ISSN: 2317-0840

Determinantes e predição de parto cesáreo utilizando técnicas de aprendizado de máquina

Tabi Thuler Santos[1], Paulo Henrique Ferreira[2†]

[1] Universidade de São Paulo; Instituto de Ciências Matemáticas e de Computação; MBA em Ciências de Dados; São Carlos – São Paulo, Brasil.
[2] Universidade Federal da Bahia; Instituto de Matemática e Estatística; Departamento de Estatística; Salvador – Bahia, Brasil.

Resumo: *As evidências de que a realização de partos cesáreos é mundialmente pautada por fatores socioeconômicos e de que essa prática pode aumentar os riscos para a saúde materna e infantil impõem uma discussão global sobre a utilização adequada das cesarianas. Este trabalho se propõe a contribuir com essa discussão com a busca pelos fatores mais importantes na determinação de parto cesáreo nos Estados Unidos em 2019 e com o desenvolvimento de modelos de predição de tipo de parto para o pré-natal. Foram testadas as performances preditivas dos algoritmos Random Forest e k-Nearest Neighbors em subconjuntos dos dados do National Center for Health Statistics. Os determinantes de parto cesáreo foram estudados com o uso de regressão logística a partir de variáveis selecionadas pela Random Forest. A análise dos resultados mostrou, por exemplo, que a intervenção de aumento do parto, a apresentação cefálica do feto, a quantidade de partos anteriores de nascidos vivos e a assistência de enfermeira ou obstetriz reduzem a probabilidade de parto cesáreo; enquanto que as faixas de idade mais altas da mãe (a partir de 35 anos), o parto gemelar, a obesidade da puérpera e a existência de cesáreas prévias aumentam a chance de o parto ser uma cesariana. Os resultados obtidos vão de encontro com os estudos existentes na literatura para o caso americano. Dentre as principais contribuições do presente trabalho à literatura está o enfoque econômico, pautando as técnicas e algoritmos de aprendizado de máquina em prol da discussão global sobre a utilização adequada de cesarianas.*

Palavras-chave: *Aprendizagem de máquina, classificação, modelos de predição, parto cesáreo, tomada de decisão.*

Determinants and prediction of cesarean delivery using machine learning techniques

Abstract: *Evidences that performing cesarean deliveries is guided worldwide by socioeconomic factors and that this practice can increase risks to maternal and child health imposes a global discussion on the proper use of cesarean sections. The aim of this study is to contribute to this discussion through: (i) identification of the most important factors for determining cesarean deliveries in the United States in 2019; and (ii) development of a model to predict the mode of delivery for prenatal care. The predictive performances of the algorithms Random Forest and k-Nearest Neighbors were tested on subsets of the National Center for Health Statistics data. The determinants of cesarean delivery were studied using logistic regression using features selected by Random Forest. The analysis of the results showed, for instance, that the intervention of labor augmentation, the cephalic presentation of the fetus, the number of previous live births, and the assistance of a nurse or midwife reduce the probability of a cesarean delivery; whereas higher maternal age groups (from 35 years old), twin births, maternal obesity, and previous cesarean sections increase the chance of the delivery being a cesarean section. The results reinforce the literature for the American case. Among the main contributions of this study to the literature is the economic focus, guiding machine learning techniques and algorithms in support of the global discussion on the appropriate use of cesarean sections.*

Keywords: *Machine learning, classification, prediction model, cesarean delivery, decision making.*

[†] Autor correspondente: paulohenri@ufba.br

Manuscrito recebido em: 08/04/2024
Manuscrito revisado em: 07/10/2024
Manuscrito aceito em: 08/10/2024

Introdução

Segundo a Organização Mundial da Saúde (OMS), partos cesáreos são cada vez mais frequentes em todo o mundo. A literatura indica que este aumento parece estar relacionado a fatores socioeconômicos, a despeito dos clínicos, e que essa prática pode aumentar os riscos negativos para a saúde materna e infantil. Tem-se, portanto, uma preocupação global, médica e governamental, em direcionar esforços para a utilização adequada da cesariana, principalmente através do repasse de informações adequadas às parturientes e da utilização de protocolos clínicos padronizados (WHO, 2015).

Neste contexto, a partir de dados sobre partos norte-americanos realizados em 2019, pretende-se aplicar técnicas de classificação estatística e de aprendizado de máquina para encontrar os fatores mais importantes na decisão do tipo de parto e ajustar o melhor modelo preditivo para o tipo de parto.

Com este estudo, procura-se então responder às seguintes perguntas:

1) Quais são os fatores que determinam a escolha por parto cesáreo atualmente?

2) Como a prática está alinhada às recomendações clínicas?

3) Quão assertiva pode ser a predição do tipo de parto a partir de fatores clínicos e demográficos?

Portanto, este trabalho se propõe a contribuir com a discussão global sobre a utilização adequada das cesarianas, com a busca pelos fatores mais importantes na determinação de parto cesáreo nos Estados Unidos em 2019 e com o desenvolvimento de um modelo de predição de tipo de parto para o pré-natal. A ideia é que o modelo de predição auxilie no processo de decisão do tipo de parto, tornando a tomada de decisão mais pautada por dados e orientando melhor todos os agentes: provedores, governo e pacientes. Por exemplo, o sistema de saúde pode atuar mais fortemente na orientação dos casos que, durante o pré-natal, o modelo classifica como futuros partos cesáreos. Assim, haveria um enfrentamento mais efetivo às cesáreas eletivas/desnecessárias, focando os esforços de orientação.

O trabalho está organizado da seguinte forma: a Seção "Revisão da Literatura" resume os trabalhos relacionados mais recentes, a Seção "Metodologia" descreve a metodologia empregada, a Seção "Aplicação" expõe os resultados da aplicação aos dados reais e a Seção "Considerações Finais" finaliza com considerações sobre este trabalho de pesquisa.

