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

Jorge Rafael Hara Moreira

Classificação de Risco Neonatal Usando Aprendizado de Máquina e Dados dos Sistemas de Informação de Saúde Pública e de Censo Demográfico Brasileiro

Juiz de Fora
2023

Jorge Rafael Hara Moreira

Classificação de Risco Neonatal Usando Aprendizado de Máquina e Dados dos Sistemas de Informação de Saúde Pública e de Censo Demográfico Brasileiro

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

Coorientador: Dr. Heder Soares Bernardino

Juiz de Fora

2023

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

Moreira, Jorge Rafael Hara.
Classificação de Risco Neonatal Usando Aprendizado de Máquina e Dados dos Sistemas de Informação de Saúde Pública e de Censo Demográfico Brasileiro / Jorge Rafael Hara Moreira. – 2023.
92 f. : il.

Orientador: Alex Borges Vieira
Coorientador: Heder Soares Bernardino
Dissertação (Mestrado) – Universidade Federal de Juiz de Fora, Instituto de Ciências Exatas. Programa de Pós-Graduação em Ciência da Computação, 2023.

1. Mineração de dados. 2. Aprendizado de máquina. 3. Óbito neonatal. 4. Sistemas de informação em saúde. 5. Vinculação de registros. I. Vieira, Alex Borges, orient. II. Bernardino, Heder Soares, coorient. III. Título.

Jorge Rafael Hara Moreira

Classificação de Risco Neonatal Usando Aprendizado de Máquina e Dados dos Sistemas de Informação de Saúde Pública e de Censo Demográfico Brasileiro

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. Área de concentração: Ciência da Computação.

Aprovada em 13 de abril de 2023.

BANCA EXAMINADORA

Prof. Dr. Alex Borges Vieira - Orientador

Universidade Federal de Juiz de Fora

Prof. Dr. Heder Soares Bernardino - Coorientador

Universidade Federal de Juiz de Fora

Profª. Dra. Priscila Capriles Goliatt

Universidade Federal de Juiz de Fora

Prof. Dr. Eduardo Krempser da Silva

Fundação Oswaldo Cruz

Juiz de Fora, 27/03/2023.

Documento assinado eletronicamente por **Alex Borges Vieira**, **Coordenador(a) em exercício**, em 19/04/2023, às 12:05, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Heder Soares Bernardino**, **Professor(a)**, em 19/04/2023, às 14:49, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Jorge Rafael Hara Moreira**, **Usuário Externo**, em 10/05/2023, às 22:55, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Eduardo Krempser da Silva**, **Usuário Externo**, em 03/07/2023, às 08:55, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Priscila Vanessa Zabala Capriles Goliatt**, **Professor(a)**, em 13/07/2023, às 09:35, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

A autenticidade deste documento pode ser conferida no Portal do SEI-Ufjf (www2.ufjf.br/SEI) através do ícone Conferência de Documentos, informando o código verificador **1204324** e o código CRC **F968B122**.

Dedico este trabalho aos meus pais e a minha esposa pelo apoio incondicional durante todo o curso de mestrado.

AGRADECIMENTOS

À Deus, por ter colocado na minha vida pessoas especiais que me apoiam e me dão força para enfrentar todos os desafios e continuar batalhando pelos meus sonhos.

Aos meus pais Severino e Samira pelo amor incondicional e por tudo que me proporcionaram durante toda a minha vida me dando condições e oportunidades de estudos sempre com muita luta e trabalho árduo ao longo dos anos. Gratidão eterna!

A minha esposa Ana Carolina pelo amor, companheirismo, apoio e compreensão durante todo o curso de mestrado.

Aos meus avós (*in memorian*) por todo carinho e ensinamentos.

Aos professores Alex Borges e Heder Bernardino pelas orientações ao longo de todo o mestrado, com sabedoria, dedicação e profissionalismo. Foi um privilégio estar ao lado dos dois.

Aos meus amigos Mayara, Airton, Genilson, Frederico, Pedro e a todos aqueles que me apoiaram e incentivaram durante toda a jornada do mestrado.

À UFJF, em especial ao PPGCC, pela oportunidade, apoio e estrutura fornecida durante o curso. Agradeço também a todos os professores e servidores que contribuíram para a minha formação.

"The new wave of artificial intelligence does not actually bring us intelligence but instead a critical component of intelligence - prediction."(Agrawal et al. Prediction Machines: The Simple Economics of Artificial Intelligence, 2018).

RESUMO

A taxa de mortalidade infantil é considerada um dos indicadores mais importantes de uma sociedade. A partir dela, é possível retratar as condições de vida de um determinado país ou região, onde a presença de índices elevados pode refletir baixo nível de desenvolvimento social e econômico da população avaliada. O Brasil, apesar da melhora nos últimos anos, ainda enfrenta desafios para reduzir esse índice, principalmente em algumas regiões do país que carecem de infraestrutura e enfrentam problemas socioeconômicos mais graves. Inserida na composição da mortalidade infantil, estudos mostram que a maior parcela dos óbitos são provenientes do componente neonatal, exigindo maior atenção do poder público neste segmento. Assim, predizer o risco de um bebê morrer nos seus primeiros dias de vida pode gerar impactos positivos ao sistema de saúde público e, consequentemente, à sociedade brasileira. Este trabalho utiliza técnicas de aprendizado de máquina e dados dos sistemas de informação em saúde e de censo demográfico brasileiro para gerar classificadores que possibilitam emitir um alerta ao sistema de saúde em caso de risco de óbito neonatal, direcionando a atenção ao acompanhamento materno e ao recém-nascido. Os classificadores foram treinados usando um grande conjunto de dados formado a partir de um processo de vinculação de registros de cinco bases de dados distintas. Além disso, o conjunto de dados foi enriquecido com novas variáveis permitindo um estudo mais amplo dos fatores influenciadores da mortalidade neonatal. Os resultados revelam taxas de *f1-score* e AUC acima de 89% na classificação de risco de óbito neonatal, mostrando que o classificador proposto é viável como mais um recurso para prevenção de óbitos neonatais e aprimoramento do sistema de saúde.

Palavras-chave: Mortalidade infantil. Óbito neonatal. Aprendizado de máquina. Classificação. Sistemas de informação de saúde. Censo demográfico. Vinculação de registros.

ABSTRACT

The infant mortality rate is considered one of the most critical indicators of society. From it, it is possible to portray the living conditions of a given country or region, where high indices may reflect a low social and economic development level of the assessed population. Despite the improvement in recent years, Brazil still needs to improve in reducing this index, especially in some regions of the country that lack infrastructure and face more severe socioeconomic problems. Inserted in the composition of infant mortality, studies show that the largest share of deaths come from the neonatal component, requiring greater attention from the public authorities in this segment. Thus, predicting the risk of a baby dying in its first days of life can positively impact the public health system and, consequently, Brazilian society. This work uses machine learning techniques and data from health information systems and the Brazilian demographic census to generate classifiers that make it possible to issue an alert to the health system in case of risk of neonatal death, directing attention to maternal and newborn follow-up. The classifiers were trained using a large dataset by linking records from five databases. In addition, the data set was enriched with new variables allowing a broader study of the factors that influence neonatal mortality. The results reveal f1-score and AUC rates above 89% in the neonatal death risk classification, showing that the proposed classifier is viable as one more resource for preventing neonatal deaths and improving the health system.

Keywords: Infant mortality. Neonatal death. Machine learning. Classification. Health information systems. Demographic census. Record linkage.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

AdaBoost	*Adaptive Boosting*
AUC	*Area Under the Curve*
CID	Classificação Estatística Internacional de Doenças e Problemas Relacionados com a Saúde
DATASUS	Departamento de Informática do Sistema Único de Saúde
ETL	*Extract, Transform and Load*
IA	Inteligência Artificial
IBGE	Instituto Brasileiro de Geografia e Estatística
IDHM	Índice de Desenvolvimento Humano Municipal
IPEA	Instituto de Pesquisa Econômica Aplicada
KDD	*Knowledge Discovery in Databases*
ODS	Objetivos de Desenvolvimento Sustentável
OMS	Organização Mundial de Saúde
PCDas	Plataforma de Ciência de Dados aplicada à Saúde
PNUD	Programa das Nações Unidas para o Desenvolvimento
RF	*Random Forest*
SIGTAP	Sistema de Gerenciamento da Tabela de Procedimentos, Medicamentos, Órteses, Próteses e Materiais Especiais do Sistema Único de Saúde
SIM	Sistema de Informação sobre Mortalidade
SINASC	Sistema de Informação sobre Nascidos Vivos
ROC	*Receiver Operating Characteristic Curve*
SHAP	*SHapley Additive exPlanations*
SIS	Sistema de Informação em Saúde
SMOTE	*Synthetic Minority Oversampling Technique*
SUS	Sistema Único de Saúde
TMI	Taxa de Mortalidade Infantil
TMN	Taxa de Mortalidade Neonatal
WHO	*World Health Organization*
XGBoost	*Extreme Gradient Boosting*

SUMÁRIO

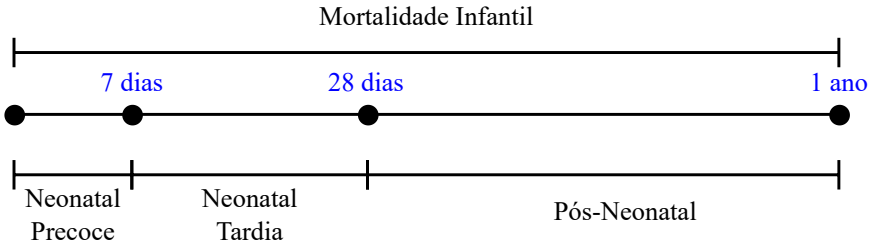
1 INTRODUÇÃO

A mortalidade infantil ainda é um grave problema de saúde pública mundial evidenciada principalmente nos países mais pobres e com baixo desenvolvimento humano (7). Ela é medida por um índice denominado de taxa de mortalidade infantil (TMI) que mede o risco que um nascido vivo tem de morrer antes de completar um ano de vida, sendo calculada pela razão entre o número de óbitos infantis com menos de um ano de vida, e a quantidade de nascidos vivos durante o ano (em determinado limite geográfico), multiplicados por mil (53, 20, 99).

A TMI é considerada um dos indicadores de saúde mais importantes de uma sociedade (53, 19). Ela é sensível à avaliação das condições socioeconômicas e assistenciais em saúde, onde podem ser sinalizados problemas como medidas sanitárias inadequadas e serviços de saúde pouco especializados e deficientes. Desta forma, a TMI permite analisar a disponibilidade, a utilização e a eficácia dos cuidados de saúde de uma região, sendo frequentemente utilizada para definir políticas públicas direcionadas à saúde materno-infantil (56).

Conforme descrito em (16), a mortalidade infantil é subdividida em dois períodos ou componentes etários: I. Neonatal, que corresponde às mortes de crianças com menos de 28 dias de vida; e II. pós-neonatal, que representam os óbitos de crianças de 28 dias a menores de um ano de idade. No escopo do óbito neonatal, é possível subdividi-lo em precoce, quando o óbito ocorre entre 0 e 6 dias completos; e tardia, quando o óbito ocorre entre 7 e 27 dias completos . A Figura 1 apresenta um esquema para visualização desses componentes.

Figura 1 – Componentes da Mortalidade Infantil.



Fonte: Elaborado pelo autor (2022).

Considera-se que o óbito no período neonatal é resultante de uma estreita relação de fatores socioeconômicos, assistenciais e biológicos (53, 4). As causas relacionadas ao óbito no primeiro ano de vida também podem estar associadas à prematuridade, às anomalias congênitas, à asfixia no parto, à sepse neonatal e à desnutrição (5). Diante dessas condições e ocorrência de casos ainda evidentes em alguns países e regiões do mundo, considera-se a

mortalidade neonatal como um problema relevante de saúde pública, pois o primeiro mês de vida é o período mais vulnerável para a sobrevivência de um bebê (4).

Técnicas de mineração de dados e aprendizado de máquina têm se mostrado potenciais ferramentas de apoio às mais variadas áreas de pesquisa (29, 78). De acordo com (11), houve uma discussão significativa nos últimos anos sobre como o aprendizado de máquina pode ser aplicado nos diversos setores da sociedade. Para (10), na epidemiologia e saúde pública, a aplicação desses métodos podem potencializar os resultados e gerar novos conhecimentos (10). Mais especificamente no estudo do óbito neonatal, observa-se um interesse da comunidade científica em aplicar o conhecimento de ciência de dados na produção de ferramentas de apoio à decisão e exploração dos fatores de riscos associados à este problema de saúde pública (45, 66, 89, 9).

Neste contexto, esta pesquisa pauta-se no uso de mineração de dados, aprendizado de máquina e grandes conjuntos de dados de saúde e censo demográfico para gerar conhecimento e recursos que possam auxiliar aos profissionais e gestores de saúde na tomada de decisão em caso de encaminhamento do recém-nascido para cuidados especiais. Desta forma, pretende-se com este trabalho: I. Formar um conjunto de dados enriquecido com variáveis socioeconômicas, assistenciais, ligadas ao parto, biológicas maternas e do recém-nascido que permitam uma análise ampla dos fatores de risco do óbito neonatal e que possam contribuir para o desfecho preditivo do risco neonatal; e II. gerar classificadores com a capacidade de identificar o risco de óbito neonatal e, com isso, possibilitar intervenções precoces e adequadas para prevenir o risco de mortalidade neonatal. Assim, busca-se disponibilizar publicamente um grande conjunto de dados compilado com diversas variáveis preditoras do risco de óbito neonatal, possibilitando aplicação em novos estudos e, além disso, propor um classificador capaz de antever o risco de óbito neonatal com alta sensibilidade e precisão, servindo como mais um recurso ao sistema de saúde e aos profissionais envolvidos na tomada de decisão para possíveis encaminhamentos e cuidados especializados ao bebê após o nascimento.

Ademais, o estudo também investiga os fatores de risco associados ao óbito neonatal através do conjunto de dados construído, permitindo uma investigação mais ampla das variáveis relacionadas ao óbito neonatal. Desta maneira, este trabalho diferenciou-se de outros da literatura pois utiliza dados que vão além daqueles presentes nos dados públicos dos sistemas de informação de saúde do SUS que são comumente usados em pesquisas que envolvem óbito infantil ou neonatal. Destaca-se que o uso do conjunto de dados enriquecido para treino e teste dos modelos ajudou a melhorar os resultados, além de revelar que as variáveis preditoras geradas, quais sejam o índice de desenvolvimento humano municipal (IDHM) e a distância entre residência da mãe e o local de nascimento, apresentam relevância na identificação do risco de óbito neonatal. Espera-se que as descobertas possam auxiliar no planejamento de ações para a reestruturação e melhoria da assistência à gestante e aos recém-nascidos visando à redução do óbito neonatal no escopo da realidade brasileira.

Os resultados revelam uma taxa de *f1-score* e área sob a curva ROC (AUC) acima de 89%, representando uma elevada taxa de acertos na classificação do risco de óbito neonatal com alta relação positiva entre a precisão e a sensibilidade dos classificadores gerados. Assim, o emprego das técnicas e abordagens metodológicas propostas neste trabalho se mostram viáveis como mais um recurso para aprimoramento do sistema de saúde e avanço no desenvolvimento de trabalhos nesta área de pesquisa.

Ressalta-se, que além deste capítulo introdutório são apresentadas no capítulo 2 as definições e aplicações dos conceitos mais importantes usados no trabalho. Em seguida, no capítulo 3, expõe-se uma análise dos trabalhos que trouxeram uma grande contribuição para o desenvolvimento do estudo. No capítulo 4, são apresentados os materiais e métodos utilizados na elaboração da pesquisa. Já no capítulo 5, os resultados alcançados e as discussões perante os trabalhos da literatura. Por fim, as considerações finais e propostas de trabalhos futuros no capítulo 6.

2 CONTEXTUALIZAÇÃO E EMBASAMENTO CONCEITUAL

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 realizada. Primeiro, na seção 2.1, são abordados os desafios da mortalidade infantil no Brasil. Em seguida, na seção 2.2, é descrito um breve histórico do Sistema Único de Saúde brasileiro. Na seção 2.3, apresentam-se o conceito de Sistema de Informação de Saúde e quais foram utilizados no trabalho. Por fim, na seção 2.4, expõe-se uma visão geral do modelo hierarquizado aplicado à investigação dos fatores de risco relacionados ao óbito neonatal.

2.1 DESAFIOS DA MORTALIDADE INFANTIL NO BRASIL

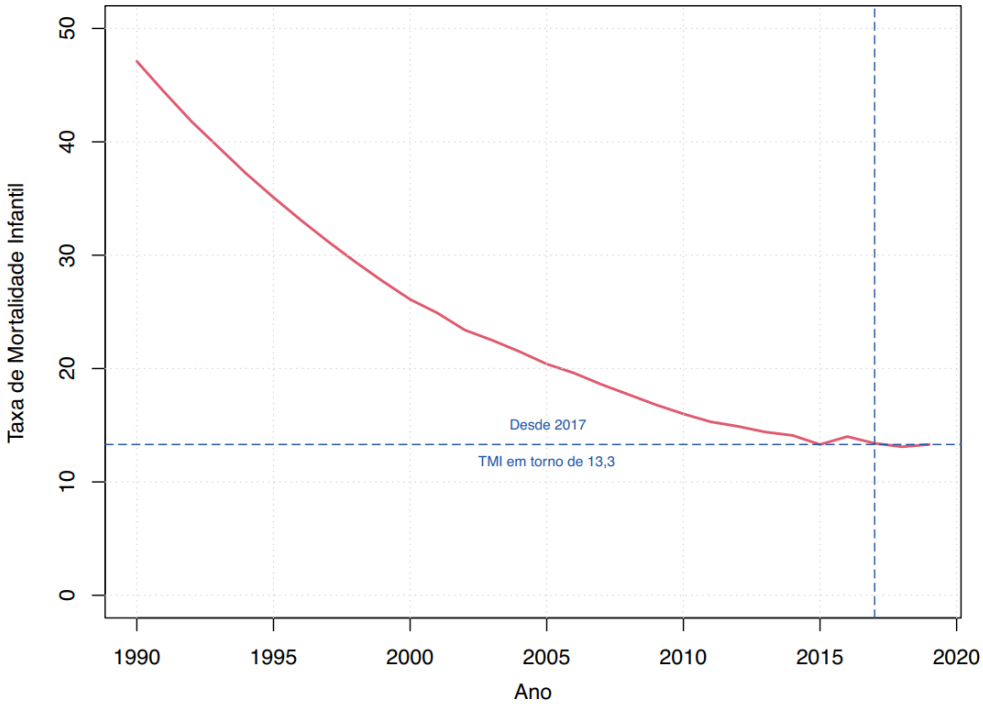
A mortalidade infantil é um grave problema de saúde pública mundial, evidenciado principalmente nos países mais pobres ou emergentes. As regiões com muita desigualdade social tendem a maiores índices de mortalidade de bebês no primeiro ano de vida, pois as mulheres e crianças geralmente enfrentam restrições ao acesso a cuidados essenciais e vitais (102).

A TMI do Brasil apresentou declínio no período de 1990 a 2015, passando de 47,1 para 13,3 o número de óbitos infantis por mil nascidos vivos (19). O índice chegou a aumentar para 14 em 2016, mas logo no ano seguinte, recuou ao patamar de 2015, perdurando até o ano de 2019. A Figura 2 mostra o histórico da TMI no Brasil entre os anos de 2000 a 2019.

Apesar do declínio das ocorrências de mortes no primeiro ano de vida nos últimos anos, o Brasil ainda enfrenta desafios para melhorar os índices de mortalidade infantil, principalmente em algumas regiões do país que não possuem infraestrutura de saúde adequada ou apresentam condições de vida desfavoráveis onde são encontrados ambientes com habitação precária contribuindo para um maior risco de doenças e complicações de saúde ao recém-nascido e à puérpera (14, 68). Além disso, muitas mulheres enfrentam o problema da peregrinação para o parto no Brasil, especialmente aquelas que vivem em áreas remotas ou regiões carentes de infraestrutura em saúde (109). A peregrinação para o parto refere-se ao deslocamento das mulheres em busca de serviços de saúde adequados para o acompanhamento da gestação e realização do parto. Esse cenário pode resultar em atrasos na assistência à saúde, complicações obstétricas e aumento do risco de mortalidade materna e neonatal.

Intimamente inserida neste problema de saúde pública, a mortalidade neonatal é considerado o principal componente desde a década de 1990 no Brasil (57). Neste sentido, o trabalho direcionou esforços para este importante componente da mortalidade infantil. Além disso, o incentivo à pesquisa nesta área foi motivada pela importância do tema em

Figura 2 – Óbitos infantis notificados ao Sistema de Informação Sobre Mortalidade (2000 a 2019).



Fonte: BRASIL. Ministério da Saúde (2021).

âmbito mundial, vide listagem dos Objetivos de Desenvolvimento Sustentável (ODS) da Agenda 2030[1] da ONU onde, dentre outros, prevê a redução dos índices de mortalidade neonatal e materna em todos os países.

No Brasil, a meta[2] é enfrentar as mortes evitáveis de recém-nascidos e crianças menores de cinco anos, objetivando reduzir a mortalidade neonatal para no máximo cinco por mil nascidos vivos e a mortalidade de crianças menores de cinco anos para no máximo oito por mil nascidos vivos. Sendo assim, o monitoramento da ocorrência dos óbitos infantis representa uma estratégia fundamental para avaliar a situação de saúde das populações e desenvolver medidas direcionadas à redução do risco de óbito no primeiro ano de vida.

Embora o Brasil tenha diminuído as disparidades sociais, econômicas e de indicadores de saúde nas últimas décadas, as diferenças intra e inter-regionais das taxas de mortalidade infantil persistem em todas as regiões do país (61). Por este motivo, torna-se relevante avançar nos estudos epidemiológicos e que envolvam saúde pública para melhorar as condições de vida da população brasileira.

[1] IBGE. Indicadores Brasileiros para os Objetivos de Desenvolvimento Sustentável. 2023. Disponível em: `https://odsbrasil.gov.br/home/agenda`.

[2] IPEA. Objetivos de Desenvolvimento Sustentável da ONU. Meta 3.2. 2019. Disponível em: `https://www.ipea.gov.br/ods/ods3.html`.

Assim, este trabalho vai ao encontro desta temática, pois ao reconhecer que a mortalidade neonatal é um indicador essencial que demonstra a qualidade de vida e desenvolvimento de uma população, torna-se fundamental buscar soluções que auxiliem na redução dos índices dessa relevante componente da mortalidade infantil.

2.2 SISTEMA ÚNICO DE SAÚDE

No Brasil vigora o Sistema Único de Saúde (SUS), considerado um dos maiores sistemas de saúde pública do mundo (2). Sua formação está pautada no art. 196 da Constituição Federal de 1988 sob o preceito que a saúde é um direito de todos e um dever do Estado. Desta forma, a União, os estados e os municípios são responsáveis pela saúde dos cidadãos brasileiros. Assim, a partir desse dispositivo da Constituição, o SUS foi criado sob a perspectiva de atender aos princípios da universalidade, integralidade, equidade, regionalização, descentralização e participação social[3].

O Ministério da Saúde define o SUS como um dos maiores e mais complexos sistemas de saúde pública do mundo garantindo acesso integral, universal e gratuito para toda a população brasileira (21). Desta maneira, a forma de organização a qual foi pautado o SUS busca adequá-lo à diversidade regional de um país continental como o Brasil, onde são encontradas, geralmente, realidades econômicas, sociais e sanitárias distintas entre as regiões[4].

Ademais, o SUS é composto por sistemas de informação em saúde que auxiliam na gestão dos processos nas áreas da epidemiologia e saúde pública no Brasil. No contexto deste trabalho, tornam-se ferramentas essenciais não só para compreensão e entendimento da mortalidade neonatal nas regiões brasileiras, como para o uso no treinamento e testes dos classificadores propostos. Os sistemas de informação em saúde utilizados neste trabalho são apresentados na próxima seção, sendo suas bases de dados relacionadas abordadas com mais detalhes no capítulo 4.

2.3 SISTEMAS DE INFORMAÇÕES EM SAÚDE

Um sistema de informação engloba um conjunto de componentes inter-relacionados, os quais possuem a capacidade de coletar, recuperar, processar, armazenar e distribuir informações pertinentes com o propósito de apoiar a tomada de decisões, a coordenação e o controle no contexto organizacional (58).

---

[3] BRASIL. Ministério da Saúde. Portaria nº 399, de 22 de fevereiro de 2006. Aprova a Política Nacional de Saúde. Diário Oficial da República Federativa do Brasil, Brasília, DF, 2006.

[4] BRASIL. Ministério da Saúde. Portaria nº 570, de 1º de junho de 2000. Instituir o Componente I do Programa de Humanização no Pré-natal e Nascimento - Incentivo à Assistência Pré-natal no âmbito do Sistema Único de Saúde. Brasília, DF, 2000.

No âmbito da saúde, os sistemas de informação, devem ser capazes de oferecer suporte à produção de informações para a melhor compreensão dos problemas e tomada de decisão no âmbito das políticas públicas e dos cuidados em saúde (110).

No Brasil, o Departamento de Informática do SUS (DATASUS) foi criado em 1991 com a finalidade de coordenar o desenvolvimento dos Sistemas de Informação de Saúde (SIS) para atender às necessidades do SUS, e com isso, estar en conformidade com as diretrizes constitucionais. Os SIS do SUS agregam diversos tipos de bases de dados de abrangência nacional com função e objetos de registro de diferentes naturezas, dentre eles cita-se três grandes sistemas muito utilizados na epidemiologia e gestão da saúde pública que são o Sistema de Informação sobre Mortalidade (SIM), o Sistema de Informação de Nascidos Vivos (SINASC) e o Cadastro Nacional de Estabelecimentos de Saúde (CNES).

O SIM tem a finalidade de reunir dados quantitativos e qualitativos sobre óbitos ocorridos no Brasil. É considerado uma importante ferramenta de gestão na área da saúde subsidiando a tomada de decisão em diversas áreas de assistência à saúde.

Com o propósito de quantificar nascidos vivos e fornecer informações sobre a gravidez, o parto e as condições da criança ao nascer, o SINASC produz informações que propiciam a construção de diagnóstico das condições de nascimento, possibilitando a realização de ações de promoção, prevenção e planejamento em saúde.

Já o CNES mantém o cadastro de todos os estabelecimentos de saúde do país, independentemente de sua natureza jurídica ou integração com o SUS. É uma ferramenta auxiliar, que proporciona o conhecimento da realidade da rede assistencial existente e suas potencialidades para apoiar o planejamento em saúde das três esferas de Governo, para uma gestão eficaz e eficiente.

Esses três sistemas de informação em saúde (SINASC, SIM e CNES) foram utilizados no trabalho associados a outros conjuntos de dados de cunho demográfico que serão apresentados na seção 4.1.

2.4 MODELO HIERARQUIZADO APLICADO À INVESTIGAÇÃO DE FATORES DE RISCO DE ÓBITO NEONATAL

Em 1984, Mosley e Chen propuseram um modelo teórico-conceitual que fornece uma estrutura analítica hierarquizada para o estudo da mortalidade infantil em países em desenvolvimento (76). Segundo este modelo os fatores socioeconômicos determinam comportamentos, os quais impactam um conjunto de fatores biológicos, que são os responsáveis diretos pela mortalidade no primeiro ano de vida. Esse modelo proporcionou uma base teórica sólida para o estudo da mortalidade infantil em países em desenvolvimento, ajudando a orientar pesquisas e políticas de saúde voltadas para a promoção da sobrevivência infantil.

O modelo hierárquico é utilizado como uma forma de entender a complexa interação entre os diferentes níveis de determinantes da mortalidade neonatal, desde fatores estruturais e socioeconômicos até fatores proximais relacionados ao recém-nascido. O modelo proposto trouxe um grande avanço para o desenvolvimento de políticas públicas, uma vez que informações oriundas de estudos que se limitam a apenas um grupo ou dimensão de fatores de risco resultam em recomendações inadequadas para avaliar os óbitos infantis, por apresentarem uma visão limitada do fenômeno (5).

Neste sentido, os fatores associados à mortalidade neonatal são complexamente articulados e influenciados pelas características biológicas maternas e do recém-nascido, pelas condições sociais e pelos cuidados prestados pelos serviços de saúde (5). Assim, têm-se como hipótese neste trabalho, que a aplicação de aprendizado de máquina usando conjunto de dados que abarquem diferentes características socioeconômicos, assistenciais e biológicos maternos e do recém-nascido, pode gerar modelos sensíveis na identificação precoce de riscos de óbito neonatal. Acredita-se também que a união de novas variáveis produzidas a partir dos dados já presentes nos sistemas de informação de saúde e fontes de dados demográficos brasileiros possam relevar algum novo conhecimento no contexto da mortalidade neonatal.

Trabalhos que abrangem estudos demográficos e epidemiológicos no Brasil que exploram a conexão entre fatores específicos relacionados à mortalidade infantil vem crescendo, porém, em sua maioria, utilizando modelos de regressão como em (35, 77, 60, 72, 44). Sendo assim, diante da oportunidade atual de exploração de grandes conjuntos de dados e maior poder computacional disponível, torna-se importante reconhecer a necessidade de explorar a utilização de ferramentas especializadas de mineração de dados e aprendizado de máquina para aumentar o poder dos estudos nessa área de pesquisa, assim como proposto neste trabalho.

3 TRABALHOS RELACIONADOS

Nota-se um interesse em estudos demográficos e epidemiológicos que aplicam tecnologias de *big data*, mineração de dados e aprendizado de máquina na classificação do risco de óbito neonatal (85, 30). A seguir são destacados os trabalhos que trouxeram maiores contribuições nessa área de pesquisa e que estão relacionados com o desenvolvimento deste trabalho.

Em (4), os autores investigam as características relacionadas ao risco de mortalidade neonatal no Brasil usando aprendizado de máquina. Os autores usam os conjuntos de dados públicos do SIM e SINASC entre os anos de 2006 e 2016 para formação de um grande conjunto de dados. Por meio de uma amostra de 30 milhões de registros, o trabalho gerou um classificador *Extreme Gradient Boosting* (XGBoost) para prever o risco de óbito neonatal com desempenho superior a 90% de área sob à curva ROC (AUC). O classificador foi usado também na investigação da importância das variáveis usadas no estudo. Os resultados apontaram as variáveis peso do recém-nascido, índice de Apgar de primeiro e quinto minuto, malformações congênitas, semanas gestacionais e número de consultas de pré-natal as seis características mais expressivas para a classificação do risco de óbito neonatal.

O estudo realizado em (90), desenvolveu um sistema de análise de saúde que utiliza mineração de dados para gerar alerta de risco de óbito de recém-nascidos. Os autores usam dados de óbitos e nascimentos do estado do Ceará registrados no SIM e SINASC entre 2013 e 2014. Dentre os métodos de aprendizado de máquina empregados no estudo, o método *naive bayes* se destacou com uma AUC de 92,1% e acurácia de 98,2% na classificação de risco de óbito de recém-nascidos.

Também em (98), os autores usaram dados dos Sistemas de Informação de Saúde do SUS (SIM e SINASC) para descoberta de conhecimento no contexto da mortalidade infantil. A diferença é que os pesquisadores aplicaram o algoritmo baseado em árvore C5.0 para gerar um classificador de risco de óbito infantil. Para tratar o desbalanceamento das classes usaram o *Synthetic Minority Oversampling Technique* (SMOTE) como técnica de sobreamostragem (*oversampling*). Usaram a estratégia de validação cruzada estratificada (K*-fold*) para divisão dos dados em treino e teste adotando o valor de K=10. Os experimentos se baseiam na média da execução de 30 vezes do C5.0, chegando a uma acurácia de 98,58% e AUC de 72,6%. Conforme o estudo, concluíram ainda que as três principais características relacionadas à morte de uma criança antes de completar o primeiro ano de vida são o peso ao nascer, o índice de Apgar de cinco minutos e as semanas de gestação.

Já em (9), os autores usaram dados dos nascidos vivos e óbitos do município de São Paulo no período de 2012 a 2017 para gerar um modelo usando o método XGBoost,

obtendo uma AUC de 97% e *f1-score* de 55% para prever a mortalidade neonatal.

O trabalho desenvolvido em (75) também gerou classificadores capazes de predizer o risco de óbito neonatal porém, além dos dados sistemas de informação do SUS, também usou os dados do censo demográfico brasileiro. As bases do SINASC, SIM, CNES e do índice de desenvolvimento humano municipal (IDHM) foram integradas para formar um grande conjunto de dados de nascidos vivos e óbitos de neonatais de 2012 a 2014. Ademais, os dados foram enriquecidos com as variáveis: distância entre residência da gestante e local de nascimento; e o IDHM de residência da gestante. Os resultados mostraram uma acurácia e sensibilidade na previsão de óbitos neonatais acima de 89%. Conforme o estudo, os classificadores regressão logística e os baseados em comitê (*random forest* e *AdaBoost*) se mostraram os mais eficientes na predição do óbito neonatal. O trabalho ainda mostrou que as variáveis peso ao nascer, índice de Apgar de um e cinco minutos, prematuridade, semanas de gestação, anomalia congênita, número de consultas, número de semanas de gestação e idade maternas são as variáveis que mais influenciam na predição do risco de óbito neonatal. O estudo de (75) compõem os resultados iniciais desta dissertação.

Este trabalho adota uma abordagem mais ampla para modelar um classificador de risco neonatal em relação aos estudos apresentados, pois abarca mais variáveis do que aquelas presentes nos sistemas de informação de saúde do SUS e que são comumente usadas nos estudos epidemiológicos com foco nos fatores do óbito neonatal. Os dados utilizados foram enriquecidos com possíveis fatores influenciadores do desfecho neonatal como o índice de desenvolvimento humano municipal (IDHM), presença e capacidade de centros obstétricos e pediátricos dos estabelecimentos de saúde e a distância entre a residência da mãe e o local de nascimento. Este cenário já se diferencia do trabalho apresentado em (4). E como proposta de extensão do estudo apresentado em (75) foi adicionada a variável denominada de classificação do recém-nascido, categorizada em: pequeno, adequado ou grande para a idade gestacional. Esta variável é importante para identificar problemas específicos que cada um desses grupos costuma apresentar (17).

