UNIVERSIDADE ESTADUAL DO OESTE DO PARANÁ

CAMPUS DE FOZ DO IGUAÇU

PROGRAMA DE PÓS-GRADUAÇÃO EM
ENGENHARIA ELÉTRICA E COMPUTAÇÃO

DISSERTAÇÃO DE MESTRADO

PREDIÇÃO DE MORTE DE CRIANÇAS ABAIXO DE 1 ANO NO ESTADO DO PARANÁ

WAGNER JORCUVICH NUNES DA SILVA

FOZ DO IGUAÇU
2023

Wagner Jorcuvich Nunes da Silva

Predição de morte de crianças abaixo de 1 ano no estado do Paraná

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Engenharia Elétrica e Computação da Universidade Estadual do Paraná como parte dos requisitos para obtenção do título de Mestre em Engenharia Elétrica e Computação. Área de concentração: Sistemas Biomédicos.

Orientador: Renato Bobsin Machado

Foz do Iguaçu
2023

Ficha de identificação da obra elaborada através do Formulário de Geração Automática do Sistema de Bibliotecas da Unioeste.

Jorcuvich Nunes da Silva, Wagner
PREDIÇÃO DE MORTE DE CRIANÇAS ABAIXO DE 1 ANO NO ESTADO DO PARANÁ / Wagner Jorcuvich Nunes da Silva; orientador Renato Bobsin Machado. -- Foz do Iguaçu, 2023.
68 p.

Dissertação (Mestrado Acadêmico Campus de Foz do Iguaçu) -- Universidade Estadual do Oeste do Paraná, Centro de Engenharias e Ciências Exatas, Programa de Pós-Graduação em Engenharia Elétrica e Computação, 2023.

1. PREDIÇÃO. 2. APRENDIZADO COMPUTACIONAL. 3. CLASSIFICAÇÃO. 4. EPIDEMIOLOGIA. I. Bobsin Machado, Renato, orient. II. Título.

Predição de morte de crianças abaixo de 1 ano no estado do Paraná

Wagner Jorcuvich Nunes da Silva

Esta Dissertação de Mestrado foi apresentada ao Programa de Pós-Graduação em Engenharia Elétrica e Computação e aprovada pela Banca Examinadora assim constituída:

Prof. Dr. **Renato Bobsin Machado** - (Orientador)
Universidade Estadual do Oeste do Paraná - UNIOESTE

Prof. Dr. **Romeu Reginatto**
Universidade Estadual do Oeste do Paraná - UNIOESTE

Profª. Dra. **Isabel Fernandes de Souza**
Centro Universitário União das Américas - UNIAMÉRICA

Data da defesa: 07 de junho de 2023.

Resumo

Este trabalho aborda a importância da utilização de técnicas de aprendizado de máquina na área da saúde, especificamente na predição da mortalidade de crianças com menos de um ano de idade. A mortalidade infantil é um problema significativo que afeta milhões de crianças em todo o mundo e requer uma abordagem efetiva para reduzir essas mortes evitáveis. No estudo, foram utilizados algoritmos de aprendizado de máquina, como Máquinas de Vetores de Suporte (*Support Vector Machines* - SVM) , k-Vizinhos Mais Próximos (*k-Nearest Neighbors* - kNN) e *Naive Bayes* (NB), para desenvolver modelos preditivos. Esses modelos foram treinados com base em dados demográficos e relacionados à saúde, coletados de uma grande amostra de base pública. A aplicação de técnicas de redução de dimensionalidade, como o teste qui-quadrado e o teste *t-Student*, permitiu selecionar os atributos mais relevantes e reduzir a complexidade do conjunto de dados. Para avaliar o desempenho dos modelos, foram utilizadas métricas como acurácia, taxa de erro, sensibilidade, especificidade, precisão e pontuação F1. Além disso, a área sob a curva característica de operação do receptor (AUC-ROC) foi empregada como medida de desempenho para avaliar a capacidade de discriminação dos modelos. A utilização de técnicas de aprendizado de máquina na área da saúde, como a predição da mortalidade infantil, pode ter um impacto significativo no direcionamento de recursos e na implementação de intervenções adequadas. Ao identificar precocemente os fatores de risco e prever o risco de mortalidade, é possível adotar medidas preventivas e estratégias de intervenção de modo mais eficiente. Os resultados deste estudo podem contribuir para a compreensão da aplicação do aprendizado de máquina na saúde, fornecendo informações valiosas para profissionais da área e auxiliando na tomada de decisões para melhorar a saúde e o bem-estar das crianças.

Palavras-chave: Predição de mortalidade infantil, Aprendizado de máquina, Modelos de classificação, Tomada de decisão em saúde, Conjunto de dados de saúde pública.

Abstract

This study addresses the importance of utilizing machine learning techniques in the healthcare field, specifically in predicting mortality in children under one year of age. Infant mortality is a significant problem that affects millions of children worldwide and requires an effective approach to reduce these preventable deaths. In this study, machine learning algorithms such as Support Vector Machines (SVM), k-Nearest Neighbors (kNN), and Naive Bayes (NB) were used to develop predictive models. These models were trained based on demographic and health-related data collected from a large publicly available dataset. The application of dimensionality reduction techniques, such as the chi-square test and Student's t-test, allowed for the selection of the most relevant attributes and reduction of dataset complexity. Performance metrics such as accuracy, error rate, sensitivity, specificity, precision, and F1 score were employed to evaluate the models' performance. Additionally, the area under the receiver operating characteristic curve (AUC-ROC) was used as a performance measure to assess the models' discrimination capability. The utilization of machine learning techniques in healthcare, such as the prediction of infant mortality, can have a significant impact on resource allocation and the implementation of appropriate interventions. By early identifying risk factors and predicting mortality risk, preventive measures and intervention strategies can be adopted more efficiently. The results of this study can contribute to the understanding of the application of machine learning in healthcare, providing valuable insights for healthcare professionals and aiding in decision-making to improve the health and well-being of children.

Keywords: Infant mortality prediction, Machine learning, Classification models, Healthcare decision-making, Public health dataset.

Agradecimentos

Gostaria de expressar minha profunda gratidão a todos que contribuíram para a conclusão desta dissertação de mestrado. Esta jornada acadêmica foi desafiadora, mas também enriquecedora, e sou sinceramente grato às pessoas que me apoiaram ao longo do caminho.

Primeiramente, estendo meu mais sincero agradecimento ao meu orientador, Professor Dr. Renato Bobsin Machado, pela orientação inestimável, apoio incansável em momento extremamente difíceis. Sua mentoria foi fundamental para moldar a trajetória da minha pesquisa e crescimento acadêmico.

Também agradeço aos membros da banca de defesa da tese, Professor Dr. Romeu Reginatto e Professora Dra. Isabel Fernandes Souza, pelo tempo dedicado, expertise e feedback construtivo. Suas contribuições cuidadosas enriqueceram a qualidade deste trabalho.

Agradeço especialmente à Professora Dra. Adriana Tokuhashi Kauati pelo apoio contínuo e assistência ao longo deste empreendimento. Seu estímulo e apoio cuidadoso foram cruciais para trilhar desafios.

Por fim, dedico esta dissertação à minha mãe e à memória de meu pai.

A todos os mencionados e aos inúmeros outros que desempenharam um papel em minha jornada acadêmica, sou verdadeiramente grato por suas contribuições, incentivo e apoio. Obrigado por serem parte integrante deste capítulo significativo em minha vida.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 88887.502540/2020-00

Atenciosamente,

Wagner Jorcuvich

"Tudo que podemos fazer é superar a tristeza e aprender alguma coisa com ela, mas isso de nada nos adianta para enfrentar a tristeza seguinte que nos atinge sem aviso."

Haruki Murakami

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Siglas e Abreviaturas

AUC-ROC	*Area Under the Receiver Operating Characteristic Curve*
CEP	Comitê de Ética em Pesquisa
CONEP	Comissão Nacional de Ética em Pesquisa
CNS	Conselho Nacional de Saúde
CID	Classificação Internacional de Doenças
DATASUS	Departamento de Informática do Sistema Único de Saúde
FP	Falsos Positivos
FN	Falsos Negativos
IPEA	Instituto de Pesquisa Econômica Aplicada
KDD	*Knowledge-Discovery in Databases*
LGPD	Lei Geral de Proteção de Dados
RBF	*Radial-basis Function*
SIM	Sistema de Informação sobre Mortalidade
SINAN	Sistema de Informação de Agravos de Notificação
SINASC	Sistema de Informação sobre Nascidos Vivos
SI-PNI	Sistema de Informação do Programa Nacional de Imunizações
SISCAN	Sistema de Informação de Câncer
SISPNCD	Sistema do Programa Nacional de Controle da Dengue
SMO	*Sequential Minimal Optimization*
SUS	Sistema Único de Saúde
SVM	*Support Vector Machine*
UN IGME	*United Nations Inter-agency Group for Child Mortality Estimation*
VP	Verdadeiros Positivos
VN	Verdadeiros Negativos

Capítulo 1

Introdução

A aplicação da ciência de dados tem experimentado um crescimento significativo em diversas áreas, incluindo a saúde pública. Cada vez mais, essa disciplina tem sido utilizada para aprimorar os cuidados prestados à população. O aumento no volume e na variedade dos dados disponíveis, bem como a velocidade com que são obtidos, criaram um ambiente propício para o surgimento de soluções baseadas em Inteligência Artificial.

A intersecção entre ciência de dados e saúde pública tem se tornado tão relevante que universidades ao redor do mundo desenvolveram cursos específicos voltados para a aplicação de ciência de dados nessa área. Hospitais e instituições de saúde têm buscado cada vez mais profissionais capacitados nesse campo, com o objetivo de desenvolver soluções que aprimorem o diagnóstico de doenças e o tratamento de seus pacientes.

A aplicação de algoritmos de inteligência artificial e análise de dados pode fornecer *insights* valiosos para auxiliar na prevenção, diagnóstico e tratamento de doenças em crianças. Essas técnicas são capazes de processar esses grandes volumes de informações e identificar padrões complexos, possibilitando a identificação de fatores de risco e a predição de eventos adversos.

Por exemplo, a análise de dados pode ser utilizada para identificar correlações entre variáveis demográficas, socioeconômicas e de saúde, auxiliando na identificação de grupos de maior vulnerabilidade e direcionando esforços de intervenção preventiva. Além disso, algoritmos de aprendizado de máquina podem ser empregados no desenvolvimento de sistemas de diagnóstico automatizados, que auxiliam os profissionais de saúde na detecção precoce de doenças e na adoção de medidas de tratamento adequadas.

Adicionalmente, a aplicação de aprendizado de máquina viabiliza a análise de dados epidemiológicos e de saúde coletados em larga escala, permitindo a identificação de tendências, padrões e fatores de risco associados à mortalidade infantil.

Desde 1990, houve uma substancial redução na taxa de mortalidade infantil (menores de 5 anos) em todo o mundo. A taxa global de mortalidade na infância diminuiu de 93 para 39 mortes por 1.000 nascidos vivos entre 1990 e 2017, representando uma queda de aproximadamente 58,06%. Embora o Brasil esteja entre os países que obtiveram progressos significativos na redução dessa taxa, com uma diminuição de 76,1% de 63 para 15 mortes por 1.000 nascidos

vivos no mesmo período (UNICEF, 2023), é necessário realizar mais esforços para garantir esses avanços e continuar reduzindo as disparidades na sobrevivência infantil entre as populações. É importante ressaltar que mais de 85% das mortes de crianças com menos de 5 anos ocorrem durante o primeiro ano de vida no Brasil (IPEA, 2014).

Os principais fatores responsáveis pelo progresso no Brasil estão relacionados à melhoria das condições sanitárias e sociais, mudanças demográficas, implementação do Sistema Único de Saúde (SUS), que estabeleceu um sistema de saúde universal após a Constituição Federal de 1988, e a expansão da cobertura da atenção primária, principalmente por meio da estratégia de Saúde da Família (IPEA, 2014). Além disso, outras ações estratégicas estão sendo implementadas para enfrentar o problema, como a Rede Cegonha, uma proposta do Governo Federal que visa ampliar o acesso e melhorar a qualidade da atenção pré-natal, assistência ao parto, puerpério e cuidados com crianças até 24 meses de vida, com o objetivo de reduzir a mortalidade materna e infantil (Giovanni, 2014).

Apesar do progresso alcançado no Brasil, a taxa de mortalidade infantil ainda é elevada em comparação com regiões da Europa e América do Norte (6 óbitos na infância por mil nascidos vivos) e Austrália e Nova Zelândia (4 óbitos na infância por mil nascidos vivos) em 2017, evidenciando a gravidade da situação. A monitorização da mortalidade infantil no Brasil é realizada por meio dos dados disponibilizados pelo Sistema de Informação sobre Mortalidade (SIM) e Sistema de Informação sobre Nascidos Vivos (SINASC), ambos implementados pelo Ministério da Saúde para suprir a carência de conhecimento sobre o perfil epidemiológico dos óbitos e nascimentos no país (Laurenti, Jorge & Gotlieb, 2008). No entanto, questões como sub-registro e a presença de variáveis desconhecidas ou não preenchidas ainda limitam a confiabilidade dos dados coletados e comprometem a representatividade da realidade da mortalidade infantil no Brasil (Romero & Cunha, 2006).

Nesse contexto, é fundamental buscar compreender as causas subjacentes da mortalidade infantil no Brasil, identificar as variáveis associadas e analisar seus efeitos sobre as crianças e suas famílias. Além disso, é importante promover ações efetivas para prevenir e reduzir a mortalidade infantil, baseadas em evidências científicas e melhores práticas internacionais.

Uma das principais causas de mortalidade infantil no Brasil é a prematuridade, que ocorre quando um bebê nasce antes das 37 semanas de gestação. Outras causas comuns incluem malformações congênitas, infecções, complicações durante o parto e doenças respiratórias (Careti, Scarpelini & de Carvalho Furtado, 2014). Para prevenir a mortalidade infantil, é crucial investir em cuidados pré-natais de qualidade para as gestantes, especialmente aquelas em situação de vulnerabilidade social. Isso pode envolver exames regulares, orientações sobre alimentação adequada e atividade física, além de acompanhamento por profissionais qualificados.

Também é importante promover o acesso a serviços de saúde de qualidade para o atendimento de emergência e tratamento de doenças em crianças. Isso pode incluir campanhas de vacinação, incentivo ao aleitamento materno, diagnóstico e tratamento precoce de doenças in-

fecciosas, entre outras medidas. Além disso, é fundamental implementar políticas públicas que abordem as causas subjacentes da mortalidade infantil, como a pobreza, a falta de acesso à educação e a desigualdade social. Isso pode englobar políticas de redistribuição de renda, acesso universal à educação de qualidade e ações para promover a igualdade de gênero e raça.

1.1 Uso de Aprendizado de Máquina

No contexto atual, o uso de técnicas de aprendizado de máquina tem se mostrado promissor no combate à mortalidade infantil. A aplicação de algoritmos de inteligência artificial e análise de dados pode fornecer *insights* valiosos para auxiliar na prevenção, diagnóstico e tratamento de doenças em crianças. Essas técnicas são capazes de processar grandes volumes de informações e identificar padrões complexos, possibilitando a identificação de fatores de risco e a previsão de eventos adversos.

Por exemplo, a análise de dados pode ser utilizada para identificar correlações entre variáveis demográficas, socioeconômicas e de saúde, auxiliando na identificação de grupos de maior vulnerabilidade e direcionando esforços de intervenção preventiva. Além disso, algoritmos de aprendizado de máquina podem ser empregados no desenvolvimento de sistemas de diagnóstico automatizados, que auxiliam os profissionais de saúde na detecção precoce de doenças e na adoção de medidas de tratamento adequadas.

A aplicação de inteligência artificial também pode contribuir na interpretação de exames e imagens médicas, como radiografias e tomografias, auxiliando na identificação de anomalias e condições de risco. Isso possibilita um diagnóstico mais preciso e uma intervenção terapêutica mais efetiva.

Além disso, o aprendizado de máquina pode ser empregado para analisar dados epidemiológicos e de saúde coletados em larga escala, permitindo a identificação de tendências, padrões e fatores de risco associados à mortalidade infantil. Com base nesses *insights*, políticas públicas podem ser formuladas e implementadas de forma mais direcionada, visando a redução das desigualdades e o aprimoramento dos serviços de saúde.

No entanto, é importante ressaltar que a tecnologia, por si só, não é a solução definitiva para a redução da mortalidade infantil. Ela deve ser vista como uma ferramenta complementar, que deve ser utilizada em conjunto com abordagens multidisciplinares e políticas públicas eficazes.

É fundamental que os profissionais de saúde e os responsáveis pela formulação de políticas compreendam e apliquem devidamente os resultados gerados pelo aprendizado de máquina. A interpretação correta dos dados e a tomada de decisões embasadas em evidências continuam sendo essenciais para a implementação de estratégias efetivas de prevenção e tratamento.

Além disso, é necessário considerar questões éticas e de privacidade relacionadas ao uso

de dados de saúde das crianças. A segurança e a proteção das informações pessoais devem ser priorizadas para garantir que o uso da tecnologia não comprometa a confidencialidade e a integridade dos dados.

Outro desafio é a acessibilidade e a disponibilidade igualitária da tecnologia e dos recursos relacionados em diferentes regiões do Brasil. É importante garantir que todas as comunidades, independentemente de sua localização ou condições socioeconômicas, tenham acesso aos benefícios proporcionados pela tecnologia para a redução da mortalidade infantil.

Para alcançar uma redução significativa na mortalidade infantil, é crucial combinar a utilização de tecnologia avançada com investimentos em infraestrutura de saúde, capacitação de profissionais, educação da população e políticas públicas abrangentes. Somente dessa forma será possível obter resultados duradouros e sustentáveis na melhoria da saúde e no bem-estar das crianças no Brasil.

1.2 Justificativa

Este trabalho tem como objetivo investigar a predição da mortalidade infantil com base em dados socioeconômicos e de saúde materno-infantil contido em bancos de dados públicos do Sistema Único de Saúde (SUS). A escolha desse tema se deve à sua relevância como um problema de saúde pública que afeta muitos países, especialmente os em desenvolvimento. No Brasil, apesar dos avanços alcançados, ainda existem regiões com taxas elevadas de mortalidade infantil, o que pode ser atribuído a fatores socioeconômicos, culturais e de acesso aos serviços de saúde.

