto que para as classificações que serão apresentadas nessa tese foram considerados conjuntos de exames com no máximo 50 campos, somente as variáveis categóricas que fizeram parte dos testes foram convertidas para valores numéricos. A Tabela 5 fornece um exemplo da conversão de alguns campos: `Codsexo` (variável referente ao sexo do paciente e com dois valores possíveis), `Raça` (cinco valores possíveis) e `Estágio` (seis valores possíveis).

Tabela 5 – Exemplos da conversão de três campos categóricos para valores numéricos.

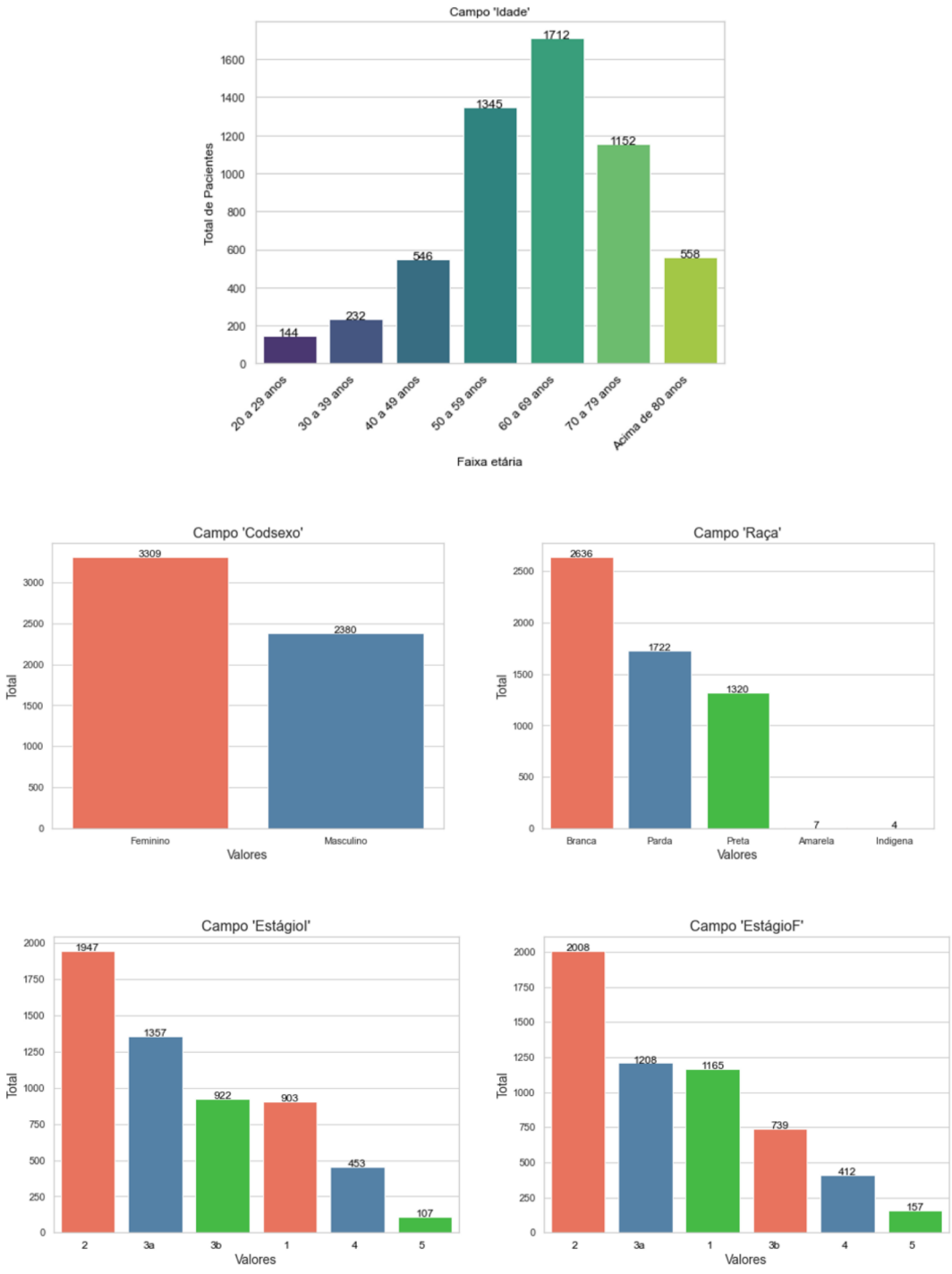
Campo	Valores originais	Valores convertidos
Codsexo	{Feminino, Masculino}	{1, 2}
Raça	{Branca, Parda, Preta, Amarela, Indígena}	{1, 2, 3, 4, 5}
Estágio	{Estágio 1 - >= 90 ml, Estágio 2 - 60-89 ml, Estágio 3a - 45-59 ml, Estágio 3b - 30-44 ml, Estágio 4 - 15-29 ml, Estágio 5 - < 15 ml}	{0, 1, 2, 3, 4, 5}

Os valores mínimo, máximo e da média de alguns campos da base de dados estão disponíveis na Tabela 6 e nas Figuras 17 e 18.

Tabela 6 – Dados estatísticos sobre alguns campos da base de dados. Cada campo se refere, respectivamente, aos dados de idade, peso inicial, peso final e altura; total de consultas nos ambulatórios de DM, DRC e HAS; tamanho da família e a renda familiar em função do total de salários mínimos; valores inicial e final de creatinina; e o valor de TFG por semestre entre 2011 e 2014.

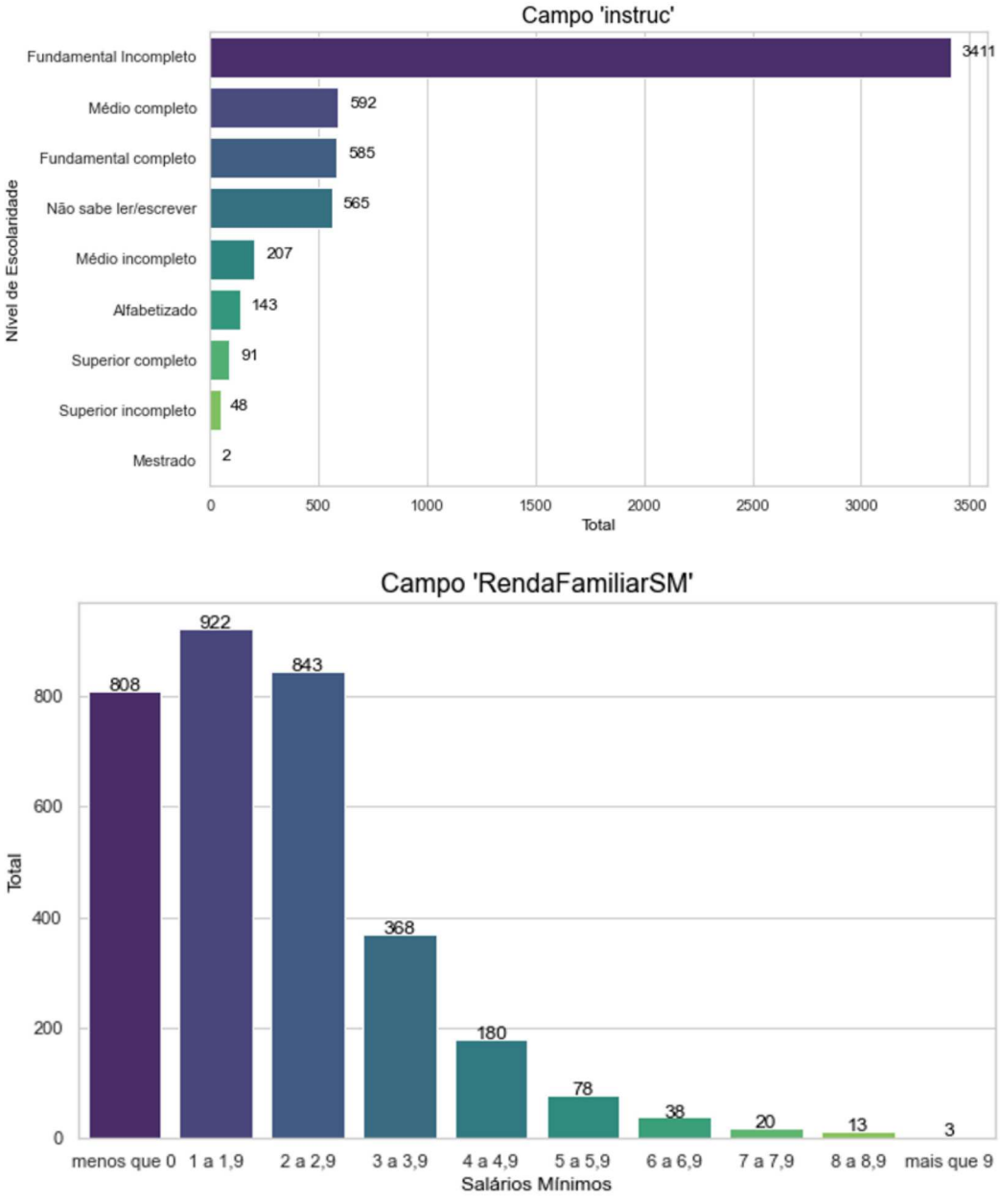
Campo	Mínimo	Máximo	Média	Total
Idade	4	102	61,96	5.689
pesoi	14,05	243,75	77,00	5.639
pesof	14,05	240,45	77,40	5.689
Alt	1,11	199,00	159,88	5.584
consultasDM	1	124	2,85	3.002
consultasDRC	1	20	2,82	2.396
consultasHAS	1	26	2,90	2.329
TamFamilia	1	18	3,00	3.295
RendaFamiliarSM	0	23.9	1,94	3.291
CREATININAI	0,02	230,01	1,52	5.689
CREATININAF	0,02	230,01	1,44	5.689
TFG_1_2011	0,25	189,54	53,52	1.283
TFG_2_2011	0,18	236,12	55,24	1.604
TFG_1_2012	0,13	205,53	54,87	1.694
TFG_2_2012	0,19	342,03	54,63	1.724
TFG_1_2013	2,26	826,76	54,84	1.960
TFG_2_2013	6,04	868,73	56,25	2.093
TFG_1_2014	0,16	298,57	61,98	2.310
TFG_2_2014	0,37	5607,59	70,14	2.038

Figura 17 – Total de pacientes em cada classe, por exame, distribuídos conforme os estágios.



Fonte: Elaborada pelo autor.

Figura 18 – O primeiro gráfico exibe os diferentes graus de instrução dos pacientes da base de dados. Já o segundo exibe a renda familiar em função do valor de salário mínimo.



Fonte: Elaborada pelo autor.

5.1.2 Reorganização da base de dados

Com o objetivo de proporcionar diferentes perspectivas acerca da organização da base de dados, foram implementadas duas propostas de reorganização, que serão utilizadas em parte dos testes descritos nesta tese.

5.1.2.1 Em função das datas

Para o desenvolvimento dos dois primeiros cenários de aplicação de algoritmos de AM na base de dados, esta foi reorganizada de forma a ser obtido um entendimento temporal dos dados clínicos e laboratoriais, isto é, uma visão da evolução da condição clínica do paciente ao longo do seu período de tratamento no CH.

Em substituição à organização inicial da base tendo um paciente por linha (Tabela 3), a nova estruturação proposta considerou uma linha para cada exame realizado por paciente entre 2010 e 2014, a partir das datas de realização. Consequentemente, todos os exames aplicados a um paciente em uma data comum passaram a ser dispostos em uma mesma linha. Dados considerados inicialmente imutáveis, como data de nascimento, número de registro e raça, foram replicados em todas as linhas, como exemplificado na Tabela 7. Em decorrência dessa nova organização, o total de registros (linhas) da base de dados aumentou de 5.689 (um por paciente) para 40.100.

Tabela 7 – Trecho da base de dados exemplificando a reorganização dos dados em função das datas de realização dos exames laboratoriais e clínicos.

Data	**ID**	**Sexo**	**...**	**Colest. HDLI**	**...**	**Triglic. I**	**...**	**PotassioI**	**...**
03/08/10	27	Masc.	...	Nulo	...	Nulo	...	Nulo	...
20/08/10	27	Masc.	...	47	...	294	...	4,8	...
08/02/11	27	Masc.	...	Nulo	...	Nulo	...	Nulo	...
16/06/11	27	Masc.	...	Nulo	...	Nulo	...	Nulo	...
21/09/12	27	Masc.	...	Nulo	...	Nulo	...	Nulo	...
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
26/08/14	27	Masc.	...	Nulo	...	Nulo	...	Nulo	...
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
22/07/10	40	Fem.	...	36	...	153	...	4,7	...
09/08/10	40	Fem.	...	Nulo	...	Nulo	...	Nulo	...
17/09/10	40	Fem.	...	Nulo	...	Nulo	...	Nulo	...
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
10/11/14	40	Fem.	...	Nulo	...	Nulo	...	Nulo	...
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

5.1.2.2 Classes

Uma segunda proposta de reorganização da base de dados considerou uma estratégia para a redução da complexidade e a mitigação de questões problemáticas associadas à distribuição dos dados, a qual nem sempre é representativa do estágio e da evolução clínica de um paciente a partir dos valores dos exames feitos durante o tratamento no CH. Dessa forma e com base em valores clínicos e laboratoriais de referência (98), os dados foram separados segundo uma divisão em três classes: valores inferiores ao esperado (classe 1), valores correspondentes ao esperado (classe 2) e valores superiores ao esperado (classe 3). Com essa definição, uma transição de classe de um paciente ao longo do tratamento poderia indicar, de maneira mais clara, a alteração em seu estado clínico.

Com o objetivo de identificar o melhor conjunto de exames para caracterizar os estágios da DRC, para a divisão em classes foram selecionados apenas os valores de exames relativos à primeira coleta, garantindo que os pacientes estivessem em fases iniciais do tratamento. Para minimizar a interferência de medicamentos no acompanhamento da doença, foram considerados apenas os exames realizados no mesmo semestre do início do tratamento. Contudo, como não é obrigatório que todos os exames contenham valores associados, foi necessário lidar com a ausência de dados. Uma abordagem adotada foi selecionar apenas pacientes atendidos no ambulatório de DRC ao menos duas vezes, a fim de que fossem identificados os exames mais relevantes, o que reduziu o conjunto de dados de 5.689 para 1.427 pacientes (98).

Pela observação dos dados expressos nas tabelas na Figura 19, é possível notar, por exemplo, que a distribuição entre as classes do exame inicial de creatinina (`CREATININAI`) passou por alterações ao longo da evolução dos estágios da DRC. Pode ser observada a prevalência integral de valores das classes 1 e 2 no primeiro estágio, que se refere ao início do tratamento, quando as funções renais ainda não estão gravemente comprometidas. Com o avanço para o estágio 2, os valores predominantes passam a pertencer integralmente à classe 2, e a partir do estágio 3a, a classe 3 se torna predominante, indicando maior comprometimento renal. Para a maioria dos exames é esperado que os valores normais da classe 2 evoluam para as classes 1 e 3 com a progressão da gravidade da DRC, reforçando a expectativa de deterioração dos resultados.

Por meio da análise das tabelas da Figura 19, pode ser observado que, além do valor inicial de creatinina, os seguintes exames também exibiram alterações na quantidade de valores em cada uma das classes, evidenciando uma progressão conforme o agravamento do estágio da DRC: `GlicemiadeJejumI`, `ColesterolHDLI`, `AcidoUricoI`, `TriglicerídesI`, `UreiaI`, `Proteinuria24hsI` e `HemoglobinaI`. E para evitar que possíveis impactos na classificação possam acontecer devido à grande quantidade de dados ausentes — como evidencia a Figura 20 — uma outra estratégia considerou apenas os pacientes com 90% dos dados registrados para os valores iniciais de creatinina e de todos os sete exames citados.

Tabela 8 – Classificação dos intervalos de valores de referência para cada exame (25) (26) (27) (60) (74) (104) (117) (159) (179) (185) (250) (251) (262). Adaptado de (98).

Exame	**Classe 1**	**Classe 2**	**Classe 3**
Ácido Úrico (U/L)	Homens $<$ 3,4 Mulheres $<$ 2,4 Idosos $<$ 2,9	3,4 $\leq$ Homens $\leq$ 7 2,4 $\leq$ Mulheres $\leq$ 6 2,9 $\leq$ Idosos $\leq$ 6	Homens $>$ 7 Mulheres $>$ 6 Idosos $>$ 6
Cálcio Total (mg/dL)	$valor <$ 8,8	$8,8 \leq valor \leq 10,4$	$valor >$ 10,4
Colesterol HDL (mg/dL)	$35 \leq valor <$ 50	$valor \geq$ 50	$valor <$ 35
Colesterol Total (mg/dL)	$200 \leq valor <$ 240	$valor <$ 200	$valor \geq$ 240
Creatinina (mg/dL)	Homens $\leq$ 0,7 Mulheres $\leq$ 0,6	0,7 $<$ Homens $\leq$ 1,3 0,6 $<$ Mulheres $\leq$ 1,2	Homens $>$ 1,3 Mulheres $>$ 1,2
Fósforo (mg/dL)	Homens $<$ 2,4 Mulheres $<$ 3,8	2,4 $\leq$ Homens $\leq$ 4,6 0,6 $\leq$ Mulheres $\leq$ 1,2	Homens $>$ 1,3 Mulheres $>$ 1,2
Glicemia de Jejum (mg/dL)	$valor \leq$ 70	70 $<$ $valor \leq$ 99	$valor >$ 99
Hemoglobina (g/dL)	Homens $\leq$ 14 Mulheres $\leq$ 12	14 $<$ Homens $\leq$ 18 12 $<$ Mulheres $\leq$ 16	Homens $>$ 18 Mulheres $>$ 16
Hemoglobina Glicada (%)	$valor <$ 4,7	$4,7 \leq valor \leq 9,9$	$valor >$ 9,9
PAD (mmHg)	$valor <$ 80	80 $\leq$ $valor <$ 110	$valor \geq$ 110
PAS (mmHg)	$valor <$ 130	130 $\leq$ $valor <$ 180	$valor \geq$ 180
Potássio (mg/dL)	$valor \leq$ 3,5	3,5 $<$ $valor \leq 5,2$	$valor >$ 5,2
Proteinúria 24h (mg/24h)	150 $<$ $valor \leq$ 300	$valor \leq$ 150	$valor >$ 300
Sódio Sérico (mEq/L)	$valor <$ 138	138 $\leq$ $valor \leq$ 142	$valor >$ 142
Sódio Uninário (mEq/L)	$valor <$ 25	Homens - 40 anos 25 $\leq$ $valor \leq$ 301	$valor >$ 301
	$valor <$ 18	Homens + 40 anos 18 $\leq$ $valor \leq$ 214	$valor >$ 214
	$valor <$ 15	Mulheres - 40 anos 15 $\leq$ $valor \leq$ 267	$valor >$ 267
	$valor <$ 15	Mulheres + 40 anos 15 $\leq$ $valor \leq$ 237	$valor >$ 237
TGP (U/L)	$valor <$ 7	7 $\leq$ $valor \leq$ 56	$valor >$ 56
Triglicérides (mg/dL)	$valor \leq$ 50	50 $<$ $valor <$ 200	$valor \geq$ 200
TSH (μUI/mL)	$valor \leq$ 0,3	0,3 $<$ $valor \leq$ 4	$valor >$ 4
Ureia (mg/dL)	$valor \leq$ 13	13 $<$ $valor \leq$ 43	$valor >$ 43

Figura 19 – Total de pacientes em cada classe, por exame, distribuídos conforme os estágios.

a) Estágio 1

	Classe 1(%)	Classe 2(%)	Classe 3(%)	Total por estágio
TSHI	7.14	64.29	28.57	14
Hemoglobinal	21.74	78.26	0.0	23
AcidoUricol	5.0	70.0	25.0	20
CalcioTotall	9.09	90.91	0.0	11
SodioUrinariol	0.0	66.67	33.33	12
ColesterolHDLl	43.48	47.83	8.7	23
ColesterolTotall	21.74	60.87	17.39	23
HemoglobinaGlicadal	0.0	88.89	11.11	9
TGPl	0.0	100.0	0.0	8
Trigliceridesl	8.7	69.57	21.74	23
Potassiol	0.0	85.71	14.29	21
GlicemiadeJejuml	4.35	52.17	43.48	23
Fosforol	44.44	55.56	0.0	9
Proteinuria24hsl	14.29	78.57	7.14	14
SodioSericol	20.0	60.0	20.0	15
Ureial	0.0	80.0	20.0	15
PAS_inicial	40.0	60.0	0.0	15
PAD_inicial	40.0	60.0	0.0	15
CREATININAl	52.17	47.83	0.0	23

b) Estágio 2

	Classe 1(%)	Classe 2(%)	Classe 3(%)	Total por estágio
TSHI	3.51	84.21	12.28	57
Hemoglobinal	27.5	70.0	2.5	80
AcidoUricol	4.35	68.12	27.54	69
CalcioTotall	8.62	75.86	15.52	58
SodioUrinariol	0.0	71.15	28.85	52
ColesterolHDLl	45.0	38.75	16.25	80
ColesterolTotall	34.18	48.1	17.72	79
HemoglobinaGlicadal	4.0	70.0	26.0	50
TGPl	0.0	97.96	2.04	49
Trigliceridesl	1.25	75.0	23.75	80
Potassiol	2.82	92.96	4.23	71
GlicemiadeJejuml	0.0	36.25	63.75	80
Fosforol	20.37	72.22	7.41	54
Proteinuria24hsl	20.69	58.62	20.69	58
SodioSericol	18.97	62.07	18.97	58
Ureial	1.67	71.67	26.67	60
PAS_inicial	21.05	71.93	7.02	57
PAD_inicial	21.05	71.93	7.02	57
CREATININAl	0.0	100.0	0.0	80

c) Estágio 3a

	Classe 1(%)	Classe 2(%)	Classe 3(%)	Total por estágio
TSHI	2.52	85.71	11.76	119
Hemoglobinal	31.03	67.49	1.48	203
AcidoUricol	1.71	53.14	45.14	175
CalcioTotall	10.24	76.38	13.39	127
SodioUrinariol	0.0	66.13	33.87	124
ColesterolHDLl	47.78	36.45	15.76	203
ColesterolTotall	29.35	54.73	15.92	201
HemoglobinaGlicadal	2.59	82.76	14.66	116
TGPl	0.77	98.46	0.77	130
Trigliceridesl	0.99	71.92	27.09	203
Potassiol	3.26	83.7	13.04	184
GlicemiadeJejuml	1.48	31.53	67.0	203
Fosforol	20.0	75.65	4.35	115
Proteinuria24hsl	20.0	59.26	20.74	135
SodioSericol	22.05	47.24	30.71	127
Ureial	0.74	52.94	46.32	136
PAS_inicial	17.27	75.54	7.19	139
PAD_inicial	17.27	75.54	7.19	139
CREATININAl	0.0	41.38	58.62	203

d) Estágio 3b

	Classe 1(%)	Classe 2(%)	Classe 3(%)	Total por estágio
TSHI	1.8	69.46	28.74	167
Hemoglobinal	44.81	54.44	0.74	270
AcidoUricol	1.29	42.92	55.79	233
CalcioTotall	10.11	71.81	18.09	188
SodioUrinariol	0.0	82.86	17.14	140
ColesterolHDLl	46.3	37.04	16.67	270
ColesterolTotall	26.42	60.0	13.58	265
HemoglobinaGlicadal	2.61	81.05	16.34	153
TGPl	1.15	96.55	2.3	174
Trigliceridesl	2.22	73.33	24.44	270
Potassiol	1.62	83.0	15.38	247
GlicemiadeJejuml	3.33	45.19	51.48	270
Fosforol	27.96	66.67	5.38	186
Proteinuria24hsl	23.67	51.48	24.85	169
SodioSericol	28.26	44.57	27.17	184
Ureial	0.0	23.88	76.12	201
PAS_inicial	20.57	70.86	8.57	175
PAD_inicial	20.57	70.86	8.57	175
CREATININAl	0.0	0.37	99.63	270

e) Estágio 4

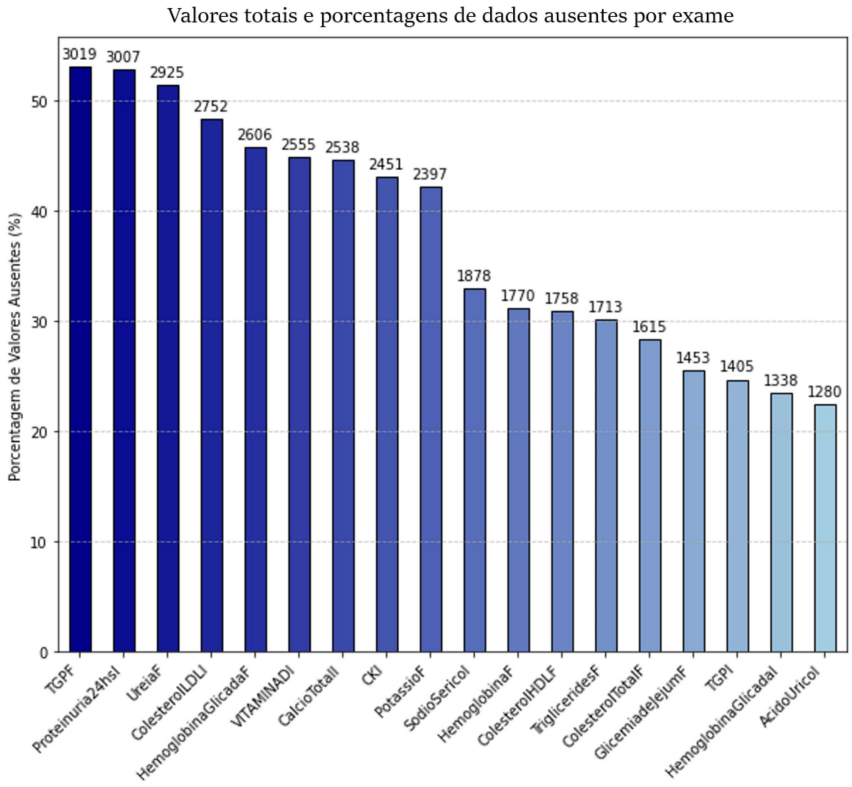
	Classe 1(%)	Classe 2(%)	Classe 3(%)	Total por estágio
TSHI	4.08	76.53	19.39	98
Hemoglobinal	63.78	35.68	0.54	185
AcidoUricol	1.23	40.12	58.64	162
CalcioTotall	13.99	73.43	12.59	143
SodioUrinariol	0.0	81.48	18.52	108
ColesterolHDLl	45.95	35.14	18.92	185
ColesterolTotall	15.38	64.29	20.33	182
HemoglobinaGlicadal	7.92	81.19	10.89	101
TGPl	4.39	94.74	0.88	114
Trigliceridesl	2.16	72.43	25.41	185
Potassiol	1.73	68.79	29.48	173
GlicemiadeJejuml	7.57	41.08	51.35	185
Fosforol	26.21	63.45	10.34	145
Proteinuria24hsl	17.07	38.21	44.72	123
SodioSericol	28.46	47.69	23.85	130
Ureial	0.7	3.5	95.8	143
PAS_inicial	27.5	64.17	8.33	120
PAD_inicial	27.5	64.17	8.33	120
CREATININAl	0.0	0.0	100.0	185

f) Estágio 5

	Classe 1(%)	Classe 2(%)	Classe 3(%)	Total por estágio
TSHI	13.33	66.67	20.0	15
Hemoglobinal	75.76	24.24	0.0	33
AcidoUricol	0.0	48.15	51.85	27
CalcioTotall	30.77	69.23	0.0	26
SodioUrinariol	0.0	85.0	15.0	20
ColesterolHDLl	48.48	27.27	24.24	33
ColesterolTotall	36.36	48.48	15.15	33
HemoglobinaGlicadal	0.0	94.12	5.88	17
TGPl	14.29	85.71	0.0	21
Trigliceridesl	3.03	57.58	39.39	33
Potassiol	0.0	62.5	37.5	32
GlicemiadeJejuml	6.06	51.52	42.42	33
Fosforol	20.0	56.0	24.0	25
Proteinuria24hsl	8.7	21.74	69.57	23
SodioSericol	12.5	50.0	37.5	24
Ureial	0.0	4.55	95.45	22
PAS_inicial	33.33	66.67	0.0	15
PAD_inicial	33.33	66.67	0.0	15
CREATININAl	0.0	0.0	100.0	33

Fonte: Adaptado de (127).