Revisão da Literatura

Segundo Ullah *et al.* (2021), as técnicas de classificação de aprendizado de máquina mais utilizadas na literatura para predição de tipos de parto são: árvores de decisão (DT), florestas aleatórias (RF), AdaBoost, *Support Vector Machines* (SVM), k-vizinhos mais próximos (k-NN) e *Naive Bayes*. Este estudo recente comparou a performance de modelos de predição de tipo de parto nas bases de dados original e ampliada pela técnica de sobreamostragem SMOTE (do inglês *Synthetic Minority Oversampling Technique*). Os algoritmos utilizados foram DT, RF, AdaBoost, *bootstrap* (ou *bagging*) e k-NN sobre variáveis preditoras de sintomas e sinais de pré-natal (idade, número de gravidezes, maturidade fetal, pressão sanguínea e problemas no coração) de uma base pública disponibilizada pela Universidade da Califórnia, Irvine, no repositório de *Machine Learning* (*Caesarian Section Classification Dataset*, https://archive.ics.uci.edu/ml/datasets/Caesarian+Section+Classification+Dataset. Segundo os autores, o número de observações é insuficiente para predição (80 partos, sendo 34 cesáreos), o que justifica a utilização da técnica SMOTE. Para avaliação da performance foi utilizada validação

cruzada com 10 *folds*, curva ROC (do inglês *Receiver Operating Characteristic*), estatística kappa, matriz de confusão e suas métricas (acurácia, precisão, sensibilidade, especificidade e medida F1). Os autores concluíram que a técnica SMOTE é uma importante estratégia na busca por modelos preditivos de tipo de parto, dado que os melhores resultados foram encontrados para a base ampliada. O k-NN foi o algoritmo que atingiu melhores resultados para a estatística kappa (0,68) e acurácia (84%). O atributo mais influente para a realização de cesariana foi a presença de doença crônica cardíaca na parturiente (identificado por testes de correlação).

Sobre a mesma base de dados, Rahman *et al.* (2021) aplicaram validação cruzada com 10 *folds* e sete classificadores de aprendizado de máquina - k-NN, DT, SVM, RF, *Gradient Boosting*, *Extreme Gradient Boosting* (XGBoost) e *Extreme Learning Machine* - após enfrentar o desbalanceamento dos dados com sobreamostragem aleatória, SMOTE e ROSE (do inglês *Random Over-Sampling Examples*). Os algoritmos foram avaliados pelas métricas: acurácia, medida F1, média geométrica, sensibilidade, especificidade, área sob a curva ROC, taxa de alarme falso e taxa de perda sobre os resultados da matriz de confusão. A eficiência dos resultados de cada algoritmo nas quatro bases foi comparada utilizando o teste de Friedman. O objetivo do estudo era comparar o desempenho dos algoritmos em predizer o risco de cesarianas considerando dados escassos e desbalanceados. De forma geral, os resultados indicaram o k-NN como o melhor classificador de partos cesáreos, alcançando as melhores performances. A técnica RF também registrou boa performance. Olhando separadamente para as bases, nenhum algoritmo alcançou nível satisfatório de acurácia na base original (desbalanceada), sendo o maior o do k-NN (68%). Os resultados melhoraram apenas marginalmente para a base gerada por sobreamostragem aleatória. Para a base gerada por SMOTE, os resultados melhoraram sensivelmente, com o k-NN atingindo 95% de acurácia e de medida F1. Por fim, para a base gerada por ROSE, os resultados foram semelhantes, porém o maior nível atingido para a acurácia foi de 90%.

Islam *et al.* (2021) trouxeram o foco da temática para as variáveis, conduzindo uma pesquisa para encontrar os atributos mais relevantes para predizer o tipo de parto. O questionário foi montado a partir das variáveis indicadas pela literatura e por especialistas selecionados, totalizando 111 variáveis, sendo aplicado posteriormente a um conjunto de 21 médicos. As variáveis mais frequentes nas respostas, 32 ao total, foram ordenadas em três grupos prioritários por suas correlações com a variável dependente (*univariate feature selection*), ficando 11 variáveis no primeiro e no segundo grupo e 10 no terceiro. O primeiro grupo continha altura, peso, índice de massa corporal (IMC), indicador de pré-natal, ganho de peso na gestação, comorbidades, abortos, número de gestações anteriores, idade gestacional e cardiotocografia fetal. Os três grupos foram então combinados para gerar mais quatro grupos, sendo três uma combinação dois a dois e o último a combinação dos três grupos. Os modelos de predição foram propostos com DT, RF, SVM, k-NN e classificação *stacking*, minimizando o número de variáveis e maximizando a acurácia. Foram utilizados dados abertos de partos de 2014 em quatro hospitais públicos espanhóis. Como resultado, os autores mostraram que os modelos performaram muito bem para o conjunto de variáveis selecionadas. Considerando apenas o grupo com as 11 variáveis principais, o algoritmo *stacking* alcançou o melhor resultado, com 84,8% na medida F1. O modelo de melhor performance do artigo utilizou o grupo de variáveis mais completo (32 variáveis) e o classificador *stacking* e superou o modelo de melhor performance de toda a literatura consultada ao registrar 97,9% para a medida F1.

Kavitha & Balasubramanian (2021) compararam a eficácia dos algoritmos k-NN, SVM e C5.0 para obter a melhor predição de parto utilizando o coeficiente kappa, sensibilidade, especificidade e acurácia. Foram selecionados os seis atributos mais relacionados a cesarianas, de acordo com a literatura, de 33 atributos coletados de 2021 parturientes em hospitais públicos e privados, entre 2015 e 2017: idade, altura da mãe, gestação múltipla, peso ao nascer, nível de líquido amniótico no momento do parto e parada de progressão do trabalho de parto. O algoritmo

C5.0 apresentou a maior acurácia (100%), porém para dados de treinamento. Esse resultado pode ser uma indicação de sobreajuste do modelo. Seria importante que o estudo tivesse incluído outras formas de validação, como validação cruzada, para que a performance preditiva dos modelos pudesse ser testada em dados desconhecidos ao treinamento.

Outros trabalhos que merecem destaque e citação são: Alam *et al.* (2021), Abbas *et al.* (2020), Kowsher *et al.* (2020), Desyani *et al.* (2020).

Dentre as principais contribuições do presente trabalho à literatura está o enfoque econômico, pautando as técnicas e algoritmos de aprendizado de máquina em prol da discussão global sobre a utilização adequada de cesarianas. A base de dados utilizada se diferencia dos estudos prévios e reflete essa abordagem, por ser robusta e relevante ao tema. Apesar da escolha clássica pelos algoritmos, a utilização da RF como seletor de variáveis para a análise de determinantes de partos cesáreos também se diferencia dos trabalhos apresentados nesta seção.

Metodologia

Como visto na seção anterior, os métodos de aprendizado de máquina mais frequentes na literatura de predição de tipo de parto são: DT, RF, AdaBoost, SVM, k-NN e *Naive Bayes*. Os estudos relatados mostram melhores resultados com os algoritmos RF e k-NN, sendo portanto os algoritmos escolhidos para a predição de partos cesáreos neste trabalho. Para o estudo dos determinantes de partos cesáreos será utilizada a regressão logística. O k-NN e a regressão logística serão aplicados sobre os atributos selecionados previamente pelo algoritmo de RF.

As técnicas utilizadas serão brevemente descritas nas subseções seguintes, construídas tendo o trabalho de Faceli *et al.* (2021) como referência principal.