A utilização de modelos de aprendizado de máquina para a predição da mortalidade infantil pode ser uma ferramenta importante para identificar fatores de risco e, assim, adotar medidas preventivas e intervenções precoces visando a redução dessa taxa.

A escolha de utilizar um conjunto limitado de atributos, dentre os disponíveis nos dados, justifica-se pela necessidade de otimizar o modelo de predição. O uso de um grande número de atributos pode levar a problemas de sobreajuste e reduzir a capacidade do modelo de generalizar para novos dados. A seleção dos atributos mais relevantes também contribui para o entendimento dos principais fatores que influenciam a mortalidade infantil, direcionando as políticas públicas para a redução desses fatores de risco específicos.

Assim, este trabalho contribui para a compreensão dos fatores associados à mortalidade infantil e para o desenvolvimento de modelos preditivos mais eficientes. A aplicação de técnicas de aprendizado de máquina tem o potencial de auxiliar na tomada de decisões em saúde pública, colaborando para a redução da taxa de mortalidade infantil.

Espera-se que os resultados obtidos neste estudo possam fornecer *insights* importantes para profissionais de saúde, pesquisadores e responsáveis pela formulação de políticas públi-

cas. Ao utilizar técnicas avançadas de análise de dados e aprendizado de máquina, espera-se identificar padrões e correlações ocultas nos dados que possam contribuir para uma compreensão mais profunda dos fatores que impactam a mortalidade infantil.

A utilização desses modelos preditivos como ferramentas complementares pode apoiar a implementação de estratégias preventivas direcionadas, intervenções efetivas e alocação de recursos de forma mais precisa. Dessa forma, espera-se que o uso de técnicas de aprendizado de máquina possa desempenhar um papel significativo na melhoria dos resultados de saúde infantil e na redução da taxa de mortalidade infantil no Brasil.

1.3 Hipótese

A hipótese deste estudo é que é possível desenvolver modelos de aprendizado de máquina capazes de prever a mortalidade infantil a partir de um conjunto de variáveis socioeconômicas, e de saúde. Essa hipótese é embasada em estudos anteriores que evidenciaram a relação entre essas variáveis e a mortalidade infantil, assim como em pesquisas que demonstraram a eficácia de algoritmos de aprendizado de máquina na previsão de resultados em saúde.

Através da seleção de um conjunto de variáveis relevantes e da aplicação de técnicas de pré-processamento de dados e modelagem de aprendizado de máquina, espera-se que os modelos desenvolvidos possam oferecer uma ferramenta útil para a identificação indivíduos com maior risco de mortalidade infantil. A hipótese é que a identificação precoce desses indivíduos de risco pode contribuir para a implementação de políticas públicas e intervenções mais direcionadas e eficazes.

Além disso, a hipótese deste estudo é que a utilização de um conjunto reduzido de variáveis pode ser suficiente para construir modelos com boa precisão na previsão da mortalidade infantil, o que representaria uma vantagem. Essa abordagem mais simplificada pode tornar a ferramenta mais acessível e fácil de implementar em contextos com recursos limitados

Assim, a validação da hipótese proposta neste estudo pode contribuir para o avanço do conhecimento científico na área da mortalidade infantil e fornecer subsídios para aprimorar as estratégias de prevenção e redução desse problema de saúde pública. A confirmação da hipótese poderá fortalecer a utilização de modelos de aprendizado de máquina como uma ferramenta complementar e efetiva para auxiliar na tomada de decisões em saúde pública, contribuindo para a redução da taxa de mortalidade infantil.

1.4 Objetivos

Este trabalho tem como objetivo principal desenvolver modelos de classificação capazes de prever a mortalidade infantil com alta acurácia, utilizando apenas um conjunto reduzido de variáveis.

- Coletar e pré-processar os dados relacionados à mortalidade infantil;
- Selecionar um conjunto reduzido de variáveis relevantes para o problema, utilizando técnicas de seleção de atributos;
- Explorar diferentes técnicas de modelagem de dados, como algoritmos de aprendizado de máquina e técnicas estatísticas;
- Avaliar o desempenho dos modelos propostos utilizando métricas adequadas;
- Comparar o desempenho dos modelos desenvolvidos com estudos anteriores que utilizaram um conjunto maior de variáveis, a fim de avaliar a eficiência da abordagem com um conjunto reduzido de variáveis;
- Discutir as implicações práticas dos modelos propostos na previsão da mortalidade infantil e sugerir possíveis aplicações em políticas públicas e práticas clínicas;
- Identificar possíveis limitações do estudo e fornecer recomendações para pesquisas futuras nessa área.

Ao atingir esses objetivos, espera-se que este trabalho contribua para o avanço do conhecimento na área da previsão da mortalidade infantil, fornecendo *insights* valiosos na construção de modelos eficientes. Além disso, os resultados obtidos podem ser utilizados para embasar a tomada de decisões em saúde pública, contribuindo para a redução da taxa de mortalidade infantil e o desenvolvimento de estratégias mais direcionadas e efetivas.

Capítulo 2

Fundamentação Teórica

2.1 Morte Infantil

Desde 1990, a mortalidade infantil, para crianças menores de 5 anos caiu substancialmente no mundo. No entanto, são necessários esforços para assegurar tais progressos e continuar reduzindo a disparidade na sobrevivência na infância entre as populações (Chao, You, Pedersen, Hug & Alkema, 2018). Conforme UN IGME (*United Nations Inter-agency Group for Child Mortality Estimation*), globalmente, a taxa de mortalidade na infância reduziu de 93 para 39 mortes por 1.000 nascidos vivos, entre 1990 e 2017, respectivamente, um declínio em torno de 58,06%.

O Brasil é uma das nações que apresentaram avanços significativos na redução da taxa de mortalidade infantil. Entre 1990 e 2017, a taxa de mortalidade na infância no país caiu de 63 para 15 mortes por 1.000 nascidos vivos, o que corresponde a uma redução de 76,19%, segundo o UN IGME. É possível observar uma queda mais acentuada no Brasil em comparação com a média mundial. O Instituto de Pesquisa Econômica Aplicada (IPEA) destaca que mais de 85% das mortes de crianças com menos de 5 anos no país ocorrem no primeiro ano de vida. (do Milênio, 2014).

O Governo Brasileiro tem propostas para assistência materno-infantil, uma delas é a Rede Cegonha, que tem como objetivo melhorar a assistência materno-infantil no país, garantindo um acompanhamento integral às mulheres e crianças até 24 meses de vida. A iniciativa prevê uma série de ações e propostas para a promoção da saúde materna e infantil, incluindo o pré-natal adequado, o parto seguro, o cuidado com o recém-nascido e a atenção à saúde da mulher no pós-parto. Entre as ações e propostas da Rede Cegonha estão a ampliação da oferta de serviços obstétricos e neonatais de qualidade, a implementação de centros de parto normal, a criação de casas de gestante, parto e puerpério, o incentivo ao aleitamento materno, o fortalecimento do planejamento reprodutivo e a capacitação dos profissionais de saúde para a atenção humanizada à gestação, parto e puerpério (de Vaconcelos & Guerrero, 2013).

O problema de dados ignorados e sub-registros é um desafio significativo na coleta e análise de dados em diversas áreas. Esse problema ocorre quando algumas informações relevantes

não são registradas ou são omitidas durante a coleta de dados. Isso pode resultar em uma análise incompleta e distorcida da realidade. Os registros são particularmente importantes para monitorar e avaliar a eficácia dos sistemas de saúde e para identificar problemas de saúde pública. Quando ocorre sub-registro, por exemplo, pode haver uma subestimação do número de casos de uma doença, o que pode levar a uma falta de recursos e esforços para prevenir e tratar a doença. É importante que os responsáveis pela coleta e análise de dados estejam cientes desse problema e tomem medidas para minimizá-lo. Isso pode incluir treinamento de pessoal para a coleta adequada de dados, uso de sistemas de registro eletrônico e monitoramento regular da qualidade dos dados coletados. O uso de técnicas estatísticas também pode ajudar a avaliar o impacto dos dados ignorados ou sub-registros na análise e fornecer estimativas mais precisas (Agranonik & Jung, 2019).

2.2 Sistemas de Informações em Saúde

Normalmente a informação é entendida como o significado que se atribui ao dado elaborado por meio de convenções e representações (Adriaans, Van Benthem et al., 2008; of Electrical & Engineers, n.d.). Em geral, no contexto de decisões de assistência à saúde ou pesquisa, a informação é importante objeto para minimizar as incertezas. O conceito de informação em saúde abrange diferentes áreas de conhecimento, em especial a demografia, a epidemiologia e a economia. A capacidade operacional e de produção dos serviços de saúde dependem fortemente da informação qualificada (da Silva Leandro, Rezende & da Conceição Pinto, 2020)

O DATASUS é o Departamento de Informática do Sistema Único de Saúde (SUS) do Brasil, criado em 1991 com o objetivo de desenvolver e gerenciar as informações em saúde. Ele é responsável por coletar, processar, armazenar e disseminar informações de saúde de todo o país, disponibilizando dados e indicadores para apoiar a gestão e o planejamento em saúde.

O Sistema de Informações sobre Mortalidade (SIM) é um dos sistemas gerenciados pelo DATASUS e tem como objetivo coletar, processar e divulgar informações sobre mortalidade no país, incluindo a causa do óbito, a idade e o sexo do falecido, bem como a região e o município onde ocorreu o óbito. O SIM foi criado em 1975 e é uma das principais fontes de dados para o monitoramento da saúde da população.

Outro sistema gerenciado pelo DATASUS, o Sistema de Informações sobre Nascido Vivo (SINASC), tem como objetivo coletar informações sobre os nascidos vivos no país, incluindo a idade e a escolaridade da mãe, o tipo de parto, o peso do recém-nascido e a região e o município onde ocorreu o nascimento. O SINASC foi criado em 1990 e é uma importante fonte de dados para o planejamento e a gestão em saúde materna e infantil.

O Sistema de Informação de Agravos de Notificação (SINAN) é responsável por coletar, armazenar e transmitir informações sobre doenças e agravos de notificação compulsória no país,

incluindo a dengue, a tuberculose, a meningite, a hepatite, a AIDS, entre outras. O SINAN foi criado em 1990 e é utilizado para monitorar a ocorrência de doenças e orientar as ações de vigilância em saúde.

O Sistema de Informação do Programa Nacional de Imunizações (SI-PNI) tem como objetivo coletar e registrar informações sobre as vacinas aplicadas no país, bem como as doses disponíveis e as campanhas de vacinação. O SI-PNI foi criado em 1973 e é utilizado para o controle e a vigilância da imunização da população brasileira.

O Sistema do Programa Nacional de Controle da Dengue (SISPNCD) é responsável por coletar e armazenar informações sobre a ocorrência da dengue no país, incluindo o número de casos suspeitos, confirmados e óbitos, bem como a localização dos focos do mosquito transmissor. O SISPNCD foi criado em 2002 e é utilizado para orientar as ações de combate e prevenção da doença.

O Sistema de Informação de Câncer (SISCAN) tem como objetivo coletar e armazenar informações sobre o diagnóstico e o tratamento do câncer no país, incluindo o tipo de câncer, o estágio da doença, o tipo de tratamento realizado e a evolução do paciente. O SISCAN foi criado em 1998 e é utilizado para monitorar a ocorrência do câncer no país e orientar as políticas de prevenção e controle da doença. (da Silva Leandro et al., 2020)

Normalmente a informação é entendida como o significado que se atribui ao dado elaborado por meio de convenções e representações (Adriaans et al., 2008). Em geral, no contexto de decisões de assistência à saúde ou pesquisa, a informação é importante objeto para minimizar as incertezas. O conceito de informação em saúde abrange diferentes áreas de conhecimento, em especial a demografia, a epidemiologia, e a economia. A capacidade operacional e de produção dos serviços de saúde depende fortemente da informação qualificada (da Silva Leandro et al., 2020).

A construção do SIM envolve várias etapas e processos, desde a notificação inicial até a disponibilização dos dados para análise e tomada de decisões. Vamos explorar essas etapas:

- **Notificação do óbito**: Quando ocorre um óbito, a informação é notificada às autoridades de saúde competentes. Os hospitais, unidades de saúde, serviços funerários e cartórios têm a responsabilidade de registrar e notificar os óbitos de acordo com os procedimentos estabelecidos.
- **Coleta de dados**: Após a notificação, os dados sobre o óbito são coletados. Isso inclui informações básicas, como data, hora e local do óbito, bem como dados sobre a pessoa falecida, como idade, sexo, causa básica e causas associadas de morte, além de informações sobre a ocorrência de doenças preexistentes ou fatores de risco relevantes.
- **Verificação e certificação do óbito**: É realizada uma verificação do óbito para garantir sua autenticidade e precisão. Um médico ou profissional de saúde autorizado é responsável por certificar a causa da morte com base nas informações disponíveis, seguindo as

orientações e critérios estabelecidos pela Classificação Internacional de Doenças (CID).

- **Codificação dos dados**: Após a certificação, os dados sobre a causa do óbito são codificados de acordo com a CID ou outra classificação padronizada. Isso permite a análise e o agrupamento dos óbitos de acordo com categorias específicas, facilitando a identificação de padrões e a comparação entre diferentes regiões e períodos de tempo.
- **Armazenamento e segurança dos dados**: Os dados do SIM são armazenados em um banco de dados seguro, com medidas adequadas de segurança e privacidade. A confidencialidade das informações pessoais dos falecidos é preservada, garantindo a conformidade com as regulamentações e leis de proteção de dados.
- **Disponibilização dos dados**: Os dados do SIM são disponibilizados para uso público, sendo acessíveis a pesquisadores, profissionais de saúde, gestores e outros interessados. Essas informações são utilizadas para monitorar as condições de saúde, avaliar a eficácia de políticas e programas de saúde, investigar surtos e epidemias, e contribuir para a tomada de decisões informadas em saúde pública.

A construção do SIM requer a colaboração e cooperação entre diversas entidades e profissionais, incluindo hospitais, unidades de saúde, serviços funerários, cartórios, órgãos de saúde pública e demais envolvidos na notificação e registro de óbitos. A precisão e a qualidade dos dados são fundamentais para a análise e interpretação adequadas, garantindo a confiabilidade das informações.

Capítulo 3

Inteligência Artificial

3.1 Considerações Iniciais

3.2 Mineração de Dados

O modelo tradicional de transformação de dados em informação envolve um processo manual em que especialistas processam os elementos e produzem relatórios para análise. No entanto, com o crescimento exponencial do volume de dados armazenados, acabamos nos deparando com uma situação em que somos ricos em dados, mas pobres em informação. A descoberta manual de padrões nesses repositórios se tornou inviável, tornando-se essencial a capacidade de explorá-los de maneira mais eficiente. Nesse contexto, a mineração de dados tem ganhado cada vez mais relevância, oferecendo técnicas que apoiam essa tarefa (Fayyad, Piatetsky-Shapiro & Smyth, 1996; Tan, Steinbach & Kumar, 2016).

A mineração de dados combina conhecimentos de diversas áreas, aproveitando métodos estatísticos tradicionais de análise de dados, bem como técnicas e algoritmos de inteligência artificial, aprendizado de máquina e reconhecimento de padrões. Seu objetivo principal é descobrir propriedades e relações úteis nos dados. O processo geral de transformação de dados brutos em informações úteis é conhecido como Descoberta de Conhecimento em Bases de Dados (*Knowledge-Discovery in Databases* - KDD), sendo a mineração de dados uma parte integrante desse processo. De acordo com Fayyad (Fayyad et al., 1996), o KDD é definido como um processo não trivial de identificação de novos padrões válidos, úteis e compreensíveis.

A Figura 3.1 ilustra uma representação visual do processo de KDD, mostrando as etapas envolvidas em sua execução. Inicialmente, são selecionados os conjuntos de dados relevantes para análise. Em seguida, ocorre o pré-processamento dos dados, onde são realizadas tarefas como limpeza, normalização e tratamento de valores ausentes. Após o pré-processamento, os dados são submetidos à mineração de dados, que utiliza algoritmos e técnicas para descobrir padrões, associações, tendências ou classificações relevantes. Por fim, na etapa de pós-processamento, os resultados da mineração são interpretados, avaliados e visualizados, permitindo a extração de conhecimento valioso para a tomada de decisões.

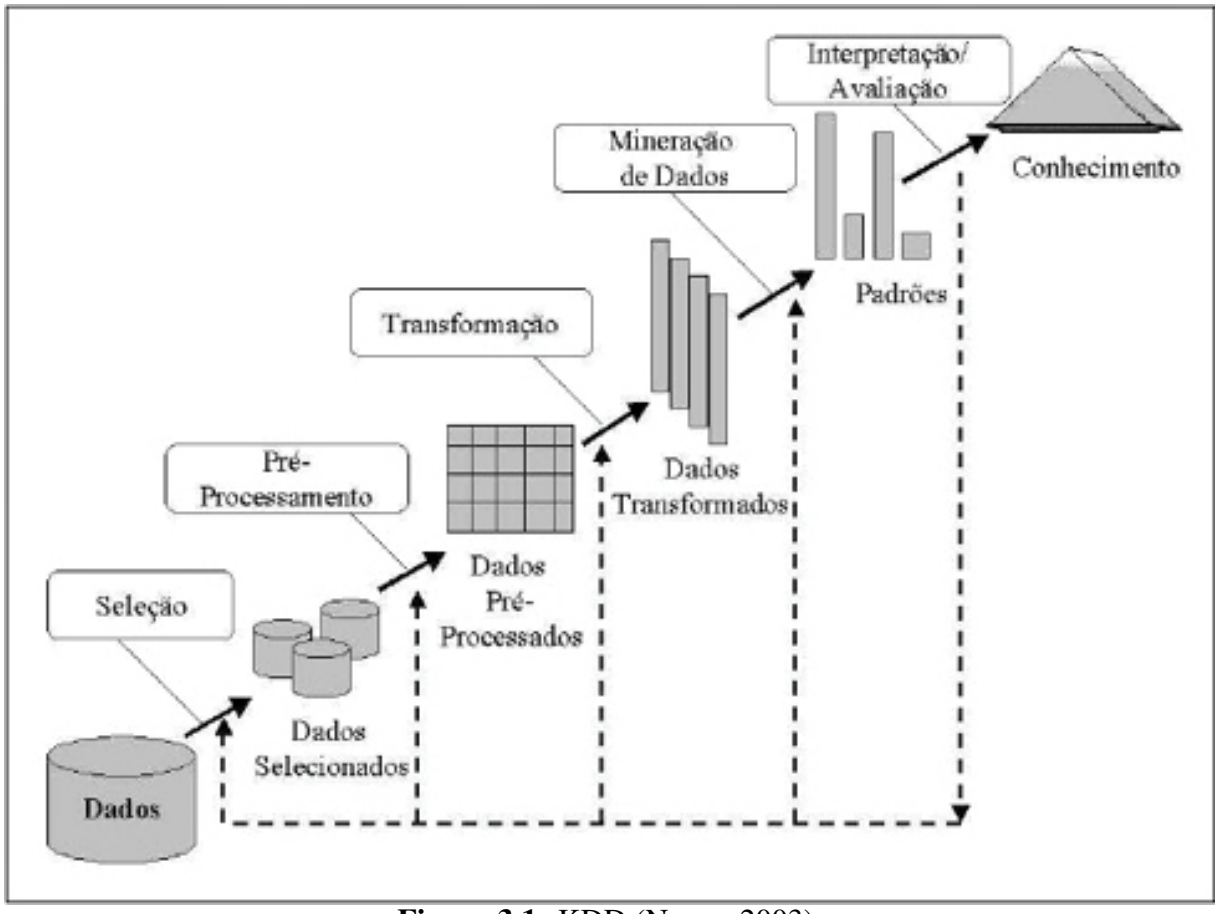

Figura 3.1: KDD (Neves, 2003).