Figura 20 – Porcentagem de valores ausentes para alguns dos principais exames da base de dados.



Fonte: Elaborada pelo autor.

Como resultado da estratégia, o total de pacientes novamente foi reduzido, de 1.427 para 850. Adicionalmente, todos os exames com mais de 40% dos valores ausentes foram retirados, processo que resultou na permanência de 19 exames, que estão representados nas tabelas da Figura 19: `AcidoUricoI`, `CalcioTotalI`, `CREATININAI`, `ColesterolHDLI`, `ColesterolTotalI`, `FosforoI`, `GlicemiadeJejumI`, `HemoglobinaGlicadaI`, `HemoglobinaI`, `PAD_inicial`, `PAS_inicial`, `PotassioI`, `ProteinuriaI24hsI`, `SodioSericoI`, `SodioUrinarioI`, `TGPI`, `TSHI`, `TriglicerideSI` e `UreiaI` (127).

A porcentagem de valores ausentes para cada um dos 19 exames pode ser vista na Tabela 9, que evidencia ainda um volume considerável de dados faltantes na base de dados. Tomando como referência a avaliação da distribuição dos valores dos 19 exames e a comparação com os padrões clínicos estabelecidos, foram removidos os valores considerados atípicos em cada exame mencionado, resultando em um total final de 794 registros de pacientes. Os valores utilizados como referência para cada exame foram (98):

- `ColesterolTotalI` $\geq$ 370 mg/dL;
- `CREATININAI` $\geq$ 10 mg/dL;
- `GlicemiadeJejumI` $\geq$ 500 mg/dL;

- `Proteinuria24hsI` $\geq$ 5.000 mg/24h;
- `SodioUrinarioI` $\geq$ 400 mEq/L;
- `TGPI` $\geq$ 74 U/L;
- `TSHI` $\geq$ 12 μUI/mL;
- `UreiaI` $\geq$ 200 mg/dL.

Exame	**Percentual (%)**
`HemoglobinaGlicadaI`	25,6
`SodioUrinarioI`	25,2
`TSHI`	23,6
`TGPI`	21,9
`Proteinuria24hsI`	20,3
`FosforoI`	19,4
`SodioSericoI`	19,1
`CalcioTotalI`	18,0
`UreiaI`	16,4
`AcidoUricoI`	8,0
`PotassioI`	5,0
`ColesterolTotalI`	0,8
`PAS_inicial`	0,6
`PAD_inicial`	0,6
`ColesterolHDLI`	0,0
`CreatininaI`	0,0
`GlicemiadeJejumI`	0,0
`HemoglobinaI`	0,0
`TrigliceridesI`	0,0

Tabela 9 – Porcentagem de valores ausentes por exame. Adaptado de (98).

5.2 Configurações para os testes

Em todos os cenários e testes descritos nesta tese, foi utilizada a linguagem de programação `Python` (198), nas versões 3.7 e 3.8, juntamente com diversas de suas bibliotecas para leitura e tratamento de dados, implementação de algoritmos, geração de gráficos, entre outros propósitos. As principais bibliotecas empregadas incluem: `copulas` (37), `fancyimpute` (211), `imblearn` (141), `Matplotlib` (122), `NumPy` (107), `Pandas` (162), `Scikit-learn` (188) e `Seaborn` (261).

E todas as simulações foram realizadas em uma máquina equipada com processador *Intel Core* i7-10510U, com uma de CPU 1.80-2.30 *gigahertz*, com 16 *gigabytes* de memória RAM e sistema operacional Windows, versão 11 *Home Single Language* 23H2.

6 CENÁRIOS, RESULTADOS E DISCUSSÃO

Seguindo os objetivos gerais e específicos delineados neste estudo, foram elaborados diferentes cenários de aplicação de algoritmos de aprendizado de máquina para a predição dos estágios da doença renal crônica utilizando o conjunto de dados selecionado. A criação de cada cenário visou não apenas a resposta aos objetivos mencionados, mas também o desenvolvimento de abordagens que possam contribuir para a literatura sobre a predição da DRC com o auxílio de métodos de AM. Essa é uma importante contribuição para o auxílio à tomada de decisões estratégicas por gestores e profissionais de saúde, aprimorando o controle de qualidade no diagnóstico e no tratamento da DRC. E mesmo nos casos em que os resultados não atingiram as expectativas, as abordagens desenvolvidas podem contribuir para a discussão sobre as estratégias mais adequadas a serem adotadas, bem como para a identificação das limitações inerentes à base de dados utilizada nesta tese.

Para alcançar esses objetivos, o conjunto de dados foi organizado de diversas maneiras, múltiplos algoritmos foram implementados, diferentes métodos de inferência de dados faltantes foram empregados, os atributos da base de dados foram ajustados conforme cada objetivo e variadas métricas de avaliação foram aplicadas para a apresentação e discussão dos resultados.

6.1 Cenário 1

6.1.1 Definição

Para a predição do URE da DRC na base de dados para cada paciente, o primeiro cenário de aplicação [1] considerou a configuração da base de dados em função das datas de realização dos exames laboratoriais e clínicos de cada paciente, como evidencia a Tabela 7, permitindo a compreensão cronológica da evolução da condição clínica do paciente ao longo do seu tratamento. É importante destacar que o URE corresponde ao último registro de estágio da DRC presente na base de dados para um dado paciente. Assim, o URE pode se referir a qualquer um dos seis estágios da doença: 1, 2, 3a, 3b, 4 ou 5.

Conforme detalhado no Capítulo 3, o algoritmo supervisionado RF foi utilizado por uma relevante parte dos estudos revisados nesta tese (10) (65) (82) (96) (105) (126) e (199), e com resultados satisfatórios. Adicionalmente, o RF foi apontado nos levantamentos realizados pelas duas revisões sistemáticas (135) (215) como um dos métodos mais utilizados na literatura da aplicação de algoritmos de AM na predição da progressão da DRC.

No cenário 1, foram propostas quatro abordagens distintas para o agrupamento de 50 das variáveis disponíveis na base, as quais estão especificadas na Tabela 10. Cada uma das abordagens buscou a elaboração de testes distintos envolvendo diferentes totais

[1] O estudo proposto no cenário 1 e os seus resultados foram publicados em 2021 no volume 1355, páginas 255-265, da *Advances in Intelligent Systems and Computing* (221).

de variáveis para a predição do URE, de forma que fosse desenvolvida uma compreensão preliminar dos resultados nos processos de classificação aplicados à base de dados do CH.

Como detalhado na Tabela 10, as variáveis preditoras selecionadas são compostas pelos quatro componentes principais de determinação da TFG pela Equação 2.2 (MDRD): creatinina sérica, idade, raça e sexo; pelas variáveis referentes às frequências semestrais dos pacientes nos ambulatórios de DM, DRC e HAS; e pelos exames clínicos que exercem impacto direto na evolução dos pacientes (10) (65) (96) (201), como ácido úrico, colesterol, glicemia, hemoglobina, hemoglobina glicada, potássio, pressão arterial, transaminase glutâmico pirúvica (TGP), hormônio tireoestimulante (TSH, do inglês *thyroid stimulating hormone*) e ureia, além do peso.

As variáveis preditoras numeradas de 1 a 26 na Tabela 10 também foram selecionadas em função de suas frequências na base de dados, já que correspondem aos 26 campos com o maior número de registros individuais. Neste contexto, embora a albuminúria seja um dos dois principais marcadores para a avaliação de risco da DRC, conforme explicado na Subseção 2.2, seus registros iniciais e finais não foram incluídos nos processos de classificação em nenhum dos cenários desenvolvidos nesta tese. Tal exclusão deve-se, principalmente, à disponibilidade extremamente reduzida desses valores na base de dados, com mais de 97% de registros ausentes em cada caso. As quatro abordagens estão detalhadas a seguir:

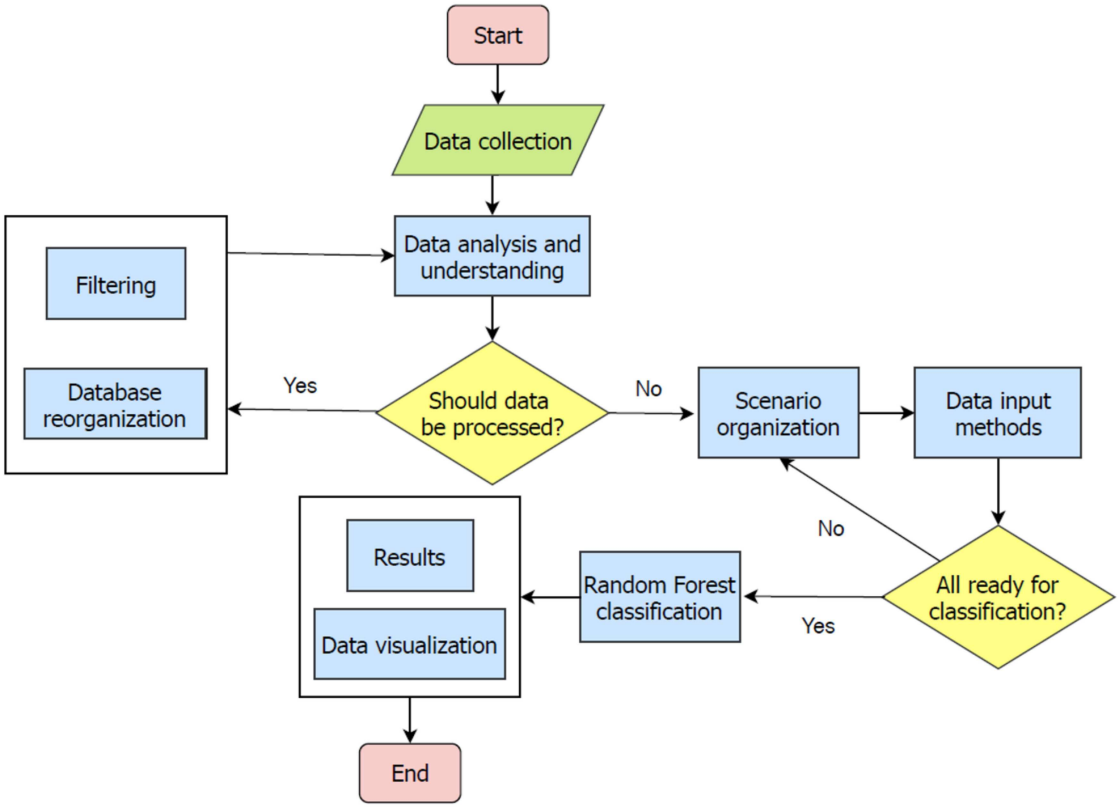
- **Abordagem 1**: foram selecionadas como variáveis preditoras somente os quatro valores determinantes para o cálculo da TFG: o primeiro registro de creatinina sérica (PRCS), idade, raça e sexo, respectivamente, as variáveis 1, 2, 3 e 4 da Tabela 10.
- **Abordagem 2**: tomando como referência o total de atributos da base de dados da UCI (210), utilizada em estudos como os de Ogunleye e Wang (2020) (180) e Rady *et al.* (2019) (201), foram adotados os 25 dados com mais registros de pacientes na base: variáveis 1, 2, 3 e de 5 a 26 da Tabela 10. São cinco dados pessoais — idade, peso inicial, peso final, raça e sexo — e 20 valores de exames clínicos e laboratoriais. Este é a única das abordagens do cenário 1 a não incluir o PRCS.
- **Abordagem 3**: foram contabilizadas as visitas semestrais dos pacientes a cada um dos três ambulatórios do CH no quadriênio 2011-2014: diabetes *mellitus* (DM), doença renal crônica (DRC) e hipertensão arterial sistêmica (HAS), correspondentes às variáveis 27 a 50 da Tabela 10, identificadas por um dígito referente ao semestre, seguido por quatro dígitos referentes ao ano. O PRCS também foi incluído nesta abordagem.
- **Abordagem 4**: a combinação das abordagens 2 e 3 resultou em todas as 50 variáveis preditoras da Tabela 10: os 25 dados e exames mais frequentes, o PRCS e a contabilização das visitas semestrais aos três ambulatórios.

A Tabela 10 é formada por 50 variáveis preditoras e 1 variável alvo comum a todos os cenários: o URE. Para cada atributo, é especificado o tipo de dado — numérico ou categórico — e o total de registros ausentes na base. Esse fator constitui um dos principais desafios da base de dados, devido à elevada quantidade de informações ausentes, especialmente relacionadas aos registros dos exames clínicos e laboratoriais dos pacientes ao longo do período de tratamento no CH, cujo processo, à época, era realizado de forma manual. Portanto, é essencial que diferentes estratégias para o tratamento de dados faltantes sejam cuidadosamente analisadas e implementadas.

Segundo as definições de Azur e colaboradores (14) e de Pedersen e colaboradores (187), a base de dados utilizada neste trabalho pode ser caracterizada como do tipo *missing at random* (MAR), uma vez que a ausência de registros não é totalmente aleatória. Uma importante parcela dos dados faltantes pode ser explicada por outras variáveis, como a condição clínica do paciente e os exames de realização fundamental em cada etapa da progressão e do tratamento da doença renal crônica nos diferentes ambulatórios de atendimento.

O fluxograma com todas as etapas propostas no cenário 1 está ilustrado na Figura 21.

Figura 21 – Fluxograma com as etapas do cenário 1.



Fonte: Extraído de (221).

Tabela 10 – Descrição das variáveis utilizadas nas diferentes abordagens do cenário 1. Adaptado de (221).

№	Valor	Variáveis	Classe	Tipo	Registros Ausentes
1	-	Idade	Preditora	Numérica	0
2	-	Raça	Preditora	Categórica	0
3	-	Sexo	Preditora	Categórica	0
4	Primeiro	Creatinina Sérica	Preditora	Categórica	0
5	Inicial	P. A. Sistêmica	Preditora	Numérica	21
6	Final	P. A. Sistêmica	Preditora	Numérica	21
7	Inicial	P. A. Diastólica	Preditora	Numérica	21
8	Final	P. A. Diastólica	Preditora	Numérica	21
9	Inicial	Peso	Preditora	Numérica	50
10	Final	Peso	Preditora	Numérica	50
11	Inicial	Hemoglobina	Preditora	Numérica	280
12	Final	Hemoglobina	Preditora	Numérica	1770
13	Inicial	Colesterol Total	Preditora	Numérica	285
14	Final	Colesterol Total	Preditora	Numérica	1615
15	Inicial	Glicose de Jejum	Preditora	Numérica	328
16	Final	Glicose de Jejum	Preditora	Numérica	1453
17	Inicial	Triglicérides	Preditora	Numérica	335
18	Final	Triglicérides	Preditora	Numérica	1713
19	Inicial	Colesterol HDL	Preditora	Numérica	389
20	Final	Colesterol HDL	Preditora	Numérica	1758
21	Inicial	Potássio	Preditora	Numérica	819
22	Inicial	Ureia	Preditora	Numérica	1132
23	Inicial	TSH	Preditora	Numérica	1229
24	Inicial	Ácido Úrico	Preditora	Numérica	1280
25	Inicial	Hemoglobina Glicada	Preditora	Numérica	1338
26	Inicial	TGP	Preditora	Numérica	1405
27	-	DRC_1_2011	Preditora	Numérica	5197
28	-	DRC_2_2011	Preditora	Numérica	5000
29	-	DRC_1_2012	Preditora	Numérica	4899
30	-	DRC_2_2012	Preditora	Numérica	4852
31	-	DRC_1_2013	Preditora	Numérica	4585
32	-	DRC_2_2013	Preditora	Numérica	4566
33	-	DRC_1_2014	Preditora	Numérica	4562
34	-	DRC_2_2014	Preditora	Numérica	4633
35	-	HAS_1_2011	Preditora	Numérica	5293
36	-	HAS_2_2011	Preditora	Numérica	4962
37	-	HAS_1_2012	Preditora	Numérica	4903
38	-	HAS_2_2012	Preditora	Numérica	4844
39	-	HAS_1_2013	Preditora	Numérica	4805
40	-	HAS_2_2013	Preditora	Numérica	4566
41	-	HAS_1_2014	Preditora	Numérica	4557
42	-	HAS_2_2014	Preditora	Numérica	4295
43	-	DM_1_2011	Preditora	Numérica	5156
44	-	DM_2_2011	Preditora	Numérica	4863
45	-	DM_1_2012	Preditora	Numérica	4860
46	-	DM_2_2012	Preditora	Numérica	4737
47	-	DM_1_2013	Preditora	Numérica	4419
48	-	DM_2_2013	Preditora	Numérica	4240
49	-	DM_1_2014	Preditora	Numérica	3923
50	-	DM_2_2014	Preditora	Numérica	3760
51	-	URE	Alvo	Categórica	0

Dado o caráter experimental da proposta do cenário 1, que consiste em análises preliminares do conjunto de dados do CH, foram aplicados cinco métodos distintos para a inferência de dados faltantes: a substituição por zeros, pela média e pela mediana, três dos métodos mais comuns de inferência (128), sendo a substituição pela média amplamente utilizada nos estudos revisados nesta tese no Capítulo 3 (9) (65) (105) (126) (136) (199) (201) (204) (243) e (259); a inferência pelo algoritmo KNN, utilizado por (126), cuja robustez neste tipo de uso foi destacada por (71) (128) (171) e (272); e pelo método MICE, adotado por (30) e (95), sendo especialmente adequado para conjuntos do tipo MAR, uma vez que em outros tipos de composição de dados, o processo de inferência resultante pode conduzir a estimativas enviesadas (14) (187).

Para a predição do URE, todos os cinco métodos de inferência de dados faltantes foram aplicados uma única vez em cada abordagem. Para o tratamento adequado dos dados, foi aplicado o processo de conversão das variáveis categóricas para valores numéricos, conforme exemplificado pela Tabela 5. Para a implementação da classificação com o algoritmo RF, foi selecionada a função `train_test_split` (219), da biblioteca `Scikit-learn` (188), da linguagem de programação `Python`, que foi utilizada na sua versão 3.7. Conforme o *framework* da biblioteca mencionada, 70% dos dados foram destinados à etapa de treinamento, e os 30% restantes foram reservados para a etapa de validação. Com essa configuração, foram realizadas 100 iterações do processo de classificação para cada abordagem de forma individual, a fim de obter o valor médio das acurácias para cada caso.

6.1.2 Resultados e Discussão

Ao se considerar apenas os quatro dados e exames necessários para o cálculo da TFG, na abordagem 1 foi possível predizer, em média, o URE para mais de 80% dos pacientes analisados conforme demonstrado na Tabela 11. Apesar da utilização de apenas quatro atributos, a abordagem 1 apresentou resultados significativamente superiores aos da abordagem subsequente, o que destaca a influência determinante da creatinina sérica na predição do URE.

Já na abordagem 2, as acurácias médias apresentaram os resultados mais baixos desta análise, sobretudo pelo fato dessa ter sido a única abordagem a não ter considerado a creatinina sérica como variável preditora. Embora esta abordagem tenha sido fundamentada em estudos que utilizaram a base de dados da UCI (210) e suas 25 variáveis, como os de Ogunleye e Wang (2020) (180) e Rady *et al.* (201), os resultados demonstraram uma perda considerável na acurácia média em comparação às demais abordagens. Essa discrepância pode ser parcialmente explicada pelas diferenças entre a base de dados utilizada nesta tese e a da UCI (210): mais de 50% dos atributos são distintos e o total de pacientes na base do CH é mais do que 14 vezes superior ao total de registros na base da UCI (210).

Tabela 11 – A acurácia média da classificação com o algoritmo RF para as quatro abordagens de acordo com cada método de inferência de dados. Adaptado de (221).

	ACURÁCIA MÉDIA (%)			
	Abordagem 1	*Abordagem 2*	*Abordagem 3*	*Abordagem 4*
Zero	83,75	68,77	95,17	98,53
Média	83,76	69,82	95,17	98,82
Mediana	83,76	69,85	90,58	97,24
KNN	83,76	69,83	95,17	96,42
MICE	83,75	69,76	95,13	98,05

A combinação, na terceira abordagem, entre o primeiro registro de creatinina sérica (PRCS) e o número total de vezes que um paciente frequentou as clínicas ambulatoriais de DM, DRC e HAS aumentou significativamente a acurácia média, alcançando os segundos maiores valores. Pode se concluir, portanto, que saber quantas vezes e quais clínicas ambulatoriais um paciente frequentou durante seu tratamento pode ser fundamental para a predição do URE e para a compreensão da evolução das condições clínicas. Até o momento, essa abordagem não possui equivalência conhecida na literatura revisada nesta tese.

Por meio da combinação das abordagens 2 e 3, a de número 4 apresentou os maiores valores médios de acurácia, com resultados próximos a 100%. Esta abordagem também não possui equivalência conhecida na literatura revisada. Portanto, os 25 exames e dados mais frequentes, juntamente com o PRCS e as informações relacionadas às clínicas ambulatoriais previamente descritas, demonstram capacidade de predição do URE para quase todos os 5.689 pacientes considerados neste estudo.

Os resultados de acurácia do cenário 1 apresentaram variação aproximada entre 69 e 99%, dependendo da abordagem utilizada e do método de inferência aplicado. Pela análise comparativa dos valores detalhados na Tabela 12 entre este cenário e cinco trabalhos revisados no Capítulo 3, é importante salientar que um dos principais diferenciais da abordagem do cenário 1 foi a utilização de uma base de dados significativamente maior que a de cada um dos estudos revisados, além de ser uma classificação multiclasse, em contraste com as classificações binárias de todos os outros trabalhos. Como já era esperado, os resultados também evidenciam a influência da creatinina sérica na predição dos estágios. A abordagem 2 foi a única a não considerar essa variável como preditora, resultando em valores de acurácia entre 68 e 69%. Já quando a creatinina foi inclusa nas outras três abordagens, os valores de acurácia ficaram entre 83 e 99%, aproximadamente, resultados bastante semelhantes aos dos demais trabalhos revisados.

Tabela 12 – Comparação entre o cenário 1 e cinco estudos revisados nesta tese, considerando as abordagens implementadas, os conjuntos de dados empregados e os resultados de acurácia obtidos na classificação utilizando o RF.

RESULTADOS COMPARATIVOS DE ACURÁCIA PARA O RF

Trabalho	**Total de registros**	**Creatinina**	**Tipo**	**Acurácia**
Cenário 1	5.689	Sim e Não	Multiclasse	69 a 99%
Debal e Sitote (2022) (65)	1.718	Sim	Binário	99%
Ghosh e Khandoker (2024) (96)	491	Sim	Binário	93%
Islam *et al.* (2023) (126)	400	Sim	Binário	98%
Qin *et al.* (2020) (199)	400	Sim	Binário	98 a 99%
Halder *et al.* (2024) (105)	400	Sim	Binário	96 a 100%

6.2 Cenário 2

6.2.1 Definição

Ao contrário do cenário 1, que considerou o PRCS em três das suas quatro abordagens, o segundo cenário de aplicação[2] propôs a classificação do URE sem a inclusão da creatinina sérica como variável preditora, devido ao seu valor estar intrinsecamente relacionado à obtenção da taxa de filtração glomerular e, por conseguinte, ao estágio clínico da doença renal crônica de um paciente. Este cenário utilizou a mesma configuração da base de dados do primeiro, na qual os dados estão organizados de acordo com a data de realização dos exames clínicos e laboratoriais, permitindo que um paciente tenha seus registros distribuídos em uma ou mais linhas da base de dados, conforme exemplificado na Tabela 7.

Novamente seguindo o total de atributos da base da UCI (210), foram selecionadas 25 variáveis para os processos de classificação: as três variáveis relacionadas ao cálculo da TFG, com exceção da creatinina sérica — idade, raça e sexo — e os 22 exames com maior número de registros na base de dados, como detalhado na Tabela 13. Em síntese, o cenário 2 utiliza a segunda abordagem do cenário 1 com o objetivo de expandir a aplicação dos algoritmos de aprendizado de máquina para a classificação do último registro de estágio, buscando, assim, a obtenção de mais informações acerca do comportamento dos dados nas classificações aplicadas.

[2] Uma parte da abordagem e dos resultados do cenário 2 foram publicados em 2021 no volume 1351, páginas 901-910, da *Advances in Intelligent Systems and Computing* (222).

Tabela 13 – Descrição das variáveis utilizadas no cenário 2. Adaptado de (222).

№	Valor	Variáveis	Classe	Tipo	Registros Ausentes
1	-	Idade	Preditora	Numérica	0
2	-	Raça	Preditora	Categórica	0
3	-	Sexo	Preditora	Categórica	0
4	Inicial	P. A. Sistêmica	Preditora	Numérica	21
5	Final	P. A. Sistêmica	Preditora	Numérica	21
6	Inicial	P. A. Diastólica	Preditora	Numérica	21
7	Final	P. A. Diastólica	Preditora	Numérica	21
8	Inicial	Peso	Preditora	Numérica	50
9	Final	Peso	Preditora	Numérica	50
10	Inicial	Hemoglobina	Preditora	Numérica	280
11	Final	Hemoglobina	Preditora	Numérica	1770
12	Inicial	Colesterol Total	Preditora	Numérica	285
13	Final	Colesterol Total	Preditora	Numérica	1615
14	Inicial	Glicose de Jejum	Preditora	Numérica	328
15	Final	Glicose de Jejum	Preditora	Numérica	1453
16	Inicial	Triglicérides	Preditora	Numérica	335
17	Final	Triglicérides	Preditora	Numérica	1713
18	Inicial	Colesterol HDL	Preditora	Numérica	389
19	Final	Colesterol HDL	Preditora	Numérica	1758
20	Inicial	Potássio	Preditora	Numérica	819
21	Inicial	Ureia	Preditora	Numérica	1132
22	Inicial	TSH	Preditora	Numérica	1229
23	Inicial	Ácido Úrico	Preditora	Numérica	1280
24	Inicial	Hemoglobina Glicada	Preditora	Numérica	1338
25	Inicial	TGP	Preditora	Numérica	1405
26	-	URE	Alvo	Categórica	0

Para a classificação, foram definidos cinco algoritmos: ELM, KNN, MLP, RF e SVM, cujos resultados foram comparados aos obtidos pelo algoritmo XGBoost, utilizado em estudos como os de (65) (96) (105) (126) e (180), revisados no Capítulo 3. O XGBoost é amplamente reconhecido nas revisões sistemáticas (135) e (215) como um dos algoritmos mais utilizados e eficazes na predição dos estágios da DRC.