Além disso, outro diferencial do trabalho em relação aos demais, foi realizar o estudo por grupo ou componente da mortalidade neonatal, onde foi possível verificar, separadamente, o desempenho dos classificadores e a relevância dos fatores de risco do óbito neonatal.

4 MATERIAIS E MÉTODOS

Neste capítulo, são apresentados os materiais e métodos utilizados no presente estudo. Na seção 4.1, é apresentada uma descrição completa dos dados e a metodologia usada em todo o processo de tratamento dos dados. Nessa seção ainda é abordada a qualidade e limitações dos dados utilizados no trabalho. Já na seção 4.2, são expostos os métodos e técnicas utilizadas para construção e avaliação dos classificadores gerados, onde são apresentados os métodos de aprendizado de máquina usados e as estratégias adotadas para divisão dos dados para treinamento, validação e teste. Nesta última, ainda são apresentadas as métricas adotadas para avaliação dos classificadores propostos.

4.1 CONJUNTO DE DADOS

Ao longo das últimas décadas, tem-se presenciado uma revolução tecnológica em várias áreas do conhecimento, impulsionada pelo avanço da capacidade computacional e, especialmente, pela disponibilidade de grandes conjuntos de dados (28, 49, 54). Na área da saúde pública a aplicação desse ferramental tem se mostrado cada vez mais relevante e promissora contribuindo para os avanços no ramo da vigilância epidemiológica e na formulação de políticas públicas (78, 107).

Assim como em (4), este trabalho considera a mortalidade neonatal um problema de saúde amplo que envolve a interação entre diferentes fatores determinantes de cunho socioeconômico, assistencial e relacionado ao aspecto biológico materno e do recém-nascido. Desta forma, para melhor entendimento das ocorrências de óbitos dos nascidos vivos no primeiro ano de vida, entendem-se que é necessário utilizar um grande volume de dados que englobe dados das gestantes, dos nascidos vivos e das ocorrências de óbitos da população brasileira. Ademais, faz-se necessária a garantia da qualidade dos dados para análise, treinamento e teste dos modelos, além do maior aproveitamento para o processo de vinculação de registros entre os conjuntos de dados utilizados.

O trabalho foi desenvolvido utilizando cinco conjuntos de dados: dois principais e três secundários. Os principais, formados pelos Sistemas de Informações sobre Nascidos Vivos (SINASC) e de Mortalidade (SIM); e os secundários, compostos pelos dados do Cadastro Nacional de Estabelecimentos de Saúde (CNES), Índice de Desenvolvimento Humano Municipal (IDHM)[5], e os dados dos municípios registrados pelo Instituto Brasileiro de Geografia e Estatística (IBGE)[6].

Os conjuntos de dados principais foram obtidos na Plataforma de Ciência de Dados da Fiocruz (PCDas), que reúne, dentre outros, os dados dos sistemas de nascidos

[5] PNUD. Atlas do Desenvolvimento Humano no Brasil. 2010. Disponível em: `http://www.atlasbrasil.org.br/`

[6] IBGE. Cidade e Estados do Brasil. 2022. Disponível em: `https://cidades.ibge.gov.br`

vivos, mortalidade e os cadastros dos estabelecimentos de saúde do SUS. A plataforma PCDas tem por objetivo disponibilizar serviços tecnológicos e computação científica para armazenamento, gestão e análise de grande volume de dados para pesquisadores, docentes e discentes de instituições de ensino e pesquisa[7]. Foi esta plataforma que balizou o trabalho e deu a sustentação necessária para formação do conjunto de dados com registros dos nascidos vivos e os casos de óbitos neonatais.

Em relação aos dados secundários, o CNES também foi obtido na PCDas e os demais extraídos de tabelas fornecidas nas plataformas eletrônicas de cada Instituição ou Agência de Pesquisa[8] e auxiliaram na integração dos dados, além de servirem como fonte de enriquecimento de dados para composição do conjunto de dados final. As variáveis, tipos e descrições presentes em todos os conjuntos de dados utilizados no trabalho são apresentados nas Tabelas 6, 7, 8, 9 e 10 no Apêndice A.

Ainda sobre os conjuntos de dados principais, a base de Nascidos Vivos da PCDas apresenta um total de 119 atributos com dados sobre os nascidos vivos no Brasil, que apresenta, dentre outros, dados importantes para o presente estudo como o peso ao nascer, idade gestacional, número de consultas pré-natal realizada e tipo de parto. Já o SIM, possui um total de 165 atributos contendo dados básicos da morte do indivíduo junto com as causas do óbito, contando com a Classificação Estatística Internacional de Doenças e Problemas Relacionados com a Saúde (CID) preenchida. Uma característica fundamental para integração dessas duas bases de dados foi a presença do atributo em comum denominado de declaração de nascido vivo que possibilitou a vinculação dos registros e a formação do conjunto de dados inicial. Todo o processo de vinculação dos registros dessas duas bases e demais conjunto de dados são abordados na seção 4.1.1.

Desda forma, foi gerado um grande conjunto de dados[9] com mais de 8 milhões de registros de nascidos vivos, no período de 2012 a 2014, com variáveis que podem influenciar o desfecho neonatal para os primeiros dias de vida do bebê. Com isso, possibilitou um estudo mais amplo dos fatores de risco do óbito neonatal, além de servir como principal insumo para treinamento e teste dos classificadores propostos. Todas as variáveis do conjunto de dados compilado cujos dados foram utilizados para treinar, validar e testar os classificadores são apresentadas no Apêndice B.

---

[7] FIOCRUZ. Plataforma de Ciência de Dados aplicada à Saúde (PCDaS). 2022. Disponível em: `https://pcdas.icict.fiocruz.br`

[8] IBGE. Tabela de Códigos de Municípios. 2022. Disponível em: `https://www.ibge.gov.br/explica/codigos-dos-municipios.php`
UNDP. Índice de Desenvolvimento Humano Municipal (2010). 2022. Disponível em: `https://www.br.undp.org/content/brazil/pt/home/idh0/rankings/idhm-municipios-2010.html`

[9] UFJF. Networks and Distributed Systems Laboratory (NetLab). Neonatal Death Prediction. 2022. Disponível em: `http://netlab.ice.ufjf.br/index.php/datasets/neonataldeathsprediction/`

4.1.1 Seleção, Pré-processamento e Formatação

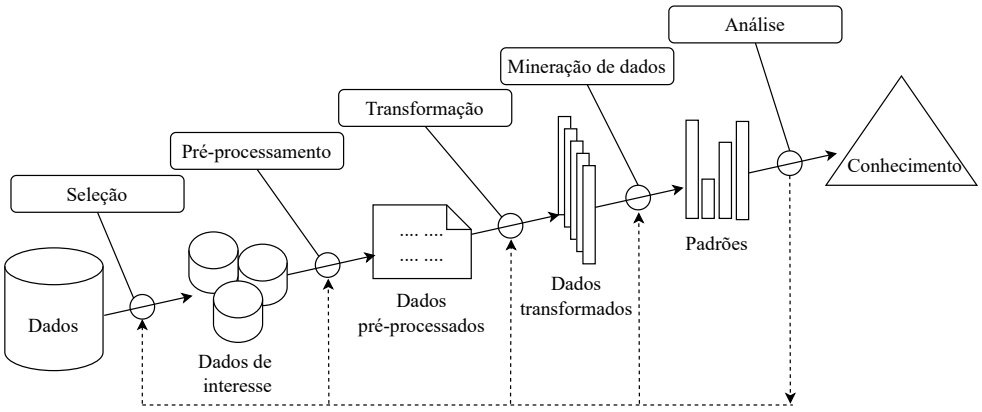
O conjunto de dados utilizado no trabalho foi proveniente da integração de cinco bases de dados distintas conforme mencionado na seção anterior. Para chegar no conjunto de dados compilado foram empregadas técnicas do processo de descoberta do conhecimento a partir dos dados. Este processo foi abordado em (36), sendo definido como:

> *[...] overall process of discovering useful knowledge from data. Data mining is a particular step in this process—application of specific algorithms for extracting patterns (models) from data. The additional steps in the KDD process, such as data preparation, data selection, data cleaning, incorporation of appropriate prior knowledge, and proper interpretation of the results of mining ensure that useful knowledge is derived from the data.* (Fayyad et al., 1996, p.28)

Trata-se de um processo bem estabelecido e comumente usado em trabalhos de mineração de dados que possuem o foco na descoberta de conhecimento em grandes conjuntos de dados. Na epidemiologia, o processo de descoberta de conhecimento em bancos de dados pode auxiliar na implementação de políticas públicas, como melhorias na assistência neonatal (8).

O processo de descoberta do conhecimento é composto por cinco etapas, conforme ilustrado na Figura 3. O processo inicia-se com o entendimento do domínio da aplicação e dos objetivos a serem atingidos. Em seguida, os dados de interesse são selecionados para posterior pré-processamento como tratamento de dados faltantes e inconsistentes. Na próxima etapa ocorre uma tarefa de formatação, onde há uma preparação dos dados como, por exemplo, transformação dos dados numéricos em dados categóricos. Após essas etapas são executadas tarefas de mineração de dados para posterior interpretação e avaliação dos classificadores gerados que serão abordadas no próximo capítulo.

Figura 3 – Etapas do processo de descoberta do conhecimento.



Fonte: Adaptada de FAYYAD et al.(1996).

Considera-se que as primeiras etapas do processo que envolvem a preparação e formatação dos dados são fundamentais para o sucesso do processo de descoberta do conhecimento em conjunto de dados. Porém, são as etapas que mais consomem tempo dentro de todo o fluxo KDD. Considera-se que as etapas de pré-processamento e formatação podem consumir até 80% do tempo necessário para todo o processo de descoberta do conhecimento (100). No desenvolvimento deste trabalho, a aplicação dessas etapas cadenciou a organização, formatação e padronização dos dados para posterior análise, exploração, treinamento de modelos e descoberta de conhecimento no escopo do estudo do óbito neonatal.

Após contextualizar a abordagem metodológica para formação do conjunto de dados compilado, seguem os passos aplicados em cada etapa do processo no contexto deste trabalho. Inicialmente, os conjuntos de dados principais (SINASC e SIM) foram coletados na plataforma de Ciência de Dados da Fiocruz (PCDas). Os dados são disponibilizados de forma tratada e enriquecida, usando uma metodologia própria de extração, transformação e carga (ETL) da Fiocruz a qual é realizada em cada um dos conjuntos de dados disponibilizados[10].

Apesar dos registros de declarações de nascidos vivos (SINASC) e óbitos (SIM) serem de 1996 a 2017, este trabalho optou por realizar os estudos utilizando uma janela temporal menor, abrangendo os anos de 2012 a 2014. Isso se deve a problemas na qualidade e limitações dos dados nos demais anos. O período utilizado possibilitou o uso de dados de maior qualidade e melhor representatividade dos grupos e classes geradas. A seção 4.1.2 aborda os problemas relacionados às bases de dados que não compõem a janela temporal utilizada no trabalho. Ressalta-se que embora a escolha tenha sido delimitada a dois anos de registros de nascidos vivos e óbitos neonatais, não inviabilizou a formação de um grande conjunto de dados representativo com mais de 8 milhões de registros.

Destaca-se ainda uma particularidade desses dois sistemas de informação de saúde é o fato deles não serem interligados, impedindo a realização de uma análise longitudinal com mais variáveis e centrada na gestante e/ou recém-nascido. Com isso, para realizar o filtro dos óbitos neonatais e posterior enriquecimento dos dados com atributos oriundos de outras bases de dados foi necessário realizar um processo de vinculação de registros. Esse processo, também chamado de *record linkage*, é muito utilizado quando há a necessidade de se ter uma visão mais ampla e integrada dos dados que estão presentes em fontes de dados distintas.

Neste trabalho foi utilizada a vinculação de registros sob pareamento determinístico, onde os pares de registros precisam ser iguais em um determinado conjunto de indexadores para que dois deles sejam considerados um par. No contexto das bases SINASC e SIM

[10] Exemplo do processo ETL da PCDas/Fiocruz aplicado ao conjunto de dados do Sistema de Informação sobre Nascidos Vivos. Disponível em: `https://pcdas.icict.fiocruz.br/conjunto-de-dados/sistema-de-informacao-sobre-nascidos-vivos/documentacao/`

o atributo possível de utilização para indexação e integração das bases é o Número de Declaração de Nascido Vivo (NUMERODN).

No intuito de abranger mais variáveis e aspectos das dimensões da mortalidade neonatal foram utilizados os conjuntos de dados secundários: CNES, IDHM e IBGE - assim como mencionado na seção 4.1. Como são bases que não tem integração com os conjuntos de dados principais e nem entre si, optou-se por realizar a vinculação dos registros também de forma determinística. No caso do CNES foi utilizado o código do estabelecimento para integração do conjunto de dados formados pelo SINASC e SIM; para integração do IDHM e IBGE foi utilizado o nome do município e código da unidade federativa; e por fim, o uso do código do município para integração à base de dados formada nas etapas anteriores.

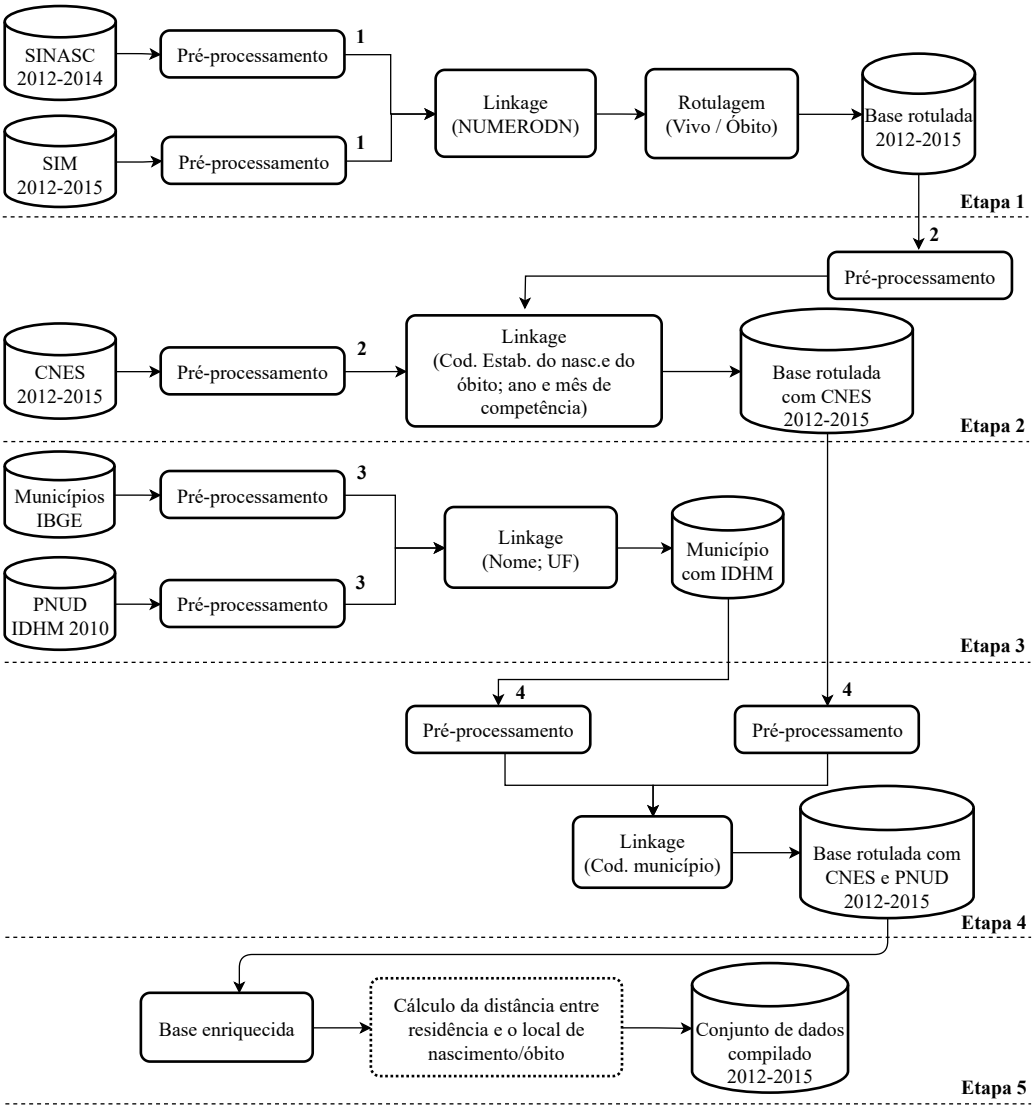
O processo de vinculação de todas as bases de dados para formação de um conjunto único de dados seguiu o fluxo presente no diagrama ilustrado na Figura 4. Conforme o diagrama, foram realizadas as seguintes etapas:

1. Vinculação dos registros de nascimento (SINASC) e óbitos neonatais (SIM) por meio do atributo NUMERODN.
2. Utilizando o atributo código do estabelecimento de saúde (CNES), bem como o mês e ano de competência do CNES, foi possível realizar a vinculação dos registros entre os conjuntos de dados. Com isso, formou-se uma base rotulada de nascidos vivos e óbitos neonatais com os respectivos dados do estabelecimento de saúde do nascimento;
3. Vinculação dos registros do IDHM com os municípios registrados na base do IBGE por meio da Unidade Federativa;
4. Vinculação dos registros dos conjuntos de dados gerados nas etapas 2 e 3 através do atributo código do município; e
5. Enriquecimento do conjunto de dados gerado adicionando um atributo que corresponde à distância entre a residência da gestante e o local de nascimento.

Destaca-se que além da ampliação do estudo com a adição de atributos presentes no CNES visando uma investigação das condições dos estabelecimentos de saúde no aspecto da obstetrícia, também foram criados três novos atributos: I. distância entre a residência da gestante e o local de nascimento; II. classificação do índice de desenvolvimento humano do município de residência da gestante; e III. classificação do recém-nascido conforme definido em (67). Este cenário permitiu uma investigação mais ampla dos fatores de riscos relacionados ao desfecho neonatal. Após o processo de vinculação das bases, os dados foram separados por classe e região, aplicando os seguintes procedimentos nos dados faltantes:

- **Categórico nominal e numérico inteiro:** preenchimento ou substituição do campo pelo valor mais frequente do atributo; e

Figura 4 – Etapas realizadas no processo de vinculação de registros.



1. Exclusão de dados ausentes e inconsistentes do atributo NUMERODN das bases SINASC e SIM.
2. Exclusão de dados omissos e inconsistentes dos atributos Código do Estabelecimento/CNES e; ano e mês de competência do CNES para vinculação de registros.
3. Adequação dos nomes, unidades federativas e códigos dos municípios para vinculação de registros entre as Bases do IBGE e do PNUD.
4. Adequação ou exclusão dos códigos dos municípios.
5. Enriquecimento de dados e formação do dataset final.

Fonte: Moreira et al. (2022).

- **Numérico contínuo:** preenchimento ou substituição do campo pela média dos valores do atributo.

Foi observado que o atributo idade materna apresentou valores inteiros diversos e que, supostamente, não condizem com a idade fértil da mulher. Assim, optou-se definir um intervalo entre 10 e 49 anos para este atributo seguindo o mesmo conceito presente na Política Nacional de Atenção Integral à Saúde da Mulher definida em (15). Para este atributo também não foram encontrados valores discrepantes e que se distanciasse do valor médio (*outlier*).

Na fase de transformação, todas as variáveis foram categorizadas e enquadradas como não ordinais e, na sequência, modificada para variáveis do tipo *dummy*. Este processo gera N novos atributos, onde N é dado pelo número de valores únicos. A categoria do atributo de correspondência de posição é preenchida com o valor 1 (um), enquanto o restante das posições são preenchidas com o valor 0 (zero). Este procedimento visou atender a uma particularidade da biblioteca (*Scikit-Learn*) utilizada no trabalho, que nem todas as implementações dos métodos de aprendizado de máquina aceitam variáveis categóricas. A Tabela 1 resume o total de registros de nascidos vivos e óbitos neonatais, por região, encontrados após o processo de vinculação de registros.

Tabela 1 – Total de registros de nascidos vivos e óbitos neonatais, por região, presentes no conjunto de dados compilado.

	NORTE		**NORDESTE**		**CENTRO OESTE**		**SUDESTE**		**SUL**	
	VIVOS	**ÓBITOS**	**VIVOS**	**ÓBITOS**	**VIVOS**	**ÓBITOS**	**VIVOS**	**ÓBITOS**	**VIVOS**	**ÓBITOS**
2012	285136	2040	711927	4793	225590	1419	1125583	6439	372439	2514
2013	293487	2192	699486	4757	229921	1545	1120922	6721	378826	2449
2014	302197	2176	709846	4776	236962	1611	1155839	6937	387863	2660
TOTAL	**880820**	**6408**	**2121259**	**14326**	**692473**	**4575**	**3402344**	**20097**	**1139128**	**7623**

Por fim, o conjunto de dados final foi desmembrado em cinco grupos de variáveis para uso na construção e avaliação dos classificadores, assim como segue:

1. Socioeconômicas maternas;
2. Comportamentais e de uso dos serviços de saúde;
3. Biológicas maternas e ligadas ao recém-nascido;
4. Presentes exclusivamente nos conjuntos de dados da PCDas; e
5. Todas variáveis, incluindo as enriquecidas (conjunto de dados compilado).

O intuito dessa divisão foi explorar cada dimensão da mortalidade neonatal e entender, separadamente, quais são os fatores de riscos mais relevantes para a classificação do risco neonatal. Além disso, buscou-se investigar se as variáveis enriquecidas trazem algum ganho no poder preditivo dos classificadores.

4.1.2 Qualidade e limitações

No processo de aprendizado de máquina, o uso de dados de treinamento incompletos, errôneos ou inadequados podem levar a modelos não confiáveis (24). De acordo com (86), a garantia da qualidade dos dados disponibilizadas pelo DATASUS é condição essencial para a análise objetiva da situação sanitária, assim como para a tomada de decisões e para a programação de ações de saúde. Ademais, a qualidade dos dados pode afetar também o processo de vinculação dos registros entre os conjuntos de dados. Segundo (112), o relacionamento de registros é altamente sensível à qualidade dos dados que estão sendo ligados, então todos os conjuntos de dados em análise devem passar por uma avaliação da qualidade dos dados antes de realizar o processo de ligação entre eles.

Neste sentido, a escolha da janela temporal (2012-2014) visou buscar um cenário de menos inconsistências e mais correspondências na vinculação dos registros entre os atributos presentes nos conjuntos de dados. São as seguintes motivações existentes para essa escolha:

I. A publicação da portaria n.º 116 de 11 de fevereiro de 2009, que regulamenta a coleta de dados, fluxo e periodicidade de envio das informações sobre óbitos e nascidos vivos para os Sistemas de Informação em Saúde, trazendo diretrizes quanto à obrigatoriedade da emissão de Declaração de Nascido Vivo (DN) para todo nascido vivo, independente da duração da gestação, peso e estatura do recém-nascido;
II. a vigência da lei n.º 12.662, de 5 de junho de 2012, que assegura validade nacional a Declaração de Nascido Vivo (DNV);
III. as mudanças em variáveis e na forma de coleta da Declaração de Nascido Vivo a partir de 2011, conforme documento elaborado pelo Ministério da Saúde[11] prevendo aumento na cobertura em todas as regiões, conferindo uma maior representatividade as informações geradas a partir do SINASC; e
IV. a presença do atributo com o número da declaração de nascido vivo para vinculação de registros entre os conjuntos de dados dos sistemas SINASC e SIM do DATASUS.

Sendo assim, foi conferido um conjunto de dados de maior qualidade para realização das análises e geração de modelos de aprendizado de máquina mais acurados.

4.1.3 Análise exploratória

A partir do conjunto de dados formado foi possível explorar com mais detalhes e extrair algumas características presentes nos dados que pudessem relevar qualquer padrão ou característica no contexto do trabalho. Ressalta-se que o estudo considera a divisão

[11] BRASIL. Ministério da Saúde. Consolidação do Sistema de Informações sobre Nascidos Vivos - 2011. Disponível em: `http://tabnet.datasus.gov.br/cgi/sinasc/Consolida_Sinasc_2011.pdf`

regional brasileira em vigor conforme IBGE (2022)[12]. A Tabela 2 apresenta a quantidade de nascidos vivos, óbitos neonatais e a taxa de mortalidade neonatal extraída do conjunto de dados compilado. Os dados apresentados corroboram a hipótese da existência de possíveis diferenças entre as regiões do país em relação aos fatores relacionados ao óbito neonatal. Esse quadro motivou avançar para uma maior investigação das questões ligadas às variáveis demográficas, socioeconômicas e de assistência à saúde na gestação e parto sob a visão regional brasileira.

Tabela 2 – Total de nascidos vivos (#NV), quantidade de óbitos neonatais (#ON) e taxa de mortalidade neonatal (#TMN), por região, entre 2012-2014.

		2012	**2013**	**2014**
NORTE	**#NV**	308375	313272	321682
	#ON	3338	321682	3320
	TMN(%)	10,82	10,70	10,32
NORDESTE	**#NV**	832631	821458	833090
	#ON	8970	8999	8594
	TMN(%)	10,77	10,95	10,32
CENTRO OESTE	**#NV**	230279	234687	245076
	#ON	2203	2185	2199
	TMN(%)	9,57	9,31	8,97
SUDESTE	**#NV**	1152846	1147627	1182949
	#ON	9659	9385	9634
	TMN(%)	8,38	8,18	8,14
SUL	**#NV**	381658	386983	396462
	#ON	2943	2808	2999
	TMN(%)	7,71	7,26	7,56

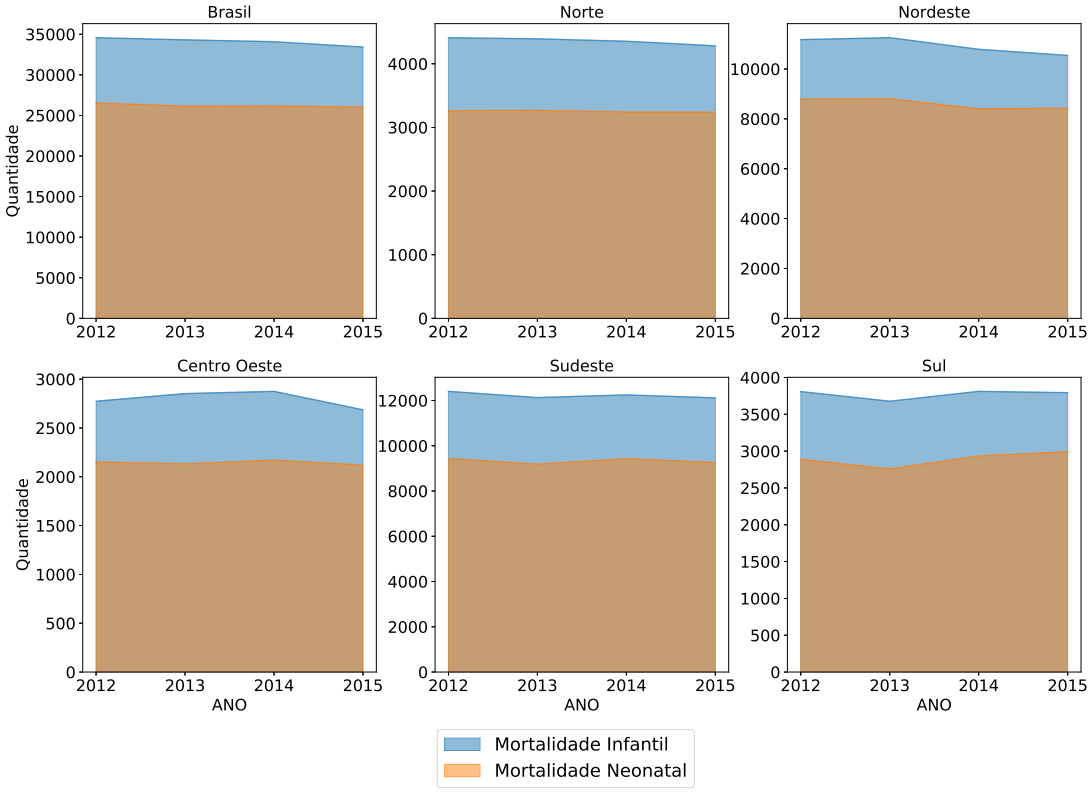
A Figura 5 mostra as quantidades de óbitos neonatais entre 2012 e 2015. Ressalta-se que o conjunto de dados do ano de 2015, foi utilizado para realizar a vinculação dos registros de gestações que se iniciaram no ano anterior. Ainda na Figura 5, observa-se uma quantidade maior de óbitos neonatais em relação aos infantis. Além disso, essas quantidades tendem a permanecer constantes ao longo do período estudado. Assim, é possível notar a importância de canalizar estudos direcionados a este importante componente da mortalidade infantil.

Visando observar os atributos separados por região e por rótulo da classe em busca de diferenças na relação entre as classes de vivos e casos de óbitos, foi realizada uma análise exploratória nos dados. Nessa fase, destacam-se:

I. Em todas as regiões, o maior número de casos, sejam de vivos ou óbitos neonatais, fica entre mães na faixa de idade de 20 a 30 anos, representando mais de 22% de toda a amostra de dados desse atributo.

[12] IBGE. Divisão Regional do Brasil. 2022. Disponível em: `https://www.ibge.gov.br/geociencias/organizacao-do-territorio/divisao-regional/15778-divisoes-regionais-do-brasil.html`

Figura 5 – Quantidades de óbitos neonatais e infantis de 2012 a 2015.



Fonte: Moreira et al. (2022).

II. Em relação ao número de consultas durante o pré-natal, foi observado uma diferença de frequência entre as classes. Mulheres que perderam seus filhos nas primeiras semanas de vida fizeram menos de sete consultas pré-natais durante a gestação.
III. Observou-se também, em todas as regiões, que os óbitos neonatais ocorreram principalmente quando a gestação ocorreu em menos de 22 semanas.
IV. A frequência relativa do tipo de apresentação pélvica ou podálica do recém-nascido na classe óbito superou em aproximadamente 10% os casos da classe vivo.
V. Outro atributo que incorreu em diferenças significativas entre as classes foram os índices de vitalidade Apgar de um e cinco minutos. O nível grave de ambos os índices marcou presença nos casos de óbito, com uma diferença de aproximadamente 30% em relação à mesma categoria da classe dos bebês que não foram a óbito.
VI. O atributo peso ao nascer também apresentou diferenças marcantes entre as classes. Os casos de óbitos onde o peso ao nascer eram menores que 2.500g, mostrou aproximadamente 60% a mais de casos em relação à mesma categoria da classe vivo.
VII. Em todas as regiões, o fator prematuridade se mostrou expressivo em valores absolutos na classe óbito. Foi observada uma diferença de aproximadamente 50% dos casos de óbitos neonatais em relação à classe vivo.

A análise realizada traz uma ideia geral sobre os dados e foram essenciais para compreender os pontos que devem ser observados na fase de treinamento dos algoritmos propostos como o tratamento do desequilíbrio entre as classes, assunto abordado na seção 4.2.1.

4.2 CONSTRUÇÃO DOS CLASSIFICADORES

Na literatura são encontrados diversos conceitos para o termo aprendizado de máquina. Sucintamente, pode ser definido como o desenvolvimento de algoritmos que podem aprender com os dados disponíveis (1), dando aos computadores a habilidade de aprender sem serem explicitamente programados (95). A aprendizagem de máquina pode ser descrita ainda como um processo que explora grandes volumes dados à procura de padrões consistentes, anomalias e correlações (104).

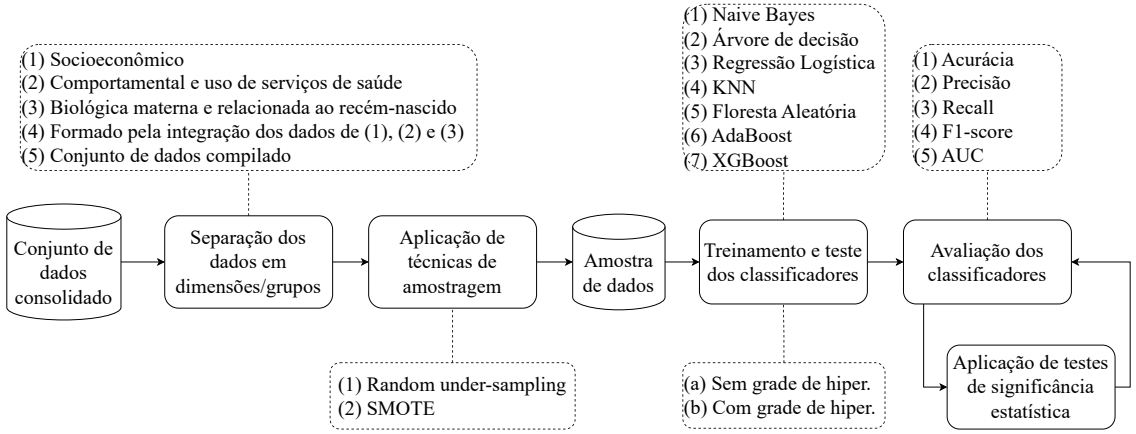
O aprendizado de máquina é uma das áreas da computação que mais cresceu nas últimas décadas (23). O surgimento de grandes conjuntos de dados foi decisivo para este crescimento, pois a aplicação de um alto volume de dados históricos para treinamento e teste dos modelos é uma característica essencial dentro desta área de pesquisa.

No geral, existem três abordagens de aprendizado de máquina: supervisionada, não-supervisionada e por reforço (63). A abordagem supervisionada utiliza rótulos pré-existentes das variáveis dependentes (uma classe, no caso de classificação). No caso da abordagem não-supervisionada, consiste no aprendizado onde não são apresentados, previamente, os rótulos para as instâncias. Já na abordagem por reforço, o processo de aprendizado é baseado na tentativa e erro para encontrar a solução de um problema.