3.2.1 Seleção de Dados

A seleção de dados desempenha um papel crucial na análise, especialmente quando se trata de gerar conhecimento para abordar problemas específicos. Essa etapa envolve uma cuidadosa identificação das bases de dados disponíveis e a determinação das informações que são relevantes para atender aos objetivos da análise.

Ao selecionar as bases de dados adequadas, é essencial considerar a qualidade e a confiabilidade dos dados. Isso inclui avaliar a fonte dos dados, a precisão das informações registradas e a abrangência dos registros. Além disso, é importante analisar a relevância dos dados para o problema em questão, identificando as variáveis-chave que são essenciais para obter *insights* significativos.

Uma abordagem estratégica na seleção de dados também envolve a consideração de diferentes tipos de dados. Isso pode incluir dados estruturados, como tabelas de banco de dados, e não estruturados, como textos, imagens ou áudio. Dependendo do problema em análise, diferentes fontes de dados podem ser necessárias para obter uma visão mais abrangente e precisa.

A seleção de dados deve levar em consideração questões éticas e legais relacionadas à privacidade e à proteção dos dados. É essencial garantir que a obtenção e o uso estejam em

conformidade com as regulamentações aplicáveis, como a Lei Geral de Proteção de Dados (LGPD). Isso inclui obter o consentimento adequado dos indivíduos envolvidos e garantir o anonimato dos dados quando necessário.

Uma vez que as bases de dados relevantes tenham sido selecionadas, é possível prosseguir com a coleta e a integração dos dados, preparando-os para análise. Isso pode envolver o desenvolvimento de procedimentos de extração, transformação e carga (ETL) para consolidar os dados de diferentes fontes em um único formato adequado para análise (Fayyad et al., 1996).

3.2.2 Pré-processamento

A etapa de pré-processamento desempenha um papel fundamental no processo de KDD, pois a qualidade dos dados tem um impacto significativo na eficácia dos algoritmos de mineração. Nessa fase, são aplicadas técnicas e procedimentos para garantir que os dados estejam limpos, coerentes e prontos para análise, visando obter resultados confiáveis e relevantes.

Uma das principais tarefas durante o pré-processamento é a eliminação de dados redundantes e inconsistentes. Isso envolve identificar e remover registros duplicados, atributos irrelevantes ou repetitivos que não contribuem para a análise. Além disso, é importante lidar com dados inconsistentes que possam estar presentes devido a erros de entrada, falhas de comunicação ou problemas de integração de diferentes fontes de dados. Essa etapa de limpeza dos dados é essencial para evitar distorções nos resultados e garantir a confiabilidade das análises.

Outro aspecto do pré-processamento de dados é a identificação e tratamento de *outliers*, que são valores atípicos ou discrepantes que podem distorcer as análises estatísticas. Esses valores podem surgir devido a erros de medição, falhas de coleta de dados ou eventos raros e incomuns. É importante avaliar cuidadosamente esses *outliers* e decidir se eles devem ser removidos, corrigidos ou tratados de modo especial durante a análise. O objetivo é garantir que esses valores discrepantes não comprometam a interpretação dos resultados e não distorçam as conclusões do processo de mineração.

O pré-processamento também envolve lidar com dados faltantes, que são valores ausentes ou incompletos em algumas observações. Esses dados podem ocorrer devido a vários motivos, como erros de coleta, informações não disponíveis ou registros incompletos. É necessário aplicar técnicas adequadas para lidar com esses dados faltantes, como a imputação de valores estimados ou a exclusão seletiva de observações com dados ausentes. O objetivo é minimizar o impacto dos dados faltantes na análise e garantir que as informações sejam as mais completas possível.

3.2.3 Transformação dos Dados

Após a etapa de pré-processamento dos dados, é comum realizar a padronização dos mesmos como parte do processo de preparação para a análise. A padronização consiste em transformar os dados de forma a facilitar a manipulação e comparação entre eles pelos algoritmos de análise. Essa técnica é especialmente útil quando os dados estão em diferentes escalas ou unidades de medida, o que pode dificultar a correta interpretação e comparação dos resultados obtidos.

A padronização dos dados não apenas facilita a manipulação e comparação dos mesmos pelos algoritmos de análise, mas também pode contribuir para melhorar a estabilidade e a performance desses algoritmos. Além disso, a padronização permite que variáveis com diferentes unidades de medida sejam comparadas de modo mais direto e significativo, evitando que uma variável com valores maiores domine a análise em detrimento das outras ().

3.2.4 Processamento

Embora todas as etapas do processo de KDD sejam importantes, a etapa de processamento ou análise recebe destaque significativo neste trabalho. É nessa fase que o conjunto de dados é submetido a técnicas de mineração de dados, também conhecidas como *data mining*.

Não existe uma técnica única que seja ideal para todos os problemas de mineração de dados. Cada método possui suas vantagens e desvantagens, e a escolha do método mais adequado depende das características específicas do problema e dos objetivos da análise. Dentre as técnicas comumente utilizadas, destacam-se as Redes Neurais Artificiais, que se inspiram no funcionamento do cérebro humano para realizar análises complexas.

Outra técnica amplamente empregada é a utilização de Árvores de Decisão, que são estruturas de decisão hierárquicas que permitem a representação visual e interpretação dos resultados. Além disso, os algoritmos de classificação baseados em exemplos têm sido bastante aplicados na identificação de padrões e na categorização de dados.

Já para classificação de dados complexos e na separação de padrões em espaços de alta dimensionalidade, uma técnica que se destaca por sua eficácia é a Máquina de Vetor de Suporte (*Support Vector Machine* - SVM).

Essas são apenas algumas das diversas técnicas disponíveis no campo da mineração de dados, e a escolha adequada depende da natureza dos dados, dos objetivos da análise e das necessidades específicas do problema em questão.

3.2.5 Pós-processamento

O pós-processamento, sendo uma das últimas fase do processo de KDD, desempenha um papel essencial na análise e avaliação dos resultados obtidos a partir da aplicação de métodos de análise sobre o conjunto de dados. Nessa etapa, os resultados gerados pela mineração de dados são cuidadosamente examinados, interpretados e validados para garantir sua relevância e confiabilidade.

Durante o pós-processamento, é comum utilizar técnicas de interpretação e avaliação dos modelos gerados. Essas técnicas permitem compreender melhor os padrões descobertos e a relação entre os atributos, além de fornecer *insights* valiosos para tomada de decisões. É nessa fase que se busca a compreensão dos resultados obtidos, sua interpretação em termos de conhecimento relevante e sua aplicação prática em um contexto específico.

Ao analisar os resultados do KDD, é importante considerar a validade e a eficácia dos modelos gerados. Pode-se avaliar a precisão dos modelos usando métricas como taxa de acerto, sensibilidade, especificidade e outras medidas relevantes para o problema em questão. Além disso, é importante considerar a interpretabilidade dos modelos, ou seja, a capacidade de entender e explicar as relações encontradas entre os dados.

Após a apresentação detalhada de todas as fases do processo de KDD, é relevante explorar algumas técnicas de seleção de atributos na próxima seção. Essas técnicas fazem parte da etapa de pré-processamento e têm como objetivo identificar os atributos mais relevantes e informativos para a análise, reduzindo a dimensionalidade do conjunto de dados. A seleção adequada de atributos pode melhorar a eficiência e a eficácia dos algoritmos de mineração de dados, facilitando a descoberta de padrões e o desenvolvimento de modelos mais precisos e compreensíveis.

3.3 Seleção de Atributos

A seleção de atributos é uma etapa crucial no processo de modelagem e análise de dados, especialmente em problemas de aprendizado de máquina. Consiste em identificar e escolher os atributos mais relevantes e informativos para a tarefa de predição ou classificação, descartando os atributos menos importantes. Essa técnica desempenha um papel fundamental na redução da dimensionalidade dos dados, eliminando atributos desnecessários ou redundantes, e pode levar a modelos mais eficientes e interpretações mais claras.

O objetivo principal da seleção de atributos é encontrar um subconjunto ótimo de atributos que melhor represente a relação entre as variáveis preditoras e a variável de resposta. Ao selecionar atributos relevantes, buscamos maximizar a capacidade preditiva do modelo e minimizar o impacto de atributos irrelevantes, ruidosos ou colineares. Existem várias abordagens e

técnicas disponíveis para realizar a seleção de atributos, que podem ser divididas em métodos baseados em filtro, *wrapper* e incorporação.

Os métodos baseados em filtro utilizam medidas estatísticas e métricas de importância para avaliar a relação entre cada atributo e o desfecho. Essas medidas podem ser aplicadas individualmente a cada atributo, independentemente do algoritmo de aprendizado de máquina utilizado posteriormente. Dentre os métodos de filtro mais comumente empregados, destacam-se o teste Qui-Quadrado (χ^2) e o teste *t-Student*.

O teste Qui-Quadrado é amplamente utilizado em problemas de classificação, quando tanto a variável preditora quanto a variável de resposta são categóricas. Ele mede a dependência estatística entre os atributos e a variável de resposta, avaliando se as distribuições observadas diferem significativamente das distribuições esperadas. A fórmula para calcular o valor de Qui-Quadrado é dada por (Agresti, 2012):

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \tag{3.1}$$

onde O_{ij} representa as frequências observadas e E_{ij} representa as frequências esperadas para cada combinação de categorias do atributo e da variável de resposta. O valor de Qui-Quadrado é então comparado com uma distribuição χ^2 para determinar o p-valor associado. Quanto menor o p-valor, mais evidências temos para rejeitar a hipótese nula de independência entre o atributo e o desfecho.

Já o teste *t-Student* é utilizado em problemas de regressão ou classificação quando a variável preditora é numérica e a variável de resposta é categórica. Ele mede se as médias das variáveis preditoras são estatisticamente diferentes entre as categorias do desfecho. A fórmula para calcular o valor de t é dada por:

$$t = \frac{\mu_1 - \mu_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \tag{3.2}$$

onde μ_1 e μ_2 representam as médias das variáveis preditoras nas diferentes categorias do desfecho, s_1 e s_2 representam os desvios-padrões e n_1 e n_2 representam os tamanhos das amostras em cada categoria. O valor de t é comparado com uma distribuição t de *Student* com $n_1 + n_2 - 2$ graus de liberdade para determinar o p-valor associado. Um valor de p baixo indica uma diferença estatisticamente significativa entre as médias das variáveis preditoras em diferentes categorias do desfecho.

A determinação do p-valor em testes estatísticos, como o Qui-Quadrado e o *t-Student*, desempenha um papel essencial na interpretação dos resultados e na tomada de decisões estatísticas. Para calcular o p-valor, utilizamos a função de distribuição acumulada (CDF) correspondente à distribuição estatística específica (χ^2 ou *t-Student*), ajustada aos graus de liberdade

apropriados. O p-valor é então calculado subtraindo-se o valor da CDF obtido do número 1. Quanto menor o p-valor, maior a evidência contra a hipótese nula, indicando uma associação mais forte entre os atributos e o desfecho em questão.

No contexto do teste Qui-Quadrado e do teste *t-Student*, o p-valor desempenha um papel fundamental na interpretação dos resultados. Um p-valor inferior a um determinado nível de significância (normalmente estabelecido como 0,05) indica que a associação entre os atributos e o desfecho é estatisticamente significativa, permitindo a rejeição da hipótese nula de independência. Por outro lado, um p-valor acima do nível de significância indica que não há evidências suficientes para rejeitar a hipótese nula, sugerindo que a associação entre os atributos e o desfecho não é estatisticamente significativa.

Ao utilizar adequadamente os testes Qui-Quadrado e *t-Student*, juntamente com a interpretação dos respectivos p-valores, podemos realizar uma seleção de atributos mais robusta e embasada.

Para o teste Qui-Quadrado:

$$p{-}valor = 1 - CDF_{\chi^2}(\chi^2, \text{graus de liberdade}) \tag{3.3}$$

Para o teste *t-Student*:

$$p{-}valor = 1 - CDF_t(t, \text{graus de liberdade}) \tag{3.4}$$

Onde CDF_{χ^2} e CDF_t representam as funções de distribuição acumulada para as distribuições χ^2 e *t-Student*, respectivamente. Os valores χ^2 e t são obtidos nos cálculos das estatísticas de teste e os "graus de liberdade"são parâmetros específicos para cada teste, dependendo do número de categorias ou do tamanho das amostras.

Utilizando essas fórmulas, é possível calcular os p-valores correspondentes a cada teste estatístico, permitindo a avaliação da significância estatística da associação entre os atributos e o desfecho analisado.

3.4 *Support Vector Machine* (SVM)

3.4.1 Introdução ao SVM

O *Support Vector Machine* (SVM), ou Máquina de Vetores de Suporte, é uma técnica de aprendizado de máquina amplamente utilizada para tarefas de classificação e regressão. Ela é baseada em princípios da teoria de aprendizado estatístico e se destaca por sua eficácia na

separação de classes em conjuntos de dados complexos.

O SVM tem sido amplamente aplicado em diversos domínios, como reconhecimento de padrões, processamento de imagens, bioinformática e finanças, devido à sua capacidade de lidar com problemas de alta dimensionalidade e não-lineares. Essa técnica também é conhecida por sua robustez em relação a ruídos e sua capacidade de generalização, ou seja, sua habilidade de classificar corretamente novos exemplos que não foram utilizados no treinamento.

A ideia principal por trás do SVM é encontrar o hiperplano de separação ótimo que maximize a margem entre as classes. O hiperplano é uma superfície de decisão que separa as classes, e a margem é a distância entre o hiperplano e os exemplos mais próximos de cada classe, chamados de vetores de suporte. Ao encontrar o hiperplano que maximiza a margem, o SVM busca encontrar a melhor fronteira de decisão entre as classes, resultando em um bom desempenho de classificação.

3.4.2 Formulação Matemática do SVM

A formulação matemática do SVM é fundamental para entender como essa técnica de aprendizado de máquina funciona. O objetivo do SVM é encontrar um hiperplano de separação ótimo que maximize a margem entre as classes de um conjunto de dados.

Suponha que tenhamos um conjunto de treinamento com exemplos rotulados, onde cada exemplo é representado por um vetor de características x e associado a uma classe y. O objetivo do SVM é encontrar um hiperplano de separação que maximize a margem entre os vetores de suporte, que são os exemplos mais próximos do hiperplano.

Matematicamente, o hiperplano de separação é definido como uma função linear da forma:

$$w^2Tx + b = 0 \tag{3.5}$$

onde w é um vetor de pesos que define a direção e a orientação do hiperplano, b é o termo de viés (ou interceptação) e x é o vetor de características.

Agora, considerando que temos dois hiperplanos paralelos que separam as classes positiva e negativa, o SVM busca encontrar o hiperplano ótimo que maximiza a margem entre esses dois hiperplanos. Essa margem é definida como a distância entre os dois hiperplanos paralelos e é denotada por $2d$.

O objetivo é encontrar os vetores de suporte que estão mais próximos desses hiperplanos, pois eles têm influência direta na determinação do hiperplano ótimo. Esses vetores de suporte são aqueles que estão na margem ou que estão incorretamente classificados.

Para encontrar o hiperplano ótimo, pode-se formular o problema como um problema de otimização convexa, buscando minimizar a função de custo sujeita a restrições. Essa função de custo pode ser formulada de diferentes maneiras, sendo a forma mais comum a função de minimização do erro hinge:

onde $||w||^2$ é a norma euclidiana ao quadrado do vetor de pesos w, C é um parâmetro de regularização que controla o equilíbrio entre a margem e a classificação correta dos exemplos, y_i é o rótulo do exemplo i e x_i é o vetor de características do exemplo i.

Essa formulação do SVM é conhecida como SVM de margem rígida, pois não permite erros de classificação. No entanto, existe também o SVM de margem suave, que permite erros de classificação, mas penaliza-os na função de custo.

A formulação matemática do SVM permite resolver o problema de encontrar o hiperplano ótimo de separação através de métodos de otimização convexa, como o algoritmo *Sequential Minimal Optimization* (SMO) e outros algoritmos eficientes.

3.4.3 Maximização da Margem

A maximização da margem é um conceito fundamental no SVM e está relacionada à busca pelo hiperplano de separação ótimo que maximiza a distância entre os vetores de suporte das classes.

Como já mencionado, a ideia principal do SVM é encontrar um hiperplano que seja capaz de separar os exemplos de diferentes classes de forma mais eficaz. Para isso, busca-se um hiperplano que possua a maior margem possível entre as classes, ou seja, a maior distância possível entre o hiperplano e os vetores de suporte.