Se no cenário 1 o KNN foi utilizado para a inferência de dados faltantes, no cenário 2 ele foi aplicado para a classificação do URE, como também realizado por (180) (199) e (259). Ademais, segundo (135), o KNN é amplamente empregado em estudos encontrados na literatura sobre AM e DRC. Da mesma forma, o SVM foi destacado tanto pela revisão sistemática (135), quanto pela revisão (215), e implementado, com resultados relevantes, na maioria dos estudos revisados nesta tese (9) (10) (65) (82) (105) (136) (180) (191) (199) (201) (204) e (243). O MLP, além de ter sido utilizado nas publicações (82) (136) e

(201), foi selecionado por sua fundamentação no conceito de redes neuronais artificiais, assim como o ELM, diferenciando-se dos outros algoritmos. Sua relevância também foi destacada na revisão sistemática (215), juntamente com o algoritmo LogR, aplicado por (96) e ressaltado por (135).

A Tabela 14 fornece uma comparação entre os seis algoritmos selecionados. Considerando que cada um deles possui fundamentação matemática, estrutura e funcionamento distintos, foi possível, no cenário 2, aplicar e avaliar diferentes processos de classificação sobre a base de dados do CH. Assim, os resultados obtidos podem servir como referência para o desenvolvimento de novos cenários de aplicação.

Tabela 14 – Comparação entre a estrutura e o funcionamento dos seis algoritmos utilizados no cenário 2 (48) (57) (58) (120) (152) (212).

Algoritmo	**Estrutura**	**Funcionamento**
ELM	Rede neural de camada única	Treinamento rápido com ajuste dos pesos da camada de saída, sem ajuste da camada oculta
KNN	Baseado em vizinhos	Classificação com base nos k vizinhos mais próximos do conjunto de variáveis
MLP	Rede neural com múltiplas camadas	Utilização de camadas ocultas e de ajuste de pesos para o aprendizado de representações complexas
LogR	Função logística/sigmoide linear	Modelagem da probabilidade de um evento binário com uma combinação linear de variáveis
SVM	Hiperplano linear ou não linear	Determinação do hiperplano ótimo que maximiza a margem entre classes
XGBoost	Conjunto de árvores de decisão	Utilização do conceito de *boosting* para a combinação de várias árvores fracas a fim de potencializar os resultados

Mais uma vez, as variáveis categóricas foram convertidas em valores numéricos. Considerando o caráter exploratório e preliminar do cenário 2, e dado que os melhores resultados do cenário 1 foram obtidos utilizando a inferência pela substituição por zeros e por média, o primeiro método foi o escolhido para a inferência dos dados faltantes, também devido à sua simplicidade e objetividade. E para os seis algoritmos de classificação, foi mais uma vez adotada a função `train_test_split` (219) do pacote `Scikit-learn` (188). O conjunto de dados foi separado em 70% para o treinamento e 30% para a validação. E um total de 100 iterações foi executado para a classificação de cada algoritmo.

6.2.2 Resultados e Discussão

Para cada um dos seis algoritmos de classificação, o resultado da acurácia média, após 100 iterações, está ilustrado na Tabela 15.

Tabela 15 – Acurácia média para cada um dos seis algoritmos de classificação. Adaptado de (222).

ACURÁCIA MÉDIA (%)						
Algoritmo	**XGBoost**	**KNN**	**ELM**	**MLP**	**LogR**	**SVM**
Valor	96	48	36	35	35	25

Já os resultados detalhados para cada algoritmo encontram-se na Tabela 16. Além da acurácia média global, são apresentados os valores das métricas de precisão, revocação e F1-*score*. Todos os dados se referem a uma única iteração, selecionada aleatoriamente como exemplo entre as 100 iterações realizadas para a obtenção dos resultados. Em todas as situações, o conjunto de suporte foi composto por 12.030 registros, distribuídos entre os estágios 1, 2, 3a, 3b, 4 e 5 da DRC, conforme a seguinte ordem: 2.285, 3.913, 2.404, 1.847, 1.146 e 435 registros.

Pela análise das Tabelas 15 e 16, é possível verificar que entre todos os algoritmos implementados, o XGBoost foi o único que alcançou uma acurácia superior a 50%, destacando-se consideravelmente em relação aos demais métodos. Quando comparados aos resultados reportados por Ogunleye e Wang (2020) (180), que obtiveram uma acurácia de 98,7%, e por Islam et al. (2023) (126), com uma acurácia de 98,3%, os dados obtidos no cenário 2 mostraram valores de acurácia similares, reforçando a eficácia do XGBoost no contexto de classificação proposto.

Embora os valores de acurácia e o total de variáveis sejam muito próximos, existem diferenças essenciais entre as abordagens dos dois trabalhos mencionados e o cenário 2. A primeira reside no uso da variável mais influente no cálculo da TFG: a creatinina sérica. Em contrapartida, a abordagem desenvolvida neste cenário não incluiu a creatinina, conforme descrito na Tabela 13.

A segunda diferença consiste na utilização de um banco de dados com 5.689 pacientes doentes renais crônicos ou com condições clínicas prévias à doença. Os estudos de Ganie *et al.* (2023) (90), Ogunleye e Wang (2020) (180) e Islam *et al.* (2023) (126) consideraram a base de dados da UCI (210), um conjunto com somente 400 pacientes com DRC, valor mais do que 14 vezes inferior ao total de pacientes da base do Centro Hiperdia. Não obstante a base utilizada nesta tese tenha consideravelmente mais pacientes e uma quantidade significativa de valores ausentes, a implementação do algoritmo XGBoost permitiu alcançar resultados muito próximos aos obtidos por Ogunleye e Wang (2020) (180), Ganie *et al.* (2023) (90) e Islam *et al.* (2023) (126).

Tabela 16 – Resultados da acurácia, precisão, revocação e F1-*score* da classificação com os algoritmos SVM, LogR, MLP, ELM, KNN e XGBoost para cada um dos seis estágios da DRC. Em todos os casos, o conjunto suporte possui 12.030 registros.

	SVM			**LogR**			
Estágios	**Precisão**	**Revocação**	**F1-*score***	**Precisão**	**Revocação**	**F1-*score***	**Suporte**
1	0,94	0,06	0,12	0,50	0,18	0,26	2.285
2	0,98	0,06	0,12	0,35	0,84	0,50	3.913
3a	0,99	0,10	0,18	0,23	0,04	0,06	2.404
3b	0,95	0,21	0,34	0,27	0,21	0,24	1.847
4	0,11	0,99	0,19	0,20	0,02	0,04	1.146
5	0,99	0,39	0,56	0,38	0,02	0,04	435
Acurácia	*25%*			*35%*			12.030
	MLP			**ELM**			
Estágios	**Precisão**	**Revocação**	**F1-*score***	**Precisão**	**Revocação**	**F1-*score***	**Suporte**
1	0,45	0,23	0,30	0,59	0,06	0,11	2.285
2	0,35	0,80	0,49	0,34	0,93	0,50	3.913
3a	0,25	0,03	0,05	0,15	0,02	0,03	2.404
3b	0,32	0,27	0,29	0,30	0,15	0,20	1.847
4	0,20	0,00	0,00	0,14	0,00	0,00	1.146
5	0,34	0,04	0,07	0,98	0,00	0,01	435
Acurácia	*35%*			*36%*			12.030
	KNN			**XGBoost**			
Estágios	**Precisão**	**Revocação**	**F1-*score***	**Precisão**	**Revocação**	**F1-*score***	**Suporte**
1	0,39	0,41	0,40	0,89	0,82	0,86	2.285
2	0,42	0,58	0,49	0,80	0,91	0,85	3.913
3a	0,35	0,30	0,32	0,86	0,83	0,85	2.404
3b	0,49	0,35	0,41	0,93	0,88	0,90	1.847
4	0,52	0,32	0,40	0,95	0,84	0,89	1.146
5	0,59	0,31	0,41	0,98	0,86	0,92	435
Acurácia	*48%*			*96%*			12.030

Ao comparar os resultados com os quatro estudos revisados no Capítulo 3, pode ser observado que o XGBoost, no cenário 2, alcançou uma acurácia semelhante à dos demais trabalhos, conforme indicado na Tabela 17. Assim como no cenário 1, deve ser enfatizado que o cenário 2 considerou uma base de dados muito mais ampla em comparação com cada um dos estudos revisados, além de aplicar processos de classificação multiclasse, diferindo dos demais trabalhos que, em sua maioria, optaram por classificações binárias. Ademais, é relevante notar que, diferentemente do cenário 2, todos os outros estudos utilizaram a creatinina sérica como variável preditora. Portanto, pela análise dos resultados comparativos da Tabela 17, o cenário 2 foi capaz de oferecer uma solução robusta mesmo em um ambiente com uma maior quantidade de variáveis e sem a dependência de marcadores padrão.

Tabela 17 – Comparação do cenário 2 com quatro trabalhos revisados nesta tese considerando as abordagens implementadas, a base de dados utilizada e os resultados de acurácia da classificação com XGBoost.

RESULTADOS COMPARATIVOS DE ACURÁCIA PARA O XGBoost

Trabalho	**Total de registros**	**Creatinina**	**Tipo**	**Acurácia**
Cenário 2	5.689	Não	Multiclasse	96%
Debal e Sitote (2022) (65)	1.718	Sim	Binário e multiclasse	83 e 99%
Ganie *et al.* (2023) (90)	400	Sim	Binário	95,93%
Ghosh e Khandoker (2024) (96)	491	Sim	Binário	93%
Islam *et al.* (2023) (126)	400	Sim	Binário	98%
Ogunleye e Wang (2020) (180)	400	Sim	Binário	99%

Além da acurácia, a análise da revocação é igualmente fundamental, por avaliar a proporção de pacientes cujo estágio da doença é corretamente identificado pelos modelos, em relação ao número total de pacientes em cada estágio real. Portanto, a capacidade do modelo de identificar corretamente o estágio da DRC em que os pacientes se encontram, fato fundamental, sobretudo à medida que a doença evolui. A análise da precisão também tem a sua importância já que denota a proporção de acerto dos modelos com relação a todas as previsões realizadas para um mesmo estágio, corretas ou incorretas, minimizando os casos de falsos positivos.

A LogR demonstrou um valor elevado de revocação para o estágio 2 (0,84), o que indica que o modelo foi eficaz em capturar a maioria dos pacientes nesse estágio. No entanto, para os demais quatro estágios, os valores de revocação foram baixos (variando de 0,02 a 0,21), sugerindo uma alta taxa de falsos negativos nesses estágios. Os valores de precisão para a LogR foram moderados, com um desempenho razoável no estágio 1, mas valores baixos nos demais, sobretudo no estágio 4, o que pode indicar uma alta taxa de falsos positivos.

O SVM teve um comportamento semelhante ao do LogR na revocação, com valor elevado apenas para o estágio 4 (0,99), consideravelmente maior do que os valores para os demais estágios, podendo indicar um caso de sobreajuste do modelo e que muitos pacientes não foram corretamente identificados nos estágios iniciais. Esses resultados são preocupantes, especialmente para os estágios 1 e 2, nos quais a identificação precoce é crucial para o encaminhamento adequado do paciente. Com relação à precisão, o SVM apresentou valores elevados nos estágios 1, 2, 3a, 3b e 5, indicando que a maioria das

previsões positivas nesses estágios foi correta. No entanto, para o estágio 4, a precisão foi consideravelmente baixa (0,11), sugerindo que, embora o modelo tenha identificado corretamente muitos pacientes no estágio 4, ele também classificou incorretamente muitos outros que estavam em outros estágios.

O MLP, por sua vez, apresentou revocação de 0,27 no estágio 3b, indicando que apenas 27% dos pacientes foram corretamente identificados. O valor da revocação foi ainda mais baixo para os estágios 1, 3a, 4 e 5. No entanto, no estágio 2, a revocação aumentou consideravelmente para 0,80, o que demonstra uma capacidade elevada de identificar corretamente os pacientes nesse estágio específico. E, para todos os estágios, os valores de precisão foram baixos para o MLP, sendo o melhor resultado para o estágio 1, 45% e o pior, assim como para a maioria dos outros algoritmos, o estágio 4, com apenas 20%.

O estágio 2 se destacou entre os resultados do ELM, com uma revocação de 0,93, indicando uma capacidade elevada de identificar corretamente pacientes nesse estágio. No entanto, o desempenho do ELM foi insatisfatório nos demais casos, com valores de revocação próximos ou iguais a zero nos estágios 1, 4 e 5, o que sugere uma elevada taxa de falsos negativos. Já com relação à precisão, o ELM obteve valor elevado no estágio 5, 98%. Contudo, assim como no SVM, esse resultado deve ser interpretado com cautela devido aos valores baixos de revocação em ambos os casos, isto é, uma pequena proporção de previsões positivas.

Com exceção do estágio 2, em que apresentou um valor moderado de revocação, 0,58, para todos os demais o KNN foi insatisfatório, variando de 0,30 a 0,41. Também para a precisão, o algoritmo não apresentou resultados muito diferentes. Nos estágios intermediários, a precisão pouco variou, com valores de 0,35 no estágio 3a e 0,49 no estágio 3b. Esses resultados indicam que o KNN foi razoavelmente eficaz em alguns estágios intermediários, mas ainda gerou um número considerável de falsos positivos. Nos estágios 4 e 5, a precisão subiu para 0,52 e 0,59, respectivamente, sugerindo que o modelo foi mais eficiente em evitar falsos positivos nos estágios mais avançados da DRC.

Além da acurácia, o XGBoost também demonstrou ser o mais eficaz dos algoritmos testados, com ótimos resultados de revocação e de precisão em todos os estágios, com valores entre 0,82 e 0,91, e de 0,80 a 0,98, respectivamente, o que demonstra uma capacidade robusta de identificar corretamente pacientes com DRC em diferentes estágios, minimizando falsos negativos e, consequentemente, favorecendo o diagnóstico preciso. Também para a precisão, o XGBoost obteve valores elevados em todos os estágios, variando de 82% a 91%. Portanto, o modelo foi eficaz em evitar falsos positivos e, ao mesmo tempo, manteve um bom desempenho na identificação correta dos estágios dos pacientes com DRC.

Com base na análise das métricas de revocação e precisão para os seis algoritmos avaliados, pode ser concluído que o XGBoost foi o algoritmo mais robusto e confiável para a predição dos estágios da DRC. Sua alta revocação em todos os estágios da doença pode

garantir que a maioria dos pacientes com DRC seja corretamente identificada, minimizando o número de falsos negativos, o que é crucial para assegurar um tratamento precoce e adequado. Além disso, a alta precisão em todos os estágios pode reduzir o risco de falsos positivos, evitando diagnósticos incorretos e intervenções desnecessárias. Esse equilíbrio entre acurácia, revocação e precisão coloca o XGBoost em uma posição vantajosa no contexto clínico, no qual a correta identificação dos estágios da DRC é essencial para o manejo eficaz da doença. Os resultados apresentados na Tabela 17 fornecem uma análise objetiva da eficácia do XGBoost na predição da DRC no cenário considerado.

Em contrapartida, os demais algoritmos apresentaram desempenho variável e, em muitos casos, inconsistências significativas em termos de revocação e precisão. O SVM e a LogR tiveram bons desempenhos apenas em alguns estágios específicos, comprometendo sua aplicação prática em cenários clínicos sensíveis. Excetuando o estágio 2 para o MLP e o ELM e, juntamente ao KNN em todos os estágios, esses três algoritmos demonstraram limitações substanciais, com baixa revocação e precisão em vários estágios, o que os torna menos confiáveis para predição precisa. Dada a natureza complexa e multifatorial da DRC, na qual a detecção precoce e a progressão precisa da doença são fundamentais, o XGBoost proporcionou resultados muito promissores, se colocando como a melhor opção entre os algoritmos avaliados para a classificação correta de todos os estágios dos pacientes da base do Centro Hiperdia.

6.3 Cenário 3

6.3.1 Definição

Conforme evidenciado pela revisão sistemática realizada por Sanmarchi *et al.* (2023) (215), a creatinina sérica é uma das principais variáveis empregadas na literatura de AM aplicado à DRC. Estudos como os de Rady *et al.* (2019) (201), Wang *et al.* (2021) (259), Debal *et al.* (2022) (65) e Ghosh *et al.* (2024) (96), destacaram a relevância dessa variável na predição dos estágios da DRC, especialmente porque seu valor está intrinsecamente relacionado à TFG de um indivíduo sendo, portanto, um indicador do estágio da doença. A base de dados da UCI (210), utilizada pela maior parte dos estudos revisados nesta tese, considerou a creatinina entre as suas variáveis preditoras.

Entre as implementações propostas nos cenários 1 e 2, apenas a abordagem 2 do primeiro cenário, que foi replicada no segundo, fez uso da creatinina como variável preditora. Foram consideradas 25 variáveis para a classificação do URE, um total que também é observado em outros estudos, como os de Ogunleye e Wang (2020) (180) e Islam *et al.* (2023) (126). No entanto, esse número elevado de dados pessoais e exames clínicos representa um desafio, uma vez que o processo de coleta nem sempre é viável ou trivial em rotinas laboratoriais.

Com base nos cenários analisados e em alinhamento com os objetivos iniciais estabelecidos na Seção 1.3, foi delineada a proposta central para o cenário 3, que também serviu de base para os cenários 4 e 5, que serão definidos. Essa proposta visa à identificação de um conjunto mínimo de variáveis, que seja viável, clinicamente adequado e que se diferencie dos padrões clínicos e laboratoriais convencionais, para a predição do URE dos pacientes da base de dados do CH, sem a necessidade de utilizar a creatinina sérica como parâmetro.

Uma vez que um dos objetivos do cenário 3 é propor uma nova abordagem para a classificação dos dados, todos os métodos selecionados foram aplicados sobre a base em sua formatação original, com os dados dispostos em uma linha para cada um dos 5.689 registros de pacientes. Entre os cinco algoritmos selecionados, encontram-se os dois com melhor desempenho nos cenários anteriores, RF e XGBoost. Como este último obteve a maior acurácia no segundo cenário e, seguindo as propostas de Ganie *et al.* (2023) (90), foram inclusos no terceiro outros dois métodos de *ensemble*, o AdaBoost, empregado por (90) (105) e (126), e o GB, adotado por (10) (90) (105) (126) e (180) e destacado na revisão sistemática (135). O quinto algoritmo escolhido foi o SVM, amplamente utilizado na maioria dos trabalhos revisados nesta tese (9) (10) (65) (82) (105) (136) (180) (191) (199) (201) (204) e (243), e destacado nas revisões (135) e (215).

Para o tratamento do desbalanceamento dos dados foram utilizados cinco métodos: ADASYN, SMOTE — assim como Silveira *et al.* (2021) (228), Su *et al.* (2022) (238) e Ganie *et al.* (2023) (90) — e três de suas principais variações: o SMOTE-Tomek , o *Borderline*-SMOTE e o SMOTE-ENN, detalhados na Seção 4.5.

Como no cenário 1, os dados ausentes foram inferidos pelos métodos KNN e MICE, e pela substituição por zeros, pela média e pela mediana. Da mesma forma, a função `train_test_split` (219) foi empregada para realizar a divisão do conjunto de dados em uma parcela com 70% das informações para o treinamento e 30% para a validação. Novamente, um total de 100 iterações foi executado para a classificação de cada algoritmo.

6.3.2 Resultados e Discussão

Os resultados aproximados da acurácia média obtidos pelos algoritmos RF e XGBoost na predição do URE, após 100 iterações, são apresentados na Tabela 18.

Tabela 18 – Resultados aproximados da acurácia média para os algoritmos RF e XGBoost com cinco métodos de inferência de dados.

	ACURÁCIA MÉDIA (%)				
Algoritmo	**Zero**	**Média**	**Mediana**	**KNN**	**MICE**
RF	43	43	43	43	43
XGBoost	42	42	42	42	42

Mesmo que só tenham sido utilizadas três das quatro variáveis necessárias para o

cálculo da TFG pela Equação 2.2 (MDRD), os resultados dos dois algoritmos acabaram sendo praticamente iguais, independentemente de qual método de inferência foi aplicado. No entanto, todos os valores de acurácia ficaram abaixo do esperado, o que significa que os algoritmos não conseguiram classificar corretamente o URE dos pacientes da base de dados do CH. Por isso, é preciso considerar outras técnicas de análise e maneiras de tratar os dados, de forma que os resultados sejam melhorados para a obtenção de classificações mais precisas.

A avaliação do balanceamento dos dados é crucial para a construção de um modelo de aprendizado de máquina. Um conjunto de dados desbalanceado pode gerar modelos enviesados, que favorecem a classe majoritária em detrimento da minoritária, como exemplificado no trabalho de Wang *et al.* (259). Portanto, as métricas de desempenho podem não refletir adequadamente a verdadeira eficácia do modelo, além de prejudicar a sua generalização.

A Tabela 19 evidencia o desbalanceamento da base de dados em relação aos estágios da DRC em que se encontram os pacientes. Há clara predominância dos estágios iniciais (1, 2 e 3a), ao passo que os demais apresentam amostras menores. É necessário aumentar a representação das classes minoritárias (3b, 4 e 5) por meio de métodos de *oversampling*. Para tratar o desbalanceamento, foram selecionados os algoritmos ADASYN, SMOTE, SMOTE-Tomek (ST), *Borderline*-SMOTE (BS) e SMOTE-ENN (SE). A partir da análise da Tabela 19, pode ser observado que o balanceamento realizado pelo algoritmo SMOTE foi eficaz em igualar a distribuição das classes referentes aos diferentes estágios da doença, tomando como referência a classe minoritária, correspondente ao estágio 1.

Tabela 19 – Distribuição das classes por estágio antes e depois do balanceamento com SMOTE.

Estágio	**Distribuição original**	**Distribuição com SMOTE**
1	1406	1405
2	846	1406
3a	815	1405
3b	517	1405
4	288	1405
5	110	1406

A Tabela 20 apresenta os resultados dos testes realizados com os cinco algoritmos de AM e os cinco métodos de balanceamento de dados. A coluna “S/B” (sem balanceamento) apresenta os resultados dos algoritmos sem a aplicação de qualquer método de balanceamento. Além disso, o valor médio da métrica ROC AUC também é fornecido para todos os seis estágios da DRC, sendo detalhado tanto por algoritmo quanto por método de balanceamento.

Tabela 20 – Resultados da acurácia e da ROC AUC média para os cinco algoritmos de AM utilizados com cada um dos cinco métodos de balanceamento dos dados.

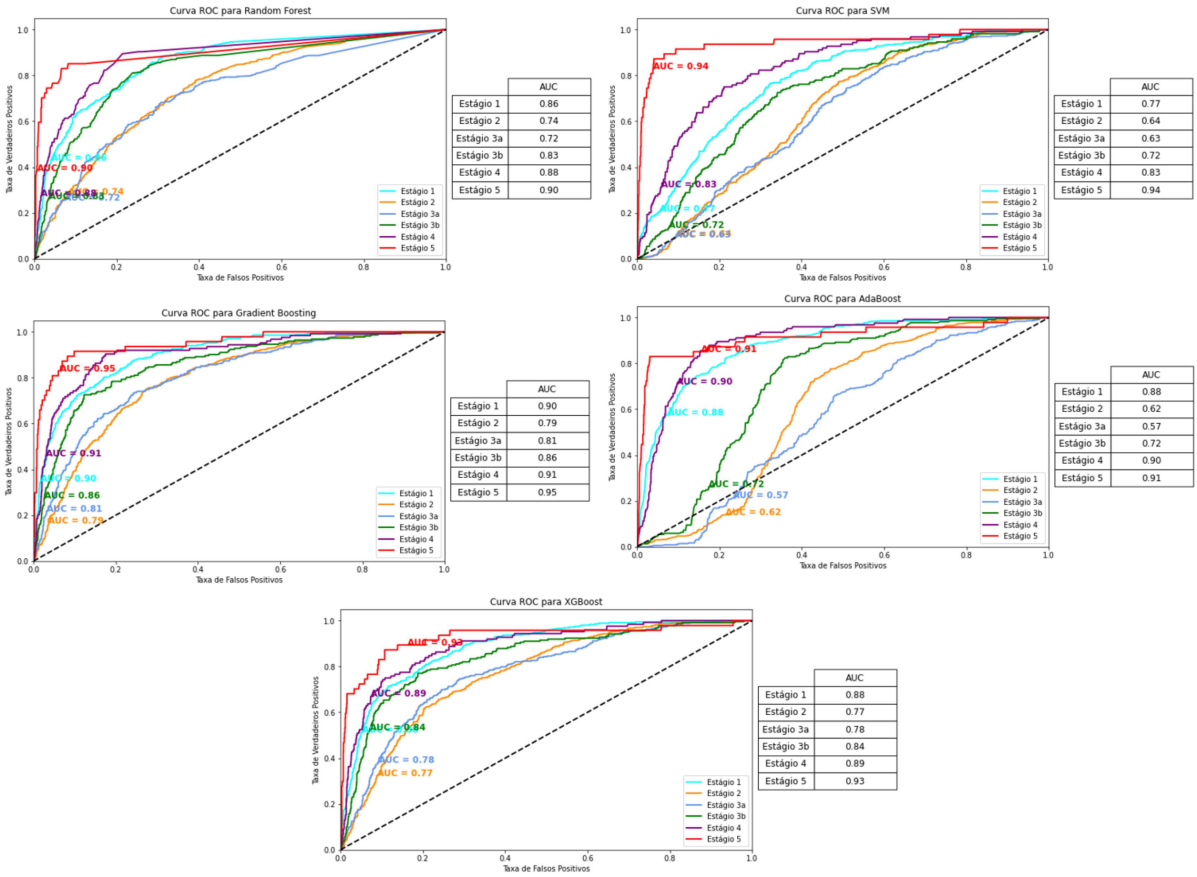
VALORES DE ACURÁCIA							
Algoritmos	**S/B**	**ADASYN**	**SMOTE**	**ST**	**BS**	**SE**	**AUC**$_{média}$
RF	0,35	0,37	0,39	0,39	0,37	0,46	0,83
SVM	0,21	0,32	0,34	0,34	0,31	0,18	0,76
GB	0,36	0,45	0,46	0,46	0,45	0,47	0,87
AdaBoost	0,36	0,41	0,43	0,43	0,42	0,25	0,77
XGBoost	0,22	0,39	0,42	0,42	0,40	0,45	0,85

Apesar dos resultados, em geral, terem sido insatisfatórios, a aplicação de técnicas de balanceamento de dados proporcionou uma melhor acurácia em quase todos os casos. Neste conjunto de testes, o GB pode ser destacado como o algoritmo mais robusto, apresentando consistentemente os melhores resultados de acurácia, especialmente após a aplicação do SMOTE-ENN, tendo atingido o valor de 0,47. Além disso, o resultado de ROC AUC de 0,87 denota a capacidade de discriminação entre as classes — *i.e.* os diferentes estágios — pelo algoritmo. O XGBoost e o RF apresentaram desempenho razoável, sobretudo por terem proporcionado uma melhoria significativa na acurácia quando as técnicas de balanceamento foram aplicadas. No entanto, os resultados também mostram que o impacto do balanceamento de dados é sensível ao algoritmo específico, como evidenciado pela queda de acurácia para o SVM e o AdaBoost com o uso de SMOTE-ENN.

