UNIVERSIDADE FEDERAL DE JUIZ DE FORA
INSTITUTO DE CIÊNCIAS EXATAS
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

Mayara Amanda da Silva

Título: Avaliação de Modelos de Aprendizado de Máquina para Classificação de Gestantes e Predição de Gravidez de Risco Usando o Histórico de Consultas Médicas

Juiz de Fora
2021

Mayara Amanda da Silva

Título: Avaliação de Modelos de Aprendizado de Máquina para Classificação de Gestantes e Predição de Gravidez de Risco Usando o Histórico de Consultas Médicas

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de Juiz de Fora como requisito parcial à obtenção do título de Mestre em Ciência da Computação.

Orientador: Dr. Alex Borges Vieira

Coorientadores: Dr. Artur Ziviani, Dr. Heder Soares Bernardino

Juiz de Fora

2021

Ficha catalográfica elaborada através do Modelo Latex do CDC da UFJF
com os dados fornecidos pelo(a) autor(a)

Silva, Mayara Amanda da.

Título : Avaliação de Modelos de Aprendizado de Máquina para Classificação de Gestantes e Predição de Gravidez de Risco Usando o Histórico de Consultas Médicas / Mayara Amanda da Silva. – 2021.

79 f. : il.

Orientador: Alex Borges Vieira

Coorientadores: Dr. Artur Ziviani, Dr. Heder Soares Bernardino

Dissertação (Mestrado) – Universidade Federal de Juiz de Fora, Instituto de Ciências Exatas. Programa de Pós-Graduação em Ciência da Computação, 2021.

1. Redes de atenção à saúde. 2. Aprendizado de máquina aplicado à saúde. 3. Sistemas de apoio à decisão clínica. 4. Predição de gravidez de risco. I. Sobrenome, Nome do orientador, orient. II. Título.

Mayara Amanda da Silva

Título: Avaliação de Modelos de Aprendizado de Máquina para Classificação de Gestantes e Predição de Gravidez de Risco Usando o Histórico de Consultas Médicas

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de Juiz de Fora como requisito parcial à obtenção do título de Mestre em Ciência da Computação.

Aprovada em 06 de outubro de 2021

BANCA EXAMINADORA

Dr. Alex Borges Vieira - Orientador
Universidade Federal de Juiz de Fora

Dr. Heder Soares Bernardino - Coorientador
Universidade Federal de Juiz de Fora

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

Dra. Márcia Ito
Programa de Mestrado Profissional em Sistemas Produtivos/CEETEPS

Dedico este trabalho a toda minha família e amigos.

AGRADECIMENTOS

Primeiramente gostaria de agradecer a Deus por me permitir todos os momentos bons que vivi, pelos aprendizados que tive, pelas pessoas que conheci e por me mostrar que quando acreditamos, todas as coisas podem dar certo.

Agradeço a minha família pela paciência e ausência durante todos esses anos. Ao meu pai Robson *(in memorian)* por ter me incentivado e minha mãe Léa por batalhar muito para me permitir estudar e investir na minha carreira. Agradeço ao meu noivo Paulo pelo incentivo, pelos momentos de compreensão, pela ajuda e paciência comigo durante esses dois anos.

Agradeço a todos os amigos que fiz no PPGCC e que estiveram comigo durante esses dois anos. Em especial, agradeço a Airton, Genilson, Jorge, Frederico e Tales pelas conversas, pelas trocas de conhecimento e experiências, pelas ajudas e pelos momentos de companheirismo.

Dedico este trabalho ao meu orientador, Professor Doutor Alex Borges Vieira. Agradeço por confiar em mim e acreditado que eu conseguiria realizar esse trabalho. Obrigada pelos ensinamentos, pela paciência e compreensão em todos os momentos em que tive dificuldades, jamais conseguiria sem você. Dedico este trabalho também ao Professor Doutor Artur Ziviani (*in memorian*) por contribuir muito para o desenvolvimento deste trabalho, pelas ajudas e por todo incentivo que me deu.

Agradeço a UFJF pela estrutura, pelos momentos bons e ruins que tive durante esse período. A CAPES, pelo apoio financeiro sem o qual este trabalho não poderia ser realizado. Ao PPGCC pela oportunidade, por todo apoio, por toda confiança depositada no meu potencial e a todos os professores que contribuíram para a minha formação.

Intelligence is the ability to adapt to change.
Stephen Hawking

RESUMO

Atualmente, a tecnologia disponível permite a criação e o armazenamento de quantidades cada vez maiores de dados. Além disso, a ampla disponibilidade de métodos de aprendizado de máquina auxilia na análise e interpretação do grande volume de dados disponíveis, o que traz claros benefícios para a sociedade. Sistemas de saúde notoriamente geram muitos dados. Há eventos importantes, sensíveis e que a análise de informações do volume de dados armazenados, podem contribuir para diagnósticos precoces. Por exemplo, gestantes tendem a ser acompanhadas por um longo período, o que gera informações importantes para tomadas de decisões. Portanto, a identificação de uma gestação de alto risco ou mesmo a detecção precoce de possíveis complicações pode garantir à gestante uma gestação com mais cuidado e conforto. Assim, neste trabalho temos como objetivo a avaliação de modelos de aprendizado de máquina para classificação de gestantes de risco. Inicialmente caracterizamos os atendimentos da gestante, onde são analisadas as trajetórias de atendimento de pacientes na rede de atenção básica de saúde. Avaliar como é a trajetória dos pacientes do SUS, seja entre os atendimentos, seja entre as unidades de atendimento, é uma atividade importante para o entendimento e aprimoramento do sistema. Com isso, aplicamos modelos de classificação para identificar gestantes em risco. Além disso, a partir dos dados existentes, criamos técnicas para agrupar automaticamente as gestantes baseado nas consultas realizadas. Tanto o modelo de classificação quanto a aplicação desse modelo em dados temporais foram testados em um grupo de gestante cadastradas no sistema público de saúde do município de São Paulo-Brasil. Os resultados demonstram que o modelo de classificação tem alto desempenho, e consegue classificar uma gravidez de risco, com acurácia de quase 74%, com apenas 4 semanas de dados de uma gestação. Os agrupamentos que o modelo automático gerou foram validadas por especialistas da área médica e as informações desses grupos são coerentes com os procedimentos realizados pelas gestantes. Por fim, acreditamos que os modelos criados e as discussões dessa dissertação poderão ser utilizadas como meio para auxiliar na tomada de decisão, tanto de médicos quanto de gestores de saúde.

Palavras-chave: Redes de atenção à saúde. Aprendizado de máquina aplicado à saúde. Sistemas de apoio à decisão clínica. Predição de gravidez de risco.

ABSTRACT

Today, modern technology allows the creation and storage of ever-increasing amounts of data. Furthermore, the wide availability of machine learning methods aids in the analysis and interpretation of the large volume of data available, which brings clear benefits to society. Health systems, notoriously, generate a lot of data. There are important sensitive events, and that the analysis of information from the volume of stored data can contribute to early diagnoses. For example, pregnant women tend to be followed for a long period, which generates important information for decision-making. Therefore, the identification of a high-risk pregnancy or even the early detection of possible complications can guarantee the pregnant woman a pregnancy with more care and comfort. Thus, in this work we aim to evaluate machine learning models for classifying pregnant women at risk. Initially, we characterize the care provided to pregnant women, where the trajectories of care for patients in the basic health network are analyzed. Assessing the trajectory of SUS patients, whether between care units or service units, is an important activity for understanding and improving the system. Thus, we apply classification models to identify pregnant women at risk. Furthermore, based on existing data, we created techniques to automatically group pregnant women based on medical appointments. Both the classification model and the application of this model in temporal data had their effectiveness tested in a group of pregnant women registered in the public health system in the city of São Paulo-Brazil. The results demonstrate that the classification model has high performance, and that is able to classify a pregnancy at risk, with an accuracy of almost 74%, with only 4 weeks of pregnancy data. The groupings generated by the automatic model were validated by specialists in the area, and the information from these groups is consistent with the procedures performed by the pregnant women. Finally, we believe that the models created and the discussions in this dissertation can be used as a way to assist in decision making, both for physicians and health managers.

Keywords: Healthcare networks. Machine learning applied to health. Clinical decision support systems. Prediction of risky pregnancy.

LISTA DE ILUSTRAÇÕES

Figura 1 – Organização das redes de atenção à saúde (RAS) e a interação entre três elementos principais 19
Figura 2 – Modelo ilustrativo de uma árvore de decisão. 23
Figura 3 – Diagrama das etapas metodológicas adotadas 38
Figura 4 – Filtragem temporal para seleção de gestantes da base de dados de modo a obter somente gestações completas registradas no período total da base de dados (2014-2015) 39
Figura 5 – Distribuições dos atendimentos por gestante, onde apresenta-se (a) uma função de distribuição cumulativa (CDF) e (b) uma distribuição geométrica com fitting dos dados sobre o histograma de dados reais. 40
Figura 6 – CIDs da categoria *Supervisão de Gravidez Normal* presentes na base . 41
Figura 7 – Fluxo teórico de atendimento à gestante recomendado pelo SUS 43
Figura 8 – Gráfico de pontos temporais dos atendimentos a gestantes 44
Figura 9 – Trajetórias de atendimentos mais frequentes pelas gestantes 45
Figura 10 – Trajetórias das gestantes pelas categorias de CIDs, onde têm-se os vértices que representam as categorias de CIDs mais frequentes, vértices de início e fim que representam o começo e fim de cada sequência e a porcentagem nas arestas que corresponde a parcela de registros que saem do vértice de origem ate o vértice de destino. 45
Figura 11 – Matriz de precedência das regiões dos atendimentos 47
Figura 12 – Função de distribuição cumulativa (CDF) de todos os registros vinculados por gestante 48
Figura 13 – Distribuição de frequências dos procedimentos da base de dados. . . . 49
Figura 14 – Quantidade de registros de cada classe na base de dados 51
Figura 15 – Acurácia x alfas para os dados de treino e de teste 54
Figura 16 – Metodologia utilizada para identificar o início da gestação. 57
Figura 17 – Acurácia do modelo de classificação temporal 59
Figura 18 – Avaliação do modelo de acordo com as métricas de avaliação e com intervalos de confiança 60
Figura 19 – Variação dos principais componentes gerados pelo PCA 62
Figura 20 – Imagem do gráfico gerada pela aplicação do método *elbow* 63
Figura 21 – Visualização 3D do agrupamento das gestantes, onde temos quatro grupos representados por cores diferentes para melhor visualização dos pontos pertencentes a cada grupo 64

LISTA DE TABELAS

Tabela 1 – Diagnósticos mais frequentes por categoria de CID. 41
Tabela 2 – Diagnósticos mais frequentes como primeiro e último registro. 42
Tabela 3 – Especialidades que mais registraram os respectivos CIDs. 43
Tabela 4 – Rendimento nominal médio mensal de pessoas com 10 anos ou mais de idade segundo a CRS do Município de São Paulo em 2010. 46
Tabela 5 – Região dos atendimentos de gestantes com uma consulta. 46
Tabela 6 – Matriz de confusão do modelo com a classificação enviesada. 51
Tabela 7 – Resultado médio das métricas geradas pelo modelo de classificação usando técnicas de balanceamento. 52
Tabela 8 – Matriz de confusão do modelo de classificação das gestantes sem poda. 54
Tabela 9 – Matriz de confusão do modelo de classificação das gestantes após a poda. 54
Tabela 10 – Valores dos alfas escolhidos por período de data 58
Tabela 11 – Número médio de procedimentos/diagnósticos realizados por grupo, onde intuitivamente os grupos representam: gestações de médio risco, gestações normais, gestações de baixo risco ou normal com atenção e gestações de alto risco respectivamente. 65

LISTA DE ABREVIATURAS E SIGLAS

AI	Artificial Intelligence
CART	*Classification and Regression Trees*
CID	Classificação Estatística Internacional de Doenças e Problemas Relacionados com a Saúde
FFQ	*Food Frequency Questionnaire*
MLP	*Perceptron Multicamadas*
PC	*Principal Component*
PCA	*Principal Component Analysis*
RF	*Random Forest*
RAS	Redes de Atenção à Saúde
SIGTAP	Sistema de Gerenciamento da Tabela de Procedimentos, Medicamentos e OPM do SUS
SUS	Sistema Único de Saúde
SVM	*Support Vector Machine*
TMLE	*Targeted Maximum Likelihood Estimation*

LISTA DE SÍMBOLOS

$\forall$	Para todo
$\in$	Pertence

SUMÁRIO

1 INTRODUÇÃO 15
1.1 MOTIVAÇÃO 15
1.2 OBJETIVOS DO TRABALHO 16
1.3 ORGANIZAÇÃO 17
2 FUNDAMENTAÇÃO TEÓRICA 18
2.1 REDES DE ATENÇÃO À SAÚDE 18
2.2 ASSISTÊNCIA PRÉ-NATAL E GESTAÇÃO DE RISCO 19
2.3 APRENDIZADO DE MÁQUINA 20
2.3.1 Aprendizado supervisionado 21
2.3.1.1 Classificação dos dados 22
2.3.2 Aprendizado não supervisionado 25
2.3.2.1 Medidas de similaridade 26
2.3.2.2 Agrupamento dos dados 28
2.3.3 Análise de componentes principais (PCA) 32
3 TRABALHOS RELACIONADOS 34
3.1 ANÁLISE E TRAJETÓRIA ESPACIAL 34
3.2 TECNOLOGIAS DE BIG DATA APLICADO À ASSISTÊNCIA MÉDICA 35
3.3 APRENDIZADO DE MÁQUINA COMO AUXÍLIO À GESTANTES . 35
4 CONJUNTO DE DADOS 38
4.1 DESCRIÇÃO DOS DADOS 38
4.2 CARACTERIZAÇÃO INICIAL DOS DADOS 39
4.3 TRAJETÓRIA DE ATENDIMENTOS DAS GESTANTES 42
4.4 TRAJETÓRIA ESPACIAL DAS GESTANTES 45
5 CLASSIFICAÇÃO DE GESTANTES 49
5.1 CONJUNTO DE DADOS 49
5.2 MODELO PROPOSTO 50
5.2.1 Descrição do modelo 50
5.2.2 Balanceamento dos dados 50
5.2.3 Poda da árvore de classificação 53
5.3 AVALIAÇÃO DO MODELO DE CLASSIFICAÇÃO DE GESTANTES 53
6 PREDIÇÃO DE GESTAÇÕES 56
6.1 FILTRAGEM TEMPORAL DAS CONSULTAS 56
6.2 PODA DAS ÁRVORES TEMPORAIS 58
6.3 AVALIAÇÃO DO MODELO DE PREDIÇÃO 59
7 AGRUPAMENTO DE GESTANTES 61
7.1 CONJUNTO DE DADOS 61
7.2 MODELO PROPOSTO 61

7.3 ANÁLISES DOS AGRUPAMENTOS GERADOS PELO MODELO . . 63
8 CONCLUSÕES 66
REFERÊNCIAS 67
APÊNDICE A – Lista de CIDs Exclusivos de Gestantes . . . 77
APÊNDICE B – Lista Com os Procedimentos Mais Frequentes 79

1 INTRODUÇÃO

1.1 MOTIVAÇÃO

O sistema público de saúde no Brasil, denominado por Sistema Único de Saúde (SUS), fornece acesso integral, universal e gratuito para toda a população do país (105). O SUS é organizado a partir do conceito de regionalização de atendimentos. Quanto à distribuição espacial de centros de atenção à saúde, existem no Brasil 438 regiões de saúde. Porém, em alguns estados, foram conformadas mesorregiões de saúde de modo a agrupar conjuntos de regiões em busca de oferta de serviços com maior grau de complexidade (95). Dada essa estrutura baseada em regionalização, conhecer as trajetórias dos pacientes através de seus atendimentos permite entender a oferta de serviços do SUS nas regiões de saúde, sobretudo sob a perspectiva dos pacientes.

Além disso, a organização dos serviços pode apresentar falhas de comunicação e de integração entre os níveis de atenção à saúde, problemas os quais as análises da trajetória do paciente podem ajudar a mitigar (108). Por isso, o desenvolvimento de técnicas que permitam trabalhar com os dados de forma gerencial, pode ser importante para acelerar o processo de avaliação da assistência hospitalar. Como solução, têm-se as redes de atenção à saúde que podem ser definidas como um conjunto de sistemas e mecanismos utilizados para auxiliar a tomada de decisões em um ambiente hospitalar (70). Em geral, essas redes têm componentes que podem ser descritos pelos tipos de serviços prestados, especialidade médica, unidades de saúde, entre outros. Por exemplo, o atendimento de um paciente em uma unidade de saúde pode conduzir esse paciente para um novo atendimento, que pode ser entre especialidades ou para um novo atendimento entre unidades, gerando interconexões entre os componentes acima, o que transforma essas redes de atenção à saúde em um sistema complexo (48).

Outro ponto a ser considerado é que avaliar como é a trajetória dos pacientes do SUS, seja entre os procedimentos, seja entre as unidades de atendimento, é uma atividade importante. É possível através dessas trajetórias, analisar como o paciente faz o uso do sistema de saúde e como os recursos são gastos. Porém, mesmo com grande parte dos dados abertos, avaliar o funcionamento de tal sistema complexo não é trivial. Em um país com dimensão continental, como o Brasil, com metrópoles como São Paulo e Rio de Janeiro, a visualização e a compreensão da dinâmica dos atendimentos aos pacientes do SUS passa por uma série de tratamentos de dados, como a subdivisão de registros em casos específicos.

Um exemplo que a análise de trajetória de pacientes pode ter alta relevância no SUS, do ponto de vista social, é o acompanhamento dos atendimentos, permitindo um melhor tratamento e um auxílio na tomada de decisão sobre os diagnósticos. Quando focadas em gestantes, pode-se acompanhar desde a identificação da gravidez, ou da primeira consulta

pré-natal, até a última consulta puerperal. Com esse acompanhamento bem definido, pode-se avaliar a sequência completa de procedimentos e realizar uma comparação com o modelo recomendado pelo SUS, por exemplo.

Além disso, salienta-se que a gestação é um fenômeno importante na vida de uma mulher e durante a mesma ocorre uma série de alterações fisiológicas (26). Com isso, é necessário que haja uma atenção com a saúde e um acompanhamento médico para ser possível ter uma gestação com mais conforto e cuidado. Note-se que no Brasil, a maior parte das gestações não tem intercorrências, mas existe uma parcela pequena de gestantes que podem apresentar algum problema durante a gestação por serem portadores de alguma comorbidade ou condição sociobiológica (72) como, por exemplo, hipertensão arterial, diabetes, obesidade, alcoolismo (64). Essa parcela caracteriza um grupo denominado gestantes de risco. No manual técnico de gestação de alto risco do ministério da saúde do Brasil descreve a gestação de alto risco como "aquela em que a vida ou a saúde da mãe e/ou do feto e/ou do recém-nascido têm maiores chances de serem atingidas que as da média da população considerada" (18). Para esse grupo de gestantes, o ideal é que exista alguma atenção especializada, seja com cuidados mínimos às gestantes com poucos riscos até atendimentos específicos para gestantes que necessitam o máximo de atenção (72). Vale salientar que uma gestação classificada como normal pode se tornar de risco a qualquer momento durante o decorrer da gestação ou no trabalho de parto. Nesses casos, a identificação precoce e adequada de uma gestação de risco pode identificar morbidades graves e até mesmo evitar agravamentos e prevenir morte materna e fetal.

Nos últimos anos, houve uma discussão significativa sobre como o aprendizado de máquina pode ser usado em diversos setores da sociedade (11), destacando-se a área da saúde. Com isso, têm-se aplicações de aprendizado de máquina para melhorias na saúde, como previsão de início de doenças (19), tomada de decisões baseadas em previsões que informem diagnósticos, caminhos de atendimento clínico e estratificação de risco paciente (2), entre muitas outras aplicações. Sendo assim, tem-se como objetivo deste trabalho a proposta de aplicação de métodos de aprendizado de máquina para criar um modelo que seja capaz, de maneira automática, agrupar e classificar gestantes de acordo com suas características, possibilitando a predição de uma possível gestação de risco.

1.2 OBJETIVOS DO TRABALHO

Foi criado um modelo de classificação de gestação de risco usando os registros de atendimentos de gestantes durante todo o período de gestação. Além disso, foram realizadas classificações automáticas e predições baseadas nos procedimentos e consultas que as gestantes realizavam com o passar do tempo de gestação, sempre adicionando atendimentos a cada duas semanas. Esses modelos e as abordagens foram avaliadas com dados reais, do Sistema Único de Saúde da cidade de São Paulo - Brasil, com registros

referentes as 24.916 gestantes atendidas entre janeiro de 2014 e dezembro de 2015. Os resultados indicam que o método utilizado para a classificação das gestantes, classificou corretamente mais de 90% das gestantes, o que a princípio é um resultado satisfatório para a proposta deste trabalho. Além disso, os resultados mostram que quando o classificador foi aplicado ao conjunto de dados dividido por semana, foi possível detectar próximo da 15ª semana quase 80% de gestações de alto risco. Por fim, o método proposto para realizar o agrupamento gerou grupos coerentes quando foram analisadas as médias dos procedimentos realizados por cada grupo.

1.3 ORGANIZAÇÃO

Além deste capítulo introdutório, apresentamos no próximo capítulo a fundamentação teórica, onde se têm as definições e aplicações dos conceitos mais importantes usados neste trabalho. No capítulo 3, apresentamos a descrição e uma análise sucinta dos trabalhos que trouxeram uma grande contribuição para o desenvolvimento deste estudo. No capítulo 4, mostramos uma descrição e caracterização dos dados usados, com a descrição das trajetórias percorridas pelas gestantes, seja com foco nos atendimentos realizados por ela, ou então, baseado nas regiões percorridas pelas gestantes. No capítulo 5 são detalhadas as metodologias utilizadas para realizar a classificação das gestantes em normal ou de risco. Já no capítulo 6 descrevemos o modelo usado para predição de gravidez de risco. No capítulo 7 apresentamos a metodologia utilizada para realizar a clusterização das gestantes e por fim, no capítulo 8 mostramos as conclusões da dissertação e trabalhos futuros.

2 FUNDAMENTAÇÃO TEÓRICA

Neste capítulo, são apresentados os conceitos fundamentais para o melhor entendimento da proposta deste trabalho, de modo a inserir o leitor no universo da pesquisa. Primeiro, serão conceituadas as redes de atenção à saúde na seção 2.1. Em seguida, apresentamos conceitos importantes de ciência de dados, com tópicos relacionados aos modelos de classificação usando algoritmos supervisionados e não supervisionados apresentados na seção 2.2.

2.1 REDES DE ATENÇÃO À SAÚDE

O conceito de redes de atenção à saúde já estava contido na Constituição Federal, no seu artigo 198, onde diz que "as ações e os serviços públicos de saúde integram uma rede regionalizada e hierarquizada e constituem um sistema único (...)" (51).

Mendes (70) conceitua as redes de atenção à saúde (RAS) como:

> (...) organizações poliárquicas de conjuntos de serviços de saúde, vinculados entre si por uma missão única, por objetivos comuns e por uma ação cooperativa e determinada população, coordenada pela atenção primária à saúde - prestada no tempo certo, no lugar certo, com o custo certo, com a qualidade certa e de forma humanizada, e com responsabilidades sanitárias e econômicas por esta população.