Aprendizado de Máquina

Random Forests

O algoritmo de *Random Forests* combina a predição de várias árvores de decisão para gerar a predição final. A combinação de estimadores, em geral, aumenta o desempenho preditivo e reduz a variância em relação aos estimadores isolados, deixando o desempenho mais estável e reduzindo a possibilidade de um superajuste do modelo aos dados de treinamento (*overfittting*). A classificação ocorre por votação, ou seja, a moda das predições das árvores para cada objeto.

As árvores são induzidas usando *bagging*. Isso significa que elas utilizam amostras diferentes dos dados de treinamento. Por ser uma amostragem com reposição, podem conter dados duplicados e possuem o mesmo tamanho da base de treinamento. Os estimadores são independentes, pois as árvores são rodadas paralelamente.

Árvores de decisão dividem recursivamente os dados em busca de subconjuntos homogêneos. Para isso, utilizam a estratégia de divisão e conquista: com a divisão do espaço em duas partes, simplificam o problema complexo em problemas mais simples que são resolvidos localmente. Medidas de impureza (entropia, Gini etc.) dos subconjuntos podem ser utilizadas como critério de parada para a busca, definindo o nível de impureza aceitável. É composta por: raiz, onde é feita a primeira partição dos dados e, portanto, onde encontra-se o atributo mais importante do modelo; nós internos, onde são feitos os testes de atributo que resultam em novas partições; e nós externos ou folhas, onde estão as previsões feitas pela árvore.

Além de amostragem com reposição para os dados de treinamento, em RF as árvores também são induzidas com subamostragem local dos atributos preditivos. A cada tentativa de

partição dos dados, em cada nó, um novo subconjunto de atributos é amostrado (sem reposição) e o melhor ponto dentre o melhor dos atributos é selecionado para criar o nó de decisão. A amostragem evita a dominância total de atributos, dando oportunidade a atributos que podem ser importantes em menor escala, o que colabora para a melhora do desempenho do algoritmo.

A fragilidade do algoritmo está em sua sensibilidade à definição dos hiperparâmetros do modelo. Os hiperparâmetros podem ser escolhidos por métodos de otimização e os mais importantes definem a quantidade de árvores que comporá a RF e o número máximo de atributos preditivos que comporá cada uma das árvores.

As árvores de decisão e as florestas aleatórias podem ser usadas para selecionar as variáveis do modelo através do ordenamento da importância de cada atributo na predição. Cada árvore utiliza um subconjunto de atributos e, a cada nó, busca o atributo que melhor separa as classes naquele ponto (reduzindo a impureza do nó). Ao armazenar essas informações, o algoritmo de RF possibilita o acesso à média e ao desvio-padrão da métrica de redução da impureza utilizada, para cada atributo, considerando todas as árvores que o compõe. Portanto, a importância do atributo pode ser medida como o desempenho obtido por ele na redução da impureza, e o ordenamento dessa medida normalizada facilita a visualização das importâncias relativas.

A seleção de atributos a partir dessa importância pode seguir alguns critérios de corte definidos pelo pesquisador, como por exemplo, importância relativa maior do que a média das importâncias de todos os atributos. O pesquisador deve estar atento à quantidade de variáveis retiradas do modelo, pois uma grande quantidade poderia incorrer em grande de perda de informação; e que características consideradas como fundamentais pelos especialistas ou pela literatura devem ser retiradas com cautela, mesmo que o modelo não encontre importância relevante.

k-Nearest Neighbors

O k-NN é um algoritmo simples, muito utilizado, que considera que objetos de classes semelhantes tendem a se agrupar no espaço de entrada. O algoritmo então aprende com a proximidade entre os objetos e assim os classifica com base na informação local. Como envolve o cálculo de distância entre objetos, o k-NN é sensível à escala dos atributos, que devem ser padronizados para evitar ponderações implícitas aos atributos decorrentes da escala. Ainda na mesma linha, há sensibilidade do algoritmo a *outliers* e atributos irrelevantes. A seleção de atributos é recomendada também para reduzir o alto tempo computacional de classificação, dado que o algoritmo não aprende um modelo dos dados de treino, mas sim classifica de acordo com a moda da classe dos k objetos do conjunto de treinamento mais próximos (algoritmo preguiçoso/*lazy*). A ausência de um modelo explícito torna o k-NN não interpretável.

De forma geral, o k-NN apresenta bons resultados preditivos para vários conjuntos de dados. Apesar disso, a escolha do número de vizinhos, hiperparâmetro do modelo, pode alterar substancialmente o resultado. A medida de distância utilizada também pode influenciar o desempenho preditivo do algoritmo. Uma das alternativas é testar diferentes valores de k e tipos de medidas diferentes pelo método de validação cruzada, que será descrito na Subseção “Métodos de Validação”.

Regressão Logística

A regressão logística é um método de modelagem estatística muito utilizado, que alia interpretabilidade dos resultados a baixo custo computacional. Mais especificamente, é um caso

particular de modelos lineares generalizados, com variável de resposta com distribuição binomial e com ligação canônica logito.

Portanto, um modelo de regressão logística múltipla aplicado a um problema de classificação binária retornará uma resposta:

$$Y_i = \begin{cases} 1, \textit{se a } i-\textit{ésima observação pertence à classe } 1, \\ 0, \textit{caso contrário}, \end{cases}$$

para $i = 1, \ldots, n$, sendo n o tamanho amostral.

A probabilidade de uma observação i pertencer à classe 0 ou 1 depende do vetor de atributos independentes x_i e é definida como:

$$P(Y_i = 1 \vee x_i) = \pi(x_i),$$
$$P(Y_i = 0 \vee x_i) = 1 - \pi(x_i).$$

A função de ligação canônica é a transformação logarítmica da fração dessas probabilidades, conhecida como razão de chances (em inglês, *odds ratio*):

$$\eta_i = \log\left(\frac{\pi(x_i)}{1-\pi(x_i)}\right) = x_i^T \beta.$$

Considerando as estimativas dos parâmetros, pode-se então definir a razão de chances como:

$$\log\left(\frac{\hat{\pi}(x_i)}{1-\hat{\pi}(x_i)}\right) = x_i^T \hat{\beta} \Rightarrow \frac{\hat{\pi}(x_i)}{1-\hat{\pi}(x_i)} = \exp\left\{x_i^T \hat{\beta}\right\}.$$

Os parâmetros são estimados no enfoque clássico por estimadores de máxima verossimilhança, encontrados por métodos iterativos como o de mínimos quadrados.