Neste estudo, é utilizada a abordagem supervisionada, em especial, de classificação binária. Por esta abordagem, há a presença da variável dependente com classes rotuladas que, no caso deste trabalho, são definidos por dois rótulos, quais sejam: um para bebês que foram ao óbito nos seus 27 primeiros dias de vida e outro para os que sobreviveram após este período.

A construção dos classificadores foi pautada por uma sequência de etapas, as quais abrangeram: a separação dos grupos de variáveis; a aplicação de técnicas de amostragem para tratamento do desbalanceamento das classes; execução de sete métodos de aprendizado de máquina - com e sem grade (*grid*) de hiperparâmetros; e a avaliação dos modelos com cinco métricas de desempenho, além da aplicação de testes de significância estatística para comparação dos resultados. A Figura 6 sintetiza o fluxo dos procedimentos adotados para a construção dos classificadores. No decorrer desta seção, cada um desses procedimentos são apresentados com mais detalhes, abordando todas as técnicas e ferramentas utilizadas na construção, avaliação e seleção dos classificadores mais acurados e eficientes na classificação do risco de óbito neonatal.

Figura 6 – Processo adotado na construção dos classificadores.



Fonte: Elaborado pelo autor (2022).

4.2.1 Desbalanceamento das classes

O processo de construção dos classificadores se iniciou com o tratamento do problema de desbalanceamento das classes. Segundo (27), um conjunto de dados é considerado desbalanceamento se as classes não são igualmente representadas. No caso deste estudo, ressalta-se a existência do alto desequilíbrio entre as classes, onde há a presença de mais casos de bebês que não foram a óbito (classe 0) do que da classe que representa os bebês que foram a óbito nos primeiros 27 dias de vida (classe 1). Esta característica tende a produzir modelos de classificação que favorecem a classe com maior probabilidade de ocorrência (majoritária), resultando em uma baixa taxa de reconhecimento para o grupo minoritário (26).

Diante desta característica presente no conjunto de dados e, no intuito de equilibrar a representatividade das classes, foram aplicadas as seguintes técnicas aos dados: I. *Synthetic Minority Oversampling Technique* (SMOTE) para sobreamostragem (*oversampling*); e II. *Random Undersampling* para subamostragem (*undersampling*). A primeira técnica consiste na criação de instâncias sintéticas da classe minoritária visando equilibrá-las com os registros da classe majoritária, e a segunda, que remove as instâncias da classe majoritária, de forma aleatória e sem reposição, para realizar o balanceamento das classes.

4.2.2 Métodos de aprendizado de máquina

Após o tratamento do desbalanceamento das classes, os grupos de dados com os componentes neonatais foram separados para treinamento e teste dos classificadores. Os métodos de aprendizado de máquina aplicados neste trabalho foram: naive bayes, árvore de decisão, regressão logística, k-vizinhos mais próximos (KNN), além dos métodos de comitês (Ensemble) floresta aleatória (Random Forest), estímulo adaptativo (AdaBoost) e

aumento extremo de gradiente (XGBoost). A seguir, é apresentada uma breve descrição do processo de aprendizado de cada um desses métodos.

• **Naive Bayes**

É um método probabilístico baseado no teorema de Bayes. Ele calcula a probabilidade de um evento a partir de um conhecimento *a priori* de fatores observados nos dados e associados à ocorrência do evento (106). De forma geral, o *naive bayes* pode ser expresso como na equação 4.1.

$$P(y|x_1, x_2, \ldots, x_n) = \frac{P(x_1, x_2, \ldots, x_n|y) \cdot P(y)}{P(x_1, x_2, \ldots, x_n)} \tag{4.1}$$

Onde, $P(y|x_1, x_2, \ldots, x_n)$ representa a probabilidade condicional da classe y dado o conjunto de atributos $x_1, x_2, \ldots, x_n$; A expressão $P(x_1, x_2, \ldots, x_n|y)$ a probabilidade conjunta dos atributos $x_1, x_2, \ldots, x_n$ dada a classe y; $P(y)$ a probabilidade a priori da classe y; e $P(x_1, x_2, \ldots, x_n)$ a probabilidade marginal dos atributos $x_1, x_2, \ldots, x_n$

• **Regressão Logística**

Outro método utilizado no trabalho e que é largamente usado em problemas de classificação foi a regressão logística. Esse método adapta técnicas de regressão linear para medir a relação entre a variável dependente, em geral binária, e outras variáveis independentes, usando uma função logística para estimar as probabilidades (105, 47). A regressão logística pode ser representada pela equação (4.2):

$$P(y = 1|x_1, x_2, \ldots, x_n) = \frac{1}{1 + e^{-z}} \tag{4.2}$$

Onde, $P(y = 1 \mid x_1, x_2, \ldots, x_n)$ representa a probabilidade condicional da variável dependente ser igual a 1, dado o conjunto de variáveis independentes $x_1, x_2, \ldots, x_n$. O valor z é o valor linear preditivo, calculado como uma combinação linear das variáveis independentes ponderadas pelos coeficientes do modelo. A função logística $\frac{1}{1+e^{-z}}$ transforma o valor linear preditivo em uma probabilidade entre 0 e 1, permitindo a interpretação dos resultados.

• **Árvore de Decisão**

Este método apresenta uma maneira simples de visualizar um modelo de tomada de decisão por meio de uma estrutura em árvore. Sua composição básica consiste em um nó raiz e nós terminais, estes últimos também chamados de folhas. As decisões se bifurcam nessa estrutura até que uma decisão de previsão seja feita para um determinado registro. Conforme abordado em (103), esse método tem sido amplamente utilizado em várias áreas do conhecimento por ser de fácil aplicação, livre de ambiguidade e robusto mesmo na presença de valores ausentes. Outra característica deste método é poder ser utilizado tanto com dados categóricos, quanto numéricos.

• **K-vizinhos mais próximos**

Também conhecido como *K-Nearest Neighbor* (KNN) é um método que utiliza um critério de similaridade definido usualmente por uma medida de distância no espaço multidimensional das observações. O tipo de medida de distância pode variar, sendo a mais usual a distância Euclidiana. No caso de uma nova instância, as previsões são feitas usando todo o conjunto de treinamento para as K instâncias mais próximas. A previsão é feita pesquisando as K instâncias mais semelhantes (vizinhos) do item a ser classificado (83). Em problemas de classificação, dada uma nova instância, são recuperados os K vizinhos mais próximos, sendo atribuída a nova instância, a classe mais frequente entre esses K vizinhos.

• **Métodos de Comitê**:

É uma técnica de aprendizado de máquina que combina a decisão de vários modelos para melhorar os resultados preditivos. Comumente são divididos em duas abordagens: *bagging* e *boosting*. Na abordagem de *bagging* ocorre um ajuste dos pesos dos modelos em paralelo. No caso da *boosting*, a estratégia é do ajuste do peso é iterativo com base na classificação anterior. A seguir são destacados os métodos enquadrados nessas abordagens e que foram utilizados neste trabalho.

◇ **Random Forest**

O método *random forest* cria várias árvores de decisão usando aleatoriedade na amostragem dos dados de treinamento e nos atributos que podem ser explorados. A floresta decide a escolha da classe a partir da combinação das respostas das suas árvores de decisão. Outra característica do *random forest* é que ele é robusto contra *overfitting* (22).

◇ **AdaBoost**

O *AdaBoost* combina vários classificadores para aumentar a precisão do modelo que está sendo construído. Ele minimiza os erros identificados na cadeia de execução dos modelos por meio de pesos. Conforme (48), esse método também tem sido usado como uma ferramenta de pré-processamento para selecionar automaticamente as características mais importantes de conjunto de dados de alta dimensionalidade.

◇ **XGBoost**

O *XGBoost* é um algoritmo de aprendizado de máquina de comitê, baseado em árvore de decisão, que utiliza uma estrutura de *Gradient boosting* (87, 51). De acordo com (92), o *XGBoost* é um método amplamente utilizado para problemas de classificação e conforme (29), entre os métodos de aprendizado de máquina, é o que mais se destaca em muitas aplicações em relação ao desempenho. Outro aspecto deste método é que ele aborda o processo de construção sequencial de árvores usando implementação paralelizada (65).

4.2.3 Otimização de hiperparâmetros

Expostos os métodos de aprendizado de máquina utilizados no trabalho, outro ponto importante a se destacar, é a utilização de um método de pesquisa para buscar o conjunto de hiperparâmetros que gera os melhores resultados dos classificadores. Ressalta-se que, frequentemente, os algoritmos de aprendizado de máquina exigem a escolha de um conjunto de parâmetros ideais, também conhecidos como ajuste (*tuning*) ou otimização de hiperparâmetros (88).

Cabe destacar que em um modelo preditivo, parâmetros e hiperparâmetros são considerados conceitos distintos e fundamentais do processo de aprendizagem. No caso dos parâmetros, são ajustados pelo próprio processo de aprendizado. Já os hiperparâmetros, são variáveis do algoritmo definidas antes da etapa de treinamento, que podem gerar ganhos no desempenho preditivo do modelo.

Neste trabalho foi utilizada a pesquisa em grade *GridSearch* que é uma técnica que realiza a busca completa em um determinado subconjunto do espaço de hiperparâmetros do algoritmo de treinamento (59). Essa abordagem testou exaustivamente todas as combinações possíveis dos hiperparâmetros definidos na fase de treinamento dos modelos propostos neste trabalho.

4.2.4 Métricas de avaliação

Neste estudo foram utilizadas as métricas acurácia, precisão, *recall* (sensibilidade), *f1-score* e a área sob a curva ROC (AUC) para avaliação dos classificadores. Essas métricas são formuladas a partir da matriz de confusão que mostra a contagem de previsões corretas e incorretas feitas pelo classificador em relação às classes reais do conjunto de dados (34, 94). Por se tratar de um problema de classificação binária, a matriz de confusão moldada no presente estudo é de ordem 2x2 assim como apresentado na Tabela 3 e será utilizada como base para uma breve descrição de cada métrica adotada.

Tabela 3 – Matriz de confusão.

		Classe prevista	
		Positivo	**Negativo**
Classe real	**Positivo**	Verdadeiro Positivo (VP)	Falso Negativo (FN)
	Negativo	Falso Positivo (FP)	Verdadeiro Negativo (VN)

Em relação a acurácia, esta consiste em quantificar a taxa de acertos do modelo em relação ao conjunto total de amostras avaliadas. É calculada dividindo a soma dos verdadeiros positivos e verdadeiros negativos pela soma de todas as entradas da matriz de confusão, assim como apresentado na expressão (4.3). De forma geral, é uma métrica que indica o desempenho geral do classificador.

$$Acurácia = \frac{VP + VN}{VP + VN + FP + FN} \tag{4.3}$$

No caso da precisão, ela identifica as proporções de previsões corretas da classe dependente (verdadeiros positivos), em relação aos casos previstos pelo modelo da classe positiva (casos de óbitos neonatais). A precisão é calculada dividindo o número de verdadeiros positivos pelo total de exemplos classificados como positivos (soma de verdadeiros positivos e falsos positivos) assim como expresso em (4.4).

$$Precisão = \frac{VP}{VP + FP} \tag{4.4}$$

Já o *recall*, trata-se da proporção de previsões corretas da classe dependente (verdadeiros positivos) em relação à soma dos verdadeiros positivos com os casos de falsos negativos, assim como formulado na expressão (4.5). Ela foi usada, pois é plausível considerar que a presença de falsos negativos seja mais prejudicial que os falsos positivos em um cenário de predição de risco de óbito neonatal.

$$Recall = \frac{VP}{VP + FN} \tag{4.5}$$

No tocante a *f1-score*, compreende da média harmônica entre a precisão e o *recall*. É utilizada para se ter um resumo da qualidade do modelo em aspectos de *trade-off* entre essas duas métricas, sendo expressa conforme apresentado em (4.6).

$$F1\text{-}score = 2 * \frac{Precisão \ * \ Recall}{Precisão \ + \ Recall} \tag{4.6}$$

Por fim, a área sob a curva ROC (AUC) que representa a área formada por uma curva de probabilidade que traça a taxa de verdadeiro positivo contra a taxa de falsos positivos.

Além desses métricas, os classificadores gerados também foram avaliados aplicando os testes de significância estatística de Friedman e *post-hoc* de Nemenyi para investigar possíveis diferenças estatísticas entre os classificadores com os melhores resultados. Neste trabalho, os testes foram executados a um nível de confiança de 95% para os resultados das métricas de AUC e *f1-score*.

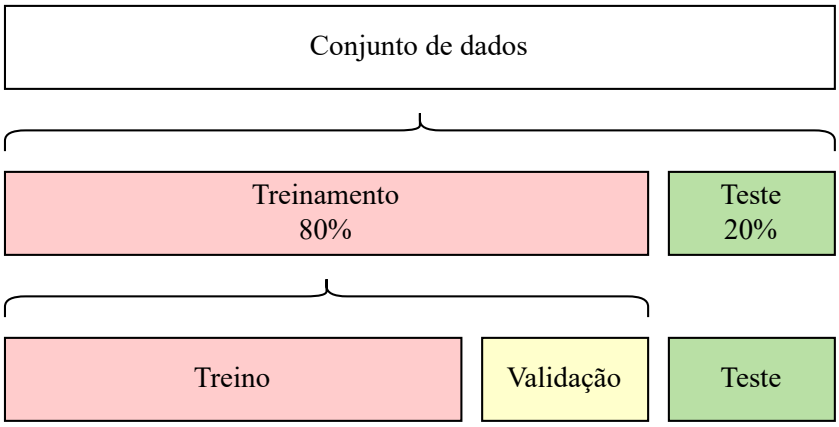
4.2.5 Divisão dos dados em treinamento, validação e teste

Inicialmente cada conjunto de dados utilizado no experimento foi dividido em dois subconjuntos: um com 80% dos dados para treinamento e 20% para teste. O treinamento e validação dos modelos ocorreu utilizando este primeiro subconjunto de dados onde foi aplicada a estratégia de validação cruzada do tipo K-*fold* estratificada. No caso deste trabalho foi utilizado o valor de K=10, ou seja, os dados são divididos em 10 subconjuntos

(*folds*) e, a cada iteração, um desses conjuntos é usado para testar o modelo gerado usando os demais dados para o treinamento. Como foi adotada a versão estratificada do *K-fold*, foi mantida a proporção original das classes nos *folds*.

Já o subconjunto de dados de teste permaneceu inalterado desde a divisão inicial visando aplicá-lo exclusivamente para o teste e avaliação dos modelos gerados. Ressalta-se que as estratégias adotadas visaram evitar o sobreajuste *overfitting* dos modelos. As estratégias adotadas nesta etapa do trabalho são apresentadas na Figura 7.

Figura 7 – Estratégia de divisão dos dados para treinamento, validação e teste dos classificadores.



Fonte: Elaborado pelo autor (2022).

4.3 EXPERIMENTOS

O ambiente computacional utilizado nos experimentos foi composto por uma máquina com CPU de 20 *threads* e 96 GB de RAM, rodando o sistema operacional *linux* (distribuição Ubuntu 20.04 de 64 bits). Foram utilizados também a linguagem de programação *Python* versão 3.6 com as bibliotecas *Numpy* (1.20.3), *Pandas* (1.3.2), *Scikit-Learn* (0.24.2) e *MatplotLib* (3.3.4), sendo os algoritmos elaborados e organizados no *Jupyter-notebook* (6.3.0).

A condução dos experimentos foi realizada sob dois cenários de testes: um utilizando os valores de hiperparâmetros padrão dos métodos de aprendizado de máquina da biblioteca *Scikit-learn* e outro com uma grade de hiperparâmetros pré-definida, conforme os valores apresentados na Tabela 4.

Cada um dos cinco conjuntos de dados formado na seção 4.1.1, foi dividido em 80% dos registros para treino e validação dos classificadores e 20% para teste. Nessa divisão, tomou-se o cuidado de preservar as mesmas proporções das classes em relação ao conjunto

Tabela 4 – Hiperparâmetros usados na grade do *GridSearch* por método de classificação.

ALGORITMO	MÉTODO SKLEARN	HIPERPARÂMETRO	DESCRIÇÃO	VALORES
Naive Bayes	GaussianNB	var_smoothing	Utilizado no cálculo de estabilidade para ampliar (ou suavizar) a curva de probabilidade	[1e-08,1e-09,1e-10]
Árvore de Decisão	DecisionTreeClassifier	criterion	Função para encontrar a melhor estratégia de divisão da árvore de decisão	['gini', 'entropy']
Regressão Logística	LogisticRegression	C	força da regularização	[0.1, 1, 10]
Random Forest	RandomForestClassifier	n_estimators	Número de árvores utilizadas no treinamento	[100, 150, 200]
KNN	KNeighborsClassifier	n_neighbors	Número de vizinhos que serão usados para consulta	[3,4,5]
AdaBoost	AdaBoostClassifier	n_estimators	Número de árvores utilizadas no treinamento	[50, 100, 150]
XGBoost	XGBClassifier	n_estimators	Número de árvores utilizadas no treinamento	[100, 150, 200]

de dados original. No tocante aos algoritmos, foi utilizado o *Grid Search* com a seguinte estrutura:

I. Aplicação de um método de classificação: todos os abordados na seção 4.2.2;
II. Validação cruzada estratificada do tipo *K-fold*: estratégia adotada para treinar e avaliar a capacidade de generalização do classificador gerado. Neste trabalho foi utilizado o valor de K=10, ou seja, os dados são divididos em 10 subconjuntos (*folds*) e, a cada iteração, um desses conjuntos é usado para testar o modelo gerado usando os demais dados para o treinamento. Como foi adotada a versão estratificada, foi mantida a proporção original das classes nos *folds*;
III. Conjunto de métricas para avaliação do classificador: as métricas usadas foram a acurácia, precisão, *recall*, *f1-score* e AUC. Todas abordadas com maiores detalhes na seção 4.2.4; e
IV. Grade de hiperparâmetros: conforme cada método de classificação empregado foi utilizado um conjunto de hiperparâmetros em consonância com a Tabela 4.

Ao todo, foram gerados 72 classificadores: 70 usando os sete métodos de aprendizado de máquina apresentados e aplicando a abordagem de subamostragem aleatória *random undersampling* com e sem grade de hiperparâmetros pré-definida; e 2 gerados aplicando sobreamostragem com o SMOTE e usando os métodos e grupos de dados dos classificadores que apresentaram os menores tempos e que também foram possíveis a utilização de todo recurso computacional disponível (memória e *threads*) no emprego de paralelismo para treinamento, avaliação e teste dos classificadores.

A avaliação dos classificadores foi pautada pela extração da média e desvio padrão da execução de 30 vezes de cada algoritmo e conjunto de dados aplicado, incluindo análises de sensibilidade e especificidade sob testes estatísticos para validação dos melhores resultados. Além disso, foram extraídas as variáveis mais relevantes do classificador com melhor desempenho que será apresentado e discutido no próximo capítulo.

5 RESULTADOS E DISCUSSÕES

Neste capítulo são apresentados os resultados da pesquisa e as considerações relativas aos experimentos realizados sobre os dados apresentados no capítulo 4. Este capítulo foi dividido em duas seções. Na seção 5.1 são apresentados os resultados de desempenho obtidos pelos classificadores, além de abordar a interpretabilidade do modelo selecionado; e na seção 5.2, a discussão sobre os resultados alcançados, comparando-os com os trabalhos da literatura.

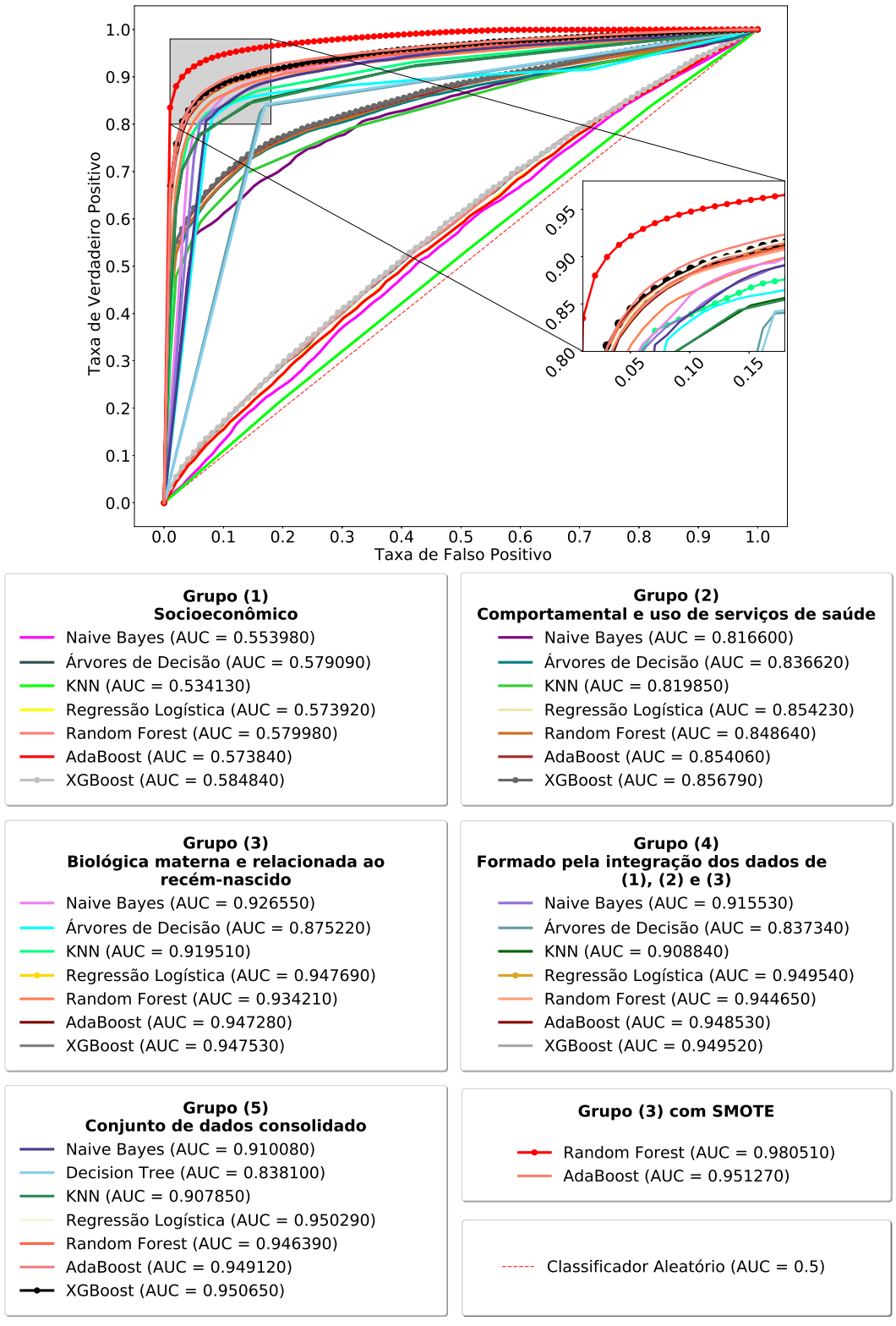
5.1 Desempenho dos classificadores

Conforme apresentado na seção 4.2, os classificadores foram avaliados pela análise descritiva usando a média e o desvio padrão da execução de 30 vezes de cada algoritmo para as métricas de acurácia, precisão, *recall*, *f1-score* e AUC. Além desses indicadores, os classificadores também foram avaliados em relação ao tempo total de execução para treinamento, avaliação e teste. O Apêndice C apresenta o quadro resumo com os resultados obtidos de todos os classificadores gerados. Como observado, os classificadores apresentaram desempenho médio, desvios padrão e tempos de execução diferentes para cada dimensão ou grupo de dados usados. É um cenário esperado considerando a particularidade de aprendizagem e ajuste dos parâmetros de cada um dos métodos aplicados.

Os resultados também revelam que os classificadores gerados usando o conjunto de dados com subamostragem aleatória com variáveis relacionadas às questões exclusivamente socioeconômicas maternas (grupo de dados 1), obtiveram resultados inferiores a 60% no poder preditivo, independente da métrica avaliada. Este resultado mostra que classificadores gerados aplicando esse grupo de dados não tem eficácia na predição do risco de óbito neonatal. A única vantagem na aplicação desse grupo de dados é em relação ao tempo de execução total para treinamento, validação e teste dos classificadores, pois conforme revelado nos resultados apresentados no Apêndice C, foram os que apresentaram os menores tempos independente do método aplicado. Em contrapartida, os classificadores treinados com o grupo composto pelas variáveis comportamentais e de uso dos serviços de saúde (grupo de dados 2) obtiveram melhores resultados em relação ao grupo anterior, com todas as métricas apresentando valores acima de 70%. Contudo, ainda com resultados inferiores aos treinados e testados com os demais grupos de dados (3, 4 e 5). Observa-se também que os classificadores gerados aplicando os métodos *naive bayes*, árvore de decisão e o KNN, apresentaram os piores resultados, sendo este último com os maiores tempos de execução com os dados subamostrados.

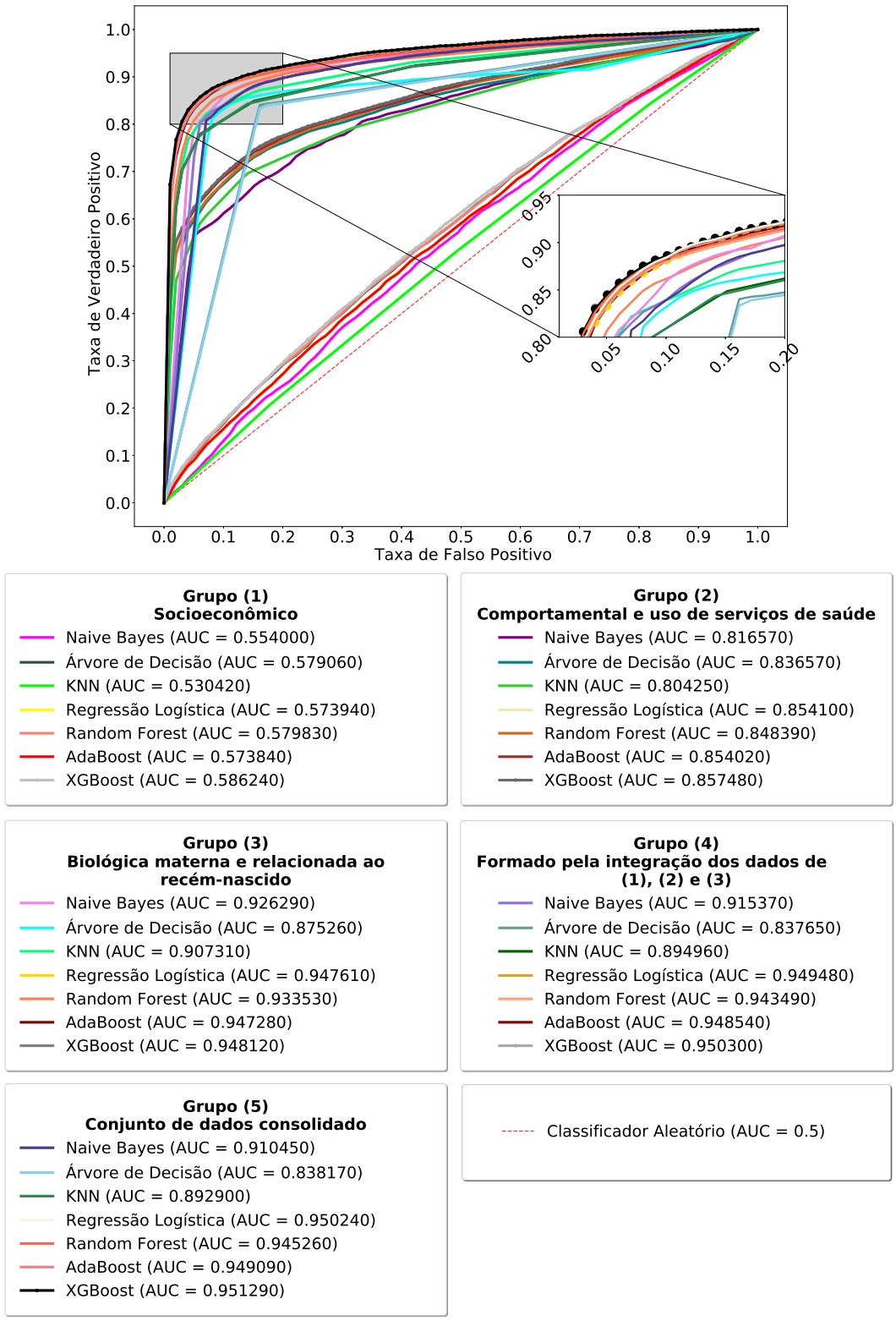
A primeira etapa de avaliação dos classificadores foi balizada pela análise comparativa das áreas sob à curva ROC. Desta forma, visando ilustrar a comparação de todos os resultados foram traçadas as curvas com suas respectivas áreas nas Figuras 8 e 9.

Figura 8 – Comparação entre as curvas ROC: usando subamostragem (*Random Undersampling*) e sobreamostragem (SMOTE) sem grade hiperparâmetros.



Fonte: Elaborado pelo autor (2022).

Figura 9 – Comparação entre as curvas ROC: usando subamostragem (*Random Undersampling*) com grade de hiperparâmetros pré-definida.



Fonte: Elaborado pelo autor (2022).

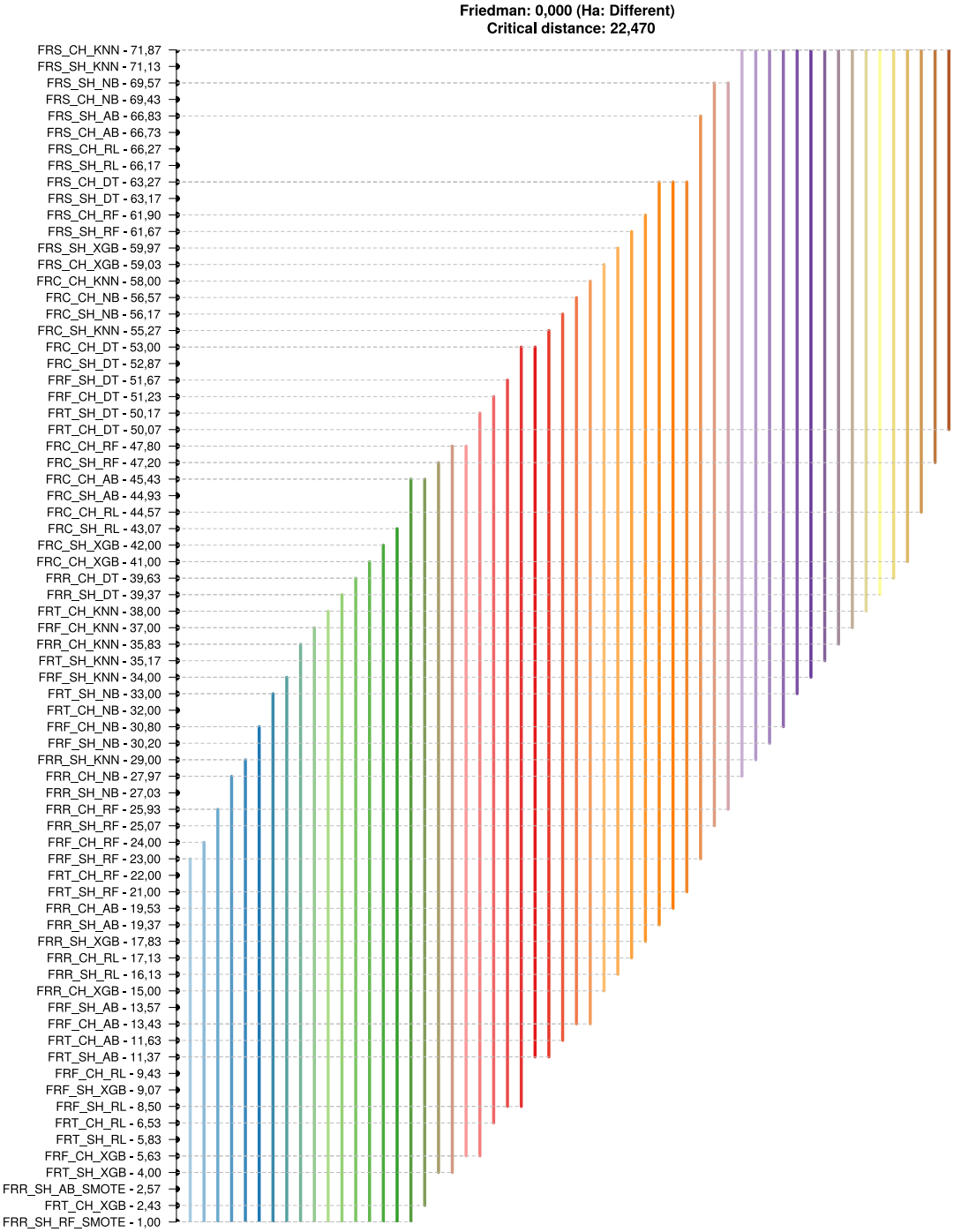
De forma complementar, para uma visualização segmentada, os gráficos com as curvas ROC e as respectivas áreas são apresentadas no Apêndice D para os algoritmos sem grade de hiperparâmetros; e no Apêndice E para os classificadores com grade de hiperparâmetros pré-definida.

Por meio dos gráficos é possível notar que os classificadores gerados aplicando regressão logística e os métodos de comitê: Random Forest, AdaBoost e XGBoost - todos usando os grupos de dados 3, 4 e 5, seja com subamostragem aleatória ou com sobreamostragem (SMOTE), possuem nítida sobreposição das curvas e valores de AUC muito próximos. Visando identificar possíveis diferenças estatísticas significativas entre os resultados foi conduzido o teste de significância estatística não paramétrico de Friedman e *post-hoc* de Nemenyi (55). Cabe destacar que o teste não paramétrico é um tipo de teste onde o modelo não especifica as condições, como a distribuição dos dados, que devem atender aos parâmetros da população da qual a amostra foi extraída (50). Destaca-se ainda que a aplicação desses testes visa fornecer um melhor discernimento na verificação da relevância de desempenho dos classificadores (96). Ademais, salienta-se que os resultados da métrica *f1-score* também foram avaliados em termos de inferência estatística, pois inserido neste tipo de estudo, onde se tem um grande número de casos de não-óbitos em relação aos casos de óbitos, é relevante determinar o modelo com a melhor relação entre a sensibilidade e a precisão no processo de classificação.