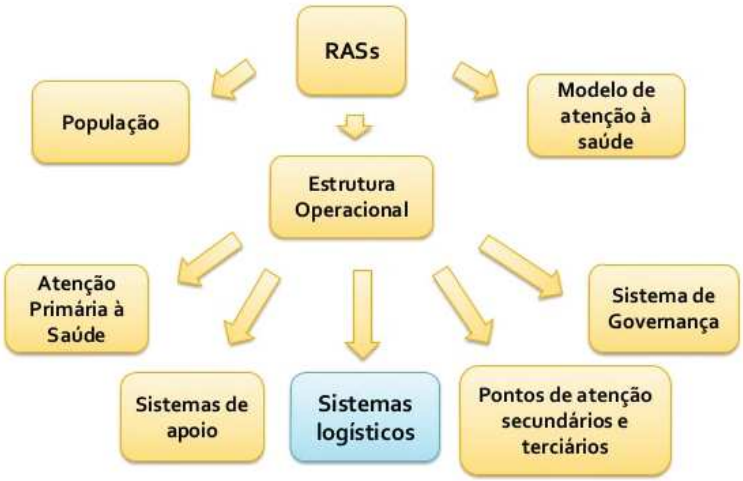
De forma simplificada, elas podem ser definidas como a organização dinâmica e horizontal de conjuntos de sistemas e serviços vinculados entre si (75) por um princípio fundamental, onde operam de maneira cooperativa e independente, permitindo oferecer atenção contínua e integral a determinada população de um território (69) e (51).

As redes de atenção à saúde constituem-se de três elementos: a população, a estrutura operacional e o modelo de atenção à saúde ilustrado pela Figura 1. Ao se referir ao primeiro elemento, à população, esta é colocada sob responsabilidade sanitária e econômica e deve ser organizada sob forma de gerenciamento, com ênfase à gestão da oferta da atenção à saúde (70) e (109).

No que se refere à estrutura operacional, afirma-se que ela é constituída pelos "nós" das redes e pelas ligações materiais e imateriais que comunicam esses diferentes nós. Pode-se definir os nós como o centro de comunicação e a atenção primária à saúde; os pontos de atenção secundários e terciários; os sistemas de apoio; os sistemas logísticos e o sistema de governança da rede de atenção à saúde (70) e (109).

O último elemento que compõe as RAS são os modelos de atenção à saúde, que são sistemas lógicos que organizam o funcionamento das RAS articulando relações entre

Figura 1 - Organização das redes de atenção à saúde (RAS) e a interação entre três elementos principais.



Fonte: `https://redehumanizasus.net/92656-operacionalizacao-das-redes-de-atencao-a-saude-rass` (2015).

os componentes da rede e intervenções sanitárias, de modo a prevalecer as melhorias de saúde (70) e (109).

A definição mais geral de uma rede pode ser caracterizada como uma abstração que permite codificar de alguma forma tipos de relacionamentos entre pares e objetos. De maneira informal, sua estrutura é definida pelo conjunto de relacionamentos existentes, ou seja, todas as ligações entre os pares e objetos (37). Cada atendimento realizado gera um fluxo de acesso aos serviços, criando relacionamentos entre os componentes da rede.

A relação entre esses componentes tem uma grande complexidade, visto que, cada conexão entre esses componentes possui características distintas e diversas interconexões, o que dificulta analisar o comportamento desta rede (48). Assim, é possível afirmar que o sistema de redes de atenção à saúde se assemelham a um sistema complexo, visto que um sistema complexo é um grupo composto por muitas partes que interagem, essas partes individuais são chamadas componentes (79).

2.2 ASSISTÊNCIA PRÉ-NATAL E GESTAÇÃO DE RISCO

A assistência pré-natal é o primeiro passo para o parto e nascimento humanizados (73) e seu principal objetivo é acolher a mulher desde o início da gestação. Pode ser definido como o acompanhamento da gestante, desde o momento confirmado da gravidez até o período do parto (118). Compreende um conjunto de procedimentos para prevenir, diagnosticar e tratar eventos indesejáveis à gestação, ao parto e ao recém-nascido (39) e (85). As práticas realizadas rotineiramente durante essa assistência estão associadas a

melhores desfechos perinatais (118).

Alguns trabalhos realizados no Brasil revelam que mulheres com menor renda familiar, menor escolaridade e não brancas são as que ingressam tardiamente no pré-natal e, quando o realizam, este é de mais baixa qualidade, revelando iniquidades sociais presentes na assistência (4) e (39).

Visto que a gestação é um fenômeno muito importante na vida de uma mulher, é possível observar que em sua grande maioria, muitas mulheres desfrutarão de experiências saudáveis e simples de gravidez e nascimento. No entanto, para algumas mulheres, essas experiências serão complicadas por condições médicas ou obstétricas que ameaçam seu bem-estar e/ou o bem-estar de seus bebês (57). Assim, qualquer condição médica ou obstétrica inesperada ou imprevista associada à gravidez com um perigo real ou potencial para a saúde ou o bem-estar da mãe ou do feto é considerada uma gravidez de alto risco (44).

Nesse caso, é necessário um acompanhamento frequente desse grupo de gestantes de alto risco. A ausência de um acompanhamento pré-natal e/ou deficiência está relacionada a maiores índices de morbimortalidade materna e perinatal (39) e (85). Outro ponto a ser considerado é que a avaliação do pré-natal pode contribuir para melhorar a assistência às gestantes, diminuindo os índices de morbimortalidade materna e perinatal (4).

Nos últimos 30 anos, o Brasil avançou muito na melhoria da atenção ao parto e ao nascimento, fruto de uma série de esforços e iniciativas do governo e da sociedade. Porém, a redução da morbimortalidade materna e infantil permanece um desafio (73). Além disso, de maneira geral, pode-se afirmar haver uma fragilidade na rede no que tange ao seguimento da mulher e da criança no pós-parto, assim como no acompanhamento do desenvolvimento da criança para que ela alcance todo seu potencial intelectual, cognitivo e motor (73).

2.3 APRENDIZADO DE MÁQUINA

Nas últimas décadas, o aprendizado de máquina se tornou um dos pilares da tecnologia da informação e é uma das áreas da computação que mais cresce (16), visto que, embora muitas vezes oculta, está presente em nosso cotidiano. É uma área que pode ser descrita como um processo que explora grandes volumes dados a procura de padrões consistentes, anomalias e correlações (107) e (67). Alguns problemas que o aprendizado de máquina trata são os problemas de classificação, regressão e agrupamento. A classificação baseia-se em fazer uma estimativa de categoria dos dados observados. A regressão funciona similarmente a classificação, diferenciando-se na estimativa feita, visto que a regressão estima um valor numérico. Já o agrupamento agrupa os dados observados em grupos também conhecidos como ‘clusters’. A grande diferença entre a classificação e regressão para o agrupamento é que, no agrupamento, os dados não precisam ser rotulados. A classificação

e regressão pertencem ao grupo de aprendizado supervisionado e o agrupamento pertence ao grupo de aprendizado não supervisionado.

Arthur Samuel (1959) parafraseou o aprendizado de máquina como: "*campo de estudo que dá aos computadores a habilidade de aprender sem serem explicitamente programados*" (96). Esse processo procura uma detecção automática de padrões nos dados para que assim, seja possível dotar programas com a habilidade de aprendizagem e adaptação, possibilitando através de técnicas e métodos a descoberta de outras informações não explícitas presentes nos bancos de dados (87). Além disso, explora a criação de algoritmos que podem aprender a partir de um grande corpo de dados e com seus erros e, com isso, realizar previsões sobre dados a partir de duas abordagens de aprendizagem: supervisionada, não supervisionada e por reforço. Isso permite produzir decisões e resultados confiáveis e repetíveis (88). Neste trabalho, utilizamos aprendizado supervisionado descrito na subseção **2.3.1** e não supervisionado que será descrito na subseção **2.3.2**.

2.3.1 Aprendizado supervisionado

Métodos supervisionados permitem que, a partir de um conjunto de dados previamente rotulados, seja possível encontrar um jeito de predizer rótulos já conhecidos (33). Ou seja, o conjunto de dados usado neste processo já possui as classes pré-definidas, denominado dados rotulados, pois se sabe de antemão a saída esperada para cada entrada de dados. Resumidamente, para cada conjunto de dados com rótulos previamente rotulados, o objetivo do algoritmo é aprender uma regra capaz de mapear as entradas baseando-se nas saídas corretamente, ou seja, para os dados, que seja possível predizer as classes já conhecidas.

Nesta categoria de aprendizado, a base de dados é dividida em dois grupos, um grupo contendo as características, também conhecidos como atributos previsores; e outro grupo com os rótulos que deseja realizar a predição. Como em qualquer modelo de aprendizado de máquina, é necessário dividir os dados em conjunto de treino, que permitirá criar o modelo e conjunto de teste responsável por verificar como o modelo se comporta com dados não vistos anteriormente.

Existem duas maneiras de predizer os rótulos, usando classificação e usando regressão (15). Se o rótulo é um número real, a tarefa chama-se regressão. Se o rótulo vem de um conjunto finito, então a tarefa chama-se classificação. A classificação é um método de mineração de dados muito utilizados para classificar elementos de um conjunto de dados em diferentes classes, baseando-se em características comuns (107) para obter um modelo de classificação e a partir deste modelo, predizer classes de novos elementos de um conjunto de dados (67).

2.3.1.1 Classificação dos dados

A classificação representa uma tarefa importante em projetos de aprendizado de máquina e mineração de dados. Seus objetivos principais consistem em (49): 1) induzir um classificador a partir de um conjunto de exemplos, chamado conjunto de treinamento, com valores de classe conhecidos e então 2) usando o classificador induzido para prever o valor da classe ou categoria de novos objetos dados os valores conhecidos de seus atributos.

Atualmente, os métodos mais conhecidos para classificação de dados são: Árvores de Decisão, Classificação de Naive Bayes, Regressão Linear de Mínimos Quadrados, Regressão Logística, Support Vector Machine e Ensemble Methods. Assim, apresentaremos o método de classificação por árvores de decisão que foi o método utilizado neste trabalho para realizar a classificação das gestantes.

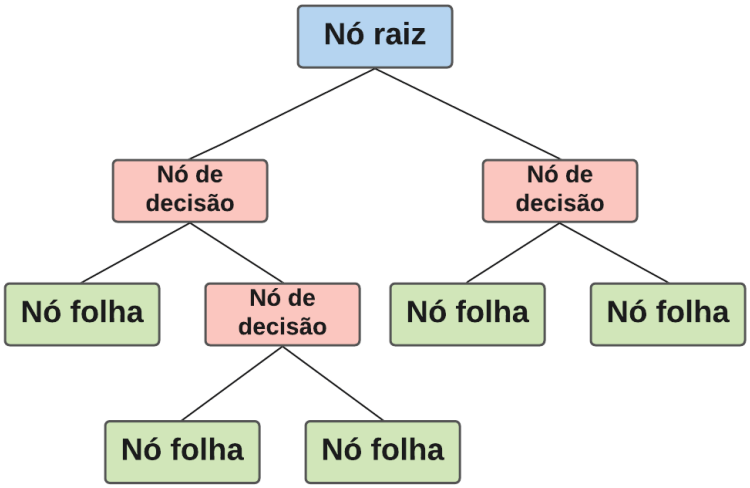
- **Árvore de decisão**

Uma árvore de decisão é uma representação de uma Tabela de decisão sob a forma de árvore (122), que particiona recursivamente um conjunto de dados em subdivisões menores baseado em um conjunto de testes em cada ramo da árvore (38). Tem uma estrutura hierárquica como um fluxograma (42), ilustrada pela Figura 2, composta por três elementos básicos: nó raiz, que representa o atributo que melhor divide o conjunto de dados, nós de decisão a atributos que correspondem aos diferentes valores de atributo possíveis e as folhas incluindo objetos que normalmente pertencem à mesma classe ou que são semelhantes. Essa representação permite induzir regras de decisão que serão utilizadas para classificar novas instâncias. Ou seja, cada caminho da raiz até uma folha corresponde a uma conjunção de atributos de teste e a árvore é considerada uma disjunção dessas conjunções (49).

De maneira mais simplificada, uma árvore de decisão é uma estrutura binária de ramificação usada para classificar um vetor de entrada arbitrário X. Cada nó na árvore normalmente contém uma comparação de características simples contra algum campo $x_i \in X$, como "é $x_i \geq 23.7$". O resultado de cada comparação é verdadeiro ou falso, determinando se devemos prosseguir para o filho esquerdo ou direito de um determinado nó. Essas estruturas às vezes são chamadas árvores de classificação e regressão (CART) porque podem ser aplicadas a uma classe mais ampla de problemas (106). Na literatura, a maioria das árvores de decisão são compostas de dois procedimentos principais: os procedimentos de construção (indução) e de classificação (inferência) (49).

Procedimento de construção: a construção de uma árvore de decisão é feita inicialmente com uma árvore vazia e selecionando para cada nó um atributo de teste apropriado usando uma medida de seleção de atributo. O princípio é selecionar o atributo que diminua ao máximo a mistura de classes entre cada subconjunto de treinamento criado pelo teste, facilitando assim a determinação das classes do objeto. O processo continua

Figura 2 - Modelo ilustrativo de uma árvore de decisão.



Fonte: Elaborado pela autora (2021).

para cada sub-árvore de decisão até chegar às folhas e fixar suas classes correspondentes (49).

Para a criação de um nó da árvore de decisão é necessário usar critérios de partição. O critério utilizado para realizar as partições é o da utilidade do atributo para a classificação (). Assim, é aplicado através deste critério um determinado ganho de informação a cada atributo. Com isso, o atributo escolhido para o corrente nó é aquele que possui o maior ganho de informação. Logo após essa aplicação, é iniciado um novo processo de partição (103). Os critérios mais utilizados para a classificação são: “Entropia” e “Índice de Gini”.

A entropia de um conjunto pode ser definida como sendo o grau de pureza desse conjunto (58) e (68). Partindo da entropia, o algoritmo confere o ganho de informação de cada atributo. Aquela que apresentar maior ganho de informação será o atributo do primeiro nó da árvore (121).

Para um conjunto nítido $E_0 \subseteq E$, a entropia de $E0$ pode ser definida pela equação 2.1.

$$\text{Entropia } (E_0) = -\sum_{i=1}^{L} p_i \log_2 p_i \tag{2.1}$$

onde p_i é a proporção do número de elementos que pertencem à classe C_i ao número total de todos os elementos em E_0. Às vezes E_0 é representado por $(p_1, p_2, \ldots, p_L)$ onde $\sum_{i=1}^{L} p_i = 1, 0 \leq p_i \leq 1$ para cada $i(1 \leq i \leq L)$ (121).

Assim, a entropia aumenta à medida que a proporção de cada classe chega ao equivalente. Quando todos os elementos em E0 pertencem à mesma classe, a entropia

é mínima; quando os elementos de todas as classes têm a mesma proporção, a entropia atinge seu máximo (121) e (29).

O índice de Gini foi desenvolvido com o objetivo de medir o grau de heterogeneidade dos dados (103), (56) e (25). Este indice em um determinado nó pode ser definido pela equação 2.2.

$$\text{Índice Gini } = 1 - \sum_{i=1}^{c} p_i^2 \tag{2.2}$$

onde *pi* é a frequência relativa de cada classe em cada nó e *c* é o número de classes.

Quando este índice é igual a zero, o nó é puro. Entretanto, quando o índice se aproxima do valor um, o nó é impuro (vai aumentando o número de classes uniformemente distribuídas neste nó) (103). Em exemplos de árvores de classificação com partições binárias, quando se utiliza o critério de Gini tende-se a isolar num ramo os registros que representam a classe mais frequente. Quando se utiliza a entropia, balanceia-se o número de registros em cada ramo (103).

Para classificar um novo objeto tendo apenas valores de todos os seus atributos, a partir da raiz da árvore construída segue-se o caminho correspondente ao valor observado do atributo no nó interno da árvore. Este processo continua até que uma folha seja encontrada. Finalmente, é usado o rótulo associado para obter o valor de classe previsto da instância em questão (49).

Após gerar as árvores de decisão, é comum que o classificador seja induzido para valores muito específicos no conjunto de treinamento (81), podendo gerar nós redundantes e sem informação, visto que o classificador super-ajustou os dados no processo de teste gerando árvores muito grandes quando não for identificado algum padrão nos dados.

Para lidar com esses problemas, existem técnicas de poda. O objetivo principal da poda consiste em eliminar os ruídos e o *overfitting* usando uma hipótese generalizada a partir do conjunto de testes, para que assim, melhore o desempenho em exemplos não vistos (82) e encontre uma forma para reduzir folhas redundantes e a profundidade da árvore gerada (10). As podas podem ser de dois tipos: pré-poda e pós-poda.

A técnica de pré-poda conta com regras para prevenir a geração de ramos que não melhoram a precisão da árvore. Já a técnica de pós-poda, que é o método mais comum (54) e (80) usado em árvores de decisão, consiste em substituir os nós e subárvores por folhas, diminuindo a complexidade final da árvore. Vale salientar que mesmo diminuindo significativamente o tamanho da árvore, a poda melhora a precisão da classificação de objetos invisíveis. Para (122) pode ser que a precisão da atribuição no conjunto de teste fique deteriorada, mas a precisão das propriedades de classificação da árvore aumenta.

- **Poda de custo-complexidade mínima**

É um método proposto por Breiman (15) e implementado no método de poda de complexidade de custo mínimo do algoritmo de aprendizagem da árvore de decisão CART. Inicialmente, ele calcula uma árvore tão grande quanto possível T_0 para ajustar os dados de treino, permitindo que o processo continue até todas as instâncias pertencerem à mesma classe. Após essa etapa, ele aplica o método de poda de complexidade de custo para calcular um conjunto de subárvores consecutivas $T_i, i \in \{1, 2, \ldots, R\}$ de tamanho decrescente da grande árvore original T_0 por poda progressiva para cima até seu nó raiz, onde T_R corresponde a subárvore que consiste apenas o nó raiz (91).

Para Breiman (15), a ideia por trás da poda de complexidade de custo mínimo é esta:

Para qualquer subárvore $T \preceq T_{max}$, defina sua complexidade como $|\bar{T}|$, o número de folhas em T. Seja $\alpha \geq 0$ um número real denominado o parâmetro de complexidade e defina a medida de custo-complexidade $R\alpha(T)$ como:

$$R_\alpha(T) = R(T) + \alpha|\bar{T}| \tag{2.3}$$

Assim, $R_\alpha(T)$ é uma combinação linear do custo da árvore e sua complexidade. O problema central do método é encontrar, para cada valor de α, a sub-árvore $T(\alpha) \preceq T_{max}$ que minimiza $R_\alpha(T)$, isto é,

$$T(\alpha) = \arg\min_{T \preceq T_{\max}} R_\alpha(T) \tag{2.4}$$

Embora α seja contínuo, há um número finito de subárvores de T_{max}. Assim, o processo de poda produz um número finito de subárvores $T_1, T_2, T_3, ...$ com progressivamente menos nós terminais. Por causa da finitude, o que acontece é que se $T(\alpha)$ é a árvore de minimização para um dado valor de α, então ela continua a minimizar conforme α aumenta até que o ponto de salto α' seja alcançado, e uma nova árvore $T(\alpha')$ torna-se minimizador e continua sendo o minimizador até o próximo ponto de salto α.

2.3.2 Aprendizado não supervisionado

Diferente dos algoritmos supervisionados, os algoritmos não supervisionados não possuem informação dos rótulos, ou seja, não existem marcadores pré-definidos sobre as classes que deseja-se prever. Por isso, os dados são considerados não-rotulados. Neste processo, o algoritmo não recebe os resultados esperados, devendo descobrir de forma automática, explorando os dados, os possíveis relacionamentos entre eles (100). Resumidamente, o processo identificará padrões entre os dados, com objetivo de agrupá-los baseado nas suas similaridades. Como não se tem rótulos definidos, não é necessário realizar o particionamento do conjunto de dados em conjunto de treinamento e conjunto de teste.

2.3.2.1 Medidas de similaridade

Na clusterização, um processo muito importante está na tentativa de identificar grupos de observações que podem estar presentes nos dados, e com isso, descobrir o quão próximos os objetos do conjunto de dados estão uns dos outros, ou quão distantes eles estão (86). Muitos métodos têm como ponto de partida uma matriz que reflete uma medida quantitativa de proximidade entre os elementos de um conjunto de dados. Quanto maior a similaridade dos objetos, ou então, quanto menor a dissimilaridade ou distância, mais próximos esses objetos se encontram. Essa matriz é chamada matriz de similaridade (36).

O conceito similaridade torna-se fundamental para a definição de um cluster (47), uma medida da similaridade entre dois padrões extraídos do mesmo espaço de recursos é essencial para a maioria dos procedimentos de clusterização. Ou seja, se existem dois pontos com características similares, baseado em algum critério utilizado na técnica de clusterização, então, estes dois pontos serão agrupados em um mesmo cluster. Caso as características não sejam similares, serão agrupados em clusters diferentes (55).

Vale salientar que diferentes medidas de similaridade ou dissimilaridade calculadas sobre um mesmo conjunto de dados podem e frequentemente irão resultar em soluções diferentes quando usadas como base para uma análise de cluster. Com isso, tem-se a necessidade de que as medidas de similaridades sejam escolhidas baseadas no tipo de variáveis utilizadas (36). A seguir serão apresentados quatro tipos de medidas de similaridade. Primeiro será apresentado medidas de similaridade entre os dados, em seguida medidas de similaridade para dados categóricos, após será conceituado medidas de similaridade entre clusters e por último similaridade euclidiana usando pontuações de quadrados mínimos no PCA.

- **Similaridade entre dados**
  Definir uma distância representa o mesmo que definir regras para atribuir números positivos entre dois dados. Sejam, portanto a, b e c três vetores com j elementos cada (120). A distância da função que associa dois dados a um número real positivo é denotada por $d(a, b)$, e para ser considerado uma métrica, deve possuir as seguintes propriedades:

  - $d\,(a, b) \geq 0$
  - $d\,(a, b) = 0$, se e somente se $a = b$
  - $d\,(a, b) = d\,(b, a)$
  - $d\,(a, c) \leq d\,(a, b) + d\,(b, c)$ onde $a, b, c \in X$.

  O cálculo de distância mais utilizado nas métricas de similaridade entre dois dados é a *distância de Minkowski* (11), e pode ser calculada através da Equação 2.5, onde d

é o número de atributos do dado (86).

$$d\left(x_i, x_j\right) = \sqrt[p]{\sum_{k=1}^{d}\left(\left|x_{ik} - x_{jk}\right|\right)^p}, p \geq 1 \tag{2.5}$$

Neste caso, quando $p = 2$ é calculada a *distância euclidiana*, representada na Equação 2.7. A *distância euclidiana* calcula a raiz da diferença quadrada entre as coordenadas de um par de objetos (102). Variações dos parâmetro p define distâncias diferentes. As variações mais comuns da distância de Minkowski são calculadas nas Equações 2.6, 2.7 e 2.8.