Em contraste com os resultados de acurácia, os valores de ROC AUC obtidos indicam potenciais aplicações promissoras dos algoritmos na classificação proposta, especialmente para o GB, RF e XGBoost, que obtiveram os valores de 0,87, 0,83 e 0,85, respectivamente. A Figura 22 apresenta uma análise dos resultados de ROC AUC obtidos após o balanceamento dos dados com o método SMOTE, considerando os cinco algoritmos e com a separação dos valores por estágio.

A análise dos resultados apresentados nos gráficos e nas tabelas da Figura 22 revela que os maiores valores de ROC AUC foram observados nos estágios extremos da DRC: o estágio inicial (1) e os estágios finais (4 e 5). Este comportamento é consistente com o esperado, pois no estágio 1 da DRC o paciente ainda apresenta poucos indícios de comprometimento renal, podendo inclusive ser assintomático e sem alterações significativas nos exames clínicos. Nos estágios 4 e 5, por outro lado, as funções renais já se encontram severamente prejudicadas, conforme exemplificado na Figura 1, e o paciente está em processo de iniciar ou já se encontra em tratamento por meio de alguma TRS. Portanto, para esses três estágios, os padrões clínicos dos pacientes são bem estabelecidos e amplamente reconhecidos. Entretanto, nos estágios intermediários (2, 3a e 3b), o comportamento clínico pode variar consideravelmente, uma vez que a progressão ou regressão da doença pode ocorrer em função de intervenções terapêuticas, como tratamentos medicamentosos.

Figura 22 – Gráficos e tabelas exibindo os resultados da curva ROC AUC, por estágio, para cada um dos cinco algoritmos de classificação, após o processo de balanceamento dos dados com SMOTE.



	AUC
Estágio 1	0.86
Estágio 2	0.74
Estágio 3a	0.72
Estágio 3b	0.83
Estágio 4	0.88
Estágio 5	0.90

	AUC
Estágio 1	0.77
Estágio 2	0.64
Estágio 3a	0.63
Estágio 3b	0.72
Estágio 4	0.83
Estágio 5	0.94

	AUC
Estágio 1	0.90
Estágio 2	0.79
Estágio 3a	0.81
Estágio 3b	0.86
Estágio 4	0.91
Estágio 5	0.95

	AUC
Estágio 1	0.88
Estágio 2	0.62
Estágio 3a	0.57
Estágio 3b	0.72
Estágio 4	0.90
Estágio 5	0.91

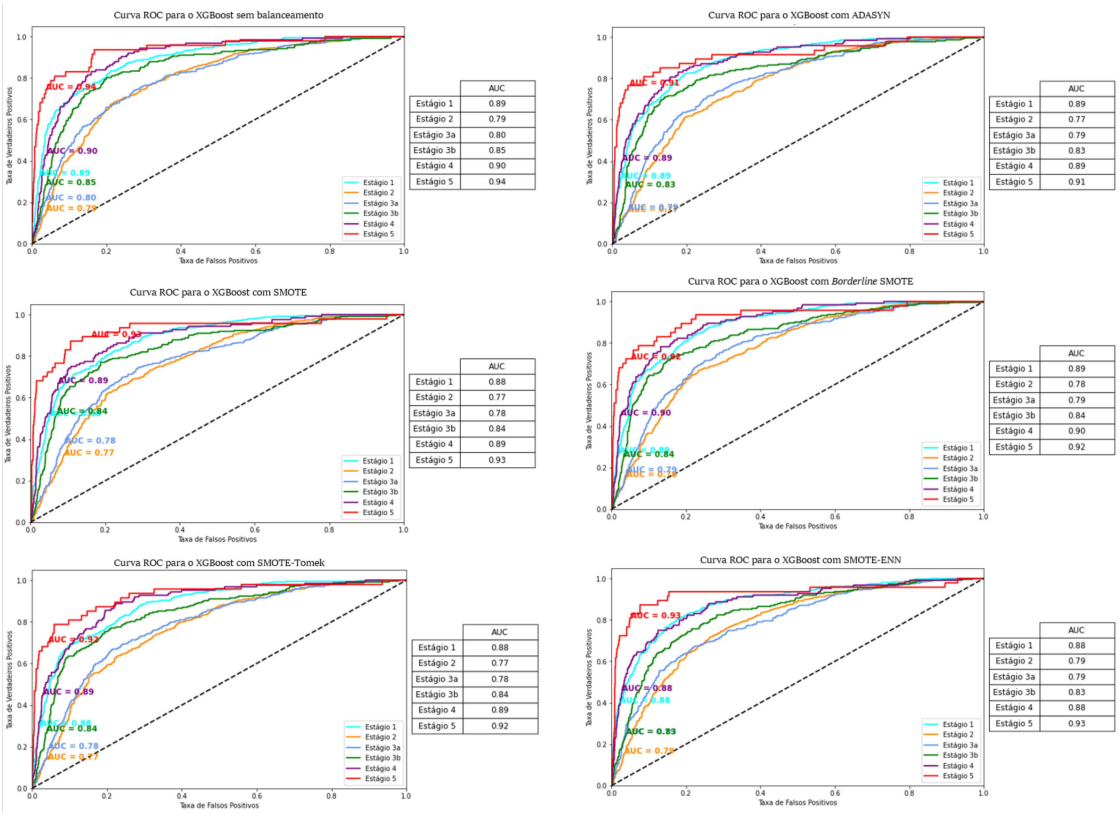
	AUC
Estágio 1	0.88
Estágio 2	0.77
Estágio 3a	0.78
Estágio 3b	0.84
Estágio 4	0.89
Estágio 5	0.93

Fonte: Elaborada pelo autor.

Para que sejam realizadas análises mais específicas e detalhadas, a Figura 23 e a Tabela 21 apresentam os resultados obtidos especificamente pelo algoritmo XGBoost. Nessas representações, são reportadas as métricas de acurácia, precisão, revocação e F1-*score*, além dos valores correspondentes à curva ROC AUC para cada estágio da doença renal crônica.

Na ausência de técnicas de balanceamento, o XGBoost apresentou uma acurácia global limitada, alcançando apenas 22%, com um desempenho moderado nas demais métricas de avaliação. No entanto, destacou-se pela obtenção de valores elevados para a métrica ROC AUC, especialmente no estágio 5, com um valor de 0,94. De maneira similar à etapa inicial deste cenário, os resultados sugerem um desempenho superior do modelo na correta classificação do estágio inicial e dos dois estágios mais avançados da DRC. Adicionalmente, para os estágios intermediários, o algoritmo também apresentou resultados relevantes.

Figura 23 – Resultados, por estágio, da curva ROC AUC para o algoritmo XGBoost quando implementado após o processo de balanceamento por cada um dos cinco métodos.



	AUC
Estágio 1	0.89
Estágio 2	0.79
Estágio 3a	0.80
Estágio 3b	0.85
Estágio 4	0.90
Estágio 5	0.94

	AUC
Estágio 1	0.89
Estágio 2	0.77
Estágio 3a	0.79
Estágio 3b	0.83
Estágio 4	0.89
Estágio 5	0.91

	AUC
Estágio 1	0.88
Estágio 2	0.77
Estágio 3a	0.78
Estágio 3b	0.84
Estágio 4	0.89
Estágio 5	0.93

	AUC
Estágio 1	0.89
Estágio 2	0.78
Estágio 3a	0.79
Estágio 3b	0.84
Estágio 4	0.90
Estágio 5	0.92

	AUC
Estágio 1	0.88
Estágio 2	0.77
Estágio 3a	0.78
Estágio 3b	0.84
Estágio 4	0.89
Estágio 5	0.92

	AUC
Estágio 1	0.88
Estágio 2	0.79
Estágio 3a	0.79
Estágio 3b	0.83
Estágio 4	0.88
Estágio 5	0.93

Fonte: Elaborada pelo autor.

As métricas de precisão e revocação apresentaram variações expressivas, com valores inconsistentes e relativamente baixos entre os diferentes estágios clínicos da DRC. Esse comportamento pode sinalizar uma limitação do modelo em identificar corretamente instâncias de classes menos representadas, refletindo a dificuldade em lidar com a distribuição desequilibrada dos dados. Em consequência, pode ser observada a presença de uma quantidade considerável de classificações equivocadas, especialmente na forma de falsos positivos, nos estágios 1, 3a, 3b e 4. A imprecisão na definição do estágio da doença renal crônica pode acarretar sérios prejuízos ao tratamento adequado e ao correto encaminhamento clínico dos pacientes, comprometendo significativamente a eficácia das intervenções terapêuticas.

Com a aplicação de métodos de balanceamento, é possível observar uma melhora discreta em algumas métricas de desempenho. A acurácia aumentou consideravelmente, alcançando 39% com ADASYN e 42% com SMOTE e SMOTE-Tomek, com o maior valor registrado sendo 45%, com o SMOTE-ENN. O SMOTE-Tomek e o *Borderline*-SMOTE conseguiram manter a ROC AUC alta, especialmente nos estágios 1, 4 e 5, ao passo que as demais métricas mostraram uma leve variação.

Tabela 21 – Seis diferentes aplicações da classificação do XGBoost. Na primeira, não há qualquer método para tratar o desbalanceamento dos dados. Na segunda, foi aplicado o algoritmo ADASYN. Nas demais, foram utilizados os métodos SMOTE e suas variações, nesta ordem: SMOTE-Tomek, o *Borderline*-SMOTE e o SMOTE-ENN. Para todos os casos, o conjunto suporte é de 350, 602, 362, 222, 124 e 47 (total de 1.707), respectivamente para os estágios 1, 2, 3a, 3b, 4 e 5.

	XGBoost (S/B)				**XGBoost e ADASYN**			
Estágios	**Prec.**	**Revoc.**	**F1-*score***	**AUC**	**Prec.**	**Revoc.**	**F1-*score***	**AUC**
1	0,60	0,39	0,47	0,89	0,49	0,62	0,55	0,89
2	0,48	0,64	0,55	0,79	0,54	0,25	0,34	0,77
3a	0,39	0,38	0,38	0,80	0,35	0,35	0,35	0,79
3b	0,40	0,33	0,36	0,85	0,29	0,38	0,33	0,83
4	0,49	0,37	0,42	0,90	0,23	0,44	0,30	0,89
5	0,65	0,60	0,62	0,94	0,35	0,64	0,45	0,91
Acurácia	*22%*				*39%*			
	XGBoost e SMOTE				**XGBoost e SMOTE-Tomek**			
Estágios	**Prec.**	**Revoc.**	**F1-*score***	**AUC**	**Prec.**	**Revoc.**	**F1-*score***	**AUC**
1	0,51	0,55	0,53	0,88	0,51	0,58	0,54	0,89
2	0,51	0,35	0,42	0,77	0,52	0,34	0,41	0,78
3a	0,39	0,44	0,41	0,78	0,38	0,43	0,41	0,79
3b	0,31	0,36	0,34	0,84	0,32	0,38	0,35	0,84
4	0,26	0,35	0,30	0,89	0,27	0,35	0,31	0,90
5	0,35	0,62	0,44	0,93	0,36	0,64	0,46	0,92
Acurácia	*42%*				*42%*			
	XGBoost e *Borderline-SMOTE*				**XGBoost e SMOTE-ENN**			
Estágios	**Prec.**	**Revoc.**	**F1-*score***	**AUC**	**Prec.**	**Revoc.**	**F1-*score***	**AUC**
1	0,50	0,63	0,56	0,88	0,51	0,57	0,54	0,88
2	0,53	0,30	0,38	0,77	0,52	0,34	0,41	0,79
3a	0,33	0,32	0,32	0,78	0,36	0,39	0,37	0,79
3b	0,29	0,40	0,33	0,84	0,32	0,38	0,35	0,83
4	0,26	0,41	0,32	0,89	0,27	0,38	0,31	0,88
5	0,35	0,60	0,44	0,92	0,38	0,68	0,49	0,93
Acurácia	*40%*				*45%*			

As métricas de revocação e precisão indicaram que o balanceamento dos dados não resultou em uma melhoria significativa na capacidade do modelo de identificar as classes minoritárias sem causar um aumento excessivo no número de falsos positivos. De maneira geral, o estágio 5 apresentou os maiores valores de revocação, embora estes se mantivessem próximos dos valores originais. Em contrapartida, o estágio 1 mostrou um aumento substancial.

Apesar das discretas melhorias observadas, os resultados demonstraram que alguns métodos de balanceamento apresentaram comportamentos inconsistentes, melhorando a revocação em certos estágios, mas diminuindo-a drasticamente em outros. Um exemplo claro é o estágio 2, no qual a revocação foi significativamente reduzida com a aplicação de todos os métodos de balanceamento. Portanto, embora o balanceamento dos dados possa melhorar o desempenho geral do modelo em certo grau, a escolha do método adequado deve ser realizada com cautela. É crucial assegurar que o modelo mantenha uma performance robusta em todos os estágios da doença, sem comprometer a precisão nem introduzir um número excessivo de falsos positivos.

Embora os resultados tenham sido, em certa medida, insatisfatórios, foi possível observar melhorias modestas nos estágios 1, 4 e 5 após a aplicação das técnicas de balanceamento. Assim, a abordagem desenvolvida no cenário 3 apresentou potencial para ser expandida e aprimorada, especialmente em relação ao estágio inicial. Com a utilização de apenas três variáveis — idade, raça e sexo — foi possível obter resultados promissores em certo nível, particularmente no que se refere à curva ROC AUC e à revocação. Dessa forma, com apenas três dados pessoais de fácil obtenção, há a possibilidade de serem alcançados bons resultados na predição da DRC, o que permitiria a identificação de pacientes no estágio 1 e o consequente encaminhamento para tratamentos iniciais apropriados, possivelmente focados em intervenções medicamentosas.

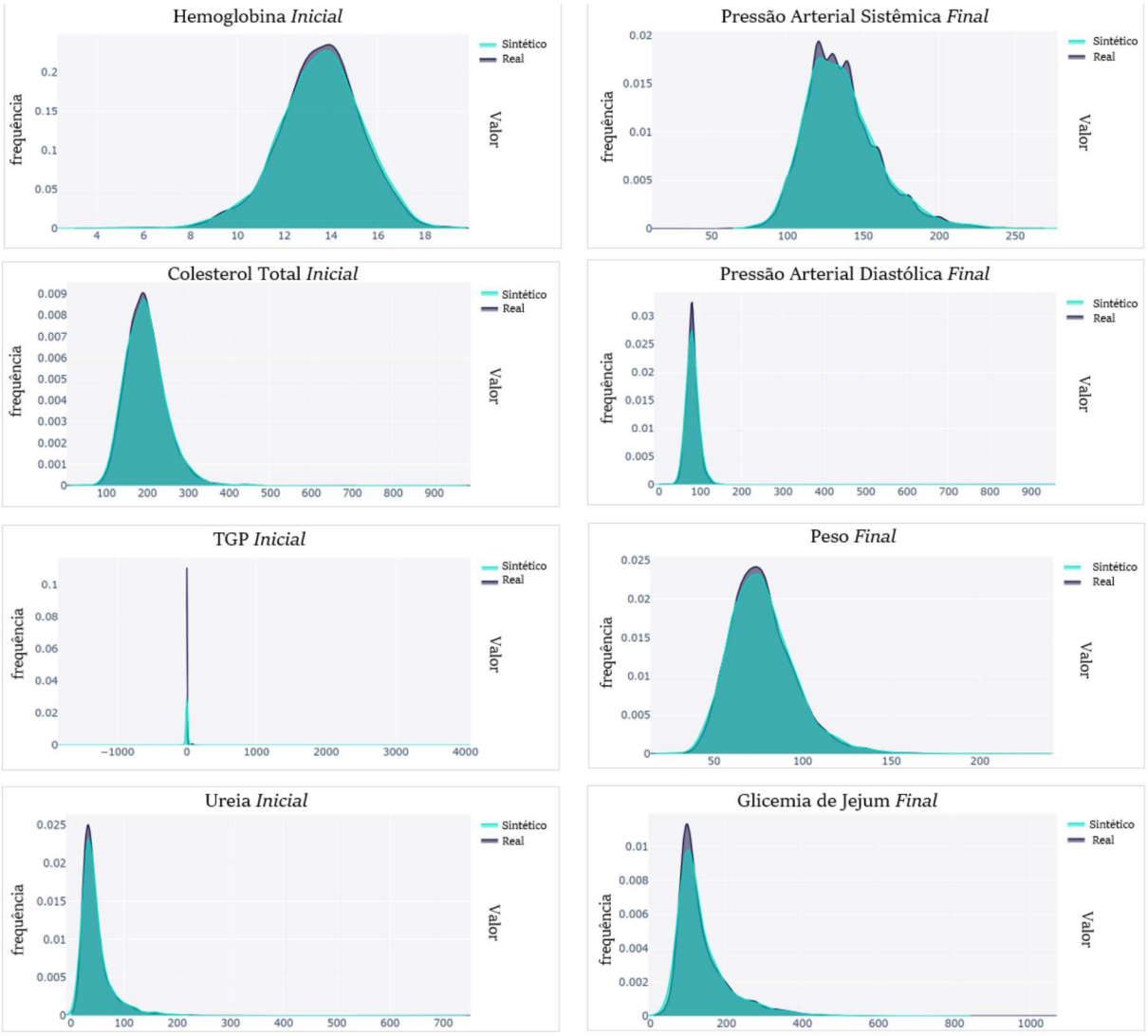
6.4 Cenário 4

6.4.1 Definição

Conforme ilustram a Tabela 10 e a Figura 20, há uma quantidade expressiva de registros faltantes na base de dados do Centro Hiperdia. Além das técnicas já utilizadas nos outros cenários como substituição por zero, pela média e pela mediana, bem como o uso do KNN e do MICE, métodos alternativos podem ser aplicados para a inferência. A quarta proposta de cenário de aplicação considerou a utilização do conceito estatístico de cópulas para a inferência de dados faltantes, de forma que as variáveis possam ter seus valores ausentes inferidos com base em distribuições estatísticas, como explicado na Subseção **4.4.7**.

Para o cenário 4, foram previamente selecionadas as 25 variáveis do cenário 2, indicadas na Tabela 13, para as quais a função `Univariate` da biblioteca `copulas` (37) foi aplicada com o objetivo de ser obtido um modelo de cópula univariada com ajuste ótimo aos dados. A partir deste modelo, foi gerada uma base com dados sintéticos do mesmo tamanho da original: 5.689 registros únicos de pacientes. Na Figura 24 são exibidos os gráficos com os valores reais e com os valores sintéticos gerados por cópulas para oito variáveis da base de dados.

Figura 24 – Gráficos com as distribuições normais para oito variáveis do conjunto de dados. Em cinza estão os valores reais e em verde os valores sintéticos gerados por cópulas.



Fonte: Elaborada pelo autor.

Os gráficos mostram que o modelo de cópula univariada reproduziu de maneira satisfatória a distribuição dos valores originais dos exames, com forte sobreposição nas regiões mais comuns. No entanto, há pequenas divergências nas caudas esquerda e direita, sobretudo nesta última, na qual os valores sintéticos são ligeiramente mais altos que os originais, indicando que ajustes nos valores gerados são necessários.

Dessa forma, para todas as variáveis consideradas, as amostras sintéticas que apresentaram valores superiores aos valores máximos observados na base de dados original foram devidamente identificadas e removidas da base de dados sintética. O total de amostras avaliadas como atípicas para cada um dos exames considerados está descrito na Tabela 22.

Tabela 22 – Total de amostras atípicas por exame na base de dados sintética.

Exame	**Amostras Atípicas**
TSHI	18
ColesterolHDLI	4
ColesterolTotalF	3
ColesterolHDLF	3
AcidoUricoI	2
HemoglobinaGlicadaI	2
PAS_inicial	1
PAS_final	1
PAD_inicial	1
PAD_final	1
GlicemiadeJejumF	1
TrigliceridesF	1
PotassioI	1
UreiaI	1
pesoi	0
pesof	0
HemoglobinaI	0
HemoglobinaF	0
ColesterolTotalI	0
GlicemiadeJejumI	0
TrigliceridesI	0
TGPI	0

Do total de 142.225 valores da base original referentes aos 25 campos considerados (Tabela 13), 17.313 ($\approx 12\%$) são nulos. O processo de substituição dos dados ausentes pelos sintéticos considerou o mapeamento linha-coluna entre as bases. Seja um valor faltante, denotado por x, e posicionado na linha i e na coluna j da base de dados original. A substituição de x é realizada por um valor sintético, y, também localizado na posição $[i, j]$, mas da base sintética. Com a finalização da nova versão da base de dados, foi aplicada a Equação 2.2 (MDRD) para todos os pacientes cujos dados foram inferidos por cópulas, com o objetivo de ser calculado o valor de TFG e, consequentemente, o campo correspondente ao URE.

Após a inferência dos dados realizada por cópulas, foram utilizados os mesmos cinco algoritmos de classificação do cenário 3 — RF, SVM, GB, AdaBoost e XGBoost — assim como as três variáveis preditoras: idade, raça e sexo. A fundamentação destas escolhas se deve ao objetivo de que seja encontrado o menor conjunto factível para a classificação dos estágios da DRC sem a consideração da creatinina sérica.

Como proposta alternativa de classificação do URE, foi implementado o método RFE para identificar as variáveis mais relevantes. Esse método, que elimina de forma recursiva as variáveis menos significativas para o modelo, foi amplamente utilizado em

estudos anteriores, como os de Debal e Sitote (2022) (65), Islam *et al.* (2023) (126) e Ogunleye e Wang (2020) (180), discutidos no Capítulo 3. No cenário 4, o objetivo foi estabelecer um conjunto ótimo de cinco variáveis por algoritmo, buscando otimizar os resultados de classificação.

Por fim, similarmente aos cenários previamente descritos, a divisão do conjunto de dados foi realizada utilizando a função `train_test_split` (219), alocando 70% dos dados para treinamento e 30% para validação em uma única execução.

O Algoritmo 9 fornece um detalhamento das etapas que compõem o processo de inferência de dados com cópulas.

Algoritmo 9 Processo de inferência de dados com cópulas

1: **Entrada**: Conjunto de dados selecionado
2: Aplicação da função `Univariate` da biblioteca `copulas` para gerar uma distribuição normal para cada um dos dados
3: Criação da base de dados sintética
4: Análise e remoção dos valores atípicos (*outliers*) da base sintética para cada exame
5: Parte da base sintética é inserida na base original para substituir os dados faltantes
6: Os valores de TFG para os dados sintéticos são calculados
7: O URE para os dados sintéticos é criado
8: **Saída**: Base de dados com todos os valores presentes

6.4.2 Resultados e Discussão

Os algoritmos RF, SVM, GB, AdaBoost e XGBoost foram novamente utilizados no processo de classificação, cujos resultados, por estágio e por métrica de avaliação, estão detalhados na Tabela 23.

A análise dos resultados de classificação revela um desempenho geral modesto e inferior aos obtidos no cenário 3. Com relação à acurácia, os melhores desempenhos foram observados com AdaBoost (35%) e GB (33%), ao passo que XGBoost, SVM e RF apresentaram valores inferiores, com 32%, 31% e 30%, respectivamente. Essas diferenças sugerem uma variação no desempenho geral dos algoritmos em função da capacidade de cada um de lidar com a estrutura dos dados após a inferência por cópulas.

Em termos de revocação, o desempenho dos algoritmos destacou-se no estágio 2, com valores significativamente superiores aos observados nos demais estágios. O SVM apresentou a melhor performance, alcançando uma revocação de 0,83, seguido pelo GB e AdaBoost, com 0,69 e 0,68, respectivamente. Esses resultados indicam que o SVM foi capaz de classificar corretamente mais de 80% dos pacientes que realmente estavam no estágio 2 da DRC. No entanto, para os estágios mais avançados, como os estágios 4 e 5, todos os algoritmos apresentaram valores de revocação extremamente baixos, podendo ser considerados irrisórios e negligenciáveis.

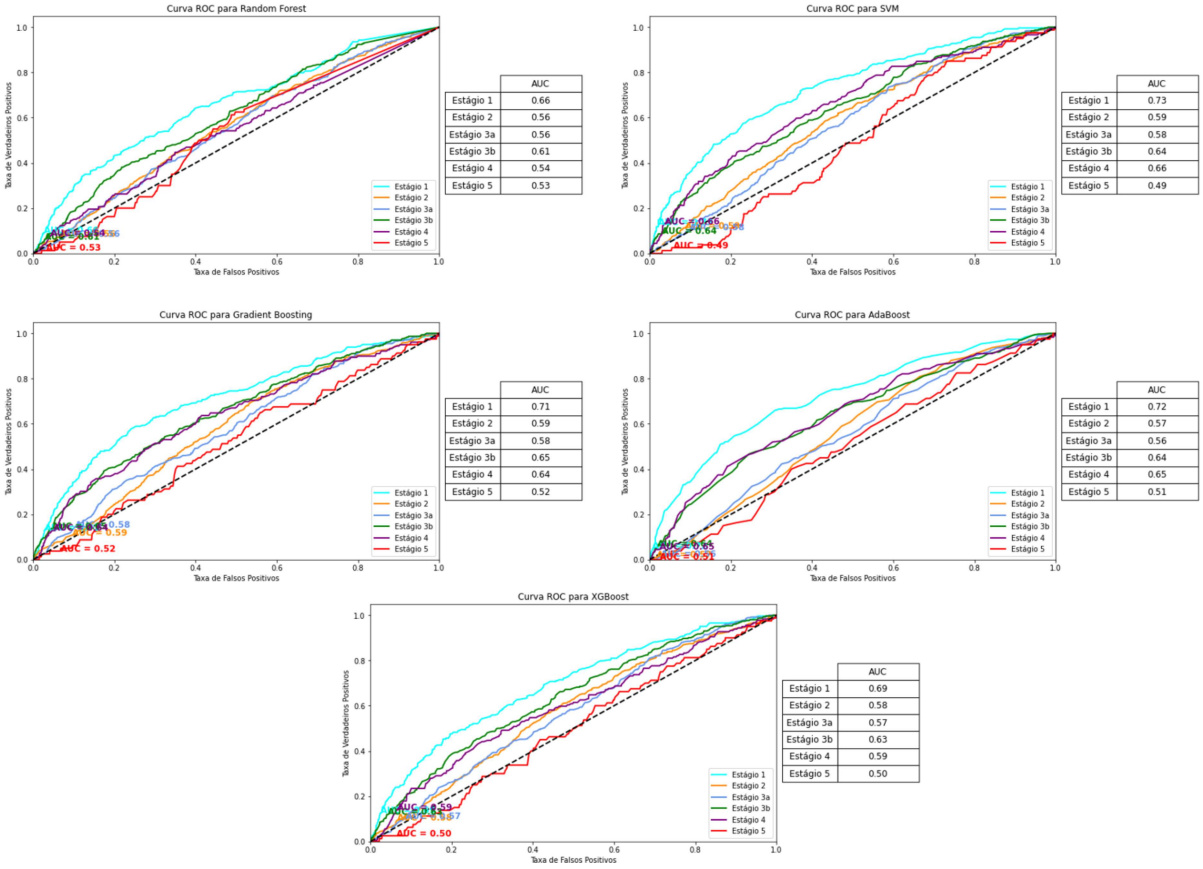
Tabela 23 – Resultados da classificação com cinco algoritmos diferentes após a inferência de dados realizada com cópulas. Para todos os casos, o conjunto suporte é composto de 272, 502, 387, 298, 171 e 77 registros (total de 1.707), respectivamente para os estágios 1, 2, 3a, 3b, 4 e 5.