Posto isso, o primeiro passo foi realizar o teste de Friedman, a um nível de confiança de 95%, para detectar a possível ocorrência de diferenças estatísticas significativas entre os resultados. Ressalta-se que esse teste classifica os algoritmos, em ordem, do melhor desempenho para o pior. De acordo com (96), se o teste revelar a presença de significância estatística entre os resultados, segue-se para a aplicação do procedimento *post-hoc* para apontar quais são os pares de algoritmos que diferem significativamente entre os resultados. No caso deste estudo, o teste aplicado foi o *post-hoc* de Nemenyi (55). Os resultados dos testes para as métricas AUC e *f1-score* são apresentados nas Figuras 10 e Figura 11, respectivamente. Ademais no intuito de oferecer mais um forma de visualização desses resultados, o Apêndice F apresenta dois gráficos complementares para as métricas avaliadas.

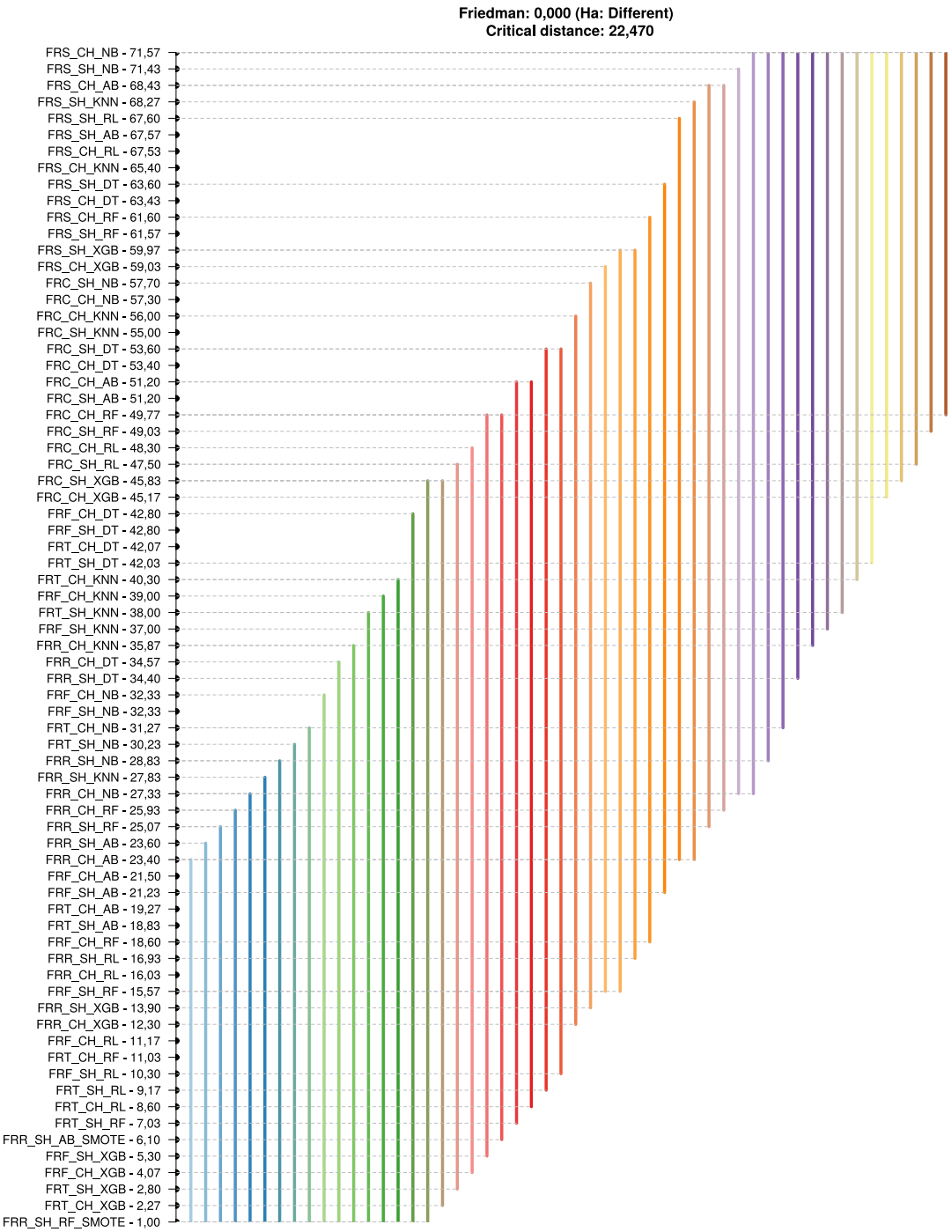
Como observado, o teste de Friedman resultou um p-valor igual a zero para ambos os casos, mostrando a existência de diferença entre as amostras dos resultados de teste com as métricas AUC e *f1-score*. Em relação ao teste *post-hoc* de Nemenyi, percebe-se que a distância crítica foi de 22,470, indicando que as distâncias entre os resultados dos classificadores devem ser maiores do que este valor para caracterizar uma diferença estatística significativa entre eles. Portanto, neste cenário, considera-se que os classificadores expostos na Tabela 5 não possuem diferença estatística significativa entre os resultados, ou seja, seus desempenhos em relação à área sob à ROC e a taxa de *f1-score* são considerados equivalentes.

Figura 10 – Testes de Friedman e *post-hoc* de Nemenyi para a métrica AUC.



Fonte: Elaborado pelo autor (2022).

Figura 11 – Testes de Friedman e *post-hoc* de Nemenyi para a métrica *f1-score*.



Fonte: Elaborado pelo autor (2022).

Tabela 5 – Melhores resultados após o teste de significância estatística para as métricas AUC e *f1-score*.

Grupo ou dimensão da mortalidade neonatal	**Com e sem grade de hiperparâmetros**
Relacionados ao recém-nascido, parto e biológicos maternos	Random Forest (SMOTE) AdaBoost (SMOTE) Regressão Logística AdaBoost XGBoost
Presentes na base da PCDas Fiocruz	Regressão Logística AdaBoost XGBoost Random Forest
Presentes na base do PCDas Fiocruz com enriquecimento de dados	Regressão Logística AdaBoost XGBoost Random Forest

Após aplicação dos testes estatísticos e separando os classificadores com as melhores resultados para as duas métricas, chegou-se a um total de 23 classificadores:

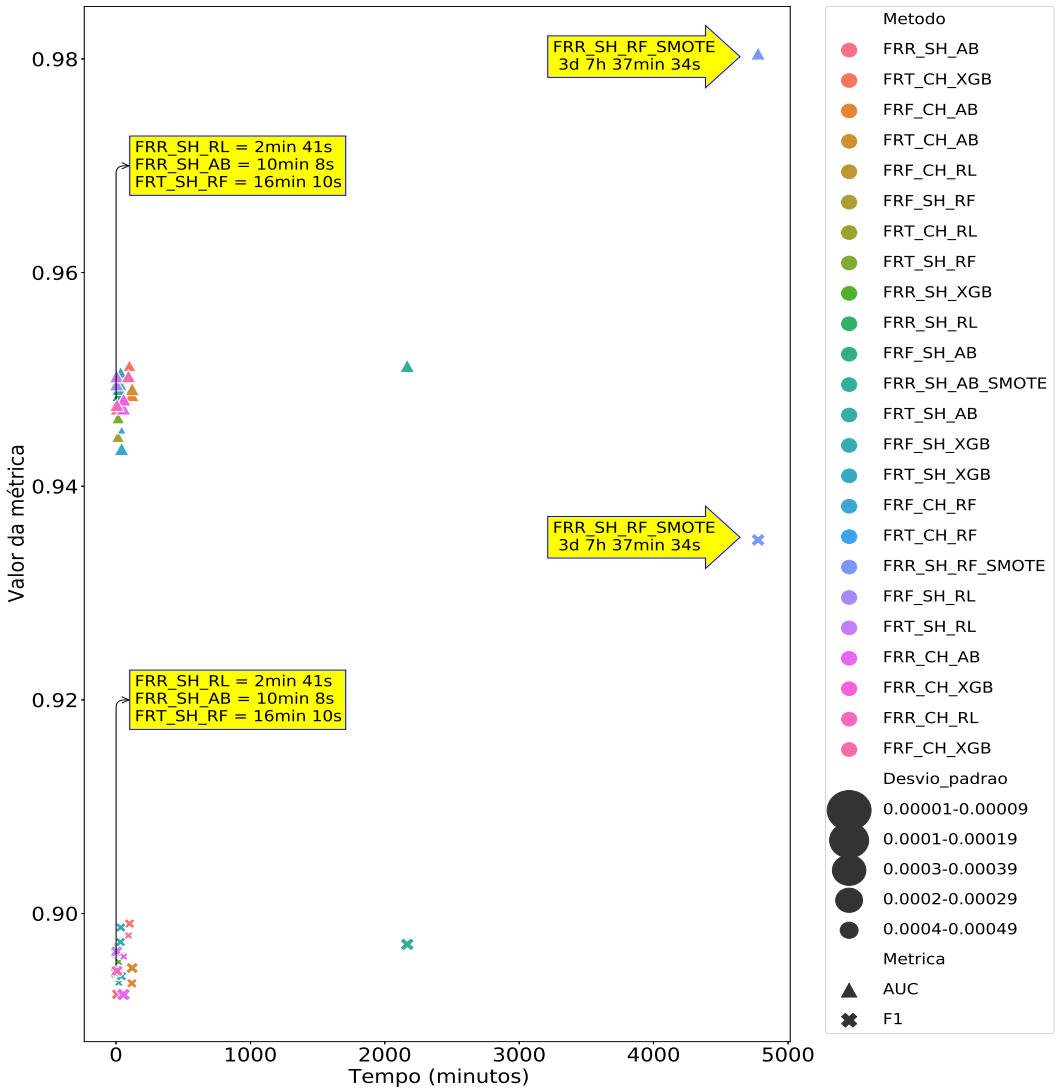
I. Dois com a técnica de sobreamostragem (SMOTE) usando o grupo de variáveis relacionados ao recém-nascido, parto e biológicos maternos (grupo de dados 3); e
II. 21 com subamostragem aleatória (*Random Undersampling*), usando os grupos de dados 3, 4 e 5.

Diante dos resultados, concluiu-se que os algoritmos apresentados na Tabela 5 não possuem diferença estatística significativa entre eles para os valores de AUC e *f1-score*. Assim, independente dos valores obtidos das métricas, foi considerado também como critério de seleção dos classificadores, o tempo de execução total para treinamento, validação e testes. Desta forma, buscou-se equalizar a relação entre os melhores resultados das métricas avaliadas com o tempo de execução total dos algoritmos. Logo, considerando-se essa relação, nota-se na Figura 12 que o método de Regressão Logística usando variáveis biológicas maternas e ligadas ao recém-nascido com subamostragem aleatória(Grupo de dados 1), mostra-se o mais indicado para aplicação no contexto da predição de óbito neonatal.

Apesar dos bons resultados e menor tempo apresentado, é importante salientar um ponto que inviabiliza a seleção do classificador de regressão logística no contexto deste trabalho que é o fato dele ser sensível a multicolinearidade (71), problema não tratado neste estudo. Portanto, considerou-se mais adequado a escolha do classificador AdaBoost usando variáveis biológicas maternas e ligadas ao recém-nascido com subamostragem aleatória, pois além de estar na relação dos melhores resultados, apresentou o menor tempo de execução.

Visando entender melhor o comportamento e o poder preditivo das variáveis do classificador selecionado foi utilizado o método *SHapley Additive exPlanations* (SHAP) para extrair as variáveis mais relevantes na predição do risco de óbito neonatal (64). Conforme a Figura 13, identificou-se que as variáveis índice de Apgar de um e cinco

Figura 12 – Comparação entre os tempos de execução para treinamento, validação e teste dos classificadores.

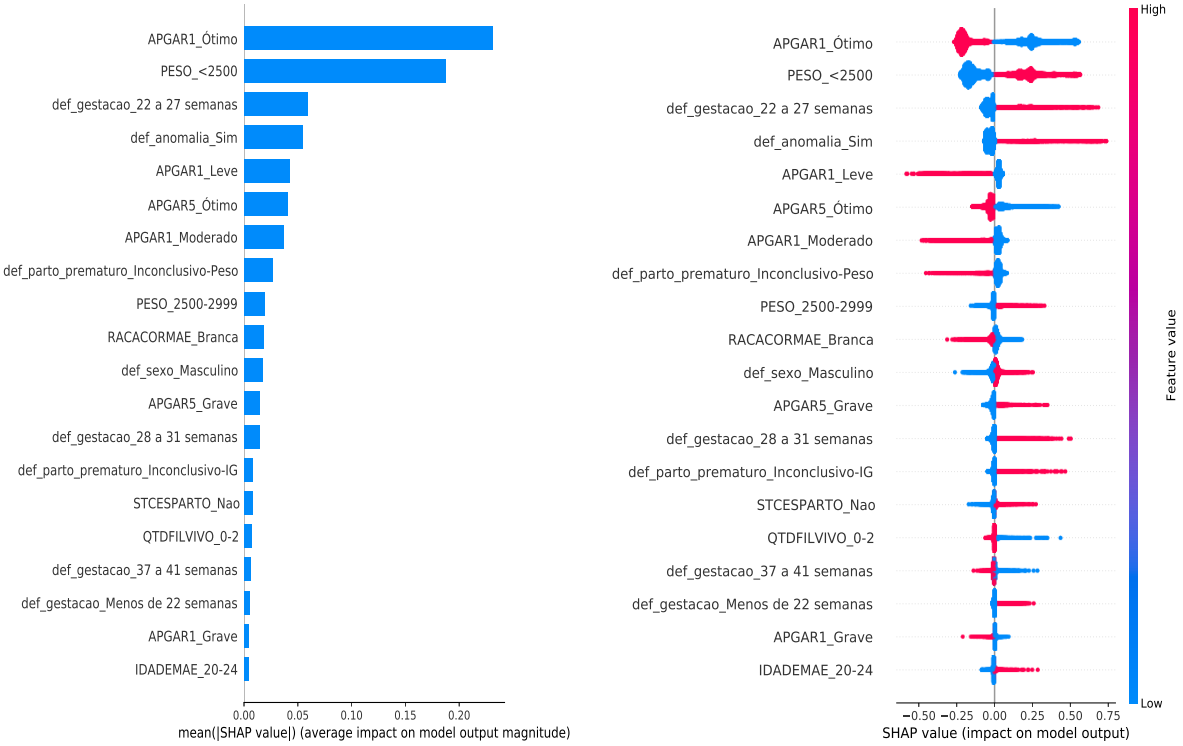


Fonte: Elaborado pelo autor (2022).

minutos, peso ao nascer, prematuridade, idade gestacional, anomalia congênita, raça da mãe e quantidade de filhos vivos são as que mais influenciam, a nível global, no desfecho de risco neonatal para esse classificador.

Em relação ao grau de relevância global das variáveis foi traçado o gráfico da Figura 14 que combina a importância das 20 principais variáveis identificadas pelo classificador AdaBoost com seus efeitos na predição no rótulo da classe. Neste gráfico, as variáveis são ordenadas de forma decrescente de acordo com sua importância e força preditiva. A posição no eixo y é determinada pela variável e no eixo x pelo valor de importância *Shapley*. Neste sentido, o gráfico mostra as relações positivas e negativas das variáveis preditoras em relação a variável alvo. Destaca-se que a interpretação e discussão dessas

Figura 13 – As 20 variáveis mais relevantes, a nível global, na predição do óbito neonatal do classificador AdaBoost usando variáveis relacionadas ao recém-nascido, parto e biológicos maternos (Grupo de dados 3).



(a) Relevância preditiva.

(b) Força preditiva.

Fonte: Elaborado pelo autor (2022).

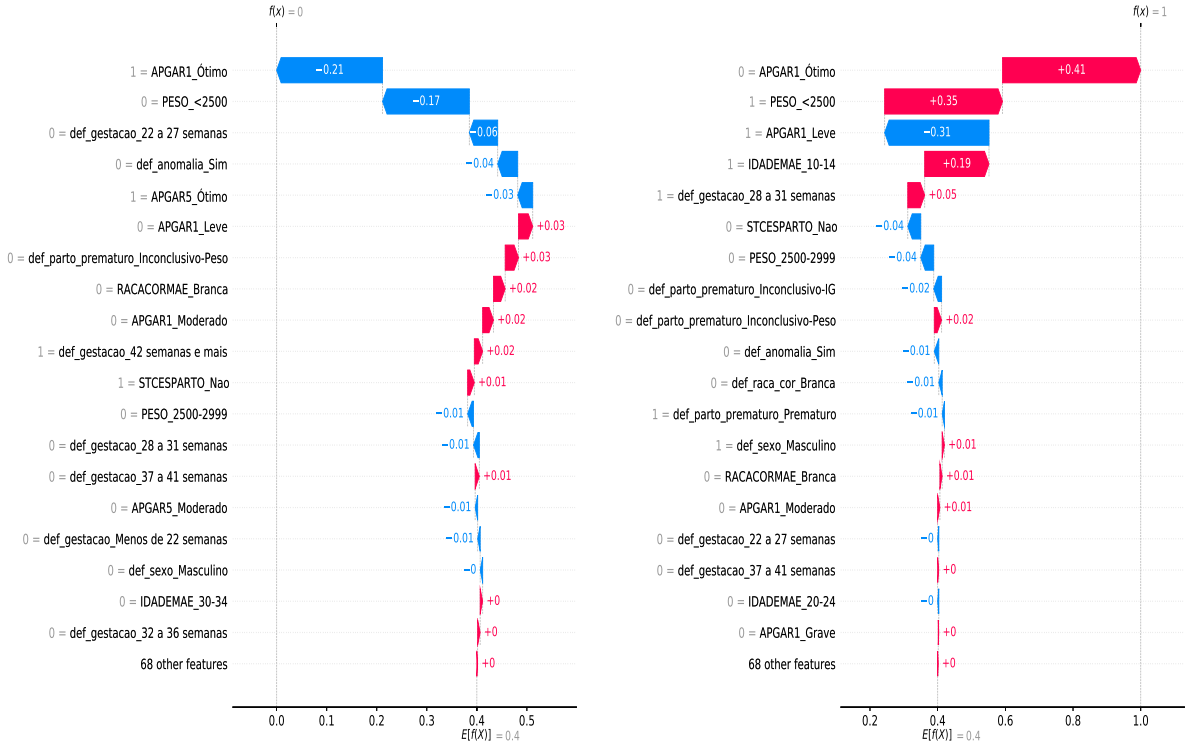
variáveis serão abordadas na seção 5.2.

Ressalta-se também a interpretação local do classificador AdaBoost visando identificar o comportamento na decisão do classificador para uma observação individual, ou seja, para uma instância ou observação do conjunto de dados. Assim, nas Figuras 14.a e 14.b são apresentados os comportamentos de decisão para um exemplo de cada caso: um de predição de sobrevivência e outro de caso de óbito, respectivamente.

Observa-se a partir da Figura 14 que o índice de Apgar de um e cinco minutos, peso e idade gestacional são realmente os atributos que mais influenciam na predição do risco de óbito neonatal. Nota-se que outros atributos aparecem com força preditiva, porém com menor expressividade.

Por outro prisma, sob a perspectiva de identificação da força preditiva das variáveis enriquecidas, extraímos as características mais relevantes do classificador AdaBoost gerado com o grupo de dados enriquecido (Grupo 5). Salienta-se que este classificador também encontra-se no rol dos melhores resultados e menor tempo de execução. As características e força preditiva das 50 variáveis mais influentes na predição são apresentadas no Apêndice G.

Figura 14 – Interpretação local das 20 variáveis mais relevantes do classificador AdaBoost usando variáveis relacionadas ao recém-nascido, parto e biológicos maternos (Grupo de dados 3).



(a) Previsão de sobrevivência. (b) Previsão de óbito.

Fonte: Elaborado pelo autor (2022).

Nota-se que além das variáveis identificadas no conjunto de dados ligados aos fatores biológicos maternos e do recém-nascido, também apresentaram força preditiva as variáveis relacionadas ao número de consultas pré-natal, se o município de residência da mãe faz parte da Amazônia Legal, o mês de gestação em que a gestante iniciou o pré-natal, além da escolaridade e da raça materna. Em relação às variáveis enriquecidas, mostraram-se com força preditiva o IDHM e a distância entre a residência e o local de nascimento. Salienta-se que a discussão sobre as variáveis identificadas como as mais relevantes, bem como os resultados obtidos perante aos trabalhos da literatura serão discutidos na próxima seção.

5.2 DISCUSSÃO

Este trabalho executou um conjunto de etapas e usou dados de diferentes fontes para construção de classificadores especializados na predição de óbito neonatal. Por este ângulo, esta seção apresenta as discussões dos pontos mais relevantes para interpretação e discussão dos resultados alcançados.

Inicialmente, menciona-se os resultados obtidos aplicando os conjuntos de dados

das dimensões da mortalidade neonatal separadamente conforme explorado no capítulo 4. Evidenciou-se que os classificadores treinados usando os conjuntos de dados que possuem variáveis ligadas aos fatores biológicos maternos e do recém-nascido, apresentam os melhores resultados. Em contrapartida, o estudo revelou que a aplicação dos dados que abarcam apenas aspectos socioeconômicos ou comportamentais e de uso dos serviços de saúde não trazem ganhos expressivos nos resultados dos classificadores, independente da métrica adotada para avaliação.

Revelou-se ainda, no escopo deste trabalho, que o classificador AdaBoost treinado com variáveis ligadas aos fatores biológicos maternos e do recém-nascido foi o mais eficiente em termos de valores de métricas e tempo de execução do algoritmo. Neste classificador, constatou-se que peso ao nascer, índice de Apgar no primeiro e quinto minuto, anomalia congênita e idade gestacional são as variáveis que mais impulsionam o poder de predição global do modelo. No caso do classificador AdaBoost com o conjunto de dados enriquecido, além das variáveis citadas, destacam-se a distância entre a residência e o local de nascimento, a quantidade de consultas pré-natal, a informação se o município de residência da mãe faz parte da Amazônia Legal, o mês de gestação em que a gestante iniciou o pré-natal, IDHM da residência da mãe, além da escolaridade e da raça materna. Todas essas variáveis se mostraram relevantes na classificação do rótulo da classe, porém, com menor relevância se comparadas aos fatores exclusivamente biológicos maternos e relacionados ao recém-nascido.

Nota-se na literatura que as variáveis identificadas realmente são expressivas e devem ser consideradas pelo sistema de saúde para atenção à gestante e ao recém-nascido. O baixo peso ao nascer e índices ruins de Apgar de um e cinco minutos, por exemplo, são alertas para possíveis complicações futuras do bebê, gerando atenção ao risco do recém-nascido ir a óbito nos primeiros dias de vida.

Em relação a variável baixo peso ao nascer, é considerada um marcador de risco de óbito neonatal e está relacionado às condições socioeconômicas precárias e ao comportamento materno em relação aos cuidados com a saúde. A Organização Mundial da Saúde (OMS) define o baixo peso ao nascer como os casos de nascimentos com menos de 2.500g (111). Ainda conforme a OMS, essa variável continua sendo um problema de saúde pública significativo em todo o mundo. Tem-se então que o baixo peso ao nascer é considerado um dos principais fatores de risco para a sobrevivência do recém-nascido (25). De acordo com (39), essa variável é apontada como o fator de maior influência na determinação da morbimortalidade neonatal, podendo estar associado a baixos níveis de desenvolvimento socioeconômico e de assistência materno-infantil. Além disso, o baixo peso ao nascer pode ser entendido como um evento sentinela[13] para os serviços de saúde, indicando a baixa qualidade da assistência pré-natal. Nos classificadores AdaBoost gerados

[13] Detecção de doença prevenível, incapacidade, ou morte inesperada, cuja ocorrência serve como um sinal de alerta de que a qualidade terapêutica ou prevenção deve ser questionada (Ministério da Saúde, 2009).

esta variável apresentou alta relevância de predição onde um peso menor que 2.500g do recém-nascido revela um impacto alto e positivo na classificação do óbito neonatal. Assim, mais um vez cita-se a necessidade de melhorar as condições assistenciais da mulher durante a gestação e na fase puerperal.

No que diz respeito ao índice de Apgar, estudos como em (3, 69, 79) mostram que valores baixos dos índices de Apgar, tanto no primeiro quanto no quinto minuto, estão relacionados a ocorrência de óbito neonatal. Considera-se que índices menores que 7, seja no primeiro ou quinto minuto, configura fator de risco para a mortalidade neonatal, sinalizando alerta para atenção especial ao recém-nascido (13, 81).

A variável anomalia congênita ou malformação congênita como também é chamada, pode ser causada por fatores genéticos, ambientais ou mesmo está ligada a fatores desconhecidos (70, 31). De acordo com (6) as malformações congênitas representam uma importante causa de mortalidade infantil em diversos países, sendo em sua grande maioria devido a ocorrências durante o primeiro ano de vida da criança. Características maternas como idade, estilo de vida, tipo de gestação e saúde materna, entre outros, têm sido pesquisados e relacionados à ocorrência de anomalias congênitas (73). Cita-se ainda a falta de assistência ou atenção adequada às mulheres na fase reprodutiva também como causas das malformações congênitas (91). Neste sentido, torna-se relevante melhorar as assistências pré-natais, assistência à saúde materno-infantil, além de ampliar as campanhas de conscientização focado neste tema.

Acerca do número de consultas pré-natal, o seu controle é considerado imprescindível para detecção precoce de intercorrências e realização de ações de prevenção de doenças durante a gestação (97). O Ministério da Saúde[14,15] preconiza que é necessário realizar no mínimo seis consultas de acompanhamento pré-natal durante toda a gravidez, considerando como ideal iniciar o acompanhamento nos primeiros três meses de gestação. Neste trabalho, constatou-se que esta variável realmente tem um forte impacto na predição do óbito neonatal, com tendência de rótulo de óbito para os casos de risco de óbito.

Outra variável relevante para a predição do óbito neonatal é a prematuridade que corresponde aos nascimentos antes de 37 semanas de gestação[16], também é reconhecida como uma variável de forte impacto no contexto do óbito neonatal, especialmente a neonatal precoce (74). A prematuridade é preocupação em saúde pública principalmente nos países menos desenvolvidos, devido às condições precárias de saúde da gestante (42). A sobrevivência de recém-nascidos prematuros e de muito baixo peso reflete a qualidade

[14] BRASIL. Ministério da Saúde. Portaria nº 570, de 1º de junho de 2000. Instituir o Componente I do Programa de Humanização no Pré-natal e Nascimento - Incentivo à Assistência Pré-natal no âmbito do Sistema Único de Saúde. Brasília, DF, 2000.

[15] BRASIL. Ministério da Saúde. Pré-Natal e Parto. Disponível em: `https://www.gov.br/saude/pt-br/assuntos/saude-de-a-a-z/s/saude-da-crianca/pre-natal-e-parto`

[16] WORLD HEALTH ORGANIZATION. Preterm birth. 2018. Disponível em: `https://www.who.int/en/news-room/fact-sheets/detail/preterm-birth`

do atendimento antenatal, do cuidado ao trabalho de parto e parto e a estrutura de atendimento neonatal das diversas regiões e países do mundo. Para os partos hospitalares dos países em desenvolvimento, a prematuridade também é o principal determinante da morbidade e mortalidade neonatal (3).

A idade gestacional também foi revelada como expressiva na predição do óbito neonatal usando o classificador Adaboost. Conforme (18) os recém-nascidos são classificados quanto a idade gestacional em I. Pré-termo: Nascido com menos de 37 semanas de gestação; II. À termo: Nascido entre 37 semanas e 41 semanas e 6 dias de gestação; e III. Pós-termo: Nascido com 42 semanas ou mais de gestação. De acordo coom (101), recente revisão sistemática da literatura apontou o nascimento pré-termo como principal causa de morte não apenas infantil, mas na infância. O nascimento de um bebê pré-termo é um evento que geralmente traz implicações de risco ao desenvolvimento saudável. Os bebês pré-termo estão sob maior risco para deficiências no desenvolvimento em relação aos bebês nascidos a termo (41). Os nascidos pré-termo têm risco aumentado de adoecer e morrer em consequência do incompleto desenvolvimento fetal e de sua maior suscetibilidade às infecções, complicadas pela manipulação e grande período de permanência nas unidades neonatais (46). Muitos evoluem com sequelas neurológicas, oftalmológicas ou pulmonares. Neste sentido é necessário melhorar a qualidade da assistência pré-natal e das ações de educação em saúde às gestantes no intuito de evitar a morbimortalidade nesse grupo de risco (62).

Para a variável intitulada de desenvolvimento humano municipal, nota-se pelo gráfico de importância apresentado no Apêndice G que mesmo um município com IDHM alto, há uma tendência da classe ser rotulada como chance de óbito. Este cenário, dependendo da localização, deve-se aos fatores socioeconômicos e assistenciais serem menos relevantes em relação aos fatores biológicos materno e ligados ao recém-nascido, refletindo o efeito protetor do elevado desenvolvimento econômico e social da região (56).

Quanto a variável enriquecida que categoriza a distância entre a residência da gestante e o local de nascimento, essa também se mostrou relevante no poder de classificação da variável alvo. É fato que a peregrinação para o parto é um problema enfrentado por muitas mulheres no Brasil, principalmente às residentes em locais remotos e carentes de infraestrutura em saúde (109). Além disso, muitas mulheres enfrentam obstáculos financeiros, sociais, culturais e de falta de informações adequadas que dificultam o acesso à assistência médica durante a gravidez e no parto. Diante disso, torna-se essencial à atenção do poder público na promoção de políticas públicas que visem garantir o acesso universal aos serviços de saúde para às gestantes em todas regiões do país.

Os classificadores também sinalizaram que o município de residência da mãe estar inserida, ou não, na região da Amazônia Legal também influencia no poder de predição da classe alvo. Considera-se que a região da Amazônia Legal é dinâmica e heterogênea, onde

prevalece um precário processo de urbanização e projetos de desenvolvimento voltados para a exploração de recursos naturais. Nessa região são encontrados diversos problemas de saúde, principalmente de doenças infecciosas e parasitárias que podem acarretar em problemas às gestantes e ao recém-nascido (108).

O mês de gestação de início do pré-natal também foi apontado como uma variável importante na predição global do risco neonatal. Segundo o Programa de Humanização no Pré-natal e Nascimento (PHPN)[17], a primeira consulta de pré-natal deve acontecer até o 4º mês de gestação. Destaca-se que quanto mais cedo se iniciar o acompanhamento pré-natal melhor é a avaliação médica e identificação antecipada de possíveis problemas durante a gestação, possibilitando ainda, a realização do número de consultas e exames pré-natais necessários (84). É um aspecto fundamental para uma boa qualidade de acompanhamento do pré-natal e que como revelado no estudo impacta o poder de decisão do classificador gerado.

No que tange a importância do grau de escolaridade da gestante em relação ao risco de óbito neonatal, nota-se no presente estudo que também foi revelada como importante no comportamento de decisão dos classificadores. Um nível educacional materno elevado permite a capacidade de aquisição de conhecimento em assuntos de saúde e o uso otimizado dos serviços de saúde (43). Estudos como em (40) mostram que adolescentes e mulheres com mais de 35 anos, de baixa escolaridade, são grupos mais vulneráveis em relação ao risco de óbito neonatal. O estudo de (52) também revelou desigualdades entre as taxas de mortalidade neonatal para filhos de mães com alta e baixa escolaridade, expondo este último como mais vulnerável aos riscos neonatais.

Em relação ao desempenho dos classificadores conforme mostrado na seção 5.1, um total de 24 classificadores apresentaram resultados com equivalência estatística entre os resultados. Foram selecionados o de menor tempo para treinamento, validação e teste (Regressão Logística usando dados com fatores biológicos maternos e ligadas ao recém-nascido aplicando subamostragem aleatória) e, também, o que apresentou os maiores valores das métricas (Random Forest usando conjunto de dados integrado e enriquecido aplicando SMOTE). Ambos, apresentaram valores acima de 94% de AUC, resultado que se mostrou superior ao trabalho de (90), (4) e (98). Em termos de *f1-score*, os resultados se revelaram superiores ao trabalho de (9) e equivalentes ao apresentado em (75). Além disso, ressalta-se que adota uma abordagem mais abrangente para modelar um classificador de risco neonatal, pois abrange mais variáveis do que aquelas presentes nos sistemas de informação de saúde do SUS e que são comumente usadas nos estudos demográficos e epidemiológicos com foco em mortalidade neonatal. Este cenário já se diferencia do

[17] Programa instituído pela Portaria nº 569, de 1º de junho de 2000 do Ministério da Saúde, com o objetivo de assegurar a melhoria no acompanhamento do parto e puerpério é o direito à humanização da assistência obstétrica e neonatal. Disponível em: `https://bvsms.saude.gov.br/bvs/saudelegis/gm/2000/prt0569_01_06_2000_rep.html`

trabalho apresentado em (4). E como proposta de extensão do estudo apresentado em (75), foi adicionada a classificação do recém-nascido em pequeno, adequado ou grande para a idade gestacional. Esta variável é importante para identificar problemas específicos que cada um desses grupos costuma apresentar (17). Além disso, outro diferencial do trabalho em relação aos demais, foi realizar o estudo por grupo ou componente da mortalidade neonatal, onde foi possível verificar separadamente e de maneira unificada, o desempenho dos classificadores e a relevância dos fatores de risco do óbito neonatal.