Distância de Manhattan, se $p = 1$

$$d\left(x_i, x_j\right) = \sum_{k=1}^{d}\left(\left|x_{ik} - x_{jk}\right|\right) \tag{2.6}$$

Distância Euclidiana, se $p = 2$

$$d\left(x_i, x_j\right) = \sqrt{\sum_{k=1}^{d}\left(\left|x_{ik} - x_{jk}\right|\right)^2} \tag{2.7}$$

Distância "sup", se $p \to \infty$

$$d\left(x_i, x_j\right) = \max_{1 \leq k \leq d}\left|x_{ik} - x_{jk}\right| \tag{2.8}$$

- **Similaridade para dados categóricos**
  Em um cenário onde os dados possuem todas as variáveis categóricas, medidas de similaridade também podem ser usadas. Neste caso, espera-se que as medidas sejam utilizadas para estar no intervalo [0,1], embora, também sejam ocasionalmente expressas no intervalo de 0-100%. Dois indivíduos i e j tem um coeficiente de similaridade s_{ij} de unidade se ambos tiverem valores idênticos para todas as variáveis. Caso o valor de similaridade for zero, isso indica que os dois indivíduos se diferem ao máximo em todas as variáveis (36).

- **Similaridade entre clusters**
  As medidas de similaridade acima são focadas na proximidade entre dois dados e não entre grupos de dados. Entretanto, em alguns algoritmos de clusterização, é necessário unir dois clusters similares. Uma maneira de medir essa similaridade é calculando a distância entre todos os pontos dos dois clusters, onde cada ponto pertence a um cluster diferente. Podem ser escolhidas a distância mínima entre todos os pares de dados *(distância do vizinho mais próximo)*, a distância máxima *(distância do vizinho mais distante)* ou a média das distâncias entre os pares de dados *(distância média)* (36).

- **Similaridade euclidiana no PCA**
  Atualmente, é possível encontrar uma ampla quantidade de soluções e interpretações resultantes de um conjunto de dados, baseado em diferentes medidas de similaridade e que é, brevemente revisada antes da demonstração de recuperação de padrão usando pontuações derivadas da similaridade euclidiana (32). Em muitos casos, a equação fundamental do PCA é expressa como:

$$\mathbf{Z} = \mathrm{FA}^{\mathrm{T}}, \tag{2.9}$$

  onde Z é a matriz de dados $(p \times n)$ e, utilizando a nomenclatura do PCA, A é a matriz $(n \times n)$ e F é a matriz $(p \times n)$ de pontuações. No PCA, os autovalores (V) são escalados pela raiz quadrada de seus respectivos autovalores $\left(\mathbf{\Lambda}^{1/2}\right)$, o que produz a matriz de carregamento (A). Apesar da fórmula fundamental, a maneira tradicional pela qual (V) é derivado através da diagonalização de uma matriz de similaridade, E, como:

$$\mathbf{E} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^{\mathrm{r}} \tag{2.10}$$

  Assim, tem-se que qualquer vetor de carregamento do componente principal a_j pode ter seus elementos multiplicados por -1, visto que, os sinais dos carregamentos se comportam de forma arbitrária. Esses carregamentos são considerados pesos que identificam combinações lineares das pontuações que se comportam de forma semelhante conforme definido na matriz de similaridade dos pares superiores (32).

  Geometricamente, os componentes principais criam um sistema de coordenadas ortogonais, onde as cargas são projetadas. Com isso, se todos os autovetores forem retidos, os dados originais podem ser recuperados exatamente, embora possam ser padronizados para certas análises. Se nem todos forem retidos, os dados originais e sua similaridade total podem ser recuperados apenas aproximadamente, mas o modelo de componente principal produz a maneira mais eficiente em que os dados podem ser expressos em um número menor de dimensões (32).

2.3.2.2 Agrupamento dos dados

O agrupamento de dados, também conhecido como *clustering* vem sendo usado com frequência em muitas áreas para realizar tarefas de exploração de dados e extração de padrões (71) e (90). Pode ser definida como um conjunto de técnicas computacionais com o objetivo de separar objetos em grupos, também denominados clusters (3) e (89).

Na literatura é possível encontrar diversas definições de cluster, onde (35) define um cluster como "*um conjunto de entidades semelhantes, e entidades pertencentes a clusters diferentes não são semelhantes*", ou "um agrupamento de pontos no espaço tal que a distância entre quaisquer dois pontos no cluster é menor que a distância entre qualquer

ponto no cluster e qualquer ponto fora deste" e por último "como regiões conectadas de um espaço multidimensional contendo uma densidade de pontos relativamente alta, separada de outras tais regiões por uma região contendo uma densidade relativamente baixa de pontos". O processo de clusterização consiste em uma técnica considerada de exploração de dados que se baseia no conceito de similaridade, isto é, possibilitar agrupar itens semelhantes de acordo com seus atributos e características, de modo a facilitar seu processamento posterior. Em outras palavras, são diversos métodos matemáticos com a finalidade de classificar observações dentro de um espaço p-dimensional, de forma que, essas observações sejam divididas em um número de grupos convenientemente, de modo que seja possível notar similaridades presentes em cada grupo (30).

De acordo com (21), espera-se que o objetivo seja de maximizar a homogeneidade em um determinado cluster e maximizar a heterogeneidade entre os demais clusters, ou seja, os elementos de um determinado cluster sejam bem parecidos entre si e dissimilares entre os elementos dos outros clusters. Os resultados obtidos na clusterização são conjuntos de agrupamentos de dados denominados clusters e são dependentes dos parâmetros utilizados, como, por exemplo, medidas de similaridade e métodos de agrupamento (71).

Para determinar a similaridade entre os dados, é necessário efetuar um cálculo de distâncias. Atualmente, existem diversos métodos para realizar esse cálculo, e, saber qual método utilizar, dependerá do tipo da aplicação e a medida em que forem gerados os dados. Neste trabalho, utilizamos algoritmos particionais. Os algoritmos de agrupamentos particionais dividem o conjunto de dados em um número determinado de cluster uma única vez (21). Esses algoritmos que geram partições de uma única vez se destacam por apresentarem vantagens nas aplicações em que a quantidade de dados é muito grande. Assim, nestes casos, a construção de um dendrograma é computacionalmente proibitiva, em outras palavras, quando o uso da clusterização hierárquica torna-se muito custosa para esses casos (86) e (55).

- **Algoritmo K-means**

O algoritmo *K-means* proposto por (63) está entre os algoritmos por particionamento mais famosos utilizados atualmente e será o algoritmo de agrupamento particionais utilizado neste trabalho. É baseado no conceito de similaridade entre dados, ou seja, a ideia principal é encontrar itens semelhantes de acordo com seus atributos através do cálculo de distâncias. Pode ser definido como um algoritmo de agrupamento de dados não-hierárquico que utiliza uma técnica iterativa para particionar um conjunto de dados (89). No funcionamento do os dados são divididos em K clusters, e, para usá-lo, é necessário pré-especificar o número de K (86). Os clusters são representados por uma informação, conhecida como centro do cluster ou centroides. Encontrar o número ideal de clusters para determinado conjunto de dados muitas vezes é um processo de tentativa e

erro, e pode ser considerado um processo subjetivo de constituir um agrupamento correto (90).

O K-means agrupa os pontos de dados de entrada em vários grupos com base na distância uns dos outros. O algoritmo assume que os recursos de dados formam um espaço vetorial e encontra um agrupamento neles (111). Os pontos estão agrupados em torno dos centroides $\mu_i \forall i = 1 \ldots k$ que são obtidos minimizando o objetivo através da Equação 2.11.

$$V = \sum_{i=1}^{k} \sum_{x_j \in S_i} (x_j - \mu_i)^2 \tag{2.11}$$

onde há k clusters $S_{i_1}, i = 1, 2, \ldots, k$ e μ_i é o centroide ou ponto médio de todos os pontos $x_j \in S_i$.

De maneira simplificada, o algoritmo inicialmente escolhe k centros de clusters para coincidir com k padrões escolhidos aleatoriamente ou k pontos definidos aleatoriamente no hiper volume que contém o conjunto de padrões. Em seguida, é atribuído cada padrão ao centro do cluster mais próximo e com isso, recalcula os centros de cluster usando as associações de clusters atuais (47). As etapas são repetidas até que o ponto central do cluster permaneça inalterado ou atinja o número definido de iterações. Os resultados do algoritmo mudam com a escolha do ponto central, resultando em uma instabilidade dos resultados (127).

Um grande desafio do processo de agrupamento particionais é encontrar um valor para k. Neste trabalho, foi usado o método *elbow* que será apresentado a seguir.

- **Método Elbow**

O método elbow, também conhecido como método do cotovelo, é um método visual que analisa a porcentagem de variância explicada em função do número de clusters (110) e sua lógica principal está apresentada através do Algoritmo 1.

A intuição é que aumentar o número de clusters irá naturalmente melhorar o ajuste (ou seja, vai explicar mais da variação), visto que existem mais clusters para usar, mas que em algum ponto o ajuste será muito grande e o cotovelo reflete isso (122). Esse método baseia-se na ideia de que é necessário escolher um número de clusters de forma que a adição de outro cluster não fornece uma modelagem muito melhor nos dados (13). Os primeiros clusters adicionam muitas informações, mas em algum ponto o ganho marginal diminuirá drasticamente e fornecerá um ângulo no gráfico. O k correto, ou seja, o número

de clusters é escolhido neste ponto, portanto, o "critério de cotovelo" (13).

Algoritmo 1: Método Elbow para determinar o K no k-means

```
início
    k = 1;
    repita
        calcule o custo da solução de qualidade ideal;
        if custo cair drasticamente then
            k = k ideal;
        else
            k = k+1;
        end
    até encontrar o k ideal;
fim
```

A ideia central é configurar a função de custo calculada através da Equação 2.12.

$$J = \sum_{i=1}^{k} \sum_{x \in C_i} |x - C_i|^2 \tag{2.12}$$

onde J é a função de custo, x é o elemento do cluster C_i e k é o número de clusters $|C_i|$. Com o aumento do número k de agrupamento, a partição da amostra será mais refinada. O grau de cada cluster aumentará gradualmente, de modo que, a função de custo J se tornará naturalmente menor. Quando k for menor que o verdadeiro número do agrupamento, a queda de J será grande. No entanto, quando k atinge seu verdadeiro valor, a queda do custo será relativamente grande e para cada aumento de k, seu custo será reduzido de forma que, se tornará quase plana. Ou seja, o gráfico de relação de J e k tem a forma de um cotovelo, e o valor k correspondente desse cotovelo é visto como o número real de grupos dos dados (60). Em outras palavras, a ideia do método é de que ele inicie com um $k = 2$ e vá aumentando a cada etapa o valor de k em 1 e, com isso, calculando seus clusters e o custo respectivo que vem com o treinamento do k naquele momento. Em algum valor de k o custo cai drasticamente e depois disso, nota-se que conforme o valor de k aumenta, o custo não tem grandes alterações. Nesse momento, ao ilustrar graficamente o funcionamento desse método é possível identificar uma curva. Quando isso acontecer, o k ideal é aquele em que houve a queda drástica do custo.

Vale ressaltar que um dos poucos problemas existentes no método é que nem sempre o "cotovelo" pode ser identificado claramente, ou seja, pode não existir um cotovelo ou então pode existir mais de um cotovelo. Ainda assim, é o método mais antigo e tradicional usado para determinar o verdadeiro número de clusters em um conjunto de dados (52).

2.3.3 Análise de componentes principais (PCA)

A análise de componentes principais (PCA) é um procedimento matemático para transformar um número de variáveis correlacionadas em um número menor de variáveis (94). É muito utilizada em diversas áreas científicas e provavelmente é a técnica estatística multivariada mais popular (94), (124) e (101). Seu principal objetivo é extrair as informações importantes do conjunto de dados e expressar essas informações como um conjunto de novas variáveis ortogonais (1), ou seja, procura combinações de variáveis originais para construir um conjunto com novos eixos (31), também chamadas componentes principais (66).

Estes componentes principais são a transformação do espaço vetorial com a maior variabilidade dos dados originais (94). O primeiro componente principal tem a maior variação possível, ou seja, inércia e com isso este componente irá explicar a maior parte da inércia da Tabela de dados. Já o segundo componente é calculado baseado na restrição de ser ortogonal ao primeiro componente e ter a maior inércia possível. Os demais componentes são calculados da mesma forma. Os valores das novas variáveis para as observações são chamados pontuações de fatores, e as pontuações de fatores podem ser interpretadas geometricamente como as projeções das observações sobre os componentes principais (1).

Os dados da matriz original são decompostos e projetados em duas partes, a amostra é classificada na parte de pontuação e os descritores em termo de separação das amostras na parte de carregamento seguindo o eixo dos componentes principais (28). O PCA pode ser descrito detalhadamente a partir das equações a seguir.

A Equação 2.13 refere-se a matriz original de qualidade de múltiplas características.

$$X^0 = \begin{bmatrix} X_{11} & \cdots & X_{1n} \\ \vdots & \ddots & \vdots \\ X_{m1} & \cdots & X_{mn} \end{bmatrix} \tag{2.13}$$

onde m é o número de tentativas e n é o número de características de qualidade. A Equação 2.14 representa a etapa de padronização dos dados.

$$\bar{X}_{ij} = \frac{X_j - \bar{X}_j}{S_j} \tag{2.14}$$

Em que $\overline{\mathrm{X}}_j$ é o valor médio de X_j e S_j é o desvio padrão de X_j. Com isso, a matriz padrão é obtida através da Equação 2.15.

$$X = \begin{bmatrix} \bar{X}_{11} & \cdots & \bar{X}_{1n} \\ \vdots & \ddots & \vdots \\ \bar{X}_{m1} & \cdots & \bar{X}_{mn} \end{bmatrix} \tag{2.15}$$

Através da Equação 2.16 é possível obter a matriz de coeficiente de correlação representada pela Equação 2.17.

$$r_{jk} = \frac{\sum_{i-1}^{n} \left(X_{ij} - \bar{X}_j\right)\left(X_{ik} - \bar{X}_k\right)}{\sqrt{\sum_{i-1}^{n} \left(X_{ij} - \bar{X}_j\right)^2 \sum_{i-1}^{n} \left(X_{ik} - \bar{X}_k\right)^2}} \tag{2.16}$$

Em que $\tilde{X}_j$ e $\tilde{X}_k$ são valores médios de $\mathrm{X_j}$ e $\mathrm{X_k}$ respectivamente.

$$R = \begin{bmatrix} r_{11} & \cdots & r_{1n} \\ \vdots & \ddots & \vdots \\ r_{m1} & \cdots & r_{mn} \end{bmatrix} \tag{2.17}$$

A determinação de autovalores e autovetores são representados através da Equação 2.18.

$$(\mathrm{R} - \lambda_\mathrm{j})\,\mathrm{V_j} = 0 \tag{2.18}$$

onde λ_j são os autovalores $\sum_{j=1}^{n} \lambda_j = n, j = 1, 2, \ldots, m$ e $V_j = (a_{j1}, a_{j2}, \ldots, a_{jn})$ são os autovetores correspondentes aos autovalores. As variações explicadas são representadas pelas Equações 2.19 e 2.20.

$$b_j = \frac{\lambda_j}{\sum_{j=1}^{n} \lambda_j} \tag{2.19}$$

$$a_p = \frac{\sum_{j=1}^{p} \lambda_j}{\sum_{j=1}^{n} \lambda_j} \tag{2.20}$$

Em que b_j representa a taxa de contribuição do j-ésimo autovalor e a_p representa a taxa de contribuição cumulativa do p-ésimo autovalor. Os autovetores correspondem aos autovalores. Com isso, tem-se os dados finais formulados como na Equação 2.21.

$$Y_{mj} = \sum_{i=1}^{n} X_m(i) V_{ij} \tag{2.21}$$

Em que Y_{mj} é o j-ésimo dado final formulado a partir do vetor próprio $\mathrm{V_{ij}}$ de componentes principais na análise. Sendo assim, com base na sua taxa de contribuição e taxa de contribuição cumulativa para a variância total, os componentes principais são ordenados. Isso significa que, o primeiro componente principal tem a maior variância possível e cada componente principal sucessivo é determinado com a propriedade de serem ortogonais ao primeiro componente principal e a maior variância possível (28).

3 TRABALHOS RELACIONADOS

Na literatura é possível identificar diversos estudos relacionados contendo os principais assuntos abordados por este trabalho. Estes trabalhos estão divididos pelos seguintes tópicos: análises e trajetórias espaciais na seção 3.1, tecnologias de *Big data* usadas para auxílio na assistência médica na seção 3.2 e por último o aprendizado de máquina aplicado à saúde e aplicado às gestantes na seção 3.3.

3.1 ANÁLISE E TRAJETÓRIA ESPACIAL

Atualmente, diferentes métodos de análise espacial na saúde coletiva vêm sendo desenvolvidos para detecção de aglomerados espaciais ou espaço-temporais e aplicados ao planejamento e avaliação de uso de serviços de saúde (7).

O trabalho de (50) sugere que a análise espacial pode agregar na pesquisa em saúde pública de duas maneiras. Inicialmente, a análise espacial pode sugerir fatores causais na patogênese de doenças, onde a associação entre doenças e locais pode implicar que a população que vive ali possuem características inerentes que a tornam mais suscetível as doenças ou experimenta elevada exposição a um fator de risco, como a poluição do ar. Outro ponto a ser considerado é que a análise espacial pode ajudar a identificar como as populações se adaptam e se relacionam com seu ambiente.

E, tratando-se de melhorias na assistência médica, (5) afirma que as simulações vêm ganhando aceitação dentro dessa área. A proposta do autor foi modelar fluxos de processos de trabalho dos funcionários e de encaminhamento dos pacientes ou eventos no departamento de emergência da Irlanda de modo a investigar o sistema de funcionamento desse estabelecimento, sendo possível determinar eventuais melhorias.

Já (34) propõem o uso de simulações em tempo real para analisar as operações que são realizadas diariamente nas salas de emergência, onde eventos não planejados podem ocorrer. Esse modelo retrata adequadamente o estado atual do sistema, onde se tem que os sistemas de gerenciamento de fluxo de trabalho das salas de emergência geralmente fornecem informações limitadas. Sendo assim, essa proposta visa estudar o impacto do uso de simulações que preveem o desempenho da sala de emergência com uma quantidade limitada de informações, ou seja, o quanto isso influencia na tomada de decisões. Além disso, (113) propõe com o uso de modelagem de simulação detalhada em métodos de modelagem estruturados, um modelo abrangente para caracterizar a utilização do tempo dos enfermeiros em tal ambiente de carga de trabalho dinâmico flexível.

3.2 TECNOLOGIAS DE BIG DATA APLICADO À ASSISTÊNCIA MÉDICA

Nos últimos anos, *Big Data* se tornou um termo onipresente (11), que representa grandes quantidades de dados que não são gerenciáveis usando software tradicional ou plataformas baseadas na internet, onde supera a quantidade tradicionalmente usada de armazenamento, processamento e poder analítico. Sabendo-se que o *Big Data* não é gerenciável com software tradicional, existe a necessidade de aplicações e softwares tecnicamente avançados capazes de utilizar poder computacional de ponta, rápido e econômico para essas tarefas (27). Assim, tem-se a implementação de inteligência artificial (AI) e novos algoritmos de fusão para dar sentido a essa grande quantidade de dados.

O ramo da assistência médica é uma das áreas que podem se beneficiar de modo significativo com o aumento da quantidade de dados e sua disponibilidade. Existem algumas áreas do setor de saúde que podem se beneficiar do uso de análises de *Big Data*. Atualmente, a medicina de precisão é a principal área em comum entre prestadores de cuidados de saúde. Ao combinar o *Big Data* com ferramentas de aprendizado de máquina, diversos setores da saúde podem obter benefícios significativos. De acordo com (92) esses benefícios potenciais podem incluir detecção de doenças em estágios iniciais, podendo ser tratadas com mais facilidade e eficácia. Inclui também a gestão de saúde individual e populacional até a detecção de fraude de saúde de forma imediata.

3.3 APRENDIZADO DE MÁQUINA COMO AUXÍLIO À GESTANTES

Atualmente, a aplicação de técnicas de aprendizado de máquina na área da saúde têm se tornado uma área de pesquisa promissora, visto que, a área da saúde pode se beneficiar muito com as aplicações. Em uma geração onde a indústria de relógios e dispositivos inteligentes está em crescimento, existe uma facilidade em coletar dados relacionados à saúde, e com isso, aumentando a aplicabilidade do aprendizado de máquina.

No âmbito de melhorar a gestão de um hospital, (78) propõe analisar fluxos dos pacientes usando abordagens de redes complexas e técnicas de mineração de processos. Para isso, foi utilizado técnicas de mineração de processos para rastrear os caminhos percorridos por pacientes ambulatoriais entre os diferentes departamentos de um hospital, de modo a construir redes capazes de identificar padrões sazonais e com isso, projetar ações preventivas para otimizar a gestão de processos do hospital.

No trabalho de (46), utilizou-se um conjunto de dados de procedimentos médicos realizados em todo o Brasil no ano de 2014, para a construção de uma rede de conexões entre médicos e pacientes. A rede construída foi analisada de modo a verificar a importância relativa dos médicos na rede e a possibilidade de colaborações entre eles no tratamento de pacientes. Os autores concluem que a utilização das métricas propostas no trabalho foram promissoras para definir prováveis equipes de atendimento, encontrando subconjuntos de

médicos com um alto índice de pacientes comuns.

Quando focados em doenças crônicas, tem-se o trabalho de (83), que propôs um classificador capaz de predizer a presença de diabetes do tipo 1 em gestantes sem diagnóstico prévio baseado nas características de altura, massa corporal e histórico de diabetes na família e histórico de atendimento da gestante. Assim, sendo possível separar as gestantes no sistema que tenha predisposição a essa doença, permitindo um auxílio para garantir um melhor acompanhamento da gestante no âmbito dessa condição de saúde.

Já o trabalho de (20), propôs um modelo baseado em técnicas de aprendizado de máquina para a identificação automática de beneficiários com propensão a doenças crônicas. Para isso, o modelo foi aplicado sobre dados de uma operadora de plano de saúde, extraídos usando técnicas de mineração de dados, avaliados por especialistas da área médica para a classificação de beneficiários em relação ao diabetes mellitus do tipo 2. A eficácia do modelo teve como acurácia próximo dos 90%.

Além das aplicações mencionadas anteriormente, vale salientar que na última década houve um aumento significativo em pesquisas relacionadas a algoritmos de predição de susceptibilidade, recorrência e sobrevivência ao câncer. Para isso, tem-se que na maioria desses estudos são usados diferentes tipos de dados de entrada: genômicos, clínicos, histológicos, de imagem, demográficos, dados epidemiológicos ou combinação destes (6), (53), (59) e (119).

Quando o assunto é relacionado às gestantes, existem diversos trabalhos que propõe a aplicação das técnicas de aprendizado de máquina para melhorar a qualidade de vida de uma gestante e do feto durante todo o período gestacional. De modo a garantir uma melhor gestação, é necessário que exista uma identificação correta dos problemas que podem ocorrer em uma gestação, para que assim, consiga aplicar intervenções clínicas e tentar prevenir ou amenizar esses problemas (115). Esses problemas variam de soluções que rastreiam os caminhos de pacientes ambulatoriais para melhorar o gerenciamento de processos de um hospital (78) a problemas que podem ocorrer durante gestações (115).

No trabalho (22) tem-se a proposta de um algoritmo baseado em visão computacional e aprendizado de máquina que prevê a gravidez usando dois bancos de dados com imagens do teste beta de gonadotrofina coriônica humana (b-HCG) da morfologia de um embrião e da idade das gestantes. Foram avaliadas usando cinco classificadores diferentes: probabilística bayesiana, máquinas de vetores de suporte (SVM), rede neural profunda, árvore de decisão e random forest (RF), usando uma validação cruzada k-fold para avaliar as capacidades de generalização do modelo. Como resultado, tem-se que para a base de dados A, o classificador SVM obteve uma pontuação F1 de 0.74 e AUC de 0.77 e para a base de dados B, o classificador RF obteve uma pontuação F1 de 0.71 e AUC de 0.75. Sendo assim, os resultados sugerem que o sistema consegue prever um teste de gravidez positivo a partir de uma única micrografia.