A margem é definida como a distância entre o hiperplano de separação e os exemplos mais próximos de cada classe, que são chamados de vetores de suporte. Esses vetores de suporte são aqueles que ficam na fronteira entre as classes e têm maior influência na determinação do hiperplano ótimo.

A maximização da margem é importante porque, ao escolher um hiperplano que possui uma margem maior, aumenta-se a capacidade de generalização do modelo. Isso significa que o modelo terá um melhor desempenho na classificação de novos exemplos que não foram utilizados no treinamento.

Para alcançar a maximização da margem, o SVM utiliza um processo de otimização. O objetivo é encontrar os parâmetros do hiperplano (os pesos w e o termo de viés b) que minimizam a função de custo, ao mesmo tempo em que mantêm a margem o mais ampla possível.

A função de custo utilizada no SVM é baseada no princípio de minimização do erro hinge, que penaliza a classificação incorreta dos exemplos que estão dentro da margem ou que estão

do lado errado do hiperplano. Essa função de custo busca equilibrar a maximização da margem com a necessidade de classificar corretamente os exemplos.

Ao resolver o problema de otimização, o SVM encontra os parâmetros do hiperplano que maximizam a margem, resultando em uma melhor capacidade de separação das classes. O SVM de margem rígida busca um hiperplano que não permite erros de classificação, enquanto o SVM de margem suave permite erros dentro da margem, mas penaliza-os na função de custo.

3.4.4 Seleção do Hiperplano

O hiperplano é uma superfície de decisão que divide o espaço de características em regiões correspondentes a cada classe. Para realizar essa separação, o SVM procura um hiperplano que esteja o mais distante possível dos exemplos de treinamento mais próximos, conhecidos como vetores de suporte.

Existem diferentes abordagens para selecionar o hiperplano ótimo, dependendo do tipo de problema e das características dos dados. As duas principais estratégias são:

- **Hiperplano de margem rígida:** Nessa abordagem, busca-se um hiperplano que separe as classes de forma exata, sem permitir erros de classificação. Isso significa que todos os exemplos de treinamento devem estar corretamente classificados e fora da margem de separação. Essa técnica é adequada quando os dados são linearmente separáveis e não há ruídos nos dados.
- **Hiperplano de margem suave:** Essa abordagem permite erros de classificação dentro de uma margem específica. Ela é mais flexível e adequada para dados que não são linearmente separáveis ou que possuem ruídos. O objetivo é encontrar um hiperplano que maximize a margem, mas também minimize a quantidade de erros de classificação. Essa abordagem utiliza um parâmetro de regularização C para controlar o equilíbrio entre a margem e os erros.

Para encontrar o hiperplano ótimo, o SVM utiliza técnicas de otimização. O problema de seleção do hiperplano pode ser formulado como um problema de otimização convexa, buscando minimizar uma função de custo sujeita a restrições. A função de custo geralmente envolve o equilíbrio entre a maximização da margem e a minimização dos erros de classificação.

3.4.5 *Kernel Trick*

O *Kernel Trick* é uma técnica fundamental no SVM que permite lidar com problemas de classificação não lineares, transformando-os em problemas lineares em um espaço de maior

dimensionalidade. Essa técnica é aplicada ao introduzir funções de kernel, que são responsáveis por mapear os dados para um espaço de características de maior dimensionalidade.

A ideia por trás do *Kernel Trick* é permitir que o SVM trabalhe em espaços de características mais complexos sem a necessidade de calcular explicitamente as transformações de alta dimensionalidade. Em vez disso, o truque do kernel permite que o SVM realize cálculos eficientes diretamente no espaço original de baixa dimensionalidade.

Os kernels são funções que medem a similaridade entre pares de exemplos de treinamento em um espaço de características transformado. Essas funções são escolhidas de forma a corresponder a uma determinada transformação de características. Os kernels mais comuns são o linear, polinomial e o kernel de função de base radial (*Radial-basis Function* - RBF).

Ao usar o *Kernel Trick*, o SVM encontra um hiperplano ótimo no espaço de características transformado, que é equivalente a encontrar um hiperplano no espaço original de dados. Isso permite que o SVM lide com problemas de classificação não lineares de forma eficiente, sem a necessidade de mapear explicitamente os dados para um espaço de alta dimensionalidade.

Além disso, o *Kernel Trick* evita o problema da maldição da dimensionalidade, que ocorre quando o número de características aumenta significativamente e pode levar a problemas de desempenho e sobreajuste. Ao trabalhar no espaço de características transformado, o SVM pode encontrar separações complexas mesmo em problemas de alta dimensionalidade.

A escolha adequada do *kernel* é crucial para obter bons resultados no SVM. Cada tipo de *kernel* tem suas próprias características e propriedades, e a escolha depende do tipo de problema e da natureza dos dados. Por exemplo, o *kernel* linear é adequado para problemas linearmente separáveis, enquanto o de RBF é mais flexível e pode lidar com problemas não lineares.

O *Kernel Trick* é uma das principais vantagens do SVM em relação a outras técnicas de classificação. Ele permite que o SVM seja aplicado em uma ampla gama de problemas, fornecendo maior flexibilidade e capacidade de generalização. Com o uso adequado do *Kernel Trick*, o SVM pode alcançar resultados de classificação de alta precisão mesmo em problemas complexos e não lineares.

3.4.6 Seleção de Parâmetros no SVM

A escolha correta dos parâmetros é essencial, pois eles afetam diretamente a capacidade de generalização e o desempenho do SVM. Uma seleção inadequada de parâmetros pode levar a um modelo superajustado (*overfitting*) ou subajustado (*underfitting*), resultando em uma baixa precisão de classificação.

A seleção de parâmetros no SVM é um processo iterativo que requer experimentação e análise cuidadosa dos resultados. É importante considerar o contexto do problema, a natureza

dos dados e a capacidade computacional disponível ao selecionar os parâmetros. Uma escolha adequada dos parâmetros pode levar a um modelo de SVM com desempenho otimizado e maior precisão de classificação.

3.5 *K-Nearest Neighbor* (KNN)

O *K-Nearest Neighbor* (KNN) é um algoritmo de aprendizado de máquina supervisionado amplamente utilizado para problemas de classificação e regressão. Ele se baseia no princípio de que objetos semelhantes tendem a estar próximos uns dos outros no espaço de características.

O funcionamento do KNN se baseia em dado um conjunto de dados de treinamento com rótulos conhecidos o algoritmo armazena essas informações para uso posterior. Quando um novo exemplo de teste é apresentado, o KNN calcula a distância entre esse exemplo e todos os exemplos de treinamento. Em seguida, seleciona os K exemplos de treinamento mais próximos com base na distância.

Para problemas de classificação, o KNN atribui ao exemplo de teste a classe que é mais frequente entre os K vizinhos mais próximos. Para problemas de regressão, o KNN calcula uma média ou uma média ponderada dos valores dos K vizinhos mais próximos e atribui esse valor ao exemplo de teste.

O parâmetro K no KNN define a quantidade de vizinhos mais próximos a serem considerados na classificação ou regressão. A escolha de um valor adequado para K é importante, pois um valor muito baixo pode resultar em instabilidade e sensibilidade ao ruído, enquanto um valor muito alto pode levar a uma perda de detalhes e informações locais.

Uma das principais vantagens do KNN é a sua simplicidade e facilidade de implementação. Além disso, o KNN é um algoritmo não paramétrico, o que significa que não faz suposições sobre a distribuição dos dados. Isso permite que o KNN lide com dados complexos e não lineares.

O KNN tem uma ampla gama de aplicações em áreas como reconhecimento de padrões, processamento de imagens, análise de dados, recomendação de itens, medicina, entre outros. Sua eficácia depende da qualidade dos dados, escolha adequada do parâmetro K e consideração das características dos dados específicos do problema.

3.5.1 Fundamentos Teóricos

O KNN é um algoritmo baseado na ideia de que objetos semelhantes tendem a pertencer à mesma classe. Ele opera no espaço de características, onde cada objeto é representado por um conjunto de atributos ou características. O algoritmo se baseia na premissa de que objetos com

características semelhantes estão próximos uns dos outros no espaço de características.

Um dos principais conceitos utilizados no KNN é a medida de distância. O cálculo da distância entre objetos é fundamental para determinar a proximidade entre eles. A distância euclidiana é frequentemente usada como medida de distância, mas outras medidas, como a distância de *Manhattan*, a distância de *Minkowski* ou a distância de *Hamming*, podem ser aplicadas dependendo da natureza dos dados.

Outro conceito importante é a escolha do parâmetro K, que define o número de vizinhos mais próximos a serem considerados na classificação ou regressão. A escolha de K afeta diretamente o desempenho do algoritmo. Um valor muito baixo pode resultar em instabilidade e superajuste (*overfitting*), enquanto um valor muito alto pode levar a uma perda de detalhes e subajuste (*underfitting*).

O método é um algoritmo de aprendizado preguiçoso (*lazy learning*), pois não constrói um modelo explícito durante a fase de treinamento. Em vez disso, ele armazena os exemplos de treinamento e realiza cálculos sob demanda durante a fase de teste. Isso torna o KNN adequado para problemas com grandes volumes de dados, mas também pode torná-lo computacionalmente mais caro durante o teste.

3.5.2 Processo de Classificação

O processo de classificação no contexto do KNN envolve a atribuição de rótulos de classe a novos exemplos de teste com base nos exemplos de treinamento existentes. A classificação ocorre em três etapas principais: cálculo de distâncias, seleção dos K vizinhos mais próximos e atribuição do rótulo de classe.

- **Cálculo de Distâncias:** Na primeira etapa, o KNN calcula a distância entre o exemplo de teste e todos os exemplos de treinamento. A distância pode ser calculada usando diferentes medidas, como a distância euclidiana, a distância de Manhattan ou a distância de *Minkowski*. O objetivo é determinar a proximidade entre os exemplos de teste e treinamento no espaço de características.
- **Seleção dos K Vizinhos Mais Próximos:** Após calcular as distâncias, o KNN seleciona os K exemplos de treinamento mais próximos do exemplo de teste com base nas distâncias calculadas. Esses K vizinhos mais próximos são considerados os vizinhos imediatos do exemplo de teste. A escolha de K é um parâmetro do algoritmo e pode variar de acordo com o problema e o conjunto de dados.
- **Atribuição do Rótulo de Classe:** Com os K vizinhos mais próximos selecionados, o próximo passo é atribuir um rótulo de classe ao exemplo de teste. Para problemas de classificação, o rótulo de classe mais comum entre os K vizinhos é atribuído ao exemplo de teste. Isso é conhecido como "princípio da maioria dos votos". Se houver um empate

na contagem de votos, pode ser necessário adotar uma estratégia adicional, como atribuir o rótulo da classe do vizinho mais próximo ou ponderar os votos com base nas distâncias.

Em alguns casos, pode ser necessário considerar pesos diferentes para os vizinhos mais próximos com base em sua distância. Isso significa que os vizinhos mais próximos podem ter mais influência na classificação do que os vizinhos mais distantes.

Uma vez concluída a etapa de classificação para o exemplo de teste atual, o processo pode ser repetido para cada exemplo de teste no conjunto de dados de teste, até que todos os exemplos tenham sido classificados.

3.5.3 Seleção do Parâmetro K

A seleção do parâmetro K depende de diversos fatores, incluindo a natureza dos dados, a distribuição das classes, o número de exemplos de treinamento disponíveis e a complexidade do problema. Não há uma regra fixa para determinar o valor ideal de K, e geralmente é necessário realizar experimentação e validação para encontrar a melhor escolha.

Um valor muito pequeno de K, como $K = 1$, pode levar a uma classificação excessivamente sensível ao ruído e a *outliers*, resultando em um modelo instável e propenso a *overfitting*. Isso significa que o modelo pode se ajustar demasiadamente aos exemplos de treinamento específicos, mas não generalizar bem para novos exemplos. Por outro lado, um valor muito grande de K pode levar a uma perda de detalhes e informações locais, resultando em um modelo menos sensível a padrões específicos.

Uma abordagem comum para selecionar o valor de K é realizar uma validação cruzada (*cross-validation*) ou um conjunto de validação (*validation set*). A ideia é treinar e testar o modelo com diferentes valores de K e avaliar sua performance usando métricas como acurácia, precisão, *recall* ou *F1-score*. Em seguida, escolhe-se o valor de K que oferece o melhor equilíbrio entre desempenho e generalização.

Outra técnica utilizada é a busca em grade (*grid search*), onde se testa um conjunto pré-definido de valores de K e se avalia o desempenho do modelo para cada valor. Essa abordagem permite encontrar o valor ótimo de K dentro do conjunto pré-definido.

Além disso, é importante considerar a natureza do problema e a distribuição das classes. Por exemplo, em problemas com classes desbalanceadas, pode ser necessário ajustar o valor de K para lidar com a falta de representação de uma classe minoritária.

3.5.4 Medidas de Similaridade

Existem medidas de similaridade comumente utilizadas, dependendo da natureza dos dados e do problema em questão. Algumas das medidas de similaridade mais comuns incluem:

- **Distância Euclidiana**: É uma medida de distância comumente usada em espaços euclidianos. Ela calcula a distância entre dois pontos em um espaço multidimensional, usando a fórmula matemática da distância euclidiana. Quanto menor a distância euclidiana entre dois pontos, maior é a sua similaridade.
- **Distância de *Manhattan***: Também conhecida como distância de cidade, é uma medida de distância que calcula a soma das diferenças absolutas entre as coordenadas dos pontos. Ela é útil quando as dimensões dos dados são discretas ou categóricas.
- **Distância de *Minkowski***: É uma generalização das distâncias euclidiana e de Manhattan. Ela calcula a distância entre dois pontos em um espaço multidimensional, permitindo ajustar o grau de importância atribuído a cada dimensão. A distância euclidiana e a distância de Manhattan são casos especiais da distância de *Minkowski*.
- **Coeficiente de Correlação:** É uma medida de similaridade frequentemente usada em dados contínuos. Ela mede o grau de relação linear entre duas variáveis, variando entre -1 e 1. Um valor próximo de 1 indica uma alta similaridade, enquanto um valor próximo de -1 indica uma baixa similaridade.
- **Coeficiente de *Jaccard*:** É uma medida de similaridade amplamente utilizada em conjuntos ou dados binários. Ela calcula a interseção dividida pela união dos conjuntos, fornecendo uma medida de sobreposição entre os conjuntos. Quanto maior o coeficiente de *Jaccard*, maior a similaridade entre os conjuntos.

É importante escolher a medida de similaridade adequada para o tipo de dados e o problema. Além disso, em alguns casos, pode ser necessário ajustar ou personalizar a medida de similaridade para atender às necessidades específicas do domínio.

3.6 *Naive Bayes*

Naive Bayes é baseado no teorema de *Bayes*, que utiliza conceitos da teoria da probabilidade para realizar classificação ou estimativa probabilística. O algoritmo é chamado de "*naive*"(ingênuo) porque faz uma suposição simplificada e independente de que todos os atributos são condicionalmente independentes entre si, dada a classe.

O *Naive Bayes* é amplamente utilizado em tarefas de classificação de textos, como análise de sentimento, categorização de documentos e detecção de spam. Ele também é aplicado em

outras áreas, como diagnóstico médico, detecção de fraudes e recomendação de produtos.

A principal vantagem do *Naive Baye*s é sua simplicidade e eficiência computacional. Ele requer uma quantidade relativamente pequena de dados de treinamento para estimar os parâmetros do modelo. Além disso, o algoritmo é rápido e pode lidar com conjuntos de dados grandes e de alta dimensionalidade.

O *Naive Bayes* estima a probabilidade *a posteriori* da classe de um exemplo de teste com base nas probabilidades a priori da classe e nas probabilidades condicionais dos atributos, usando o teorema de Bayes. Ele assume independência entre os atributos, o que pode ser uma simplificação excessiva, mas em muitos casos funciona bem na prática.

Durante a fase de treinamento, o *Naive Bayes* calcula as probabilidades a priori de cada classe com base na frequência das classes no conjunto de treinamento. Em seguida, ele estima as probabilidades condicionais dos atributos para cada classe, utilizando diferentes distribuições de probabilidade, como a distribuição Gaussiana para atributos contínuos e a distribuição de *Bernoulli* ou multinomial para atributos discretos.

Durante a fase de teste, o *Naive Bayes* utiliza as probabilidades estimadas para calcular a probabilidade a posteriori da classe para um novo exemplo de teste. A classe com a maior probabilidade é então atribuída ao exemplo de teste.

Embora o *Naive Bayes* seja um algoritmo simples e eficaz em muitos cenários, ele também apresenta algumas limitações. A suposição de independência entre os atributos pode ser inadequada para alguns conjuntos de dados, e o algoritmo pode sofrer com atributos correlacionados. Além disso, o *Naive Bayes* pode ser sensível a atributos irrelevantes ou pouco informativos.

3.6.1 Fundamentos Teóricos

Naive Bayes é um algoritmo de aprendizado de máquina baseado na teoria das probabilidades e no teorema de Bayes. Seu objetivo é realizar a classificação de dados em categorias ou classes com base na probabilidade condicional dos atributos dado cada classe.

O algoritmo assume que todos os atributos são independentes entre si, uma suposição conhecida como independência condicional *naive*. Essa suposição simplificadora permite que o algoritmo calcule as probabilidades condicionais de cada atributo de forma separada, tornando o processo de aprendizado mais rápido e eficiente.

O teorema de *Bayes* é fundamental para o *Naive Bayes*, pois estabelece uma relação entre a probabilidade a posteriori de uma classe dado um conjunto de atributos e as probabilidades a priori da classe e dos atributos. A fórmula geral do teorema de *Bayes* é:

P(classe | atributos) = (P(classe) * P(atributos | classe)) / P(atributos)

Onde:

- $P(classe|atributos)$ é a probabilidade a posteriori da classe dado os atributos.
- $P(classe)$ é a probabilidade a priori da classe.
- $P(atributos|classe)$ é a probabilidade condicional dos atributos dado a classe.
- $P(atributos)$ é a probabilidade dos atributos.