	RF				**SVM**			
Estágios	**Prec.**	**Revoc.**	**F1-*score***	**AUC**	**Prec.**	**Revoc.**	**F1-*score***	**AUC**
1	0,32	0,22	0,26	0,66	0,47	0,10	0,16	0,73
2	0,33	0,51	0,40	0,56	0,32	0,83	0,47	0,59
3a	0,29	0,34	0,31	0,56	0,26	0,21	0,23	0,58
3b	0,28	0,19	0,23	0,61	0,34	0,07	0,12	0,64
4	0,17	0,09	0,12	0,54	0,00	0,00	0,00	0,66
5	0,00	0,00	0,00	0,53	0,00	0,00	0,05	0,49
Acurácia	*30%*				*31%*			

	GB				**AdaBoost**			
Estágios	**Prec.**	**Revoc.**	**F1-*score***	**AUC**	**Prec.**	**Revoc.**	**F1-*score***	**AUC**
1	0,37	0,19	0,25	0,71	0,45	0,22	0,29	0,72
2	0,36	0,69	0,47	0,59	0,37	0,68	0,48	0,57
3a	0,29	0,29	0,29	0,58	0,29	0,35	0,31	0,56
3b	0,26	0,15	0,15	0,65	0,30	0,19	0,24	0,64
4	0,14	0,02	0,02	0,64	0,75	0,02	0,03	0,65
5	0,00	0,00	0,00	0,52	0,00	0,00	0,00	0,51
Acurácia	*33%*				*35%*			

	XGBoost			
Estágios	**Prec.**	**Revoc.**	**F1-*score***	**AUC**
1	0,30	0,18	0,23	0,69
2	0,35	0,61	0,44	0,58
3a	0,27	0,30	0,28	0,57
3b	0,25	0,16	0,20	0,63
4	0,23	0,06	0,10	0,59
5	0,00	0,00	0,00	0,50
Acurácia	*32%*			

Figura 25 – Resultados da curva ROC AUC para todos os cinco algoritmos de classificação utilizados quando aplicados após a inferência realizada por cópulas.



Fonte: Elaborada pelo autor.

Com exceção do valor 0,75 obtido pelo AdaBoost no estágio 4, os valores de precisão também foram insatisfatórios para as três variáveis preditoras consideradas neste cenário, destacando a dificuldade desses algoritmos em manter um equilíbrio adequado entre falsos positivos e verdadeiros positivos. Os valores de F1-*score* e AUC reforçam a incapacidade dos algoritmos em equilibrar adequadamente precisão e revocação, especialmente nos estágios mais avançados da DRC. Os valores de AUC, que variam de 0,49 a 0,73, como ilustrado na Figura 25, indicam que, embora os modelos possuam algum nível de discriminação, seu desempenho na diferenciação das classes é limitado, particularmente nos estágios mais graves da doença. Portanto, a variação significativa no desempenho dos algoritmos entre os diferentes estágios da DRC revela uma limitação expressiva na aplicabilidade dessas abordagens em cenários clínicos reais quando considerados como variáveis preditoras apenas os campos de idade, raça e sexo.

Uma vez que os resultados para classificação com somente três variáveis foram insatisfatórios, uma outra proposta desenvolvida foi a utilização de todos os 25 exames

previamente selecionados para o cenário 2 (Tabela 13). Assim, com os resultados gerados, o foco da proposta foi avaliar a importância da contribuição de cada variável para a classificação. Seguindo os trabalhos supracitados de Ogunleye e Wang (2020) (180), Debal e Sitote (2022) (65) e Islam *et al.* (2023) (126), foi aplicada a RFE, técnica de seleção e eliminação recursiva de variáveis que, ao remover os atributos irrelevantes ou redundantes, pode melhorar a generalização de um modelo e reduzir o risco de sobreajuste.

Para cada modelo gerado, a RFE foi configurada para selecionar as cinco variáveis com maior impacto na capacidade preditiva do modelo, representadas na Tabela 24. Além das variáveis previamente consideradas e formadoras do conjunto mínimo — idade, sexo e raça — diversos exames, do conjunto de 25, foram identificados pela técnica de RFE como relevantes para a classificação, com ênfase em hemoglobina (RF, SVM, GB e XGBoost), triglicérides (RF, GB e AdaBoost) e ureia (RF, GB, AdaBoost e XGBoost).

Tabela 24 – Os cinco conjuntos de variáveis mais importantes selecionadas pelo método RFE para cada um dos algoritmos utilizados.

1) RF	**2) SVM**	**3) GB**	**4) AdaBoost**	**5) XGBoost**
UreiaI	Codsexo	Idade	TGPI	Idade
TrigliceridesF	Raça	UreiaI	TSHI	Codsexo
GlicemiadeJejumI	AcidoUricoI	TrigliceridesI	UreiaI	Raça
HemoglobinaF	PotassioI	HemoglobinaF	TrigliceridesF	UreiaI
pesof	HemoglobinaI	HemoglobinaI	GlicemiadeJejumI	HemoglobinaF

Com base nos resultados apresentados na Tabela 24, cada um dos métodos de AM foi submetido ao processo descrito no Algoritmo 9, visando à inferência de dados por meio de cópulas para a classificação do URE dos pacientes. Esse procedimento levou em consideração exclusivamente as cinco variáveis identificadas como as mais relevantes pela técnica de RFE para cada caso. Dessa forma, AdaBoost, GB, RF, SVM e XGBoost foram aplicados não apenas com suas cinco variáveis mais importantes, mas também com os outros quatro conjuntos de variáveis relevantes identificados para os demais algoritmos. Os resultados de acurácia estão apresentados na Tabela 25.

Tabela 25 – Resultados de acurácia dos algoritmos RF, SVM, GB, AdaBoost e XGBoost para cada um dos cinco conjunto de dados considerados.

	ACURÁCIA				
	RF	**SVM**	**GB**	**AdaBoost**	**XGBoost**
Conjunto 1	0,31	0,32	0,33	0,34	0,31
Conjunto 2	0,26	0,30	0,29	0,30	0,27
Conjunto 3	0,34	0,33	0,37	0,34	0,34
Conjunto 4	0,31	0,32	0,31	0,34	0,28
Conjunto 5	0,32	0,35	0,38	0,36	0,32

Em comparação aos resultados dos testes anteriores, evidenciados na Tabela 23,

em que nenhum método de seleção de variáveis foi implementado, os resultados descritos na Tabela 25 demonstraram-se insatisfatórios, não apresentando diferenças substanciais com relação aos testes anteriores. Mesmo quando foi considerado o seu próprio conjunto de cinco variáveis ótimas apontado pela RFE, cada algoritmo não foi capaz de oferecer valores resultantes significativos.

O conjunto suporte dos testes descritos na Tabela 23 foi composto, para cada um dos estágios — 1, 2, 3a, 3b, 4 e 5 — respectivamente, pelos valores: 272, 504, 384, 300, 174 e 73, totalizando 1.707 registros. A análise desses números permite concluir que o desbalanceamento do total de valores por estágio é evidente, sobretudo para os mais avançados, 4 e 5. Para contornar esse problema e fomentar a construção de mais uma abordagem, foram novamente seguidos os passos do Algoritmo 9, com a adição de uma nova etapa: a inclusão, na base resultante, de registros sintéticos de pacientes dos estágios 4 e 5, por meio de cópulas. Assim, foram gerados 135 registros para o estágio 4 e 130 registros para o estágio 5, o que ocasionou o incremento do conjunto suporte exibido na Tabela 26 para esses dois casos.

Tabela 26 – A distribuição do conjunto suporte na base original e na base com dados com registros sintéticos de pacientes dos estágios 4 e 5.

Estágio	**Suporte original**	**Suporte sintético**
1	272	272
2	504	504
3a	384	384
3b	300	300
4	174	309
5	73	203
Total	1.707	1.972

Uma vez mais, também foram selecionados os cinco algoritmos de AM utilizados neste cenário e a variável alvo foi o URE de cada paciente. Também foi aplicada a RFE para que os conjuntos ótimos com cinco variáveis fossem determinados para cada algoritmo. O resultado da RFE para as cinco variáveis mais importantes, a partir do conjunto de 25 já utilizado (Tabela 13), está detalhado na Tabela 27.

Para garantir uma exibição objetiva dos resultados, foram calculados os valores médios exclusivamente da acurácia para a classificação, considerando tanto os cinco conjuntos ótimos (Tabela 27), quanto as três variáveis do conjunto mínimo. Dessa forma, seis processos de classificação distintos foram consolidados nos resultados apresentados na Tabela 28.

Tabela 27 – Os cinco conjuntos de variáveis mais importantes selecionadas pelo método RFE para cada um dos algoritmos utilizados.

1) RF	2) SVM	3) GB	4) AdaBoost	5) XGBoost
Idade	TrigliceridesI	Idade	Codsexo	TGPI
TrigliceridesI	UreiaI	Raça	Raça	Codsexo
PotassioI	Codsexo	ColesterolTotalF	UreiaI	GlicemiadeJejumF
Codsexo	GlicemiadeJejumF	TGPI	PotassioI	GlicemiadeJejumI
Raça	PotassioI	Codsexo	TriglicéridesI	PAD_final

Tabela 28 – Resultados médios de acurácia para a classificação com RF, SVM, GB, AdaBoost e XGBoost considerando os cinco conjuntos ótimos obtidos via RFE e o conjunto mínimo de três variáveis, após a inferência de dados realizada com cópulas. Para todos os casos, o conjunto suporte está quantificado na Tabela 26.

ACURÁCIA MÉDIA (%)				
RF	**SVM**	**GB**	**AdaBoost**	**XGBoost**
35	28	36	37	35

Os valores resultantes evidenciam a incapacidade das abordagens empregadas em classificar adequadamente os estágios da DRC. Mesmo após todos os tratamentos de dados e aplicação dos métodos detalhados, com destaque para o conceito de cópulas, as acurácias médias alcançadas foram inferiores às observadas em todos os cenários e abordagens anteriores.

6.5 Cenário 5

6.5.1 Definição

Os estudos de Sharma *et al.* (2016) (224) e Iftikhar *et al.* (2023) (123) ressaltaram que os métodos diagnósticos atuais, que se baseiam em marcadores tradicionais como a creatinina sérica, albuminúria e a TFG, embora eficazes na identificação dos estágios da DRC, apresentam sensibilidade limitada para identificar sinais iniciais de danos renais, comprometendo o diagnóstico precoce. Essa lacuna ressalta a necessidade de abordagens inovadoras que permitam a identificação de indivíduos em risco de DRC em fases iniciais, quando intervenções podem ser mais eficazes.

A partir da reorganização da base de dados em função das classes e das etapas de filtragem dos dados descritas na Subseção *5.1.2.2*, a proposta do cenário de aplicação 5[3] consistiu no desenvolvimento de processos de tratamento e de classificação que incorporassem a maioria dos métodos aplicados nos demais cenários. Considerando a redução da

[3] A reorganização da base de dados descrita na Subseção *5.1.2.2* fez parte do trabalho publicado em 2023 no volume 1052 da *Lecture Notes in Networks and Systems* (98).

complexidade da base proporcionada pela divisão em três classes e a diminuição do total de atributos, o cenário 5 foi o único a não considerar o total integral de 5.689 pacientes, já que apenas 794 restaram após os processamentos realizados.

Como o uso da creatinina sérica foi desconsiderado, este foi o único cenário que não fez uso de alguma variável diretamente relacionada à obtenção do valor da TFG, por meio da Equação 2.2 (MDRD). Dessa forma, o cenário 5 inicialmente empregou 18 dos 19 exames descritos na Tabela 9. Ademais, como todos os valores das variáveis se referem a registros iniciais, a variável alvo deste cenário foi o primeiro estágio registrado (PRE) de um paciente, diferentemente das abordagens anteriores que consideraram o URE.

E para a elaboração da abordagem de classificação do PRE, os métodos KNN, MICE, cópulas e substituição pela média foram utilizados para a inferência dos dados ausentes. E, para o balanceamento das classes envolvidas, o SMOTE foi o selecionado após os testes realizados no cenário 3. Com relação aos 18 exames, novamente foi utilizada a RFE para a determinação das cinco variáveis mais influentes para cada um deles e de acordo com cada um dos cinco algoritmos de AM utilizados: RF, SVM, GB, AdaBoost e XGBoost.

Em oposição aos quatro cenários anteriores, para a separação dos conjuntos de treinamento e teste foi aplicado o método *stratified k-fold* (SKF), por meio da função homônima `StratifiedKFold` (218), da biblioteca `Scikit-learn` (188). O objetivo foi maximizar a eficiência da divisão dos conjuntos de dados e aprimorar a efetividade dos modelos gerados. Embora os estudos revisados no Capítulo 3 (10) (9) (65) (105) (126) e (191) tenham implementado a validação cruzada com 10 subconjuntos, o SKF é recomendado para problemas de classificação com classes desbalanceadas, uma vez que garante que o modelo seja testado em subconjuntos (*folds*) representativos da distribuição real das classes (247). Logo, em todas as abordagens de classificações desenvolvidas no cenário 5, os conjuntos de treinamento e teste foram separados por meio do SKF com 10 subconjuntos. Além disso, o parâmetro `shuffle` da função `StratifiedKFold` (218) foi ativado em todos os casos, de forma a ocorrer o embaralhamento dos dados antes da divisão em subconjuntos, evitando qualquer ordem subjacente nos dados que possa afetar o desempenho do modelo.

De maneira similar aos cenários 3 e 4, o principal objetivo do cenário 5 foi investigar uma possível solução para um dos objetivos deste trabalho: identificar um conjunto mínimo de exames clínicos capazes de predizer os estágios dos pacientes com DRC. Para alcançar esse objetivo, a base de dados foi reorganizada (como detalhado na Subseção *5.1.2.2*), foram aplicadas várias técnicas de inferência de dados, realizadas divisões dos conjuntos com SKF para a classificação, identificadas as variáveis mais influentes em todos os algoritmos e conduzida a classificação do PRE dos pacientes do CH. Ademais, o cenário 5 ficou restrito à análise do primeiro registro dos exames clínicos na base de dados, os quais também não

apresentam relação direta com a TFG. Finalmente, após os resultados da aplicação da RFE para os 18 exames, os cinco algoritmos de classificação foram aplicados na base de dados com 794 pacientes, tendo como variáveis preditoras os 5 exames identificados como os mais influentes para a classificação do PRE.

6.5.2 Resultados e Discussão

Na etapa inicial de implementação da proposta, a classificação do PRE foi realizada por meio da utilização de todas as 18 variáveis. Os resultados obtidos estão detalhados na Tabela 29. Cabe destacar que, além do SMOTE, os outros quatro métodos de balanceamento utilizados nos demais cenários também foram implementados no cenário 5. Contudo, como os resultados obtidos com esses quatro métodos foram equivalentes aos obtidos com o SMOTE em todos os casos, e para tornar a exibição dos achados e das respectivas análises mais objetiva, somente os indicadores obtidos com o SMOTE foram inclusos na Tabela 29.

Os resultados evidenciam que os melhores desempenhos baseados nas métricas utilizadas ficaram com os algoritmos RF e SVM, sobretudo para acurácia e revocação. Ademais, a introdução de cópulas não gerou o efeito esperado para nenhum algoritmo, uma vez que os resultados gerados para todas as métricas possuem valores praticamente idênticos aos dos demais métodos de inferência. Nesta abordagem, o pior desempenho ficou por conta dos três algoritmos de *ensemble*, com valores não promissores para nenhuma métrica.

Os resultados das métricas de avaliação indicam que, considerando os 18 exames analisados, os cinco algoritmos de classificação demonstraram uma variação mínima de desempenho, exibindo um comportamento praticamente indiferente aos métodos de inferência de dados e de balanceamento aplicados, mesmo com o uso do SKF na divisão dos conjuntos de treinamento e teste. Essa constatação sugere que as técnicas de tratamento de dados faltantes e de balanceamento de classes não influenciaram significativamente o desempenho dos modelos no contexto do cenário 5.

Uma vez que os resultados obtidos na primeira abordagem foram mais uma vez insatisfatórios, a RFE foi novamente aplicada para todas as abordagens implementadas na etapa anterior. Assim, para cada algoritmo e os respectivos métodos de inferência ou de balanceamento empregados, foram identificadas as cinco variáveis mais revelantes para as classificações implementadas com os cinco algoritmos. Os conjuntos de variáveis para cada caso estão detalhados na Tabela 30.

A Tabela 31 apresenta os resultados de desempenho de cada algoritmo de classificação aplicado aos respectivos conjuntos de cinco variáveis selecionadas por meio do método RFE, considerando também os cinco métodos distintos de inferência de dados e de balanceamento.

Tabela 29 – Resultados médios das métricas de avaliação para os algoritmos RF, SVM, GB, AdaBoost e XGBoost quando utilizadas as inferências por média, KNN, MICE e Cópulas, e balanceamento dos dados por SMOTE, no conjunto de dados com 18 exames. Os valores das métricas estão acompanhados dos respectivos desvios padrão.

RF - VALORES MÉDIOS POR MÉTODO					
	Acurácia	**Precisão**	**Revocação**	**F1-*score***	**AUC**
Média	0,37 ± 0,03	0,35 ± 0,02	0,37 ± 0,03	0,35 ± 0,02	0,61 ± 0,03
KNN	0,37 ± 0,04	0,35 ± 0,05	0,37 ± 0,04	0,35 ± 0,04	0,61 ± 0,04
MICE	0,37 ± 0,04	0,35 ± 0,04	0,37 ± 0,04	0,35 ± 0,03	0,61 ± 0,03
Cópulas	0,38 ± 0,04	0,34 ± 0,04	0,38 ± 0,04	0,35 ± 0,04	0,61 ± 0,03
SMOTE	0,34 ± 0,04	0,35 ± 0,02	0,37 ± 0,03	0,35 ± 0,02	0,62 ± 0,04
SVM - VALORES MÉDIOS POR MÉTODO					
	Acurácia	**Precisão**	**Revocação**	**F1-*score***	**AUC**
Média	0,41 ± 0,04	0,34 ± 0,04	0,41 ± 0,04	0,37 ± 0,04	0,66 ± 0,05
KNN	0,41 ± 0,04	0,34 ± 0,04	0,41 ± 0,04	0,37 ± 0,04	0,66 ± 0,05
MICE	0,41 ± 0,04	0,34 ± 0,04	0,41 ± 0,04	0,37 ± 0,04	0,66 ± 0,05
Cópulas	0,41 ± 0,05	0,34 ± 0,05	0,41 ± 0,05	0,36 ± 0,05	0,67 ± 0,05
SMOTE	0,34 ± 0,03	0,38 ± 0,04	0,34 ± 0,03	0,35 ± 0,03	0,63 ± 0,04
GB - VALORES MÉDIOS POR MÉTODO					
	Acurácia	**Precisão**	**Revocação**	**F1-*score***	**AUC**
Média	0,35 ± 0,04	0,33 ± 0,03	0,35 ± 0,04	0,34 ± 0,03	0,62 ± 0,03
KNN	0,35 ± 0,04	0,33 ± 0,03	0,35 ± 0,04	0,33 ± 0,03	0,62 ± 0,03
MICE	0,35 ± 0,04	0,33 ± 0,04	0,35 ± 0,04	0,33 ± 0,03	0,62 ± 0,03
Cópulas	0,36 ± 0,04	0,34 ± 0,04	0,36 ± 0,04	0,34 ± 0,04	0,63 ± 0,03
SMOTE	0,33 ± 0,03	0,35 ± 0,03	0,33 ± 0,03	0,33 ± 0,03	0,62 ± 0,04
AdaBoost - VALORES MÉDIOS POR MÉTODO					
	Acurácia	**Precisão**	**Revocação**	**F1-*score***	**AUC**
Média	0,30 ± 0,07	0,35 ± 0,07	0,30 ± 0,07	0,31 ± 0,07	0,56 ± 0,03
KNN	0,30 ± 0,07	0,35 ± 0,07	0,30 ± 0,07	0,31 ± 0,07	0,56 ± 0,03
MICE	0,30 ± 0,07	0,35 ± 0,07	0,30 ± 0,07	0,31 ± 0,07	0,56 ± 0,03
Cópulas	0,30 ± 0,05	0,33 ± 0,07	0,30 ± 0,05	0,31 ± 0,05	0,57 ± 0,02
SMOTE	0,27 ± 0,07	0,33 ± 0,06	0,27 ± 0,05	0,29 ± 0,05	0,56 ± 0,03
XGBoost - VALORES MÉDIOS POR MÉTODO					
	Acurácia	**Precisão**	**Revocação**	**F1-*score***	**AUC**
Média	0,33 ± 0,03	0,32 ± 0,03	0,33 ± 0,03	0,32 ± 0,03	0,60 ± 0,04
KNN	0,33 ± 0,03	0,32 ± 0,03	0,33 ± 0,03	0,32 ± 0,03	0,60 ± 0,04
MICE	0,33 ± 0,03	0,32 ± 0,03	0,33 ± 0,03	0,32 ± 0,03	0,60 ± 0,04
Cópulas	0,33 ± 0,03	0,32 ± 0,03	0,33 ± 0,03	0,32 ± 0,03	0,60 ± 0,04
SMOTE	0,34 ± 0,03	0,35 ± 0,04	0,34 ± 0,04	0,34 ± 0,04	0,62 ± 0,04

Tabela 30 – Os conjuntos de cinco exames selecionados por meio da técnica de RFE e que foram obtidos para cada método de inferência de dados faltantes, em combinação com cada um dos algoritmos de classificação utilizados.

	RF	**SVM**	**GB**	**AdaBoost**	**XGBoost**
Média	CalcioTotalI, ColesterolHDLI, FosforoI, PAD_inicial e SodioSericoI	AcidoUricoI, HemoglobinaI, PotassioI, TGPI e UreiaI	HemoglobinaI, PAD_inicial, Proteinuria24hsI, TSHI e UreiaI	ColesterolHDLI, FosforoI, HemoglobinaI, SodioSericoI e UreiaI	CalcioTotalI, Proteinuria24hsI, SodioUrinarioI, TSHI e UreiaI
KNN	CalcioTotalI, ColesterolHDLI, FosforoI, Proteinuria24hsI e SodioSericoI	AcidoUricoI, HemoglobinaI, PotassioI, TGPI e UreiaI	CalcioTotalI, HemoglobinaI, Proteinuria24hsI, TSHI e UreiaI	ColesterolHDLI, FosforoI, HemoglobinaI, SodioSericoI e UreiaI	CalcioTotalI, Proteinuria24hsI, SodioUrinarioI, TSHI e UreiaI
MICE	ColesterolHDLI, FosforoI, PAD_inicial, Proteinuria24hsI e SodioSericoI	AcidoUricoI, HemoglobinaI, PotassioI, TGPI e UreiaI	HemoglobinaI, PAD_inicial, Proteinuria24hsI, TSHI e UreiaI	ColesterolHDLI, FosforoI, HemoglobinaI, SodioSericoI e UreiaI	CalcioTotalI, Proteinuria24hsI, UreiaI, TSHI e UreiaI
Cópulas	CalcioTotalI, ColesterolHDLI, FosforoI, PAD_inicial e SodioSericoI	AcidoUricoI, HemoglobinaI, PotassioI, TGPI e UreiaI	AcidoUricoI, ColesterolTotalI, HemoglobinaI, PAS_inicial e UreiaI	AcidoUricoI, HemoglobinaI, Proteinuria24hsI, SodioSericoI e UreiaI	AcidoUricoI, CalcioTotalI, GlicemiadeJejumI, HemoglobinaI e UreiaI
SMOTE	ColesterolHDLI, FosforoI, PAD_inicial, Proteinuria24hsI e SodioSericoI	AcidoUricoI, HemoglobinaI, PotassioI, TGPI e UreiaI	HemoglobinaI, PAD_inicial, Proteinuria24hsI, TSHI e UreiaI	ColesterolHDLI, FosforoI, HemoglobinaI, SodioSericoI e UreiaI	CalcioTotalI, Proteinuria24hsI, SodioUrinarioI TSHI e UreiaI

Os resultados obtidos nesta etapa não apresentam diferenças substanciais para os resultados da etapa anterior. Mesmo que em uma modesta proporção, os valores de acurácia apresentaram crescimento para os três algoritmos de *ensemble*: GB, AdaBoost e XGBoost. Possivelmente, a seleção de variáveis mais importantes foi responsável pela redução da propagação de erros ao longo das etapas de execução desses métodos, fato que pode ter prejudicado o desempenho na primeira etapa.

Uma vez mais, a aplicação do SMOTE originou os piores resultados para a maior parcela dos algoritmos. Os valores obtidos na primeira etapa são muito semelhantes aos valores da segunda etapa, com variações irrisórias para todas as métricas e dentro dos intervalos de desvios padrão indicados.

Em comparação aos resultados da Tabela 29, nesta etapa, o RF teve uma redução considerável para todas métricas na abordagem de inferência por cópulas. Uma possível explicação para esse comportamento pode ser o fato das cópulas serem capazes de gerar dependências não lineares complexas, gerando padrões difíceis de serem capturadas pelas árvores de decisão individuais do RF. Como consequência, esse algoritmo pode gerar um modelo com baixa eficiência em suas regras de classificação e, portanto, limitar a capacidade de realizar previsões precisas. Por outro lado, o SVM destacou-se ao apresentar o melhor desempenho global, atingindo uma acurácia de até 44% e uma revocação de 41%. Esses resultados foram especialmente evidentes quando a inferência foi realizada com KNN e com cópulas.

Tabela 31 – Resultados médios das métricas de avaliação para os algoritmos RF, SVM, GB, AdaBoost e XGBoost quando utilizadas as inferências por média, KNN, MICE e Cópulas, e balanceamento dos dados por SMOTE, no conjunto de dados com 5 exames determinados pela RFE. Os valores das métricas estão acompanhados dos respectivos desvios padrão.