Métodos de Validação

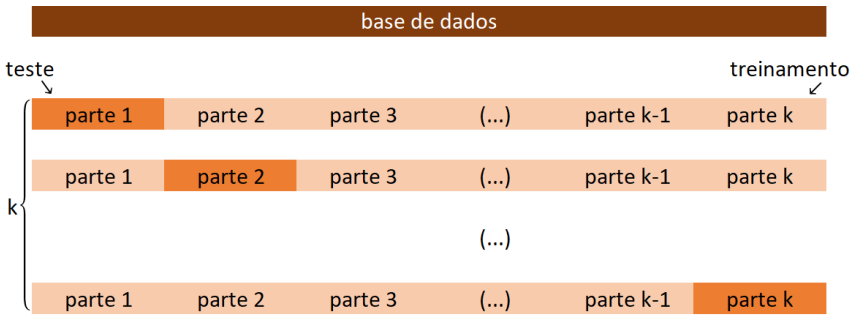
A escolha do melhor modelo envolve a busca pela redução tanto da variância, para que seja generalizável (funcione para dados inéditos), quanto do viés, para que o erro de predição seja o menor possível. Muito utilizada nos estudos da área de predição de tipos de parto, como visto na revisão da literatura (Seção "Revisão da Literatura"), a validação cruzada tem como objetivo validar o aprendizado do modelo treinado, evitando subajuste ou superajuste aos dados de treinamento.

No método *hold-out*, os dados são separados em conjuntos de treinamento e de teste. Os dados de treino são utilizados para o algoritmo induzir o modelo preditivo, enquanto os dados de teste, inéditos ao modelo, servem para avaliar o desempenho com as métricas de avaliação.

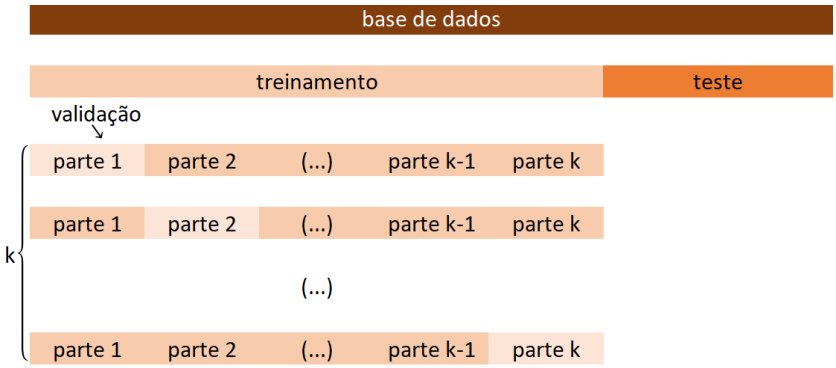
Porém, o desempenho final é influenciado por como a partição dos dados é realizada. Caso o *hold-out* seja realizado N vezes, a cada partição os conjuntos extraídos aleatoriamente para treino e teste variam, o que faz variar a métrica de desempenho.

O método de validação cruzada *k-fold* pode ser utilizado para avaliar a performance do modelo, melhorando os resultados do *hold-out* ao variar o conjunto de teste. Os dados são divididos em k partes de mesmo tamanho e k rodadas são realizadas - em geral, os valores mais utilizados para k são 5, *default* de alguns pacotes e 10, conforme visto na Seção "Revisão da Literatura". A cada rodada, uma das k partes é separada das demais como conjunto de teste e as k - 1 partes restantes formam o conjunto de treinamento, conforme ilustrado na Figura 1(a). O desempenho do modelo rodado pode ser sintetizado como a média dos resultados de teste das k rodadas.

Figure 1: Division of data in the *k-fold* cross-validation method.



(a) Para avaliação da performance do modelo.



(b) Para calibragem de hiperparâmetros.

Source: from the authors (2024).

Diferentemente dos parâmetros do modelo, que são aprendidos durante o treinamento, os hiperparâmetros do modelo são definidos pelo pesquisador. Algoritmos de aprendizado de máquina podem ser muito sensíveis aos hiperparâmetros, como o k-NN. Por isso, antes de avaliar a performance final é interessante utilizar o método *k-fold* de validação cruzada também para calibragem dos hiperparâmetros.

A busca pelos melhores hiperparâmetros começa com a divisão dos dados de treino em k partes de igual tamanho. A cada rodada, uma das k partes é separada das demais para validação, conforme mostra a Figura 1(b). Os hiperparâmetros são definidos, um modelo é induzido das k - 1 partes e posteriormente avaliado na parte dos dados que ficou de fora do treinamento. Após a realização de k rodadas, cada parte terá sido utilizada como validação 1 vez e como treino k - 1 vezes. Assim, o modelo com a melhor média de desempenho no conjunto de validação indicará quais hiperparâmetros devem ser selecionados.

É importante ressaltar que a comparação de performance entre os classificadores só será justa se o conjunto de treinamento, o conjunto de teste e a(s) métrica(s) de avaliação utilizados forem os mesmos para todos.

Métricas de Avaliação

Uma das formas de avaliar o desempenho dos algoritmos é organizar os resultados em uma matriz de confusão, que relaciona as predições com os valores reais de classificação. Na Tabela 1 é possível observar a matriz com os acertos (verdadeiros positivos, ou VP, e verdadeiros negativos, ou VN) e os erros (falsos positivos, ou FP, e falsos negativos, ou FN) de predição.

Table 1: Confusion matrix for a binary classifier.

		classes preditas	
		1	0
classes	1	VP	FN
verdadeiras	0	FP	VN

Source: from the authors (2024).

Uma das métricas mais utilizadas para avaliação de modelos de predição de tipos de parto é a acurácia. Derivada da matriz de confusão, ela mede a proporção dos acertos na predição (verdadeiros positivos e negativos) dentre todas as predições feitas pelo modelo (verdadeiros positivos e negativos, mais falsos positivos e negativos). A acurácia é definida como:

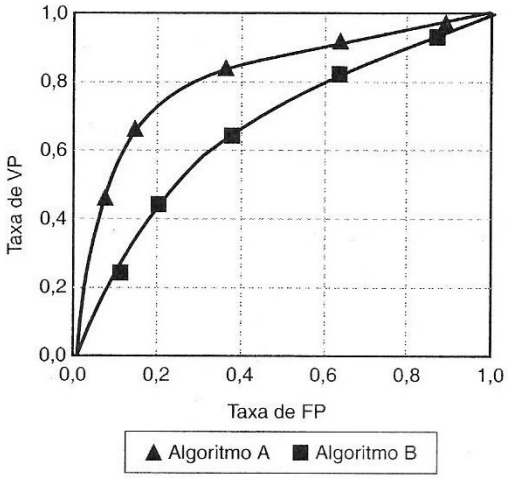
$$Acurácia = \frac{VP + VN}{VP + VN + FP + FN}.$$

Segundo Ullah *et al.* (2021), a curva ROC é amplamente utilizada na tomada de decisão em saúde: "*ROC curves are highly useful for establishing the classifiers and envisioning their performance and are commonly used in health care decision making because it visualizes the entire scenario (...) and is considered an effectual measure of inherent validity of a diagnostic test*".