Durante o processo de treinamento do *Naive Bayes*, são estimadas as probabilidades a priori de cada classe com base na frequência das classes no conjunto de treinamento. Além disso, são estimadas as probabilidades condicionais dos atributos para cada classe, utilizando diferentes distribuições de probabilidade, dependendo da natureza dos atributos (por exemplo, distribuição Gaussiana para atributos contínuos, distribuição de Bernoulli para atributos binários, distribuição multinomial para atributos discretos).

Na etapa de teste, o *Naive Bayes* utiliza as probabilidades estimadas para calcular a probabilidade a posteriori de cada classe para um novo exemplo de teste. A classe com a maior probabilidade a posteriori é atribuída ao exemplo de teste.

3.6.2 Processo de Classificação

O processo de classificação no contexto do algoritmo *Naive Bayes* envolve atribuir uma classe ou categoria a um novo exemplo com base nas estimativas de probabilidade calculadas durante a fase de treinamento. Esse processo pode ser dividido em etapas principais:

- **Preparação dos dados:** Antes de aplicar o algoritmo *Naive Bayes*, é necessário preparar os dados de treinamento e teste. Isso envolve garantir que os dados estejam em um formato adequado e que todos os atributos relevantes estejam presentes.
- **Estimação das probabilidades *a priori*:** Durante a fase de treinamento, o Naive Bayes calcula as probabilidades *a priori* de cada classe. Essas probabilidades representam a frequência relativa de cada classe nos dados de treinamento.
- **Estimação das probabilidades condicionais:** O *Naive Bayes* também estima as probabilidades condicionais dos atributos para cada classe. Essas probabilidades condicionais indicam a probabilidade de um determinado valor de atributo ocorrer, dado que a classe é verdadeira.
- **Cálculo da probabilidade *a posteriori*:** Com as probabilidades *a priori* e condicionais estimadas, o *Naive Bayes* pode calcular a probabilidade *a posteriori* de cada classe para um novo exemplo de teste. Isso é feito multiplicando as probabilidades a priori das classes pelas probabilidades condicionais dos atributos observados para cada classe.
- **Escolha da classe atribuída:** A classe atribuída ao novo exemplo de teste é determinada com base na probabilidade *a posteriori*. A classe com a maior probabilidade *a posteriori* é selecionada como a classe atribuída ao exemplo de teste.

3.6.3 Suavização e Prevenção de *Overfitting*

Naive Bayes é um algoritmo de aprendizado de máquina que pode ser suscetível ao *overfitting*, que ocorre quando o modelo se ajusta muito aos dados de treinamento e não consegue generalizar bem para novos dados. Para lidar com esse problema, podem ser aplicadas técnicas de suavização (*smoothing*) e prevenção de *overfitting*.

A suavização é utilizada para lidar com o desafio da ocorrência de probabilidades zero. Em alguns casos, durante a estimação das probabilidades condicionais, pode acontecer de um atributo não ser observado em uma classe específica, resultando em uma probabilidade condicional de zero para essa combinação. Isso pode levar a problemas durante a classificação.

Uma técnica comum de suavização é a suavização de *Laplace* (também conhecida como suavização de adição de um). Essa técnica consiste em adicionar uma quantidade fixa (geralmente 1) a todas as contagens de ocorrências de atributos em cada classe. Isso garante que todas as combinações possuam uma probabilidade não nula, mesmo que seja uma pequena probabilidade.

A suavização de *Laplace* ajuda a evitar a ocorrência de probabilidades zero e permite que o modelo faça predições razoáveis para casos não vistos no conjunto de treinamento. No entanto, é importante encontrar um equilíbrio na quantidade de suavização aplicada, pois uma suavização excessiva pode levar a uma perda de informação.

Além da suavização, existem outras técnicas que podem ser aplicadas para prevenir o *overfitting* no *Naive Bayes*, como:

- **Redução da dimensionalidade:** Reduzir a dimensionalidade do conjunto de atributos pode ajudar a evitar o *overfitting*, pois reduz a complexidade do modelo e evita a incorporação de ruído desnecessário nos dados.
- **Seleção de atributos:** Selecionar os atributos mais relevantes e informativos para o problema em questão pode melhorar o desempenho do modelo e reduzir a tendência ao *overfitting*.
- **Validação cruzada:** Utilizar técnicas de validação cruzada, como *k-fold cross-validation*, pode ajudar a avaliar o desempenho do modelo em dados não vistos durante o treinamento e identificar possíveis problemas de *overfitting*.
- **Regularização:** Em alguns casos, técnicas de regularização, como a regressão logística com regularização L1 ou L2, podem ser aplicadas ao Naive Bayes para controlar a complexidade do modelo e evitar o *overfitting*.

3.6.4 Considerações Especiais do Naive Bayes

O algoritmo *Naive Bayes* possui algumas considerações especiais que devem ser levadas em conta ao utilizá-lo:

- **Suposição de independência condicional:** assume que todos os atributos são independentes entre si, dado o valor da classe. Essa suposição simplificadora pode não ser verdadeira em alguns casos, especialmente quando há dependências entre os atributos. Portanto, é importante considerar a adequação dessa suposição em relação aos dados específicos do problema.
- **Sensibilidade a atributos irrelevantes:** pode ser sensível a atributos irrelevantes ou redundantes, pois considera todos os atributos igualmente importantes. A presença de atributos irrelevantes pode afetar negativamente a precisão do modelo. Portanto, é recomendado realizar uma análise cuidadosa dos atributos e considerar técnicas de seleção de atributos para melhorar o desempenho do *Naive Bayes*.
- **Tratamento de dados faltantes:** lida de forma eficiente com dados faltantes, pois utiliza apenas as informações disponíveis para calcular as probabilidades. No entanto, é importante considerar como lidar com dados ausentes durante a fase de pré-processamento para evitar distorções nas estimativas de probabilidade.
- **Alta eficiência computacional:** é conhecido por sua simplicidade e eficiência computacional. Ele pode ser aplicado em conjuntos de dados grandes sem requerer muitos recursos computacionais. Essa eficiência torna o *Naive Bayes* uma escolha popular em cenários com grandes volumes de dados.
- **Aplicabilidade em problemas de classificação:** é comumente utilizado em problemas de classificação, onde o objetivo é atribuir uma classe a um exemplo com base em seus atributos. Ele tem demonstrado bons resultados em várias áreas, como processamento de texto, detecção de spam, diagnóstico médico e reconhecimento de padrões.
- **Requisitos de independência estatística:** O desempenho é melhor quando os atributos são independentes estatisticamente, dado o valor da classe. Em cenários onde existem dependências complexas entre os atributos, outros algoritmos podem ser mais apropriados.

3.6.5 Aplicações e Considerações Finais

O algoritmo *Naive Bayes* tem sido amplamente utilizado em várias aplicações de aprendizado de máquina devido à sua simplicidade, eficiência e bom desempenho em muitos cenários. Algumas das aplicações mais comuns do *Naive Bayes* incluem:

- **Classificação de texto:** é frequentemente aplicado em problemas de classificação de texto, como filtragem de spam, análise de sentimento, categorização de documentos e classificação de notícias. Ele pode lidar com grandes volumes de texto de forma eficiente e fornecer classificações precisas.
- **Diagnóstico médico:** na área da saúde tem sido utilizado para auxiliar no diagnóstico médico. Pode ser aplicado para identificar doenças com base em sintomas e características clínicas, ajudando os médicos a tomar decisões mais precisas.
- **Recomendação de produtos:** pode ser utilizado em sistemas de recomendação para prever as preferências dos usuários e oferecer recomendações personalizadas de produtos ou serviços. Isso é especialmente útil em plataformas de comércio eletrônico e *streaming* de conteúdo.
- **Detecção de fraudes:** pode ser aplicado para detecção de fraudes em diversas áreas, como sistemas de cartões de crédito, segurança cibernética e detecção de atividades fraudulentas em transações financeiras.
- **Reconhecimento de padrões:** pode ser usado em problemas de reconhecimento de padrões, como reconhecimento facial, detecção de objetos em imagens e identificação de padrões em sinais de áudio.

Ao utilizar o *Naive Bayes*, é importante considerar algumas limitações:

- **Suposição de independência:** assume que os atributos são independentes, dado o valor da classe. Essa suposição nem sempre é verdadeira, e pode haver dependências entre os atributos que podem afetar a precisão do modelo.
- **Sensibilidade a atributos irrelevantes:** considera todos os atributos igualmente importantes, o que pode levar a uma sensibilidade a atributos irrelevantes. É essencial realizar uma análise cuidadosa dos atributos e considerar técnicas de seleção de atributos para melhorar o desempenho do modelo.
- **Adequação aos dados:** O desempenho pode variar dependendo do conjunto de dados e do problema em questão. É recomendado avaliar diferentes algoritmos e técnicas para encontrar a abordagem mais adequada para cada cenário específico.

Capítulo 4

Trabalhos Relacionados

4.1 Trabalhos Relacionados

O estudo conduzido por Silva (da Silva, Quintino Alves, Crispim Braga, Pereira Júnior, de Andrade & de Oliveira, 2017) teve como objetivo investigar óbitos infantis em crianças com idade de até um ano, empregando técnicas de mineração de dados. Para isso, utilizaram-se as bases de dados SIM e SINASC, referentes ao Estado do Ceará, entre os anos de 2013 e 2014, relacionando 1.182 indivíduos que sofreram óbito e 124.876 sobreviventes. O estudo considerou 16 atributos: idade da mãe, estado civil da mãe, escolaridade da mãe, local de nascimento, quantidade de filhos vivos em gestações anteriores, quantidade de filhos mortos em gestações anteriores, quantidade de semanas de gestação, tipo de gravidez, tipo de parto, sexo da criança, peso ao nascer, quantidade de consultas no pré-natal, Apgar1 no primeiro minuto de vida, Apgar5 no quinto minuto de vida, presença ou ausência de anomalias e cor da criança. As técnicas utilizadas no estudo foram *Random Forest*, KNN, *Naive Bayes*, SVM, RNA e J48. Os resultados da avaliação foram expressos por meio da curva ROC, que permite aferir a capacidade discriminatória dos algoritmos em termos de sensibilidade e especificidade. O J48 apresentou uma área sob a curva (AUC) de 0,888, seguido pelo *Random Forest* (AUC = 0,913), Naive Bayes (AUC = 0,924), IBK (AUC = 0,843), V. Perceptron (AUC = 0,875), MLP (AUC = 0,898) e SMO (AUC = 0,865).

Em outro estudo (Ramos, Silva, Moreira, Rodrigues, Oliveira & Monteiro, 2017), a equipe de pesquisa propôs o desenvolvimento do GISSA, um sistema inteligente de governança para o Sistema Único de Saúde (SISSU), com o objetivo de melhorar a assistência à saúde de mulheres grávidas e recém-nascidos. O GISSA utiliza técnicas de mineração de dados para gerar alertas de risco de mortalidade infantil, fornecendo suporte aos tomadores de decisão na implementação de ações preventivas. Para gerar os alertas de risco de morte de recém-nascidos, foi desenvolvido o LAIS, um sistema de análise de saúde inteligente que utiliza técnicas de mineração de dados. O LAIS emprega um conjunto de 16 atributos para identificar o risco de mortalidade infantil, dentre eles: peso ao nascer, idade gestacional, tipo de parto, presença de malformações congênitas, e outras informações relevantes. Os resultados preliminares indicam que o classificador *Naive Bayes* foi o mais eficaz em identificar o risco de mortalidade infantil,

apresentando uma precisão de 0,982, sensibilidae de 0,607, um F1-score de 0,396 e uma AUC de 0,921.

Foi realizado um estudo na cidade de Teerã (Saroj, Yadav, Singh, Chilyabanyama et al., 2022) com o objetivo de predizer a mortalidade infantil em crianças com idade inferior a cinco anos. Para a análise, foram utilizadas diversas técnicas, como *Random Forest*, *Naive Bayes*, KNN, regressão logística, SVM e rede neural, tendo sido utilizados 17 atributos. Os resultados indicaram que a rede neural foi o modelo mais eficiente para a previsão da mortalidade infantil abaixo de cinco anos. No entanto, a regressão logística também apresentou bons resultados, com precisão variando entre 94% a 95% e alcance de AUC de 0,934 a 0,94,8. Além disso, foram identificados alguns fatores importantes que influenciam a mortalidade infantil abaixo de cinco anos, como o número de crianças vivas, tempo de sobrevivência, índice de riqueza, tamanho da criança ao nascer, nascimento nos últimos cinco anos, número total de filhos já nascidos, nível de escolaridade da mãe e ordem de nascimento.

O estudo conduzido por Alves (Alves, Beluzo, Arruda, Bresan & Carvalho, 2020) se baseou em uma grande amostra de dados, composta por informações de nascimentos ocorridos entre os anos de 2000 a 2016, contabilizando um total de 30.873.500 amostras, das quais 208.391 foram óbitos infantis. Para a construção dos modelos preditivos, foram considerados 22 atributos e aplicados métodos de aprendizagem de máquina, tais como SVM, *Random Forest* e *Xgboost*. O método proposto se mostrou capaz de fornecer respostas precisas sobre o risco de morte neonatal, além de oferecer uma interpretação dos resultados obtidos. Os modelos apresentaram um desempenho satisfatório, tendo sido obtida uma AUC de 0,939 para o modelo *XGBoost*, além de AUC de 0,926 e 0,924 para os métodos *Random Forest* e SVM, respectivamente. Na análise das características mais importantes para a predição da mortalidade neonatal, foi possível identificar a relevância do peso do recém-nascido, do índice de Apgar no quinto minuto após o nascimento, da presença de malformações congênitas, do índice de Apgar no primeiro minuto após o nascimento, das semanas de gestação e do número de consultas pré-natais.

O estudo realizado por Soares (Soares, Zárate, Song & Nobre, 2021) visou predizer o óbito de crianças gêmeas com menos de um ano de idade, com base em dados provenientes do DATASUS. A amostra do estudo incluiu crianças nascidas entre os anos de 2012 e 2016, totalizando 303.379 registros de nascimentos múltiplos e 11.868 óbitos. O estudo utilizou algoritmos de aprendizado de máquina, incluindo Regressão Logística, Árvore de Decisão, SVM, *Random Forest* e *Gradient Boosting*, tendo obtido valores de AUC de 0,93, 0,72, 0,88, 0,93 e 0,95, respectivamente. Para as análises de classificação, foram selecionados 45 atributos. É importante ressaltar que o estudo buscou investigar a eficácia de métodos para lidar com pares de gêmeos, sendo que dois conjuntos de dados foram preparados e testados. O primeiro conjunto de dados, denominado "Dataset 1", continha todos os pares de gêmeos extraídos do banco de dados de recém-nascidos brasileiros (SINASC), agrupados em pares e considerados como registros independentes. O segundo conjunto de dados, "Dataset 2", foi construído com base em uma

estratégia que selecionava aleatoriamente um gêmeo do par para compor um novo conjunto de dados.

Capítulo 5

Material e Métodos

Neste capítulo, serão abordados os aspectos éticos envolvidos, destacando a obtenção dos dados a partir de fontes de domínio público. Em seguida, será descrita a seleção dos dados utilizados, com ênfase nos registros de nascidos vivos no estado do Paraná de 2017 e 2018 e mortes entre 2017 e 2019. Será explicado o processo de transformação dos dados, incluindo o pareamento das fontes de dados, a criação de variáveis resposta e a padronização das variáveis quantitativas. Além disso, serão apresentadas as etapas de pré-processamento realizadas para tratar dados faltantes e codificar variáveis categóricas. Por fim, serão discutidos os métodos de análise descritiva e exploratória, incluindo testes estatísticos para avaliar associações entre variáveis e o desfecho de mortalidade infantil. Também será abordado o desafio do desbalanceamento de classes e a divisão do conjunto de dados em treinamento e teste para o desenvolvimento de modelos de aprendizado de máquina.

5.1 Aspectos Éticos

Todas as informações utilizadas neste projeto foram obtidas a partir de fontes de domínio público, o que significa que os dados estão disponíveis livremente para o público em geral. Dessa forma, não há necessidade de registro ou aprovação junto aos Comitês de Ética em Pesquisa (CEP) e Comissão Nacional de Ética em Pesquisa (CONEP), conforme estabelecido pela resolução nº 510 de 7 de abril de 2016 (Guerriero, 2016).

Essa resolução, emitida pelo Conselho Nacional de Saúde (CNS), estabelece as diretrizes e normas regulamentadoras para a pesquisa em saúde no Brasil. Segundo o documento, as pesquisas que utilizam exclusivamente dados de domínio público e que não envolvem a identificação dos participantes ou o acesso a informações pessoais ou sensíveis estão isentas de análise pelos CEPs e pela CONEP.

No caso deste projeto, os dados foram coletados a partir de fontes públicas, ou seja, não foi necessário obter consentimento informado ou autorização específica para acessá-los. Além disso, todas as informações foram utilizadas de forma anonimizada, garantindo a privacidade dos indivíduos envolvidos.

Vale ressaltar que a ética em pesquisa é um tema fundamental e deve ser considerado em todas as etapas de um projeto de pesquisa. Mesmo quando a pesquisa é considerada isenta de análise pelos CEPs e pela CONEP, é importante garantir que os dados sejam utilizados de forma responsável e que a privacidade e os direitos dos indivíduos sejam respeitados. Para isso, é recomendável seguir as boas práticas em pesquisa e adotar medidas de segurança e anonimização dos dados.