6 CONCLUSÃO

Apesar do declínio da taxa de mortalidade infantil ao longo dos últimos anos no Brasil, ainda continua sendo um desafio significativo diminuí-lo principalmente em algumas regiões do país que carecem de infraestrutura de saúde, presença de peregrinação para o parto e falta de informações adequadas em saúde para a população. Neste sentido torna-se relevante direcionar estudos visando diminuir ainda mais os índices de mortalidade infantil, principalmente do componente neonatal, pois ainda é o mais prevalente entre os óbitos no primeiro ano de vida. Conforme apresentado no trabalho, o risco de óbito neonatal é evidenciado pela relação entre os fatores biológicos, socioeconômicos e assistenciais relacionados à gestante e ao recém-nascido. Assim, faz-se necessário avançar em estudos que possam auxiliar e dar apoio a este crítico problema de saúde pública.

O presente estudo visou construir classificadores capazes de predizer o risco de óbito neonatal com alta taxa de sensibilidade e precisão. O trabalho também visou identificar as principais variáveis ligadas ao óbito neonatal com base em abordagens de mineração de dados e aprendizado de máquina usando dados públicos do sistema de saúde e censo demográfico brasileiro. O trabalho foi pautado na formação de um conjunto de dados abrangendo mais de 8 milhões de registros de nascidos vivos da população brasileiras nos anos de 2012 a 2014. Ademais, o conjunto de dados foi enriquecido com novas variáveis para uma investigação mais ampla dos fatores influenciadores do desfecho de risco neonatal.

Os resultados revelam que os classificadores propostos possibilitam a identificação precoce do risco de óbito neonatal com uma AUC de 94% e taxa de *f1-score* de 89%, auxiliando na identificação antecipada de possíveis óbitos neonatais. Esses resultados podem gerar impactos positivos ao sistema de saúde público brasileiro servindo como mais um recurso para aprimoramento do sistema. Os resultados mostram ainda que as variáveis relacionadas aos fatores biológicos maternos e ao recém-nascido são as que mais influenciam na predição de risco de óbito neonatal com destaque ao índice de Apgar de um e cinco minutos, peso ao nascer, prematuridade, idade gestacional, anomalia congênita, raça da mãe e quantidade de filhos vivos.

Evidenciou-se também que as variáveis enriquecidas: índice de desenvolvimento humano municipal e distância da residência da gestante até o local de nascimento, possuem força preditiva e devem ser levadas em consideração não só para construção do classificador mas para direcionar atenção a assistência à saúde da gestante e aplicação de políticas públicas.

Como proposta de trabalhos futuros, pretende-se investigar com mais detalhes as correlações entre as variáveis presentes em cada componente da mortalidade neonatal e o impacto que podem gerar no poder preditivo dos classificadores. Pretende-se também pesquisar os efeitos na classificação quando são aplicadas técnicas de redução de

dimensionalidade aos dados. Almeja-se ainda buscar mais fontes de dados com novas variáveis influenciadoras da mortalidade neonatal que possam melhorar os resultados dos classificadores e os entendimentos deste importante indicador de saúde pública.

REFERÊNCIAS

1 AKHIL, J.; SAMREEN, S.; ALUVALU, R. **The Future of Health care: Machine Learning**. International Journal of Engineering and Technology, v. 7, p. 23-25, 2018.

2 ALBUQUERQUE, J. P. de; PRADO, E. P. V.; MACHADO, G. R. **Ambivalent implications of health care information systems: a study in the Brazilian public health care system**. Revista de Administração de Empresas, v. 51, n. 1, p. 58-71, 2011.

3 ALMEIDA, M. F. B. de et al. **Fatores perinatais associados ao óbito precoce em prematuros nascidos nos centros da Rede Brasileira de Pesquisas Neonatais**. Jornal de Pediatria, v. 84, n. 4, p. 300-307, 2008.

4 ALVES, L. C. et al. **Assessing the Performance of Machine Learning Models to Predict Neonatal Mortality Risk in Brazil, 2000-2016**. medRxiv, 2020.

5 ALVES, T. F.; COELHO, A. B. **Mortalidade infantil e gênero no Brasil: uma investigação usando dados em painel**. Ciência & Saúde Coletiva, v. 26, n. 4, p. 1259–1264, 2021.

6 AMORIM, M. M. R. de et al. **Impacto das malformações congênitas na mortalidade perinatal e neonatal em uma maternidade-escola do Recife**. Revista Brasileira de Saúde Materno Infantil, v. 6, p. 19-25, 2006.

7 ANELE, C. R. et al. **The influence of the municipal human development index and maternal education on infant mortality: an investigation in a retrospective cohort study in the extreme south of Brazil**. BMC Public Health, v. 21, n. 194, 2021.

8 BARONI, L. et al. **Neonatal mortality rates in Brazilian municipalities: from 1996 to 2017**. BMC Research Notes, v. 14, n. 1, p. 55, 2021.

9 BATISTA, A. F. M. et al. **Neonatal mortality prediction with routinely collected data: a machine learning approach**. BMC Pediatrics, v. 21, n. 1, p. 322, 2021.

10 BELUZO, C. E. et al. **Machine Learning to Predict Neonatal Mortality Using Public Health Data from São Paulo - Brazil**. medRxiv, 2020.

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

12 BITTAR, O. et al. **Sistemas de informação em saúde e sua complexidade**. Revista de Administração em Saúde, v. 18, 2018.

13 BITTENCOURT, R. M.; GAÍVA, M. A. M. **Mortalidade neonatal precoce relacionada a intervenções clínicas**. Revista Brasileira de Enfermagem, v. 67, n. 2, p. 195-201, 2014.

14 BORGES, G. M.; SANTOS, R. V. **Desigualdades territoriais na mortalidade infantil no Brasil: revisão sistemática da literatura**. Revista Brasileira de Estudos de População, v. 36, n. 3, p. 1-23, 2019.

15 BRASIL, Ministério da Saúde. **Política nacional de atenção integral à saúde da mulher: princípios e diretrizes**. Brasília, 2004. Disponível em: `https://conselho.saude.gov.br/ultimas_noticias/2007/politica_mulher.pdf`. Acesso em: 6 jun. 2022.

16 BRASIL, Ministério da Saúde. **Manual de Vigilância do Óbito Infantil e Fetal e do Comitê de Prevenção do Óbito Infantil e Fetal**. 2. ed. Brasília, 2009. Disponível em: `https://bvsms.saude.gov.br/bvs/publicacoes/manual_obito_infantil_fetal_2ed.pdf`. Acesso em: 2 mai. 2022.

17 BRASIL, Ministério da Saúde. **Guia de Orientações para o Método Canguru na Atenção Básica: Cuidado Compartilhado**. Brasília, 2016. Disponível em: `https://bvsms.saude.gov.br/bvs/publicacoes/guia_orientacoes_metodo_canguru.pdf`. Acesso em: 10 de jul. 2022.

18 BRASIL, Ministério da Saúde. Método Canguru. **Guia de Orientações para o Método Canguru na Atenção Básica: Cuidado Compartilhado**. Brasilia, 2016.

19 BRASIL, Ministério da Saúde. **Mortalidade infantil no Brasil**. 2021. Disponível em: `https://www.gov.br/saude/pt-br/centrais-de-conteudo/publicacoes/boletins/epidemiologicos/edicoes/2021/boletim_epidemiologico_svs_37_v2.pdf`. Acesso em: 2 mai. 2022.

20 BRASIL, Ministério da Saúde. Departamento de Informática do SUS (DATASUS). **Indicadores e Dados Básicos 2012 (IDB-2012)**. Brasília, 2022a. Disponível em: `https://datasus.saude.gov.br/informacoes-de-saude-tabnet/`. Acesso em: 8 mar. 2022.

21 BRASIL, Ministério da Saúde. **Estrutura, princípios e como funciona**. Brasília, 2022b. Disponível em: `https://www.gov.br/saude/pt-br`. Acesso em: 2 mai. 2022.

22 BREIMAN, L. **Random forests. Machine Learning**, v. 45, n. 1, p. 5-32, 2001.

23 BREUEL, C. **Carreiras em Machine Learning: O Presente e o Futuro**. Sociedade Brasileira de Computação. Revista Horizontes, 2020.

24 BUDACH, L. et al. **The Effects of Data Quality on Machine Learning Performance**. ArXiv, 2022.

25 CAPELLI, J. de C. S. et al. **Baixo peso ao nascer e fatores associados ao prénatal: estudo seccional em uma maternidade de referência de Macaé**. Saúde em Redes. v. 6, n. 1, p. 163-173, 2020.

26 CASTRO, C. L.; BRAGA, A. P. **Aprendizado supervisionado com conjuntos de dados desbalanceados**. Sba: Controle & Automação Sociedade Brasileira de Automatica, v. 22, n. 5, p. 441–466, 2011.

27 CHAWLA, N. V. et al. **SMOTE: synthetic minority over-sampling technique**. Journal of Artificial Intelligence Research. v. 16, p. 321-357, 2002.

28 CHEN, M.; MAO, S.; LIU, Y. **Big Data: A Survey**. Mobile Networks and Applications, v. 19, p. 171-209, 2014.

29 CHEN, T.; GUESTRIN, C. **XGBoost: A Scalable Tree Boosting System**. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, p. 785–794, 2016.

30 CHUNG, C. et al. **Predicting neonatal mortality using machine learning and electronic health records**. PLoS One, v. 14, n. 12, p. e0225823, 2019.

31 COSME, H. W.; LIMA, L. S.; BARBOSA, L. G. **Prevalência de anomalias congênitas e fatores associados em recém-nascidos do município de São Paulo no período de 2010 a 2014**. Revista Paulista De Pediatria, v. 35, n. 1, p. 33-38, 2017.

32 CUNHA, E. M. da; VARGENS, J. M. da C. **Sistemas de informação do Sistema Único de Saúde**. *In*: GONDIM, Grácia Maria de Miranda; CHRISTÓFARO, Maria Auxiliadora Córdova; MIYASHIRO, Gladys Miyashiro (Org.). Técnico de vigilância em saúde: fundamentos. V. 2. Rio de Janeiro: EPSJV, p. 71-112, 2017.

33 DEMSAR, J. **Statistical comparisons of classifiers over multiple data sets**, Journal of Machine Learning Research, v. 7, p. 1-30, 2006.

34 DENG, X. et al. **An improved method to construct basic probability assignment based on the confusion matrix for classification problem**. Information Sciences, v. 340-341, p. 250-261, 2016.

35 DUARTE, J. L. M. B.; MENDONÇA, G. A. S. **Fatores associados à morte neonatal em recém-nascidos de muito baixo peso em quatro maternidades no Município do Rio de Janeiro, Brasil**. Cadernos De Saúde Pública, v. 21, n. 1, p. 181-191, 2005.

36 FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. **The KDD process for extracting useful knowledge from volumes of data**. Communications of the ACM, v. 39, n. 11, p. 27-34, 1996.

37 FAYYAD et al. **Advances in knowledge discovery and data mining**. American Association for Artificial Intelligence. USA, 1996.

38 FERNANDES, F. T.; CHIAVEGATTO FILHO, A. D. P. **Perspectivas do uso de mineração de dados e aprendizado de máquina em saúde e segurança no trabalho**. Revista Brasileira De Saúde Ocupacional, v. 44, e13, 2019.

39 FERRAZ, Thaise da Rocha; NEVES, Eliane Tatsch. **Fatores de risco para baixo peso ao nascer em maternidades públicas: um estudo transversal**. Revista Gaúcha De Enfermagem, v. 32, 2011.

40 FONSECA, S. C. et al. **Maternal education and age: inequalities in neonatal death**. Revista De Saúde Pública, v. 51, p. 94, 2017.

41 FORMIGA, C. K. M. R.; LINHARES, M. B. M. **Avaliação do desenvolvimento inicial de crianças nascidas pré-termo**. Revista Da Escola De Enfermagem Da USP, v. 43, n. 2, p. 472-480, 2009.

42 FREITAS, B. A. C. et al. **Características epidemiológicas e óbitos de prematuros atendidos em hospital de referência para gestante de alto risco**. Revista Brasileira De Terapia Intensiva, v. 24, n. 4, p. 386-392, 2012.

43 GAKIDOU, E. et al. **Increased educational attainment and its effect on child mortality in 175 countries between 1970 and 2009: a systematic analysis**. The Lancet, v. 376, n. 9745, p. 959-974, 2010.

44 GARCIA, L. P.; FERNANDES, C. M.; TRAEBERT, J. **Risk factors for neonatal death in the capital city with the lowest infant mortality rate in Brazil**. Jornal De Pediatria, v. 95, n. 2, p. 194-200, 2019.

45 GEBREMARIAM, A. D. et al. **Development and Validation of a Clinical Prognostic Risk Score to Predict Early Neonatal Mortality, Ethiopia: A Receiver Operating Characteristic Curve Analysis**. Clinical epidemiology, v. 13, p. 637-647, 2021.

46 GUIMARÃES, E. A. de A. et al. **Prevalência e fatores associados à prematuridade em Divinópolis, Minas Gerais, 2008-2011: análise do Sistema de Informações sobre Nascidos Vivos**. Epidemiologia e Serviços de Saúde, v. 26, n. 1, p. 91-98, 2017.

47 HARRIS, J. K. **Primer on binary logistic regression**. Family medicine and community health, v. 9, Suppl 1, e001290, 2021.

48 HATWELL, J.; GABER, M. M.; ATIF AZAD, R. M. **Ada-WHIPS: explaining AdaBoost classification with applications in the health sciences**. BMC Medical Informatics and Decision Making, v. 20, p. 250, 2020.

49 HILBERT, M. **Big Data for Development: A Review of Promises and Challenges**. Development Policy Review, v. 34, p. 135-174, 2016.

50 HOLLANDER, M.; WOLFE, D. A.; CHICKEN, E. **Nonparametric Statistical Methods**. ed. 3. Wiley. 2014.

51 INOUE, T. et al. **XGBoost, a Machine Learning Method, Predicts Neurological Recovery in Patients with Cervical Spinal Cord Injury**. Neurotrauma Reports, v. 1, n. 1, p. 8-16, 2020.

52 KALE, P. L. et al. **Fetal and infant mortality trends according to the avoidability of causes of death and maternal education**. Revista Brasileira De Epidemiologia, v. 24, e210008, 2021.

53 KASSAR, Samir Buainain. **Mortalidade Neonatal em Maceió-AL: Evolução e Fatores de Risco**. 2010. Tese (Doutorado em Saúde da Criança e do Adolescente) - Centro de Ciências da Saúde, Universidade Federal de Pernambuco, Pernambuco, 2010.

54 KITCHIN, R. **Big Data, new epistemologies and paradigm shift**. Big Data & Society, v. 1, n. 1, 2014.

55 KOURENTZES, N. **Intermittent demand forecasts with neural networks**. International Journal of Production Economics, v. 143, n. 1, p. 198-206, 2013.

56 KROPIWIEC, M. V. et al. **Fatores associados à mortalidade infantil em município com índice de desenvolvimento humano elevado**. Revista Paulista De Pediatria, v. 35, n. 4, p. 391-398, 2017.

57 LANSKY, S. et al. **Pesquisa Nascer no Brasil: perfil da mortalidade neonatal e avaliação da assistência à gestante e ao recém-nascido**. Cadernos De Saúde Pública, v. 30, p. 192-207, 2014.

58 LAUDON, K. C.; LAUDON J. P. **Sistemas de Informação Gerenciais**. Pearson. 11 ed. São Paulo. Pearson Education do Brasil, 2014.

59 LIASHCHYNSKYI, P.; LIASHCHYNSKYI, P. **Grid Search, Random Search, Genetic Algorithm: A Big Comparison for NAS**. arXiv, 2019.

60 LIMA, E. de F. A. et al. **Fatores de risco para mortalidade neonatal no município de Serra, Espírito Santo**. Revista Brasileira De Enfermagem, v. 65, n. 4, p. 578-585, 2012.

61 LIMA, J. C. et al. **Estudo de base populacional sobre mortalidade infantil**. Ciência & Saúde Coletiva, v. 22, n. 3, p. 931-939, 2017.

62 LOPES, E. L.; BEZERRA, M. M. M. **Condutas Preventivas ao Parto Prematuro na Atenção Primária a Saúde**. Revista Multidisciplinar de Psicologia. v. 14, n. 53, p. 1154-1164, 2020.

63 LUDERMIR, T. B. **Inteligência Artificial e Aprendizado de Máquina: estado atual e tendências**. Revista Estudos Avançados, São Paulo, v. 35, n. 101, p. 85-94, 2021.

64 LUNDBERG, S. M.; LEE, S. **A Unified Approach to Interpreting Model Predictions**. Advances in Neural Information Processing Systems, v. 30, 2017.

65 MA, J. et al. **Application of the XGBoost Machine Learning Method in PM2.5 Prediction: A Case Study of Shanghai**. Aerosol and Air Quality Research, v. 20, 2019.

66 MANGOLD, C. et al. **Machine Learning Models for Predicting Neonatal Mortality: A Systematic Review**. Neonatology, v. 118, n. 4, p. 394-405, 2021.

67 MARGOTTO, P. R. **Assistência ao Recém-Nascido de Risco**. 4. ed. Hospital Materno Infantil de Brasília, Brasília, 2021.

68 MELO, E. C. P.; SILVA, K. S.; SALDANHA, A. A. W. **Desigualdades regionais na mortalidade neonatal no Brasil**. Revista de Saúde Pública, São Paulo, v. 54, p. 69, 2020.

69 MELO, W. A. de et al. **Anomalias congênitas: fatores associados à idade materna em município sul brasileiro, 2000 a 2007**. Revista Eletrônica de Enfermagem. v. 12, n. 1, 2010.

70 MENDES, I. C. et al. **Anomalias congênitas e suas principais causas evitáveis: uma revisão**. Revista Médica de Minas Gerais. v. 28, 2018.

71 MIDI, H.; SARKAR, S.; RANA, S. **Collinearity diagnostics of binary logistic regression model**. Journal of Interdisciplinary Mathematics. v. 13, p. 253-267, 2013.

72 MIGOTO, M. T. et al. **Early neonatal mortality and risk factors: a case-control study in Paraná state**. Revista Brasileira de Enfermagem, v. 71, 2018.

73 MOORE, K. L.; PERSAUD, T. V. **Defeitos Congênitos Humanos**. In: MOORE, K. L.; PERSAUD, T. V. Embriologia Clínica. 6. ed. Rio de Janeiro: Guanabara Koogan, p. 161-193, 2000.

74 MORAIS, A.; PEREIRA, A. **Mortalidade neonatal precoce em um hospital terciário do nordeste brasileiro**. Revista da Sociedade Brasileira de Enfermeiros Pediatras. v. 19, p. 89-96, 2020.

75 MOREIRA, J. R. H.; BERNARDINO, H. S.; VIEIRA, A. B. **Predição de Óbito Neonatal usando Dados dos Sistemas de Informação do SUS e de Censo Demográfico**. Simpósio Brasileiro de Computação Aplicada à Saúde (SBCAS), v. 22, p. 234-245, 2022.

76 MOSLEY, W. H.; LINCOLN C. C. **An Analytical Framework for the Study of Child Survival in Developing Countries**. Population and Development Review, v. 10, p. 25–45, 1984.

77 NASCIMENTO, Renata Mota do et al. **Determinantes da mortalidade neonatal: estudo caso-controle em Fortaleza, Ceará, Brasil**. Cadernos de Saúde Pública. v. 28, 2012.

78 OBERMEYER, Z.; EMANUEL, E. J. **Predicting the Future — Big Data, Machine Learning, and Clinical Medicine**. The New England Journal of Medicine, v. 375, n. 13, p. 1216-1219, 2016.

79 OLIVEIRA, A. R. R.; LLERENA JUNIOR, J. C.; COSTA, M. F. S. **Perfil dos óbitos de recém-nascidos ocorridos na sala de parto de uma maternidade do Rio de Janeiro, 2010-2012**. Epidemiologia e Serviços de Saúde, Brasília, v. 22, n. 3, p. 501-508, 2013.

80 OLIVEIRA, G. P. de et al. **Accuracy of probabilistic and deterministic record linkage: the case of tuberculosis**. Revista de Saúde Pública. v. 50, 2016.

81 OLIVEIRA, T. G. et al. **Apgar score and neonatal mortality in a hospital located in the southern area of São Paulo city, Brazil**. Einstein (São Paulo), v. 10, n. 1, p. 22-28, 2012.

82 PAIXAO, A. B. da; MARTINS, M. M. F. **Perfil de óbitos neonatais em uma região do estado da Bahia**. Revista Saúde.Com, v. 17, n. 2, 2021.

83 PARK, J.; LEE, D. H. **Privacy Preserving k-Nearest Neighbor for Medical Diagnosis in e-Health Cloud**. Journal of Healthcare Engineering, 2018.

84 PEDRAZA, D. F.; ROCHA, A. C. D.; CARDOSO, M. V. L. M. L. **Assistência pré-natal e peso ao nascer: uma análise no contexto de unidades básicas de saúde da família**. Revista Brasileira De Ginecologia E Obstetrícia, v. 35, n. 8, p. 349–356, 2013.

85 PENNA, L. H. G. et al. **Uso de big data e inteligência artificial para aprimorar a assistência à saúde materno-infantil**. Epidemiologia e Serviços de Saúde, v. 28, n. 1, 2019.

86 PICCOLO, D. M. **Qualidade de dados dos sistemas de informação do datasus: análise crítica da literatura**. Ciência da Informação em Revista, v. 5, n. 3, p. 13-19, 2018.

87 PRASETYO, H. D.; HOGANTARA, P. A.; ISNAINIYAH, I. N.. **A Web-Based Diabetes Prediction Application Using XGBoost Algorithm**. Data Science: Journal of Computing and Applied Informatics, v. 5, n. 2, p. 49-59, 2021.

88 PRIYADARSHINI, I.; COTTON, C. **A novel LSTM-CNN-grid search-based deep neural network for sentiment analysis**. The Journal of Supercomputing, v. 77, n. 12, p. 13911-13932, 2021.

89 RAMAKRISHNAN, R.; RAO, S.; HE, J. R. **Perinatal health predictors using artificial intelligence: A review**. Women's Health (London, England), v. 17, 2021.

90 RAMOS, R. et al. **Using predictive classifiers to prevent infant mortality in the Brazilian northeast**. In: 2017 IEEE 19th International Conference on e-Health Networking, Applications and Services (Healthcom), p. 1-6, 2017.

91 RODRIGUES, L. S. et al. **Características das crianças nascidas com malformações congênitas no município de São Luís, Maranhão, 2002-2011**. Epidemiologia e Serviços de Saúde, v. 23, n. 2, p. 295-304, 2014.

92 RUSDAH, D. A.; MURFI, H. **XGBoost in handling missing values for life insurance risk prediction**. SN Applied Sciences, v. 2, n. 8, p. 1336, 2020.

93 SAINI, A.; MEITEI, A. J.; SINGH, J. **Machine Learning in Healthcare: A Review**. In: Proceedings of the International Conference on Innovative Computing & Communication (ICICC), 2021.

94 SALMON, B. P. et al. **Proper comparison among methods using a confusion matrix**. In: 2015 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Milan, Italy, p. 3057-3060, 2015.

95 SAMUEL, A. 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.

96 SETTOUTI, N.; BECHAR, M.; CHIKN, M. **Statistical Comparisons of the Top 10 Algorithms in Data Mining for Classi cation Task**. International Journal of Interactive Multimedia and Artificial Inteligence, v. 4, p. 46-51. 2016

97 SILVA, A. C. F. A. da et al. **Índice de morte neonatal precoce: Uma análise do perfil materno**. Revista Eletrônica Acervo Saúde. n. 26, p. e690, 2019.

98 SILVA, A. M.; RODRIGUES, Y. R.; ISHII, R. P. **RIGOR: A New Proposal for Predicting Infant Mortality in Government Health Systems Using Artificial Intelligence in Brazil**. In Computer, v. 53, n. 10, p. 69-76, 2020.

99 SILVA, V. A. et al. **Desigualdades socioeconômicas: uma análise sobre os determinantes da taxa de mortalidade infantil nos municípios brasileiros**. Revista Brasileira de Estudos Regionais e Urbanos, v. 13, n. 1, p. 73–97, 2019.

100 SILVER, D. L. **Knowledge discovery and data mining**. Technical Report MBA6522. CogNova Technologies London Health Science Center, 1996.

101 SLEUTJES, F. C. M. et al. **Fatores de risco de óbito neonatal em região do interior paulista, Brasil**. Ciência & Saúde Coletiva, Rio de Janeiro, v. 23, n. 8, p. 2713-2720, 2018.

102 SOARES, W. L. G. et al. **Caracterizando a Mortalidade Infantil utilizando técnicas de Machine Learning: um Estudo de Caso em dois Estados Brasileiros - Santa Catarina e Amapá**. Brazilian Journal of Development, v. 7, n. 5, p. 45269–45290, 2021.

103 SONG, Y. Y.; LU, Y. **Decision tree methods: applications for classification and prediction**. Shanghai archives of psychiatry, v. 27, n. 2, p. 130-135, 2015.

104 SOUSA, M.; MATTOSO, M.; EBECKEN, N. F. **Data Mining: A Database Perspective**. WIT Transactions on Information and Communication Technologies, v. 22, 1970.

105 STOLTZFUS, J. C. **Logistic regression: a brief primer**. Academic emergency medicine : official journal of the Society for Academic Emergency Medicine. v. 18, p. 1099–1104, 2011.

106 UDDIN, S. et al. **Comparing different supervised machine learning algorithms for disease prediction**. BMC Medical Informatics and Decision Making, v. 19, n. 1, p. 281, 2019.

107 VASSILIOU, A. G. et al. **Health in All Policy Making Utilizing Big Data**. Acta Informatica Medica, v. 28, n. 1, p. 65-70, 2020.

108 VIANA, R. L.; FREITAS, C. M. de; GIATTI, L. L. **Saúde ambiental e desenvolvimento na Amazônia legal: indicadores socioeconômicos, ambientais e sanitários, desafios e perspectivas**. Saúde e Sociedade, v. 25, n. 1, p. 233-246, 2016.

109 VIELLAS, E. F. et al. **Assistência pré-natal no Brasil**. Cadernos de Saúde Pública, v. 30, p. 85-100, 2014.

110 WORLD HEALTH ORGANIZATION. **Monitoring the building blocks of health systems: a handbook of indicators and their measurement strategies**. Genebra: WHO, 2010. Disponível em: `https://apps.who.int/iris/handle/10665/258734`. Acesso em: 12 de jul. 2022.

111 WORLD HEALTH ORGANIZATION. **Global nutrition targets 2025: low birth weight policy brief**. Genebra: WHO, 2014. Disponível em: `https://www.who.int/publications-detail-redirect/WHO-NMH-NHD-14.5`. Acesso em: 5 de ago. 2022.

112 XAVIER, R. M. A. **Análise da sobrevida dos pacientes submetidos à cirurgia cardíaca valvar no Brasil no Sistema Único de Saúde entre 2001 e 2007**. 2015. Tese (Doutorado do Programa de Pós-graduação em Medicina), Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2015.

APÊNDICE A – Variáveis dos sistemas de informações de saúde e de censo demográfico brasileiro.

Tabela 6 – Variáveis do Sistema de Informações sobre Nascidos Vivos presentes no conjunto de dados disponibilizado na Plataforma de Ciência de Dados aplicada à Sáude da Fiocruz.

Variável	**Tipo**	**Descrição**
NUMERODN	int8	Número de declaração de nascido vivo
CODINST	int8	Sem descrição
ORIGEM	int8	Sem descrição
NUMERODV	int8	Sem descrição
PREFIXODN	int8	Sem descrição
CODESTAB	text	Código de estabelecimento
CODMUNNASC	int8	Município de ocorrência, em codificação idêntica a de CODMUNRES, conforme tabela TABMUN
LOCNASC	int8	Local de ocorrência do nascimento, conforme a tabela: 9: Ignorado; 1: Hospital; 2: Outro Estab Saúde; 3: Domicílio; 4: Outros
def_loc_nasc	text	Local de nascimento (Nominal, com as seguintes classificações: Hospital; Outros estabelecimentos de saúde; Domicílio; Via pública; Outros; Ignorado)
IDADEMAE	int8	Idade da mãe em anos
ESTCIVMAE	int8	Estado civil, conforme a tabela: 1: Solteira; 2: Casada; 3: Viuva; 4: Separado judicialmente/Divorciado; 5: União consensual (versões anteriores); 9: Ignorado
def_est_civil	text	Estado civil (Situação conjugal: Solteiro; Casado; Viúvo; Separado judicialmente/divorciado; União estável; Ignorado)
ESCMAE	int8	Escolaridade, anos de estudo concluídos: 1: Nenhuma; 2: 1 a 3 anos; 3: 4 a 7 anos; 4: 8 a 11 anos; 5: 12 e mais; 9: Ignorado
def_escol_mae	text	Escolaridade da mãe (Nenhuma; de 1 a 3 anos; de 4 a 7 anos; 8 a 11 anos; 12 anos e mais; Ignorado)
CODOCUPMAE	text	Ocupação, conforme a Classificação Brasileira de Ocupações (CBO-2002)
QTDFILVIVO	text	Número de filhos vivos
QTDFILMORT	text	Número de filhos mortos
CODMUNRES	int8	Município de residência da mãe, em codificação idêntica a de CODMUNNASC, conforme tabela TABMUN
GESTACAO	int8	Semanas de gestação, conforme a tabela: 9: Ignorado; 1: Menos de 22 semanas; 2: 22 a 27 semanas; 3: 28 a 31 semanas; 4: 32 a 36 semanas; 5: 37 a 41 semanas; 6: 42 semanas e mais
def_gestacao	text	Semana de gestação (Nominal, com as seguintes classificações: Menos de 22 semanas; 22 a 27 semanas; 28 a 31 semanas; 32 a 36 semanas; 37 a 41 semanas; 42 semanas ou mais)
GRAVIDEZ	int8	Tipo de gravidez, conforme a tabela: 9: Ignorado; 1: Única; 2: Dupla; 3: Tripla e mais
def_gravidez	text	Tipo de gravidez (Nominal, com as seguintes classificações: Única; Dupla; Tripla e mais e Ignorada)
PARTO	int8	Tipo de parto, conforme a tabela: 9: Ignorado; 1: Vaginal; 2: Cesáreo
def_parto	text	Tipo de parto (Nominal, com as seguintes classificações: Vaginal; Cesáreo; Ignorado)
CONSULTAS	int8	Número de consultas de pré-natal: 1: Nenhuma; 2: de 1 a 3; 3: de 4 a 6; 4: 7 e mais; 9: Ignorado
def_consultas	text	Número de consultas durante o pré-natal (Nenhuma; de 1 a 3; de 4 a 6; 7 e mais; Ignorado)
DTNASC	text	Data do nascimento, no formato ddmmaaaa
data_nasc	date	Data de nascimento
ano_nasc	int8	Ano do nascimento
dia_semana_nasc	text	Dia da semana em que ocorreu o nascimento (dom; seg; ter; qua; qui; sex; sáb)
HORANASC	text	Horário de nascimento

SEXO	int8	Sexo, conforme a tabela: 0: Ignorado, não informado; 1: Masculino; 2: Feminino
def_sexo	text	Sexo (Nominal, com as seguintes classificações: Masculino; Feminino; Ignorado)
APGAR1	text	Apgar no primeiro minuto 00 a 10
APGAR5	text	Apgar no quinto minuto 00 a 10
RACACOR	int8	Raça/Cor: 1: Branca; 2: Preta; 3: Amarela; 4: Parda; 5: Indígena
def_raca_cor	text	Raça/cor (Nominal, com as seguintes classificações: Branca; Preta; Amarela; Parda; Indígena)
PESO	text	Peso ao nascer, em gramas
IDANOMAL	int8	Anomalia congênita: 9: Ignorado; 1: Sim; 2: Não
def_anomalia	text	Anomalia congênita (Ignorado; Sim; Não)
DTCADASTRO	text	Data do cadastro da DN no sistema
CODANOMAL	text	Código de malformação congênita ou anomalia cromossômica, de acordo com a CID-10
NUMEROLOTE	int8	Número do lote
VERSAOSIST	text	Versão do sistema
DTRECEBIM	text	Data de recebimento no nível central, data da última atualização do registro.
DIFDATA	text	Diferença entre a data de óbito e data do recebimento original da DO ([DTNASC] – [DTRECORIG])
DTRECORIGA	text	Data do 1o recebimento do lote, dada pelo Sisnet.
NATURALMAE	int8	Se a mãe for estrangeira, constará o código do país de nascimento.
CODMUNNATU	int8	Código do município de naturalidade da mãe
CODUFNATU	int8	Código da UF de naturalidade da mãe
ESCMAE2010	int8	Escolaridade 2010. Valores: 0 – Sem escolaridade; 1 –Fundamental I (1a a 4a série); 2 – Fundamental II (5a a 8a série); 3 – Médio (antigo 2o Grau); 4 – Superior incompleto; 5 –Superior completo; 9 – Ignorado.
SERIESCMAE	int8	Série escolar da mãe. Valores de 1 a 8.
DTNASCMAE	text	Data de nascimento da mãe
RACACORMAE	int8	Raça/cor da mãe
QTDGESTANT	text	Número de gestações anteriores
QTDPARTNOR	text	Número de partos vaginais
QTDPARTCES	text	Número de partos cesáreos
IDADEPAI	int8	Idade do pai
DTULTMENST	text	Data da última menstruação (DUM): dd mm aaaa
SEMAGESTAC	int8	Número de semanas de gestação.
TPMETESTIM	int8	Método utilizado. Valores: 1– Exame físico; 2– Outro método; 9– Ignorado.
CONSPRENAT	text	Número de consultas pré-natal
MESPRENAT	text	Mês de gestação em que iniciou o pré-natal
TPAPRESENT	int8	Tipo de apresentação do RN. Valores: 1– Cefálico; 2– Pélvica ou podálica; 3– Transversa; 9– Ignorado.
STTRABPART	int8	Trabalho de parto induzido? Valores: 1– Sim; 2– Não; 3– Não se aplica; 9– Ignorado.
STCESPARTO	int8	Cesárea ocorreu antes do trabalho de parto iniciar? Valores: 1–Sim; 2– Não; 3– Não se aplica; 9– Ignorado.
TPNASCASSI	int8	Nascimento foi assistido por? Valores: 1– Médico; 2–Enfermeira/obstetriz; 3– Parteira; 4– Outros; 9– Ignorado.
TPFUNCRESP	int8	Tipo de função do responsável pelo preenchimento. Valores:1– Médico; 2– Enfermeiro; 3– Parteira; 4– Funcionário docartório; 5– Outros.
TPDOCRESP	int8	Tipo do documento do responsável. Valores: 1-CNES; 2-CRM; 3-COREN; 4-RG; 5-CPF.
DTDECLARAC	text	Data da declaração: dd mm aaaa
ESCMAEAGR1	text	Escolaridade 2010 agregada. Valores: 00 – Sem Escolaridade; 01 – Fundamental I Incompleto; 02 – Fundamental I Completo; 03 – Fundamental II Incompleto; 04 – Fundamental II Completo; 05 – Ensino Médio Incompleto; 06 – Ensino Médio Completo; 07 – Superior Incompleto; 08 – Superior Completo; 09 – Ignorado; 10 – Fundamental I Incompleto ou Inespecífico; 11 – Fundamental II Incompleto ou Inespecífico; 12 – EnsinoMédio Incompleto ou Inespecífico.
STDNEPIDEM	int8	Status de DO Epidemiológica. Valores: 1 – SIM; 0 – NÃO.
STDNNOVA	int8	Status de DO Nova. Valores: 1 – SIM; 0 – NÃO.
CODPAISRES	int8	Código do país de residência
TPROBSON	text	Código do Grupo de Robson, gerado pelo sistema

