CENTRO UNIVERSITÁRIO CHRISTUS

SISTEMAS DE INFORMAÇÃO

DAVID GOMES WANDERLEY

CLASSIFICAÇÃO DE VIA DE PARTO UTILIZANDO APRENDIZADO DE MÁQUINA

FORTALEZA

2023

DAVID GOMES WANDERLEY

CLASSIFICAÇÃO DE VIA DE PARTO UTILIZANDO APRENDIZADO DE MÁQUINA

Trabalho de Conclusão de Curso (TCC) apresentado ao curso de Sistemas de Informação do Centro Universitário Christus, como requisito parcial para obtenção do grau de bacharel em Sistemas de Informação.

Orientador: Prof. Me. Felipe Timbó Brito

FORTALEZA

2023

Dados Internacionais de Catalogação na Publicação
Centro Universitário Christus - Unichristus
Gerada automaticamente pelo Sistema de Elaboração de Ficha Catalográfica do
Centro Universitário Christus - Unichristus, com dados fornecidos pelo(a) autor(a)

W245c Wanderley, David Gomes.
CLASSIFICAÇÃO DE VIA DE PARTO UTILIZANDO APRENDIZADO DE MÁQUINA / David Gomes Wanderley. - 2023.
30 f. : il. color.

Trabalho de Conclusão de Curso (Graduação) - Centro Universitário Christus - Unichristus, Curso de Sistemas de Informação, Fortaleza, 2023.
Orientação: Prof. Me. Felipe Timbó Brito.

1. Via de Parto. 2. Parto Normal. 3. Cesariana. 4. Aprendizado de Máquina. 5. Classificação. I. Título.

CDD 004.07

DAVID GOMES WANDERLEY

CLASSIFICAÇÃO DE VIA DE PARTO UTILIZANDO APRENDIZADO DE MÁQUINA

Trabalho de Conclusão de Curso (TCC) apresentado ao curso de Sistemas de Informação do Centro Universitário Christus, como requisito parcial para obtenção do grau de bacharel em Sistemas de Informação.

Aprovada em:

BANCA EXAMINADORA

Prof. Me. Felipe Timbó Brito (Orientador)
Centro Universitário Christus (Unichristus)

Prof. Dr. Daniel Nascimento Teixeira
Centro Universitário Christus (Unichristus)

Prof. Me Iago Castro Chaves
Centro Universitário Christus (Unichristus)

Quem me apoiou, quem me ajudou, quem me fez chegar até aqui, eu agradeço de coração.

AGRADECIMENTOS

Gostaria de expressar meus sinceros agradecimentos em primeiro lugar a Deus pela oportunidade de viver, respirar e poder estudar.

Agradeço aos meus pais, Carlos e Vanessa, por me proporcionarem a oportunidade de estudar e por me auxiliarem em minha trajetória desde o início até o presente. Sou imensamente grato pelo apoio emocional e por tudo que já fizeram por mim. Amo vocês, pai e mãe.

À minha irmã Sarah, que me ajudou com conselhos e me apoiou, e a toda a minha família em geral.

Gostaria de agradecer especialmente à minha tia Beta, que me ajudou durante o último ano da faculdade, orientando-me e incentivando-me nos estudos, na busca por estágio, entre outras coisas.

Ao professor Celso, que quando eu estava no ensino médio e não fazia ideia do que fazer após o colégio, foi fundamental ao dar uma palestra que me fez decidir ingressar no curso de Sistemas de Informação.

Agradeço imensamente ao meu orientador, professor Felipe Timbó. Ele foi um professor que me ajudou muito durante as disciplinas que ministrou e, principalmente, durante a elaboração do meu Trabalho de Conclusão de Curso. Sem o senhor, eu não teria conseguido. Muito obrigado de coração, professor. Desejo-lhe muito sucesso em sua jornada, junto à sua família e em sua vida. Que Deus o abençoe e o auxilie em todas as suas empreitadas.

Ao meu Coordenador, professor Daniel, meu muito obrigado por me acompanhar e auxiliar ao longo de toda a minha formação. Também agradeço pelos e-mails que o senhor envia sobre oportunidades de estágio, pois foram de grande ajuda. Por tudo isso, muito obrigado.

Por fim, gostaria de agradecer aos meus amigos, que me apoiaram e acompanharam em minha trajetória. Meus sinceros agradecimentos a todos vocês.

"A vantagem é recíproca, pois os homens, enquanto ensinam, aprendem."

(Lucius Annaeus Seneca)

RESUMO

A classificação de via de parto é uma tarefa clínica importante que busca determinar se um parto será realizado de forma vaginal ou por meio de uma cesariana. Essa decisão tem implicações significativas para a saúde da mãe e do bebê, bem como para os recursos médicos disponíveis. Neste contexto, o uso de técnicas de aprendizado de máquina tem se mostrado promissor para auxiliar os profissionais de saúde na tomada de decisão. Este trabalho apresenta um estudo sobre a aplicação de algoritmos de aprendizado de máquina na classificação de via de parto. O objetivo é mostrar que é possível desenvolver um modelo computacional de aprendizado de máquina, capaz de fornecer uma previsão sobre a via de parto, utilizando apenas informações relacionadas à mãe, tais como idade, ocupação da mãe, quantidade de partos que a mãe já realizou, entre outros. Os resultados experimentais mostram que os modelos de aprendizado de máquina alcançam um desempenho promissor na classificação de via de parto, com acurácia média superior a 70%, apenas utilizando dados da mãe.

Palavras-chave: Via de Parto. Parto Normal. Cesariana. Aprendizado de Máquina. Classificação.

ABSTRACT

Classification of the route of delivery is an important clinical task that seeks to determine whether a delivery will be performed vaginally or through a cesarean section. This decision has significant implications for the health of the mother and baby, as well as available medical resources. In this context, the use of machine learning techniques has shown promise to assist health professionals in decision making. This work presents a study on the application of machine learning algorithms in the classification of the type of delivery. The objective is to show that it is possible to develop a computer model of machine learning, capable of providing a prediction about the mode of delivery, using only information related to the mother, such as age, mother's occupation, number of deliveries that the mother has already performed, between others. The experimental results show that the machine learning models achieve a promising performance in classifying the type of delivery, with an average accuracy greater than 70%, using only the mother's data.