5.2 Mineração de Dados

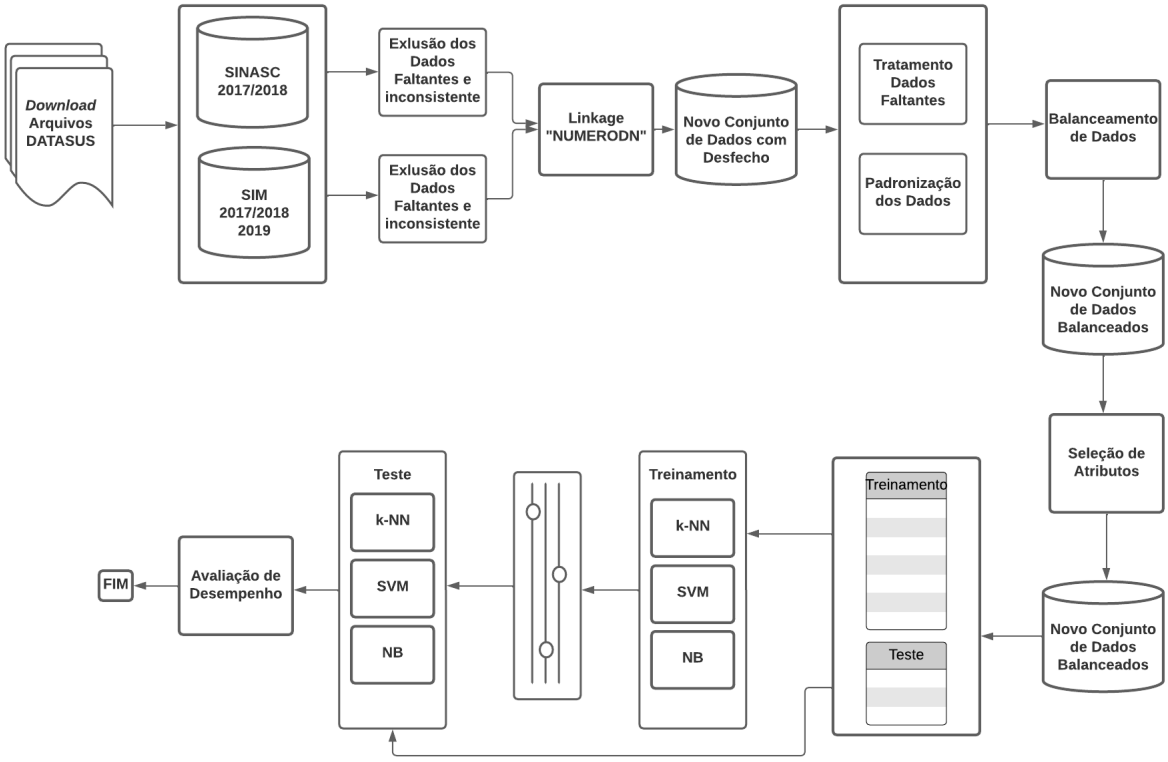

Figura 5.1: Aplicação das técnicas de *machine learning* para a análise preditiva de morte de crianças até um ano.

Para esta pesquisa, os dados foram obtidos a partir dos registros de nascidos vivos e mortes, tendo como referência o estado do Paraná no período 2016 a 2018, em hospitais que pertencem à rede do SUS. Os dados utilizados foram obtidos por meio das bases dos sistemas SINASC e SIM do DATASUS. Vale ressaltar que, para o estudo, os dados de óbito fetal foram desconsiderados, visto que o objetivo principal da pesquisa é analisar o perfil da mortalidade infantil, que abrange os óbitos nos períodos neonatal precoce e tardio, e pós-neonatal.

Os dados foram baixados do site do DATASUS, que é o departamento de informática do Sistema Único de Saúde do Brasil. A transferência foi realizada por região e período, e foram descarregados. Para facilitar a análise e a manipulação de tais informações, foram criados dois

bancos de dados distintos, um contendo todos os dados do SINASC e outro com todos os dados do SIM.

No banco de dados do SINASC, há 64 variáveis, sendo que 34 delas contém informações descontinuadas, relacionadas a cadastros e informações paternas. Em particular, exite a variável "NUMERODN", que indica o número único de registro de nascimento, foi mantida para ser confrontada com a base de dados do SIM. Já as 30 variáveis únicas, que descrevem, foram mantidas e estão listadas na tabela 5.1.

A análise dos dados do SIM foi realizada a partir da construção de uma variável resposta para todo o conjunto de dados do SINASC, considerando a presença na base do SIM como óbito e os demais registros como não óbito. É importante ressaltar que a mortalidade infantil é o foco deste estudo, o que justifica a exclusão dos óbitos fetais.

No contexto em questão, uma abordagem utilizada foi a preservação de todos os registros da classe minoritária "óbito"e a seleção aleatória de uma quantidade equivalente de registros da classe majoritária "não óbito"em uma base de dados binária.

Como resultado dessa estratégia, foi obtida uma nova base de dados balanceada, na qual ambas as classes possuem o mesmo número de exemplos. Essa abordagem é benéfica quando o objetivo é manter todos os exemplos da classe minoritária e garantir que o algoritmo de aprendizado de máquina leve em consideração devidamente ambas as classes durante o treinamento.

Por fim, para permitir a utilização dos algoritmos *Support Vector Machine* (SVM), que lidam apenas com dados numéricos, foi necessário converter os atributos simbólicos em numéricos e normalizar os atributos numéricos.

5.3 Transformação dos Dados

No contexto dos dados de saúde pública brasileiros, é essencial compreender a importância das variáveis relacionadas ao índice de Apgar no SINASC. O índice de Apgar é um indicador crucial para avaliar a vitalidade do recém-nascido, fornecendo informações sobre seu estado de saúde no momento do nascimento. No entanto, é comum encontrar dados faltantes nessas variáveis, principalmente devido a problemas no preenchimento de formulários manuscritos. Com o objetivo de minimizar o impacto dessas lacunas na análise e no treinamento de algoritmos de aprendizado de máquina, foi adotada a estratégia de preencher os dados faltantes utilizando a média dos dados existentes. Essa abordagem permite preservar a integridade da base de dados e aproveitar ao máximo as informações disponíveis, contribuindo para uma análise mais robusta e confiável dos dados de saúde pública. O estudo realizado no Espírito Santo evidencia que as variáveis relacionadas ao índice de Apgar apresentam uma baixa frequência de dados faltantes, com menos de 10% de ocorrência (da Silva, Moreira, Amorim, de Castro & Zandonade, 2014).

Para integrar e relacionar as fontes de dados utilizadas, foi empregada a técnica de parea-

mento determinístico. Essa abordagem consiste em identificar uma variável chave em comum entre as bases de dados, que, neste caso, foi o "NUMERODN". No processo de união das bases, quando a informação de data de nascimento estava ausente na base do SINASC, essa lacuna foi preenchida com a data correspondente proveniente da base do SIM confome Figura 5.2. Essa etapa permitiu obter uma base de dados integrada e mais completa. Posteriormente, com base nessas informações unificadas, foi criada uma nova variável denominada "desfecho", a qual indicava se a pessoa veio a óbito dentro de um ano após o nascimento ou não. Essa nova variável desempenhou um papel crucial na análise e compreensão dos desfechos relacionados à saúde dos indivíduos estudados.

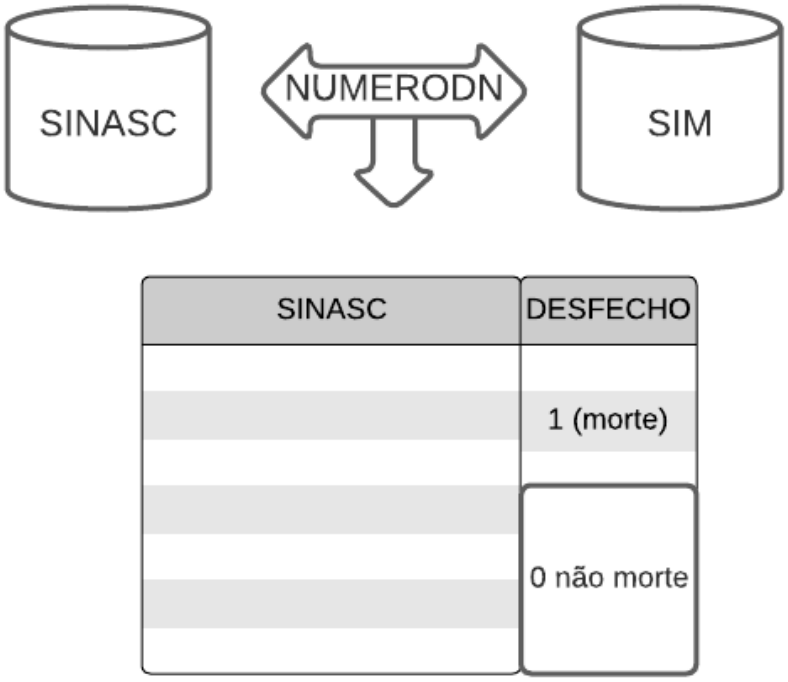

Figura 5.2: Relacionamento dos dados SINASC e SIM.

Para a análise, as bases com os óbitos foram filtradas para incluir apenas os nascimentos ocorridos entre 2018 e 2019, o que totalizou 309.670 nascimentos no estado do Paraná. A base dos nascimentos também foi filtrada para incluir apenas aqueles cujo "NUMERODN"estava presente na base de óbitos. Isso permitiu a formação de uma nova coluna na base de nascimentos com os desfechos de óbito e não óbito.

Com as bases pareadas, a variável resposta foi construída para todo o conjunto de dados do SINASC, considerando a presença na base do SIM como indicador de óbito e os demais registros como não óbito. Como algoritmos como o SVM lidam apenas com dados numéricos, foi necessário converter os atributos discretos em numéricos e normalizar os atributos numéricos para a análise.

Foram empregados métodos de pré-processamento com o objetivo de assegurar a adequa-

ção dos dados para serem utilizados em modelos de aprendizado de máquina.

Inicialmente, procedeu-se ao tratamento dos dados faltantes de acordo com a natureza das variáveis. No caso das variáveis quantitativas, optou-se por preencher os valores ausentes utilizando a média das observações existentes. Já para as variáveis qualitativas, recorreu-se à moda como forma de substituir os dados faltantes. Dessa maneira, todas as variáveis foram completadas com dados consistentes.

Uma etapa relevante do pré-processamento foi a transformação das variáveis quantitativas. Com o intuito de assegurar que todos os atributos possuíssem a mesma importância e não exercessem um impacto desproporcional sobre o modelo, procedeu-se à padronização das variáveis quantitativas. Essa abordagem envolveu a redimensionamento dos atributos para uma escala uniforme.

Adicionalmente, realizou-se a transformação dos atributos categóricos a fim de torná-los facilmente utilizáveis em modelos de aprendizado de máquina. Considerando que as variáveis categóricas podem ser classificadas como nominais ou ordinais, foram adotadas as devidas transformações. Os atributos nominais foram convertidos em variáveis binárias, representando assim as diferentes categorias. Por outro lado, as variáveis ordinais foram codificadas em uma escala numérica, preservando a ordem de classificação.

Na fase de análise descritiva e exploratória, é importante entender que o conjunto de dados e detectar possíveis associações entre as variáveis e a presença de observações discrepantes. Para tanto, foram calculadas as frequências absolutas e relativas para as variáveis qualitativas em relação ao desfecho (morte/não morte). Além disso, foram realizados testes para avaliar a significância das associações encontradas.

O teste do Qui-Quadrado foi selecionado para avaliar a associação entre variáveis categóricas. Essa técnica é amplamente utilizada nesse contexto, pois permite verificar se existe uma relação estatisticamente significativa entre duas variáveis, levando em consideração a frequência observada e a frequência esperada. Ao aplicar o teste do qui-quadrado, foi possível identificar possíveis associações entre os parâmetros de interesse e o desfecho. Essa análise é particularmente relevante para compreender a relação entre variáveis categóricas e identificar associações importantes no contexto do estudo (Armitage, Berry & Matthews, 2008).

Por sua vez, o teste *t-Student* foi escolhido para avaliar a diferença das médias entre grupos de amostras independentes. Essa técnica é especialmente adequada quando se deseja comparar as médias de um parâmetro em relação ao desfecho entre grupos distintos. O teste *t-Student* permite verificar se a diferença observada nas médias é estatisticamente significativa ou se pode ter ocorrido por acaso. A utilização desse teste proporcionou uma avaliação estatística robusta das diferenças entre os grupos, fornecendo evidências de que as diferenças observadas não foram resultantes do acaso (Armitage et al., 2008).

Adicionalmente, o valor de p (valor-p) desempenha um papel fundamental nesse contexto. Ele é uma medida estatística que indica a probabilidade de observar uma associação tão forte ou

uma diferença tão grande entre os grupos, considerando-se apenas o acaso. Um valor-p menor indica uma maior evidência estatística de uma associação significativa ou de uma diferença significativa entre os grupos. Portanto, ao interpretar os valores de p obtidos nos testes estatísticos, é possível avaliar a robustez e a significância estatística dos resultados.

Dessa forma, os métodos de redução de dimensionalidade selecionados, o teste Qui-Quadrado e o teste *t-Student*, foram justificados por sua capacidade de avaliar associações entre variáveis categóricas e diferenças nas médias, respectivamente, além da importância do valor de p na interpretação estatística.

Um problema comum que afeta o desempenho dos algoritmos é o desbalanceamento das classes, onde uma classe possui muito mais instâncias do que outra, resultando em uma sobreposição estatística entre as classes majoritária e minoritária. Para contornar esse problema, uma amostra aleatória com o mesmo tamanho da base de registros de óbito foi gerada a partir dos registros de não óbito, conforme Figura 5.3.

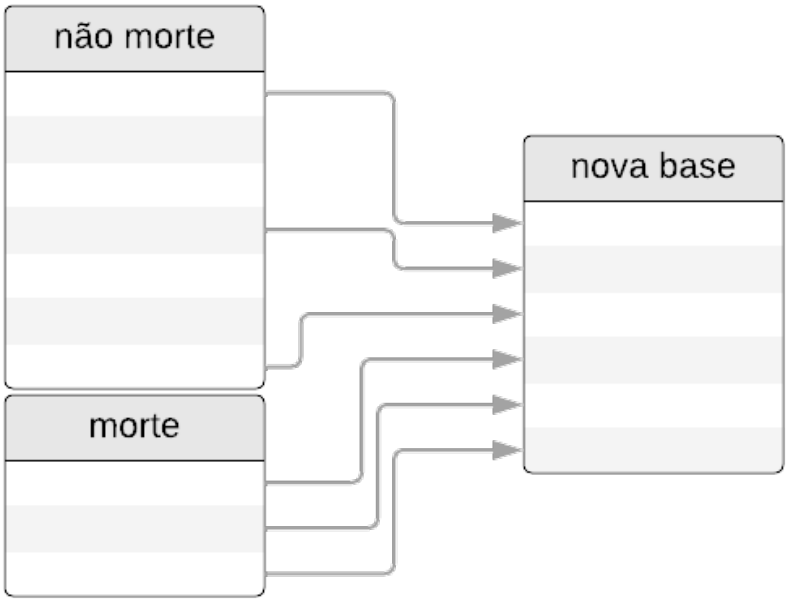

Figura 5.3: Balanceamento dos dados.

Uma prática comum em experimentos de aprendizado de máquina é a divisão do conjunto de dados em treinamento e teste (Hastie, Tibshirani, Friedman & Friedman, 2009). Nesse sentido, a proporção adotada pode influenciar diretamente nos resultados obtidos. No presente estudo, foi adotada a divisão 70:30, em que 70% dos dados foram destinados para o conjunto de treinamento e 30% para o conjunto de teste. Isso significa que o modelo foi ajustado utilizando 70% dos dados e testado com os 30% restantes, de forma a avaliar sua capacidade de generalização para novas observações. Essa proporção é amplamente utilizada na literatura e pode ser considerada como uma prática recomendada em experimentos de aprendizado de máquina.

5.4 Métodos de Classificação

Neste estudo, foram utilizados três modelos de aprendizado de máquina - SVM (*Support Vector Machine*), KNN (*K-Nearest Neighbors*) e NB (*Naive Bayes*) - em uma base de dados contendo inicialmente 16 atributos. O objetivo foi investigar a capacidade desses modelos em realizar a classificação com base nos atributos disponíveis.

Primeiramente, os três modelos foram treinados e avaliados utilizando todos os 16 atributos. Em seguida, uma análise estatística foi conduzida para identificar os atributos mais relevantes para o desfecho em estudo. Com base nos resultados dos testes estatísticos realizados para cada atributo, foram selecionados os 8 atributos com os menores valores de p, indicando uma maior associação estatística com o desfecho.

Esses 8 atributos selecionados foram então utilizados para treinar e avaliar novamente os modelos SVM, KNN e NB. Essa abordagem permitiu investigar se a redução da dimensionalidade para os atributos mais relevantes poderia melhorar o desempenho dos modelos em relação à classificação do desfecho.

O modelo SVM foi treinado com base no conjunto de dados de treinamento, utilizando a variável "Desfecho"como a variável de resposta e as demais variáveis disponíveis como variáveis preditoras. A função *kernel* utilizada foi "linear"e o parâmetro C, que controla o erro de classificação permitido, assumiu seu valor padrão.

O modelo KNN foi aplicado no conjunto de treinamento, utilizando o parâmetro "k"que varia de 1 a 20 para determinar o número de vizinhos considerados na classificação.

Já o modelo *Naive Bayes* foi aplicado no conjunto de treinamento, excluindo a variável resposta "Desfecho", e utilizado para prever as classes dos dados de teste.

Essas etapas de aplicação dos modelos e treinamento foram realizadas visando explorar a capacidade de classificação dos modelos com base nos atributos disponíveis na base de dados.

5.5 Pós Processamento

A análise dos resultados envolveu a construção das matrizes de confusão para cada modelo, tanto para os conjuntos de dados com 8 atributos quanto para os conjuntos de dados com 16 atributos. As matrizes de confusão permitiram avaliar o desempenho dos modelos na classificação das amostras em relação às classes positivas e negativas. Com base nas matrizes de confusão, foram calculadas diversas métricas de desempenho para cada modelo.

Uma tabela de métricas foi construída para apresentar de forma clara e concisa os resultados obtidos. Essa tabela incluiu métricas como acurácia, taxa de erro, sensibilidade, es-

pecificidade, precisão e F1-score. Essas métricas forneceram informações detalhadas sobre o desempenho de cada modelo em relação à classificação das amostras.

A medida de AUC-ROC (*Area Under the Receiver Operating Characteristic Curve*) também foi calculada para cada modelo. Essa medida é amplamente utilizada para avaliar o desempenho de modelos de classificação binária. A AUC-ROC fornece uma medida da capacidade discriminativa do modelo, ou seja, sua habilidade de distinguir entre as classes positiva e negativa. Valores mais altos de AUC-ROC indicam um melhor desempenho do modelo na classificação das amostras.