A curva ROC relaciona custo e benefício de um modelo. Seguindo essa análise, o classificador ideal combinaria o maior benefício possível, com taxa de VP = 100%, com o menor custo possível, ou seja, nenhum FP. Portanto, o modelo que mais se aproxima desse cenário é o melhor classificador. Como um exemplo fictício da comparação gráfica, pode-se observar que as curvas ROC da Figura 2 indicam melhor desempenho para o algoritmo A que para o B.

A área sob a curva ROC, AUC (do inglês *Area Under the Curve*), mede o grau de separação entre as classes que o modelo consegue atingir, variando entre 0 e 1. Quando assume o valor 1, o modelo é capaz de distinguir perfeitamente entre as classes; quando 0,5, o modelo não é capaz de distinguir, pois tem o mesmo desempenho que um classificador aleatório; e quando não há área (AUC = 0), o modelo é perfeitamente incapaz, errando todas as predições (trocando as classes). Assim, quanto maior a AUC, melhor a performance do modelo avaliado.

Figure 2: Example of a ROC curve.



Source: Faceli *et al.* (2021).

Também utilizado para avaliar os modelos encontrados neste trabalho, o *Matthews Correlation Coefficient* (MCC) é um coeficiente de correlação entre as classes preditas e reais. Quando não há correlação (MCC = 0), a classificação é aleatória. Valores de MCC iguais a 1 e -1 indicam correlações perfeitas direta e inversa, respectivamente. Quanto maior a correlação, melhor o classificador (Baldi *et al.*, 2000). O MCC é definido da seguinte forma:

$$MCC = \frac{(VP \times VN) - (FP \times FN)}{\sqrt{(VP+FN)(VP+FP)(VN+FP)(VN+FN)}}.$$

Especificamente para o modelo de regressão logística foi utilizado o critério AIC (do inglês *Akaike Information Criterion*). A ideia do método de Akaike de seleção de modelos é encontrar a combinação entre um bom ajuste e um número reduzido de parâmetros. Para que haja uma melhora no ajuste, ou seja, uma redução no AIC, é preciso que o aumento de complexidade do modelo com o aumento do número de parâmetros seja compensado por uma redução considerável oriunda da primeira parcela da equação, que contém o logaritmo da função de verossimilhança, conforme abaixo:

$$AIC = -2l(\hat{\beta}) + 2p,$$

em que $l(\hat{\beta})$ é o logaritmo da verossimilhança maximizada e p é o número de parâmetros (Paula, 2023).

Linguagem de Programação e *Hardware*

Os algoritmos foram executados em linguagem Python, na versão 3.7, utilizando o pacote *scikit-learn* (Pedregosa *et al.*, 2011) para RF e k-NN, e o módulo *statsmodels* para regressão logística (Seabold & Perktold, 2010).

A configuração do *hardware* é de processador Intel64 Family 6 Model 142 Stepping 9 GenuineIntel ~400 MHz, com memória física total de 8,090 MB e memória virtual máxima de 18,842 MB.

Aplicação

Conjunto de Dados

A base de dados utilizada contém microdados públicos de natalidade (NCHS, 2020). O *National Center for Health Statistics* (NCHS) disponibiliza cerca de 120 características dos 3,75 milhões de partos realizados em 2019, no Estados Unidos, com informações sobre saúde materna, saúde do bebê, utilização de serviços de saúde, características geográficas e características demográficas dos pais.

Destas, foram selecionadas 55 características (54 preditoras e 1 resposta) para este estudo, considerando os objetivos do estudo e a revisão da literatura descrita na Seção "Revisão da Literatura", que são apresentadas na Tabela 2.

Table 2: Characteristics of the selected births for the study.

(continue)

dia do parto	-	parto no final de semana
utilização de serviço de saúde	-	período do dia do parto | parto no hospital
demográficas	mãe	idade | nacionalidade | local de residência e do parto | raça | origem hispânica | escolaridade
	-	paternidade reconhecida | mãe casada
	pai	idade | raça | origem hispânica | escolaridade
	partos anteriores	filhos falecidos | ordenamento do parto dentre os de nascidos vivos | ordenamento dentre todos os partos | tempo após o último nascido vivo
utilização de serviços de saúde	pré-natal	trimestre de início | mais do que 10 consultas | programa de nutrição suplementar
saúde materna	-	fumante (antes e durante a gravidez) | altura – *outlier* | índice de massa corporal | ganho de peso durante a gravidez
	fatores de risco da gravidez	diabetes (pré e gestacional) | hipertensão (pré e gestacional) | eclâmpsia
	fatores de risco da gravidez	partos prematuros prévios | tratamento de infertilidade | cesariana prévia | número de cesarianas prévias | sem risco
	-	infecções

Table 2: Characteristics of the selected births for the study.

(continued)

utilização de serviços de saúde	parto e trabalho de parto	manobra de versão cefálica externa | trabalho de parto induzido | trabalho de parto prolongado | esteroides | antibióticos | infecção intra-amniótica | anestesia
	método do parto	posição cefálica | método | trabalho de parto atendido | cesariana | acompanhamento profissional
	-	mãe transferida de outras unidades | forma de pagamento
saúde do bebê	-	múltiplos | gestação combinada | idade gestacional
	-	peso ao nascer | condições anormais | anomalia congênita

Source: from the authors (2024).

Pré-processamento

Durante o tratamento dos dados algumas variáveis foram recategorizadas com o objetivo de facilitar a estimação, a interpretação e a comparabilidade dos resultados. O primeiro passo foi mapear as alterações necessárias para a transformação de atributos em variáveis categóricas numéricas. Depois foi feito o tratamento de dados faltantes (*missing data*), em duas etapas:

i) exclusão de observações com muitos dados faltantes;

ii) imputação do valor da moda para os dados faltantes restantes.

A etapa (i) foi derivada da investigação sobre ausência de dados na base, que revelou um padrão de concentração nas variáveis paternas; sendo assim, foram excluídos os cerca de 840 mil partos que não possuíam essas informações (aproximadamente 22% do total de partos). O próximo passo foi analisar a distribuição dos dados para que categorias menos frequentes ou de menor relevância para o estudo fossem agregadas. Por último, foram removidas as duplicatas de observações. Ao final do processo, restavam ainda quase 2,9 milhões de partos na base de dados.