Keywords: Delivery mode. Vaginal birth. Cesarean section. Machine learning. Classification.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

SINASC	Sistema de Informações sobre Nascidos Vivo
SVS	Secretaria de Vigilância em Saúde
SVM	Support Vector Machine
ROC	Receiver Operating Characteristic Curve
SO	Sistema Operacional
CPU	Central processing unit
RAM	Random Access Memory
SSD	Solid-state Drive
VP	Verdadeiro Positivo
FN	Falso Negativo
FP	Falso Positivo
VN	Verdadeiro Negativo

SUMÁRIO

1 INTRODUÇÃO

1.1 Contextualização e delimitação do tema

O nascimento de um bebê é um momento de grande expectativa e emoção para pais e familiares. Ao planejar a chegada do pequeno ser, uma das decisões mais importantes a serem tomadas é o tipo de parto: vaginal ou cesáreo. Ambos os métodos têm suas próprias características, indicações e implicações para a mãe e o bebê (WEIDLE *et al.*, 2014).

O parto vaginal, conhecido como parto normal, é o método de parto mais comum e natural. Nesse tipo de parto, o bebê é expelido do útero através do canal de parto de forma espontânea, impulsionado pelas contrações uterinas e pela pressão exercida pelo movimento do bebê. O parto vaginal traz uma série de vantagens, como uma recuperação geralmente mais rápida para a mãe, menor risco de complicações cirúrgicas, menor tempo de internação hospitalar e menor probabilidade de infecção. Além disso, o parto vaginal proporciona benefícios para o bebê, como a compressão do tórax durante o trajeto pelo canal de parto, que ajuda a eliminar líquidos dos pulmões e estimula o sistema imunológico (FERREIRA *et al.*, 2014).

Por outro lado, o parto cesáreo é uma intervenção cirúrgica em que o bebê é retirado do útero através de uma incisão abdominal e uterina. Esse tipo de parto é geralmente realizado por indicações médicas, como complicações durante a gestação ou trabalho de parto, riscos para a mãe ou o bebê, ou opção pessoal da gestante. O parto cesáreo apresenta vantagens em casos específicos, como quando há problemas de saúde que tornam o parto vaginal arriscado ou quando o bebê está em posição desfavorável. No entanto, é importante ressaltar que o parto cesáreo é uma cirurgia maior, com riscos associados, como infecções, sangramento excessivo e complicações respiratórias para o bebê. Além disso, a recuperação pós-operatória pode ser mais lenta e dolorosa para a mãe (BARBOSA *et al.*, 2003).

Com o avanço das tecnologias e o crescente volume de dados disponíveis, o campo do aprendizado de máquina tem se mostrado uma ferramenta promissora para lidar com problemas de classificação complexos (ZHOU, 2021). O aprendizado de máquina é uma subárea da inteligência artificial que envolve o desenvolvimento de algoritmos e modelos capazes de aprender padrões e relações a partir de dados, sem a necessidade de programação explícita. Ao aplicar algoritmos de aprendizado de máquina a conjuntos de dados relevantes, é possível explorar informações e identificar características que podem ser utilizadas para classificar, por exemplo, a via de parto de forma eficiente.

1.2 Problematização

A classificação da via de parto, que envolve a distinção entre parto vaginal e cesáreo, é uma tarefa complexa e crucial para os profissionais de saúde. A precisão na determinação da via de parto tem implicações diretas na saúde da mãe e do bebê, no planejamento dos recursos hospitalares e na tomada de decisões clínicas.

A disponibilidade de dados para a predição da via de parto utilizando técnicas de aprendizado de máquina desempenha um papel fundamental na performance de modelos. Contudo, existem situações em que dados importantes, como por exemplo dados clínicos, podem não estar disponíveis para utilização em modelos de aprendizado de máquina. Essas circunstâncias, podem incluir contextos clínicos desatualizados, onde os registros de saúde são mantidos em formato físico ou em sistemas desatualizados, dificultando a coleta e o uso desses dados para análise. A falta de digitalização dos dados limita sua disponibilidade para alimentar os modelos de aprendizado de máquina.

Além disso, restrições legais e éticas podem ser um obstáculo para a utilização de dados clínicos em modelos de aprendizado de máquina. Regulamentações de proteção de dados, como a Lei Geral de Proteção de Dados (LGPD) (BRASIL, 2018), estabelecem diretrizes estritas sobre o uso e compartilhamento de informações pessoais. Isso significa que, em alguns casos, os dados clínicos podem não estar disponíveis para pesquisa ou análise devido a requisitos de privacidade e consentimento dos pacientes.

Outra situação em que os dados clínicos podem não estar disponíveis é quando ocorrem problemas de interoperabilidade entre sistemas de saúde (BITTAR *et al.*, 2018). Em muitos lugares, os registros de saúde são mantidos por diferentes instituições e sistemas, o que pode dificultar a integração e o compartilhamento de informações entre eles. Isso pode levar a lacunas nos dados, tornando-os incompletos ou fragmentados, o que compromete sua utilização efetiva em modelos de aprendizado de máquina.

Nesse contexto, utilizar informações da mãe do bebê pode ser promissor para o problema de classificar a via de parto. Ao aplicar técnicas de aprendizagem de máquina nesse cenário, abre-se uma perspectiva promissora de desenvolver modelos de classificação precisos. Dessa forma, a problemática reside em utilizar efetivamente a aprendizagem de máquina para classificar a via de parto com base apenas em dados maternos isoladamente.

1.3 Pressupostos

Os pressupostos deste estudo têm como base a premissa de que os dados maternos podem fornecer informações relevantes para a classificação da via de parto. Acredita-se que características físicas e histórico obstétrico da mãe desempenham um papel significativo na determinação da via de parto.

Com base nos pressupostos mencionados anteriormente, formula-se a seguinte hipótese para este estudo: a utilização exclusiva de dados da mãe, por meio de técnicas de aprendizagem de máquina, é capaz de fornecer uma classificação precisa e confiável da via de parto. Acredita-se que as informações maternas, como quantidade de partos, idade da mãe, escolaridade e número de consultas realizadas antes do parto, contêm padrões e relações que podem ser identificados por algoritmos de aprendizagem de máquina, permitindo uma classificação adequada da via de parto. Supõe-se que esses modelos sejam capazes de capturar a complexidade do problema, considerando a importância relativa de cada variável materna na decisão final da via de parto.