RF - VALORES MÉDIOS POR MÉTODO					
	Acurácia	**Precisão**	**Revocação**	**F1-*score***	**AUC**
Média	0,35 ± 0,04	0,35 ± 0,04	0,37 ± 0,04	0,37 ± 0,04	0,61 ± 0,04
KNN	0,37 ± 0,03	0,35 ± 0,04	0,39 ± 0,03	0,39 ± 0,03	0,64 ± 0,03
MICE	0,39 ± 0,04	0,38 ± 0,04	0,39 ± 0,04	0,37 ± 0,04	0,62 ± 0,04
Cópulas	0,28 ± 0,05	0,24 ± 0,06	0,28 ± 0,05	0,25 ± 0,04	0,49 ± 0,03
SMOTE	0,35 ± 0,05	0,36 ± 0,05	0,35 ± 0,05	0,35 ± 0,05	0,61 ± 0,03
SVM - VALORES MÉDIOS POR MÉTODO					
	Acurácia	**Precisão**	**Revocação**	**F1-*score***	**AUC**
Média	0,39 ± 0,04	0,34 ± 0,04	0,41 ± 0,04	0,38 ± 0,04	0,67 ± 0,05
KNN	0,44 ± 0,04	0,40 ± 0,04	0,41 ± 0,04	0,39 ± 0,04	0,69 ± 0,05
MICE	0,43 ± 0,04	0,39 ± 0,04	0,40 ± 0,04	0,37 ± 0,04	0,66 ± 0,05
Cópulas	0,44 ± 0,05	0,38 ± 0,06	0,40 ± 0,05	0,35 ± 0,05	0,64 ± 0,04
SMOTE	0,37 ± 0,03	0,38 ± 0,04	0,34 ± 0,03	0,35 ± 0,03	0,63 ± 0,04
GB - VALORES MÉDIOS POR MÉTODO					
	Acurácia	**Precisão**	**Revocação**	**F1-*score***	**AUC**
Média	0,39 ± 0,03	0,33 ± 0,03	0,35 ± 0,03	0,33 ± 0,03	0,60 ± 0,03
KNN	0,38 ± 0,04	0,32 ± 0,04	0,35 ± 0,04	0,33 ± 0,03	0,61 ± 0,03
MICE	0,35 ± 0,03	0,35 ± 0,03	0,38 ± 0,03	0,33 ± 0,03	0,63 ± 0,03
Cópulas	0,37 ± 0,05	0,31 ± 0,04	0,37 ± 0,05	0,34 ± 0,05	0,60 ± 0,04
SMOTE	0,33 ± 0,03	0,34 ± 0,03	0,33 ± 0,03	0,33 ± 0,03	0,62 ± 0,04
AdaBoost - VALORES MÉDIOS POR MÉTODO					
	Acurácia	**Precisão**	**Revocação**	**F1-*score***	**AUC**
Média	0,34 ± 0,07	0,34 ± 0,07	0,30 ± 0,07	0,31 ± 0,07	0,56 ± 0,03
KNN	0,33 ± 0,07	0,34 ± 0,07	0,31 ± 0,07	0,31 ± 0,07	0,57 ± 0,03
MICE	0,31 ± 0,07	0,35 ± 0,07	0,34 ± 0,07	0,31 ± 0,07	0,56 ± 0,03
Cópulas	0,32 ± 0,03	0,32 ± 0,03	0,32 ± 0,03	0,31 ± 0,02	0,59 ± 0,03
SMOTE	0,29 ± 0,05	0,33 ± 0,06	0,27 ± 0,05	0,29 ± 0,05	0,56 ± 0,03
XGBoost - VALORES MÉDIOS POR MÉTODO					
	Acurácia	**Precisão**	**Revocação**	**F1-*score***	**AUC**
Média	0,37 ± 0,03	0,32 ± 0,03	0,33 ± 0,03	0,32 ± 0,03	0,60 ± 0,04
KNN	0,33 ± 0,03	0,34 ± 0,03	0,33 ± 0,03	0,32 ± 0,03	0,60 ± 0,04
MICE	0,35 ± 0,03	0,32 ± 0,03	0,33 ± 0,03	0,32 ± 0,03	0,60 ± 0,04
Cópulas	0,36 ± 0,05	0,35 ± 0,06	0,37 ± 0,05	0,35 ± 0,06	0,62 ± 0,05
SMOTE	0,34 ± 0,04	0,35 ± 0,04	0,34 ± 0,04	0,34 ± 0,04	0,62 ± 0,04

Os resultados das duas etapas evidenciam que o cenário 5, embora tenha se caracterizado pela implementação de uma abordagem completa que envolveu etapas de organização, análise, tratamento e classificação dos dados, não atingiu os resultados esperados. Ainda que os 18 exames tenham sido inclusos no conjunto de variáveis preditoras, os indicadores resultantes foram insatisfatórios. Da mesma forma, a abordagem que considerou os cinco conjuntos ótimos de variáveis também não foi bem sucedida.

Para este cenário, a reorganização da base de dados em função das classes de gravidade dos exames não produziu a melhoria esperada de redução da complexidade dos dados. As inferências realizadas por KNN, MICE, cópulas e por substituição pela média, e o balanceamento por SMOTE (e pelas outras quatro técnicas descritas nesta tese) não foram capazes de melhorar o desempenho de todos os algoritmos, mesmo com a seleção de atributos realizada via RFE.

É fundamental destacar que o cenário 5 foi o único a utilizar um total reduzido de pacientes, 794, e somente exames relativos aos primeiros registros para cada paciente tendo, portanto, o PRE como variável preditora. Os resultados sugerem que a predição do estágio inicial da DRC considerando somente exames primários não é eficiente para os pacientes da base de dados do CH, pelo menos em abordagens que não considerem a creatinina sérica como variável preditora.

6.6 Resumo

A Tabela 32 exibe um resumo dos principais detalhes acerca da composição de cada um dos cinco cenários:

- Organização da base de dados;
- Total de pacientes;
- Total de variáveis;
- Inclusão da creatinina sérica no conjunto de variáveis preditoras;
- Algoritmos de classificação;
- Métodos de inferência de dados faltantes;
- Detalhes adicionais acerca dos tratamentos e dos métodos utilizados;
- Métricas de avaliação dos resultados obtidos.

Tabela 32 – Resumo da composição de cada um dos cinco cenários.

RESUMO DOS RESULTADOS PARA OS CINCO CENÁRIOS

Cenário	Base	Pacientes	Variáveis	Creatinina	Algoritmos	Inferência	Outros	Resultados
1	Por data	5.689	4 a 50	Sim e Não	RF	Zeros, média, mediana, KNN e MICE	4 abordagens	Acc: 68% a 99%
2	Por data	5.689	25	Não	ELM, KNN, LogR, MLP, SVM e XGBoost	Zeros	Baseado no total de dados dos trabalhos da base da UCI	Acc: 25% a 96%
3	Original	5.689	3	Não	AdaBoost, GB, RF, SVM e XGBoost	Zeros, média, mediana, KNN e MICE	5 métodos de balanceamento	Acc: 21% a 46% Rev: 34% a 64% AUC: 0,77 a 0,94
4	Original	5.689	3 e 5	Não	AdaBoost, GB, RF, SVM e XGBoost	Cópulas	Incremento do conjunto suporte com cópulas	Acc: 28% a 37% AUC: 0,49 a 0,76
5	Por classe	794	5	Não	AdaBoost, GB, RF, SVM e XGBoost	Média, Cópulas, KNN e MICE	Exames iniciais, balanceamento, cópulas, RFE e SKF	Acc: 28% a 44% AUC: 0,49 a 0,69

7 CONCLUSÃO

A doença renal crônica apresenta alta prevalência global, caracterizando-se frequentemente pela ausência de sintomas e sendo comumente diagnosticada em estágios avançados. A predição de seus estágios é fundamental para otimizar tanto o tratamento quanto o acompanhamento clínico dos pacientes. Ademais, intervenções precoces são as estratégias mais eficazes na redução dos custos relacionados ao tratamento, consolidando a antecipação do diagnóstico como um fator essencial no enfrentamento da doença enquanto questão de saúde pública.

O principal objetivo deste trabalho foi desenvolver cenários de aplicação de algoritmos e técnicas de aprendizado de máquina para a predição dos estágios da DRC, a partir de uma base de dados proveniente do sistema de saúde público do Brasil. Esses cenários foram elaborados com o propósito de identificar conjuntos de exames e dados pessoais que fossem clinicamente viáveis para a detecção dos estágios da DRC, sem depender, de forma geral, de marcadores e exames tradicionais, como a creatinina sérica, adotando, assim, uma abordagem inovadora.

O primeiro cenário desenvolvido teve como objetivo a realização de análises preliminares da base de dados, organizada em função das datas de realização dos exames de cada paciente. Foram elaboradas quatro abordagens distintas para a classificação do URE dos pacientes com diferentes agrupamentos de variáveis — com e sem a creatinina sérica —, cinco técnicas de inferência de dados e tendo o RF como algoritmo de classificação. Em todas as abordagens que utilizaram a creatinina como variável, os valores de acurácia alcançados foram significativamente elevados, variando entre 84% e 99%. A única abordagem que não incluiu essa variável apresentou os menores índices de acurácia, situando-se na faixa de 68% a 69%, evidenciando, assim, o potencial de uso da creatinina como variável preditora para a base de dados considerada.

A partir do segundo cenário, todas as classificações desenvolvidas descartaram a creatinina do conjunto de variáveis preditoras, de forma que o objetivo geral deste trabalho fosse respondido. Mantendo-se a mesma organização dos dados utilizada no cenário 1, foram selecionadas 25 variáveis e seis algoritmos de classificação, visando uma compreensão mais aprofundada do comportamento dos dados. Embora alguns algoritmos tenham apresentado resultados de razoável valor agregado para alguns estágios específicos, apenas o XGBoost gerou indicadores consistentes e de qualidade para todos os estágios. Os valores de acurácia, precisão e revocação obtidos pelo algoritmo foram elevados, destacando-se pela equivalência com resultados documentados na literatura. O principal benefício derivado desse cenário foi a elaboração de uma abordagem preditiva robusta, sem a inclusão da creatinina sérica, uma metodologia que se destaca por não ter precedentes na literatura revisada e que atende ao objetivo central deste trabalho, no contexto da base de dados do

Centro Hiperdia.

No cenário 3, o foco foi responder um dos objetivos deste trabalho: identificar o menor conjunto factível de variáveis pessoais, clínicas e laboratoriais capazes de predizer os estágios da DRC, excluindo a creatinina sérica. O XGBoost, que apresentou os melhores resultados no cenário 2, foi complementado pelo RF, SVM e por dois algoritmos de *ensemble*: AdaBoost e GB. Foram utilizados os mesmos métodos de inferência do cenário 1 e o desbalanceamento dos dados foi tratado com cinco métodos diferentes: ADASYN, SMOTE, SMOTE-Tomek, SMOTE-ENN e *Borderline*-SMOTE. Embora os resultados gerais tenham sido insatisfatórios, as técnicas de balanceamento proporcionaram uma melhoria discreta nas métricas de desempenho em geral, sobretudo para os estágios 1, 4 e 5. Já para os demais, os resultados foram inconsistentes, sobretudo pela queda acentuada na revocação após o balanceamento. No entanto, com o uso de apenas três variáveis mínimas (idade, raça e sexo), foi possível alcançar valores promissores, particularmente para a curva ROC AUC e para a revocação com o XGBoost, especialmente no estágio 1, sugerindo que essa abordagem pode ser eficaz na identificação precoce da DRC neste estágio introdutório, permitindo um encaminhamento mais assertivo dos pacientes para tratamentos iniciais, como intervenções medicamentosas.

No quarto cenário, uma nova base de dados foi gerada por meio da aplicação do conceito de cópulas. Os dados sintéticos foram utilizados tanto para a inferência de dados ausentes, quanto para o aumento do conjunto de suporte da classificação, a qual foi realizada com os mesmos cinco algoritmos do cenário anterior. Na primeira abordagem, somente com a inferência por cópulas, os resultados para todos os métodos foram insatisfatórios, com valores baixos para as métricas consideradas e em todos os estágios. Houve, em particular, uma significativa variação nos resultados de revocação e precisão, evidenciando uma considerável dificuldade dos algoritmos empregados em equilibrar falsos positivos e falsos negativos, sendo estes últimos os casos mais preocupantes. Por meio da análise do desbalanceamento do conjunto suporte entre os seis estágios, foram inclusas na base de dados amostras referente às classes minoritárias, os estágios 4 e 5. Foi aplicado o método RFE para seleção das cinco variáveis mais relevantes para a classificação de cada algoritmo. No entanto, os resultados obtidos com a seleção de variáveis também foram insatisfatórios, não apresentando melhorias substanciais em relação aos cenários anteriores. Mesmo após a geração de registros sintéticos para os estágios 4 e 5 por meio de cópulas, as acurácias médias permaneceram abaixo das expectativas, evidenciando a limitação das abordagens empregadas na classificação adequada dos estágios da DRC quando apenas as variáveis referentes à idade, à raça e ao sexo dos pacientes foram consideradas na classificação.

Assim como nos cenários anteriores, o cenário 5 teve como objetivo identificar um conjunto mínimo e viável de variáveis preditoras. Para alcançar esse objetivo, após as tentativas realizadas nas abordagens anteriores, foi proposta uma terceira reorganização da base de dados, agora estruturada em função das classes de gravidade de 18 exames e

reduzida a um total de 794 pacientes, após um processo de tratamento dos dados. Além dos métodos de inferência, classificação e seleção de variáveis aplicados nos cenários 3 e 4, a divisão dos dados em conjuntos de treinamento e teste foi realizada através do método SKF com dez subconjuntos. Os resultados das classificações, tanto para o conjunto completo de 18 variáveis, quanto para os conjuntos reduzidos de cinco variáveis selecionadas via RFE, não atingiram as expectativas. Os valores obtidos foram insuficientes para a maioria das métricas de desempenho, independentemente do método de inferência ou de balanceamento empregado. Esses achados indicam que a abordagem de predição baseada exclusivamente nos exames iniciais não foi eficaz para os pacientes da base de dados do CH. Logo, neste contexto, o uso de exames primários, sem a inclusão da creatinina sérica, pode não ser adequado para fornecer predições precisas sobre o estágio inicial da doença.

A análise dos resultados obtidos nos cinco cenários evidencia, de forma inequívoca, a relevância da creatinina na predição dos estágios da DRC, conforme amplamente reportado na literatura especializada. No primeiro cenário, três das abordagens incorporaram o exame de creatinina, gerando resultados consistentes e em concordância com os valores encontrados em estudos correlatos. No segundo cenário, sem o uso da creatinina, mas empregando 25 variáveis, incluindo dados pessoais e exames clínicos, a classificação realizada atingiu acurácia de 96% na predição geral de todos os estágios. Este desempenho se alinha aos valores reportados na literatura, embora todos os estudos revisados neste trabalho incluam o uso da creatinina sérica em suas análises. Portanto, trata-se de uma abordagem potencialmente inovadora, que poderia ser testada na detecção dos seis estágios da DRC na prática clínica a partir da obtenção dos exames indicados na Tabela 13, possibilitando o encaminhamento e o acompanhamento adequados dos pacientes.

O conjunto mínimo de três variáveis relacionadas à TFG — idade, raça e sexo — não apresentou desempenho satisfatório na predição dos estágios da DRC, com exceção de alguns casos isolados observados no cenário 3. Nesses, especialmente em relação ao estágio 1, os resultados mostraram-se promissores em certa medida, com destaque para as métricas de revocação e ROC AUC. Esse desempenho sugere uma possível redução de falsos negativos, evitando que pacientes com a doença não recebam o diagnóstico correto. Como destacado por Sharma *et al.* (2016) (224) e Iftikhar *et al.* (2023) (123), os métodos diagnósticos atuais, baseados na creatinina, detectam a DRC em estágios avançados, mas apresentam limitações na identificação da condição em seus estágios iniciais. Contudo, mesmo considerando apenas três dados de fácil obtenção para cada paciente, e após o aprofundamento do método desenvolvido no cenário 3, é possível alcançar resultados significativos na predição da DRC em estágios iniciais. Isso possibilitaria a identificação precoce desses pacientes e o encaminhamento para tratamentos apropriados, com potencial enfoque em intervenções medicamentosas.

Os diferentes conjuntos mínimos compostos por cinco exames nos cenários 4 e 5, apesar da aplicação de métodos de reorganização da base de dados, balanceamento,

inferência e seleção de variáveis, não apresentaram resultados satisfatórios em relação à predição de nenhum dos estágios da doença. Portanto, nesses dois cenários, a proposta de identificação de um conjunto mínimo viável mostrou-se falha. Dentro desse contexto, podem ser citadas algumas limitações do uso de técnicas de AM para a detecção da DRC, segundo Delrue *et al.* (2024) (68): o sobreajuste, a necessidade de dados de alta qualidade e a complexidade dos modelos, que podem dificultar a interpretação pelos clínicos. A falta de padronização e a necessidade de validação externa são obstáculos importantes a serem superados.

Considerando a vasta e complexa gama de fatores que podem influenciar significativamente a progressão DRC, a avaliação precisa do risco individual de desenvolvimento e avanço da doença se apresenta como um desafio considerável. A multiplicidade de variáveis envolvidas, incluindo informações pessoais, socioeconômicas e clínicas, dificulta a identificação de preditores consistentes e confiáveis, conforme ponderado por Sanmarchi *et al.* (2023) (215). A base de dados do Centro Hiperdia apresenta elevada complexidade, um número significativo de pacientes e de dados faltantes, diversidade étnica e natureza classificatória multiclasse, características raramente encontradas e abordadas na literatura correlata. Ainda assim, embora não tenha sido viável a obtenção definitiva de um conjunto mínimo de variáveis preditoras, foi possível alcançar, mesmo sem o uso da creatinina, resultados promissores e com potencial inovador.

Em conclusão, com os cenários desenvolvidos e com os resultados obtidos, pode ser discutido e avaliado o uso das técnicas na prática hospitalar, de forma que predições possam ser realizadas com novos dados e na rotina clínica convencional. Essa abordagem poderia auxiliar pacientes em seus tratamentos, contribuindo para detecção precoce e a consequente redução de custos associada.

8 TRABALHOS FUTUROS

Como trabalhos futuros decorrentes do desenvolvimento desta tese, podem ser citados:

- Validação dos modelos desenvolvidos em ambientes clínicos reais, de forma a aumentar a aplicabilidade prática das soluções propostas;
- Trabalho em conjunto com especialistas em nefrologia, de forma que o desenvolvimento de novas abordagens seja potencializado;
- Desenvolvimento de novos cenários de classificação, que possam considerar outros algoritmos e métodos de inferência de dados;
- Análise detalhada acerca de possíveis casos de sobreajuste nos resultados obtidos;
- Utilização de métodos como *shapley additive explanations* (SHAP) e *local interpretable model-agnostic explanations* (LIME), com o propósito de avaliar a relevância e a contribuição de cada variável preditora nos resultados produzidos pelos modelos;
- Implementação dos cenários desenvolvidos em diferentes bases de dados;
- Utilização de modelos de aprendizado profundo para predição dos estágios da DRC.

REFERÊNCIAS

1 ACHARYA, M. S.; ARMAAN, A.; ANTONY, A. S. A Comparison of Regression Models for Prediction of Graduate Admissions. *International Conference on Computational Intelligence in Data Science (ICCIDS)*, pp. 1-5, 2019.

2 ADADI, A.; BERRADA, M. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, vol. 6, pp. 52138-52160, 2018.

3 AGUIAR, L. K.; PRADO, R. R.; GAZZINELLI, A.; MALTA, D. C. Fatores associados à doença renal crônica: inquérito epidemiológico da Pesquisa Nacional de Saúde. *Revista Brasileira de Epidemiologia*, vol. 23, e200044, 2020.

4 AHMETOGLU, H.; RESUL, D. A comprehensive review on detection of cyber-attacks: Data sets, methods, challenges, and future research directions. *Internet of Things*, vol. 20, 2022, ISSN 2542-6605.

5 AKKAYA, B.; ÇOLAKOĞLU, N. Comparison of Multi-class Classification Algorithms on Early Diagnosis of Heart Diseases, 2019.

6 ALDAUSARI, N.; SOWMYA, A.; NADINE, M.; GELAREH, M. Video Generative Adversarial Networks: A Review. *ACM Computing Surveys*, vol. 55, no. 2, Article 30, 25 pages, 2022.

7 ALEXOPOULOS, E. C. Introduction to multivariate regression analysis. *Hippokratia*, 14(Suppl 1), pp. 23–28, 2010.

8 ALI, J.; KHAN, R.; AHMAD, N.; MAQSOOD, I. Random Forests and Decision Trees. *International Journal of Computer Science Issues (IJCSI)*, vol. 9, issue 5, no. 3, 2012.

9 ALMANSOUR, N. A.; SYED, H. F.; KHAYAT, N. R.; ALTHEEB, R. K.; JURI, R. E.; ALHIYAFI, J.; ALRASHED, S.; OLATUNJI, S. O. Neural network and support vector machine for the prediction of chronic kidney disease: A comparative study. *Computers in Biology and Medicine*, vol. 109, pp. 101-111, 2019. ISSN 0010-4825.

10 ALMASOUD, M.; WARD, T. E. Detection of chronic kidney disease using machine learning algorithms with least number of predictors. *International Journal of Soft Computing and Its Applications*, vol. 10, no. 8, 2019. ISSN 2074-8523.

11 ALNUAIMI, A.; ALBALDAWI, T. An overview of machine learning classification techniques. *BIO Web of Conferences*, vol. 97, 2024.

12 ANIL, R. *et al.* PaLM 2 Technical Report, 2023.

13 ANJOS, U. U.; FERREIRA, F.; HENN, F. H.; KOLEV, N.; MENDES, B. V. M. Modelando Dependências via Cópulas. Universidade Federal de São Paulo, IME. ABE, 143 pp., 2004.

14 AZUR, M. J.; STUART, E. A.; FRANGAKIS, C.; LEAF, P. J. Multiple imputation by chained equations: what is it and how does it work? *International Journal of Methods in Psychiatric Research*, vol. 20, no. 1, pp. 40-49, 2011. PMID: 21499542; PMCID: PMC3074241.

15 BASTOS, M. G.; BREGMAN, R.; KIRSZTAJN, G. M. Doença renal crônica: frequente e grave, mas também prevenível e tratável. *Revista da Associação Médica Brasileira*, São Paulo, vol. 56, no. 2, pp. 248-253, 2010.

16 BASTOS, M. G.; OLIVEIRA, D. C.; KIRSTAJN, G. M. Doença Renal Crônica no Paciente Idoso. *Clinical & Biomedical Research*, vol. 31, no. 1, 2011. ISSN 2357-9730.

17 BASTOS, M. G.; KIRSZTAJN, G. M. Doença renal crônica: importância do diagnóstico precoce, encaminhamento imediato e abordagem interdisciplinar estruturada para melhora do desfecho em pacientes ainda não submetidos à diálise. *Jornal Brasileiro de Nefrologia*, vol. 33, no. 1, pp. 74-87, 2011.

18 BATISTA, G.; MONARD, M. C. A Study of K-Nearest Neighbour as an Imputation Method. In: *Hybrid Intelligent Systems, Frontiers in Artificial Intelligence and Applications*, vol. 30, pp. 251-260, 2002.

19 BATISTA, G.; BAZZAN, A.; MONARD, M. C. Balancing Training Data for Automated Annotation of Keywords: a Case Study. In: *Proceedings of Workshop on Bioinformatics*, pp. 10-18, 2003.

20 BECKER, R. M.; HEIDEMANN, I. T. S. B. Promoção da saúde no cuidado às pessoas com doença não transmissível: revisão integrativa. *Texto & Contexto - Enfermagem*, Florianópolis, vol. 29, e20180250, 2020.

21 BEJA-BATTAIS, P. Overview of AdaBoost: Reconciling its views to better understand its dynamics, 2023.

22 BELLO, A. K. *et al.* ISN–Global Kidney Health Atlas: A report by the International Society of Nephrology: An assessment of global kidney health care status focussing on capacity, availability, accessibility, affordability, and outcomes of kidney disease. *International Society of Nephrology*, Brussels, 2023.

23 BENTÉJAC, C.; CSÖRGO, A.; MARTÍNEZ-MUÑOZ, G. A comparative analysis of gradient boosting algorithms. *Artificial Intelligence Review*, vol. 54, pp. 1937-1967, 2019.

24 BERETTA, L.; SANTANIELLO, A. Nearest neighbor imputation algorithms: a critical evaluation. *BMC Medical Informatics and Decision Making*, vol. 16 (Suppl 3), 74, 2016.

25 BILLETT, H. H. Hemoglobin and Hematocrit. In: WALKER, H. K., editor. *Clinical Methods: The History, Physical, and Laboratory Examinations*. 3rd ed., Butterworths, 1990.

26 BINIA, A.; JAEGER, J.; HU, Y.; SINGH, A.; ZIMMERMANN, D. Daily potassium intake and sodium-to-potassium ratio in the reduction of blood pressure: a meta-analysis of randomized controlled trials. *Journal of Hypertension*, vol. 33, no. 8, 2015.

27 BODEN, W. E. High-density lipoprotein cholesterol as an independent risk factor in cardiovascular disease: assessing the data from Framingham to the Veterans Affairs High-Density Lipoprotein Intervention Trial. *The American Journal of Cardiology*, vol. 86, no. 12A, pp. 19L–22L, 2000.

28 BOSER, B. E.; GUYON, I. M.; VAPNIK, V. N. A Training Algorithm for Optimal Margin Classifiers. In: *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, Pittsburgh, pp. 144-152, 1992.

29 BIKBOV, B. *et al.* Global, regional, and national burden of chronic kidney disease, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017. *The Lancet*, vol. 395, pp. 709-733, 2020.

30 BLAZEK, K.; ZWIETEN, A.; SAGLIMBENE, V.; TEIXEIRA-PINTO, A. A practical guide to multiple imputation of missing data in nephrology. *Kidney International*, vol. 99, 2020.

31 BRADLEY, A. P. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, vol. 30, no. 7, pp. 1145-1159, 1997.

32 BRANDT, J.; LANZÉN, E. A Comparative Review of SMOTE and ADASYN in Imbalanced Data Classification (Dissertation), 42 pages, 2021.

33 BREIMAN, L. Random Forests. *Machine Learning*, vol. 45, pp. 5–32, Kluwer Academic Publishers, 2001.

34 BRITO, T. N. S.; OLIVEIRA, A. R. A.; da SILVA, A. K. C. Glomerular filtration rate estimated in adults: characteristics and limitations of equations used. *Revista Brasileira de Análises Clínicas*, Rio de Janeiro, fev. 2016.

35 BURGES, C. J. A Tutorial on Support Vector Machines for Pattern Recognition. *Data Mining and Knowledge Discovery*, vol. 2, pp. 121–167, 1998.

36 BUSUTTIL, S. Support vector machines. In: *1st Computer Science Annual Workshop*, Msida, pp. 34-39. University of Malta, Faculty of ICT, 2003.

37 CAMINO, R.; LIU, X.; RAMÍREZ, J. Copulas: Multivariate Modeling with Python. *Journal of Open Source Software*, vol. 4, no. 42, 1639, 2019.

38 CAMPOS, A. C. IBGE: pelo menos uma doença crônica afetou 52% dos adultos em 2019. *Agência Brasil*, Rio de Janeiro, 2020. Disponível em `https://agenciabrasil.ebc.com.br/saude/noticia/2020-11/ibge-pelo-menos-uma-doenca-cronica-afetou-52-dos-adultos-em-2019`. Acesso em 14 Junho de 2021.