Outra aplicação existente das técnicas de aprendizado de máquina em gestantes foi o trabalho (14) que propõe associações entre ingestão relativa à ingestão total de energia de frutas, vegetais e resultados adversos da gravidez usando estimativa de máxima verossimilhança (TMLE) emparelhada com o algoritmo de aprendizado de máquina de conjunto Super Learner comparados com os resultados gerados por regressão logística multivariável. Foi analisando a ingestão periconcepcional diária usual de frutas e vegetais totais estimadas a partir de um questionário de frequência alimentar (FFQ). Além disso, foi calculado o risco marginal de nascimento prematuro, nascimento pequeno para a idade gestacional, diabetes gestacional e pré-eclâmpsia conforme a densidade de frutas e vegetais (xícaras/1000 kcal) $\geq$ percentil 80 em comparação com $< 80^{o}$ percentil usando regressão logística multivariável e Super Learner com TMLE. O resultado apresentado mostrou que a sinergia alimentar contabilizada no aprendizado de máquina pode desempenhar um papel importante nos resultados da gravidez.

Já o trabalho (43) propõe um algoritmo de seleção de recursos de *hill climbing* com as técnicas de perceptron multicamadas (MLP), máquinas de vetores de suporte (SVM), árvores de decisão e regressão (CART) e florestas aleatórias (RF) para analisar e prever o sucesso da gravidez de fertilização in vitro. Neste trabalho, a técnica de *hill climbing* foi empregada para avaliar a influência de cada atributo do procedimento de tratamento de fertilização in vitro e selecionar os atributos mais influentes para cada classificador para aumentar a previsibilidade do sucesso da gravidez. Os resultados mostram que a abordagem proposta combinando o método de seleção de características de hill climbing com classificadores como SVM, MLP ou RF rendeu desempenho de predição melhor que os existentes na literatura.

Este capítulo apresentou os trabalhos que, em sua síntese, solucionam problemas relacionados à saúde com uso de tecnologias. Na literatura, a lacuna é que não encontramos pesquisas que classifiquem gestantes de risco, o que diferencia nosso trabalho, visto que classificamos as gestantes usando duas abordagens, usando árvores de decisão e algoritmos de agrupamento.

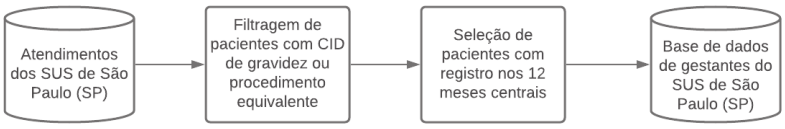
4 CONJUNTO DE DADOS

Neste capítulo, são apresentadas informações fundamentais para o entendimento do conjunto de dados que foi usado nessa pesquisa. Primeiramente, na seção 4.1, será feito uma descrição completa dos dados e metodologia usada para limpeza e filtragem de dados e população estudada e na seção 4.2, será feito uma caracterização inicial desses dados, bem como as médias de procedimentos, os procedimentos mais frequentes, entre outras características.

4.1 DESCRIÇÃO DOS DADOS

Os dados utilizados neste trabalho são referentes a registros de atendimento de uma base de dados da secretaria municipal de saúde da cidade de São Paulo (SP), no período de dois anos, compreendidos entre os meses de janeiro de 2014 e dezembro de 2015[1]. Esses registros estão divididos em atendimentos com procedimentos usando os códigos do Sistema de Gerenciamento da Tabela de Procedimentos, Medicamentos e OPM do SUS (SIGTAP) e atendimentos com diagnósticos, usando os identificadores de Classificação Estatística Internacional de Doenças e Problemas Relacionados com a Saúde (CID-10). A fim de avaliar a trajetória de um paciente no sistema de saúde, optou-se por analisar dados referentes a gestantes uma vez que se pode avaliar a sequência completa de procedimentos e realizar uma comparação com o modelo recomendado pelo SUS, além de ser um estudo de caso de alta relevância do ponto de vista social.

Figura 3 - Diagrama das etapas metodológicas adotadas.



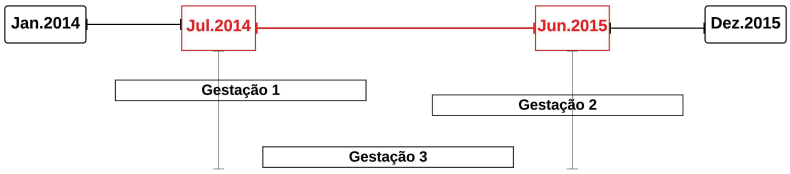
Fonte: Elaborado pela autora (2021).

A Figura 3 apresenta uma descrição do método adotado para geração da base de dados de gestantes da cidade de São Paulo, onde inicialmente tinha-se uma base com todos os registros de atendimentos e, com isso, foi feita uma filtragem com 87 CIDs exclusivos de gestantes ou procedimentos equivalentes, gerando uma base de gestantes com um total de 481.057 registros. Os CIDs selecionados estão listados no Apêndice A. A fim de obter somente registros de gestantes com gestação completa durante o período disponível, ou seja, cuja gestação teve início e fim dentro dos registros selecionados, optou-se por uma

[1] Utilização dos dados aprovada pelo Comitê de Ética e Pesquisa (CAAE: 51038515.6.3001.0086).

filtragem de tempo, como mostrado na Figura 4. Nesse método, os 24 meses da base de dados foram separados em 3 partes: os 6 meses iniciais (de jan/2014 a jun/2014); os 12 meses centrais (de jul/2014 a jun/2015); e os últimos 6 meses (de jul/2015 a dez/2015). Assim, se uma gestante teve atendimento nos 6 primeiros meses, mas não nos 12 meses centrais, então provavelmente a sua gestação se iniciou em 2013 e terminou no início de 2014, de modo que esta paciente não se enquadra neste estudo. De forma semelhante, se a paciente teve atendimento nos últimos 6 meses, mas não nos 12 meses centrais, então provavelmente sua gestação se iniciou no fim de 2015 e terminou em 2016, o que também não é proveitoso para esta pesquisa.

Figura 4 - Filtragem temporal para seleção de gestantes da base de dados de modo a obter somente gestações completas registradas no período total da base de dados (2014-2015).



Fonte: SILVA, Mayara *et al.* (104) (2020).

Por outro lado, casos como os ilustrados por “Gestação 1”, “Gestação 2” e “Gestação 3” na Figura 4 são alvo deste estudo. A “Gestação 1” corresponde aos atendimentos de gestantes cuja gestação iniciou nos 6 primeiros meses e terminaram na parte central da base. Já “Gestação 2” representa os registros de gestantes que começaram sua gestação nos meses centrais da base e cuja gestação terminou nos 6 últimos meses. Por sua vez, “Gestação 3” contempla todas as gestações cujos registros estão na parte central da base, ou seja, entre os meses de julho de 2014 e junho de 2015.

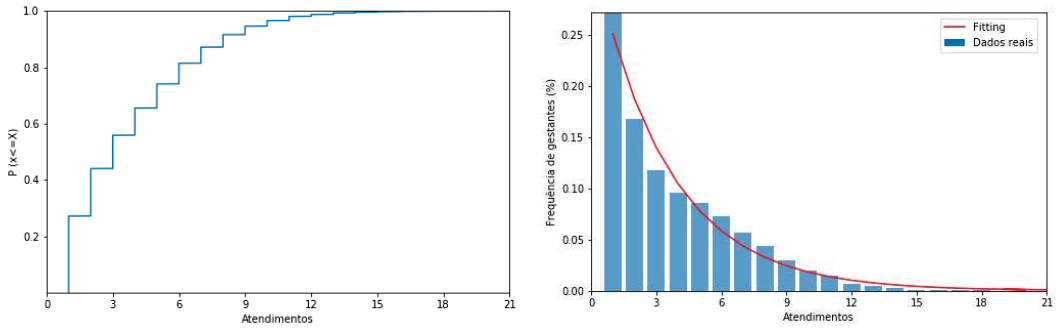
Após identificar as gestantes com ao menos um atendimento durante os 12 meses centrais, foram selecionados os registros de atendimentos com diagnósticos referentes a elas nos demais períodos, criando a base de dados de gestantes com gestação completa. Com isso, tem-se a base final usada neste estudo, referente a 24.916 gestantes.

4.2 CARACTERIZAÇÃO INICIAL DOS DADOS

Inicialmente, fizemos uma caracterização desses dados, onde foi contabilizado o número de atendimentos para cada uma das gestantes. Para ilustrar essa caracterização tem-se na Figura 5(a) a função de distribuição acumulada (CDF) do número de atendimentos por gestante, e na Figura 5 (b) a distribuição dos atendimentos realizados por gestante. Ao observar os resultados e por se tratar de um caso discreto, a distribuição geométrica se

mostrou mais eficiente, por comportar melhor os dados amostrais. Sendo assim, a Figura 3 (b) têm características dessa distribuição e, visualmente inspecionando os resultados, percebe-se que as amostras se comportam conforme a distribuição geométrica de equação $P(X = n) = (1 - p)^{n-1}p$, onde $p \approx 0{,}8$.

Figura 5 - Distribuições dos atendimentos por gestante, onde apresenta-se (a) uma função de distribuição cumulativa (CDF) e (b) uma distribuição geométrica com fitting dos dados sobre o histograma de dados reais.



(a) CDF dos atendimentos (b) Distribuição geométrica dos atendimentos

Fonte: SILVA, Mayara *et al.* (104) (2020).

De forma geral, com base nos resultados da Figura 5, as gestantes avaliadas tiveram um baixo número de atendimentos junto ao sistema de saúde. Ressalta-se que 8 atendimentos é o número indicado pelo Manual Técnico do Pré-natal e Puerpério, proposto pela Secretaria de Estado da Saúde de São Paulo (SES-SP) (9). Observa-se que 20% das gestantes identificadas fizeram pelos menos 6 atendimentos. Nesse conjunto de dados, cerca de 27% das gestantes fizeram apenas um atendimento e, na média, foram realizados 3.87 atendimentos por gestante, sendo 30 o número máximo de atendimentos feitos por uma gestante.

Alguns dos CIDs registrados na base contêm apenas a categoria do diagnóstico (CID de três dígitos), enquanto outros contêm a subcategoria (quarto dígito). A fim de comparar a frequência dos CIDs, optou-se por considerar somente a categoria dos registros. Por exemplo, se um atendimento continha o CID Z349 (“Supervisão De Gravidez Normal, Não Especificada”), neste primeiro momento considerou-se apenas que ele está ligado à categoria de mais alto nível Z34 (“Supervisão de gravidez normal”).

A Tabela 1 apresenta a distribuição dos diagnósticos por categoria de CID mais comuns no conjunto de dados. Essa Tabela mostra os cinco diagnósticos mais comuns, uma vez que eles já são responsáveis por aproximadamente 96% dos atendimentos encontrados no conjunto de dados.

Podemos observar na Tabela 1 que os CIDs de supervisão de gravidez representam

Tabela 1 – Diagnósticos mais frequentes por categoria de CID.

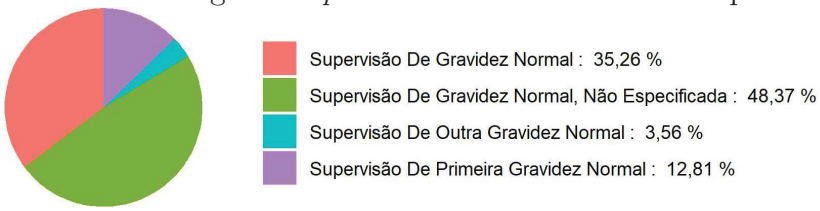
Descrição do CID	Frequência absoluta	Frequência relativa
Supervisão de gravidez normal	66.904	69,31%
Supervisão de gravidez de alto risco	13.315	13,79%
Exame ou teste de gravidez	10.783	11,17%
Gravidez como achado casual	1.142	1,18%
Gravidez ectópica	718	0,74%

Fonte: SILVA, Mayara *et al.* (104) (2020).

a maior parte dos diagnósticos, sendo que a supervisão de gravidez normal está relacionada a mais de dois terços dos atendimentos. O número de registros de supervisão de gestação de alto risco também é notável, mas deve-se destacar que a porcentagem de atendimentos não é necessariamente proporcional à porcentagem de pacientes com gestação de alto risco. Isso se dá por dois motivos: Por um lado, espera-se que o número de atendimentos por gestante seja maior nesta classe dada a necessidade de acompanhamento mais frequente da gestação; por outro lado, este não é o único CID voltado para gestações de risco (o quinto CID mais comum, "Gravidez Ectópica" é um exemplo disso). O terceiro diagnóstico mais comum relacionado é, na verdade, o procedimento de teste de gravidez. Diferentemente dos CIDs de supervisão de gravidez, espera-se que cada paciente realize no máximo um teste de gravidez e, por isso, já era esperada uma participação menor deste CID. Percebe-se, no entanto, que a frequência absoluta do teste de gravidez é inferior à metade do número de gestantes. Uma hipótese que justificaria este cenário é a de que o teste de gravidez, por ser um procedimento bem simples e de baixo custo, pode ter sido realizado pela própria gestante, sendo subestimado nessa análise.

Dada a predominância de atendimentos com as duas categorias de CIDs de supervisão de gravidez (normal e de alto risco), buscou-se identificar na base as subcategorias mais comuns em cada caso. Para a categoria relacionada à gravidez normal, tem-se quatro CIDs, distribuídos como apresentado na Figura 6. Observa-se que em mais de 80% dos casos não há mais detalhes sobre a gestação. Algo semelhante ocorre com a categoria de supervisão de gestação de alto risco, em que 90,7% dos registros não possuem o identificador de subcategoria e 6.0% constam como supervisão não especificada.

Figura 6 - CIDs da categoria *Supervisão de Gravidez Normal* presentes na base.



Fonte: SILVA, Mayara *et al.* (104) (2020).

A Tabela 2 apresenta os diagnósticos mais comuns iniciando ou finalizando o conjunto de dados avaliado, *i.e.*, mais frequentes como primeiro ou último registro de uma paciente, respectivamente. Observa-se que o diagnóstico "Supervisão de Gravidez Normal" é o mais comum, tanto como primeiro quanto como último registro. O "Exame ou teste de gravidez" é o segundo procedimento mais comum iniciando o registro de uma gestante. Como explicado anteriormente, é um resultado coerente, pois possivelmente muitas gestantes iniciam o acompanhamento pré-natal a partir de um resultado positivo do teste realizado particularmente. Em 6.39% das vezes, o teste de gravidez foi o último registro e uma hipótese que justifica isso é a de que essas pacientes podem ter recebido um resultado negativo para o teste de gravidez, encerrando assim o seu acompanhamento. A "Supervisão de gravidez de alto risco" é a terceira categoria mais comum como primeiro registro, mas passa a ser a segunda mais frequente como categoria final da trajetória da paciente, incrementando a participação de 8,51% para 13,69%. Isso se justifica não só pela participação menor do teste de gravidez como último registro, mas também pelo fato de que muitas vezes a gestação se inicia sem problemas, mas uma condição de risco se desenvolve ao longo das semanas.

Tabela 2 – Diagnósticos mais frequentes como primeiro e último registro.

Categorias de CID mais frequentes como primeiro registro	**Freq. Relativa**	**Categorias de CID mais frequentes como último registro**	**Freq. Relativa**
Supervisão de gravidez normal	55,73%	Supervisão de gravidez normal	70,66%
Exame ou teste de gravidez	27,18%	Supervisão de gravidez de alto risco	13,69%
Supervisão de gravidez de alto risco	8,51%	Exame ou teste de gravidez	6,39%
Gravidez ectópica	1,84%	Gravidez ectópica	1,83%
Gravidez como achado casual	1,49%	Gravidez como achado casual	1,81%

Fonte: SILVA, Mayara *et al.* (104) (2020).

Considerando os procedimentos de "Exame ou teste de gravide", "Supervisão de gravidez normal" e "Supervisão de gravidez de alto risco", é possível estabelecer ainda quais foram as especialidades médicas que mais frequentemente realizaram esses registros de atendimento. A Tabela 3 apresenta essas informações. Como poderia se esperar, por ser um procedimento simples, os exames ou testes de gravidez são realizados em sua maioria por um auxiliar de enfermagem. Em outro extremo, a supervisão de uma gravidez de alto risco é acompanhada por um médico especialista em fração significativamente superior à supervisão de uma gravidez considerada normal.

4.3 TRAJETÓRIA DE ATENDIMENTOS DAS GESTANTES

De acordo com Programa de Humanização no Pré-natal e Nascimento, disponibilizado pelo Ministério da Saúde (74), o fluxo teórico ideal que deve ser seguido por uma

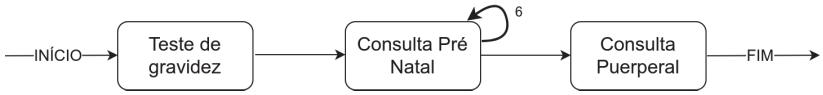
Tabela 3 – Especialidades que mais registraram os respectivos CIDs.

CID Z32	**Freq.**	**CID Z34**	**Freq.**	**CID Z35**	**Freq.**
Aux. de enfermagem	41,39%	Ginecologista e Obstetra	72,43%	Ginecologista e Obstetra	89,68%
Aux. de enfermagem da ESF	28,40%	Médico da ESF	15,81%	Médico da ESF	3,46%
Enfermeiro	12,09%	Radiologia e DI	9,78%	Radiologia e DI	3,13%
Ginecologista e Obstetra	7,92%	Médico Clínico	0,90%	Nutricionista	1,51%

Fonte: SILVA, Mayara *et al.* (104) (2020).

gestante apresenta seis atendimentos pré-natal seguido por um atendimento pós-parto, como ilustra a Figura 7.

Figura 7 - Fluxo teórico de atendimento à gestante recomendado pelo SUS.



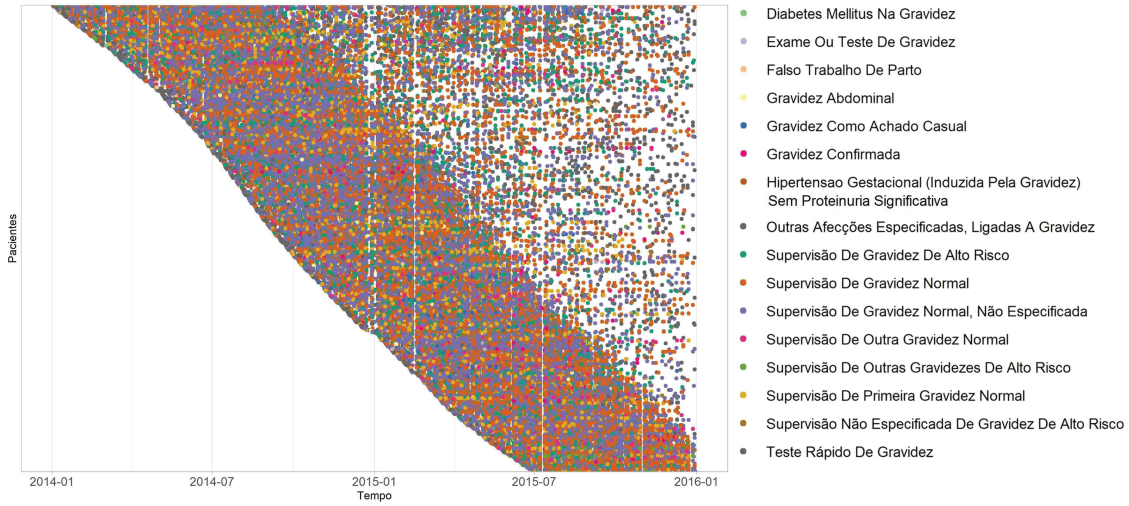
Fonte: SILVA, Mayara *et al.* (104) (2020).

Para uma análise visual da trajetória de atendimentos foi utilizada a representação através de um mapa de processos. Um mapa de processos é um grafo ponderado e direcionado cujos nós são as atividades existentes e as arestas são criadas quando um mesmo caso realiza uma atividade (nó de origem da aresta) e, em seu próximo registro, realiza outra (nó de destino da aresta). Quanto mais casos repetirem a mesma aresta, maior será o peso dessa aresta. Por exemplo, quando uma gestante realiza um procedimento "x" no tempo t, é criado um nó no mapa de processos. Quando essa gestante realiza um procedimento "y" em $t + 1$, também é criado um nó "y" no mapa. Como "y" logicamente sucede "x", é criada uma aresta direcionada "x" $\rightarrow$ "y". A informação temporal dessa aresta é capturada pelo mapa de processos.

Assim, para as gestantes com mais de um atendimento, foi criado um mapa de processos a partir dos dados do conjunto de dados do período completo (2014-2015). A Figura 8 apresenta uma visualização da sequência de atendimentos a cada gestante identificada na base. Cada linha nessa Figura representa a trajetória (sequência) de atendimentos de uma gestante. A primeira gestante encontrada no conjunto de dados é a linha apresentada mais acima, no eixo Y do gráfico. Os círculos apresentados em cada linha representam os atendimentos realizados pela gestante na data correspondente à sua coordenada. Esses círculos estão coloridos de acordo com o código CID do procedimento correspondente. Os 16 tipos de atendimentos apresentados na Figura 8 são responsáveis por 98% das atividades das gestantes na base de dados avaliada.

Nota-se que, visualmente, o conjunto de pontos (atendimentos) na Figura 8 tem um adensamento de atividades em um período de aproximadamente nove meses. Porém,

Figura 8 - Gráfico de pontos temporais dos atendimentos a gestantes. No eixo horizontal tem-se uma linha do tempo e cada "linha" do eixo vertical representa a sequência de atendimentos a uma gestante, sendo coloridos de acordo com o CID correspondente. Os dados foram filtrados para conter apenas gestantes com ao menos dois atendimentos e inclui 98% das atividades mais frequentes.



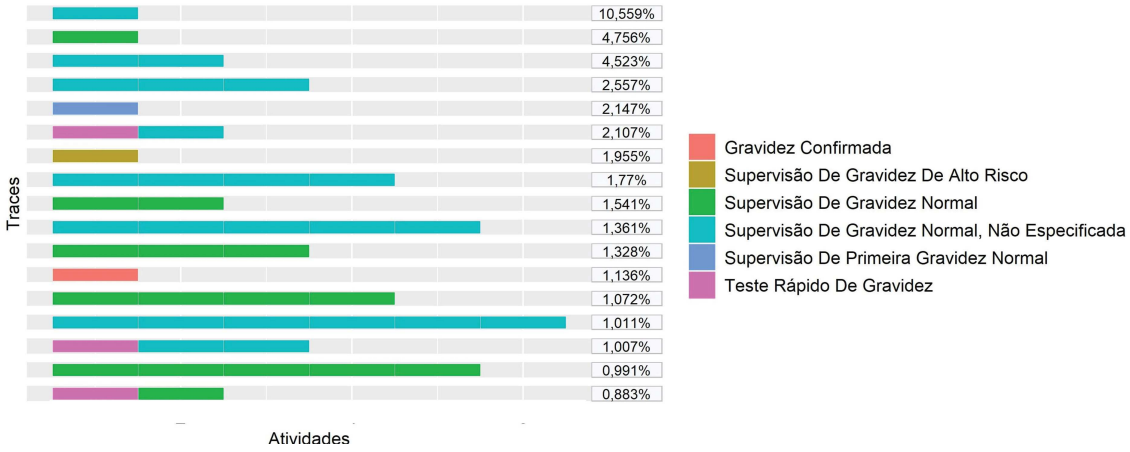
Fonte: SILVA, Mayara *et al.* (104) (2020).

em mais detalhes, nota-se que o tempo médio entre o primeiro e o último evento de uma gestante é de 23,5 semanas, com desvio-padrão de 16.33 semanas. Em outras palavras, pelo conjunto de dados analisados, o acompanhamento de uma gestante também é inferior ao desejado, assim como o número médio de atendimentos.