1.4 Objetivos

Este trabalho tem como objetivo investigar a aplicação de técnicas de aprendizado de máquina para classificar a via de um parto com base apenas em informações disponíveis sobre mãe do bebê.

Em particular, os objetivos específicos deste trabalho são:

1. Obter um conjunto de dados relevantes contendo informações sobre a mãe do bebê para aplicação de técnicas de aprendizado de máquina.
2. Selecionar os atributos mais importantes e que possam ser utilizados antes do parto para o problema de classificação.
3. Implementar algoritmos de aprendizado de máquina adequados para a tarefa de classificação de via de parto.
4. Realizar a avaliação do desempenho dos modelos de aprendizado de máquina desenvolvidos, utilizando métricas de desempenho, a fim de determinar a eficácia e a confiabilidade desses modelos na classificação da via de parto.

1.5 Estrutura do trabalho

Os próximos capítulos são estruturados da seguinte forma: No capítulo 2 é apresentado a fundamentação teórica deste trabalho. O capítulo 3 mostra a metodologia utilizada e todo o passo a passo utilizado para construção dos modelos computacionais. No capítulo 4 os resultados obtidos são apresentados. Por fim, o trabalho é concluído e direções futuras são apontadas no capítulo 5.

2 REFERENCIAL TEÓRICO

Esta seção apresenta o referencial teórico necessário para embasar a compreensão e análise do problema em questão. São abordados conceitos relacionados a dados de parto, aprendizado de máquina e técnicas de desempenho para problemas de classificação.

2.1 SINASC

O Sistema de Informações sobre Nascidos Vivos (SINASC, 2023) é um sistema que coleta dados de nascimentos relatados nacionalmente e fornece dados de natalidade para todos os níveis do sistema de saúde. Dados provenientes desse sistema foram utilizados neste trabalho.

2.1.1 Como são coletados os dados

Para coletar os dados, deve-se primeiro imprimi-los e depois preencher os detalhes pré-numerados na sequência. A Secretaria de Saúde é responsável pela produção e distribuição para o estado. A Secretaria Estadual de Saúde é responsável pela distribuição para os municípios. Os profissionais de saúde são obrigados a coletar dados e implementar serviços de entrega sob demanda e, no caso de entrega em domicílio, devem ser cadastrados e acompanhados pela Secretaria Municipal de Saúde.

2.1.2 Processamentos dos dados

Na Secretaria Municipal de Saúde, as certidões de nascimento são impressas, processadas, criticadas e consolidadas no SINASC local. Os dados de nascimento relatados pelos municípios são transferidos para bancos de dados nacionais para serem agregados e enviados para o nível federal. Essas transferências são feitas pela Internet e ocorrem simultaneamente em três níveis de controle.

No âmbito federal, a SVS (Secretaria de Vigilância em Saúde) atua na análise, avaliação e divulgação das informações do SINASC, agregando informações estado a estado e produzindo relatórios analíticos sobre taxas de fecundidade, quadros de indicadores e outras ferramentas de informação estatística e divulgando-os em âmbito nacional.

2.2 Aprendizado de máquina

Aprendizado de máquina é um ramo da inteligência artificial que se concentra no desenvolvimento de algoritmos e modelos capazes de aprender a partir de dados e realizar tarefas específicas sem serem explicitamente programados para isso (CARVALHO *et al.*, 2011). Ou seja, é um processo pelo qual uma máquina aprende a identificar padrões nos dados e usa esses padrões para fazer previsões ou tomar decisões.

A importância do aprendizado de máquina reside em sua capacidade de processar grandes quantidades de dados e encontrar padrões complexos que seriam difíceis ou impossíveis de serem identificados por um ser humano. Isso torna o aprendizado de máquina uma ferramenta poderosa para muitas áreas, incluindo reconhecimento de fala, processamento de imagens, análise de sentimentos, detecção de fraudes, previsão de demanda e muitas outras.

2.2.1 Aprendizado não supervisionado

O aprendizado de máquina não supervisionado é uma abordagem de aprendizado de máquina em que os algoritmos são treinados em um conjunto de dados não rotulados (ROSSI, 2016), ou seja, sem a presença de rótulos ou respostas conhecidas. O aprendizado de máquina não supervisionado visa encontrar padrões e estruturas ocultas nos dados sem qualquer orientação ou supervisão.

Os algoritmos de aprendizado de máquina não supervisionado são capazes de identificar grupos de dados semelhantes, chamados de clusters, e descobrir relações entre diferentes características dos dados. Isso pode ser útil em muitas aplicações, como segmentação de clientes, análise de sentimentos, detecção de anomalias, entre outras.

2.2.2 Aprendizado supervisionado

O aprendizado de máquina supervisionado é uma abordagem de aprendizado de máquina em que os algoritmos são treinados em um conjunto de dados rotulados (MONARD; BARANAUSKAS, 2003), ou seja, dados que possuem rótulos ou respostas conhecidas. O objetivo do treinamento é ensinar o algoritmo a fazer previsões precisas sobre novos dados que não foram vistos durante o treinamento.

Os algoritmos de aprendizado de máquina supervisionado são usados para resolver problemas de classificação e regressão. Em um problema de classificação, o objetivo é prever

a classe ou categoria a que um determinado exemplo pertence, como a classificação de emails como spam ou não-spam, ou a classificação de imagens como contendo um objeto ou não. Em um problema de regressão, o objetivo é prever um valor numérico, como o preço de uma casa com base em suas características.

Existem vários tipos de algoritmos de aprendizado de máquina supervisionado, incluindo árvores de decisão, redes neurais, regressão logística e algoritmos de SVM. Cada algoritmo tem suas próprias vantagens e desvantagens, dependendo da natureza do problema de aprendizado de máquina.

Neste trabalho utilizamos o aprendizado supervisionado, visto que possuímos um conjunto de dados rotulados contendo informações da mãe e a respectiva via de parto. Essa abordagem nos permite treinar um modelo de aprendizagem de máquina para aprender os padrões presentes nos dados e, posteriormente, utilizá-lo para realizar previsões de classificação da via de parto com base em novos dados maternos.