PARIDADE	int8	Sem descrição
KOTELCHUCK	int8	Sem descrição
nasc_MUNNOME	text	Nome do município de nascimento
nasc_MUNNOMEX	text	Nome do município de nascimento em maiúsculas e sem acentos
nasc_AMAZONIA	text	Indica (S/N) se o município de nascimento faz parte da Amazônia Legal (conforme IBGE)
nasc_FRONTEIRA	text	Indica (S/N) se o município de nascimento faz parte da faixa de fronteira (conforme IBGE)
nasc_CAPITAL	text	Indica (S/N) se o município de nascimento é capital de UF
nasc_MSAUDCOD	int8	Código da Macrorregional de Saúde a que o Município de nascimento pertence
nasc_RSAUDCOD	int8	Código da Regional de Saúde a que o Município de nascimento pertence
nasc_CSAUDCOD	int8	Código da Microrregional de Saúde a que o Município de nascimento pertence
nasc_LATITUDE	float8	Latitude do município de nascimento
nasc_LONGITUDE	float8	Longitude do município de nascimento
nasc_ALTITUDE	int8	Altitude do município de nascimento
nasc_AREA	float8	Área do município de nascimento
nasc_codigo_adotado	int8	Armazena o código atribuído ao município de nascimento atualmente, tratando os casos em que múltiplos códigos tenham sido utilizados para um mesmo município ao longo do tempo
res_MUNNOME	text	Nome do município de residência
res_MUNNOMEX	text	Nome do município de residência em maiúsculas e sem acentos
res_AMAZONIA	text	Indica (S/N) se o município de residência faz parte da Amazônia Legal (conforme IBGE)
res_FRONTEIRA	text	Indica (S/N) se o município de residência faz parte da faixa de fronteira (conforme IBGE)
res_CAPITAL	text	Indica (S/N) se o município de residência é capital de UF
res_MSAUDCOD	int8	Código da Macrorregional de Saúde a que o Município de residência pertence
res_RSAUDCOD	int8	Código da Regional de Saúde a que o Município de residência pertence
res_CSAUDCOD	int8	Código da Microrregional de Saúde a que o Município de residência pertence
res_LATITUDE	float8	Latitude do município de residência
res_LONGITUDE	float8	Longitude do município de residência
res_ALTITUDE	int8	Altitude do município de residência
res_AREA	float8	Área do município de residência
res_codigo_adotado	int8	Armazena o código atribuído ao município de residência atualmente, tratando os casos em que múltiplos códigos tenham sido utilizados para um mesmo município ao longo do tempo
nasc_SIGLA_UF	text	Sigla da unidade da federação de nascimento
nasc_CODIGO_UF	int8	Código da UF de nascimento
nasc_NOME_UF	text	Nome da unidade da federação de nascimento
res_SIGLA_UF	text	Sigla da unidade da federação de residência
res_CODIGO_UF	int8	Código da UF de residência
res_NOME_UF	text	Nome da unidade da federação de residência
nasc_REGIAO	text	Nome da região da unidade da federação de nascimento
res_REGIAO	text	Nome da região da unidade da federação de residência
codanomal_capitulo	text	Capítulo CID-10 da malformação congênita ou anomalia cromossômica
codanomal_grupo	text	Grupo CID-10 da malformação congênita ou anomalia cromossômica
codanomal_categoria	text	Categoria CID-10 da malformação congênita ou anomalia cromossômica
codanomal_subcategoria	text	Subcategoria CID-10 da malformação congênita ou anomalia cromossômica
nasc_coordenadas	text	Coordenadas do município de nascimento
res_coordenadas	text	Coordenadas do município de residência
parto_prematuro	int8	Indica a prematuridade do nascimento. 0: não há indícios de prematuridade; 1: há indício de prematuridade dado pela idade gestacional (GESTACAO<=4); 2: há indício de prematuridade dado pelo peso ao nascer (PESO<2500); 3: a idade gestacional e o peso ao nascer indicam prematuridade
def_parto_prematuro	text	Indica a prematuridade do nascimento. Termo: não há indícios de prematuridade; Inconclusivo-IG: há indício de prematuridade dado pela idade gestacional (GESTACAO<=4); Inconclusivo-Peso: há indício de prematuridade dado pelo peso ao nascer (PESO<2500); Prematuro: a idade gestacional e o peso ao nascer indicam prematuridade

Tabela 7 – Variáveis do Sistema de Informações sobre Mortalidade presentes no conjunto de dados disponibilizado na Plataforma de Ciência de Dados aplicada à Sáude da Fiocruz.

Variável	**Tipo**	**Descrição**
NUMERODO	int8	Número de declaração de óbito
TIPOBITO	int8	Tipo do óbito:1: óbito fetal2: óbito não fetal
def_tipo_obito	text	Tipo de óbito (Nominal, com as seguintes classificações: Fetal e Não Fetal)
DTOBITO	text	Data do óbito, no formato ddmmaaaa
data_obito	date	Data de ocorrência do óbito
ano_obito	int8	Ano do óbito
dia_semana_obito	text	Dia da semana em que ocorreu o óbito
NATURAL	int8	Naturalidade, conforme a tabela de países. Se for brasileiro,porém, o primeiro dígito contém 8 e os demais o código daUF de naturalidade
DTNASC	text	Data de nascimento do falecido no formato ddmmaaaa
data_nasc	date	Data de nascimento
idade_obito_calculado	int8	Idade do óbito calculado utilizando a data de óbito e a data denascimento
ano_nasc	int8	Ano do nascimento
dia_semana_nasc	text	Dia da semana em que ocorreu o nascimento
IDADE	int4	Idade, composto de dois subcampos. O primeiro, de 1dígito, indica a unidade da idade, conforme a tabela aseguir. O segundo, de dois dígitos, indica a quantidade deunidades:0: Idade ignorada, o segundo subcampo e1: Horas, o segundo subcampo varia de 01 a 232: Dias, o segundo subcampo varia de 01 a 293: Meses, o segundo subcampo varia de 01 a 114: Anos, o segundo subcampo varia de 00 a 995: Anos (mais de 100 anos), o segundo subcampo varia de0 a 99, exemplos:000: Idade ignorada020: 20 minutos103: 3 horas204: 4 dias305: 5 meses400: menor de 1 ano, mas não se sabe o numero de horas,dias ou meses410: 10 anos505: 105 anos
idade_obito_anos	int8	Idade do óbito (em anos) informada na declaração de óbito
idade_obito_meses	int8	Idade do óbito (em meses) informada na declaração de óbito
idade_obito_dias	int8	Idade do óbito (em dias) informada na declaração de óbito
idade_obito_horas	int8	Idade do óbito (em horas) informada na declaração de óbito
idade_obito_mins	int8	Idade do óbito (em minutos) informada na declaração de óbito
SEXO	int8	Sexo, conforme a tabela:0: Ignorado1: Masculino2: Feminino
def_sexo	text	Sexo (Nominal, com as seguintes classificações: Masculino; Feminino; Ignorado)
RACACOR	int8	Raça/Cor:1: Branca2: Preta3: Amarela4: Parda5: Indígena
def_raca_cor	text	Raça/cor (Nominal, com as seguintes classificações: Branca; Preta; Amarela; Parda; Indígena)
ESTCIV	int8	Estado civil, conforme a tabela:1: Solteiro2: Casado3: Viúvo4: Separado judicialmente5: União consensual (versões anteriores)9: Ignorado
def_est_civil	text	Estado civil (Nominal, com as seguintes classificações: Solteiro;Casado; Viúvo; Separado Judicialmente/divorciado; União Estável;Ignorado)
ESC	int8	Escolaridade, Anos de estudo concluídos:1: Nenhuma2: 1 a 3 anos3: 4 a 7 anos4: 8 a 11 anos5: 12 e mais9: Ignorado
def_escol	text	Escolaridade (Nominal, com as seguintes classificações: Nenhuma; de 1 a 3 anos; de 4 a 7 anos; de 8 a 11 anos; 12 e mais; Ignorado)
OCUP	int8	Ocupação, conforme a Classificação Brasileira deOcupações (CBO-2002)
CODBAIRES	int8	Sem descrição
CODMUNRES	int8	Município de residência do falecido, conforme códigos IBGE
LOCOCOR	int8	Local de ocorrência do óbito, conforme a tabela:9: Ignorado1: Hospital2: Outro estab saúde3: Domicílio4: Via Pública5: Outros
def_loc_ocor	text	Local de ocorrência do óbito (Nominal, com as seguintes classificações: Hospital; Outros estabelecimentos de saúde; Domicílio; Via pública; Outros; Ignorado)
CODMUNOCOR	int8	Município de ocorrência do óbito, conforme códigos IBGE
IDADEMAE	int8	Idade da mãe em anos
ESCMAE	int8	Escolaridade da mãe, Anos de estudo concluídos:1: Nenhuma2: 1 a 3 anos3: 4 a 7 anos4: 8 a 11 anos5: 12 e mais9: Ignorado
def_escol_mae	text	Escolaridade da mãe (Nominal, com as seguintes classificações: Nenhuma; de 1 a 3 anos; de 4 a 7 anos; de 8 a 11 anos; 12 e mais; Ignorado)

OCUPMAE	int8	Ocupação da mãe, conforme codificação de OCUPAÇÃO
QTDFILVIVO	int8	Número de filhos vivos
QTDFILMORT	int8	Número de filhos mortos, ignorados, não incluindo opróprio
GRAVIDEZ	int8	Tipo de gravidez, conforme a tabela:9: Ignorado1: Única2: Dupla3: Tripla e mais
def_gravidez	text	Tipo de gravidez (Nominal, com as seguintes classificações: Única; Dupla; Tripla e mais; Ignorada)
GESTACAO	int8	Semanas de gestação, conforme a tabela:9: Ignorado1: Menos de 22 semanas2: 22 a 27 semanas3: 28 a 31 semanas4: 32 a 36 semanas5: 37 a 41 semanas6: 42 semanas e mais
def_gestacao	text	Semanas de gestação (Nominal, com as seguintes classificações: Ignorado; Menos de 22 semanas; 22 a 27 semanas; 28 a 31 semanas; 32 a 36 semanas; 37 a 41 semanas; 42 semanas e mais)
PARTO	int8	Tipo de parto, conforme a tabela:9: Ignorado1: Vaginal2: Cesáreo
def_parto	text	Tipo de parto (Nominal, com as seguintes classificações: Vaginal;Cesáreo; Ignorado)
OBITOPARTO	int8	Morte em relação ao parto, conforme tabela:9: Ignorado1: Antes2: Durante3: Depois
def_obito_parto	text	Indicação de como foi a morte em relação ao parto (Nominal, com as seguintes classificações: Antes; Durante; Depois; Ignorado)
PESO	int8	Peso ao nascer, em gramas
OBITOGRAV	int8	Morte durante a Gravidez conforme tabela:9: Ignorado1: Sim2: Não
def_obito_grav	text	Indicação de ocorrência do óbito durante a gravidez (Nominal, com as seguintes classificações: Sim; Não; Ignorado)
OBITOPUERP	int8	Morte durante o puerpério, conforme tabela:9: Ignorado1: Sim, até 42 dias2: Sim, de 43 dias a 01 ano3: Não
def_obito_puerp	text	Indicação de óbito no puerpério (Nominal, com as seguintesclassificações: Sim, até 42 dias após o parto; Sim, de 43 dias a 01 anos; Não; Ignorado)
ASSISTMED	int8	Indica se houve assistência medica, conforme a tabela:9: Ignorado1: Com assistência2: Sem assistência
def_assist_med	text	Assistência médica (Nominal, com as seguintes classificações: Com assistência; Sem assistência; Ignorado)
EXAME	int8	Indica se houve exame complementar, conforme a tabela:9: Ignorado1: Sim2: Não
def_exame	text	Indicação de realização de exame (Nominal, com as seguintesclassificações: Sim; Não; Ignorado)
CIRURGIA	int8	Indica se houve cirurgia, conforme a tabela:9: Ignorado1: Sim2: Não
def_cirurgia	text	Indica se houve cirurgia (Nominal, com as seguintesclassificações: Sim; Não; Ignorado)
NECROPSIA	int8	Indica se houve necrópsia, conforme a tabela:9: Ignorado1: Sim2: Não
def_necropsia	text	Confirmação do diagnóstico por necrópsia (Nominal, com as seguintes classificações: Sim; Não; Ignorado)
CAUSABAS	text	Causa básica, conforme a Classificação Internacional de Doença (CID), 10a. Revisão
LINHAA	text	Linha A do atestado, conforme a Classificação Internacional de Doença (CID), 10a. Revisão
LINHAB	text	Linha B do atestado, conforme a Classificação Internacional de Doença (CID), 10a. Revisão
LINHAC	text	Linha C do atestado, conforme a Classificação Internacional de Doença (CID), 10a. Revisão
LINHAD	text	Linha D do atestado, conforme a Classificação Internacional de Doença (CID), 10a. Revisão
LINHAII	text	Linha II do atestado, conforme a Classificação Internacional de Doença (CID), 10a. Revisão
CIRCOBITO	int8	Indica o tipo de acidente, se cabível:9: Ignorado1: Acidente2: Suicídio3: Homicídio4: Outros
def_circ_obito	text	Indicação da provável circunstância de morte não natural (Nominal, com as seguintes classificações: Acidente; Suicídio; Homicídio; Outros; Ignorado)
ACIDTRAB	int8	Indica se foi acidente de trabalho, conforme a tabela:9: Ignorado1: Sim2: Não
def_acid_trab	text	Indicação de ocorrência de acidente de trabalho (Nominal, com as seguintes classificações: Sim; Não; Ignorado)

FONTE	int8	Fonte da informação, conforme a tabela:9: Ignorado1: Boletim de Ocorrência2: Hospital3: Família4: Outra
def_fonte	text	Fonte de informação (Nominal, com as seguintes classificações: Boletim de Ocorrência; Hospital; Família; Outra; Ignorado)
CODINST	int8	Sem descrição
NUMERODV	int8	Sem descrição
ORIGEM	int8	Sem descrição
HORAOBITO	text	Horário do óbito
NUMSUS	int8	Sem descrição
CODMUNNATU	int8	Código do município de naturalidade do falecido
ESC2010	int8	Escolaridade 2010. Valores: 0 – Sem escolaridade; 1 – FundamentalI (1a a 4a série); 2 – Fundamental II (5a a 8a série); 3 – Médio(antigo 2o Grau); 4 – Superior incompleto; 5 – Superior completo; 9– Ignorado.
SERIESCFAL	int8	Série escolar do falecido. Valores de 1 a 8.
CODESTAB	text	Código do estabelecimento
ESTABDESCR	text	Sem descrição
ESCMAE2010	int8	Escolaridade 2010. Valores: 0 – Sem escolaridade; 1 – FundamentalI (1a a 4a série); 2 – Fundamental II (5a a 8a série); 3 – Médio(antigo 2o Grau); 4 – Superior incompleto; 5 – Superior completo; 9– Ignorado.
SERIESCMAE	int8	Série escolar da mãe. Valores de 1 a 8.
SEMAGESTAC	int8	Semanas de gestação
NUMERODN	int8	Número de declaração de nascido vivo
TPMORTEOCO	int8	Informar quando a morte ocorreu: 1 – na gravidez; 2 – no parto; 3– no aborto; 4 – até 42 dias após o parto; 5 – de 43 dias a 1 anoapós o parto; 8 – não ocorreu nestes períodos; 9 – ignorado.
CB_PRE	text	Causa selecionada sem re-seleção (novo SCB)
CRM	int8	No do CRM
COMUNSVOIM	int8	Código do município do SVO ou do IML
DTATESTADO	text	Data do Atestado
NUMEROLOTE	int8	Número do lote
TPPOS	text	Óbito investigado 1-Sim, 2- Não
DTINVESTIG	text	Data de investigação
CAUSABAS_O	text	Causa básica original, a primeira informação que entra nosistema
DTCADASTRO	text	Data de cadastro do registro no sistema
ATESTANTE	int8	Indica se o médico que assina atendeu o paciente:1: Sim2: Substituto3: IML4: SVO5: Outros
STCODIFICA	text	Status de instalação: se codificadora (valor: S) ou não (valor: N)
CODIFICADO	text	Se estiver codificado (valor: S) ou não (valor: N)
VERSAOSIST	text	Versão do sistema
VERSAOSCB	float8	Versão do seletor de causa básica
FONTEINV	int8	Fonte de investigação:1: Comitê de Morte Materna e/ou Infantil2: Visita domiciliar / Entrevista família3: Estab Saúde / Prontuário4: Relacion com outros bancos de dados5: S V O6: I M L7: Outra fonte8: Múltiplas fontes9: Ignorado
DTRECEBIM	text	Data de recebimento no nível central, data da última atualização do registro
ATESTADO	text	CIDs informado no atestado
DTRECORIGA	text	Data do recebimento original: dd mm aaaa
CAUSAMAT	text	Causa externa associada a uma causa materna
ESCMAEAGR1	text	Escolaridade 2010 agregada. Valores: 00 – Sem Escolaridade; 01 –Fundamental I Incompleto; 02 – Fundamental I Completo; 03 –Fundamental II Incompleto; 04 – Fundamental II Completo; 05 –Ensino Médio Incompleto; 06 – Ensino Médio Completo; 07 –Superior Incompleto; 08 – Superior Completo; 09 – Ignorado; 10 –Fundamental I Incompleto ou Inespecífico; 11 – Fundamental IIIncompleto ou Inespecífico; 12 – Ensino Médio Incompleto ouInespecífico.
ESCFALAGR1	text	Escolaridade 2010 agregada. Valores: 00 – Sem Escolaridade; 01 –Fundamental I Incompleto; 02 – Fundamental I Completo; 03 –Fundamental II Incompleto; 04 – Fundamental II Completo; 05 –Ensino Médio Incompleto; 06 – Ensino Médio Completo; 07 –Superior Incompleto; 08 – Superior Completo; 09 – Ignorado; 10 –Fundamental I Incompleto ou Inespecífico; 11 – Fundamental IIIncompleto ou Inespecífico; 12 – Ensino Médio Incompleto ouInespecífico.
STDOEPIDEM	int8	Status de DO Epidemiológica. Valores: 1 – SIM; 0 – NÃO.

STDONOVA	int8	Status de DO Nova. Valores: 1 – SIM; 0 – NÃO.
DIFDATA	text	Diferença entre a data de óbito e data do recebimento original daDO ([DTOBITO] – [DTRECORIG])
NUDIASOBCO	text	Sem descrição
NUDIASOBIN	text	Sem descrição
DTCADINV	text	Sem descrição
TPOBITOCOR	int8	Sem descrição
DTCONINV	text	Sem descrição
FONTES	text	Sem descrição
TPRESGINFO	int8	Sem descrição
TPNIVELINV	text	Sem descrição
NUDIASINF	text	Sem descrição
DTCADINF	text	Sem descrição
MORTEPARTO	int8	Sem descrição
DTCONCASO	text	Sem descrição
FONTESINF	text	Sem descrição
ALTCAUSA	int8	Sem descrição
res_MUNNOME	text	Nome (acentuado, maiúsculas e minúsculas) do Município (padrão DOS, página de código 850) de residência da pessoa que foi à óbito
res_MUNNOMEX	text	Nome (sem acentos, em maiúsculas) do Município de residência da pessoa que foi à óbito
res_AMAZONIA	text	Indica (S ou N) se o município de residência da pessoa que foi à óbito faz parte da Amazônia Legal (conforme IBGE)
res_FRONTEIRA	text	Indica (S ou N) se o município de residência da pessoa que foi à óbito faz parte da faixa de fronteira (conforme IBGE)
res_CAPITAL	text	Indica (S ou N) se o município de residência da pessoa que foi à óbito é capital da UF
res_MSAUDCOD	int8	Código da Macrorregional de Saúde a que o Município de residência da pessoa que foi à óbito pertence
res_RSAUDCOD	int8	Código da Regional de Saúde a que o Município de residência da pessoa que foi à óbito pertence
res_CSAUDCOD	int8	Código da Microrregional de Saúde a que o Município de residência da pessoa que foi à óbito pertence
res_LATITUDE	float8	Latitude da sede do Município de residência da pessoa que foi à óbito
res_LONGITUDE	float8	Longitude da sede do Município de residência da pessoa que foi à óbito
res_ALTITUDE	int8	Altitude, em metros, da sede do Município de residência da pessoa que foi à óbito
res_AREA	float8	Área, em quilômetros quadrados, do Município de residência da pessoa que foi à óbito, segundo a Resolução 05, de 10/12/2002, do IBGE
res_codigo_adotado	int8	Armazena o código atribuído ao município de residência da pessoa que foi à óbito, tratando os casos em que múltiplos códigos tenham sido utilizados para um mesmo município ao longo do tempo
ocor_MUNNOME	text	Nome (acentuado, maiúsculas e minúsculas) do Município (padrão DOS, página de código 850) de ocorrência do óbito
ocor_MUNNOMEX	text	Nome (sem acentos, em maiúsculas) do Município de ocorrência do óbito
ocor_AMAZONIA	text	Indica (S ou N) se o município de ocorrência do óbito faz parte da Amazônia Legal (conforme IBGE)
ocor_FRONTEIRA	text	Indica (S ou N) se o município de ocorrência do óbito faz parte da faixa de fronteira (conforme IBGE)
ocor_CAPITAL	text	Indica (S ou N) se o município de ocorrência do óbito é capital da UF
ocor_MSAUDCOD	int8	Código da Macrorregional de Saúde a que o Município de ocorrência do óbito pertence
ocor_RSAUDCOD	int8	Código da Regional de Saúde a que o Município de ocorrência do óbito pertence
ocor_CSAUDCOD	int8	Código da Microrregional de Saúde a que o Município de ocorrência do óbito pertence
ocor_LATITUDE	float8	Latitude da sede do Município de ocorrência do óbito
ocor_LONGITUDE	float8	Longitude da sede do Município de ocorrência do óbito
ocor_ALTITUDE	int8	Altitude, em metros, da sede do Município de ocorrência do óbito
ocor_AREA	float8	Área, em quilômetros quadrados, do Município de ocorrência do óbito, segundo a Resolução 05, de 10/12/2002, do IBGE

ocor_codigo_adotado	int8	Armazena o código atribuído ao município de ocorrência do óbito, tratando os casos em que múltiplos códigos tenham sido utilizados para um mesmo município ao longo do tempo
res_SIGLA_UF	text	Sigla da unidade da federação de residência da pessoa que foi à óbito
res_CODIGO_UF	text	Código IBGE da Unidade da Federação de residência da pessoa que foi à óbito
res_NOME_UF	text	Nome da unidade da federação de residência da pessoa que foi à óbito
ocor_SIGLA_UF	text	Sigla da unidade da federação de ocorrência do óbito
ocor_CODIGO_UF	text	Código IBGE da Unidade da Federação de ocorrência do óbito
ocor_NOME_UF	text	Nome da unidade da federação de ocorrência do óbito
res_REGIAO	text	Nome da região da unidade da federação de residência do falecido
ocor_REGIAO	text	Nome da região da unidade da federação de ocorrência do óbito
causabas_capitulo	text	Capítulo CID-10 da causa base do óbito
causabas_grupo	text	Grupo CID-10 da causa base do óbito
causabas_categoria	text	Categoria CID-10 da causa base do óbito
causabas_subcategoria	text	Subcategoria CID-10 da causa base do óbito
res_coordenadas	text	Coordenadas do município de residência da pessoa que foi à óbito
ocor_coordenadas	text	Coordenadas do município de ocorrência do óbito

Tabela 8 – Variáveis do Cadastro Nacional de Estabelecimentos de Saúde presentes no conjunto de dados disponibilizado na Plataforma de Ciência de Dados aplicada à Saúde da Fiocruz.

Variável	**Tipo**	**Descrição**
CNES	text	Número nacional do estabelecimento de saúde
CODUFMUN	int8	Código do município do estabelecimento: UF + MUNIC (sem dígito)
COD_CEP	text	Código do CEP do estabelecimento
CPF_CNPJ	text	CPF do estabelecimento, caso pessoa física ou CNPJ, caso pessoa jurídica
PF_PJ	int8	Indicador de pessoa: 1-Física 3-Jurídica
def_pf_pj	text	Definição do indicador de pessoa
NIV_DEP	int8	Grau de dependência, onde: 1-Individual 3-Mantida
def_niv_dep	text	Definição do grau de dependência
CNPJ_MAN	text	CNPJ da mantenedora do estabelecimento.
COD_IR	text	Indica o tipo de retenção de tributos da mantenedora: 10-Estabelecimento PUBLICO; 11-Estabelecimento FILANTROPICO; 12-Estabelecimento SEM FINS LUCRATIVOS; 13-Estabelecimento PRIVADO LUCRATIVA SIMPLES; 14-Estabelecimento PRIVADO LUCRATIVA; 15-Estabelecimento SINDICAL; 16-Estabelecimento PESSOA FISICA; 19-Estabelecimento Ret.Manten.código 19; IR-Estabelecimento Ret.Manten.código IR
def_cod_ir	text	Definição da indicação o tipo de retenção de tributos da mantenedora
REGSAUDE	text	Código da região de saúde NOAS
MICR_REG	text	Código da microrregião de saúde NOAS
DISTRSAN	text	Código do distrito sanitário
DISTRADM	text	Inicialmente chamado Distrito Administrativo, atualmente Código do Módulo Assistencial conforme tabela local, em conformidade com o Plano Diretor de Regionalização do Estado/Município.
VINC_SUS	int8	Vínculo com SUS: 1-Sim 0-Não
TPGESTAO	text	Gestão de saúde: Z-Não informado; D-Dupla; E-Estadual; M-Municipal; S-Sem gestão
def_tpgestao	text	Definição da gestão de saúde
ESFERA_A	text	Código da esfera administrativa: 01-Federal; 02-Estadual; 03-Municipal; 04-Privada; 99-Esfera não informada
def_esfera_a	text	Definição do código da esfera administrativa
RETENCAO	text	Código de retenção de tributos: 00,99-Retenção estab. não informada; 10-Estabelecimento PUBLICO; 11-Estabelecimento FILANTROPICO; 12-Estabelecimento SEM FINS LUCRATIVOS; 13-Estabelecimento PRIVADO LUCRATIVA SIMPLES; 14-Estabelecimento PRIVADO LUCRATIVA; 15-Estabelecimento SINDICAL; 16-Estabelecimento PESSOA FISICA

def_retencao	text	Definição do código de retenção de tributos
ATIVIDAD	text	Código da atividade de ensino/pesquisa: 01-Unidade Universitária; 02-Unidade Escola Superior Isolada; 03-Unidade Auxiliar de Ensino; 04-Unidade SEM atividade de Ensino; 05-Hospital de Ensino; 99-Atividade Ensino não informada
def_atividad	text	Definição do código da atividade de ensino/pesquisa
NATUREZA	text	Código da natureza da organização: 01-Administração Direta da Saúde (MS, SES, e SMS); 02-Adm Direta outros orgãos (MEX, MEx, Marinha,...); 03-Adm Indireta - Autarquias; 04-Adm Indireta - Fundação Pública; 05-Adm Indireta - Empresa Pública; 06-Adm Indireta - Organização Social Pública; 07-Empresa Privada; 08-Fundação Privada; 09-Cooperativa; 10-Serviço Social Autônomo; 11-Entidade Beneficente SEM fins lucrativos; 12-Economia MIsta; 13-Sindicato; 00-Natureza inexistente; 0-Natureza inexistente; 99-Natureza não informada
def_natureza	text	Definição do código da natureza da organização
CLIENTEL	text	Código de FLUXO da clientela: 01-Atendimento de demanda espontânea; 02-Atendimento de demanda referenciada; 03-Atendimento de demanda espontânea e referenciada; 00-Fluxo de Clientela não exigido; 99-Fluxo de Clientela não informado
def_clientel	text	Definição do código de FLUXO da clientela
TP_UNID	text	Tipo de unidade (estabelecimento)
def_tp_unid	text	Definição do tipo de unidade (estabelecimento)
TURNO_AT	text	Código de turno de atendimento: 01-ATENDIMENTO TURNOS INTERMITENTES; 02-ATENDIMENTO CONTÍNUO 24 HORAS/DIA (Pl Sab Dom Fer); 03-ATENDIMENTO TURNOS MANHÃ/TARDE/NOITE; 04-ATENDIMENTO SOMENTE PELA MANHÃ; 05-ATENDIMENTO SOMENTE À TARDE; 06-ATENDIMENTO TURNOS MANHÃ/TARDE; 07-ATENDIMENTO SOMENTE À NOITE; 99-Turno não informado
def_turno_at	text	Definição do código de turno de atendimento
NIV_HIER	text	Código do nível de hierarquia: 01-NH 1-PAB-PABA; 02-NH 2-Média M1; 03-NH 3-Média M2 e M3; 04-NH 4-AltaComplex.Ambul.; 05-NH 5-Baixa M1 e M2; 06-NH 6-Média M2 e M3; 07-NH 7-Média M3; 08-NH 8-AltaCompl.Hos/Amb.; 00,99-NH não informado
def_niv_hier	text	Definição do código do nível de hierarquia
TP_PREST	int8	Tipo de Prestador: 30-PUBLICO FEDERAL; 40-PUBLICO ESTADUAL; 50-PUBLICO MUNICIPAL; 61-FILANTROPICO COM CNAS VALIDO; 80-SINDICATO; 20-PRIVADO COM FINS LUCRATIVOS; 22-PRIVADO OPTANTE PELO SIMPLES; 60-PRIVADO SEM FINS LUCRATIVOS; 99-TIPO DE PRESTADOR NÃO INFORMADO
def_tp_prest	text	Definição de Tipo de Prestador
CO_BANCO	text	Código do Banco do Estabelecimento
CO_AGENC	text	Código da Agência do Estabelecimento
C_CORREN	text	Código da Conta Corrente do Estabelecimento
CONTRATM	text	Número do contrato/convênio municipal do vínculo com o SUS
DT_PUBLM	text	Ano e Mês de publicação do contrato /convênio municipal (AAAAMM)
CONTRATE	text	Número do contrato/convênio estadual do vínculo com o SUS
DT_PUBLE	text	Ano e Mês de publicação do contrato /convênio estadual (AAAAMM)
ALVARA	text	Número do alvará
DT_EXPED	text	Ano e Mês de da expedição do alvará (AAAAMM)
ORGEXPED	int8	Órgão expedidor do alvará, onde: 1-SES 2-SMS
def_orgexped	text	Definição do órgão expedidor do alvará
AV_ACRED	text	Indica se o estabelecimento foi avaliado segundo o Manual de Acreditação Hospitalar do MS, onde: 1-Sim 2-Não
def_av_acred	text	Definição da avaliação do estabelecimento segundo o Manual de Acreditação Hospitalar do MS
CLASAVAL	text	Classificação da avaliação do estabelecimento segundo o Manual de Acreditação Hospitalar do MS: 1-ACREDITADO NO NÍV.1; 2-ACREDITADO NO NÍV.2; 3-ACREDITADO NO NÍV.3; 0-NÃO ATENDEU AOS PADRÕES MÍNIMOS; 9-Nível Avaliação não informado
def_clasaval	text	Definição da classificação da avaliação do estabelecimento segundo o Manual de Acreditação Hospitalar do MS
DT_ACRED	text	Ano e Mês da acreditação (AAAAMM)