39 CARNEY, E. F. The impact of chronic kidney disease on global health. *Nature Reviews Nephrology*, vol. 16, pp. 251, 2020.

40 CASTRO, M. C. R. *Manual de Transplante Renal.* Unidade de Transplante Renal do Hospital das Clínicas da Faculdade de Medicina da Universidade de São Paulo. Associação Brasileira de Transplante de Órgãos, 2006.

41 CENTERS FOR DISEASE CONTROL AND PREVENTION. Chronic Kidney Disease in the United States, 2021. Atlanta, GA: US Department of Health and Human Services, Centers for Disease Control and Prevention; 2021.

42 CHAN, L.; VAID, A.; NADKARNI, G. N. Applications of machine learning methods in kidney disease: hope or hype? *Current Opinion in Nephrology and Hypertension*, vol. 29, no. 3, pp. 319–326, 2020.

43 CERVANTES, J.; GARCIA-LAMONT, F.; RODRÍGUEZ-MAZAHUA, L.; LOPEZ, A. A comprehensive survey on support vector machine classification: Applications, challenges and trends. *Neurocomputing*, vol. 408, pp. 189-215, 2020.

44 CHAGAS, E. T. O. Deep Learning e suas aplicações na atualidade. *Revista Científica Multidisciplinar Núcleo do Conhecimento*, ano 04, ed. 05, vol. 04, pp. 05-26, 2019. ISSN: 2448-0959.

45 CHAN, C. T.; BLANKESTIJN, P. J.; DEMBER, L. M.; et al. Dialysis initiation, modality choice, access, and prescription: conclusions from a Kidney Disease: Improving Global Outcomes (KDIGO) Controversies Conference. *Kidney International*, vol. 96, no. 1, pp. 37–47, 2019.

46 CHAUBEY, A.; SHRESTHA, A.; GOGOI, A. Using Linear Regression Machine Learning Algorithm for the Prediction of Real Estate, 2022.

47 CHAWLA, N.; BOWYER, K.; HALL, L.; KEGELMEYER, K. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research (JAIR)*, vol. 16, pp. 321-357, 2002.

48 CHEN, T.; GUESTRIN, C. XGBoost: A scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785-794, 2016.

49 CHEN, C.; LI, K.; DUAN, M.; LI, K. Chapter 6 - Extreme Learning Machine and Its Applications in Big Data Processing. In: *Intelligent Data-Centric Systems, Big Data Analytics for Sensor-Network Collected Intelligence*. Academic Press, pp. 117-150, 2017. ISBN 9780128093931.

50 CHENGSHENG, T.; HUACHENG, L.; BING, X. AdaBoost typical Algorithm and its application research. *MATEC Web of Conferences*, vol. 139, 00222, 2017.

51 CHERUBINI, U.; LUCIANO, E.; VECCHIATO, W. *Copula Methods in Finance*. John Wiley & Sons Ltd, 2004. ISBN: 9781118673331.

52 CHOWDHERY, A. *et al.* PaLM: scaling language modeling with pathways. *Journal of Machine Learning Research*, vol. 24, no. 1, Article 240, 113 pages, 2023.

53 CHRISTOFOLETTI, M. et al. Simultaneidade de doenças crônicas não transmissíveis em 2013 nas capitais brasileiras: prevalência e perfil sociodemográfico. *Epidemiologia e Serviços de Saúde*, Brasília, vol. 29, no. 1, e2018487, 2020.

54 COCKCROFT, D. W.; GAULT, M. H. Prediction of creatinine clearance from serum creatinine. *Nephron*, vol. 16, no. 1, pp. 31–41, 1976.

55 CLARIVATE ANALYTICS. Web of Science. Disponível em `https://www.webofscience.com/`. Acesso em 2 de Abril de 2024.

56 COCKWELL, P.; FISH, L. A. The global burden of chronic kidney disease. *The Lancet*, vol. 395, no. 10225, 2020.

57 CORTES, C.; VAPNIK, V. Support-vector networks. *Machine Learning*, vol. 20, no. 3, pp. 273-297, 1995.

58 COVER, T.; HART, P. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, vol. 13, no. 1, pp. 21-27, 1967.

59 COX, D. R. The regression analysis of binary sequences. *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 20, no. 2, pp. 215-232, 1958.

60 COX, R. A.; GARCÍA-PALMIERI, M. R. Cholesterol, Triglycerides, and Associated Lipoproteins. In: WALKER, H. K.; HALL, W. D.; HURST, J. W., editors. *Clinical Methods: The History, Physical, and Laboratory Examinations*. 3rd ed., Butterworths, 1990.

61 CRISTIANINI, N.; SHAWE-TAYLOR, J. *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge University Press, 2000. ISBN: 9780521780193.

62 CRISTIANINI, N.; RICCI, E. Support Vector Machines. In: *Encyclopedia of Algorithms*, pp. 928–932, 2008.

63 CUNNINGHAM, P.; DELANY, S. k-Nearest Neighbour Classifiers - A Tutorial. *ACM Computing Surveys*, vol. 54, no. 6, pp. 1-25, 2021.

64 DAUGIRDAS, J. T.; BLAKE, P. *Manual de Diálise*. Rio de Janeiro, RJ: Guanabara Koogan, 2007.

65 DEBAL, D. A.; SITOTE, T. M. Chronic kidney disease prediction using machine learning techniques. *Journal of Big Data*, vol. 9, no. 109, pp. 1-19, 2022.

66 DE PAULI, S. T. Z.; KLEINA, M.; BONAT, W. Multilayer Perceptron Artificial Neural Networks: An Approach for Learning Through the Bayesian Framework. *Revista Brasileira de Biometria*, vol. 39, no. 1, pp. 45-59, 2021.

67 DE SOUSA, L. C. M.; SILVA, N. R.; AZEREDO, C. M.; RINALDI, A. E. M.; DA SILVA, L. S. Health-related patterns and chronic kidney disease in the Brazilian population: National Health Survey, 2019. *Frontiers in Public Health*, vol. 11, 2023.

68 DELRUE, C.; DE BRUYNE, S.; SPEECKAERT, M. M. Application of Machine Learning in Chronic Kidney Disease: Current Status and Future Prospects. *Biomedicines*, vol. 12, no. 568, 2024.

69 DEVLIN, J.; CHANG, M. W.; LEE, K.; TOUTANOVA, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, 2018.

70 DI LEVA, A. AI-augmented multidisciplinary teams: hype or hope? *The Lancet*, vol. 394, no. 10211, pp. 1801, 2019.

71 DING, Y.; ROSS, A. A comparison of imputation methods for handling missing scores in biometric fusion. *Pattern Recognition*, vol. 45, issue 3, pp. 919-933, 2012.

72 DOUPE, P.; FAGHMOUS, J.; BASU, S. Machine Learning for Health Services Researchers. *Value in Health*, vol. 22, no. 7, pp. 808-815, 2019.

73 DRAIBE, S. A. *Especialização em Nefrologia Multidisciplinar*. Módulo 3 - Análise Epidemiológica da Doença Renal. Unidade 1 - Panorama da Doença Renal Crônica no Brasil e no mundo. Universidade Aberta do SUS, Universidade Federal do Maranhão, São Luís, 2014.

74 DRISKELL, O. J.; HOLLAND, D.; WALDRON, J. L.; FORD, C.; SCARGILL, J. J.; HEALD, A.; TRAIN, M.; HANNA, F. W.; JONES, P. W.; PEMBERTON, R. J.; FRYER, A. A. Reduced testing frequency for glycated hemoglobin, HbA1c, is associated with deteriorating diabetes control. *Diabetes Care*, vol. 37, no. 10, 2014.

75 DUARTE, A.; SANTOS, S.; ARAÚJO, M.; MACÁRIO, E.; FARIS, M.; LIMA, M.; OLIVEIRA, D. Perfil epidemiológico da insuficiência renal no Brasil de 2012 a 2022. *Research, Society and Development*, 2023.

76 ELREEDY, D.; ATIYA, A. F. A Comprehensive Analysis of Synthetic Minority Oversampling Technique (SMOTE) for handling class imbalance. *Information Sciences*, vol. 505, pp. 32-64, 2019. ISSN 0020-0255.

77 ELREEDY, D.; ATIYA, A. F.; KAMALOV, F. A theoretical distribution analysis of synthetic minority oversampling technique (SMOTE) for imbalanced learning. *Machine Learning*, vol. 113, pp. 4903–4923, 2024.

78 EMBASE. *Embase database*. Elsevier. Disponível em `https://www.embase.com`. Acesso em 15 de Julho de 2024.

79 EMMANUEL, T.; MAUPONG, T.; MPOELENG, D.; SEMONG, T.; BANYATSANG, M.; TABONA, O. A survey on missing data in machine learning. *Journal of Big Data*, vol. 8, no. 140, 2021.

80 ENGELBRECHT, A. P. *Computational Intelligence: An Introduction*. John Wiley & Sons Ltd., 2007. ISBN 9780470512500.

81 EVGENIOU, T.; PONTIL, M. Support Vector Machines: Theory and Applications. In: *Advances in Large Margin Classifiers*, pp. 249-257, 2001.

82 FATIMA, M.; PASHA, M. Survey of Machine Learning Algorithms for Disease Diagnostic. *Journal of Intelligent Learning Systems and Applications*, vol. 9, pp. 1-16, 2017.

83 FAWCETT, T. An introduction to ROC analysis. *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861-874, 2006.

84 FIGUEIREDO, A. E. B.; CECCON, R. F.; FIGUEIREDO, J. H. C. Doenças crônicas não transmissíveis e suas implicações na vida de idosos dependentes. *Ciência & Saúde Coletiva*, Rio de Janeiro, vol. 26, no. 1, pp. 77-88, 2021.

85 FRANCIS, A.; HARHAY, M. N.; ONG, A. C. M.; TUMMALAPALLI, S. L.; ORTIZ, A.; FOGO, A. B.; et al. Chronic kidney disease and the global public health agenda: an international consensus. *Nature Reviews Nephrology*, vol. 20, pp. 473–485, 2024.

86 FREUND, Y.; SCHAPIRE, R. E. A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *Journal of Computer and System Sciences*, vol. 55, no. 1, pp. 119-139, 1997.

87 FREUND, Y.; SCHAPIRE, R. E. A Short Introduction to Boosting, 1999.

88 FRIEDMAN, J. Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics*, vol. 29, 1999.

89 FUKUSHIMA, R. L. M.; et al. Fatores associados à qualidade de vida de pacientes renais crônicos em hemodiálise. *Acta Paulista de Enfermagem*, São Paulo, vol. 29, no. 5, pp. 518-524, 2016.

90 GANIE, S. M.; DUTTA PRAMANIK, P. K.; MALLIK, S.; ZHAO, Z. Chronic kidney disease prediction using boosting techniques based on clinical parameters. *PloS One*, vol. 18, no. 12, 2023.

91 GAO, Q.; JIN, X.; XIA, E. H.; WU, X.; GU, L.; YAH, H.; XIA, Y.; LI, S. Identification of Orphan Genes in Unbalanced Datasets Based on Ensemble Learning. *Frontiers in Genetics*, vol. 11, no. 820, 2020.

92 GARDNER, M. W.; DORLING, S. R. Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. *Atmospheric Environment*, vol. 32, issues 14–15, pp. 2627-2636, 1998. ISSN 1352-2310.

93 GEMINI TEAM; ANIL, R. *et al.*. Gemini: A Family of Highly Capable Multimodal Models, 2024.

94 GENEST, C.; FAVRE, A. C. Everything You Always Wanted to Know about Copula Modeling but Were Afraid to Ask. *Journal of Hydrologic Engineering*, vol. 12, no. 4, pp. 347-368, 2007.

95 GEROLDINGER, A.; HRONSKY, M.; ENDEL, F.; GOTTFRIED, E.; OBERBAUER, R.; HEINZE, G. Estimation of the Prevalence of Chronic Kidney Disease in People with Diabetes by Combining Information from Multiple Routine Data Collections. *Journal of the Royal Statistical Society Series A: Statistics in Society*, vol. 184, issue 4, pp. 1260–1282, 2021.

96 GHOSH, S. K.; KHANDOKER, A. H. Investigation on explainable machine learning models to predict chronic kidney diseases. *Scientific Reports*, vol. 14, no. 3687, pp. 1-14, 2024.

97 GLAS, C. A. W. Missing Data. In: Penelope Peterson, Eva Baker, Barry McGaw, editors. *International Encyclopedia of Education* (Third Edition), pp. 283-288, 2010. ISBN: 9780080448947.

98 GOLIATT, L.; CAPRILES, P. V. S. Z.; IWASHIMA, G. C.; SCORALICK, J. P. Unsupervised Analysis of Clinical and Laboratory Parameters of Chronic Kidney Disease. In: *Intelligent Systems Design and Applications*. Lecture Notes in Networks and Systems, Springer, vol. 1052, 2024.

99 GOODFELLOW, I. J.; POUGET-ABADIE, J.; MIRZA, M.; XU, B.; WARDE-FARLEY, D.; OZAIR, S.; COURVILLE, A.; BENGIO, Y. Generative Adversarial Networks. *Advances in Neural Information Processing Systems*, vol. 3, 2014.

100 GUANG-BIN, H.; QIN-YU, Z.; CHEE-KHEONG, S. Extreme Learning Machine: A New Learning Scheme of Feedforward Neural Networks. In: *IEEE International Joint Conference on Neural Networks*, vol. 2, pp. 985-990, 2004.

101 GUANG-BIN, H.; QIN-YU, Z.; CHEE-KHEONG, S. Extreme Learning Machine: Theory and Applications. *Neurocomputing*, vol. 70, issues 1–3, pp. 489-501, 2006.

102 GUO, H.; NGUYEN, H.; VU, D. A.; BUI, X. N. Forecasting mining capital cost for open-pit mining projects based on artificial neural network approach. *Resources Policy*, 2019.

103 GUO, Y.; YU, H.; CHEN, D.; ZHOU, Y.-T. Machine learning distilled metabolite biomarkers for early stage renal injury. *Metabolomics*, vol. 16, no. 4, 2020.

104 HAIDER, M. Z.; ASLAM, A. Proteinuria. StatPearls. Treasure Island (FL): StatPearls Publishing. Disponível em `https://www.ncbi.nlm.nih.gov/books/NBK564390/`. Acesso em 15 de Junho de 2024.

105 HALDER, R. J.; UDDIN, M. N.; UDDIN, M. A.; ARYAL, S.; SAHA, S.; HOSSEN, R.; AHMED, S.; RONY, M. A. T.; AKTER, M. F. ML-CKDP: Machine learning-based chronic kidney disease prediction with smart web application. *Journal of Pathology Informatics*, vol. 15, 2024. ISSN 2153-3539.

106 HALAM, A.; MUKHERJEE, D.; CHASSAGNE, R. Multivariate imputation via chained equations for elastic well log imputation and prediction. *Applied Computing and Geosciences*, vol. 14, 2022.

107 HARRIS, C. R.; MILLMAN, K. J.; VAN DER WALT, S. J.; *et al.* Array programming with NumPy. *Nature*, vol. 585, pp. 357–362, 2020.

108 HARRISON, M. *Machine Learning – Guia de Referência Rápida: Trabalhando com dados estruturados em Python*. Novatec Editora, 2019. ISBN: 9788575228180.

109 HARTWIG, F. P.; DAVEY, G. S.; SCHMIDT, A. F.; STERNE, J. A. C.; HIGGINS, J. P. T.; BOWDEN, J. The median and the mode as robust meta-analysis estimators in the presence of small-study effects and outliers. *Research Synthesis Methods*, vol. 11, no. 3, pp. 397–412, 2020.

110 HAYKIN, S. *Redes Neurais: Princípios e Prática*. Bookman Editora, 2001. ISBN 9788577800865.

111 HEARST, M.; DUMAIS, S. T.; OSMAN, E.; PLATT, J.; SCHOLKOPF, B. Support vector machines. *IEEE Intelligent Systems and Their Applications*, vol. 13, pp. 18-28, 1998.

112 HE, H.; BAI, Y.; GARCIA, E. A.; LI, S. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In: *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, Hong Kong, pp. 1322-1328, 2008.

113 HICKS, S. A.; STRUMKE, I.; THAMBAWITA, V.; HAMMOU, M.; RIEGLER, M. A.; HALVORSEN, P.; PARASA, S. On evaluation metrics for medical applications of artificial intelligence. *Scientific Reports*, vol. 12, no. 1, 5979, 2022.

114 HINTON, G. E.; SALAKHUTDINOV, R. R. Reducing the Dimensionality of Data with Neural Networks. *Science*, vol. 313, pp. 504-507, 2006.

115 HOLLAND, J. H. Outline for a Logical Theory of Adaptive Systems. *Journal of the ACM*, vol. 9, no. 3, pp. 297–314, 1962.

116 HOLLAND, J. H. *Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence*, 1992.

117 HOSTEN, A. O. BUN and Creatinine in Clinical Methods. In: WALKER, H. K.; HALL, W. D.; HURST, J. W., editors. *The History, Physical, and Laboratory Examinations*, 3rd ed., Butterworths, Boston, 1990.

118 HUAIRA, R. M. N. N. Validação de um registro e caracterização de uma coorte de usuários com doença renal crônica pré-dialítica em um centro multiprofissional de atendimentos em doenças crônicas não transmissíveis. 2017. Dissertação (Mestrado em Saúde). Pós-graduação em Saúde da Faculdade de Medicina, Universidade Federal de Juiz de Fora. Orientadores: Natália Maria da Silva Fernandes e Marcus Gomes Bastos.

119 HUAIRA, R. M. N. N.; DE PAULA, R.; BASTOS, M.; COLUGNATI, F.; FERNANDES, N. Validated registry of pre-dialysis chronic kidney disease: description of a large cohort. *Brazilian Journal of Nephrology*, vol. 40, 2018.

120 HUANG, G.-B.; ZHU, Q.-Y.; SIEW, C.-K. Extreme learning machine: Theory and applications. *Neurocomputing*, vol. 70, no. 1-3, pp. 489-501, 2006.

121 HUANG, G.; HUANG, G. B.; SONG, S.; YOU, K. Trends in extreme learning machines: A review. *Neural Networks*, vol. 61, pp. 32-48, 2015.

122 HUNTER, J.D. Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, vol. 9, no. 3, pp. 90-95, 2007.

123 IFTIKHAR, H.; KHAN, M.; KHAN, Z.; KHAN, F.; ALSHANBARI, H. M.; AHMAD, Z. A Comparative Analysis of Machine Learning Models: A Case Study in Predicting Chronic Kidney Disease. *Sustainability*, vol. 15, no. 2754, 2023.

124 IQBAL, M. A. Application of Regression Techniques with their Advantages and Disadvantages. vol. 4, pp. 11-17, 2021.

125 ISLAM, M.; BARUA, S.; AHMED, M.; BEGUM, S.; DI FLUMERI, G. Deep Learning for Automatic EEG Feature Extraction: An Application in Drivers' Mental Workload Classification, 2019.

126 ISLAM, M. A.; MAJUMDER, M. Z. H.; HUSSEIN, M. A. Chronic kidney disease prediction based on machine learning algorithms. *Journal of Pathology Informatics*, vol. 14, 2023. ISSN 2153-3539.

127 IWASHIMA, G. C. Análise computacional de parâmetros clínicos e laboratoriais da doença renal crônica. 2024. 51f. Trabalho de Conclusão de Curso (Graduação em Ciência da Computação) – Universidade Federal de Juiz de Fora, 2023.

128 JADHAV, A.; PRAMOD, D.; RAMANATHAN, K. Comparison of Performance of Data Imputation Methods for Numeric Dataset. *Applied Artificial Intelligence*, vol. 33, no. 10, pp. 913–933, 2019.

129 JAEGER, B.; GEIGER, A. An Invitation to Deep Reinforcement Learning, 2023.

130 KAISER, J. Dealing with Missing Values in Data. *Journal of Systems Integration*, vol. 5, pp. 42-51, 2014.

131 KOTSIANTIS, S.; KANELLOPOULOS, D.; PINTELAS, P. Handling imbalanced datasets: A review. *GESTS International Transactions on Computer Science and Engineering*, vol. 30, pp. 25-36, 2005.

132 JAMES, G.; WITTEN, D.; HASTIE, T.; TIBSHIRANI, R.; TAYLOR, J. *An Introduction to Statistical Learning: with Applications in Python*. New York: Springer, 2023.

133 JOEL, L. O.; DOORSAMY, W.; PAUL, B. S. On the Performance of Imputation Techniques for Missing Values on Healthcare Datasets, 2024.

134 KALCHEVA, N.; TODOROVA, M.; MARINOVA, G. Naive Classifier, Decision Tree and AdaBoost Ensemble Algorithm - Advantages and Disadvantages, pp. 153-157, 2020.

135 KHALID, F.; ALSADOUN, L.; KHILJI, F.; MUSHTAQ, M.; EZE-ODURUKWE, A.; MUSHTAG, M. M.; ALI, H.; FARMAN, R. O.; ALI, S. M.; FATIA, R.; BOKHARI, S. F. H. Predicting the Progression of Chronic Kidney Disease: A Systematic Review of Artificial Intelligence and Machine Learning Approaches. *Cureus*, vol. 16, no. 5, 2024.

136 KHAN, B.; NASEEM, R.; MUHAMMAD, F.; ABBAS, G.; KIM, S. An Empirical Evaluation of Machine Learning Techniques for Chronic Kidney Disease Prophecy. *IEEE Access*, vol. 8, pp. 55012-55022, 2020.

137 Kidney disease is a common and serious condition, driven by the global epidemics of hypertension, atherosclerotic diseases and diabetes. *Nature Reviews Nephrology*, vol. 20, pp. 829, 421-423, 2024. Disponível em `https://www.nature.com/articles/s41581-024-00829-x`. Acesso em 11 de Maio de 2024.

138 KIRSZTAJN, G. M. *et al.* Leitura rápida do KDIGO 2012: Diretrizes para avaliação e manuseio da doença renal crônica na prática clínica. *Jornal Brasileiro de Nefrologia*, São Paulo, vol. 36, no. 1, pp. 63-73, 2014.

139 LASTER, M.; SHEN, J. I.; NORRIS, K. C. Kidney Disease Among African Americans: A Population Perspective. *American Journal of Kidney Diseases*, vol. 72, issue 5, Supplement 1, pp. S3-S7, 2018. ISSN 0272-6386.

140 LECUN, Y.; BENGIO, Y.; HINTON, G. Deep Learning. *Nature*, vol. 521, pp. 436-44, 2015.

141 LEMAÎTRE, G.; NOGUEIRA, F.; ARIDAS, C. K. Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning. *Journal of Machine Learning Research*, vol. 18, pp. 1-5, 2017.

142 LEVEY, A. S. *et al.* A more accurate method to estimate glomerular filtration rate from serum creatinine: a new prediction equation. *Annals of Internal Medicine*, vol. 130, no. 6, pp. 461–470, 1999.

143 LEVEY, A. S. *et al.* A simplified equation to predict glomerular filtration rate from serum creatinine [Abstract]. *Journal of the American Society of Nephrology*, vol. 11, A0828, pp. 115A, 2000.

144 LEVEY, A. S.; Chronic Kidney Disease Epidemiology Collaboration. Using standardized serum creatinine values in the modification of diet in renal disease study equation for estimating glomerular filtration rate. *Annals of Internal Medicine*, vol. 145, no. 4, pp. 247–254, 2006.

145 LEVEY, A. S.; CKD-EPI (Chronic Kidney Disease Epidemiology Collaboration). A new equation to estimate glomerular filtration rate. *Annals of Internal Medicine*, vol. 150, no. 9, pp. 604–612, 2009.

146 LITTLE, R. J. A.; RUBIN, D. B. *Statistical Analysis with Missing Data*. Wiley Series in Probability and Statistics. Wiley, 2014. ISBN: 9781118625880.

147 LIU, H.; YAN, G.; ZHU, D.; CHEN, C. Intelligent modeling strategies for forecasting air quality time series: A review. *Applied Soft Computing*, vol. 102, 2021.

148 LOPES, A. A.; SILVEIRA, M. A.; MARTINELLI, R. P.; ROCHA, H. Associação entre raça e incidência de doença renal terminal secundária a glomerulonefrite: influência do tipo histológico e da presença de hipertensão arterial. *Revista Da Associação Médica Brasileira*, vol. 47, n. 1, pp. 78–84, 2001.

149 LOWE-JONES, R.; ETHIER, I.; FISH, L. A.; WONG, M. M.; THOMPSON, S.; NAKHOUL, G.; SANDAL, S.; CHANCHLANI, R.; DAVISON, S. N.; GHIMIRE, A.; JINDAL, K.; OSMAN, M. A.; RIAZ, P.; SAAD, S.; SOZIO, S. M.; TUNGSANG, S.; CAMBIER, A.; ARRUEBO, S.; BELLO, A. K.; CASKEY, F. J. Regional Board and ISN-GKHA Team Authors. Capacity for the management of kidney failure in the International Society of Nephrology North America and the Caribbean region: report from the 2023 ISN Global Kidney Health Atlas (ISN-GKHA). *Kidney International Supplements*, vol. 13, no. 1, pp. 83–96, 2024.

150 LUENGO, J.; GARCÍA, S.; HERRERA, F. On the choice of the best imputation methods for missing values considering three groups of classification methods. *Knowledge and Information Systems*, vol. 32, no. 1, pp. 77–108, 2012.

151 LV, J. C.; ZHANG, L. X. Prevalence and Disease Burden of Chronic Kidney Disease. *Advances in Experimental Medicine and Biology*, vol. 1165, pp. 3–15, 2019.

152 MAALOUF, M. Logistic regression in data analysis: An overview. *International Journal of Data Analysis Techniques and Strategies*, vol. 3, pp. 281-299, 2011.

153 MADERO, M.; SARNAK, M. J. Creatinine-based formulae for estimating glomerular filtration rate: is it time to change to chronic kidney disease epidemiology collaboration equation?. *Current Opinion in Nephrology and Hypertension*, vol. 20, no. 6, pp. 622–630, 2011.

154 MALTA, D. C. *et al.* Noncommunicable diseases and the use of health services: analysis of the National Health Survey in Brazil. *Revista de Saúde Pública*, São Paulo, vol. 51, supl. 1, pp. 4s, 2017.

155 MALTA, D. C.; ANDRADE, S. S. C. A.; OLIVEIRA, T. P.; MOURA, L.; PRADO, R. R.; SOUZA, M. F. M. Probabilidade de morte prematura por doenças crônicas não transmissíveis, Brasil e regiões, projeções para 2025. *Revista Brasileira de Epidemiologia*, vol. 22, 2019.

156 MALTA, D. C.; DUNCAN, B. B.; SCHMIDT, M. I.; TEIXEIRA, R.; RIBEIRO, A. L. P.; FELISBINO-MENDES, M. S.; MACHADO, Í. E.; VELASQUEZ-MELENDEZ, G.; BRANT, L. C. C.; SILVA, D. A. S.; PASSOS, V. M. D. A.; NASCIMENTO, B. R.; COUSIN, E.; GLENN, S.; NAGHAVI, M. Trends in mortality due to non-communicable diseases in the Brazilian adult population: national and subnational estimates and projections for 2030. *Population Health Metrics*, vol. 18, supl. 1, pp. 16, 2020.