A Figura 9 apresenta as sequências mais comuns de atendimentos, encontradas entre as 4.507 sequências de atividades diferentes para todas as 24.916 gestantes. Metade dessas sequências mais comuns têm apenas uma atividade. Inclusive, a sequência mais comum entre todas têm apenas um registro, o de "Supervisão de Gravidez Normal, Não Especificada". Assim, essa Figura corrobora com os dados apresentados anteriormente na Figura 6. Como apontado anteriormente, também, "Supervisão de Gravidez Normal, Não Especificada" é um dos atendimentos mais comuns.

O grafo representado na Figura 10 foi gerado considerando apenas gestantes com ao menos dois atendimentos e atividades responsáveis por 95% dos atendimentos. O número de atendimentos com cada CID é apresentado no respectivo vértice. Os vértices "Início" e "Fim" são virtuais para representar o início e fim de uma sequencia de CIDs, respectivamente. Dado um par de vértices i e j, cada aresta direcionada indica a fração de atendimentos do vértice i que tiveram como atendimento subsequente o vértice j. Essa Figura apresenta o fluxo de atendimento geral para as gestantes entre as categorias de CIDs associadas aos atendimentos do sistema de saúde. Para tornar a Figura de mais fácil visualização, foram eliminados os registros de gestantes com apenas um atendimento e

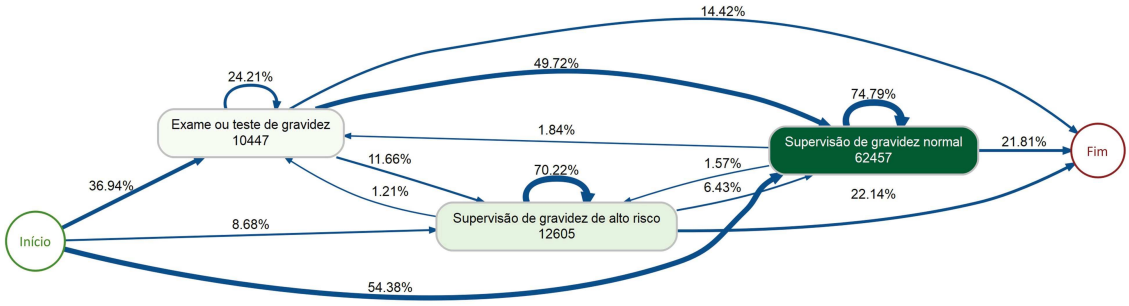
Figura 9 - Trajetórias de atendimentos mais frequentes pelas gestantes.



Fonte: SILVA, Mayara *et al.* (104) (2020).

selecionou-se apenas os CIDs mais frequentes, que são responsáveis por 95% dos registros. A porcentagem nas arestas corresponde à parcela de registros que saem do vértice de origem e o número nos vértices corresponde ao número de atendimentos com aquele CID. Deste modo, tem-se a probabilidade do próximo estado da gestante. Por exemplo, 49,7% das gestantes que realizaram "Exame ou teste de gravidez" tem como próximo registro "Supervisão de gravidez normal".

Figura 10 - Trajetórias das gestantes pelas categorias de CIDs, onde têm-se os vértices que representam as categorias de CIDs mais frequentes, vértices de início e fim que representam o começo e fim de cada sequência e a porcentagem nas arestas que corresponde a parcela de registros que saem do vértice de origem ate o vértice de destino.



Fonte: SILVA, Mayara *et al.* (104) (2020).

4.4 TRAJETÓRIA ESPACIAL DAS GESTANTES

No município de São Paulo, que contava com uma população de 11,25 milhões de habitantes em 2010 (45), atuam cinco Coordenadorias de Regiões de Saúde (CRS)(98):

Centro-Oeste, Leste, Norte, Sudeste e Sul. Essas coordenadorias são responsáveis, além pela coordenação, articulação, organização do sistema de saúde loco-regional e compatibilização dos planos, programas e projetos dos Departamentos Regionais de Saúde (DRS) em função das políticas e diretrizes da SES/SP e dos recursos disponíveis (24). A distribuição de renda entre essas regiões não é igualitária, como pode-se observar na Tabela 4. Nessa Tabela, apresenta-se o rendimento médio mensal de uma pessoa, segundo a CRS do Município de São Paulo em 2010 (99), e com isso, é possível ver que a região Centro-Oeste é a que apresenta maior renda média, cujo valor é quase quatro vezes maior do que o referente à região Leste, a de menor renda média.

Tabela 4 – Rendimento nominal médio mensal de pessoas com 10 anos ou mais de idade segundo a CRS do Município de São Paulo em 2010.

	CRS Centro-Oeste	CRS Sudeste	CRS Sul	CRS Norte	CRS Leste
Rendimento	R$ 3.846,24	R$ 2.323,73	R$ 1.568,30	R$ 1.550,09	R$ 1.054,47

Fonte: SILVA, Mayara *et al.* (104) (2020).

A Tabela 5 apresenta o número de gestantes com um único atendimento e o número total de gestantes, para cada uma das regiões de São Paulo. A região Centro-Oeste é a que tem o maior percentual de gestantes com um único atendimento, com praticamente 47% delas nessa situação. Nas demais regiões, esse número é igual ou inferior a 26%. Não se pode afirmar com certeza, mas, conjectura-se que na região Centro-Oeste, por ter uma média de renda mais elevada, como apresentado anteriormente, as gestantes procuram um serviço médico suplementar ao SUS na rede privada, o que justifica o grande número de gestantes com baixo número de atendimentos nessa região.

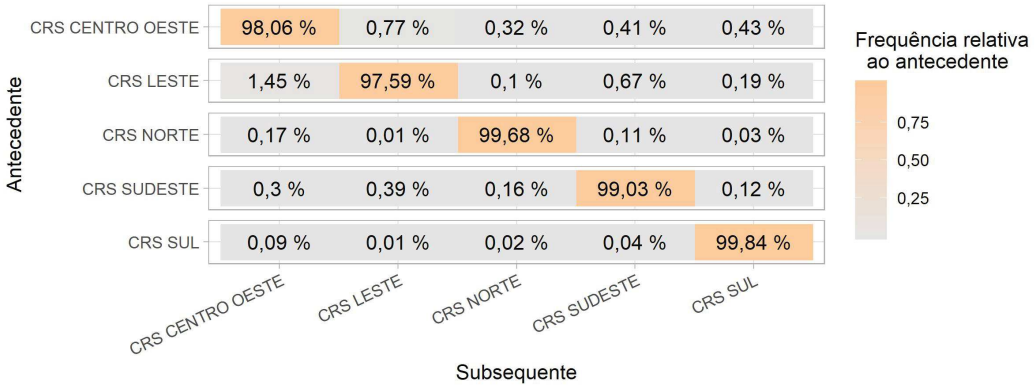
Tabela 5 – Região dos atendimentos de gestantes com uma consulta.

Regiões	Gestantes com apenas um registro	Total de gestantes	Porcentagem
Sul	3.330	12.839	25,93%
Centro Oeste	1.672	3.566	46,88%
Sudeste	696	2.889	24,09%
Norte	565	3.238	17,44%
Leste	516	2.384	21,64%

Fonte: SILVA, Mayara *et al.* (104) (2020).

Também foram avaliadas as trajetórias das gestantes entre as unidades de atendimento do sistema de saúde. Sabe-se que durante o período gestacional, a gestante pode ser atendida em diversas unidades de atendimento. A Figura 11 mostra uma matriz de precedência das regiões dos atendimentos. Neste trabalho, focamos nas regiões de saúde em que o município de São Paulo é dividido, mas tal abordagem pode ser expandida para traçar a trajetória entre as unidades de atendimento do sistema de saúde.

Figura 11 - Matriz de precedência das regiões dos atendimentos.



Fonte: SILVA, Mayara *et al.* (104) (2020).

Ao analisar a Figura 11, vê-se que a probabilidade do próximo atendimento ser na mesma região é alta para todas as zonas. Destaca-se que, neste caso, não se considera a probabilidade de um atendimento ser o primeiro ou o último. Por exemplo, dado que uma gestante foi atendida na região Sul e que ela fará um novo atendimento, a probabilidade de que esse atendimento seja feito também na região Sul é de 99,84%.

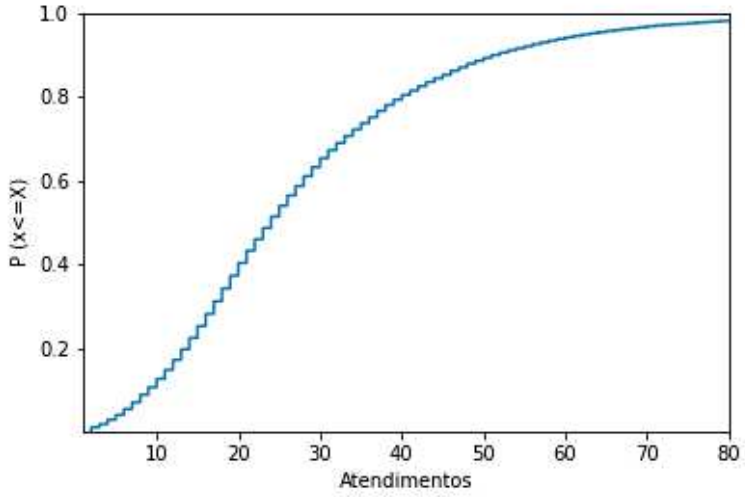
Para fazer uma análise mais detalhada das gestantes, seria necessário obter um histórico mais completo, com todas as consultas associadas a essas gestantes, para que assim, fosse possível associar uma justificativa para a gestação ser classificada com de risco, uma vez que, com o histórico da gestante, seria possível observar procedimentos antes da gestação que poderiam ter ocasionado determinada complicação.

Sendo assim, foram associados às 24.916 gestantes inicialmente identificadas todos os registros da base de dados. Após essa associação de registros, a base de dados tinha um total de 1.083.106 registros. A fim de identificar registros duplicados, foi feito uma verificação de todos os registros da base de dados, e com isso, foi possível identificar registros de atendimentos com rótulos que não iriam agregar nenhuma informação a mais ao nosso estudo. Esses registros estavam rotulados como *DESCONHECIDO*, *NÃO SE APLICA* e *NÃO PREENCHIDO*, com um total de 380.445 registros. Sendo assim, a base utilizada neste estudo possui 702.661 registros de atendimentos.

Para ilustrar essa nova caracterização, temos na Figura 12 a função de distribuição acumulada (CDF) do número de todos atendimentos por gestante. Como mencionado anteriormente, inicialmente tinha-se caracterizado que 27% das gestantes efetuaram apenas um atendimento e a média das consultas era 3.87 consultas por gestante e o número máximo de atendimentos feitos foi 30. Já nesse conjunto de dados mais completo, nenhuma gestante fez somente uma consulta e aproximadamente 1% das gestantes fizeram somente duas consultas. A média de atendimentos foi 28.20 consultas por gestante e o número máximo

de consultas foi 269.

Figura 12 - Função de distribuição cumulativa (CDF) de todos os registros vinculados por gestante.



Fonte: Elaborado pela autora (2021).

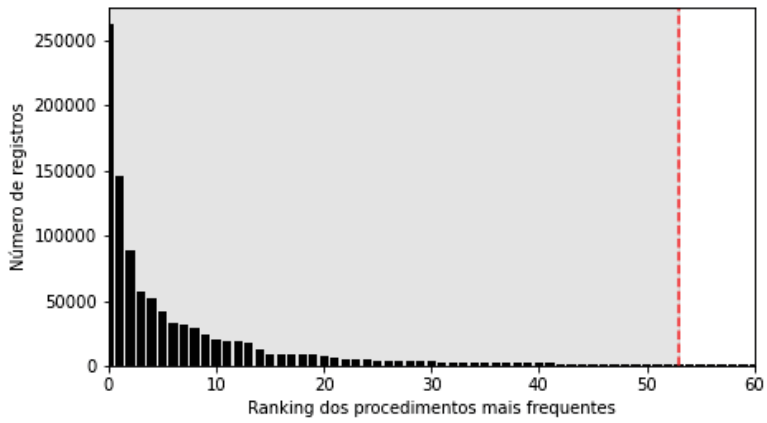
5 CLASSIFICAÇÃO DE GESTANTES

Neste capítulo, temos a aplicação dos métodos para classificação das gestantes em dois grupos, um grupo de gestações de risco e um grupo de gestações normais. Na seção 5.1 apresentamos as filtragens feitas na base de dados para obtermos as *features* usadas nessa abordagem. Na seção 5.2 apresentamos o modelo utilizado para realizar as classificações das gestantes, incluindo o balanceamento dos dados e a aplicação da técnica de poda. Por fim, na seção 5.3 apresentamos os experimentos feitos e a avaliação do modelo, com as considerações sobre os resultados obtidos.

5.1 CONJUNTO DE DADOS

O termo procedimento usado nesta seção pode ser definido como qualquer tipo de consulta realizada por uma gestante. A Figura 13 apresenta o ranking dos procedimentos mais frequentes do conjunto de dados utilizado neste estudo.

Figura 13 - Distribuição de frequências dos procedimentos da base de dados.



Fonte: Elaborado pela autora (2021).

Segundo a Figura 13, a maior parte dos registros está concentrada entre os 53 procedimentos mais frequentes. Esses procedimentos estão listados no Apêndice B. Inicialmente, a base de dados continha 702.661 registros de consultas de gestantes. Dentre todos esses registros, 26.984 registros eram referentes à consulta odontológica, o que para este modelo não é relevante, visto que, consultas odontológicas não carregam informações relacionadas ao período gestacional de uma mulher. Além disso, existiam aqueles procedimentos em que a frequência que ocorriam era muito baixa, totalizando 3.104 registros. Nesse sentido, acrescentar procedimentos pouco frequentes podem não agregar valor às avaliações a serem realizadas nesse trabalho. Assim, desconsideramos os procedimentos realizados menos

de 50 vezes. Sendo assim, foram considerados nos modelos os 53 procedimentos mais frequentes, o que corresponde a 96% do total dos registros existentes na base de dados. Vale ressaltar que para a aplicação do modelo de classificação, o conjunto de dados foi dividido em um conjunto de treinamento (70%) e um conjunto de testes (30%). O modelo foi executado várias vezes pegando partes aleatórias dos dados para treinamento e teste.

5.2 MODELO PROPOSTO

Nessa seção serão descritos a metodologia utilizada durante o desenvolvimento do modelo de classificação. Apresentamos uma breve descrição do modelo de classificação na subseção **5.2.1**. Em seguida, foi necessário realizar um balanceamento dos dados presente na subseção **5.2.2** e por fim, na subseção **5.2.3** apresentamos todo o processo de poda realizado na árvore de decisão.

5.2.1 Descrição do modelo

Para realizar a classificação, utilizamos o algoritmo de árvore de decisão. Esse algoritmo tem como entrada duas matrizes, uma matriz X com as amostras de treinamento e uma matriz Y com valores inteiros referentes aos rótulos de classe para as amostras de treinamento.

Nos parâmetros, utilizados como critério o "Índice de Gini", a estratégia usada para escolher a divisão em cada nó foi "best", inicialmente não definimos uma profundidade máxima da árvore, ou seja, ela cresceu o máximo que podia. Definimos o mínimo de amostras necessárias para dividir um nó interno sendo 2, o número mínimo de amostras necessárias para estar em um nó folha foi definido como 1. Os parâmetros foram escolhidos empiricamente.

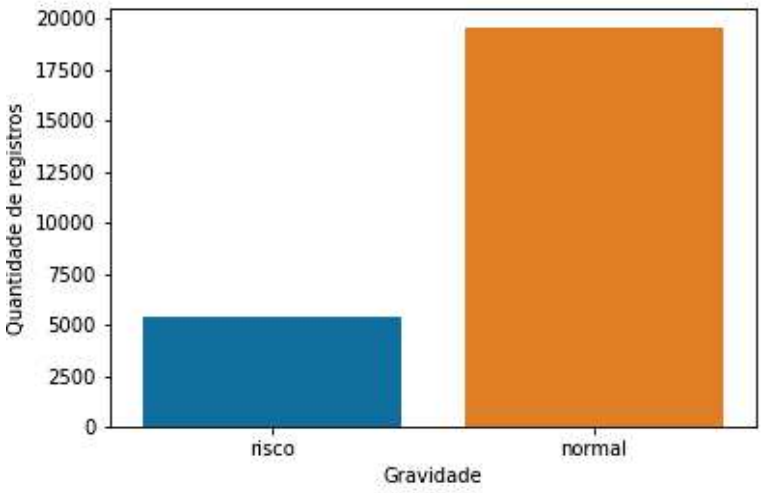
5.2.2 Balanceamento dos dados

Atualmente, é muito comum a coleta de dados para a realização de análises, aplicação de técnicas de automação ou de aprendizado de máquina. Os dados, em sua grande maioria, são desbalanceados (97) e (23).

No caso desta pesquisa, era esperado que existissem mais casos da classe 0 (gestação normal) do que da classe 1 (gestação de risco). Na base estudada, existem 5.357 gestantes consideradas com gravidez de risco e 19.559 gestantes consideradas com gravidez normal, como é ilustrado na Figura 14.

O modelo foi executado e a classificação gerada estava totalmente enviesada, visto que os dados utilizados neste estudo são desbalanceados. Caso esse desbalanceamento não seja ajustado, o resultado da classificação gerada não é confiável (62). Isso significa que o modelo tende a classificar novos dados como sendo da classe que possui mais exemplos,

Figura 14 - Quantidade de registros de cada classe na base de dados.



Fonte: Elaborado pela autora (2021).

neste caso, da classe de gestantes normais. Um problema dessa classificação incorreta, por exemplo, é que ao classificar incorretamente uma gravidez de risco, a tomada de decisão nesses casos ficará prejudicada, dado que essa gravidez é rotulada como uma gravidez normal

Para ilustrar melhor a classificação enviesada do modelo, temos na Tabela 6 a representação numérica de uma matriz de confusão. A diagonal principal da matriz representa o valor predito corretamente de cada classe. Assim, ao observar a matriz, é possível notar que, o modelo classificou 91% da classe 0 e que realmente eram da classe 0 e 9% como da classe 1, que eram da classe 0, como mostra a primeira linha da Tabela. Já, ao observar a segunda linha da Tabela, é possível identificar que o modelo classificou 38% sendo predito da classe 0 e que eram da classe 1 e 62% da classe 1 que realmente eram da classe 1.

Tabela 6 – Matriz de confusão do modelo com a classificação enviesada.

Rótulo real	**Classe 0**	0.91	0.09
	Classe 1	0.38	0.62
		Classe 0	**Classe 1**
		Rótulo predito	

Fonte: Elaborado pela autora (2021).

Para superar esse desbalanceamento e melhorar a classificação, aplicamos três técnicas de balanceamento nos dados, de modo a identificar a que melhor se adequa ao

modelo proposto e aos dados. As técnicas serão apresentadas com os respectivos resultados.

A técnica "*Under Sampling*" reduz os exemplos da classe majoritária de forma aleatória, ou seja, o algoritmo percorre o conjunto procurando exemplos da classe majoritária e removendo até que consiga equilibrar a amostra final. Isso significa que no final desse algoritmo a amostra tem o mesmo número de casos das duas classes (8) (41).

Já a técnica de balanceamento "*NearMiss*" refere-se a uma coleção de métodos de *undersampling* que selecionam exemplos com base na distância dos exemplos de classes majoritárias aos exemplos de classes minoritárias. De forma eficiente, ele percorre os pontos observando a distribuição da classe e eliminando aleatoriamente as amostras da classe maior. Isso significa que quando dois pontos de classes diferentes estão muito próximos um do outro, esse algoritmo elimina o ponto da classe majoritária, tentando assim equilibrar a distribuição (126) (8).

Outra técnica estudada foi o "*Smote*" usado para gerar exemplos sintéticos operando em 'espaço de recursos' ao invés de 'espaço de dados'. Isto é, a classe minoritária é super amostrada onde para cada amostra da classe minoritária é introduzido exemplos sintéticos ao longo dos segmentos. Essas amostras sintéticas são criadas na vizinhança considerando o caso mais próximo da classe minoritária, fazendo com que o classificador crie regiões de decisão maiores e menos específicas, em vez de regiões menores e mais específicas, como é normalmente causado por sobreamostragem com replicação. Agora, regiões mais gerais são apreendidas para a classe minoritária ao invés de serem rodeadas de amostras maiores da classe majoritária, aumentando o poder de generalização dos classificadores gerados para essa amostra de dados (23) (65).

O modelo de classificação foi executado sobre os dados utilizando cada uma das técnicas de balanceamento mencionadas anteriormente e representamos as métricas de avaliação dessas execuções na Tabela 7. Assim, é possível fazer uma comparação para detectar qual é a melhor técnica de balanceamento para ser usada neste modelo.

Tabela 7 – Resultado médio das métricas geradas pelo modelo de classificação usando técnicas de balanceamento.

Métricas	Nenhuma		UnderSampling		**Near Miss**		Smote	
	Classe 0	Classe 1	Classe 0	Classe 1	Classe 0	Classe 1	Classe 0	Classe 1
precisão	0.92	0.68	0.80	0.78	**0.90**	**0.91**	0.86	0.83
revocação	0.91	0.71	0.77	0.80	**0.91**	**0.90**	0.82	0.87
f1-score	0.91	0.70	0.78	0.79	**0.90**	**0.90**	0.84	0.85
acurácia	0.86		0.78		**0.90**		0.84	

Fonte: Elaborado pela autora (2021).

Na Tabela 7, temos na primeira coluna, as métricas de avaliação do modelo sem usar nenhuma técnica de balanceamento. Podemos ver o desempenho do modelo sem balancear, assim, temos que uma acurácia de 86%, o que não é ruim para um modelo, mas

como os dados estão desbalanceados, a acurácia não é a melhor métrica de avaliação do modelo. Ao observarmos a revocação (que se refere ao número de acertos) para a classe 0 é muito alta, por outro lado, para a classe 1 é muito baixa, reforçando assim, a ideia de enviesamento do modelo, pois ele aprendeu mais sobre a classe que tinha uma maior quantidade de exemplos.

Dentre às três técnicas avaliadas, a *Near Miss* foi a que se mostrou mais eficiente. Além de ter melhorado a acurácia do modelo, a métrica revocação dessa técnica obteve um número de acertos quase igual para às duas classes e as outras métricas obtiveram resultados equilibrados para ambas as classes. Sendo assim, a técnica *Near Miss* foi a que melhor se enquadrou nesta pesquisa.

Como os dados passaram por um processo de balanceamento, a amostra que era inicialmente composta por 24.916 gestantes foi reduzida para 10.714 gestantes, sendo 5.357 gestantes de risco e 5.357 gestantes normais. Com isso, foram separados de maneira aleatória 7.499 gestantes para o processo de treinamento e 3.215 gestantes para o processo de teste.