AV_PNASS	text	Indica se o estabelecimento foi avaliado segundo o Programa Nacional de Serviços de Saúde - PNASS, onde: 1-Sim 2-Não
def_av_pnass	text	Definição da avaliação do estabelecimento segundo o Programa Nacional de Serviços de Saúde - PNASS
DT_PNASS	text	Ano e Mês da Avaliação PNASS (AAAAMM)
GESPRG1E	int8	Indicador se a gestão da atividade Atenção Básica do Nível de Atenção Ambulatorial é estadual, onde: 1-sim 0-não
GESPRG1M	int8	Indicador se a gestão da atividade Atenção Básica do Nível de Atenção Ambulatorial é municipal, onde: 1-sim 0-não
GESPRG2E	int8	Indicador se a gestão da atividade Média Complexidade do Nível de Atenção Ambulatorial é estadual, onde: 1-sim 0-não
GESPRG2M	int8	Indicador se a gestão da atividade Média Complexidade do Nível de Atenção Ambulatorial é municipal, onde: 1-sim 0-não
GESPRG4E	int8	Indicador se a gestão da atividade Alta Complexidade do Nível de Atenção Ambulatorial é estadual, onde: 1-sim 0-não
GESPRG4M	int8	Indicador se a gestão da atividade Alta Complexidade do Nível de Atenção Ambulatorial é municipal, onde: 1-sim 0-não
NIVATE_A	int8	Indica a existência de Nível de Atenção Ambulatorial, de gestão Municipal/Estadual, para este CNES, onde: 1-sim 0-não
GESPRG3E	int8	Indicador se a gestão do programa cód. 03 é estadual, onde: 1-sim 0-não
GESPRG3M	int8	Indicador se a gestão do programa cód. 03 é municipal, onde: 1-sim 0-não
GESPRG5E	int8	Indicador se a gestão da atividade Média Complexidade do Nível de Atenção Hospitalar é estadual, onde: 1-sim 0-não
GESPRG5M	int8	Indicador se a gestão da atividade Média Complexidade do Nível de Atenção Hospitalar é municipal, onde: 1-sim 0-não
GESPRG6E	int8	Indicador se a gestão da atividade Alta Complexidade do Nível de Atenção Hospitalar é estadual, onde: 1-sim 0-não
GESPRG6M	int8	Indicador se a gestão da atividade Alta Complexidade do Nível de Atenção Hospitalar é municipal, onde: 1-sim 0-não
NIVATE_H	int8	Indica a existência de Nível de Atenção Hospitalar, de gestão Municipal/Estadual, para este CNES, onde: 1-sim 0-não
QTLEITP1	int8	Quantidade de leitos tipo 1 (cirúrgico) existentes
QTLEITP2	int8	Quantidade de leitos tipo 2 (clínico) existentes
QTLEITP3	int8	Quantidade de leitos tipo 3 (complem.) existentes
LEITHOSP	int8	Indica a existência de Leitos Hospitalares - Cirurgicos, Clínicos e Complementares, para este CNES, onde: 1-sim 0-não
QTINST01	int8	Quantidade de salas/consultórios de atendimento pediátrico (URGÊNCIA/EMERGÊNCIA)
QTINST02	int8	Quantidade de salas/consultórios de atendimento feminino (URGÊNCIA/EMERGÊNCIA)
QTINST03	int8	Quantidade de salas/consultórios de atendimento masculino (URGÊNCIA/EMERGÊNCIA)
QTINST04	int8	Quantidade de salas/consultórios de atendimento indiferenciado (URGÊNCIA/EMERGÊNCIA)
QTINST05	int8	Quantidade de salas de repouso/observação pediátrico (URGÊNCIA/EMERGÊNCIA)
QTINST06	int8	Quantidade de salas de repouso/observação feminino (URGÊNCIA/EMERGÊNCIA)
QTINST07	int8	Quantidade de salas de repouso/observação masculino (URGÊNCIA/EMERGÊNCIA)
QTINST08	int8	Quantidade de salas de repouso/observação indiferenciado (URGÊNCIA/EMERGÊNCIA)
QTINST09	int8	Quantidade de consultórios de odontologia (URGÊNCIA/EMERGÊNCIA)
QTINST10	int8	Quantidade de salas de Higienização (URGÊNCIA/EMERGÊNCIA)
QTINST11	int8	Quantidade de salas de gesso (URGÊNCIA/EMERGÊNCIA)
QTINST12	int8	Quantidade de salas de curativos (URGÊNCIA/EMERGÊNCIA)
QTINST13	int8	Quantidade de salas de pequenas cirurgias (URGÊNCIA/EMERGÊNCIA)
QTINST14	int8	Quantidade de consultórios médicos (URGÊNCIA/EMERGÊNCIA)
URGEMERG	int8	Indica a existência de instalação física de URGÊNCIA/EMERGÊNCIA para este CNES, onde: 1-sim 0-não

QTINST15	int8	Quantidade de consultórios de clínica básica (ATEND. AMBULATORIAL)
QTINST16	int8	Quantidade de consultórios de clínica especializada (ATEND. AMBULATORIAL)
QTINST17	int8	Quantidade de consultórios de clínica indiferenciada (ATEND. AMBULATORIAL)
QTINST18	int8	Quantidade de consultórios não médicos (ATEND. AMBULATORIAL)
QTINST19	int8	Quantidade de salas de repouso/observação feminino (ATEND. AMBULATORIAL)
QTINST20	int8	Quantidade de salas de repouso/observação masculino (ATEND. AMBULATORIAL)
QTINST21	int8	Quantidade de salas de repouso/observação pediátrico (ATEND. AMBULATORIAL)
QTINST22	int8	Quantidade de salas de repouso/observação indiferenciado (ATEND. AMBULATORIAL)
QTINST23	int8	Quantidade de consultórios de odontologia (ATEND. AMBULATORIAL)
QTINST24	int8	Quantidade de salas de pequenas cirurgias (ATEND. AMBULATORIAL)
QTINST25	int8	Quantidade de salas de enfermagem (ATEND. AMBULATORIAL)
QTINST26	int8	Quantidade de salas de imunização (ATEND. AMBULATORIAL)
QTINST27	int8	Quantidade de salas de nebulização (ATEND. AMBULATORIAL)
QTINST28	int8	Quantidade de salas de gesso (ATEND. AMBULATORIAL)
QTINST29	int8	Quantidade de salas de curativos (ATEND. AMBULATORIAL)
QTINST30	int8	Quantidade de salas de cirurgia ambulatorial (ATEND. AMBULATORIAL)
ATENDAMB	int8	Indica a existência de instalação física de ATENDIMENTO AMBULATORIAL para este CNES, onde: 1-sim 0-não
QTINST31	int8	Quantidade de salas de cirurgias (CENTRO CIRÚRGICO)
QTINST32	int8	Quantidade de salas de recuperação (CENTRO CIRÚRGICO)
QTINST33	int8	Quantidade de salas de cirurgia ambulatorial (CENTRO CIRÚRGICO)
CENTRCIR	int8	Indica a existência de instalação física de ATENDIMENTO HOSPITALAR - CENTRO CIRÚRGICO para este CNES, onde: 1-sim 0-não
QTINST34	int8	Quantidade de salas de pré parto (CENTRO OBSTÉTRICO)
QTINST35	int8	Quantidade de salas de parto normal (CENTRO OBSTÉTRICO)
QTINST36	int8	Quantidade de salas de curetagem (CENTRO OBSTÉTRICO)
QTINST37	int8	Quantidade de salas de cirurgias (CENTRO OBSTÉTRICO)
CENTROBS	int8	Indica a existência de instalação física de ATENDIMENTO HOSPITALAR - CENTRO OBSTÉTRICO para este CNES, onde: 1-sim 0-não
QTLEIT05	int8	Quantidade de leitos de repouso/observação pediátrico (URGÊNCIA/EMERGÊNCIA)
QTLEIT06	int8	Quantidade de leitos de repouso/observação feminino (URGÊNCIA/EMERGÊNCIA)
QTLEIT07	int8	Quantidade de leitos de repouso/observação masculino (URGÊNCIA/EMERGÊNCIA)
QTLEIT08	int8	Quantidade de leitos de repouso/observação indiferenciado (URGÊNCIA/EMERGÊNCIA)
QTLEIT09	int8	Quantidade de equipos de odontologia (URGÊNCIA/EMERGÊNCIA)
QTLEIT19	int8	Quantidade de leitos de repouso/observação feminino (ATEND. AMBULATORIAL)
QTLEIT20	int8	Quantidade de leitos de repouso/observação masculino (ATEND. AMBULATORIAL)
QTLEIT21	int8	Quantidade de leitos de repouso/observação pediátrico (ATEND. AMBULATORIAL)
QTLEIT22	int8	Quantidade de leitos de repouso/observação indiferenciado (ATEND. AMBULATORIAL)
QTLEIT23	int8	Quantidade de equipos de odontologia (ATEND. AMBULATORIAL)
QTLEIT32	int8	Quantidade de leitos de recuperação (CENTRO CIRÚRGICO)
QTLEIT34	int8	Quantidade de leitos de pré parto (CENTRO OBSTÉTRICO)
QTLEIT38	int8	Quantidade de leitos de recém-nascido normal (UNID NEONATAL)
QTLEIT39	int8	Quantidade de leitos de recém-nascido patológico (UNID NEONATAL)
QTLEIT40	int8	Quantidade de leitos de alojamento conjunto (UNID NEONATAL)
CENTRNEO	int8	Indica a existência de instalação física de ATENDIMENTO HOSPITALAR - UNIDADE NEONATAL para este CNES, onde: 1-sim 0-não

ATENDHOS	int8	Indica a existência de instalação física de ATENDIMENTO HOSPITALAR para este CNES, onde: 1-sim 0-não
SERAP01P	int8	Serviço de apoio S.A.M.E. ou S.P.P. próprio, onde: 1-sim 0-não
SERAP01T	int8	Serviço de apoio S.A.M.E. ou S.P.P. terceirizado, onde: 1-sim 0-não
SERAP02P	int8	Serviço de apoio Serviço Social próprio, onde: 1-sim 0-não
SERAP02T	int8	Serviço de apoio Serviço Social terceirizado, onde: 1-sim 0-não
SERAP03P	int8	Serviço de apoio Farmácia próprio, onde: 1-sim 0-não
SERAP03T	int8	Serviço de apoio Farmácia terceirizado, onde: 1-sim 0-não
SERAP04P	int8	Serviço de apoio Esterilização de Materiais próprio, onde: 1-sim 0-não
SERAP04T	int8	Serviço de apoio Esterilização de Materiais terceirizado, onde: 1-sim 0-não
SERAP05P	int8	Serviço de apoio Nutrição/Dietética (S.N.D.) próprio, onde: 1-sim 0-não
SERAP05T	int8	Serviço de apoio Nutrição/Dietética (S.N.D.) terceirizado, onde: 1-sim 0-não
SERAP06P	int8	Serviço de apoio Lactário próprio, onde: 1-sim 0-não
SERAP06T	int8	Serviço de apoio Lactário terceirizado, onde: 1-sim 0-não
SERAP07P	int8	Serviço de apoio Banco de Leite próprio, onde: 1-sim 0-não
SERAP07T	int8	Serviço de apoio Banco de Leite terceirizado, onde: 1-sim 0-não
SERAP08P	int8	Serviço de apoio Lavanderia próprio, onde: 1-sim 0-não
SERAP08T	int8	Serviço de apoio Lavanderia terceirizado, onde: 1-sim 0-não
SERAP09P	int8	Serviço de apoio Manutenção de Equipamento próprio, onde: 1-sim 0-não
SERAP09T	int8	Serviço de apoio Manutenção de Equipamento terceirizado, onde: 1-sim 0-não
SERAP10P	int8	Serviço de apoio Ambulância próprio, onde: 1-sim 0-não
SERAP10T	int8	Serviço de apoio Ambulância terceirizado, onde: 1-sim 0-não
SERAP11P	int8	Serviço de apoio Necrotério próprio, onde: 1-sim 0-não
SERAP11T	int8	Serviço de apoio Necrotério terceirizado, onde: 1-sim 0-não
SERAPOIO	int8	Indica a existência de algum serviço de apoio para este CNES, onde: 1-sim 0-não
RES_BIOL	int8	Existe coleta de resíduo biológico, onde: 1-sim 0-não
RES_QUIM	int8	Existe coleta de resíduo químico, onde: 1-sim 0-não
RES_RADI	int8	Existe coleta de rejeitos radioativos, onde: 1-sim 0-não
RES_COMU	int8	Existe coleta de rejeitos comum, onde: 1-sim 0-não
COLETRES	int8	Indica a existência de alguma coleta de resíduo para este CNES, onde: 1-sim 0-não
COMISS01	int8	Existe comissão de ética médica, onde: 1-sim 0-não
COMISS02	int8	Existe comissão de ética de enfermagem, onde: 1-sim 0-não
COMISS03	int8	Existe comissão de farmácia e terapêutica, onde: 1-sim 0-não
COMISS04	int8	Existe comissão de controle de infecção hospitalar, onde: 1-sim 0-não
COMISS05	int8	Existe comissão de apropriação de custos, onde: 1-sim 0-não
COMISS06	int8	Existe comissão de CIPA, onde: 1-sim 0-não
COMISS07	int8	Existe comissão de revisão de prontuários, onde: 1-sim 0-não
COMISS08	int8	Existe comissão de revisão de documentação médica e estatística, onde:1-sim 0-não
COMISS09	int8	Existe comissão de análise de óbitos e biópsias, onde: 1-sim 0-não
COMISS10	int8	Existe comissão de investigação epidemiológica, onde: 1-sim 0-não
COMISS11	int8	Existe comissão de notificação de doenças, onde: 1-sim 0-não
COMISS12	int8	Existe comissão de controle de zoonoses e vetores, onde: 1-sim 0-não
COMISSAO	int8	Indica a existência de alguma comissão para este CNES, onde: 1-sim 0-não
AP01CV01	int8	Atendimento prestado Internação/Convênio SUS, onde: 1-sim 0-não
AP01CV02	int8	Atendimento prestado Internação/Convênio Particular, onde: 1-sim 0-não
AP01CV05	int8	Atendimento prestado Internação/ Convênio Plano de Saúde Público, onde: 1-sim 0-não
AP01CV06	int8	Atendimento prestado Internação/Convênio Plano de Saúde Privado, onde: 1-sim 0-não
AP01CV03	int8	Atendimento prestado Internação/Convênio Plano-Seguro Próprio, onde: 1-sim 0-não
AP01CV04	int8	Atendimento prestado Internação/Convênio Plano-Seguro Terceiro, onde: 1-sim 0-não
AP02CV01	int8	Atendimento prestado Atendimento Ambulatorial/Convênio SUS, onde: 1-sim 0-não
AP02CV02	int8	Atendimento prestado Atendimento Ambulatorial/Convênio Particular, onde: 1-sim 0-não

AP02CV05	int8	Atendimento prestado Atendimento Ambulatorial/Convênio Plano de Saúde Público, onde: 1-sim 0-não
AP02CV06	int8	Atendimento prestado Atendimento Ambulatorial/Convênio Plano de Saúde Privado, onde: 1-sim 0-não
AP02CV03	int8	Atendimento prestado Atendimento Ambulatorial/Convênio Plano-Seguro Próprio, onde: 1-sim 0-não
AP02CV04	int8	Atendimento prestado Atendimento Ambulatorial/Convênio Plano-Seguro Terceiro, onde: 1-sim 0-não
AP03CV01	int8	Atendimento prestado SADT/ Convênio SUS, onde: 1-sim 0-não
AP03CV02	int8	Atendimento prestado SADT/ Convênio Particular, onde: 1-sim 0-não
AP03CV05	int8	Atendimento prestado SADT/ Convênio Plano de Saúde Público, onde: 1-sim 0-não
AP03CV06	int8	Atendimento prestado SADT/ Convênio Plano de Saúde Privado, onde: 1-sim 0-não
AP03CV03	int8	Atendimento prestado SADT/ Convênio Plano-Seguro Próprio, onde: 1-sim 0-não
AP03CV04	int8	Atendimento prestado SADT/ Convênio Plano-Seguro Terceiro, onde: 1-sim 0-não
AP04CV01	int8	Atendimento prestado Urgência/Convênio SUS, onde: 1-sim 0-não
AP04CV02	int8	Atendimento prestado Urgência/Convênio Particular, onde: 1-sim 0-não
AP04CV05	int8	Atendimento prestado Urgência/Convênio Plano de Saúde Público, onde: 1-sim 0-não
AP04CV06	int8	Atendimento prestado Urgência/Convênio Plano de Saúde Privado, onde: 1-sim 0-não
AP04CV03	int8	Atendimento prestado Urgência/Convênio Plano-Seguro Próprio, onde: 1-sim 0-não
AP04CV04	int8	Atendimento prestado Urgência/Convênio Plano-Seguro Terceiro, onde: 1-sim 0-não
AP05CV01	int8	Atendimento prestado Outros/Convênio SUS,onde: 1-sim 0-não
AP05CV02	int8	Atendimento prestado Outros/Convênio Particular, onde: 1-sim 0-não
AP05CV05	int8	Atendimento prestado Outros/Convênio Plano de Saúde Público, onde: 1-sim 0-não
AP05CV06	int8	Atendimento prestado Outros/Convênio Plano de Saúde Privado, onde: 1-sim 0-não
AP05CV03	int8	Atendimento prestado Outros/Convênio Plano-Seguro Próprio, onde: 1-sim 0-não
AP05CV04	int8	Atendimento prestado Outros/Convênio Plano-Seguro Terceiro, onde: 1-sim 0-não
AP06CV01	int8	Atendimento prestado Vigilância em Saúde/Convênio SUS, onde: 1-sim 0-não
AP06CV02	int8	Atendimento prestado Vigilância em Saúde/Convênio Particular, onde: 1-sim 0-não
AP06CV05	int8	Atendimento prestado Vigilância em Saúde/Convênio Plano de Saúde Público, onde: 1-sim 0-não
AP06CV06	int8	Atendimento prestado Vigilância em Saúde/Convênio Plano de Saúde Privado, onde: 1-sim 0-não
AP06CV03	int8	Atendimento prestado Vigilância em Saúde/Convênio Plano-Seguro Próprio, onde: 1-sim 0-não
AP06CV04	int8	Atendimento prestado Vigilância em Saúde/Convênio Plano-Seguro Terceiro, onde: 1-sim 0-não
AP07CV01	int8	Atendimento prestado Regulação/Convênio SUS, onde: 1-sim 0-não
AP07CV02	int8	Atendimento prestado Regulação/Convênio Particular, onde: 1-sim 0-não
AP07CV05	int8	Atendimento prestado Regulação/Convênio Plano de Saúde Público, onde: 1-sim 0-não
AP07CV06	int8	Atendimento prestado Regulação/Convênio Plano de Saúde Privado, onde: 1-sim 0-não
AP07CV03	int8	Atendimento prestado Regulação/Convênio Plano-Seguro Próprio, onde: 1-sim 0-não
AP07CV04	int8	Atendimento prestado RRegulação/Convênio Plano-Seguro Terceiro, onde: 1-sim 0-não

ATEND_PR	int8	Indica a existência de algum atendimento prestado para este CNES, onde: 1-sim / 0-não
DT_ATUAL	int8	Ano e Mês de competência da atualização da informação pelo estabelecimento (AAAAMM)
COMPETEN	text	Ano e Mês de competência da informação (AAAAMM)
def_competen	text	Ano e Mês de competência da informação (AAAA/MM)
ano_competen	text	Ano de competência da informação
mes_competen	text	Mês de competência da informação
mun_MUNNOME	text	Nome (acentuado, maiúsculas e minúsculas) do Município (padrão DOS, página de código 850) do estabelecimento
mun_MUNNOMEX	text	Nome (sem acentos, em maiúsculas) do Município do estabelecimento
mun_AMAZONIA	text	Indica (S ou N) se o município do estabelecimento faz parte da Amazônia Legal (conforme IBGE)
mun_FRONTEIRA	text	Indica (S ou N) se o município do estabelecimento faz parte da faixa de fronteira (conforme IBGE)
mun_CAPITAL	text	Indica (S ou N) se o município do estabelecimento é capital da UF
mun_MSAUDCOD	int8	Código da Macrorregional de Saúde a que o Município do estabelecimento pertence
mun_RSAUDCOD	int8	Código da Regional de Saúde a que o Município do estabelecimento pertence
mun_CSAUDCOD	int8	Código da Microrregional de Saúde a que o Município do estabelecimento pertence
mun_LATITUDE	float8	Latitude da sede do Município do estabelecimento
mun_LONGITUDE	float8	Longitude da sede do Município do estabelecimento
mun_ALTITUDE	int8	Altitude, em metros, da sede do Município do estabelecimento
mun_AREA	float8	Área, em quilômetros quadrados, do Município do estabelecimento, segundo a Resolução 05, de 10/12/2002, do IBGE
mun_codigo_adotado	int8	Armazena o código atribuído ao município do estabelecimento, tratando os casos em que múltiplos códigos tenham sido utilizados para um mesmo município ao longo do tempo
uf_SIGLA_UF	text	Sigla da unidade da federação do estabelecimento
uf_CODIGO_UF	int8	Código IBGE da Unidade da Federação do estabelecimento
uf_NOME_UF	text	Nome da unidade da federação do estabelecimento
mun_coordenadas	text	Coordenadas do município do estabelecimento

Tabela 9 – Tabela de Códigos dos Municípios do Instituto Brasileiro de Geografia e Estatística.

Variável	**Tipo**	**Descrição**
COD_MUNICIPIO	int8	Código do município
COD_UF	int8	Código da unidade federativa
UF	text	Unidade federativa
NOME_MUNICIPIO	text	Nome d Município

Tabela 10 – Índice de Desenvolvimento Humano Municipal 2010.

Variável	**Tipo**	**Descrição**
UF	text	Unidade federativa
NOME_MUNICIPIO	text	Nome d Município
Ranking_IDHM_2010	int8	Posição no ranking do IDHM 2010
IDHM_2010	float8	Valor do IDHM 2010
IDHM_Renda_2010	float8	Valor do IDHM renda 2010
IDHM_Longevidade_2010	float8	Valor do IDHM longevidade 2010
IDHM_Educacao_2010	float8	Valor do IDHM educação 2010

APÊNDICE B – Relação das variáveis e suas respectivas categorias do conjunto de dados compilado usado nos experimentos.

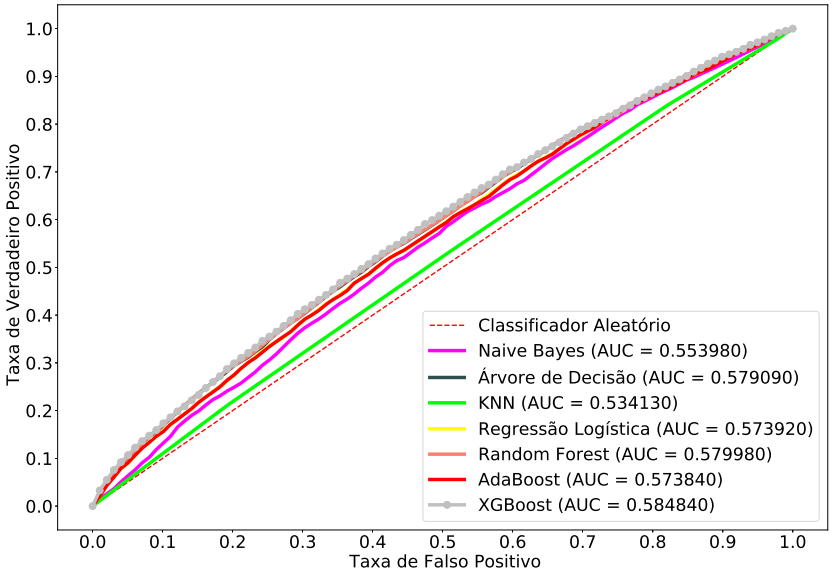
Variáveis	Categorias
Demográficas e socioeconômicas maternas	
Idade materna	(10-14, 15-19, 20-24, 25-29, 30-34, 35-39, 40-44, 45-49)
Raça/cor da pele	(Branca, Preta, Amarela, Parda, Indígena)
Estado civil	(Solteiro, Casado, Viúvo, Separado judicialmente/divorciado, União estável)
Escolaridade	(Nenhuma, de 1 a 3 anos, de 4 a 7 anos, 8 a 11 anos, 12 anos e mais)
Município de residência faz parte da Amazônia Legal	(Sim/Não)
Município de residência faz parte da faixa de fronteira	(Sim/Não)
Município de residência é capital de UF	(Sim/Não)
Índice de Desenvolvimento Humano Municipal 2010	(Alto, Baixo, Muito baixo, Médio)
Referentes à atenção/assistência à saúde na gestação e parto	
Local de nascimento	(Hospital, Outros estabelecimentos de saúde, Domicílio, Via pública, Outros)
Número de consultas durante o pré-natal	(Nenhuma, de 1 a 3, de 4 a 6, 7 e mais)
Tipo de parto	(Vaginal, Cesáreo)
Número de semanas de gestação	(<22, 22-27, 28-31, 32-36, 37-41, 42 ou mais)
Mês de gestação em que iniciou o pré-natal	(0-3, 4-7, 8-11, 12-15, 16-19, 20 ou mais)
Trabalho de parto induzido	(Sim, Não, Não se aplica)
Cesárea ocorreu antes do trabalho de parto iniciar	(Sim, Não, Não se aplica)
Assistência do nascimento	(Médico, Enfermeira/obstetriz, Parteira, Outros)
Município de nascimento faz parte da Amazônia Legal	(Sim/Não)
Município de nascimento é capital de UF	(Sim/Não)
Distância entre residência e local de nascimento (km)	(<50, 50-100, 100-150, 150-200, 200 ou mais)
Qtd. de salas de pré parto em centro obstétrico	(0-3, 4-7, 8-11, 12-15, 16-19, 20 ou mais)
Qtd. de salas de parto normal em centro obstétrico	(0-3, 4-7, 8-11, 12-15, 16-19, 20 ou mais)
Qtd. de salas de curetagem em centro obstétrico	(0-3, 4-7, 8-11, 12-15, 16-19, 20 ou mais)
Qtd. de salas de cirurgias em centro obstétrico	(0-3, 4-7, 8-11, 12-15, 16-19, 20 ou mais)
Qtd. de leitos de pré parto em centro obstétrico	(0-3, 4-7, 8-11, 12-15, 16-19, 20 ou mais)
Referentes às condições de nascimento do recém-nascido	
Semana de gestação	(Menos de 22 semanas, 22 a 27 semanas, 28 a 31 semanas, 32 a 36 semanas, 37 a 41 semanas, 42 semanas ou mais)
Tipo de gravidez	(Única, Dupla, Tripla e mais)
Tipo de apresentação do RN	(Cefálico, Pélvica ou podálica, Transversa)
Número de gestações anteriores	(0-3, 4-7, 8-11, 12-15, 16-19, 20 ou mais)
Número de partos vaginais	(0-3, 4-7, 8-11, 12-15, 16-19, 20 ou mais)
Número de partos cesáreos	(0-3, 4-7, 8-11, 12-15, 16-19, 20 ou mais)
Número de filhos vivos	(0-2, 3-5, 6-8, 9-11, 12 ou mais)
Número de filhos mortos	(0-2, 3-5, 6-8, 9-11, 12 ou mais)
Sexo	(Masculino, Feminino)
Apgar no primeiro minuto	(0-3: grave, 4-6: moderado, 7: leve, 8-10: ótimo)
Apgar no quinto minuto	(0-3: grave, 4-6: moderado, 7: leve, 8-10: ótimo)
Peso ao nascer em gramas	(<2500, 2500-2999, 3000-3999, 4000 ou mais)
Anomalia congênita	(Ignorado; Sim; Não)
Prematuridade do nascimento	(Termo, Inconclusivo-IG, Inconclusivo-Peso, Prematuro)
Qtd. salas/consultórios atend. pediátrico urg./emerg.	(0-3, 4-7, 8-11, 12-15, 16-19, 20 ou mais)
Qtd. salas de repouso/observação pediátrico urg./emerg.	(0-3, 4-7, 8-11, 12-15, 16-19, 20 ou mais)
Qtd. de salas de repouso ou obs. pediátrico de atend. ambul.	(0-3, 4-7, 8-11, 12-15, 16-19, 20 ou mais)
Qtd. de leitos repouso ou obs. pediátrico de urg./emerg.	(0-3, 4-7, 8-11, 12-15, 16-19, 20 ou mais)
Qtd. de leitos de repouso ou obs. pediátrico de atend. ambul.	(0-3, 4-7, 8-11, 12-15, 16-19, 20 ou mais)
Existência instalação física atend. hosp. em centro obstétrico	(Sim/Não)

APÊNDICE C – Resultados dos classificadores por grupo ou componente da mortalidade neonatal.

Técnica de Amostragem	ID Grupo de dados	Descrição Grupo de dados	Métrica	Naive Bayes				Árvore de Decisão				Regressão Logística				Random Forest				KNN				AdaBoost				XGBoost			
				Sem Grade Hiperparam		Com Grade Hiperparam		Sem Grade Hiperparam		Com Grade Hiperparam		Sem Grade Hiperparam		Com Grade Hiperparam		Sem Grade Hiperparam		Com Grade Hiperparam		Sem Grade Hiperparam		Com Grade Hiperparam		Sem Grade Hiperparam		Com Grade Hiperparam		Sem Grade Hiperparam		Com Grade Hiperparam	
				Média	Desvio Padrão	Média	Desvio Padrão	Média	Desvio Padrão	Média	Desvio Padrão	Média	Desvio Padrão	Média	Desvio Padrão	Média	Desvio Padrão	Média	Desvio Padrão	Média	Desvio Padrão	Média	Desvio Padrão	Média	Desvio Padrão	Média	Desvio Padrão	Média	Desvio Padrão	Média	Desvio Padrão
RANDOM UNDER SAMPLER	1	SOCIOECONÔMICOS	ACURÁCIA	0.53674	0.00033	0.53667	0.00039	0.5552	0.00068	0.55536	0.00067	0.55083	0.00028	0.55077	0.00026	0.55603	0.00077	0.55617	0.00072	0.52522	0.00227	0.52419	0.00089	0.55077	0.00028	0.55066	0.00026	0.55915	0.00057	0.55987	0.0006
			PRECISÃO	0.55513	0.00056	0.55503	0.00058	0.56131	0.00064	0.56143	0.00069	0.55707	0.00031	0.55695	0.00029	0.56098	0.00079	0.56111	0.00064	0.52565	0.00253	0.52398	0.00097	0.55698	0.00031	0.55689	0.00033	0.5652	0.00068	0.56644	0.00063
			RECALL	0.37016	0.00103	0.37007	0.00101	0.50557	0.00236	0.50626	0.00272	0.49628	0.0011	0.4967	0.00092	0.51563	0.00351	0.51609	0.00264	0.52308	0.01127	0.53025	0.00278	0.49638	0.00102	0.49601	0.0007	0.51293	0.00205	0.51071	0.00277
			F1-SCORE	0.44408	0.00076	0.44399	0.00085	0.53192	0.00148	0.53234	0.00159	0.52489	0.00063	0.52508	0.00053	0.53727	0.00197	0.5376	0.00162	0.52352	0.00543	0.52684	0.00128	0.52491	0.0006	0.52466	0.0004	0.53774	0.00115	0.53706	0.00157
			AUC	0.55398	0.00021	0.554	0.00018	0.57909	0.00044	0.57906	0.00054	0.57392	0.00015	0.57394	0.00016	0.57998	0.00065	0.57983	0.00053	0.53413	0.00279	0.53042	0.00067	0.57384	0.00016	0.57384	0.00014	0.58484	0.00043	0.58624	0.00035
			TEMPO	10.6 s		20.1 s		24.7 s		46.4 s		1min 12s		3min 17s		5min 9s		14min 27s		56min 56s		2h 8min 40s		5min 14s		29min 26s		11min 54s		34min 38s	
	2	COMPORTAMENTAIS E USO DE SERVIÇOS DE SAÚDE	ACURÁCIA	0.75374	0.00009	0.75377	0.0001	0.78812	0.00033	0.78814	0.00037	0.79452	0.00008	0.79442	0.00012	0.7908	0.00042	0.79068	0.00041	0.76586	0.00424	0.75325	0.00082	0.7941	0.00013	0.79412	0.0001	0.79588	0.0003	0.79625	0.00029
			PRECISÃO	0.8473	0.00019	0.84735	0.00024	0.84225	0.00069	0.84225	0.00062	0.85396	0.00017	0.85412	0.00019	0.83748	0.00055	0.83754	0.00061	0.80169	0.00963	0.77351	0.00147	0.85556	0.0002	0.85555	0.00022	0.84172	0.00059	0.84194	0.00044
			RECALL	0.61908	0.00013	0.6191	0.00014	0.70907	0.00084	0.70915	0.00084	0.7106	0.00014	0.71016	0.0002	0.72167	0.00081	0.72131	0.00066	0.70931	0.00572	0.71668	0.00115	0.70771	0.00022	0.70776	0.00023	0.72886	0.00067	0.72949	0.00068
			F1-SCORE	0.7154	0.0001	0.71544	0.00011	0.76992	0.00042	0.76997	0.00047	0.77569	0.00009	0.77549	0.00014	0.77525	0.00051	0.77507	0.00046	0.75201	0.0026	0.7439	0.00072	0.77462	0.00015	0.77465	0.00012	0.78121	0.00035	0.78166	0.00037
			AUC	0.8166	0.00008	0.81657	0.00009	0.83662	0.0004	0.83657	0.0004	0.85423	0.00006	0.8541	0.00006	0.84864	0.00025	0.84839	0.0003	0.81985	0.00207	0.80425	0.00057	0.85406	0,00007	0.85402	0.00009	0.85679	0.00017	0.85748	0.00017
			TEMPO	13.1 s		27.3 s		36.1 s		1min 14s		2min 6s		5min 53s		8min 35s		24min 21s		55min 48s		2h 7min 57s		7min 23s		40min 15s		15min 54s		46min 40s	
	3	RELACIONADOS AO RECÉM NASCIDO, PARTO E BIOLÓGICOS MATERNOS	ACURÁCIA	0.87823	0.00014	0.87881	0.00018	0.87167	0.00066	0.87174	0.00051	0.89839	0.00012	0.89851	0.0001	0.88795	0.00037	0.8872	0.00037	0.8785	0.00065	0.86996	0.00046	0.89691	0.00015	0.8969	0.00018	0.89893	0.00025	0.8995	0.0002
			PRECISÃO	0.92387	0.0003	0.92126	0.00027	0.90677	0.00082	0.90682	0.00075	0.93046	0.00018	0.93016	0.00015	0.91156	0.00058	0.91068	0.00062	0.92131	0.0009	0.90199	0.00068	0.93295	0.00027	0.93303	0.00026	0.92732	0.00045	0.92853	0.00038
			RECALL	0.82442	0.00036	0.82846	0.00039	0.82854	0.00116	0.82866	0.00082	0.86115	0.00015	0.86173	0.00017	0.85931	0.00044	0.85864	0.0005	0.82773	0.00093	0.83015	0.00065	0.8553	0.00029	0.8552	0.00038	0.86575	0.00031	0.86564	0.00031
			F1-SCORE	0.8713	0.00016	0.87238	0.0002	0.86588	0.00073	0.86596	0.00055	0.89446	0.00013	0.89463	0.00011	0.88465	0.00037	0.88388	0.00037	0.872	0.0007	0.86457	0.00048	0.89243	0.00015	0.89241	0.0002	0.89546	0.00025	0.89597	0.0002
			AUC	0.92655	0.0001	0.92629	0.0001	0.87522	0.00093	0.87526	0.00077	**0.94769**	**0.00003**	**0.94761**	**0.00003**	**0.93421**	**0.00027**	**0.93353**	**0.00027**	0.91951	0.00047	0.90731	0.00031	**0.94728**	**0.00004**	**0.94728**	**0.00003**	**0.94753**	**0.0001**	**0.94812**	**0.00009**
			TEMPO	15.7 s		34 s		56.9 s		1min 41s		**2min 41s**		**7min 22s**		**9min 42s**		**26min 30s**		58min 3s		2h 13min 10s		**10min 8s**		**55min 37s**		**19min 20s**		**57min 42s**	
	4	COMPOSTA PELA INTEGRAÇÃO DOS DADOS DE I, II E III	ACURÁCIA	0.87269	0.00019	0.87251	0.00024	0.83796	0.00094	0.83826	0.00088	0.90003	0.00014	0.89997	0.00011	0.89766	0.00034	0.89705	0.00035	0.86105	0.00046	0.85499	0.00037	0.89777	0.00023	0.89772	0.00022	0.90065	0.00029	0.90129	0.00028
			PRECISÃO	0.90713	0.00038	0.90555	0.00041	0.83504	0.0013	0.83539	0.00117	0.9312	0.00018	0.93142	0.00016	0.92015	0.00048	0.92018	0.0005	0.92725	0.00065	0.91143	0.00051	0.93212	0.00036	0.93227	0.00033	0.9283	0.00036	0.92947	0.0004
			RECALL	0.83046	0.00025	0.83184	0.00035	0.84237	0.00143	0.84259	0.00128	0.86391	0.00022	0.86354	0.00018	0.87093	0.00036	0.86954	0.00052	0.7836	0.00067	0.78642	0.00051	0.85803	0.0004	0.85778	0.00041	0.86838	0.0005	0.86849	0.00031
			F1-SCORE	0.86708	0.00019	0.8671	0.00025	0.83867	0.00094	0.83896	0.00087	0.89628	0.00015	0.89618	0.00012	0.89485	0.00034	0.89413	0.00036	0.84938	0.00052	0.84431	0.00041	0.89353	0.00024	0.89346	0.00024	0.89733	0.00031	0.89794	0.00028
			AUC	0.91553	0.00014	0.91537	0.00012	0.83734	0.00094	0.83765	0.00089	**0.94954**	**0.00005**	**0.94948**	**0.00004**	**0.94465**	**0.00018**	**0.94349**	**0.0002**	0.90884	0.00043	0.89496	0.00031	**0.94853**	**0,00006**	**0.94854**	**0,00006**	**0.94952**	**0.00012**	**0.9503**	**0,00008**
			TEMPO	27.1 s		1min 4s		2min 8s		4min 3s		**5min 3s**		**14min 15s**		**15min 2s**		**41min 29s**		59min 52s		2h 20min 15s		**21min 10s**		**1h 58min 7s**		**33min 26s**		**1h 33min 17s**	
	5	COMPOSTA PELOS DADOS DE IV COM VARIÁVEIS ENRIQUECIDAS	ACURÁCIA	0.87036	0.00017	0.8703	0.00014	0.83842	0.00092	0.83849	0.00083	0.90031	0.00016	0.90037	0.00012	0.89975	0.00036	0.89917	0.00045	0.85888	0.00046	0.85178	0.00035	0.89841	0.0002	0.89832	0.00016	0.90201	0.00034	0.90245	0.00031
			PRECISÃO	0.88642	0.00046	0.88774	0.0003	0.8349	0.00123	0.83485	0.00106	0.93257	0.00025	0.9328	0.00013	0.92353	0.00052	0.92356	0.00055	0.92917	0.00082	0.91104	0.00049	0.93375	0.00037	0.93362	0.0003	0.93021	0.00041	0.93139	0.00037
			RECALL	0.84964	0.00065	0.84786	0.00033	0.84372	0.00131	0.84396	0.00087	0.86305	0.00021	0.86292	0.00018	0.87169	0.00041	0.8704	0.00058	0,777	0.00068	0.77973	0.00056	0.85769	0.00042	0.85763	0.00029	0.86925	0.0004	0.86892	0.00048
			F1-SCORE	0.86761	0.00021	0.86732	0.00014	0.83927	0.00091	0.83936	0.00079	0.89645	0.00016	0.89649	0.00013	0.89685	0.00036	0.89618	0.00046	0.84628	0.0005	0.84027	0.00039	0.89409	0.00021	0.894	0.00016	0.89869	0.00035	0.89906	0.00033
			AUC	0.91008	0.00015	0.91045	0.00016	0.8381	0.0009	0.83817	0.00084	**0.95029**	**0.00004**	**0.95024**	**0.00005**	**0.94639**	**0.00018**	**0.94526**	**0.00021**	0.90785	0.00034	0.8929	0.00027	**0.94912**	**0,00008**	**0.94909**	**0,00008**	**0.95065**	**0.00014**	**0.95129**	**0.00015**
			TEMPO	28.5 s		1min 9s		2min 26s		4min 28s		**5min 27s**		**15min 22s**		**16min 10s**		**42min 47s**		1h 27s		2h 22min		**22min 37s**		**2h 31s**		**35min 17s**		**1h 40min 13s**	
SMOTE	3	RELACIONADOS AO RECÉM NASCIDO, PARTO E BIOLÓGICOS MATERNOS	ACURÁCIA													0.93629	0.00001							0.90105	0.00002						
			PRECISÃO													0.95479	0.00002							0.93435	0.00006						
			RECALL													0.91596	0.00002							0.86273	0.00006						
			F1-SCORE													0.93497	0.00001							0.89711	0.00002						
			AUC													**0.98051**	**0.00001**							**0.95127**	**0.00001**						
			TEMPO													**3d 7h 37min 34s**								**1d 12h 6min 50s**							