2.2.3 Aprendizado por reforço

O aprendizado de máquina por reforço é uma abordagem de aprendizado de máquina em que um agente aprende a tomar decisões em um ambiente dinâmico para maximizar uma recompensa acumulada ao longo do tempo (DORÇA *et al.*, 2012).

Nessa abordagem, o agente toma ações em um ambiente e recebe *feedback* imediato na forma de recompensas ou penalidades, com o objetivo de aprender a melhor sequência de ações para obter a maior recompensa total.

No aprendizado de máquina por reforço, o agente interage com o ambiente em etapas discretas. Em cada etapa, o agente observa o estado atual do ambiente, toma uma ação com base nessa observação e recebe uma recompensa. Com o tempo, o agente aprende a mapear os estados do ambiente para ações que maximizam a recompensa acumulada ao longo do tempo. O objetivo final do agente é aprender uma política, que é uma função que mapeia estados para ações, de forma a obter a maior recompensa possível.

2.3 Métricas para classificação

Métricas de classificação são medidas que são usadas para avaliar a qualidade das previsões feitas por um modelo de aprendizado de máquina em um problema de classificação.

Essas métricas são úteis para medir o desempenho do modelo em diferentes aspectos, como precisão, *recall*, *F1-score*, entre outros.

O uso de métricas de classificação ajuda a avaliar o desempenho do modelo em diferentes aspectos e pode ser útil na seleção de modelos e ajuste de parâmetros para melhorar a qualidade das revisões (HOSSIN; SULAIMAN, 2015).

2.3.1 Matriz de Confusão

A matriz de confusão é a tabela que é usada para avaliar o desempenho de um modelo de aprendizado de máquina em um problema de classificação. Ela permite avaliar a qualidade das previsões feitas pelo modelo, comparando as previsões com as classes reais dos exemplos de teste.

A matriz de confusão é uma tabela de duas dimensões, em que as linhas representam as classes reais e as colunas representam as classes previstas pelo modelo. Cada célula na tabela contém o número de exemplos que foram classificados em uma determinada combinação de classe real e prevista. Um exemplo de matriz de comfução está presente na Tabela 1, as quatro células da matriz são formadas por VP, VN, FP e FN (SAMMUT; WEBB, 2017).

Tabela 1 – Tabela que exemplifica uma matriz de confusão.

		Valor Predito	
		Sim	**Não**
Valor Real	**Sim**	VP	FN
	Não	FP	VN

Fonte: Elaborado pelo autor.

2.3.2 Acurácia

Accuracy (acurácia) é uma métrica comum de avaliação de modelos de classificação, que mede a proporção de exemplos classificados corretamente pelo modelo em relação ao número total de exemplos. Em outras palavras, a *Accuracy* indica a porcentagem de previsões corretas que o modelo faz em relação ao número total de previsões. Por exemplo, se um modelo de classificação é treinado para distinguir se o parto de uma mulher foi cesariano ou normal e é testado em um conjunto de dados contendo 100 linhas referente a partos normais e 100 linhas referente a parte cesariana, e o modelo classifica corretamente 180 casos, então a *Accuracy* do modelo seria de 90%.

A Acurácia é medida como o número de previsões corretas dividido pelo número total de itens (SAMMUT; WEBB, 2017), conforme definida a seguir:

$$\text{acurácia} = \frac{\text{número de predições corretas}}{\text{número total de objetos}}$$

2.3.3 Precision

A *Precision* é uma métrica que mede a proporção de exemplos classificados como positivos corretamente em relação a todos os exemplos classificados como positivos pelo modelo. Em outras palavras, ela representa a capacidade do modelo de identificar corretamente os verdadeiros positivos e minimizar os falsos positivos.

A fórmula para calcular a *Precision* é a seguinte:

$$\text{Precision} = \frac{\text{verdadeiros positivos}}{\text{verdadeiros positivos} + \text{falsos positivos}}$$

Uma alta *Precision* indica que o modelo tem uma baixa taxa de falsos positivos, ou seja, a maioria dos exemplos classificados como positivos é realmente positiva. No entanto, uma *Precision* alta não garante uma boa capacidade de identificar todos os verdadeiros positivos.

2.3.4 Recall

O *Recall* é uma métrica que mede a proporção de exemplos positivos corretamente identificados em relação a todos os exemplos verdadeiramente positivos no conjunto de dados. Em outras palavras, representa a capacidade do modelo de encontrar todos os verdadeiros positivos e minimizar os falsos negativos.

A fórmula para calcular o *Recall* é a seguinte:

$$\textit{Recall} = \frac{\text{verdadeiros positivos}}{\text{verdadeiros positivos} + \text{falsos negativos}}$$

Um *Recall* alto indica que o modelo tem uma baixa taxa de falsos negativos, ou seja, ele é capaz de identificar a maioria dos exemplos positivos presentes no conjunto de dados. No entanto, um *Recall* alto não garante uma boa capacidade de evitar falsos positivos.

2.3.5 F1 Score

O *F1 Score* é uma métrica que combina a *Precision* e o *Recall* em uma única medida, fornecendo uma medida de desempenho geral do modelo. Ele é especialmente útil quando há um desequilíbrio entre as classes do conjunto de dados.

A fórmula do *F1 Score* é a seguinte:

$$\text{F1 Score} = 2 * \frac{precision * recall}{precision + recall}$$

O F1 Score varia de 0 a 1, sendo 1 o melhor desempenho possível. Ele considera tanto a capacidade do modelo de evitar falsos positivos (*Precision*) quanto a capacidade de encontrar todos os verdadeiros positivos (*Recall*). Dessa forma, é uma métrica útil para encontrar um equilíbrio entre a precisão e a revocação.

3 METODOLOGIA

Esta seção apresenta a metodologia da pesquisa realizada para atingir os objetivos apresentados neste trabalho.

3.1 Dados utilizados

Conforme mencionado anteriormente, o conjunto de dados utilizado neste trabalho é oriundo do Sistema de Informação sobre Nascidos Vivos (SINASC, 2023). Em particular, utilizamos dados referentes apenas ao município de Fortaleza/CE, por ser a cidade na qual a instituição de ensino, referente a este trabalho, está situada e por ser a cidade em que o autor deste trabalho reside. Foram utilizados apenas dados do ano de 2020. Este conjunto possui um total de 40.944 registros e 60 atributos[1]. A descrição de cada atributo pertencente ao conjunto de dados utilizado é mostrada na Tabela 2.