5.6 Ferramentas

Nesta seção, são descritas as principais ferramentas e recursos utilizados durante a condução do presente estudo. Essas ferramentas desempenharam um papel fundamental na coleta, processamento, análise e visualização dos dados, bem como na implementação dos algoritmos de aprendizado de máquina. A seguir, são apresentadas as principais ferramentas e recursos utilizados:

- **Linguagem de programação *Python*:** A linguagem de programação *Python* foi amplamente empregada neste estudo devido à sua ampla gama de bibliotecas e *frameworks* voltados para a análise de dados e aprendizado de máquina. Através do uso de bibliotecas como *NumPy*, *Pandas* e *scikit-learn*, foi possível realizar tarefas como pré-processamento de dados, construção de modelos de aprendizado de máquina, avaliação de desempenho e visualização de resultados.
- ***Jupyter Notebook*:** O *Jupyter Notebook* foi utilizado como ambiente de desenvolvimento interativo, permitindo a execução de código *Python* de forma iterativa, além de fornecer uma interface intuitiva para a visualização dos resultados. Os notebooks do *Jupyter* foram essenciais para o processo de exploração dos dados, a implementação dos algoritmos de aprendizado de máquina e a documentação das etapas do estudo.
- **Biblioteca *scikit-learn*:** A biblioteca *scikit-learn* é uma das principais bibliotecas de aprendizado de máquina em Python, fornecendo uma ampla variedade de algoritmos e ferramentas para tarefas como classificação, regressão, clustering e pré-processamento de dados. Utilizamos a *scikit-learn* para implementar e avaliar os modelos de aprendizado de máquina neste estudo, bem como para realizar operações de pré-processamento, como normalização de dados e redução de dimensionalidade.
- **R:** A linguagem de programação R foi utilizada para análise estatística e visualização dos dados. Através de pacotes como *ggplot2*, *dplyr* e *tidyr*, foi possível realizar análises estatísticas descritivas, criar gráficos e plotar visualizações dos resultados. O R é amplamente utilizado em estudos estatísticos e fornece uma variedade de recursos e funcionalidades

específicas para análise de dados.

- **Computador:** O computador utilizado para a execução do estudo possui um processador Intel i7 3770, 8GB de memória RAM e um SSD de 240GB. Essas especificações foram adequadas para o processamento dos dados e a execução dos algoritmos de aprendizado de máquina. O uso de um computador com boa capacidade de processamento e memória permitiu a realização das análises de forma eficiente.

Essas foram algumas das principais ferramentas e recursos utilizados durante o presente estudo. A combinação dessas ferramentas e o uso do computador adequado contribuíram para a realização de uma pesquisa de qualidade, garantindo uma abordagem metodológica sólida e uma análise adequada dos dados coletados.

Tabela 5.1: Estrutura do SINASC

SEQ	NOME	TIPO/TAM	DESCRIÇÃO
01	NUMERODN	C(08)	Número da DN, seqüencial por UF informante e por ano.
02	LOCNASC	C(01)	Local de ocorrência do nascimento, conforme a tabela: 9: Ignorado 1: Hospital 2: Outro Estab Saúde 3: Domicílio 4: Outros
03	CODESTAB	C(07)	Código de estabelecimento de saúde.
04	CODBAINASC	C(08)	Código Bairro nascimento.
05	CODMUNNASC	C(07)	Código do município de ocorrência.
06	IDADEMAE	C(02)	Idade da mãe em anos.
07	ESTCIVMAE	C(01)	Estado civil, conforme a tabel 1: Solteira 2: Casada 3: Viúva 4: Separado judicialmente/Divorciado 9: Ignorado
08	ESCMAE	C(01)	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
09	CODOCUPMAE	C(06)	Ocupação, conforme a Classificação Brasileira de Ocupações (CBO-2002).
10	QTDFILVIVO	C(02)	Número de filhos vivos.
11	QTDFILMORT	C(02)	Número de filhos mortos.
12	CODBAIRES	C(08)	Código bairro residência.
13	CODMUNRES	C(07)	Município de residência da mãe.
14	GESTACAO	C(01)	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
15	GRAVIDEZ	C(01)	Tipo de gravidez, conforme a tabela: 9: Ignorado 1: Única 2: Dupla 3: Tripla e mais

16	PARTO	C(01)	Tipo de parto, conforme a tabela: 9: Ignorado 1: Vaginal 2: Cesáreo
17	CONSULTAS	C(01)	Número de consultas de pré-natal: 1: Nenhuma 2: de 1 a 3 3: de 4 a 6 4: 7 e mais 9: Ignorado
18	DTNASC	C(08)	Data do nascimento, no formato ddmmaaaa
19	HORANASC	C(04)	Hora do nascimento
20	SEXO	C(01)	Sexo, conforme a tabela: 0: Ignorado 1: Masculino 2: Feminino
21	APGAR 1	C(02)	Apgar no primeiro minuto 00 a 10
22	APGAR 5	C(02)	Apgar no quinto minuto 00 a 10
23	RACACOR	C(01)	Raça/Cor: 1:Branca 2:Preta 3:Amarela 4: Parda 5: Indígena
24	PESO	C(04)	Peso ao nascer, em gramas.
25	IDANOMAL	C(01)	Anomalia congênita: 9-Ignorado 1=Sim 2=Não
26	CODANOMAL	(C20)	Código de malformação congênita ou anomalia cromossômica, de acordo com a CID-10.
27	DTCADASTR0	C(08)	Data de cadastramento no sistema.
28	DTRECEBIM	C(08)	Data de recebimento no nível central, data da última atualização do registro.
29	CODINST	C(18)	Código da Instalação da geração dos Registros.
30	UFINFORM	C(02)	Código da UF que informou o registro.

Capítulo 6

Resultados e Discussões

Neste capítulo, apresentaremos os resultados experimentais, juntamente com as observações e discussões correspondentes.

A Tabela 6.1 apresenta os testes estatísticos e as análises de correlação realizadas entre os atributos coletados do SINASC e o desfecho estudado.

Os atributos foram divididos em diferentes tipos de dados, como categóricos e numéricos. Foram aplicados testes estatísticos específicos para cada tipo de dado a fim de identificar possíveis associações significativas.

Os testes estatísticos utilizados incluíram o teste qui-quadrado, aplicado para atributos categóricos, e o teste *t-Student*, aplicado para atributos numéricos. O valor de p resultante de cada teste indica a probabilidade de observar uma associação ou diferença tão grande entre as variáveis considerando apenas o acaso.

Os resultados dos testes revelaram algumas associações estatisticamente significativas. Por exemplo, o atributo "LOCNASC"apresentou uma associação significativa com o desfecho, conforme indicado pelo teste qui-quadrado ($p = 0,0303$). Da mesma forma, o atributo "GESTACAO"também demonstrou uma associação estatisticamente significativa ($p = 0,0005$).

O critério adotado de $p < 0,05$ para selecionar os atributos mais significativos baseia-se em uma prática comum na análise estatística. Esse valor é frequentemente utilizado como um limiar para determinar se uma associação entre duas variáveis é estatisticamente significativa.

O valor de p < 0,05 indica que a probabilidade de obter uma associação tão forte ou uma diferença tão grande entre os grupos, considerando-se apenas o acaso, é inferior a 5%. Em outras palavras, se o valor de p for menor ou igual a 0,05, geralmente concluímos que existe uma associação estatisticamente significativa entre o atributo em questão e o desfecho.

No contexto do estudo, os 8 atributos com valores de p < 0,05 foram selecionados porque suas associações com o desfecho foram consideradas estatisticamente significativas. Essa abordagem visa identificar os atributos que têm um impacto significativo na predição da mortalidade de crianças com menos de um ano de idade. Esses atributos podem fornecer informações valiosas para compreender os fatores de risco e desenvolver estratégias preventivas ou intervencionistas adequadas para reduzir a mortalidade infantil.

Tabela 6.1: Testes estatísticos e análises de correlação SINASC vs Desfecho

Atributo	Tipo de Dado	Teste Estatístico	Resultado (valor de p)
NUMERODN	Categórico	-	-
LOCNASC	Categórico	Qui-Quadrado	p = 0,0303
CODESTAB	Categórico	-	-
CODBAINASC	Categórico	-	-
CODMUNNASC	Categórico	-	-
IDADEMAE	Numérico	*t-Student*	p = 0,9338
ESTCIVMAE	Categórico	Qui-Quadrado	p = 0,9703
ESCMAE	Categórico	Qui-Quadrado	p = 0,0754
CODOCUPMAE	Categórico	-	-
QTDFILVIVO	Numérico	*t-Student*	p = 0,4002
QTDFILMORT	Numérico	*t-Student*	p = 0,0642
CODBAIRES	Categórico	-	-
CODMUNRES	Categórico	-	-
GESTACAO	Categórico	Qui-Quadrado	p = 0,0005
GRAVIDEZ	Categórico	Qui-Quadrado	p = 1,89e-05
PARTO	Categórico	Qui-Quadrado	p = 0,8474
CONSULTAS	Categórico	Qui-Quadrado	p = 6,612e-10
DTNASC	Categórico	-	-
HORANASC	Categórico	-	-
SEXO	Categórico	Qui-Quadrado	p = 0,4004
APGAR 1	Numérico	*t-Student*	p = 2,2e-16
APGAR 5	Numérico	*t-Student*	p = 2,2e-16
RACACOR	Categórico	Qui-Quadrado	p = 0,2387
PESO	Numérico	*t-Student*	p = 2,2e-16
IDANOMAL	Categórico	Qui-Quadrado	p = 1,316e-11
CODANOMAL	Categórico	-	-
DTCADASTRO	Numérico	-	-
DTRECEBIM	Numérico	-	-
CODINST	Categórico	-	-
UFINFORM	Categórico	-	-

É importante ressaltar que a escolha de $p < 0,05$ como critério de significância é uma convenção e pode variar dependendo do campo de estudo e do contexto específico da pesquisa.

A matriz de confusão fornece informações sobre os verdadeiros positivos (VP), verdadeiros negativos (VN), falsos positivos (FP) e falsos negativos (FN) para cada modelo. Os VP e VN representam as previsões corretas do modelo, enquanto FP e FN indicam os erros do modelo. O objetivo é maximizar os VP e VN, e minimizar os FP e FN.

As Tabelas 6.2, 6.3 e 6.4 apresentam as matrizes de confusão para os modelos de classificação SVM, KNN e Naive Bayes (NB), respectivamente, considerando 8 parâmetros. A análise dessas tabelas permite compreender os resultados alcançados pelos modelos em termos de acertos e erros.

Tabela 6.2: Matriz de Confusão SVM, para 8 parâmetros

		Valor Predito	
		Morte	Não Morte
Real	Morte	270	8
	Não Morte	70	208

Tabela 6.3: Matriz de Confusão KNN, para 8 parâmetros

		Valor Predito	
		Morte	Não Morte
Real	Morte	261	17
	Não Morte	72	206

Tabela 6.4: Matriz de Confusão NB, para 8 parâmetros

		Valor Predito	
		Morte	Não Morte
Real	Morte	261	17
	Não Morte	72	206

Observando a Tabela 6.2, notamos que o modelo SVM obteve um número maior de VP (270) e VN (208) em comparação com os modelos KNN e NB. Além disso, apresentou um menor número de FP (70) e FN (8) em relação aos demais modelos. Isso parece indicar que o modelo SVM conseguiu prever corretamente a maioria das mortes infantis e não mortes infantis.

O modelo KNN, apresentado na Tabela 6.3, obteve resultados numericamente iguais aos do modelo SVM em termos de VP (261) e VN (206), mas apresentou mais FP (72) e FN (17) em comparação com o modelo SVM. Por outro lado, o modelo NB, apresentado na Tabela 6.4, teve resultados numericamente iguais aos do modelo KNN em termos de VP (261) e FP (72), mas obteve menos VN (206) e mais FN (17) do que o modelo KNN.

Considerando agora as Tabelas 6.5, 6.6 e 6.7 para 16 parâmetros, podemos observar as seguintes alterações:

Na matriz de confusão do modelo SVM (Tabela 6.5), houve um aumento nos VP (275) e

uma redução nos FP (68). Além disso, o número de FN (3) também diminuiu em comparação com a configuração de 8 parâmetros. Para o modelo KNN (Tabela 6.6), observamos que houve um aumento numérico nos VP (263) e uma diminuição nos FN (15). No entanto, houve um pequeno aumento nos FP (70) em comparação com a configuração de 8 parâmetros. Já o modelo NB (Tabela 6.7) mostrou um aumento nos VP (265) e uma diminuição nos FN (13). O número de FP (68) também diminuiu em relação à configuração de 8 parâmetros.

Tabela 6.5: Matriz de Confusão SVM, para 16 parâmetros

		Valor Predito	
		Morte	Não Morte
Real	Morte	275	3
	Não Morte	68	210

Tabela 6.6: Matriz de Confusão KNN, para 16 parâmetros

		Valor Predito	
		Morte	Não Morte
Real	Morte	263	15
	Não Morte	70	208

Tabela 6.7: Matriz de Confusão NB, para 16 parâmetros

		Valor Predito	
		Morte	Não Morte
Real	Morte	265	13
	Não Morte	68	210

Com base nos dados obtidos, as métricas de avaliação foram aplicadas aos modelos SVM, KNN e NB. Essas métricas incluem acurácia, taxa de erro, sensibilidade, especificidade, precisão e F1-Score.

A Tabela 6.8 apresenta os valores dessas métricas para os modelos com 8 parâmetros. O modelo SVM alcançou uma acurácia de 0,908, indicando que 90,8% das previsões foram corretas, enquanto a taxa de erro foi de 0,092. A sensibilidade do SVM foi de 0,973, refletindo sua habilidade de identificar corretamente 97,3% das mortes infantis reais, enquanto a especificidade foi de 0,723, representando sua capacidade de identificar corretamente 72,3% das não mortes infantis reais. A precisão do modelo SVM foi de 0,799, e o F1-Score foi de 0,876.

Tanto o modelo KNN quanto o modelo NB obtiveram resultados numericamente iguais em relação às métricas de desempenho para a configuração de 8 parâmetros. Ambos os modelos alcançaram uma acurácia de 0,840 e uma taxa de erro de 0,160. A sensibilidade e especificidade foram iguais para os dois modelos, com um valor de 0,783 para sensibilidade e 0,924 para especificidade. A precisão foi de 0,939 para ambos os modelos e o F1-Score de 0,854.

Já na Tabela 6.9, são apresentados os valores das métricas para os modelos com 16 parâmetros. O modelo SVM obteve uma acurácia de 0,878 e uma taxa de erro de 0,122. A sensibilidade do SVM foi de 0,989, enquanto a especificidade foi de 0,755. A precisão do

Tabela 6.8: Métricas de Avaliação dos Classificadores, para 8 parâmetros

Classificador	Acurácia	Taxa de Erro	Sensibilidade	Especificidade	Precisão	F1-Score
SVM	0,908	0,092	0,973	0,723	0,799	0,876
KNN	0,840	0,160	0,783	0,924	0,939	0,854
NB	0,840	0,160	0,783	0,924	0,939	0,854

modelo SVM foi de 0,986, e o F1-Score foi de 0,944.

Tabela 6.9: Métricas de Avaliação dos Classificadores, para 16 parâmetros

Classificador	Acurácia	Taxa de Erro	Sensibilidade	Especificidade	Precisão	F1-Score
SVM	0,878	0,122	0,989	0,755	0,986	0,944
KNN	0,829	0,171	0,946	0,748	0,946	0,862
NB	0,878	0,122	0,953	0,755	0,953	0,868

No caso do modelo KNN para a configuração de 16 parâmetros, foi observada uma acurácia de 0,829 e uma taxa de erro de 0,171. A sensibilidade foi de 0,946, denotando uma capacidade razoável de identificar as mortes infantis reais, e a especificidade foi de 0,748. A precisão foi de 0,946, e o *F1-Score* foi de 0,862. Similarmente ao modelo KNN, o modelo NB para a configuração de 16 parâmetros também apresentou uma acurácia de 0,878 e uma taxa de erro de 0,122. A sensibilidade foi de 0,953, a especificidade foi de 0,755, a precisão foi de 0,953 e o F1-Score foi de 0,868.

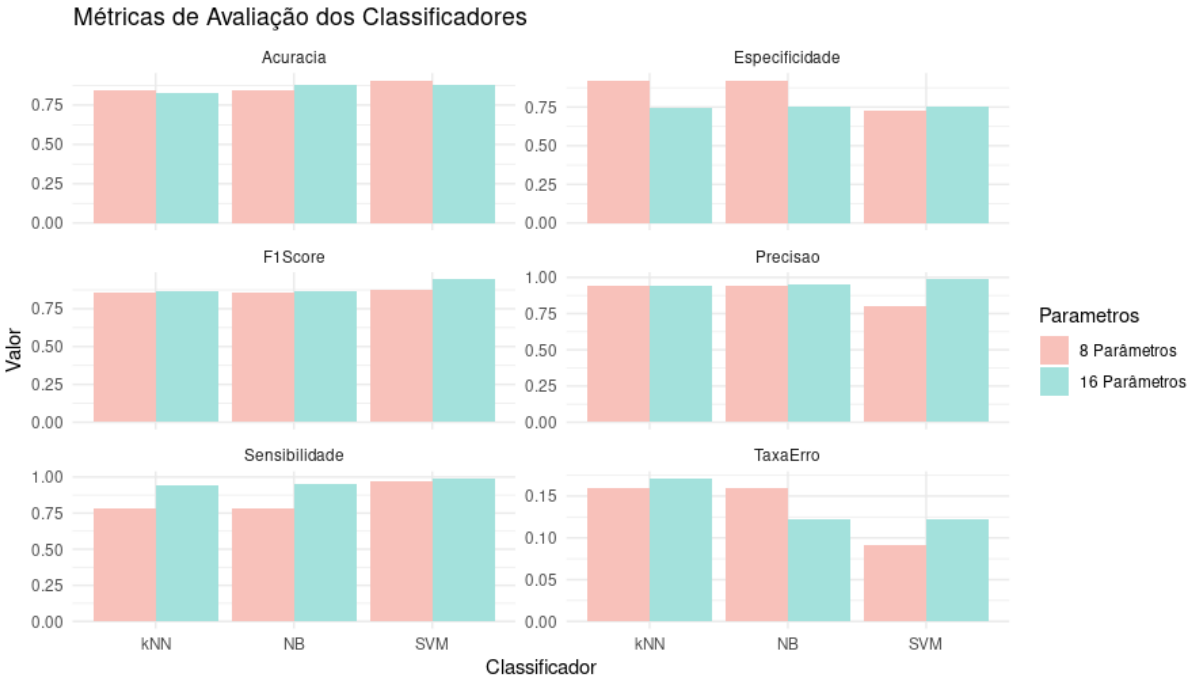

Figura 6.1: Métricas de avaliação dos Classificadores para 8 e 16 Parâmetros.