5.2.3 Poda da árvore de classificação

As árvores de decisão sofrem um problema comum de sobreajuste. Por padrão, o classificador de árvore de decisão não realiza nenhuma poda e, com isso, a árvore cresce o máximo. Ao reproduzir o classificador, obteve-se a pontuação de precisão de 0.99 e 0.83 no treino e no teste respectivamente. Por isso, podemos dizer que o classificador da árvore está sobreajustada dado que, o modelo se ajusta completamente aos dados de treino e falha ao generalizar os dados de testes não vistos.

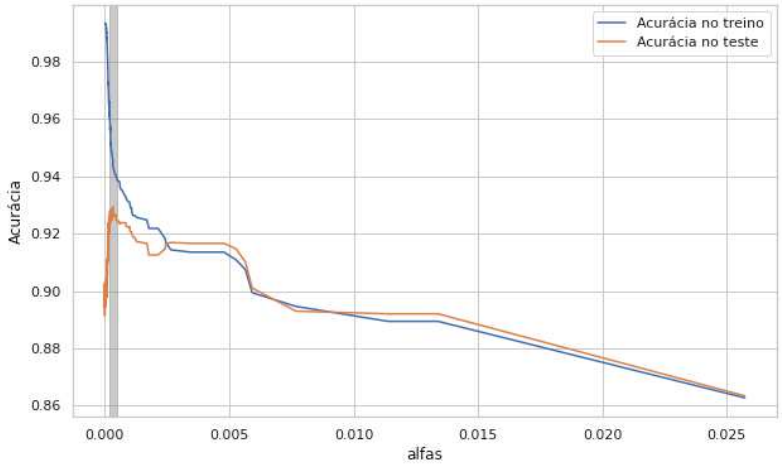
Em geral, métodos de poda servem para evitar que árvores de decisão sofram sobreajuste. Neste trabalho, foi utilizado a poda custo-complexidade mínima descrita na seção *2.3.1.1*. Nos classificadores de árvore de decisão essa técnica de poda é parametrizada pelo parâmetro de complexidade de custo alfa.

A Figura 15 representa a relação entre os valores de acurácia no treino e no teste em seus respectivos valores de alfa. Ao observar essa imagem, é possivel notar que a precisão máxima do teste é obtida entre alfa = 0.000 e 0.001. Sendo assim, o valor do alfa escolhido foi de 0.001.

5.3 AVALIAÇÃO DO MODELO DE CLASSIFICAÇÃO DE GESTANTES

Para avaliar o desempenho do modelo, repetimos a execução por 30 vezes. A cada execução, separamos os dados aleatoriamente, para que assim, partes de dados diferentes fossem usadas para efetuar o treino e o teste, gerando resultados diferentes em cada uma das vezes que executamos o modelo.

Figura 15 - Acurácia x alfas para os dados de treino e de teste.



Fonte: Elaborado pela autora (2021).

Tabela 8 – Matriz de confusão do modelo de classificação das gestantes sem poda.

Rótulo real	**Classe 0**	0.92	0.08
	Classe 1	0.11	0.89
		Classe 0	**Classe 1**
		Rótulo predito	

Fonte: Elaborada pela autora (2021).

Salienta-se que agora o modelo de árvore de decisão sofre com um problema muito comum conhecido como *overfitting*, que ocorre quando a árvore cresce até a sua profundidade máxima e pode decorar o conjunto de treinamento ocasionando em uma piora no seu poder de predição quando aplicado ao conjunto de testes. Para tratar esse *overfitting*, aplicamos a poda de custo-complexidade mínima. Após aplicar a poda no modelo, tem-se a matriz de confusão representada pela Tabela 9.

Tabela 9 – Matriz de confusão do modelo de classificação das gestantes após a poda.

Rótulo real	**Classe 0**	0.95	0.05
	Classe 1	0.09	0.91
		Classe 0	**Classe 1**
		Rótulo predito	

Fonte: Elaborada pela autora (2021).

Ao observar essa Tabela, é possível ver que o modelo previu corretamente 95% dos casos da classe 0 e 91% dos casos da classe 1, com acurácia geral do modelo de 93% e desvio padrão de 0.02. Ao comparar a matriz de confusão após o processo de poda com a matriz inicialmente gerada, tem-se que, ao realizar a poda, o modelo teve uma melhora na previsão de 3% para ambas as classes e um aumento de 3% na acurácia geral.

6 PREDIÇÃO DE GESTAÇÕES

Diante do modelo de classificação de gestantes apresentado no capítulo 5, tem-se a possibilidade de uma nova aplicação desse modelo de modo a realizar uma previsão de gestações de risco. Assim, neste capítulo, utiliza-se o modelo de classificação parametrizado no capítulo anterior com dados temporais das gestantes. Essa aplicação permite verificar com quantas semanas de gestação é possível classificar uma gestação como de risco ou não.

Na seção 6.1, apresentamos o processo de filtragem temporal para criação dos grupos. Como utilizamos os parâmetros do modelo de classificação, na seção 6.2 apresentamos o processo de poda para cada árvore de decisão gerada e por fim, na seção 6.3 apresentamos a avaliação do modelo de predição com as nossas considerações.

6.1 FILTRAGEM TEMPORAL DAS CONSULTAS

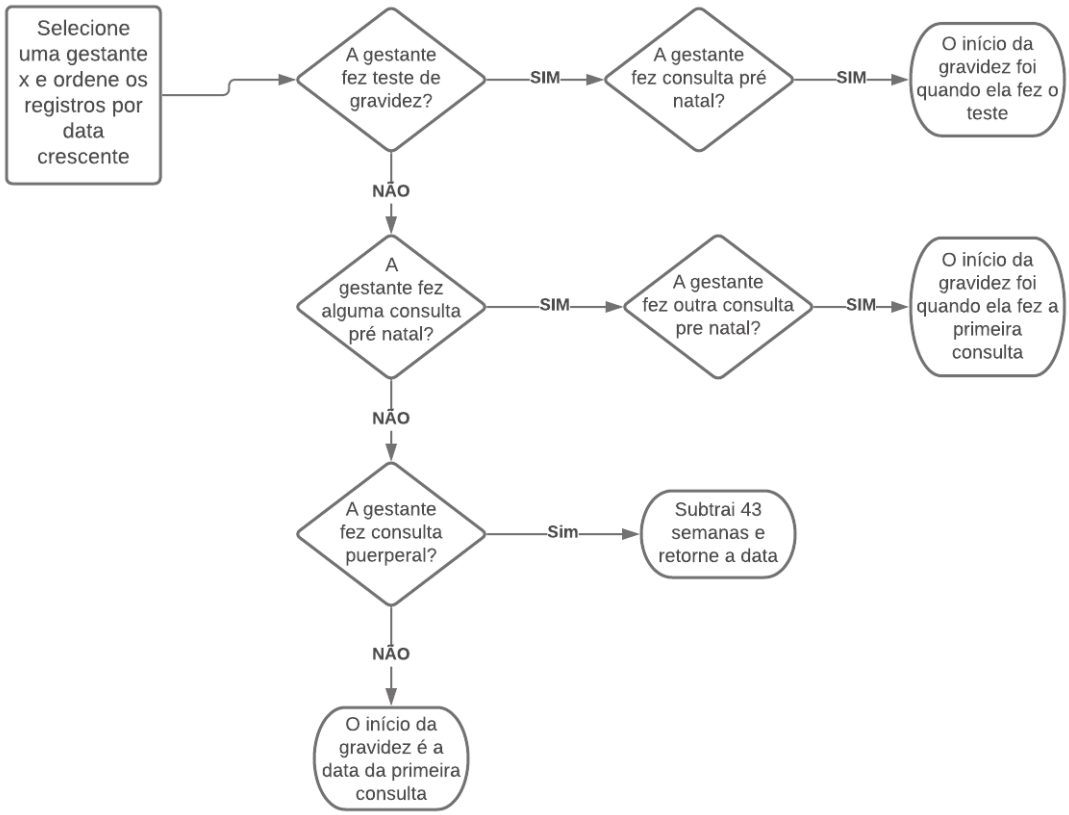
Os dados utilizados na aplicação do modelo anterior foram filtrados temporalmente. Essa filtragem temporal foi responsável por gerar grupos de dados referentes a períodos da gestação. Ou seja, inicialmente geramos um conjunto de dados referentes as 2 primeiras semanas de gestação e posteriormente, fomos acrescentando registros referentes as próximas semanas de gestação, sempre acrescentando a cada duas semanas, terminando na semana 16, visto que, na semana 16 já temos uma relevante quantidade de registros. Sendo assim, temos oito grupos de dados, cada um referente às semanas 2, 4, 6, 8, 10, 12, 14 e 16 de gestação de todas as gestantes contidas na base de dados.

Para realizar essa filtragem, primeiramente precisávamos identificar o início da gestação de cada gestante e com isso, ir adicionando os dados conforme os períodos. Para isso, temos o diagrama representado pela Figura 16, que descreve a ideia por trás do método utilizado para descobrir a data aproximada do início da gestação. Sabendo que todas as gestantes da base realizaram consultas do tipo pré-natal, usamos três casos para identificar o início da gestação. Vale salientar que todas as gestantes da base de dados se enquadram em um dos casos abaixo.

No primeiro caso, o algoritmo seleciona uma gestante da base de dados, ordena os registros por ordem cronológica e faz algumas verificações usando o código do procedimento. Inicialmente o algoritmo verifica se a gestante realizou algum procedimento do tipo “Teste ou Exame de Gravidez”, se sim, o algoritmo verifica se a gestante realizou alguma consulta do tipo “Consulta Pré Natal”, se sim, a data aproximada do início da gestação é quando a gestante realizou o teste ou exame de gravidez. Caso o algoritmo não tenha encontrado nenhum procedimento do tipo teste ou exame de gravidez, ele percorre os procedimentos da gestante em busca por dois procedimentos de “Consulta Pré Natal” e caso encontre, retorna a data da primeira consulta pré-natal registrada.

Por fim, o algoritmo busca pelo procedimento do tipo “Consulta Puerperal” e caso

Figura 16 - Metodologia utilizada para identificar o início da gestação.



Fonte: Elaborado pela autora (2021).

encontre, subtrai 43 semanas da data da consulta puerperal e com isso, tem-se a data inicial aproximada para a gestante que se enquadra nesse caso.

O algoritmo usado para acrescentar os registros referentes aos períodos de datas selecionadas para essa pesquisa está apresentado pelo pseudocódigo contido no algoritmo 2. De forma resumida, o algoritmo recebe o conjunto de dados inicial utilizado no processo de classificação do capítulo 5 e uma lista de todas as gestantes deste dataframe. Inicialmente, o algoritmo pega uma gestante da base de dados e chama o método responsável por calcular a data de início da sua gestação. Ao retornar a data de início da gestação, o algoritmo chama o método responsável por adicionar à data inicial, as semanas referentes ao período de dados desejado. Assim, temos uma data de início da gestação e uma data período que se refere a data do início da gestação mais o período acrescido. Por fim, o algoritmo realiza uma filtragem de procedimentos entre às duas datas. O processo se repete até percorrer todas as gestantes da lista.

Após as filtragens e de modo a obter características mais completas para o modelo, foram fixados os procedimentos da quarta semana e, para os conjuntos de dados das demais

semanas, apenas os procedimentos da quarta semana foram filtrados. Com isso, as *features* de todos os grupos são as mesmas da quarta semana, mudando somente a quantidade de atendimentos que foram realizados daquele procedimento.

Tendo esses novos grupos de dados, foram gerados para cada grupo um modelo de predição de gestação de risco. Assim, seria possível observar as métricas de avaliação do modelo para identificar a partir de qual semana o modelo conseguiu rotular a gestação com uma precisão.

Algoritmo 2: Filtragem temporal para criação dos grupos de dados

```
Entrada: dataframe, listagestantes
início
    dataframeFinal = novoDataframe()
    repita
        registrosGestante = filtro o dataframe com 1 gestante da lista;
        dataInicio = calculaInicioGestacao(registrosGestante);
        dataPeriodo = acrescentaSemanas(dataInicio);
        filtroSemana = registrosEntreDatas(registrosGestante, dataInicio,
         dataPeriodo);
        dataframeFinal.add(filtroSemana);
    até percorrer a lista de gestantes;
    dataframeFinal;
fim
```

6.2 PODA DAS ÁRVORES TEMPORAIS

Para cada grávida da base de dados, foram geradas informações referentes a 2, 4, 6, 8, 10, 12, 14 e 16 semanas, resultando em oito grupos de dados. Para cada um desses grupos, foi avaliado através de árvores de decisão, a capacidade preditiva do modelo apresentado pelo capítulo 5. Assim, foram geradas árvores usando os mesmo parâmetros e tornou-se necessário parametrizar um valor de alfa para realizar a poda de cada uma das árvores de decisão geradas.

Tabela 10 – Valores dos alfas escolhidos por período de data

	2 sem	4 sem	6 sem	8 sem	10 sem	12 sem	14 sem	16 sem
alfa escolhido	0.0015	0.00020	0.00023	0.00020	0.00025	0.00020	0.00030	0.00020

Fonte: Elaborada pela autora (2021).

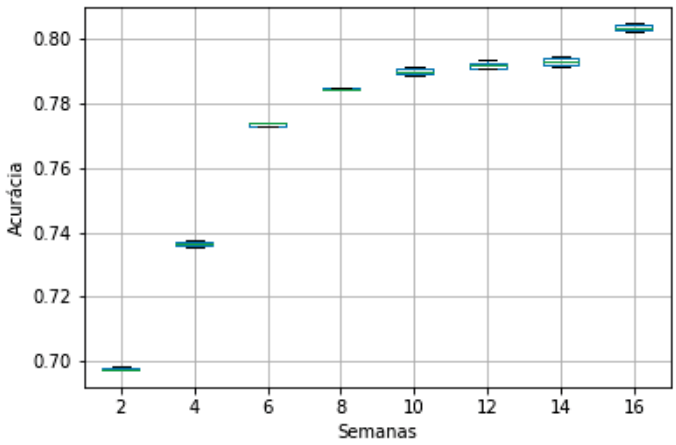
Ao reproduzir o modelo, foi realizado um cálculo que mostrava a faixa em que os alfas retornavam o melhor desempenho do modelo, isto, para cada árvore. Com isso,

temos na Tabela 10 a representação dos valores dos alfas escolhidos para cada árvore de modo a obter o melhor resultado no processo de poda.

6.3 AVALIAÇÃO DO MODELO DE PREDIÇÃO

Considerando que proposta deste modelo de avaliação temporal é de forma automática predizer em qual semana é possível classificar uma gestação, seja classificando como uma gestação normal ou classificando como gestação de risco, temos a Figura 17 que mostra a acurácia do modelo de classificação temporal. Ao observar a Figura 17, é possível notar que, a acurácia do modelo aumenta conforme são inseridos os dados na base de estudo. As acurácias que se destacam são as da semana 12, 14 e 16 respectivamente, visto que são os conjuntos de dados com mais registros das gestantes. Vale salientar que os dados foram adicionados a cada duas semanas, e que o primeiro experimento foi realizado após duas semanas de gestação. A Figura 17 mostra que mesmo após apenas duas semanas, o modelo possui uma precisão de 69%, que já representa um resultado satisfatório para o conjunto de dados com apenas duas semanas.

Figura 17 - Acurácia do modelo de classificação temporal com poda

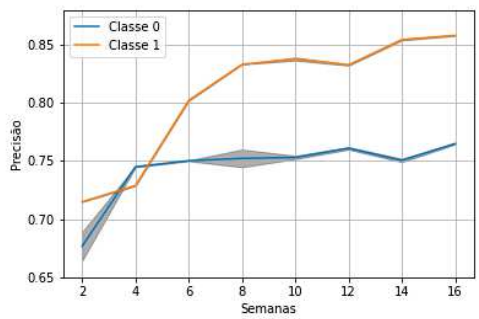


Fonte: Elaborado pela autora (2021).

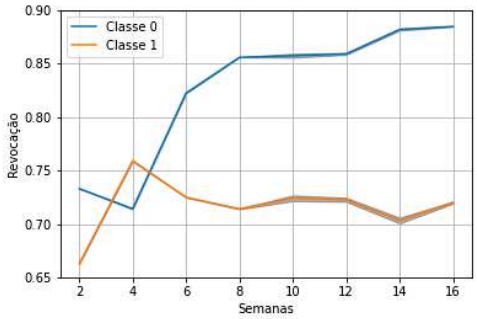
Além da acurácia, temos a imagem 18 que mostra a avaliação de acordo com as métricas do modelo de classificação temporal.

Ao analisar a Figura, podemos observar que a partir da 12ª semana a taxa f1-score é de 81% para a classe 0 e 78% para a classe 1, ou seja, a porcentagem de gestantes que estão sendo classificadas corretamente. Já os valores menores das métricas, ocorrem nos conjuntos de dados referentes a semana 2 e semana 4, e isto se justifica, considerando que

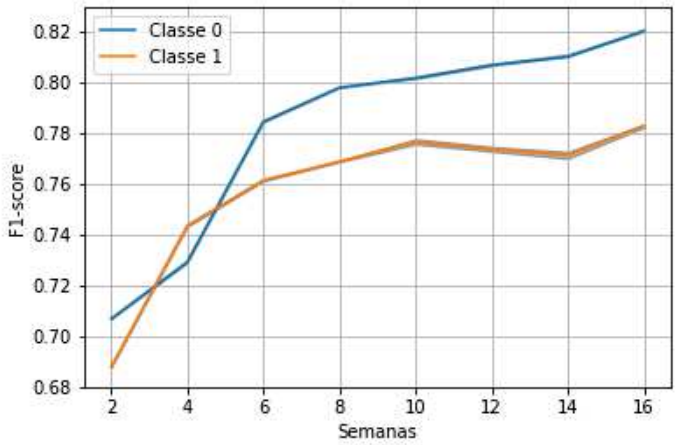
Figura 18 - Avaliação do modelo conforme as métricas de avaliação e com intervalos de confiança, onde (a) aponta dentre todas as classificações da classe positiva que o modelo fez, quantas estão corretas, (b) que indica dentre todas as situações de classe positivas como esperado, quantas estão corretas e (c) apresenta uma média harmônica entre precisão e revocação.



(a) Precisão



(b) Revocação



(c) f1-score

Fonte: Elaborado pela autora (2021).

nesses conjuntos de dados, a concentração de registros referentes ao histórico da gestante são reduzidos.

A partir da 16ª semana, os valores das métricas já representam resultados satisfatórios. Nesta semana, a taxa de revocação pode chegar a 89% para a classe 0 e 73% para a classe 1. Já os valores do f1-score se aproximam de 82% para a classe 0 e 79% para a classe 1. Além dos valores mencionados, a Figura 18 também apresenta o intervalo de confiança dos valores das respectivas métricas, que ajudam a indicar a confiabilidade da estimativa feita. Assim, salienta-se que um intervalo de confiança pequeno é confiável para as estimativas da pesquisa.

7 AGRUPAMENTO DE GESTANTES

Nesta seção, são aplicadas técnicas de agrupamento de dados, nos dados relativos às gestantes, para encontrar grupos de gestantes sem intervenção humana. Os grupos formados são analisados semanticamente e confrontados com os resultados das classificações dos capítulos anteriores. Apresentamos na seção 7.1, as filtragens realizadas na base de dados para obtermos o conjunto de dados utilizado nessa abordagem. Já na seção 7.2, descrevemos a aplicação do modelo de clusterização, incluindo as técnicas utilizadas. Por fim na seção 7.3, apresentamos os agrupamentos, com as análises e considerações sobre os grupos gerados.

7.1 CONJUNTO DE DADOS

Para realizar o agrupamento, optamos por inserir mais características ao conjunto de dados, para que assim fosse possível utilizar o máximo de informações disponíveis no processo onde são gerados os grupos. Isso, de modo a propor um modelo que complete os modelos de classificação usados nos capítulos anteriores. Assim, foram adicionados o procedimento mais frequente, o primeiro e o último procedimento realizado por cada uma das gestantes.

O procedimento mais frequente, o primeiro e o último procedimentos realizados por cada gestante podem variar de uma gestante para outra. Por se tratarem de variáveis categóricas, transformamo-as em variáveis binárias. Isso significa que, para cada gestante, criamos uma característica com o primeiro, último e procedimento mais frequente realizado por ela, sendo marcado 1 como procedimento realizado e 0 como procedimento não realizado. Assim, ao percorrer todas as gestantes, o algoritmo resultou em um conjunto de dados com 1.284 colunas. Cada coluna é responsável por caracterizar um procedimento realizado sendo consideradas dimensões, ou seja, temos uma base de dados com muitas dimensões.

7.2 MODELO PROPOSTO

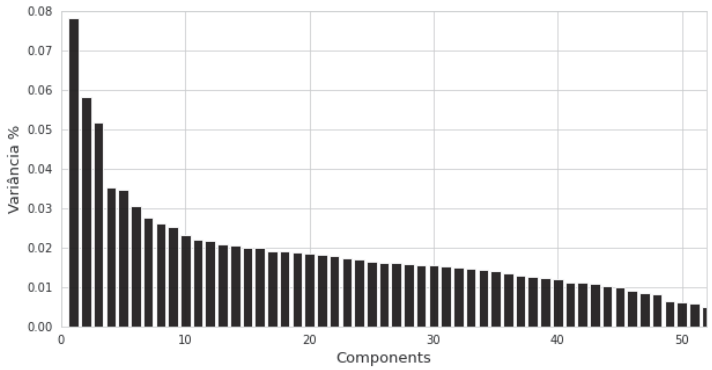
Visto que um número elevado de dimensões exige um alto poder de processamento computacional, foi necessário diminuir a dimensionalidade do dado. Para isso, foi utilizado a técnica de análise de componentes principais (PCA). É uma técnica de análise multivariada onde seu objetivo é encontrar um meio de condensar a informação contida em várias variáveis originais em um conjunto menor de variáveis estatísticas que caracterizam os objetos estudados com uma perda mínima de informação. Essas novas variáveis são chamadas componentes principais (66).

Os componentes principais apresentam propriedades importantes: cada componente

principal é uma combinação linear de todas as variáveis originais, são independentes entre si e estimados com o propósito de reter, em ordem de estimação, o máximo de informação, em termos da variação total contida nos dados (116).

As variâncias dos componentes principais representam a proporção de variância total explicada pelo componente principal (116) e estão presentes no gráfico da Figura 19. Ao observar este gráfico, é possível notar que os valores vêm em um decrescente e que os 10 primeiros componentes representam uma parte significativa dos dados contidos na base original. Sendo assim, para esse estudo, foram selecionados os 10 primeiros componentes principais.

Figura 19 - Variação dos principais componentes gerados pelo PCA



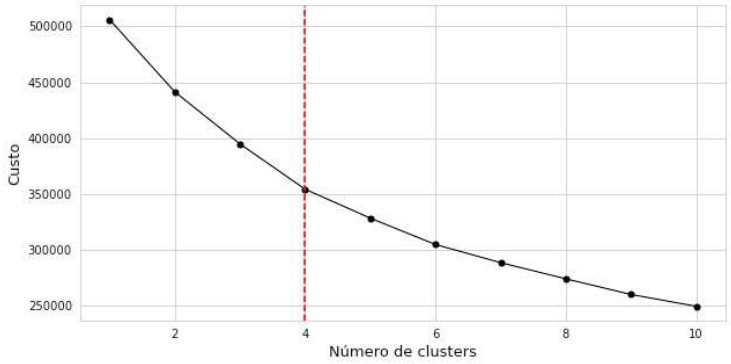
Fonte: Elaborado pela autora (2021).