Para este conjunto de dados, queremos predizer a variável "PARTO", a qual contém a informação de que o parto ocorreu por meio de cesariana ou normal. É importante observar, apesar de só utilizarmos dados da mãe para a construção de modelos de predição, quais dados específicos da mãe são completos e podem fornecer informações valiosas para a classificação da via de parto. Para isso, é necessário realizar o pré-processamento e a limpeza dos dados.

3.2 Pré-processamento e limpeza dos dados

O pré-processamento de dados consiste em um conjunto de técnicas utilizadas para preparar e limpar os dados antes de serem analisados, onde podemos incluir várias etapas, como a remoção de dados duplicados ou incompletos, o tratamento de valores faltantes, a normalização dos dados e a redução de dimensionalidade.

Neste trabalho inicialmente verificamos quais atributos possuíam valores nulos para remoção do conjunto de dados. A Figura 1 apresenta um mapa de calor dos atributos e seus respectivos valores nulos. Verificamos uma grande presença de dados nulos para os atributos: "CODANOMAL", "DTRECORIGA", "SERIESCMAE", "IDADEPAI", "QTDFILVIVO", "QTDFILMORT", "RACACOR", "RACACORMAE"e "DTULTMENSTR". Consequentemente, tais atributos foram removidos do conjunto de dados.

Em seguida, verificamos quais atributos possuíam valores únicos que não seriam de

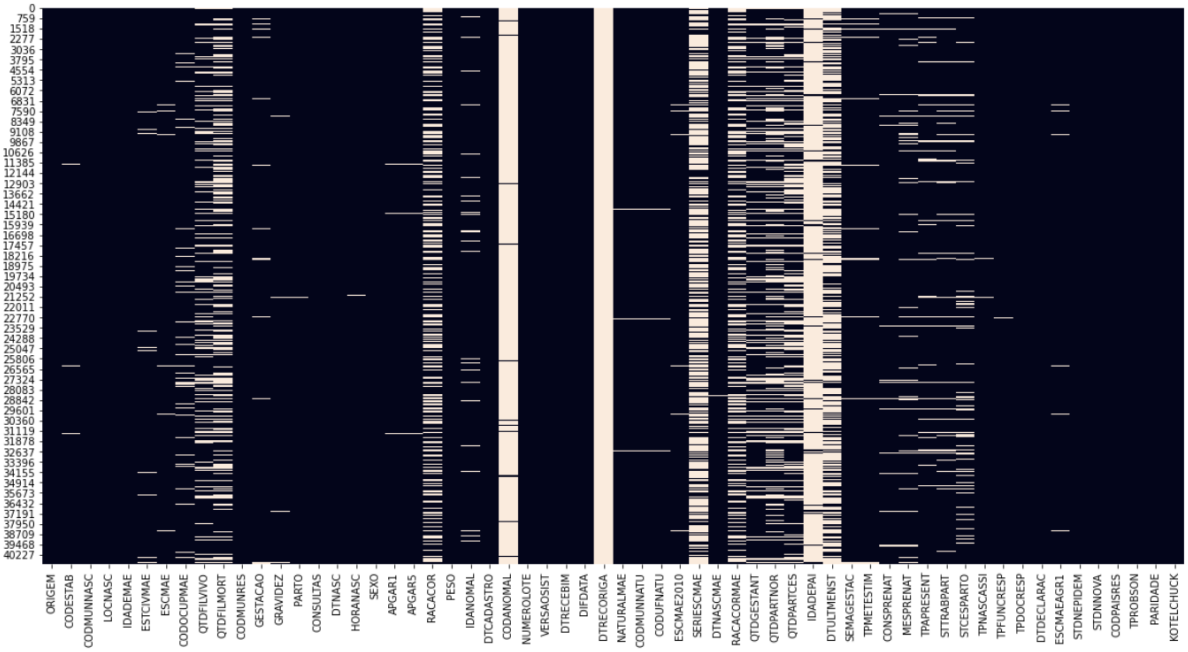
[1] https://dados.fortaleza.ce.gov.br/dataset/epidemiologia

Tabela 2 – Descrição de cada atributo pertencente ao conjunto de dados.

ORIGEM	Sem descrição
CODESTAB	Código de estabelecimento
CODMUNNASC	Município de ocorrência
LOCNASC	Local de ocorrência do nascimento
IDADEMAE	Idade da mãe em anos
ESTCIVMAE	Estado civil
ESCMAE	Escolaridade da mãe
CODOCUPMAE	Ocupação da mãe
QTDFILVIVO	Número de filhos vivos
QTDFILMORT	Número de filhos mortos
CODMUNRES	Município de residência da mãe
GESTACAO	Semanas de gestação
GRAVIDEZ	Tipo de gravidez: única, dupla, etc.
PARTO	Tipo de parto: vaginal ou cesáreo.
CONSULTAS	Número de consultas de pré-natal
DTNASC	Data do nascimento
HORANASC	Horário de nascimento
SEXO	Sexo do RN
APGAR1	Apgar no primeiro minuto
APGAR5	Apgar no quinto minuto
RACACOR	Raça/Cor do RN
PESO	Peso do RN
IDANOMAL	Anomalia congênita
DTCADASTRO	Data do cadastro
CODANOMAL	Código de malformação congênita
NUMEROLOTE	Número do lote
VERSAOSIST	Versão do sistema
DTRECEBIM	Data de recebimento no nível central
DIFDATA	Diferença entre data de óbito e data do recebimento
DTRECORIGA	Data do 1o recebimento do lote
NATURALMAE	Se a mãe for estrangeira, contém o código do país
CODMUNNATU	Código do município de naturalidade da mãe
CODUFNATU	Código da UF de naturalidade da mãe
ESCMAE2010	Escolaridade da mãe em 2010
SERIESCMAE	Série escolar da mãe.
DTNASCMAW	Data de nascimento da mãe
RACACORMAE	Raça/cor da mãe
QTDGESTANT	Número de gestações anteriores
QTDPARTNOR	Número de partos vaginais
QTDPARTCES	Número de partos cesáreos
IDADEPAI	Idade do pai
DTULTMENST	Data da última menstruação
SEMAGESTAC	Número de semanas de gestação
TPMETESTIM	Método utilizado: exame físico ou outro método
CONSPRENAT	Número de consultas pré-natal
MESPRENAT	Mês de gestação em que iniciou o pré-natal
TRAPRESENT	Tipo de apresentação do RN: cefálico, pélvico, etc.
STTRABPART	Trabalho de parto induzido?
STCESPARTO	Cesárea ocorreu antes do trabalho de parto iniciar?
TPNASCASSI	Nascimento foi assistido por?
TPFUNCRESP	Tipo de função do responsável pelo preenchimento.
TPDOCRESP	Tipo do documento do responsável.
DTDECLARAC	Data da declaração.
ESCMAEAGR1	Escolaridade da mãe em 2010 agregada.
STDNEPIDEM	Status de DO Epidemiológica.
STDNNOVA	Status de DO Nova.
CODPAISRES	Código do país de residência.
TPROBSON	Código do Grupo de Robson, gerado pelo sistema.
PARIDADE	Sem descrição.
KOTELCHUCK	Sem descrição.