O tamanho da base de dados incorreria em um alto custo computacional, especialmente na busca por hiperparâmetros e no ajuste do algoritmo k-NN. Para que esse alto custo não se tornasse impeditivo, foram feitos testes com subamostras aleatórias das observações. O método *RandomUnderSampler* da biblioteca *imbalanced-learn* permitiu também o balanceamento das subamostras na variável-alvo, mantendo-se as proporções nos demais atributos (Lemaître *et al.*, 2017). A comparação entre as métricas de teste para o modelo *default* de RF aplicado com validação cruzada de $k = 10$ a subamostras de 1 milhão, 200 mil e 100 mil observações mostrou que os resultados variavam pouco, o que reforçou a escolha pela subamostra de 100 mil observações.

Na última etapa foi aplicado o método *hold-out* com 30% dos dados reservados para o conjunto de teste, ou seja, 30 mil observações.

Análise Exploratória

Esta subsubseção apresenta uma breve descrição dos 70 mil partos que compõem a base de treino através das suas características mais relevantes.

Assim como na literatura estudada, a idade das parturientes se distribui de forma diferente entre os tipos de parto. A Figura 3(a) mostra uma associação positiva entre idade e cesarianas, visto que, conforme aumenta a idade da mãe, também aumenta a proporção de partos cesáreos.

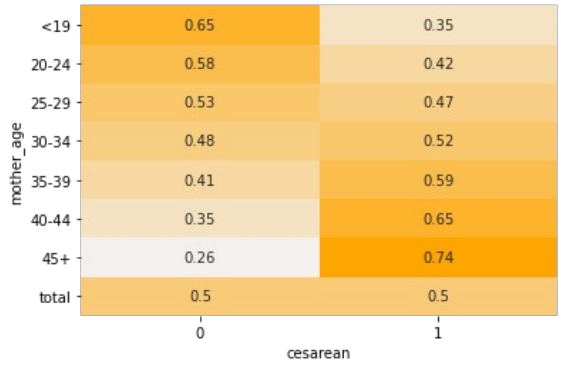
A escolaridade da parturiente também se mostra uma importante característica para predição de cesáreas. Conforme ilustrado na Figura 3(b), quase 60% das parturientes na menor faixa de escolaridade passaram por partos vaginais. Esse percentual decai com o avanço das faixas e estabiliza abaixo da média após o título de bacharel.

Um ponto importante é a investigação da relação entre educação e idade, dado que são atributos potencialmente relacionados. A distribuição cruzada entre essas variáveis na Figura 4 parece indicar alguma relação positiva entre algumas das faixas mais baixas, mas não pode se dizer o mesmo para todas as faixas. A análise dos atributos de idade e educação do pai tem resultado análogo.

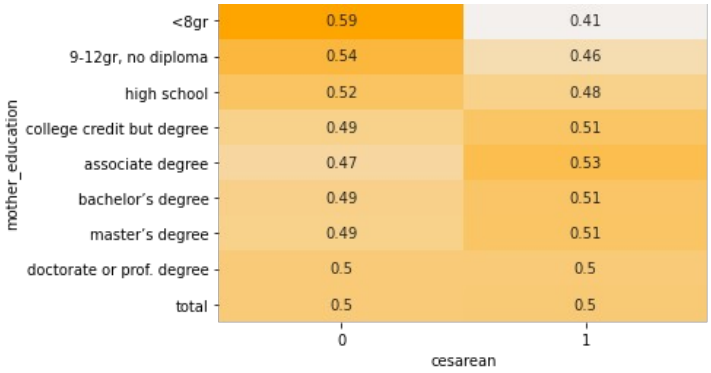
A literatura defende que a utilização de cesarianas deve estar associada ao risco do parto, conforme visto na Seção “Revisão da Literatura”. Um dos atributos que pode ser entendido como uma *proxy* para esse risco é o que relaciona estados e municípios de ocorrência do parto e de residência da mãe. A hipótese é de que um parto realizado fora da localização da residência teria maior probabilidade de ser um parto de emergência ou de a parturiente ter passado por transferências entre hospitais. A distribuição dos dados reforça essa hipótese, pois a proporção de cesarianas é de 49% para os partos que ocorrem no mesmo local de residência, contra 52% dos que ocorrem em estados diferentes e 53% para estrangeiros.

Segundo a NCHS (2020), partos de baixo risco estão associados a partos de primíparas (primeiro parto), únicos (não gemelares), a termo (idade gestacional de pelo menos 37 semanas) e cefálicos (posicionamento do feto no momento do parto). Alguns atributos denotados como importantes para os modelos preditivos deste estudo se relacionam diretamente ou indiretamente a esses fatores, como o número de partos já realizados pela parturiente, o de nascidos vivos nos partos anteriores, o tempo decorrido desde o último parto de nascido vivo e a própria posição cefálica do feto no parto. Essas características são exploradas nos próximos parágrafos.

Figure 3: Distribution.

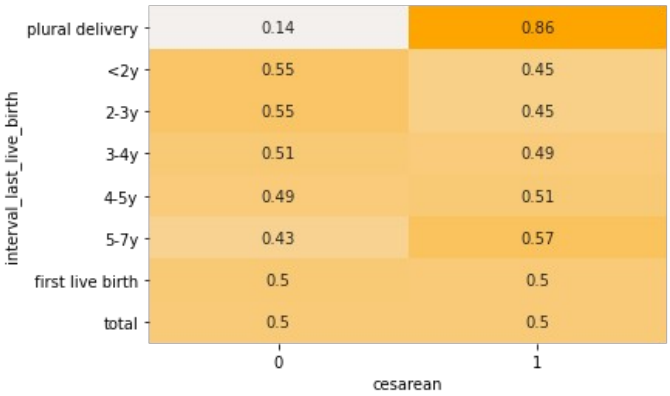


(a) Idade da parturiente por tipo de parto.

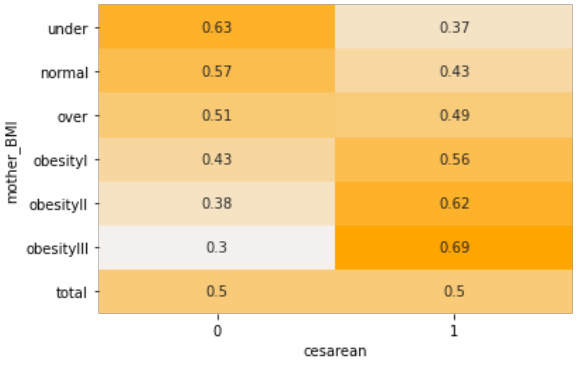


(b) Nível de escolaridade da parturiente por tipo de parto.