Sabendo que a clusterização visa fazer agrupamentos automáticos de dados, segundo o seu grau de semelhança (125), é necessário definir previamente o número de grupos. Assim, gestantes com características similares pertencerão a um mesmo grupo e gestantes com características diferentes irão pertencer a grupos diferentes.

Com isso, utilizamos o método *elbow* para nos auxiliar a encontrar um número ideal de grupos, pois o objetivo desse método é determinar o melhor número de grupos que podem ser obtidos, mesmo sem conhecer a resposta do método de agrupamento (110). Resumidamente, o método consiste em representar graficamente a variação explicada em função do número de grupos e escolher o cotovelo da curva como o número de grupos a utilizar (88).

A Figura 20 contém a função de custo com a soma dos quadrados das distâncias entre os grupos plotados no gráfico e ao observar podemos ver o momento em que o valor de k cai drasticamente criando uma leve curva no gráfico. Não é bem claro, mas pode ser um bom indicativo de que $k = 4$ pode ser o valor interessante de k a ser trabalhado e assumiremos esse valor para realizar o agrupamento.

Figura 20 - Imagem do gráfico gerada pela aplicação do método *elbow*



Fonte: Elaborado pela autora (2021).

7.3 ANÁLISES DOS AGRUPAMENTOS GERADOS PELO MODELO

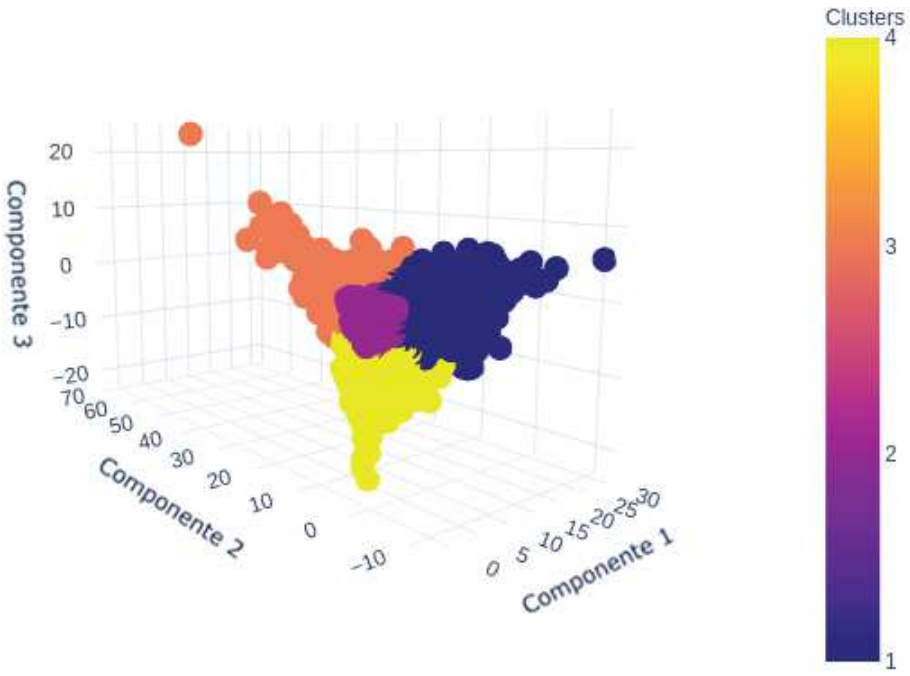
Na redução de dimensão utilizando o PCA, foram obtidas novas variáveis ou componentes que expressam reduzidamente as informações contidas no conjunto de dados original. Assumindo 4 como o número de grupos desejados e para a visualização desses grupos gerados pelo *K-means*, selecionamos os três componentes principais que representam um percentual relevante dos dados. Esses grupos gerados com base nos componentes principais são mostrados na Figura 21.

De modo a definir os grupos e identificar o que representa cada grupo, tem-se a Tabela 11 com as médias das quantidades de consultas realizadas por cada grupo.

O grupo de gestações de médio risco é composto por gestantes que realizam muitas consultas pré-natais, apenas algumas consultas de atenção e algumas consultas relacionadas com comorbidades. Ao observar a Tabela 11, temos o grupo 1 que tem em média 7.45 consultas pré-natal, tem a maior média dentre os grupos, de consultas de supervisão de gravidez normal, e, além disso, destacam-se as médias das consultas de visita domiciliar por profissional de nível médio e atendimento de urgência em atenção básica e com isso, esse grupo tem características de um gestação de médio risco, onde a gestante precisa de uma atenção especializada.

Podemos descrever o grupo de gestação normal sendo composto por gestantes que realizam consultas de pré-natal periodicamente. O grupo 2 é o grupo que tem o maior número de gestantes, destacando-se o procedimento/diagnóstico supervisão de gravidez normal com média 2.48 consultas e consulta pré-natal com média de 5.53 consultas. Por terem as médias de consultas realizadas mais equilibradas, intuitivamente, pode ser equiparado a um grupo equivalente a uma gestação normal.

Figura 21 - Visualização 3D do agrupamento das gestantes, onde temos quatro grupos representados por cores diferentes para melhor visualização dos pontos pertencentes a cada grupo



Fonte: Elaborado pela autora (2021).

O grupo de gestações normais com atenção é composto por gestantes que realizam um alto número de consultas pré-natais, algumas consultas de atenção e alguns exames. No grupo 3, destacam-se as consultas médicas em atenção básica que está entre as maiores médias e atendimento de urgência em atenção básica que tem a maior média desse procedimento realizado. Além disso, é o grupo em terceiro de média da consulta de supervisão de gravidez de alto risco, tem a segunda média no procedimento aferição de pressão arterial e 5.39 consultas de pré-natal. Baseado nessas médias, o grupo 3 tem características de um grupo de baixo risco ou normal com atenção.

Por último, tem-se o grupo de gestações de alto risco composto por gestantes que realizam muitas consultas de atenção e muitas consultas referentes à alguma comorbidade. Considerando a média dos procedimentos do grupo 4, tem-se 5.67 consultas de supervisão de gravidez de alto risco, 3.84 consultas de aferição de pressão arterial, tem a maior média de consultas pré-natal e 4.75 em consulta médica em atenção básica. Esse grupo pode ser de alto risco, pois tem características que rementem ao encontrado nos capítulos 5 e 6.

A aplicação desse algoritmo em dados da saúde poderia servir de auxílio à tomada de decisão relacionado às gestantes. Geralmente, identificar qual grupo pertence uma gestante, facilitaria em relação à quais medidas devem ser tomadas durante o período gestacional.

Tabela 11 – Número médio de procedimentos/diagnósticos realizados por grupo, onde intuitivamente os grupos representam: gestações de médio risco, gestações normais, gestações de baixo risco ou normal com atenção e gestações de alto risco respectivamente.

Procedimentos/Diagnósticos	**Grupo 1**	**Grupo 2**	**Grupo 3**	**Grupo 4**
Consulta Pre-Natal	7.45	5.53	5.39	7.63
Supervisão De Gravidez Normal	2.71	2.48	1.55	0.19
Consulta Medica Em Atenção Básica	3.29	1.93	3.55	4.75
Consulta De Profissionais De Nível Superior Na Atenção Básica (Exceto Médico)	4.90	1.51	3.01	4.12
Atendimento De Urgência Em Atenção Básica	1.76	1.12	10.30	1.67
Visita Domiciliar Por Profissional De Nível Médio	2.21	1.17	2.14	0.14
Ultra-Sonografia Obstétrica	0.91	0.73	0.60	1.77
Consulta Medica Em Atenção Especializada	0.73	0.38	1.17	6.38
Coleta De Material P/ Exame Laboratorial	1.37	0.68	0.72	0.77
Aferição De Pressão Arterial	0.85	0.45	1.16	3.83
Supervisão De Gravidez De Alto Risco	0.47	0.22	0.37	5.67

Fonte: Elaborada pela autora (2021).

Como se trata de uma abordagem envolvendo suposições médicas, foi necessário validar com especialistas da área todo o processo de agrupamento e caracterização dos grupos. Após essa validação, concluímos que o agrupamento feito é coerente, validando o nosso modelo como eficiente.

8 CONCLUSÕES

Sabendo da importância do acompanhamento médico da gestante e que muitas vezes, a identificação precoce de uma gestação de risco pode possibilitar um cuidado especial à gestante, aliado à evolução das técnicas de aprendizado de máquina, têm se como objetivo principal a proposta de classificar gestações de risco.

Este trabalho apresentou um estudo em gestantes atendidas pelo SUS da cidade de São Paulo durante 2014 e 2015 onde, inicialmente, foi realizado uma caracterização dos dados, mostrando avaliações das trajetórias dos atendimentos das gestantes. Além dessa caracterização, utilizamos modelos de árvores de decisão para classificar gestações de gestantes em duas classes, normal e de risco. Os resultados experimentais obtidos na abordagem de classificação usando a árvore de decisão mostraram que o modelo conseguiu classificar corretamente mais de 90% das gestantes, tanto as classificadas como gestantes normais quanto as classificadas como gestantes de risco, cumprindo assim a proposta inicial. Vale salientar que estes resultados foram obtidos através de dados não clínicos,

Além dessa abordagem, aplicamos esse modelo em grupos de dados referentes a períodos da gestação. Esses períodos eram separados a cada duas semanas, ou seja, foram geradas árvores de decisão para os grupos de 2, 4, 6, 8, 10, 12, 14 e 16 semanas de gestação.Nessa aplicação, foram obtidos resultados satisfatórios a partir da 12ª semana, onde se obteve acurácia de quase 80%, ou seja, a partir dessa semana era possível prever mais de 80% das gestantes normais e quase 78% das gestantes de risco de maneira correta, o que torna essa aplicação interessante, pois a partir da 12ª semana de gestação não há como reverter uma gestação de risco. Assim, quanto antes a gestação de risco for predita, melhor será a tomada de decisão.

Por fim, a última abordagem apresentada neste trabalho, propôs o agrupamento de gestantes baseado em suas características. Essas características são os procedimentos/diagnósticos realizados pela gestante. Nessa abordagem trabalhamos com dados com alta dimensão, visto que se pegou os 53 procedimentos mais frequentes, o primeiro procedimento e o último procedimento realizado por cada gestante. Para esse agrupamento, foi utilizado o PCA com o algoritmo *K-means* e, com isso, encontramos quatro grupos diferentes de gestantes. Esses grupos foram classificados baseado nas médias das quantidades de consultas referentes aos procedimentos apresentados.

Como se trata de um estudo promissor, existem diversas possibilidades que podem ser complementos da pesquisa, um estudo interessante também inclui uma abordagem de modo a descobrir a trajetória posterior da gestante, ou seja, o que aconteceu após o parto e categorizar de acordo com a saúde. Outra proposta interessante de trabalho futuro consiste em realizar análises e discussões das árvores geradas.

REFERÊNCIAS

1 ABDI, Hervé; WILLIAMS, Lynne J. **Principal component analysis**. Wiley interdisciplinary reviews: computational statistics, v. 2, n. 4, p. 433-459, 2010.

2 AHMAD, Muhammad Aurangzeb; ECKERT, Carly; TEREDESAI, Ankur. **Interpretable machine learning in healthcare.** In: Proceedings of the 2018 ACM international conference on bioinformatics, computational biology, and health informatics. 2018. p. 559-560.

3 ALSABTI, Khaled; RANKA, Sanjay; SINGH, Vineet. **An efficient k-means clustering algorithm**. 1997.

4 ANVERSA, Elenir Terezinha Rizzetti et al. **Qualidade do processo da assistência pré-natal: unidades básicas de saúde e unidades de Estratégia Saúde da Família em município no Sul do Brasil**. Cadernos de Saúde Pública, v. 28, p. 789-800, 2012.

5 ARISHA, Amr; THORWARTH, Michael. **A Simulation-Based Decision Support System to Model Complex Demand Driven Heathcare Facilities**. 2012.

6 AYER, Turgay et al. **Breast cancer risk estimation with artificial neural networks revisited: discrimination and calibration**. Cancer, v. 116, n. 14, p. 3310-3321, 2010.

7 BAILEY, Trevor C. **Spatial statistical methods in health**. Cadernos de Saúde Pública, v. 17, n. 5, p. 1083-1098, 2001.

8 BAO, Lei et al. **Boosted near-miss under-sampling on SVM ensembles for concept detection in large-scale imbalanced datasets**. Neurocomputing, v. 172, p. 198-206, 2016.

9 BATISTA, Karina Barros Calife et al. **Atenção à gestante e à puérpera no SUS-SP: manual técnico do pré-natal e puerpério**. In: Atenção à gestante e à puérpera no SUS-SP: manual técnico do pré-natal e puérperio. 2010. p. 234-234.

10 BERTOZZO, Richard Junior et al. A**plicação de Machine Learning em dataset de consultas médicas do SUS**. 2019.

11 BHARDWAJ, Rohan; NAMBIAR, Ankita R.; DUTTA, Debojyoti. **A study of machine learning in healthcare.** In: 2017 IEEE 41st Annual Computer Software and Applications Conference (COMPSAC). IEEE, 2017. p. 236-241.

12 BHARDWAJ, Ruchie; SETHI, Adhiraaj; NAMBIAR, Raghunath. **Big data in genomics: An overview**. In: 2014 IEEE International Conference on Big Data (Big Data). IEEE, 2014. p. 45-49.

13 BHOLOWALIA, Purnima; KUMAR, Arvind. **EBK-means: A clustering technique based on elbow method and k-means in WSN**. International Journal of Computer Applications, v. 105, n. 9, 2014.

14 BODNAR, Lisa M. et al. **Machine learning as a strategy to account for dietary synergy: an illustration based on dietary intake and adverse pregnancy outcomes**. The American journal of clinical nutrition, v. 111, n. 6, p. 1235-1243, 2020.

15 BREIMAN, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1993). **Classification and regression trees**. wadsworth, 1984. Google Scholar.

16 BREUEL, C., 2020. **Carreiras em Machine Learning: O Presente e o Futuro**. SBC Horizontes. ISSN: 2175-9235. Disponível em: http://horizontes.sbc.org.br/index.php/2020/07/26/carreiras-em-machine-learning/

17 BROWNE, Michael W. **Cross-validation methods**. Journal of mathematical psychology, v. 44, n. 1, p. 108-132, 2000.

18 CALDEYRO-BARCIA, R. et al. Frecuencia cardíaca y equilibrio acido base del feto. Montevideo: Centro Latinoamericanode Perinatologia y Desarrollo Humano, n. 519, 1973.

19 CALLAHAN, Alison; SHAH, Nigam H. **Machine learning in healthcare**. In: Key Advances in Clinical Informatics. Academic Press, 2017. p. 279-291.

20 CARVALHO, Deborah Ribeiro; DALLAGASSA, Marcelo Rosano; DA SILVA, Sandra Honorato. **Uso de técnicas de mineração de dados para a identificação automática de beneficiários propensos ao diabetes mellitus tipo 2**. Informação & Informação, v. 20, n. 3, p. 274-296, 2015.

21 CASSIANO, Keila Mara. **Análise de Séries Temporais Usando Análise Espectral Singular (SSA) e Clusterização de Suas Componentes Baseada em Densidade**. Pontifícia Universidade Católica do Rio de Janeiro, 2014.

22 CHAVEZ-BADIOLA, Alejandro et al. **Predicting pregnancy test results after embryo transfer by image feature extraction and analysis using machine learning**. Scientific reports, v. 10, n. 1, p. 1-6, 2020.

23 CIESLAK, David A.; CHAWLA, Nitesh V.; STRIEGEL, Aaron. **Combating imbalance in network intrusion datasets**. In: GrC. 2006. p. 732-737.

24 COORDENADORIA DE REGIÕES DE SAÚDE - CRS. **Secretaria do Estado de Saúde de São Paulo**. Disponível em: `http://www.saude.sp.gov.br/coordenadoria-de-recursos-humanos/orgaos-sub-setoriais/coordenadoria-de-regioes-de-saude-crs`. Acesso em: 16 mar. 2021.

25 DA CUNHA CAVALCANTI, George Darmiton; REN, Tsang Ing. **Árvore de Decisão**.

26 DA SILVA, Michele Silveira; ROSA, Mary Rosane Quirino Polli. **Perfil de gestantes de alto risco atendidas em um centro obstétrico de Santa Catarina**. Revista Interdisciplinar, v. 7, n. 2, p. 95-102, 2014.

27 DASH, Sabyasachi et al. **Big data in healthcare: management, analysis and future prospects**. Journal of Big Data, v. 6, n. 1, p. 1-25, 2019.

28 DJELLOULI, A. et al. **A datamining approach to classify, select and predict the formation enthalpy for intermetallic compound hydrides**. International Journal of Hydrogen Energy, v. 43, n. 41, p. 19111-19120, 2018.

29 DO NASCIMENTO JUNIOR, Luiz Antônio Ferreira. **Aplicando método do gradiente ótimo na otimização do cálculo do grau de cobertura das regras em árvores de decisão Fuzzy**. Revista Brasileira de Computação Aplicada, v. 9, n. 3, p. 31-43, 2017.

30 DOMINGUES, Matheus Assis. **Desenvolvimento de um sensor virtual para controle da resistência à tração do papel em uma planta de polpa CTMP.**

31 DRAY, Stéphane. **On the number of principal components: A test of dimensionality based on measurements of similarity between matrices**. Computational Statistics & Data Analysis, v. 52, n. 4, p. 2228-2237, 2008.

32 ELMORE, Kimberly L.; RICHMAN, Michael B. **Euclidean distance as a similarity metric for principal component analysis**. Monthly weather review, v. 129, n. 3, p. 540-549, 2001.

33 ENGELSDORFF, Tiago Simon. **Métodos em machine learning para construção de curvas de carga a partir de medições**. 2019.

34 ESPINOZA, Camila et al. **Real-time simulation as a way to improve daily operations in an emergency room**. In: Proceedings of the Winter Simulation Conference 2014. IEEE, 2014. p. 1445-1456.

35 EVERITT, B. S. (1974) **Cluster Analysis**. Heinemann Educational Books, London.

36 EVERITT, B. S.; Landau, S.; Leese, M. & Stahl, D. **Cluster Analysis Wiley**, 2011.

37 FIGUEIREDO, Daniel R. **Introdução a redes complexas**. Atualizações em Informática, p. 303-358, 2011.

38 FRIEDL, Mark A.; BRODLEY, Carla E. **Decision tree classification of land cover from remotely sensed data**. Remote sensing of environment, v. 61, n. 3, p. 399-409, 1997.

39 GONÇALVES, Carla Vitola; CESAR, Juraci Almeida; MENDOZA-SASSI, Raul A. **Qualidade e equidade na assistência à gestante: um estudo de base populacional no Sul do Brasil**. Cadernos de Saúde Pública, v. 25, p. 2507-2516, 2009.

40 GUNASEGARAN, T., & Cheah, Y. N. (2017, May). **Evolutionary cross validation**. In 2017 8th International Conference on Information Technology (ICIT) (pp. 89-95). IEEE.

41 HA, Jihyun; LEE, Jong-Seok. **A new under-sampling method using genetic algorithm for imbalanced data classification**. In: Proceedings of the 10th International Conference on Ubiquitous Information Management and Communication. 2016. p. 1-6.

42 HANSEN, Matthew C. et al. **Global land cover classification at 1 km spatial resolution using a classification tree approach**. International journal of remote sensing, v. 21, n. 6-7, p. 1331-1364, 2000.

43 HASSAN, Md Rafiul et al. **A machine learning approach for prediction of pregnancy outcome following IVF treatment**. Neural computing and applications, v. 32, n. 7, p. 2283-2297, 2020.

44 HOLNESS, Nola. **High-risk pregnancy**. Nursing Clinics, v. 53, n. 2, p. 241-251, 2018.

45 IBGE. **Panorama populacional da cidade de São Paulo**. Disponível em: `https://cidades.ibge.gov.br/brasil/sp/sao-paulo/panorama`. Acessado em: 03 de novembro de 2021.

46 ITOA, Márcia et al. **Analysis of the existence of patient care team using social network methods in physician communities from healthcare insurance companies**. 2017.

47 JAIN, Anil K.; MURTY, M. Narasimha; FLYNN, Patrick J. **Data clustering: a review**. ACM computing surveys (CSUR), v. 31, n. 3, p. 264-323, 1999.

48 JASMIM, Leonardo de O. et al. **Caracterização de Atendimentos em uma Rede de Atenção à Saúde**. In: Anais do XVII Workshop de Informática Médica. SBC, 2017.

49 JENHANI, Ilyes; AMOR, Nahla Ben; ELOUEDI, Zied. **Decision trees as possibilistic classifiers**. International journal of approximate reasoning, v. 48, n. 3, p. 784-807, 2008.

50 JERRETT, Michael et al. **Spatial analysis for environmental health research: concepts, methods, and examples**. Journal of Toxicology and Environmental Health Part A, v. 66, n. 16-19, p. 1783-1810, 2003.

51 JUNIOR, Helvécio Miranda Magalhães. **Redes de Atenção à Saúde: rumo à integralidade**. Divulgação em saúde para debate [on-line], v. 52, p. 15-37, 2014.

52 KODINARIYA, Trupti M.; MAKWANA, Prashant R. **Review on determining number of Cluster in K-Means Clustering**. International Journal, v. 1, n. 6, p. 90-95, 2013.

53 KOUROU, Konstantina et al. **Machine learning applications in cancer prognosis and prediction**. Computational and structural biotechnology journal, v. 13, p. 8-17, 2015.

54 KULKARNI, Vrushali Y.; SINHA, Pradeep K. **Pruning of random forest classifiers: A survey and future directions**. In: 2012 International Conference on Data Science & Engineering (ICDSE). IEEE, 2012. p. 64-68.

55 LACHI, Ricardo Luis, and HV da Rocha. "**Aspectos básicos de clustering: conceitos e técnicas.**" Relatório Técnico–Instituto de Computação, Universidade Estadual de Campinas, Campinas (2005).

56 LAURETTO, Marcelo S. **Árvores de Decisão**. 2010.

57 LEE, Suzanne. **Risk perception in women with high-risk pregnancies**. British Journal of Midwifery, v. 22, n. 1, p. 8-13, 2014.

58 LIMA, Christiane Ferreira Lemos; ASSIS, F. M.; SOUZA, C. P. **Árvores de Decisão baseadas nas entropias de Shannon, Rényi e Tsallis para Sistemas Tolerantes a Intrusão**. In: La Novena Conferencia Iberoamericana en Sistemas, Cibernética e Informática: CISCI. sn, 2010. p. 34.

59 LISTGARTEN, Jennifer et al. **Predictive models for breast cancer susceptibility from multiple single nucleotide polymorphisms**. Clinical cancer research, v. 10, n. 8, p. 2725-2737, 2004.

60 LIU, Fan, and Yong Deng. "**Determine the number of unknown targets in Open World based on Elbow method**". IEEE Transactions on Fuzzy Systems (2020).

61 LIU, Hongwei, et al. "**Machine learning risk score for prediction of gestational diabetes in early pregnancy in Tianjin, China**". Diabetes/Metabolism Research and Reviews (2020): e3397.

62 LIU, Xu-Ying, Jianxin Wu, and Zhi-Hua Zhou. "**Exploratory undersampling for class-imbalance learning**". IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) 39.2 (2008): 539-550.