Fonte: (SINASC, 2023)

Figura 1 – Mapa de calor dos atributos e seus respectivos valores nulos (em cor branca).



Fonte: Elaborado pelo autor.

importância para o treinamento dos modelos por meio de histogramas. Histogramas permitem verificar a distribuição de frequência de dados para identificação de padrões importantes. Assim, exibimos o histograma de cada atributo remanescente da etapa anterior, conforme Figura 2. Dessa forma, os atributos: "ORIGEM", "LOCNASC", "CODMUNNASC", "VERSAOSIST", "STDNEPIDEM", "STDNNOVA"e "CODPAISRES"foram removidos por possuírem apenas um valor único de registro.

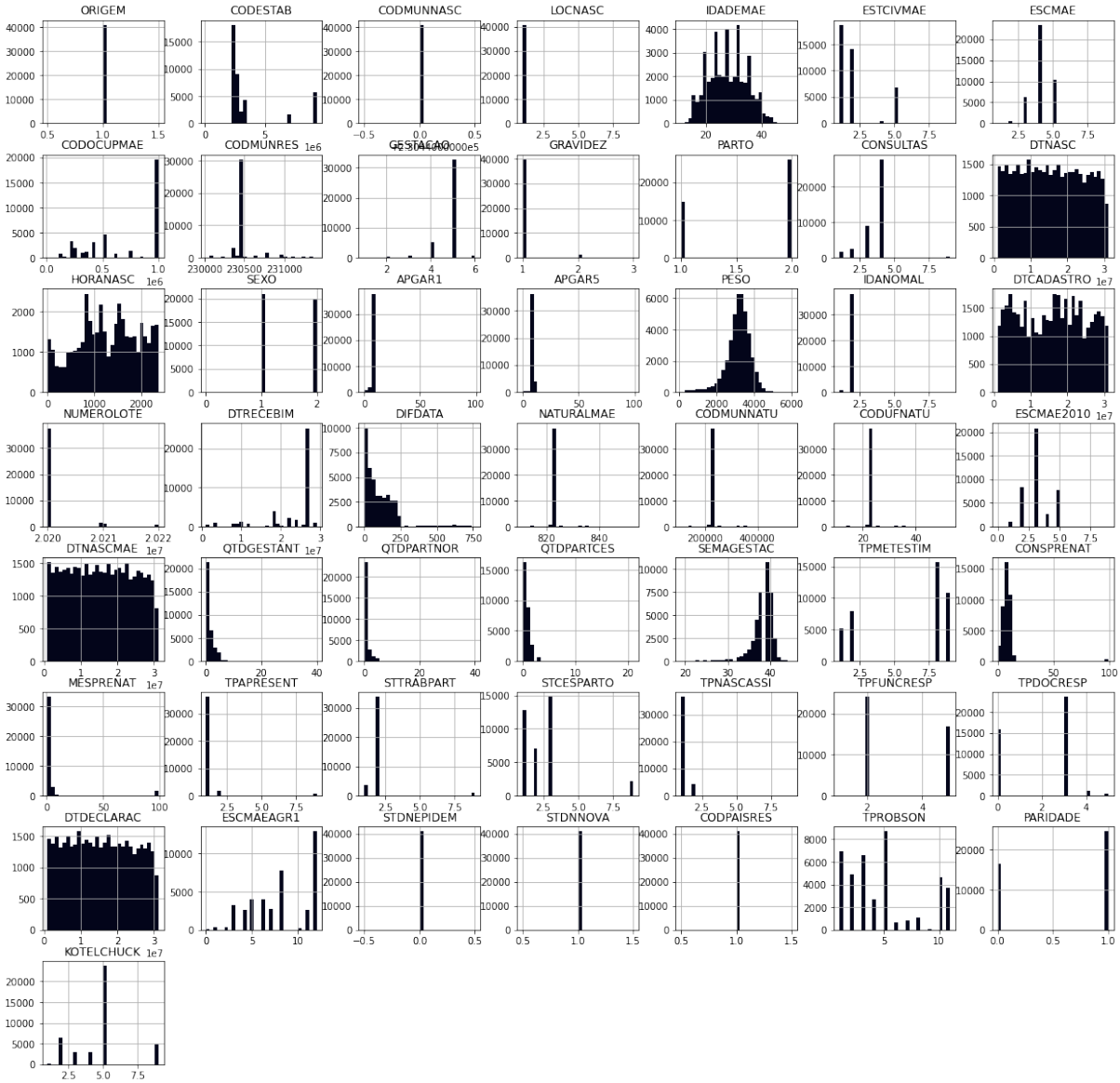
Após a remoção de atributos nas etapas iniciais, o conjunto de dados permaneceram com 40.944 registros e 44 atributos.

3.3 Seleção de atributos

Após a limpeza dos dados, realizamos a etapa de seleção de quais atributos da mãe seriam utilizados neste trabalho. A seleção de atributos pode ser realizada de diversas maneiras, dependendo da técnica utilizada e das características dos dados. Neste trabalho utilizamos o conhecimento do domínio, também conhecido como *domain knowledge*, para seleção de atributos (GROVES, 2013).

Com base na hipótese deste trabalho, apenas atributos referentes à mãe devem ser selecionados. Dessa forma, os atributos selecionados foram: "QTDPARTCES", "QTDPARTNOR", "IDADEMAE", "CODOCUPMAE", "ESCMAE"e "CONSULTAS".