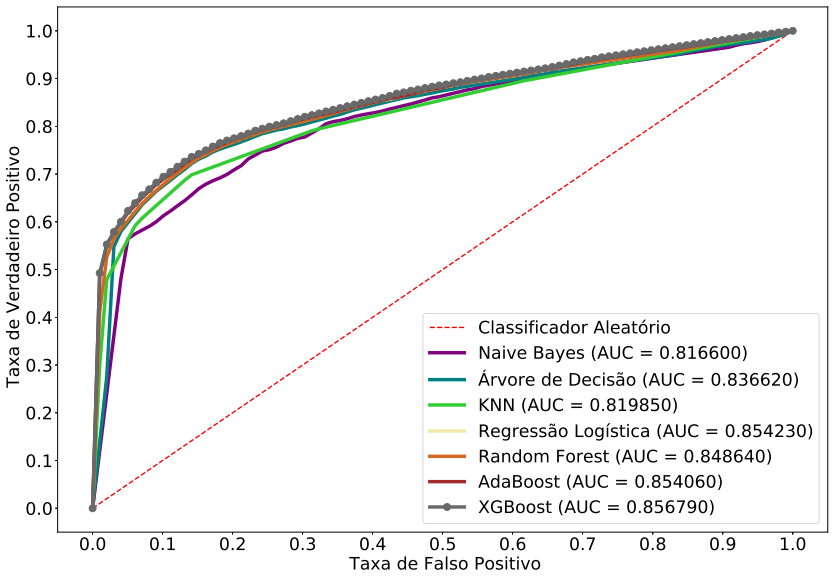
APÊNDICE D – Gráficos com as curvas ROC dos classificadores gerados aplicando subamostragem aleatória ou SMOTE aos dados - sem grade de hiperparâmetros.

Figura 15 – Classificadores usando o grupo de dados com variáveis socioeconômicas (Grupo 1) e aplicando subamostragem aleatória.



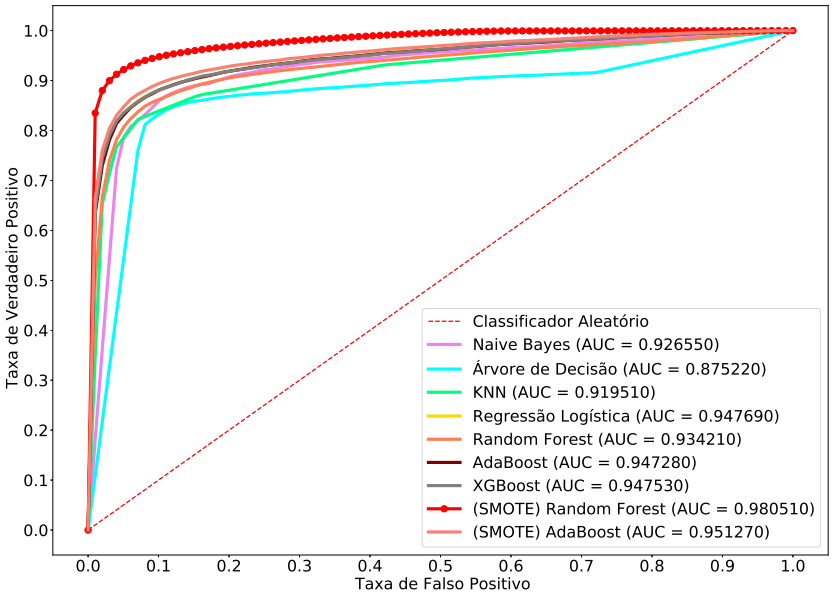
Fonte: Elaborado pelo autor (2022).

Figura 16 – Classificadores usando grupo de dados com variáveis comportamentais e de uso do serviço de saúde (Grupo 2), e aplicando subamostragem aleatória.



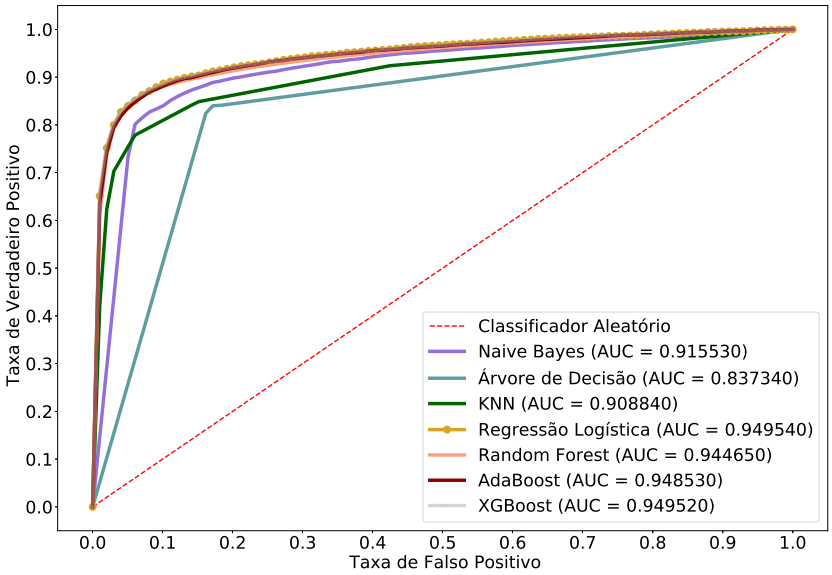
Fonte: Elaborado pelo autor (2022).

Figura 17 – Classificadores usando grupo de dados com variáveis relacionadas ao recém-nascido, parto e biológicos maternos (Grupo 3), e aplicação de subamostragem aleatória e SMOTE.



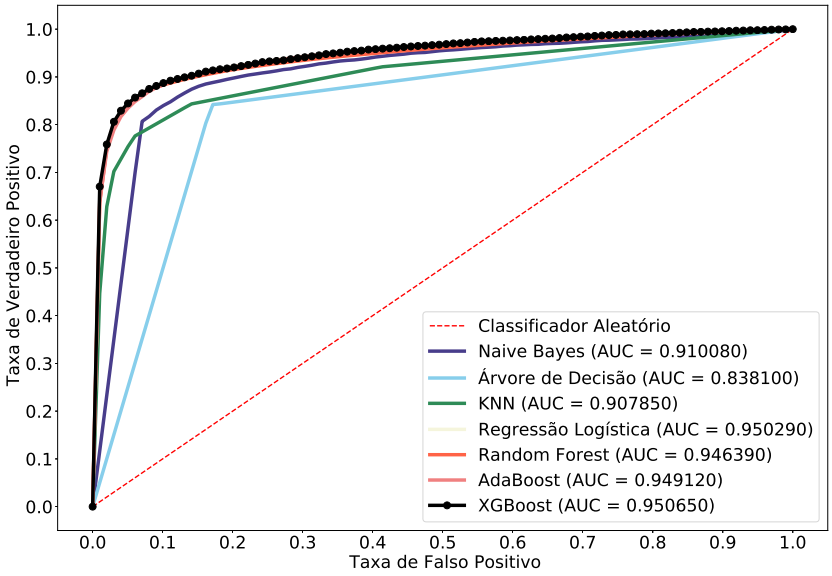
Fonte: Elaborado pelo autor (2022).

Figura 18 – Classificadores usando o grupo de dados composto pela integração dos grupos de dados 1, 2 e 3 (Grupo 4); com aplicação de subamostragem aleatória.



Fonte: Elaborado pelo autor (2022).

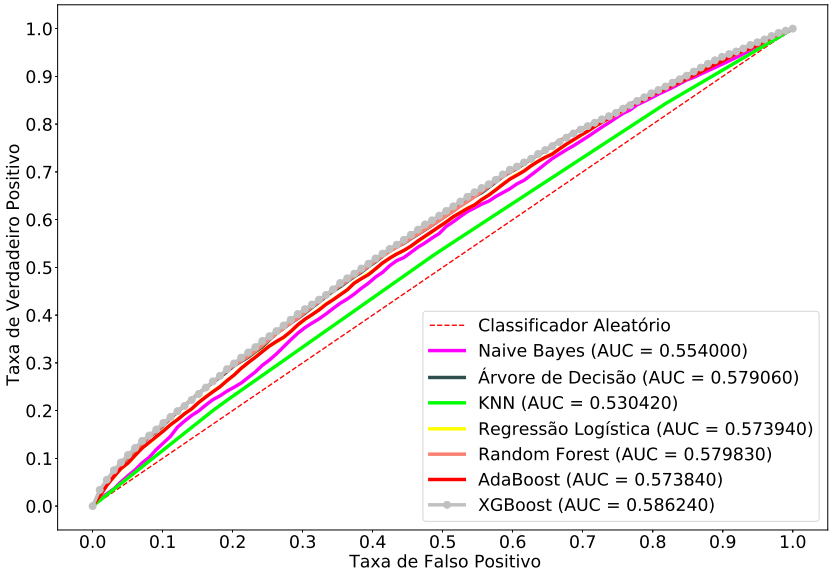
Figura 19 – Classificadores usando o grupo de dados compilado (Grupo 5) com aplicação de subamostragem aleatória.



Fonte: Elaborado pelo autor (2022).

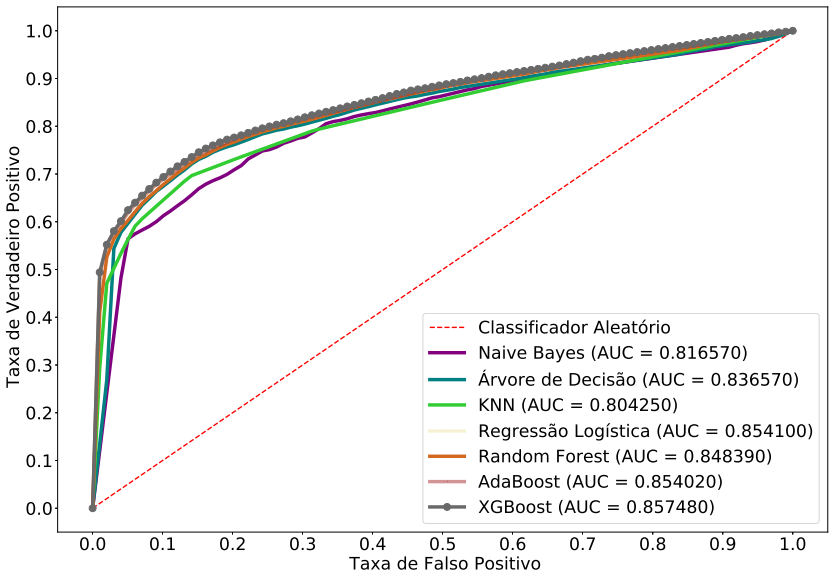
APÊNDICE E – Gráficos com as curvas ROC dos classificadores gerados aplicando subamostragem aleatória ou SMOTE - com grade de hiperparâmetros

Figura 21 – Classificadores usando grupo de dados com variáveis socioeconômicas (Grupo 1) e aplicação de subamostragem aleatória.



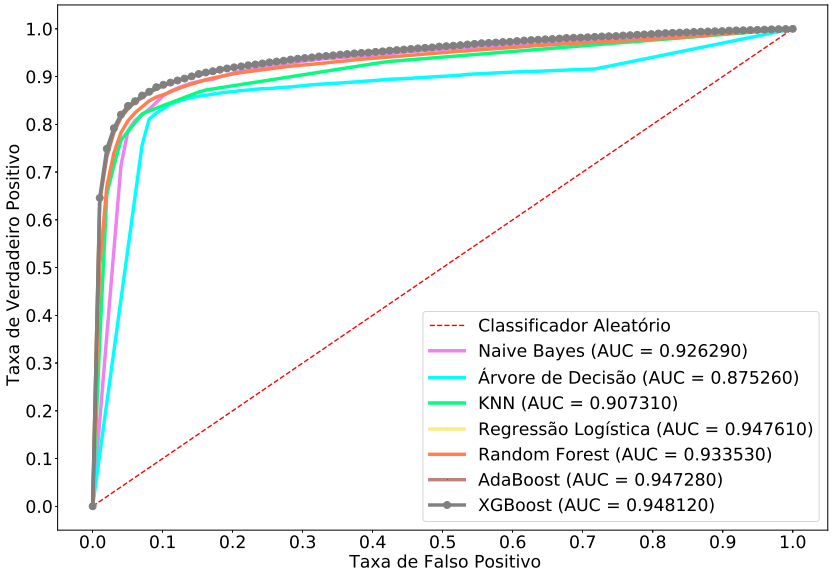
Fonte: Elaborado pelo autor (2022).

Figura 22 – Classificadores usando grupo de dados com variáveis comportamentais e de uso do serviço de saúde (Grupo 2), e aplicação de subamostragem aleatória.



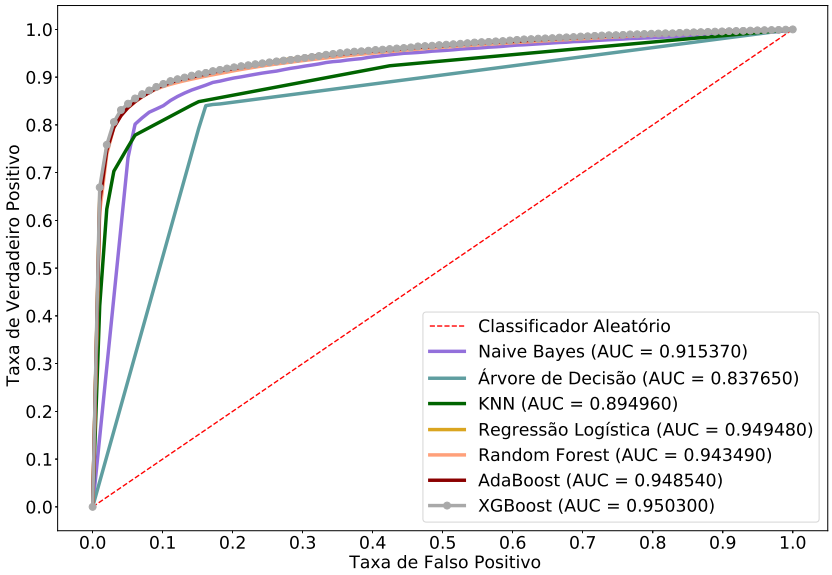
Fonte: Elaborado pelo autor (2022).

Figura 23 – Classificadores usando grupo de dados com variáveis relacionadas ao recém-nascido, parto e biológicos maternos (Grupo 3), e aplicação de subamostragem aleatória.



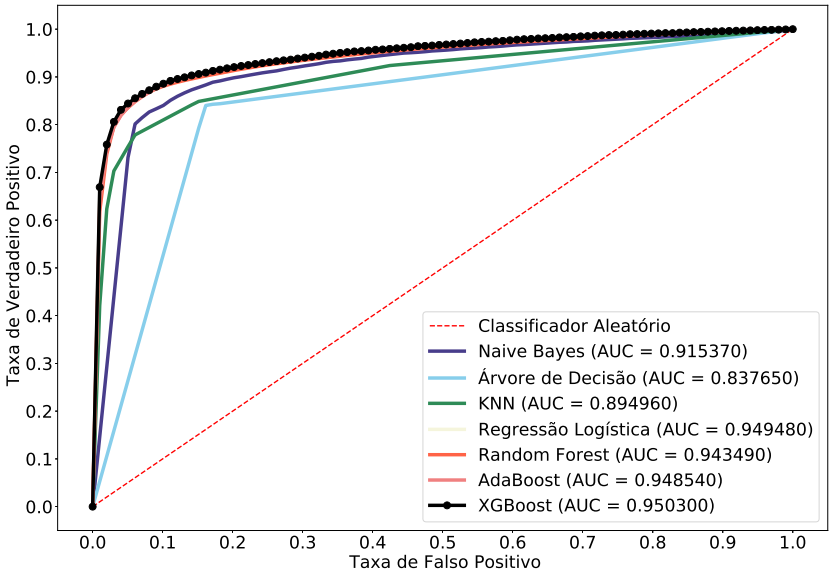
Fonte: Elaborado pelo autor (2022).

Figura 24 – Classificadores usando o grupo de dados composto pela integração dos conjuntos de dados 1, 2 e 3 (Grupo 4); e aplicação de subamostragem aleatória.



Fonte: Elaborado pelo autor (2022).

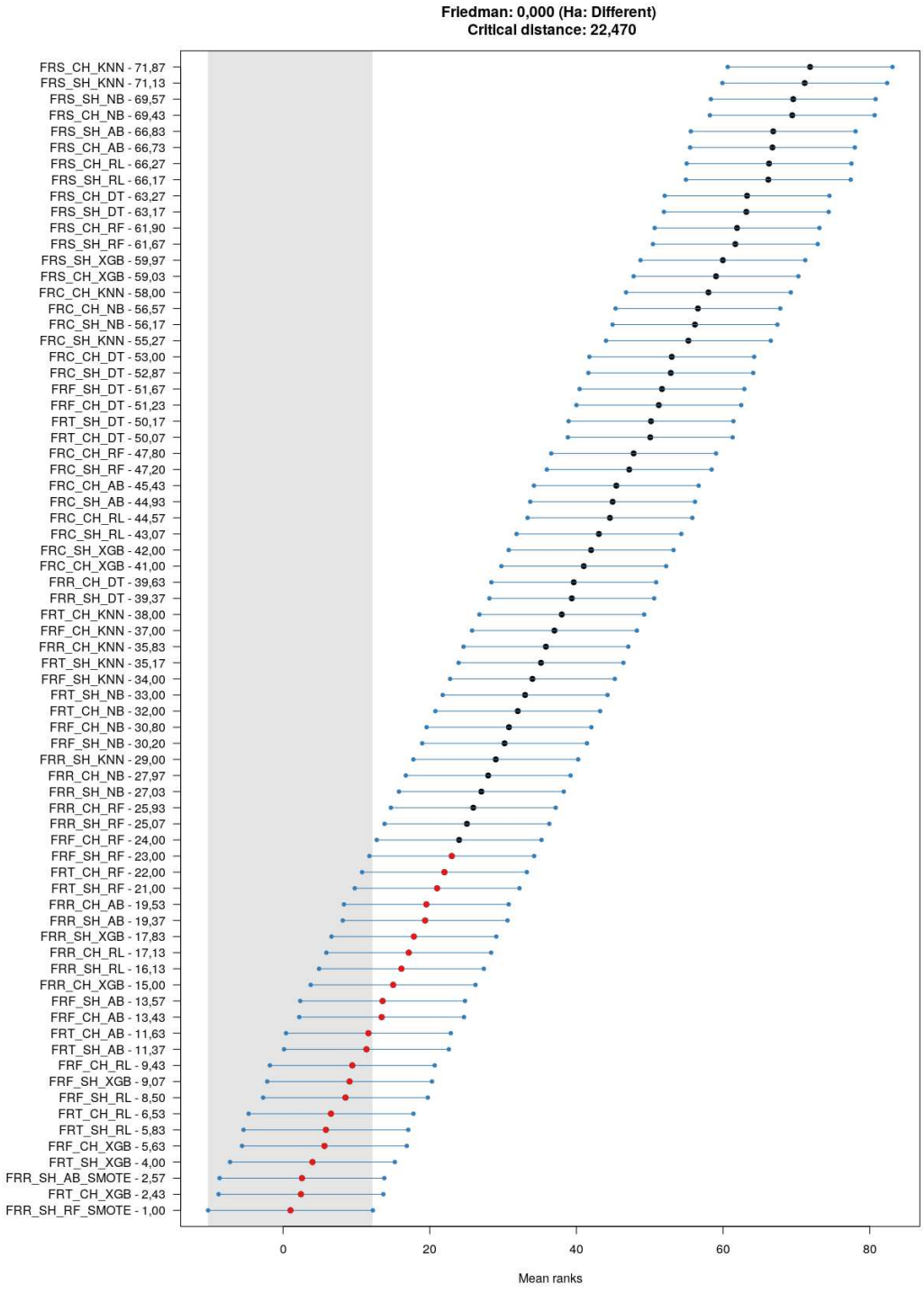
Figura 25 – Classificadores usando o grupo de dados compilado (Grupo 5) com aplicação de subamostragem aleatória.



Fonte: Elaborado pelo autor (2022).

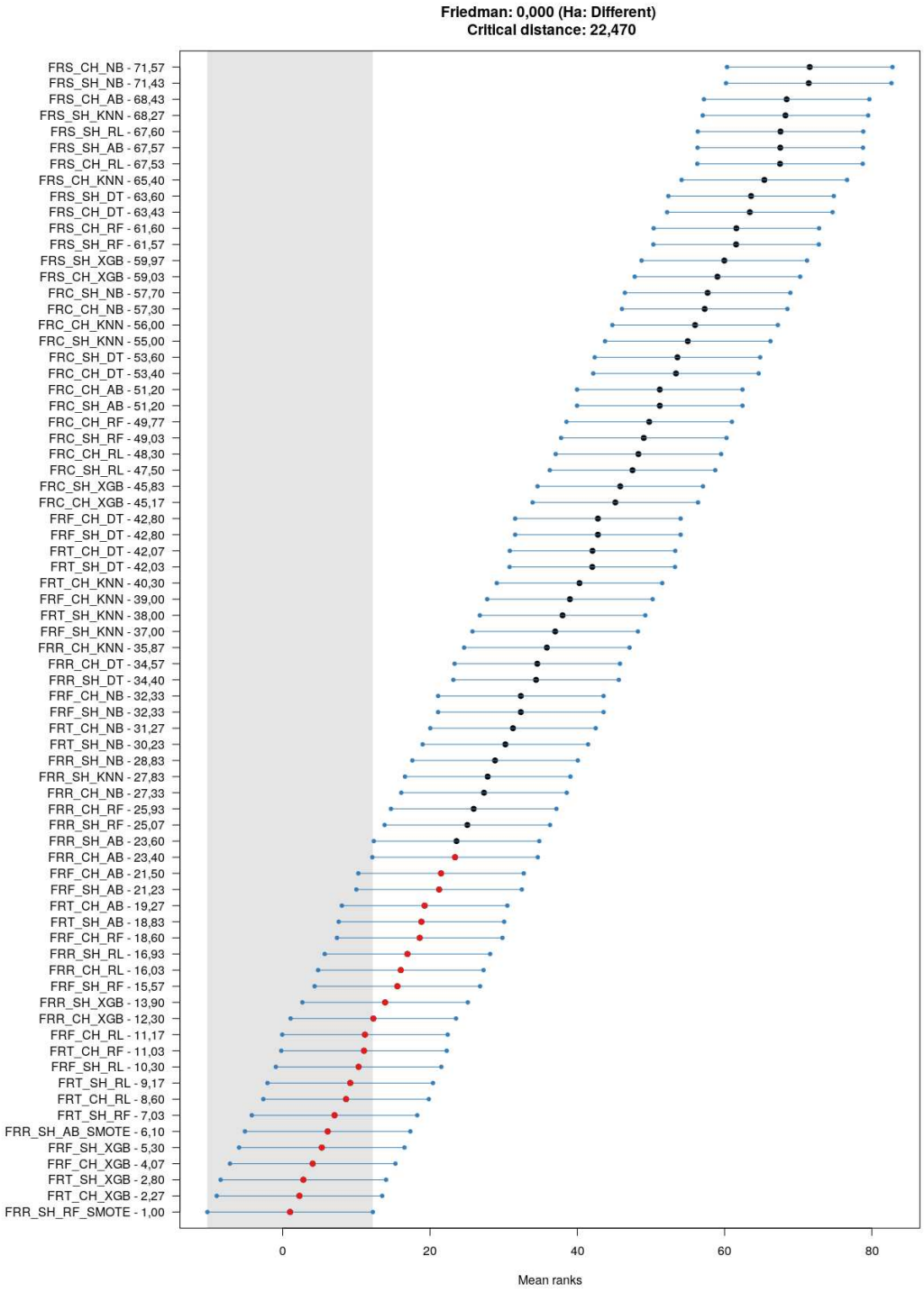
APÊNDICE F – Gráficos complementares do teste de Friedman e *post-hoc* de Nemenyi para as métricas AUC e *f1-score*.

Figura 27 – Gráfico complementar com o resultado do teste de Friedman e *post-hoc* de Nemenyi para a métrica AUC.



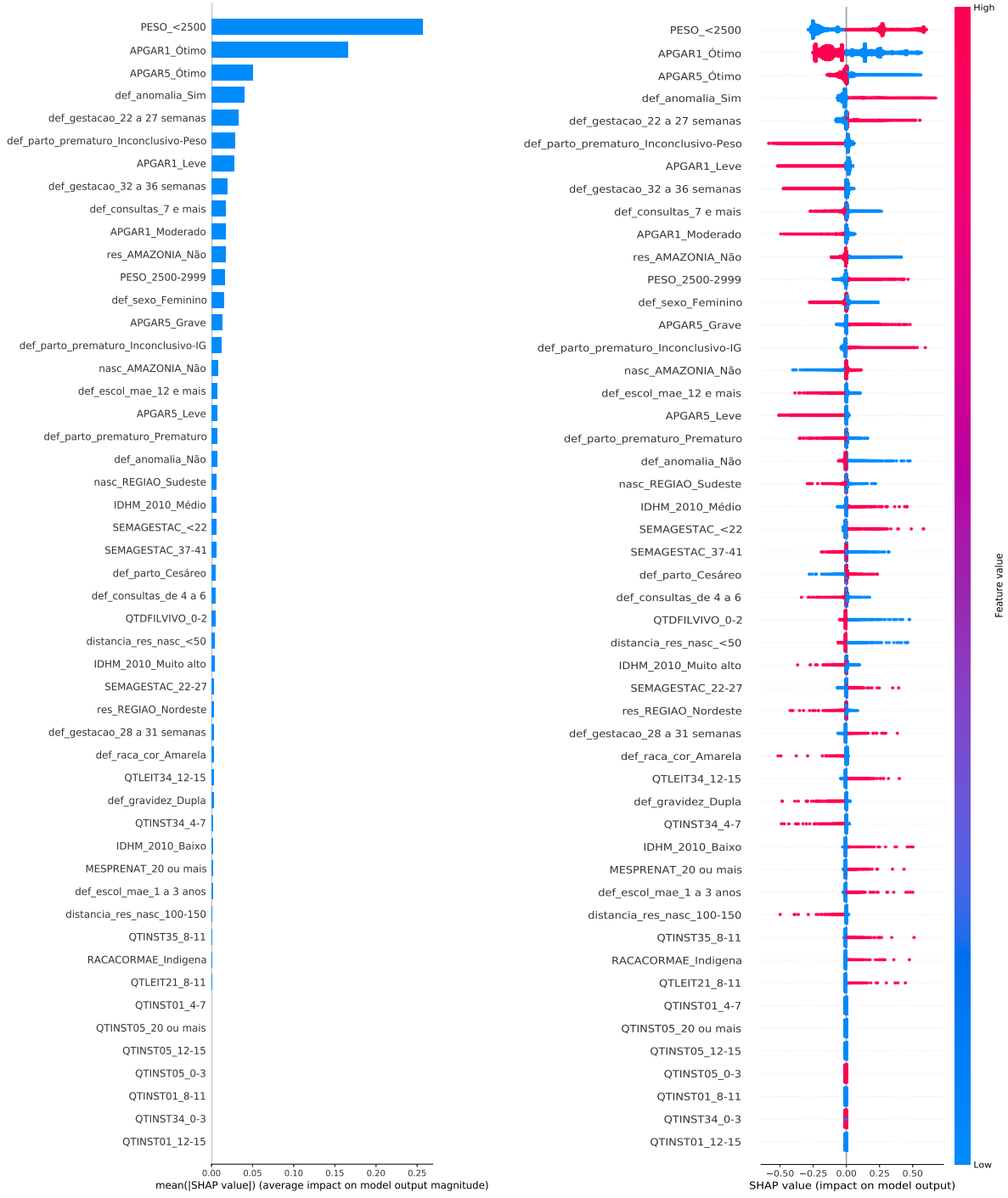
Fonte: Elaborado pelo autor (2022).

Figura 28 – Gráfico complementar com o resultado do teste de Friedman e *post-hoc* de Nemenyi para a métrica *f1-score*.



Fonte: Elaborado pelo autor (2022).

APÊNDICE G – As 50 variáveis mais relevantes a nível global na predição do óbito neonatal - Classificador AdaBoost usando o grupo de dados compilado (Grupo 5).



(a) Relevância preditiva

(b) Força preditiva

Fonte: Elaborado pelo autor (2022).