63 LLOYD, Stuart. **Least squares quantization in PCM**. IEEE transactions on information theory, v. 28, n. 2, p. 129-137, 1982.

64 LUZ, Bruna Gusman, et al. "**O perfil das gestantes de alto risco acompanhadas no pré-natal da policlínica de Divinópolis-MG, no biênio 2013/14**". Journal of Health & Biological Sciences 3.3 (2015): 137-143.

65 MACHADO, Emerson Lopes. **Um estudo de limpeza em base de dados desbalanceada e com sobreposição de classes**. 2007.

66 MAĆKIEWICZ, Andrzej; RATAJCZAK, Waldemar. **Principal components analysis (PCA)**. Computers & Geosciences, v. 19, n. 3, p. 303-342, 1993.

67 MAIA, Cynthia Moreira, Julio Cartier Maia Gomes, and Luana Dantas Chagas. "**Estudo Sobre o Uso de Árvores de Decisão na Área da Saúde.**" Anais do Encontro de Computação do Oeste Potiguar ECOP/UFERSA (ISSN 2526-7574) 1 (2017).

68 MARIN, Maikon Aloan. **Indução de Árvores de Decisão para a Inferência de Redes Gênicas**. 2013. Tese de Doutorado. Tese de Doutoramento, Universidade Tecnológica Federal do Paraná.

69 MENDES, Eugênio Vilaça. "**As redes de atenção à saúde.**" Rev Med Minas Gerais 18.4 Supl 4 (2008): S3-S11.

70 MENDES, Eugênio Vilaça. **As redes de atenção à saúde**. Ciência & saúde coletiva, v. 15, n. 5, p. 2297-2305, 2010.

71 METZ, Jean. **Interpretação de clusters gerados por algoritmos de clustering hierárquico**. Diss. Universidade de São Paulo, 2006.

72 MINISTÉRIO DA SAÚDE. "**Gestação de alto risco: manual técnico**". Estratégicas (2012). Disponível em: https://bvsms.saude.gov.br/bvs/publicacoes/manual_tecnico_gestacao_alto_risco.pdf.

73 MINISTÉRIO DA SAÚDE. **Manual prático para implementação da Rede Cegonha**. Brasília: Ministério da Saúde; 2011.

74 MINISTÉRIO DA SAÚDE. Portaria nº 570, de 1 de junho de 2000. "**Programa de Humanização no Pré-Natal e Nascimento**" (2000). Diário Oficial da União 2000; 1 jun. Disponível em: http://bvsms.saude.gov.br/bvs/saudelegis/gm/2000/prt0570_01_06_2000_rep.html.

75 MINISTÉRIO DA SAÚDE. Portaria nº 4.279, de 30 de dezembro de 2010. "**Estabelece diretrizes para a organização da Rede de Atenção à Saúde no âmbito do Sistema Único de Saúde (SUS).**" (2010). Diário Oficial da União 2010; 31 dez. Disponível em: http://bvsms.saude.gov.br/bvs/saudelegis/gm/2000/prt0570_01_06_2000_rep.html.

76 MINISTÉRIO DA SAÚDE. Portaria nº 1.130, de 5 de agosto de 2015. "**Institui a Política Nacional de Atenção Integral à Saúde da Criança (PNAISC) no âmbito do Sistema Único de Saúde (SUS)**". Diário Oficial da União 2015; 6 ago.

77 MINISTÉRIO DA SAÚDE. **Pré-Natal** (2019). Disponível em: https://www.saude.go.gov.br/biblioteca/7637-pr%C3%A9-natal. Acessado em: 02 de janeiro de 2022.

78 MIRANDA, M. A., et al. "**Characterization of the flow of patients in a hospital from complex networks**". Health Care Management Science 23.1 (2020): 66-79.

79 MITCHELL, Melanie; NEWMAN, Mark. **Complex systems theory and evolution**. Encyclopedia of evolution, v. 1, p. 1-5, 2002.

80 MOHAMED, W. Nor Haizan W.; SALLEH, Mohd Najib Mohd; OMAR, Abdul Halim. **A comparative study of reduced error pruning method in decision tree algorithms**. In: 2012 IEEE International conference on control system, computing and engineering. IEEE, 2012. p. 392-397.

81 MONARD, Maria Carolina, and José Augusto Baranauskas. "**Indução de regras e árvores de decisão**". Sistemas Inteligentes-Fundamentos e Aplicações 1 (2003): 115-139.

82 MONARD, Maria Carolina, and José Augusto Baranauskas. "**Conceitos sobre aprendizado de máquina**". Sistemas inteligentes-Fundamentos e aplicações 1.1 (2003): 32.

83 MOREIRA, Jorge RH et al. **Modelos de Aprendizado de Máquina na Predição de Diabetes Tipo 1 na Gestação usando Dados do Sistema Único de Saúde**. In: Anais do XXI Simpósio Brasileiro de Computação Aplicada à Saúde. SBC, 2021. p. 392-403.

84 MOSIER, C. I. (1951). Symposium: The need and means of cross-validation. I. **Problems and designs of cross-validation**. Educational and Psychological Measurement, 11, 5-11.

85 NEUMANN, Nelson A. et al. **Qualidade e equidade da atenção ao pré-natal e ao parto em Criciúma, Santa Catarina, Sul do Brasil**. Revista Brasileira de Epidemiologia, v. 6, p. 307-318, 2003.

86 OLIVEIRA, T. B. S. D. (2008). **Clusterização de dados utilizando técnicas de redes complexas e computação bioinspirada** (Doctoral dissertation, Universidade de São Paulo).

87 OSISANWO, F. Y. et al. S**upervised machine learning algorithms: classification and comparison**. International Journal of Computer Trends and Technology (IJCTT), v. 48, n. 3, p. 128-138, 2017.

88 PARK, Cheolsoo; TOOK, Clive Cheong; SEONG, Joon-Kyung. **Machine learning in biomedical engineering**. 2018.

89 PALMA, L. **Agrupamento de dados: k-médias**. Universidade Federal do Recôncavo da Bahia Centro de Ciências Exatas e Tecnológicas, 2018.

90 PHAM, Duc Truong; DIMOV, Stefan S.; NGUYEN, Chi D. **Selection of K in K-means clustering**. Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science, v. 219, n. 1, p. 103-119, 2005.

91 PRODROMIDIS, Andreas L.; STOLFO, Salvatore J. **Cost complexity-based pruning of ensemble classifiers**. Knowledge and Information Systems, v. 3, n. 4, p. 449-469, 2001.

92 RAGHUPATHI, Wullianallur; RAGHUPATHI, Viju. **Big data analytics in healthcare: promise and potential**. Health information science and systems, v. 2, n. 1, p. 1-10, 2014.

93 RASIA, Isabel Cristina Rosa Barros; ALBERNAZ, Elaine. **Atenção pré-natal na cidade de Pelotas, Rio Grande do Sul, Brasil**. Revista Brasileira de Saúde Materno Infantil, v. 8, p. 401-410, 2008.

94 RICHARDSON, Mark. "**Principal component analysis.**" Disponível em: `http://people.maths.ox.ac.uk/richardsonm/SignalProcPCA.pdf` (Acessado em: 10 de setembro de 2020). Aleš Hladnik Dr., Ass. Prof., Chair of Information and Graphic Arts Technology, Faculty of Natural Sciences and Engineering, University of Ljubljana, Slovenia ales. hladnik ntf. uni-lj. si 6 (2009): 16.

95 SALDANHA, Raphael de Freitas et al. **Estudo de análise de rede do fluxo de pacientes de câncer de mama no Brasil entre 2014 e 2016**. Cadernos de Saúde Pública, v. 35, p. e00090918, 2019.

96 SAMUEL, Arthur L. **Some studies in machine learning using the game of checkers**. IBM Journal of research and development, v. 3, n. 3, p. 210-229, 1959.

97 SANTOS, Pedro; Maudes, Jesús; Bustillo, Andres. **Identifying maximum imbalance in datasets for fault diagnosis of gearboxes**. Journal of Intelligent Manufacturing, v. 29, n. 2, p. 333-351, 2018.

98 S AO PAULO. **COORDENADORIAS REGIONAIS de SAÚDE DO MUNICÍPIO de SÃO PAULO | Secretaria Municipal Da Saúde | Prefeitura Da Cidade de São Paulo**. Disponível em: `https://www.prefeitura.sp.gov.br/cidade/secretarias/saude/organizacao/index.php?p=228575`. Acessado em: 03 de novembro de 2021.

99 S AO PAULO. **Escolaridade E Renda | Secretaria Municipal Da Saúde | Prefeitura Da Cidade de São Paulo**. Disponível em: `https://www.prefeitura.sp.gov.br/cidade/secretarias/saude/epidemiologia_e_informacao/geoprocessamento_e_informacoes_socioambientais/index.php?p=265358`. Acessado em: 03 de novembro de 2021.

100 SATHYA, Ramadass et al. **Comparison of supervised and unsupervised learning algorithms for pattern classification**. International Journal of Advanced Research in Artificial Intelligence, v. 2, n. 2, p. 34-38, 2013.

101 SHLENS, Jonathon. "**A tutorial on principal component analysis.**" arXiv preprint arXiv:1404.1100 (2014).

102 SINGH, Archana, Avantika Yadav, and Ajay Rana. "**K-means with Three different Distance Metrics.**" International Journal of Computer Applications 67.10 (2013).

103 SILVA, L. M. **Uma aplicação de Árvores de Decisão, Redes Neurais e KNN para a Identificação de Modelos ARMA não Sazonais e Sazonais**. Rio de Janeiro. 145p. Tese de Doutorado-Departamento de Engenharia Elétrica, Pontifícia Universidade Católica do Rio de Janeiro, 2005.

104 SILVA, Mayara et al. **Análise dos Atendimentos de Gestantes na Rede de Atenção Básica de Saúde no Município de São Paulo.** In: Anais do XX Simpósio Brasileiro de Computação Aplicada à Saúde. SBC, 2020. p. 250-261.

105 **SISTEMA ÚNICO DE SAÚDE (SUS)**. Disponível em: `http://www.saude.gov.br/sistema-unico-de-saude#principios`. Acessado em: 14 de maio de 2021.

106 SKIENA, Steven S. **The data science design manual**. Springer, 2017.

107 SOUSA, M. S., M. L. Q. Mattoso, and N. F. F. Ebecken. "**Data mining: a database perspective.**" WIT Transactions on Information and Communication Technologies 22 (1970).

108 SOUZA, F. S.; SILVA, L. M. F. R.; ROVERI, E. **Desenvolvimento de um sistema para o gerenciamento das internações e fluxo de pacientes entre hospitais e cidades de uma região.** In: Anais do XI Congresso Brasileiro de Informática em Saúde. Campos do Jordão: Universidade Federal de Minas Gerais. 2008. p. 1-6.

109 SUS, D. O. **Para entender a gestão do SUS**. Coleção Progestores, 2015.

110 SYAKUR, M. A., et al. "**Integration k-means clustering method and elbow method for identification of the best customer profile cluster.**" IOP Conference Series: Materials Science and Engineering. Vol. 336. No. 1. IOP Publishing, 2018.

111 TATIRAJU, Suman, and Avi Mehta. "**Image Segmentation using k-means clustering, EM and Normalized Cuts.**" Department of EECS 1 (2008): 1-7.

112 THORWARTH, Michael, and Amr Arisha. "**A simulation-based decision support system to model complex demand driven healthcare facilities.**" Proceedings of the 2012 Winter Simulation Conference (WSC). IEEE, 2012.

113 THORWARTH, Michael; ARISHA, Amr; HARPER, Paul. **Simulation model to investigate flexible workload management for healthcare and servicescape environment**. In: Proceedings of the 2009 Winter Simulation Conference (WSC). IEEE, 2009. p. 1946-1956.

114 TREVISAN, Maria do Rosário et al. **Perfil da assistência pré-natal entre usuárias do Sistema Único de Saúde em Caxias do Sul**. Revista Brasileira de Ginecologia e Obstetrícia, v. 24, p. 293-299, 2002.

115 VAN DYNE, M. M., et al. "**Using machine learning and expert systems to predict preterm delivery in pregnant women.**" Proceedings of the Tenth Conference on Artificial Intelligence for Applications. IEEE, 1994.

116 VARELLA, Carlos Alberto Alves. **Análise de componentes principais**. Seropédica: Universidade Federal Rural do Rio de Janeiro, 2008.

117 VASCONCELOS, Simone. "**Análise de Componentes Principais (PCA)**". Disponível em: http://www.ic.uff.br/ aconci/PCA-ACP.pdf. Acesso em: 12 de abril de 2021.

118 VIELLAS, Elaine Fernandes et al. **Assistência pré-natal no Brasil**. Cadernos de Saúde Pública, v. 30, p. S85-S100, 2014.

119 WADDELL, Michael; PAGE, David; SHAUGHNESSY JR, John. **Predicting cancer susceptibility from single-nucleotide polymorphism data: a case study in multiple myeloma**. In: Proceedings of the 5th international workshop on Bioinformatics. 2005. p. 21-28.

120 WALTERS-WILLIAMS, J., & Li, Y. (2010). **Comparative study of distance functions for nearest neighbors**. In Advanced techniques in computing sciences and software engineering (pp. 79-84). Springer, Dordrecht.

121 WANG, Xi-Zhao; DONG, Ling-Cai; YAN, Jian-Hui. **Maximum ambiguity-based sample selection in fuzzy decision tree induction**. IEEE Transactions on Knowledge and Data Engineering, v. 24, n. 8, p. 1491-1505, 2011.

122 WIKIPÉDIA. "**Decision tree learning**" (2021). Disponível em: `https://en.wikipedia.org/wiki/Decision_tree_learning`. Acessado em: 21 de janeiro de 2021.

123 WIKIPÉDIA. **"Anamnese (saúde)"** (2021). Disponível em: `https://pt.wikipedia.org/wiki/Anamnese_(sa%C3%BAde)`. Acessado em: 02 de janeiro de 2022.

124 WOLD, Svante, Kim Esbensen, and Paul Geladi. "**Principal component analysis.**" Chemometrics and intelligent laboratory systems 2.1-3 (1987): 37-52.

125 XU, JinHua, and Hong Liu. "**Web user clustering analysis based on KMeans algorithm.**" 2010 International Conference on Information, Networking and Automation (ICINA). 2010.

126 YEN, Show-Jane, and Yue-Shi Lee. "**Under-sampling approaches for improving prediction of the minority class in an imbalanced dataset.**" Intelligent Control and Automation. Springer, Berlin, Heidelberg, 2006. 731-740.

127 YUAN, Chunhui, and Haitao Yang. "**Research on K-value selection method of K-means clustering algorithm.**" J—Multidisciplinary Scientific Journal 2.2 (2019): 226-235.

APÊNDICE A – Lista de CIDs Exclusivos de Gestantes

Código CID	Nome CID
O244	Diabetes Mellitus Que Surge Durante A Gravidez
O249	Diabetes Mellitus Na Gravidez, Não Especificado
O00	Gravidez Ectópica
O000	Gravidez Abdominal
O008	Hipertensão Pré-Existente Não Especificada, Complicando A Gravidez, O Parto E O Puerpério
O008	Outras Formas De Gravidez Ectópica
O009	Gravidez Ectópica, Não Especificada
O009	Outras Doenças Da Mãe, Classificadas Em Outra Parte, Mas Que Complicam A Gravidez O Parto E O Puerpério
O03	Aborto Espontâneo
O036	Parto Espontâneo Cefálico
O036	Parto Múltiplo
O059	Outros Tipos De Aborto - Completo Ou Não Especificado, Sem Complicações
O10	Hipertensão Pre-Existente Complicando A Gravidez, O Parto E O Puerpério
O102	Falso Trabalho De Parto Na 37ª Semana Completa Ou Depois Dela
O104	Doença Renal Hipertensiva Pré-Existente Complicando A Gravidez, O Parto E O Puerpério
O109	Outras Formas De Gravidez Ectópica
O109	Falso Trabalho De Parto, Não Especificado
O13	Hipertensão Gestacional (Induzida Pela Gravidez) Sem Proteinuria Significativa
O14	Hipertensão Gestacional [Induzida Pela Gravidez] Com Proteinuria Significativa
O20	Resultado Do Parto
O20	Hemorragia Do Inicio Da Gravidez
O209	Hemorragia Do Início Da Gravidez, Não Especificada
O21	Vômitos Excessivos Na Gravidez
O218	Supervisão De Gravidez Com História De Assistência Pré-Natal Insuficiente
O218	Infecções Do Rim Na Gravidez
O219	Vômitos Da Gravidez, Não Especificados
O23	Infecções Do Trato Geniturinário Na Gravidez
O230	Outras Infecções Dos Órgãos Genitais Subsequentes Ao Parto
O231	Infecções Da Bexiga Na Gravidez
O234	Supervisão De Gravidez De Alto Risco Devido A Problemas Sociais
O239	Complicações Do Puerpério não Classificadas Em Outra Parte
O24	Diabetes Mellitus Na Gravidez
O25	Assistência À Gravidez Por Motivo De Abortamento Habitual
O26	Assistência Materna Por Outras Complicações Ligadas Predominantemente A Gravidez
O260	Ganho Excessivo De Peso Na Gravidez
O262	Desnutrição Na Gravidez
O268	Outras Afecções Especificadas, Ligadas A Gravidez
O29	Complicações De Anestesia Administrada Durante A Gravidez
O30	Gestação Múltipla
O300	Complicação Do Puerpério Não Especificada
O309	Falso Trabalho De Parto Antes De Se Completarem 37 Semanas De Gestação
O367	Assistência Prestada À Mãe Por Feto Viável Em Gravidez Abdominal
O47	Falso Trabalho De Parto
O470	Gestação Múltipla, Não Especificada
O471	Infecções Do Rim Na Gravidez
O479	Falso Trabalho De Parto, Não Especificado
O48	Gravidez Prolongada
O60	Parto Pré-Termo
O600	Trabalho De Parto Pré-Termo Sem Parto
O623	Trabalho De Parto Precipitado
O67	Doença Hemolítica Do Feto E Do Recém-Nascido
O68	Trabalho De Parto E Parto Complicados Por Sofrimento Fetal
O72	Aborto Espontâneo - Completo Ou Não Espec., Complicado P/ Hemor. Excessiva Ou Tardia
O746	Outras Infecções Dos Órgãos Genitais Subsequentes Ao Parto
O80	Parto Único Espontâneo
O800	Hemorragia Pós-Parto
O82	Parto Único Por Cesariana
O84	Anemia Complicando A Gravidez, O Parto E O Puerpério
O861	Complicações De Anestesia Administrada Durante A Gravidez
O90	Supervisão De Gravidez Com Grande Multiparidade

Código CID	Nome CID
O909	Gravidez Dupla
O912	Mastite Não Purulenta Associada Ao Parto
O92	Supervisão De Gravidez Com Outros Antecedentes De Procriação Problemática
O98	Doenças Infecciosas E Parasitarias Maternas Classificáveis Em Outra Parte Mas Que Compliquem A Gravidez, O Parto E O Puerpério
O981	Sífilis Complicando A Gravidez, O Parto E O Puerpério
O99	Outras Doenças Da Mãe, Classificadas Em Outra Parte, Mas Que Complicam A Gravidez O Parto E O Puerpério
O990	Trabalho De Parto E Parto Complicados Por Hemorragia Intra parto Não Classificados Em Outra Parte
P55	Anemia Complicando A Gravidez, O Parto E O Puerpério
P95	Sífilis Complicando A Gravidez, O Parto E O Puerpério
Z32	Exame Ou Teste De Gravidez
Z321	Gravidez Confirmada
Z33	Gravidez Como Achado Casual
Z34	Supervisão De Gravidez Normal
Z340	Supervisão De Primeira Gravidez Normal
Z348	Supervisão De Outra Gravidez Normal
Z349	Supervisão De Gravidez Normal, Não Especificada
Z35	Supervisão De Gravidez De Alto Risco
Z350	Supervisão De Gravidez Com História De Esterilidade
Z351	Supervisão De Gravidez Com História De Aborto
Z352	Supervisão De Gravidez Com Outros Antecedentes De Procriação Problemática
Z353	Outras Formas De Vômitos Complicando A Gravidez
Z354	Outras Infecções E As Não Especificadas Do Trato Urinário Na Gravidez
Z357	Infecção Não Especificada Do Trato Urinário Na Gravidez
Z358	Supervisão De Outras Gravidezes De Alto Risco
Z359	Supervisão Não Especificada De Gravidez De Alto Risco
Z37	Hemorragia Do Inicio Da Gravidez

APÊNDICE B – Lista Com os Procedimentos Mais Frequentes

	Procedimentos/Diagnósticos
1	Consulta Pré-Natal
2	Consulta Medica Em Atenção Básica
3	Consulta De Profissionais De Nível Superior Na Atenção Básica (Exceto Médico)
4	Atendimento De Urgência Em Atenção Básica
5	Visita Domiciliar Por Profissional De Nível Médio
6	Visita Domiciliar Por Profissional De Nível Médio
7	Consulta Medica Em Atenção Especializada
8	Coleta De Material P/ Exame Laboratorial
9	Aferição De Pressão Arterial
10	Coleta De Material P/ Exame Citopatológico De Colo Uterino
11	Consulta Puerperal
12	Teste Rápido De Gravidez
13	Administração De Medicamentos Em Atenção Básica (Por Paciente)
14	Consulta/Atendimento Domiciliar Na Atenção Básica
15	Exame Ginecológico (Geral) (De Rotina)
16	Consulta De Profissionais De Nível Superior Na Atenção Especializada (Exceto Médico)
17	Exame Médico Geral
18	Tratamento Inicial Na Atenção Básica
19	Exame Geral E Investigação De Pessoas Sem Queixas Ou Diagnóstico Relatado
20	Assistência Domiciliar Por Profissional De Nível Médio
21	Atendimento De Urgência Em Atenção Básica Com Observação Até 8 Horas
22	Atendimento De Urgência Em Atenção Especializada
23	Cefaleia
24	Infecção Aguda Das Vias Aéreas Superiores Não Especificada
25	Tratamento Concluído Na Atenção Básica
26	Seguimento Pós-Parto De Rotina
27	Ultra-Sonografia Transvaginal
28	Outros Transtornos Do Trato Urinário
29	Diarreia E Gastroenterite De Origem Infecciosa Presumível
30	Outros Sintomas E Sinais Gerais Especificados
31	Retirada De Pontos De Cirurgias Básicas (Por Paciente)
32	Ultra-Sonografia Doppler De Fluxo Obstétrico
33	Ultra-Sonografia Obstétrica Morfológica
34	Ação Coletiva De Exame Bucal Com Finalidade Epidemiológica
35	Glicemia Capilar
36	Hipertensão Essencial (Primária)
37	Atendimento De Urgência Em Atenção Básica Com Remoção
38	Instrumentação De Procedimentos Clínicos
39	Dor Abdominal E Pélvica
40	Assistência E Exame Pós-Natal
41	Acolhimento
42	Infecção Do Trato Urinário De Localização Não Especificada
43	Gravidez Como Achado Casual
44	Gravidez Confirmada
45	Acolhimento Com Classificação De Risco
46	Avaliação Antropométrica
47	Náusea E Vômitos
48	Rastreamento ("Screening") Pre-Natal
49	Causas Desconhecidas E Não Especificadas De Morbidade
50	Eletrocardiograma
51	Anticoncepção
52	Acolhimento Na Atenção Básica Nível Superior
53	Aconselhamento Geral Sobre Contracepção