Figura 2 – Histogramas dos atributos remanescentes do conjunto de dados.



Fonte: Elaborado pelo autor.

Vale ressaltar que, para os atributos "QTDPARTCES"e "QTDPARTNOR", que são importantes segundo a etapa de seleção de atributos, alguns registros possuíam valores nulos. Esses registros foram preenchidos com o número zero. Para os demais atributos que possuíam valores nulos, estes foram removidos do conjunto de dados.

Finalmente, o conjunto de dados final ficou com os seis atributos selecionados mais o atributo alvo "PARTO", com um total de 40.887 registros e 7 variáveis. Da classe cesárea, o conjunto de dados apresenta 63.97% dos dados, enquanto os outros 36.03% correspondem a registros da classe normal. Um exemplo do conjunto de dados utilizado para construção dos modelos é mostrado na Figura 3.

Figura 3 – Exemplo de dados utilizados para construção dos modelos.

	QTDPARTCES	QTDPARTNOR	IDADEMAE	CODOCUPMAE	ESCMAE	CONSULTAS	PARTO
0	1.00	2.00	24	999992.00	4.00	3	2.00
1	1.00	2.00	21	513205.00	3.00	2	2.00
2	2.00	0.00	31	999992.00	3.00	4	2.00
3	0.00	2.00	28	763015.00	3.00	4	2.00
4	0.00	2.00	18	999992.00	3.00	2	1.00
...	...	...	...	...	...	...	...
40882	0.00	2.00	43	763010.00	4.00	4	2.00
40883	0.00	3.00	38	631010.00	3.00	2	2.00
40884	0.00	2.00	25	354705.00	3.00	4	1.00
40885	1.00	0.00	18	999992.00	3.00	2	2.00
40886	3.00	0.00	25	999992.00	3.00	1	2.00

Fonte: Elaborado pelo autor.

3.4 Modelagem

A etapa de modelagem consistiu na aplicação de vários algoritmos de aprendizagem de máquina para tratar problemas de classificação. São eles: *XGBClassifier; LGBMClassifier; AdaBoostClassifier; BernoulliNB; LogisticRegression; RandomForestClassifier; ExtraTreesClassifier; BaggingClassifier; NuSVC; NearestCentroid; LinearDiscriminantAnalysis; LinearSVC; SGDClassifier; RidgeClassifier; RidgeClassifierCV; CalibratedClassifierCV; SVC; KNeighborsClassifier; LabelPropagation; DecisionTreeClassifier; PassiveAggressiveClassifier; ExtraTreeClassifier; Perceptron; QuadraticDiscriminantAnalysis; GaussianNB* e *DummyClassifier*.

Os conjuntos de dados foram divididos em treino e teste, utilizando a proporção 90% e 10%, respectivamente. Os resultados para cada modelo são apresentados no próximo capítulo.

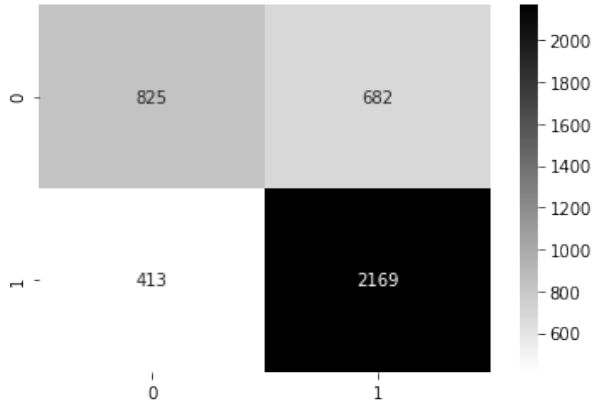
4 RESULTADOS

Este trabalho foi desenvolvido utilizando a linguagem de programação Python, SO windows 10, 64 bits, 4 cores de *CPU*, memória *RAM* de 8GB e armazenamento de dados em *SSD*. Para a execução dos modelos, utilizamos a biblioteca Python *LazyPredict* (PATEL, 2020). Ela é uma ferramenta útil para simplificar e acelerar o processo de modelagem e avaliação de algoritmos de aprendizagem de máquina, além de fornecer uma interface simples e intuitiva para treinar vários modelos de aprendizado de máquina com apenas algumas linhas de código. A principal finalidade da biblioteca *LazyPredict* é automatizar o fluxo de trabalho de modelagem de dados, permitindo aos usuários experimentar rapidamente diversos modelos sem a necessidade de escrever um código extenso para cada algoritmo.

Os resultados provenientes da execução dos modelos a partir dessa biblioteca são exibidos na Figura 5. Observa-se que o classificador *AdaBoost* apresentou os melhores resultados em comparação aos outros classificadores testados, em termos de acurácia e *F1 Score*. O tempo de execução para treinamento desse modelo com os dados testados foi baixo também, de apenas 0,70 segundos.

Em particular, a matriz de confusão para esse classificador pode ser visualizada na Figura 4, apresentando uma análise detalhada do desempenho do modelo na classificação dos partos vaginais (classe 0) e cesáreos (classe 1). Ao examinar a matriz, podemos observar que o modelo apresenta um bom desempenho na identificação dos partos cesáreos, refletido pelos valores verdadeiros positivos registrados nessa categoria. No entanto, é notável que o modelo ainda carece de melhorias para corretamente identificar os partos normais (classe 0), visto que os valores verdadeiros negativos nessa categoria são relativamente baixos.

Figura 4 – Matriz de confusão para o classificador *AdaBoost*.



Fonte: Elaborado pelo autor.

Figura 5 – Resultados obtidos a partir dos modelos testados.