Os resultados obtidos, e mostrados nas Tabela 6.8, na Tabela 6.9 e na Figura 6.1 ,em nosso estudo revelam informações valiosas sobre a predição da mortalidade infantil. Ao analisar os algoritmos utilizados, constatamos que a média harmônica entre precisão e sensibilidade, representada pela métrica *F1-score*, apresentou um valor médio de aproximadamente 90% para

ambos os estados: "Morte"e "Não morte". No entanto, foi observado que a sensibilidade da classe "Morte"é ligeiramente inferior à classe "Não morte", enquanto a precisão é ligeiramente maior para a classe "Morte". Esses resultados indicam que os modelos tendem a cometer mais erros ao classificar instâncias da classe "Morte"como "Não morte", embora o número de falsos positivos para a classe "Morte"seja menor.

Essas descobertas sugerem a existência de bebês que faleceram, mas foram erroneamente classificados como "Não morte"pelos modelos de classificação adotados. Isso sugere que alguns bebês possuem características que os assemelham à classe "Não morte", mesmo que tenham vindo a óbito. Essa análise das métricas de avaliação dos modelos proporciona uma visão esclarecedora sobre a capacidade desses modelos.

Essas observações sugerem que há casos em que bebês que faleceram foram classificados erroneamente como "Vivos"pelos modelos de classificação utilizados. Isso pode indicar que alguns bebês possuíam características que apontavam para a classe "Vivo", mas, por alguma razão, acabaram falecendo. Essa análise das métricas de avaliação dos modelos revela *insights* importantes sobre a capacidade desses modelos em distinguir entre bebês vivos e óbitos infantis.

A Tabela 6.10 apresenta uma compilação dos resultados de alguns estudos de predição de mortalidade infantil utilizando técnicas de aprendizado de máquina. Esses estudos foram selecionados com base na sua relevância para abordar o tema usando bases de dados semelhantes e na disponibilidade de métricas de desempenho, como a curva AUC-ROC.

Para o conjunto de 8 atributos do presente estudo, o SVM obteve um valor de AUC de 0,861, seguido pelo KNN com 0,848 e *Naive Bayes* com 0,850. Já para o conjunto de 16 atributos, o SVM apresentou um desempenho superior, com um valor de AUC de 0.903, enquanto o KNN obteve um valor de 0,850 e o *Naive Bayes* de 0,891.

Ao comparar os resultados do presente estudo com os estudos citados na tabela, observa-se que os modelos SVM e *Naive Bayes* apresentaram resultados competitivos em relação a outros estudos anteriores. No entanto, é importante destacar que cada estudo pode ter suas próprias características, como a seleção de atributos, o tamanho da amostra e a qualidade dos dados, o que pode influenciar nos resultados obtidos.

Após a análise dos resultados apresentados na Tabela 6.10, é possível observar o desempenho de diferentes estudos de predição de mortalidade infantil utilizando técnicas de aprendizado de máquina. Dentre os estudos analisados, nosso modelo apresentou resultados comparáveis em termos de AUC em relação a alguns estudos anteriores.

O estudo conduzido por Silva (da Silva et al., 2017) apresentou resultados interessantes ao utilizar seis técnicas de aprendizado de máquina para a predição da mortalidade infantil. Os valores de AUC obtidos variaram de 0,843 a 0,924, sendo que a abordagem mais bem-sucedida, o *Naive Bayes*, alcançou uma AUC superior à do nosso estudo (0,850 para 8 atributos e 0,891 para 16 atributos). No entanto, é importante ressaltar que o estudo de Silva utilizou uma base de dados relativa a outra região, no caso o estado do Ceará, e um período diferente.

Tabela 6.10: Resultados dos estudos de predição de mortalidade infantil utilizando técnicas de aprendizado de máquina.

Estudo	**Técnica**	**Atributos**	**AUC-ROC**
Silva (da Silva et al., 2017)	J48	16	0,888
	Random Forest	16	0,913
	Naive Bayes	16	0,924
	IBK	16	0,843
	V. *Perceptron*	16	0,875
	MLP	16	0,898
	SMO	16	0,865
Ramos (Ramos et al., 2017)	*Naive Bayes*	16	0,921
Saroj (Saroj et al., 2022)	Rede Neural	17	0,934-0,948
	Regressão Logística	17	0,934-0,948
Alves (Alves et al., 2020)	*XGBoost*	22	0,939
	Random Forest	22	0,926
	SVM	22	0,924
Soares (Soares et al., 2021)	Regressão Logística	45	0,930
	Árvore de Decisão	45	0,720
	SVM	45	0,880
	Random Forest	45	0,930
	Gradient Boosting	45	0,950
Presente Estudo	SVM	8	0,861
	KNN	8	0,848
	Naive Bayes	8	0,850
	SVM	16	0,903
	KNN	16	0,850
	Naive Bayes	16	0,891

Essas diferenças contextuais podem influenciar os resultados e, portanto, tornam difícil uma comparação direta entre os estudos. Cada região possui suas particularidades em relação a fatores socioeconômicos, acesso aos serviços de saúde e políticas de saúde implementadas, o que pode impactar a predição da mortalidade infantil.

Em contraste com o estudo conduzido por Alves (Alves et al., 2020), no qual foram empregados algoritmos *XGBoost*, *Random Forest* e SVM para a predição da mortalidade infantil, o presente estudo utilizou técnicas de aprendizado de máquina, incluindo SVM, KNN e *Naive Bayes*. Embora o estudo de Alves tenha demonstrado que o *XGBoost* obteve um desempenho superior, com uma AUC de 0,939, é importante destacar que Alves utilizou um conjunto de dados com 22 atributos, enquanto nosso estudo empregou um conjunto de dados com um número reduzido de atributos. Essas diferenças nas abordagens e nos conjuntos de dados tornam inviável uma comparação direta entre os resultados alcançados por Alves e os resultados obtidos neste estudo. Cada pesquisa possui suas próprias particularidades e, portanto, os resultados devem ser interpretados considerando-se o contexto específico de cada estudo.

Não é possível estabelecer uma comparação direta entre o estudo realizado por Saroj

(Saroj et al., 2022) e o presente trabalho, pois eles abordam contextos diferentes. O estudo mencionado por Saroj foi conduzido em hospitais específicos do Teerã, no Irã, com suas próprias características e peculiaridades, enquanto o nosso estudo foi realizado em um contexto distinto. É importante ressaltar que as condições de coleta de dados, a população de estudo e os métodos empregados podem variar entre as pesquisas, o que inviabiliza uma comparação direta dos resultados. Portanto, é necessário considerar os estudos individualmente, reconhecendo as particularidades de cada contexto e seus respectivos achados.

O estudo conduzido por Soares (Soares et al., 2021) utilizou diversas técnicas de aprendizado de máquina, incluindo Regressão Logística, Árvore de Decisão, SVM, *Random Forest* e *Gradient Boosting*, para caracterizar a predição da mortalidade infantil. Os resultados obtidos foram bastante variados, com valores de AUC variando de 0,720 a 0,950 para o conjunto de 45 atributos utilizado. É importante destacar que o desempenho do modelo de Regressão Logística foi relativamente bom, alcançando uma AUC de 0,930. No entanto, o modelo de Árvore de Decisão obteve um desempenho inferior, com uma AUC de 0,720. Por outro lado, os modelos SVM, *Random Forest* e *Gradient Boosting* apresentaram resultados promissores, com AUCs de 0,880, 0,930 e 0,950, respectivamente. Assim como mencionado anteriormente, é necessário levar em consideração que o estudo de Soares pode ter sido realizado em um contexto específico, com diferentes características populacionais, condições socioeconômicas e fatores de risco para mortalidade infantil.

Capítulo 7

Conclusão

Nesse trabalho, investigamos a predição de mortalidade infantil por meio de técnicas de aprendizado de máquina. Ao longo deste estudo, exploramos diferentes algoritmos, conjuntos de atributos e métricas de desempenho, com o objetivo de desenvolver modelos eficazes para essa importante tarefa no campo da saúde.

Os resultados obtidos demonstram o potencial promissor das abordagens de aprendizado de máquina na predição de mortalidade infantil. Os modelos SVM, kNN e *Naive Bayes* mostraram-se capazes de distinguir entre bebês vivos e óbitos infantis. Esses modelos proporcionaram informações valiosas que podem auxiliar na tomada de decisões médicas e no planejamento de políticas de saúde voltadas para a redução da mortalidade infantil.

No entanto, é fundamental destacar que os resultados desta pesquisa são específicos para o conjunto de dados e atributos utilizados. Cada contexto e conjunto de atributos podem apresentar particularidades que influenciam o desempenho dos modelos. Portanto, é necessário considerar a adequação e a generalização dos resultados para diferentes cenários antes de aplicar esses modelos em larga escala.

Ademais, é importante mencionar as limitações e desafios encontrados durante a realização desta pesquisa. A qualidade dos dados utilizados, incluindo a integridade, a completude e a representatividade das informações, desempenha um papel crucial na precisão e confiabilidade dos modelos desenvolvidos. Portanto, é imprescindível investir em coleta de dados precisa e abrangente, bem como em estratégias para lidar com dados ausentes ou inconsistentes.

Um aspecto relevante deste estudo é a utilização de um conjunto de dados real, composto por informações de pacientes reais coletadas em um hospital público brasileiro. Isso confere maior relevância e validade aos resultados obtidos, pois refletem a realidade da população estudada. Além disso, a aplicação de técnicas de pré-processamento de dados, como a seleção de atributos, desempenha um papel crucial na otimização da análise. A seleção cuidadosa dos atributos reduz o número de variáveis consideradas, evitando a inclusão de informações redundantes ou irrelevantes.

Este estudo destaca-se por fornecer evidências de que é possível predizer a mortalidade infantil, utilizando um conjunto de dados representativo e apenas oito atributos, apesar dos resultados serem inferiores quando são utilizados 16 atributos. Esses resultados têm potencial

aplicação na prática clínica, auxiliando os profissionais de saúde na tomada de decisões e na adoção de medidas preventivas direcionadas à redução da taxa de mortalidade infantil. Além disso, a seleção de atributos demonstrou ser uma técnica valiosa para otimizar a análise de grandes conjuntos de dados.

É importante ressaltar que este estudo se baseia em uma revisão de outros trabalhos relacionados à predição da mortalidade infantil, fornecendo uma visão comparativa com outras abordagens e técnicas de aprendizado de máquina e embora sejam conjuntos de dados distintos. Essa análise comparativa permite uma melhor compreensão do desempenho. No entanto, são necessárias pesquisas adicionais para confirmar e validar os resultados obtidos, explorando diferentes conjuntos de dados e aplicando técnicas avançadas de aprendizado de máquina.

7.1 Trabalhos Futuros

Com base nos resultados apresentados neste trabalho, alguns trabalhos futuros podem ser seguidos para melhorar a precisão da classificação da mortalidade infantil.

Primeiramente, é importante considerar a inclusão de outros atributos relacionados à saúde da mãe e do bebê, como fatores nutricionais, socioeconômicos, entre outros. Esses atributos podem contribuir para a construção de modelos mais robustos e precisos.

Além disso, é possível explorar outras técnicas de aprendizado de máquina e algoritmos de classificação que possam apresentar um desempenho ainda melhor na tarefa de predição da mortalidade infantil. Por exemplo, é possível utilizar redes neurais profundas, que podem aprender características mais complexas dos dados.

Outro aspecto importante a ser considerado é a validação dos modelos em diferentes conjuntos de dados, para verificar se o desempenho obtido é consistente em diferentes populações e contextos. Também é importante investigar possíveis vieses nos dados, como a falta de informações sobre determinados atributos, e buscar soluções para minimizar esses problemas.

Por fim, é importante destacar a importância da aplicação dos modelos de predição de mortalidade infantil em contextos reais, como em unidades de saúde e hospitais, com o objetivo de auxiliar os profissionais de saúde na tomada de decisão e na priorização de recursos e cuidados. Isso pode contribuir para a redução da mortalidade infantil e melhorar a qualidade de vida das mães e dos bebês.

Referências Bibliográficas

Adriaans, P., Van Benthem, J. et al. (2008). *Philosophy of information*, AmsterdamNorth Holland. Citado 2 vezes nas páginas 20 e 21.

Agranonik, M. & Jung, R. O. (2019). Qualidade dos sistemas de informações sobre nascidos vivos e sobre mortalidade no rio grande do sul, brasil, 2000 a 2014, *Ciência & Saúde Coletiva* **24**: 1945–1958. Citado na página 20.

Agresti, A. (2012). *Categorical data analysis*, Vol. 792, John Wiley & Sons. Citado na página 28.

Alves, L. C., Beluzo, C. E., Arruda, N. M., Bresan, R. C. & Carvalho, T. (2020). Assessing the performance of machine learning models to predict neonatal mortality risk in brazil, 2000-2016, *medRxiv* . Citado 2 vezes nas páginas 44 e 63.

Armitage, P., Berry, G. & Matthews, J. N. S. (2008). *Statistical methods in medical research*, John Wiley & Sons. Citado na página 50.

Careti, C. M., Scarpelini, A. H. P. & de Carvalho Furtado, M. C. (2014). Perfil da mortalidade infantil a partir da investigação de óbitos, *Revista Eletrônica de Enfermagem* **16**(2): 352–60. Citado na página 14.

Chao, F., You, D., Pedersen, J., Hug, L. & Alkema, L. (2018). National and regional under-5 mortality rate by economic status for low-income and middle-income countries: a systematic assessment, *The Lancet Global Health* **6**(5): e535–e547. Citado na página 19.

da Silva, C. L., Quintino Alves, J., Crispim Braga, O., Pereira Júnior, J. W., de Andrade, L. O. M. & de Oliveira, A. M. B. (2017). Usando o classificador naive bayes para geração de alertas de risco de óbito infantil., *Revista Eletrônica de Sistemas de Informação* **16**(2). Citado 3 vezes nas páginas 43, 62 e 63.

da Silva, L. P., Moreira, C. M. M., Amorim, M. H. C., de Castro, D. S. & Zandonade, E. (2014). Evaluation of the quality of data in the live birth information system and the information system on mortality during the neonatal period in the state of espírito santo, brazil, between 2007 and 2009, *Ciência & Saúde Coletiva* **19**(7): 2011. Citado na página 48.

da Silva Leandro, B. B., Rezende, F. A. V. S. & da Conceição Pinto, J. M. (2020). *Informações e registros em saúde e seus usos no SUS*, SciELO-Editora FIOCRUZ. Citado 2 vezes nas páginas 20 e 21.

de Vaconcelos, A. L. R. & Guerrero, A. V. P. (2013). Rede cegonha, *Physis Revista de Saúde Coletiva* **23**(4): 1297–1316. Citado na página 19.

do Milênio, O. d. D. (2014). Relatório nacional de acompanhamento, *Brasília: Ipea* . Citado na página 19.

Fayyad, U., Piatetsky-Shapiro, G. & Smyth, P. (1996). From data mining to knowledge discovery in databases, *AI magazine* **17**(3): 37–37. Citado 2 vezes nas páginas 23 e 25.

Giovanni, M. D. (2014). Rede cegonha: da concepção à implantação. Citado na página 14.

Guerriero, I. C. Z. (2016). Resolução nº 510 de 7 de abril de 2016 que trata das especificidades éticas das pesquisas nas ciências humanas e sociais e de outras que utilizam metodologias próprias dessas áreas, *Ciência & Saúde Coletiva* **21**: 2619–2629. Citado na página 46.

Hastie, T., Tibshirani, R., Friedman, J. H. & Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction*, Vol. 2, Springer. Citado na página 51.

IPEA, S. (2014). Objetivos de desenvolvimento do milênio: relatório nacional de acompanhamento.
URL: *https://repositorio.ipea.gov.br/handle/11058/3205* Citado na página 14.

Laurenti, R., Jorge, M. H. P. & Gotlieb, S. L. D. (2008). Mortes maternas e mortes por causas maternas, *Epidemiologia e Serviços de Saúde* **17**(4): 283–292. Citado na página 14.

Neves, R. d. C. D. d. (2003). Pré-processamento no processo de descoberta de conhecimento em banco de dados. Citado 2 vezes nas páginas 10 e 24.

of Electrical, I. & Engineers, E. (n.d.). *ICCCN 2016 : 2016 25th International Conference on Computer Communications and Networks (ICCCN) : August 1 - August 4, 2016*. Não citado.

Ramos, R., Silva, C., Moreira, M. W., Rodrigues, J. J., Oliveira, M. & Monteiro, O. (2017). Using predictive classifiers to prevent infant mortality in the brazilian northeast, *2017 IEEE 19th International Conference on e-Health Networking, Applications and Services (Healthcom)*, IEEE, pp. 1–6. Citado 2 vezes nas páginas 43 e 63.

Romero, D. E. & Cunha, C. B. d. (2006). Avaliação da qualidade das variáveis sócio-econômicas e demográficas dos óbitos de crianças menores de um ano registrados no sistema de informações sobre mortalidade do brasil (1996/2001), *Cadernos de Saúde Pública* **22**: 673–681. Citado na página 14.

Saroj, R. K., Yadav, P. K., Singh, R., Chilyabanyama, O. et al. (2022). Machine learning algorithms for understanding the determinants of under-five mortality, *BioData mining* **15**(1): 1–22. Citado 3 vezes nas páginas 44, 63 e 64.

Soares, W. L., Zárate, L. E., Song, M. A. & Nobre, C. N. (2021). 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* **7**(5): 45269–45290. Citado 3 vezes nas páginas 44, 63 e 64.

Tan, P.-N., Steinbach, M. & Kumar, V. (2016). *Introduction to data mining*, Pearson Education India. Citado na página 23.

UNICEF (2023). Malnutrition in children - unicef data.
URL: *https://data.unicef.org/topic/nutrition/malnutrition* Citado na página 14.