157 MARNENI, D.; VEMUKA, S. Analysis of Covid-19 using machine learning techniques. In: *Statistical Modeling in Machine Learning*, Academic Press, pp. 37-53, 2023. ISBN 9780323917766.

158 MAULUD, D.; ABDULAZEEZ, A. M. A Review on Linear Regression Comprehensive in Machine Learning. *Journal of Applied Science and Technology Trends*, vol. 1, no. 2, pp. 14-147, 2020. ISSN 2708-0757.

159 MEDSCAPE. Overview of Hypertensive Heart Disease. Medscape, 2024. Disponível em `https://emedicine.medscape.com/article/2088449-overview`. Acesso em 15 de Junho de 2024.

160 MESSAOUD, S.; BRADAI, A.; BUKHARI, S. H. R.; QUANG, P. T. A.; AHMED, O. B.; ATRI, M. A survey on machine learning in Internet of Things: Algorithms, strategies, and applications. *Internet of Things*, vol. 12, 2020. ISSN 2542-6605.

161 MCCULLOCH, W. S.; PITTS, W. A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*, vol. 5, pp. 115-133, 1943.

162 MCKINNEY, W. Data Structures for Statistical Computing in Python. In: *Proceedings of the 9th Python in Science Conference*, pp. 51-56, 2010.

163 MITCHELL, T. M. *Machine Learning*. McGraw-Hill International Editions, McGraw-Hill, 1997. ISBN 9780071154673.

164 MOHAMMED, A.; KORA, R. A comprehensive review on ensemble deep learning: Opportunities and challenges. *Journal of King Saud University - Computer and Information Sciences*, vol. 35, no. 2, pp. 757-774, 2023. ISSN 1319-1578.

165 MORAES, C.; FERNANDES, N.; COLUGNATI, F. Multidisciplinary treatment for patients with chronic kidney disease in pre-dialysis minimizes costs: a four-year retrospective cohort analysis. *Brazilian Journal of Nephrology*, vol. 43, 2021.

166 MORAES JUNIOR, C. S. Custos com os prestadores de serviço na atenção da Doença Renal Crônica em meio aos caminhos e descaminhos do SUS. 2019. Tese (Doutorado em Saúde). Pós-graduação em Saúde da Faculdade de Medicina, Universidade Federal de Juiz de Fora. Orientador: Fernando Antonio Basile Colugnati.

167 MÜLLER, S. A.; ELIMIAN, K.; RAFAMATANANTSOA, J. F.; *et al.* The burden and treatment of non-communicable diseases among healthcare workers in sub-Saharan Africa: a multi-country cross-sectional study. *Frontiers in Public Health*, vol. 12, 2024. ISSN 2296-2565.

168 MURTAGH, F. Multilayer perceptrons for classification and regression. *Neurocomputing*, vol. 2, issues 5–6, pp. 183-197, 1991. ISSN 0925-2312.

169 MUSA, A. B. Logistic Regression Classification for Uncertain Data. *Research Journal of Mathematical and Statistical Sciences*, vol. 2, pp. 1-6, 2014. ISSN 2320–6047.

170 MUZAMMIL, K.; MEHRAN, M. T.; HAQ, Z. U.; ULLAH, Z.; NAQVI, S. R.; IHSAN, M.; ABBAS, H. Applications of artificial intelligence in COVID-19 pandemic: A comprehensive review. *Expert Systems with Applications*, vol. 185, 2021. ISSN 0957-4174.

171 NAGARAJAN, G.; BABU, D. Missing data imputation on biomedical data using deeply learned clustering and L2 regularized regression based on symmetric uncertainty. *Artificial Intelligence in Medicine*, vol. 123, 2022.

172 NAEEM, S.; ALI, A.; ANAM, S.; AHMED, M. An Unsupervised Machine Learning Algorithms: Comprehensive Review. *IJCDS Journal*, vol. 13, pp. 911-921, 2023.

173 NATEKIN, A.; KNOLL, A. Gradient boosting machines, a tutorial. *Frontiers in Neurorobotics*, vol. 7, no. 21, 2013.

174 NATIONAL KIDNEY FOUNDATION. KDIGO 2012 clinical practice guideline for the evaluation and management of chronic kidney disease. *Kidney International Supplements*, vol. 3, no. 1, pp. 1-150, 2013.

175 NATIONAL KIDNEY FOUNDATION. Estimated Glomerular Filtration Rate (eGFR), 2018. Disponível em `https://www.kidney.org/atoz/content/gfr`. Acesso em 2 Junho de 2021.

176 NELSEN, R. B. *An Introduction to Copulas*. Springer Series in Statistics, Springer New York, 2007. ISBN 9780387286785.

177 NERBASS, F. B.; LIMA, H. N.; MOURA-NETO, J. A.; LUGON, J. R.; SESSO, R. Brazilian Dialysis Survey 2022. *Brazilian Journal of Nephrology*, vol. 46, no. 2, 2023.

178 NEVES, P. D. M. M.; SESSO, R. C. C.; THOMÉ, F. S.; LUGON, J. R.; NASCIMENTO, M. M. Censo Brasileiro de Diálise: análise de dados da década 2009-2018. *Brazilian Journal of Nephrology*, vol. 42, no. 2, pp. 191-200, 2020.

179 NYRNES, A.; TOFT, I.; NJØLSTAD, I.; MATHIESEN, E. B.; WILSGAARD, T.; HANSEN, J. B.; LØCHEN, M. L. Uric acid is associated with future atrial fibrillation: an 11-year follow-up of 6308 men and women–the Tromso Study. *Europace*, vol. 16, no. 3, pp. 320–326, 2014.

180 OGUNLEYE, A.; WANG, Q. G. XGBoost Model for Chronic Kidney Disease Diagnosis. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 17, no. 6, pp. 2131–2140, 2020.

181 OPENAI; ACHIAM, J. *et al.* GPT-5 Technical Report. OpenAI, 2024.

182 OPITZ, D.; MACLIN, R. Popular Ensemble Methods: An Empirical Study. *Journal of Artificial Intelligence Research*, vol. 11, pp. 169-198, 1999.

183 OSTERTAGOVÁ, E. Modelling using Polynomial Regression. *Procedia Engineering*, vol. 48, pp. 500-506, 2012. ISSN 1877-7058.

184 PADMANABHA, Y. C. A.; PULAVAIGARI, V.; ESWARA, B. Semi-supervised learning: a brief review. *International Journal of Engineering & Technology*, vol. 7, pp. 81, 2018.

185 PAPADAKIS, M. A.; ARIEFF, A. I. Unpredictability of clinical evaluation of renal function in cirrhosis. *The American Journal of Medicine*, vol. 82, no. 5, pp. 945–952, 1987.

186 PECOITS, R. F. S.; RIBEIRO, S. C. (Org). Modalidades de terapia renal substitutiva: hemodiálise e diálise peritoneal. *Nefrologia, Unidade 3*, Universidade Federal do Maranhão UNASUS/UFMA - São Luís, 2014.

187 PEDERSEN, A. B.; MIKKELSEN, E. M.; CRONIN-FENTON, D.; KRISTENSEN, N. R.; PHAM, T. M.; PEDERSEN, L.; PEDERSEN, I. Missing data and multiple imputation in clinical epidemiological research. *Clinical Epidemiology*, vol. 9, pp. 157-166, 2017.

188 PEDREGOSA, F.; VAROQUAUX, G.; GRAMFORT, A.; MICHEL, V.; THIRION, B.; GRISEL, O.; BLONDEL, M.; PRETTENHOFER, P.; WEISS, R.; DUBOURG, V.; VANDERPLAS, J.; PASSOS, A.; COURNAPEAU, D.; BRUCHER, M.; PERROT, M.; DUCHESNAY, E. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, vol. 12, pp. 2825-2830, 2011.

189 PICCININI, G. The First Computational Theory of Mind and Brain: A Close Look at Mcculloch and Pitts's "Logical Calculus of Ideas Immanent in Nervous Activity". *Synthese*, vol. 141, pp. 175–215, 2004.

190 PIRI, M. Review of Regression Algorithms. *5th International Congress on Engineering, Technology and Innovation*, 2023.

191 POLAT, H.; MEHR, H. D.; CETIN, A. Diagnosis of Chronic Kidney Disease Based on Support Vector Machine by Feature Selection Methods. *Journal of Medical Systems*, vol. 41, no. 4, pp. 55, 2017.

192 POLIKAR, R. Ensemble Learning. *Ensemble Machine Learning*, pp. 1-34, 2012.

193 PONTES, L. B. *et al.* Prevalência de insuficiência renal em pacientes idosos com câncer em um centro de tratamento oncológico. *Einstein*, vol. 12, no. 3, pp. 300-3, 2014.

194 PORTO, J. R. *et al.* Evaluation of Renal Function in Chronic Kidney Disease. *Revista Brasileira de Análises Clínicas*, vol. 49, no. 1, pp. 26-35, 2017.

195 POWERS, D. M. W. Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. *Journal of Machine Learning Technologies*, vol. 2, pp. 2229-3981, 2011.

196 PRAJWALA, T. R. A Comparative Study on Decision Tree and Random Forest Using R Tool. *International Journal of Advanced Research in Computer and Communication Engineering*, 2015.

197 PRISMA. Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA), 2024. Disponível em `http://www.prisma-statement.org`. Acesso em 12 de Janeiro de 2024.

198 PYTHON SOFTWARE FOUNDATION. *Python: A Powerful Programming Language*, Version 3.10.0, 2024.

199 QIN, J. *et al.* A Machine Learning Methodology for Diagnosing Chronic Kidney Disease. *IEEE Access*, vol. 8, pp. 20991-21002, 2020.

200 RADFORD, A.; KARTHINK, N. Improving Language Understanding by Generative Pre-Training, 2018.

201 RADY, E. H. A.; ANWAR, A. S. Prediction of Kidney Disease Stages Using Data Mining Algorithms. *Informatics in Medicine Unlocked*, vol. 15, 2019.

202 RANGANATHAN, P.; PRAMESH, C. S.; AGGARWAL, R. Common Pitfalls in Statistical Analysis: Logistic Regression. *Perspectives in Clinical Research*, vol. 8, no. 3, pp. 148–151, 2017.

203 RAGHUNATHAN, T. E.; SOLENBERGER, P. W.; VAN HOEWYK, J. IVEware: Imputation and Variance Estimation Software. Survey Research Center, Institute for Social Research, University of Michigan, 2002.

204 RESHMA, S.; SALMA, S.; SR, A.; VISHNU PRIYA, S. R.; JANISHA, A. Chronic Kidney Disease Prediction Using Machine Learning. *International Journal of Engineering Research and Technology (IJERT)*, vol. 9, no. 7, 2020.

205 ROKACH, L.; MAIMON, O. Decision Trees. *Springer US*, pp. 165-192, 2005. ISBN 978-0-387-25465-4.

206 ROMÃO JUNIOR, J. E. Doença Renal Crônica: Definição, Epidemiologia e Classificação. *Brazilian Journal of Nephrology*, vol. 26, no. 3, supl. 1, pp. 1-3, 2004.

207 ROSENBLATT, F. The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain. *Psychological Review*, vol. 65, no. 6, pp. 386–408, 1958.

208 ROSENBERG, M. Overview of the Management of Chronic Kidney Disease in Adults. UpToDate, 2024.

209 RUBIN, D. B. Inference and Missing Data. *Biometrika*, vol. 63, no. 3, pp. 581–592, 1976.

210 RUBINI, L.; SOUNDARAPANDIAN, P.; ESWARAN, P. Chronic Kidney Disease DataSet, UCI Machine Learning Repository, 2015.

211 RUBINSTEY, A.; FELDMAN, S. Fancyimpute: An Imputation Library for Python, Version 0.7.0, 2016. Disponível em `https://github.com/iskandr/fancyimpute`. Acesso em 15 de Junho de 2024.

212 RUMELHART, D. E.; HINTON, G. E.; WILLIAMS, R. J. Learning Representations by Back-Propagating Errors. *Nature*, vol. 323, no. 6088, pp. 533-536, 1986.

213 SAIDI, A.; OTHMAN, S. B.; DHOUIBI, M.; SAOUD, S. B. FPGA-based implementation of classification techniques: A survey. *Integration*, vol. 81, pp. 280-299, 2021. ISSN 0167-9260.

214 SAMUEL, A. M. Some Studies in Machine Learning Using the Game of Checkers. *IBM Journal of Research and Development*, vol. 3, no. 3, pp. 210-229, 1959.

215 SANMARCHI, F.; FANCONI, C.; GOLINELLI, D.; GORI, D.; HERNANDEZ-BOUSSARD, T.; CAPODICI, A. Predict, Diagnose, and Treat Chronic Kidney Disease with Machine Learning: A Systematic Literature Review. *Journal of Nephrology*, vol. 36, pp. 1101–1117, 2023.

216 SANTOS, K. K. *et al.* Perfil epidemiológico de pacientes renais crônicos em tratamento. *Revista de Enfermagem UFPE Online*, vol. 12, no. 9, pp. 2293-2300, set. 2018.

217 SCHENA, F. P.; ANELLI, V. W.; ABBRESCIA, D. I.; DI NOIA, T. Prediction of Chronic Kidney Disease and Its Progression by Artificial Intelligence Algorithms. *Journal of Nephrology*, vol. 35, pp. 1953-1971, 2022.

218 SCIKIT-LEARN. StratifiedKFold Documentation. 2024. Disponível em `https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.StratifiedKFold.html`. Acesso em 23 de Dezembro de 2023.

219 SCIKIT-LEARN. train_test_split Documentation. 2024. Disponível em `https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html`. Acesso em 23 de Dezembro de 2023.

220 SCOPUS. *Scopus Database*. Elsevier, 2024. Disponível em `https://www.scopus.com`. Acesso em 15 de julho de 2024.

221 SCORALICK, J. P.; IWASHIMA, G. C.; COLUGNATI, F. A. B.; GOLIATT, L.; CAPRILES, P. V. S. Z. A Random Forest Classifier Combined with Missing Data Strategies for Predicting Chronic Kidney Disease. *Advances in Intelligent Systems and Computing*, vol. 1355, pp. 255-265, 2021.

222 SCORALICK, J. P.; IWASHIMA, G. C.; COLUGNATI, F. A. B.; GOLIATT, L.; CAPRILES, P. V. S. Z. An Extreme Gradient Boosting Classifier for Predicting Chronic Kidney Disease Stages. *Advances in Intelligent Systems and Computing*, vol. 1351, pp. 901-910, 2021.

223 Secretaria de Estado de Saúde, Governo do Estado de Minas Gerais. Resolução SES nº 2.606 de 7 de Setembro de 2010. Disponível em `https://www.saude.mg.gov.br/images/documentos/Resolucao%202606_10.pdf`. Acesso em 3 Junho 2022.

224 SHARMA, S.; SHARMA, V.; SHARMA, A. Performance Based Evaluation of Various Machine Learning Classification Techniques for Chronic Kidney Disease Diagnosis. *arXiv preprint* arXiv:1606.09581, 2016.

225 SILVA, A. R. *et al.* Doenças Crônicas Não Transmissíveis e Fatores Sociodemográficos Associados a Sintomas de Depressão em Idosos. *Jornal Brasileiro de Psiquiatria*, vol. 66, no. 1, pp. 45-51, mar. 2017.

226 SILVA, T. A. Estudo do Desempenho da Combinação de Preditores Baseados em Cópulas e Máquinas de Vetor de Suporte para Séries Temporais Úteis ao Desenvolvimento Sustentável. 2020. Dissertação (Mestrado), Universidade Federal Rural de Pernambuco.

227 SILVER, D. *et al.* Mastering the Game of Go with Deep Neural Networks and Tree Search. *Nature*, vol. 529, pp. 484-489, 2016.

228 SILVEIRA, A.; SOBRINHO, A.; SILVA, L.; COSTA, E.; PERKUSICH, A.; PINHEIRO, M. Machine Learning to Early Prediction of Chronic Kidney Disease: Using Imbalanced and Limited Size Data Sets, 2021.

229 SKLAR, A. Distribution Functions of n Dimensions and Margins. *Institute of Statistics at the University of Paris*, pp. 229-231, 1959.

230 Sociedade Latino Americana de Nefrologia e Hipertensão (SLANH). Informe 2018 - Registro Latinoamericano de Diálise e Transplante Renal. SLANH, 2018. Disponível em `http://slanh.net/reporte-2018/`. Acesso em 12 de Junho de 2021.

231 SMITH, D. Chronic Kidney Disease: A Global Crisis. Siemens Healthineers, 2018. Disponível em `https://www.siemens-healthineers.com/en-id/news/chronic-kidney-disease.html`. Acesso em 25 Maio de 2022.

232 SMITH, S. *et al.* Using DeepSpeed and Megatron to Train Megatron-Turing NLG 530B, A Large-Scale Generative Language Model, 2022.

233 Sociedade Brasileira de Nefrologia. O que é um transplante renal? 2021. Disponível em`https://www.sbn.org.br/orientacoes-e-tratamentos/tratamentos/transplante-renal/`. Acesso em 7 de Junho de 2021.

234 SODRE, F. L.; COSTA, J. C. B.; LIMA, J. C. C. Avaliação da Função e da Lesão Renal: Um Desafio Laboratorial. *Jornal Brasileiro de Patologia e Medicina Laboratorial*, vol. 43, no. 5, pp. 329-337, out. 2007.

235 SONG, Y. Y.; LU, Y. Decision Tree Methods: Applications for Classification and Prediction. *Shanghai Archives of Psychiatry*, vol. 27, no. 2, pp. 130-135, 2015.

236 STERNE, J. A.; WHITE, I. R.; CARLIN, J. B.; SPRAT, M.; ROYSTON, P.; KENWARD, M. G.; WOOD, A. M.; CARPENTER, J. R. Multiple Imputation for Missing Data in Epidemiological and Clinical Research: Potential and Pitfalls. *BMJ*, vol. 338, b2393, 2009.

237 STOPA, S. R. *et al.* Prevalência da Hipertensão Arterial, do Diabetes *Mellitus* e da Adesão às Medidas Comportamentais no Município de São Paulo, Brasil, 2003-2015. *Cadernos de Saúde Pública*, vol. 34, no. 10, e00198717, 2018.

238 SU, C. T.; CHANG, Y. P.; KU, Y. T.; LIN, C. M. Machine Learning Models for the Prediction of Renal Failure in Chronic Kidney Disease: A Retrospective Cohort Study. *Diagnostics*, vol. 12, no. 10, p. 2454, 2022.

239 SUN, J.; DU, W.; SHI, N. A Survey of kNN Algorithm. *Information Engineering and Applied Computing*, vol. 1, 2018.

240 TARWIDI, D.; PUDJAPRASETYA, S. R.; ADYTIA, D.; APRI, M. An Optimized XGBoost-Based Machine Learning Method for Predicting Wave Run-Up on a Sloping Beach. *MethodsX*, vol. 10, 2023. ISSN 2215-0161.

241 TAUNK, K.; DE, S.; VERMA, S.; SWETAPADMA, A. A Brief Review of Nearest Neighbor Algorithm for Learning and Classification. pp. 1255-1260, 2019.

242 TEIXEIRA, L. Seminário da Disciplina "Bioquímica do Tecido Animal". Programa de Pós-Graduação em Ciências Veterinárias da Universidade Federal do Rio Grande do Sul, 2013. Professor responsável pela disciplina: Félix H. D. González. Disponível em `https://www.ufrgs.br/lacvet/site/wp-content/uploads/2013/10/renalLiege.pdf`. Acesso em 17 de Maio de 2021.

243 TEKALE, S.; SHINGAVI, P.; WANDHEKAR, S. Prediction of Chronic Kidney Disease Using Machine Learning Algorithm. *IJARCCE*, vol. 7, pp. 92-96, 2018.

244 The Modification of Diet in Renal Disease Study: Design, Methods, and Results from the Feasibility Study. *American Journal of Kidney Diseases*, vol. 20, no. 1, pp. 18–33, 1992.

245 TIRAPANI, L. S. Avaliação do Impacto da Renda, Educação e Cor na Hipertensão Arterial, Diabetes Mellitus e Doença Renal Crônica. 2018. Tese (Doutorado), Universidade Federal de Juiz de Fora.

246 TIRAPANI, L. S.; PINHEIRO, H. S.; MANSUR, H. N.; OLIVEIRA, D.; HUAIRA, R. M. N. H.; HUAIRA, C. C.; GRINCENKOV, F. R. S.; BASTOS, M. G.; FERNANDES, N. M. S. Impact of Social Vulnerability on the Outcomes of Predialysis Chronic Kidney Disease Patients in an Interdisciplinary Center. *Jornal Brasileiro de Nefrologia*, vol. 37, pp. 19-26, 2015.

247 MANESH, T. R.; KUMAR, V. V.; KUMAR, V. D.; GEMAN, O.; MARGALA, M.; GUDURI, M. The Stratified K-Folds Cross-Validation and Class-Balancing Methods with High-Performance Ensemble Classifiers for Breast Cancer Classification. *Healthcare Analytics*, vol. 4, 2023.

248 TU, J. V. Advantages and Disadvantages of Using Artificial Neural Networks Versus Logistic Regression for Predicting Medical Outcomes. *Journal of Clinical Epidemiology*, vol. 49, no. 11, pp. 1225-1231, 1996. ISSN 0895-4356.

249 TURING, A. M. Computing Machinery and Intelligence. *Mind*, vol. 59, no. 236, pp. 433–460, 1950.

250 UCSF HEALTH. Creatinine Blood Test, 2024. Disponível em `https://www.ucsfhealth.org/medical-tests/creatinine-blood-test#:~:text=Normal%20Results,person's%20size%20and%20muscle%20mass`. Acesso em 15 de Junho de 2024.

251 UCSF HEALTH. Phosphorus Blood Test, 2024. Disponível em `https://www.ucsfhealth.org/medical-tests/phosphorus-blood-test`. Acesso em 15 de Junho de 2024

252 UDDIN, S.; HAQUE, I.; LU, H. *et al.* Comparative Performance Analysis of K-Nearest Neighbour (KNN) Algorithm and Its Different Variants for Disease Prediction. *Scientific Reports*, vol. 12, 6256, 2022.

253 U. S. NATIONAL LIBRARY OF MEDICINE. PubMed: A Free Resource Developed and Maintained by the National Center for Biotechnology Information (NCBI), at the U.S. National Library of Medicine (NLM), 2024. Disponível em `https://pubmed.ncbi.nlm.nih.gov`. Acesso em 14 Maio de 2024.

254 UNITED STATES RENAL DATA SYSTEM (USRDS). US Renal Data System 2019 Annual Data Report: Epidemiology of Kidney Disease in the United States. National Institutes of Health, National Institute of Diabetes and Digestive and Kidney Diseases, 2019.

255 UYANIK, G. K.; GÜLER. A Study on Multiple Linear Regression Analysis. *Procedia - Social and Behavioral Sciences*, vol. 106, pp. 234-240, 2013. ISSN 1877-0428.

256 VAN BUREN, S. Flexible Imputation of Missing Data, Second Edition. Chapman and Hall/CRC, 2018.

257 VAPNIK, V.N. The Nature of Statistical Learning Theory. Springer-Verlag, New York, 1995.

258 VASWANI, A.; SHAZEER, N.; PARMAR, N.; USZKOREIT, J.; JONES, L.; GOMEZ, A.; KAISER, L.; POLOSUKHIN, I. Attention is All You Need. *Advances in Neural Information Processing Systems*, vol. 30, 2017.

259 WANG, W.; CHAKRABORTY, G.; CHAKRABORTY, B. Predicting the Risk of Chronic Kidney Disease (CKD) Using Machine Learning Algorithm. *Applied Sciences*, vol. 11, no. 1, 202, 2021.

260 WANG, J.; LU, S.; WANG, S. H.; ZHANG, Y. D. A Review on Extreme Learning Machine. *Multimedia Tools and Applications*, vol. 81, 2022.

261 WASKOM, M. Seaborn: Statistical Data Visualization. *Journal of Open Source Software*, vol. 6, no. 60, p. 3021, 2021.

262 WEBMD. Do I Need a Calcium Blood Test?, 2024. Disponível em `https://www.webmd.com/a-to-z-guides/do-i-need-a-calcium-blood-test`. Acesso em 15 de Julho de 2024.

263 WEHRMEISTER, F. C.; WENDT, A. T.; SARDINHA, L. M. V. Iniquidades e Doenças Crônicas Não Transmissíveis no Brasil. *Epidemiologia e Serviços de Saúde: Revista do Sistema Único de Saúde do Brasil*, vol. 31, spe1, 2022.

264 WEISMAN, D. S.; THAVARAJAH, S.; JAAR, B. G. Prime Time for Chronic Kidney Disease. *BMC Nephrology*, vol. 24, p. 295, 2023.

265 WHALEY-CORNECÇ, A.; NISTALA, R.; CHAUDHARY, K. The Importance of Early Identification of Chronic Kidney Disease. *Missouri Medicine*, vol. 108, no. 1, pp. 25–28, 2011.

266 WIEDERHOLD, G.; MCCARTHY, J. Arthur Samuel: Pioneer in Machine Learning. *IBM Journal of Research and Development*, vol. 36, pp. 329-331, 1992.

267 WORLD HEALTH ORGANIZATION. Noncommunicable Diseases, 2023. Disponível em `https://www.who.int/news-room/fact-sheets/detail/noncommunicable-diseases`. Acesso em 19 de Julho 2024.

268 WONG, G.; BERNIER-JEAN, A.; ROVIN, B.; RONCO, P. Time for Action: Recognizing Chronic Kidney Disease as a Major Noncommunicable Disease Driver of Premature Mortality. *International Society of Nephrology*, Elsevier, 2024.

269 WOOLDRIDGE, M.; JENNINGS, N. R. Intelligent Agents: Theory and Practice. *The Knowledge Engineering Review*, vol. 10, pp. 115-152, 1995.

270 WU, C. C.; ISLAM, M. M.; POLY, T. N.; WENG, Y. C. Artificial Intelligence in Kidney Disease: A Comprehensive Study and Directions for Future Research. *Diagnostics*, vol. 14, no. 4, p. 397, 2024.

271 ZADEH, L. A. Fuzzy Sets. *Information and Control*, vol. 8, issue 3, pp. 338-353, 1965. ISSN 0019-9958.

272 ZHANG, S. Nearest Neighbor Selection for Iteratively kNN Imputation. *Journal of Systems and Software*, vol. 85, issue 11, pp. 2541-2552, 2012.

273 ZHANG, H.; XIE, P.; XING, E. Missing Value Imputation Based on Deep Generative Models, 2018.

274 ZHANG, S. *et al.* OPT: Open Pre-trained Transformer Language Models, 2022.

275 ZOU, X.; HU, Y.; TIAN, Z.; SHEN, K. Logistic Regression Model Optimization and Case Analysis. *2019 IEEE 7th International Conference on Computer Science and Network Technology (ICCSNT)*, pp. 135-139, 2019.