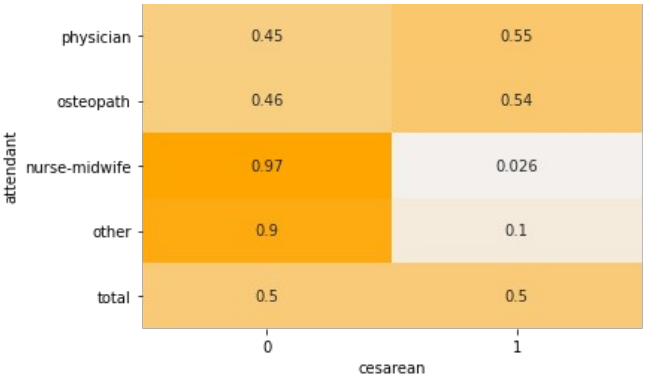
(c) Ordenamento do parto por tipo de parto.



(d) Intervalo entre o parto atual e o último parto de nascido vivo por tipo de parto.



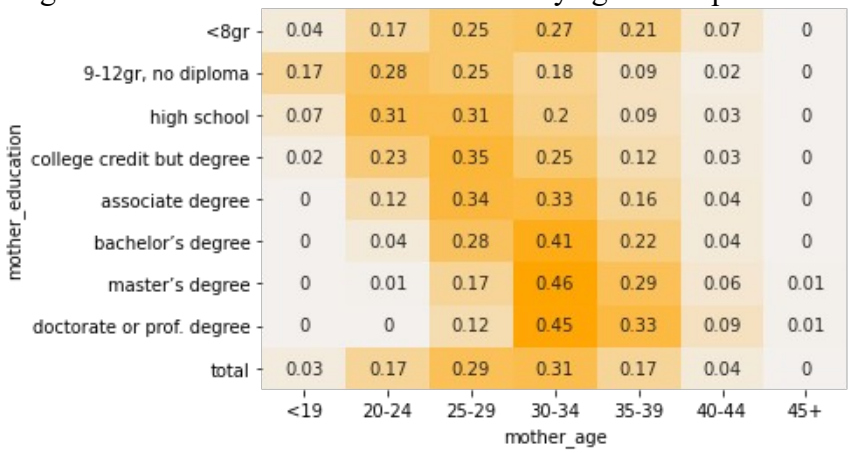
(e) IMC da parturiente por tipo de parto.



(f) Perfil do profissional de assistência ao parto por tipo de parto.

Source: from the authors (2024).

Figure 4: Distribution of education level by age of the parturient.



Source: from the authors (2024).

As cesáreas são menos frequentes entre as primíparas. Passam a ser mais frequentes quando a puérpera está entre o segundo e o quarto parto, mas voltam a cair a partir do quinto parto, como mostra a Figura 3(c).

O atributo de intervalo entre o parto anterior de nascido vivo e o atual categoriza os partos gemelares. Nos partos dessa primeira categoria, 86% são cesáreos (Figura 3(d)). Há indício de associação positiva entre o tempo decorrido e a utilização de cesarianas para os partos únicos, exceto pela categoria de primíparas, na qual não há distinção entre partos cesáreos e vaginais. Foi investigada a associação desse atributo com a idade da parturiente, mas não foram encontrados indícios de correlação.

Sobre a posição do feto, a diferença entre os tipos de parto é relevante: 94% dos partos em que o feto não estava na posição cefálica foram cesáreos, contra 47% nos casos de partos com fetos na posição cefálica. Portanto, há indícios de que a realização de cesarianas nos partos estudados possui relação com maior risco conforme a classificação da NHCS, dado que há prevalência de cesáreas em partos de multíparas (parturientes com filhos), gemelares e não cefálicos.

O risco também pode ser visto do prisma da gravidez. O atributo que identifica a presença de fatores de risco na gravidez (diabetes, hipertensão, eclâmpsia, partos prematuros anteriores, tratamento de infertilização e cesáreas prévias) também indica relação com o tipo de parto: o percentual de cesarianas passa de 34% entre gravidezes sem risco para 74% quando há pelo menos um fator de risco associado à gravidez.

Outro fator da saúde da mulher importante no acompanhamento da gestação é o IMC. A Figura 3(e) mostra a associação positiva entre o IMC e a taxa de partos cesáreos, que passam a ser maioria quando a parturiente é categorizada nas faixas de obesidade. O ganho de peso na gravidez também se mostra importante nesse estudo, sendo as cesarianas predominantes nas extremidades (pouco e muito ganho de peso).

As características do trabalho de parto mais relevantes neste estudo tratam de partos induzidos e aumentados e da administração de antibióticos para a parturiente durante o trabalho de parto. Quando necessárias, a indução e o aumento são intervenções realizadas para que o trabalho de parto evolua favoravelmente, evitando assim o uso desnecessário de cesarianas (WHO, 2015). Os dados mostram importante diferença no percentual de partos cesáreos entre os que passaram ou não por essas duas intervenções. O percentual de cesáreas passa de 56% para 34% na presença de indução e de 56% para 26% na presença de aumento. Quanto à administração de antibióticos, enquanto 45% dos partos são cesarianas quando não há essa medicação, 63% são cesarianas para o grupo de parturientes medicadas.

A literatura investiga a influência das características do profissional de assistência ao parto no tipo de parto (Santos, 2011). Nesse estudo, o tipo de profissional se mostrou importante nos modelos preditivos. A Figura 3(f) mostra a diferença dos tipos de parto entre os profissionais, com representação de apenas 2% de cesarianas entre os partos assistidos por enfermeiras obstétricas e obstetrizes.

Resultados

Modelos Preditivos

O primeiro passo foi a otimização da RF com a busca pelos melhores valores para os hiperparâmetros mais relevantes segundo a literatura, que são o número de árvores para compor a floresta e o número máximo de atributos que as árvores devem considerar em cada nó. Os valores encontrados foram de 600 árvores e da raiz quadrada do número total de atributos preditivos (no caso, $\sqrt{54}$). O resultado do modelo ajustado para esses hiperparâmetros possibilitou a primeira análise de importância dos atributos para a predição.