Modelo	Acurácia	F1 Score	Tempo de execução
AdaBoostClassifier	0.73	0.73	0.70
SVC	0.73	0.72	86.11
XGBClassifier	0.73	0.72	1.28
LGBMClassifier	0.73	0.72	0.45
NuSVC	0.72	0.71	321.95
LabelPropagation	0.72	0.71	341.08
LogisticRegression	0.72	0.71	0.14
CalibratedClassifierCV	0.72	0.71	6.26
BernoulliNB	0.72	0.72	0.06
LinearSVC	0.72	0.70	3.64
QuadraticDiscriminantAnalysis	0.72	0.72	0.11
LinearDiscriminantAnalysis	0.71	0.70	0.42
GaussianNB	0.71	0.71	0.03
RidgeClassifierCV	0.71	0.69	0.08
RidgeClassifier	0.71	0.69	0.06
SGDClassifier	0.70	0.67	0.12
RandomForestClassifier	0.70	0.70	2.85
NearestCentroid	0.70	0.70	0.05
ExtraTreesClassifier	0.70	0.69	2.12
BaggingClassifier	0.70	0.69	0.44
KNeighborsClassifier	0.69	0.68	1.30
ExtraTreeClassifier	0.69	0.69	0.05
DecisionTreeClassifier	0.68	0.68	0.07
Perceptron	0.67	0.68	0.11
PassiveAggressiveClassifier	0.65	0.65	0.13
DummyClassifier	0.63	0.49	0.02

Fonte: Elaborado pelo autor.

Por fim, os resultados em termos de *precision*, *recall* e *F1 Score* para o classificador *AdaBoost* são mostrados na Tabela 3.

Tabela 3 – Resultados em termos de precisão, *Recall* e*F1-score*.

Classe	Precision	Recall	F1-Score
0	0,69	0,61	0,64
1	0,79	0,84	0,81

5 CONCLUSÕES E TRABALHOS FUTUROS

Este trabalho investigou a aplicação de técnicas de aprendizado de máquina para classificar a via de um parto com base apenas em informações da mãe.

Ao utilizar técnicas de aprendizado de máquina, este estudo buscou desenvolver um modelo capaz de analisar informações da mãe como idade, quantidade de partos, escolaridade, para fazer a classificação da via de parto. O classificador *AdaBoost* se destacou ao ser avaliado utilizando diversas métricas de desempenho, mostrando sua eficácia na tarefa de classificação. Sua acurácia girou em torno de 73%, indicando que o modelo conseguiu realizar previsões corretas em boas proporções utilizando apenas dados da mãe do bebê.

Como trabalhos futuros, pretende-se explorar técnicas e métodos para melhorar a interpretabilidade das predições. Dessa forma, espera-se explorar abordagens que combinem vários modelos explicáveis, como árvores de decisão, onde as decisões de cada modelo individual são ponderadas e combinadas para chegar a uma decisão final. Além disso, planeja-se realizar comparações entre diferentes modelos explicáveis, como árvores de decisão, regras de associação, redes bayesianas, entre outros, para avaliar seu desempenho na predição da via de parto e na interpretabilidade dos resultados.

Por fim, é importante reconhecer que a classificação da via de parto é um processo complexo, que envolve uma interação complexa entre diversos fatores, incluindo o bem-estar fetal, condições específicas do trabalho de parto e outros elementos contextuais relevantes, como a escolha da própria mãe pela via desejada. Dessa forma, é essencial considerar que a utilização exclusiva de dados maternos pode apresentar limitações na obtenção de uma classificação abrangente e precisa.

REFERÊNCIAS

BARBOSA, G. P.; GIFFIN, K.; ANGULO-TUESTA, A.; GAMA, A. d. S.; CHOR, D.; D'ORSI, E.; REIS, A. C. G. V. d. Parto cesáreo: quem o deseja? em quais circunstâncias? **Cadernos de Saúde Pública**, SciELO Brasil, v. 19, p. 1611–1620, 2003.

BITTAR, O. J. N.; BICZYK, M.; SERINOLLI, M. I.; NOVARETTI, M. C. Z.; MOURA, M. M. N. de. Sistemas de informação em saúde e sua complexidade. **Revista de Administração em Saúde**, v. 18, n. 70, 2018.

BRASIL. **Lei Geral de Proteção de Dados Pessoais (LGPD)**. 2018. Disponível em: <http://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/L13709.htm>. Acesso em: 05 de junho de 2023.

CARVALHO, A.; FACELI, K.; LORENA, A.; GAMA, J. **Inteligência Artificial–uma abordagem de aprendizado de máquina**. [S.l.]: LTC, 2011.

DORÇA, F. A. *et al.* Uma abordagem estocástica baseada em aprendizagem por reforço para modelagem automática e dinâmica de estilos de aprendizagem de estudantes em sistemas adaptativos e inteligentes para educação a distância. Universidade Federal de Uberlândia, 2012.

FERREIRA, K. M.; MACHADO, L. V.; MESQUITA, M. do A. Humanização do parto normal: uma revisão de literatura/humanization normal child birth: a review of literature. **Saúde em Foco**, v. 1, n. 2, p. 134–148, 2014.

GROVES, W. Using domain knowledge to systematically guide feature selection. In: CITESEER. **Twenty-Third International Joint Conference on Artificial Intelligence**. [S.l.], 2013.

HOSSIN, M.; SULAIMAN, M. A review on evaluation metrics for data classification evaluations. **International Journal of Data Mining & Knowledge Management Process**, Academy & Industry Research Collaboration Center (AIRCC), v. 5, n. 2, p. 1, 2015.

MONARD, M. C.; BARANAUSKAS, J. A. Conceitos sobre aprendizado de máquina. **Sistemas inteligentes-Fundamentos e aplicações**, v. 1, n. 1, p. 32, 2003.

PATEL, S. **LazyPredict: Exploring and comparing multiple models for classification and regression in Python**. 2020. <https://github.com/shankarpandala/lazypredict>. Acesso em: 05 de junho de 2023.

ROSSI, R. G. **Classificaçao automática de textos por meio de aprendizado de máquina baseado em redes**. Tese (Doutorado) — Universidade de São Paulo, 2016.

SAMMUT, C.; WEBB, G. I. **Encyclopedia of machine learning and data mining**. [S.l.]: Springer Publishing Company, Incorporated, 2017.

SINASC. **Sistema de Informações sobre Nascidos Vivos**. 2023. Disponível em: https://svs.aids.gov.br/daent/cgiae/sinasc/. Acesso em: 05 de junho de 2023.

WEIDLE, W. G.; MEDEIROS, C. R. G.; GRAVE, M. T. Q.; BOSCO, S. M. D. Escolha da via de parto pela mulher: autonomia ou indução? **Cadernos Saúde Coletiva**, SciELO Brasil, v. 22, p. 46–53, 2014.

ZHOU, Z.-H. **Machine learning**. [S.l.]: Springer Nature, 2021.