Os dados foram então divididos em subconjuntos de acordo com a distribuição da importância relativa dos atributos:

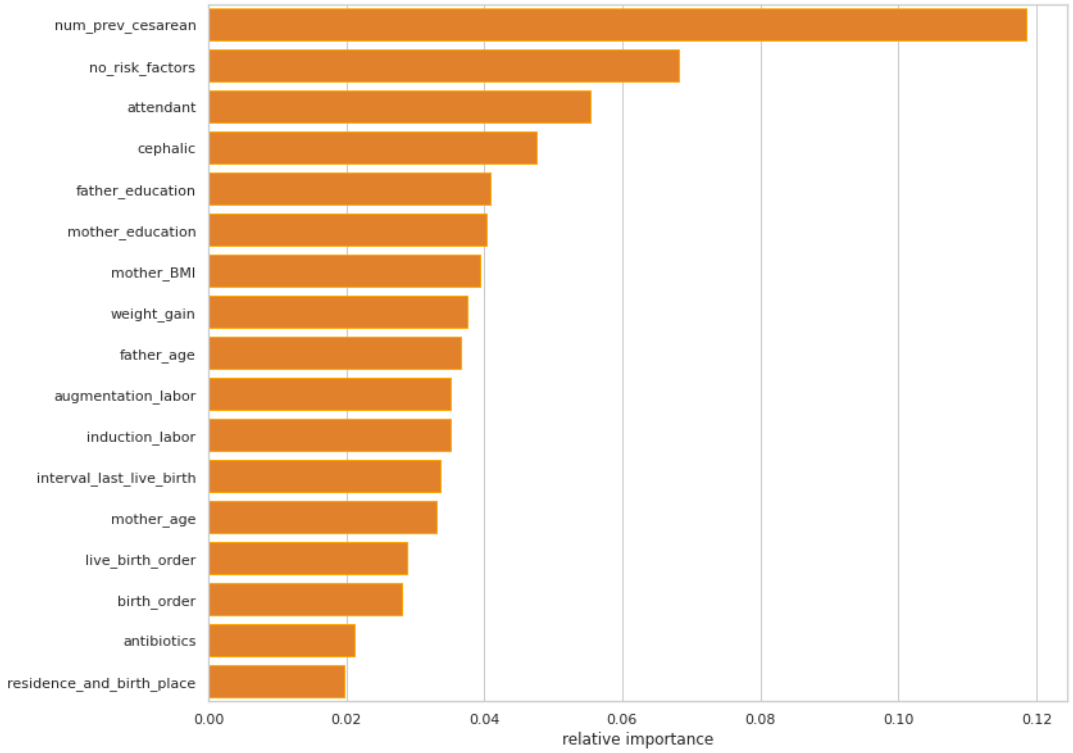
i) importância acima do primeiro decil (48 atributos);

ii) importância acima do primeiro quartil (40 atributos);

iii) importância acima da média (17 atributos).

A Figura 5 apresenta os atributos selecionados desse último subconjunto.

Figure 5: Attributes selected by mean, in order of importance.



Source: from the authors (2024).

Assim como para a RF, foi feita uma busca pelos melhores hiperparâmetros para o k-NN com o total de atributos. Para o número de vizinhos, a função de ponderação dos vizinhos e o parâmetro da distância de Minkowski, foram encontrados os valores de 19 vizinhos a ser ponderados pela distância (vizinhos mais próximos têm maior influência) e o parâmetro 2, o que torna a medida de Minkowski equivalente à euclidiana.

Os algoritmos RF e k-NN foram ajustados aos quatro conjuntos de dados (total de atributos e subconjuntos) e os resultados são expostos na Tabela 3. O algoritmo de DT foi incluído como mais uma referência de *baseline* para os modelos. As melhores métricas foram alcançadas pelo algoritmo RF, em especial, o modelo com hiperparâmetros otimizados e utilizando todos os atributos disponíveis, alcançando cerca de 81% de acurácia e AUC, e 0,62 no MCC.

Table 3: Performance of the predictive models.

	Acurácia	AUC	MCC
DT			
Default			
Total de atributos	0,7365	0,7365	0,4731
RF			
Default			
Total de atributos	0,8064	0,8063	0,6143
Otimizada			
Total de atributos	0,8096	0,8094	0,6203
Selecionados pelo decil	0,8089	0,8088	0,6192
Selecionados pelo quartil	0,8082	0,8080	0,6175
Selecionados pela média	0,7797	0,7796	0,5601
k-NN			
Default			
Total de atributos	0,7294	0,7293	0,4594
Otimizada			
Total de atributos	0,7544	0,7542	0,5111
Selecionados pelo decil	0,7541	0,7540	0,5105
Selecionados pelo quartil	0,7552	0,7551	0,5123
Selecionados pela média	0,7592	0,7590	0,5198

Source: from the authors (2024).

Fatores Determinantes pelo Modelo Logístico

A aplicação da regressão logística partiu do subconjunto de 17 atributos selecionados pela média de importância segundo a RF. Para melhor análise dos resultados e estimação, algumas categorias foram agregadas antes da transformação dos atributos em binários.

A busca pelo melhor modelo passou pela inclusão de intercepto e retirada gradual de blocos de *dummies* correspondentes a um atributo (método *backward*), analisando a significância marginal dos coeficientes estimados, pelo p-valor, e o ajuste do modelo aos dados pelo AIC.

Como resultado, a Tabela 4 (saída do Python) mostra as 14 variáveis que permaneceram no modelo, com suas categorias abertas em variáveis binárias, além do intercepto (*constant*). Dentre os atributos mais importantes, a intervenção de aumento do parto, a apresentação cefálica do feto, a quantidade de partos anteriores de nascidos vivos e a assistência de enfermeira ou obstetriz reduzem a probabilidade de parto cesáreo. Já as faixas de idade mais altas da mãe (a partir de 35 anos), o parto gemelar, a obesidade da puérpera e a existência de cesáreas prévias aumentam a chance de o parto ser uma cesariana. Mesmo que em menor medida, o uso de antibióticos, a existência de fatores de risco e o volume de peso ganho pela parturiente durante a gravidez também contribuem para o aumento da probabilidade de parto cesáreo. A Tabela 5 (saída do Python) apresenta a razão de chances com intervalo de confiança de 95% para todos os atributos do modelo.

Table 4: Results of the selected model for logistic regression (Python output).

(continue)

Dep. Variable:	y	No. Observations:	70.000
Model:	Logit	Df Residuals:	69.956
Method:	MLE	Df Model:	43
converged:	True	Pseudo R-squ.:	0,3877
Covariance Type:	nonrobust	Log-Likelihood:	-29.711,00
		LL-Null:	-48.520,00
		LLR p-value:	0,000

	coef	std err	z	P > |z|	[0,025, 0,975]
induction_labor	-0,6678	0,023	-29,195	0,000	[-0,713, -0,623]
augmentation_labor	-1,0433	0,026	-39,718	0,000	[-1,095, -0,992]
antibiotics	0,7521	0,023	32,693	0,000	[0,707, 0,797]
cephalic	-2,9526	0,065	-45,591	0,000	[-3,080, -2,826]
risk_factors	0,4127	0,026	15,891	0,000	[0,362, 0,464]
mother_age_20-24	0,2779	0,064	4,344	0,000	[0,153, 0,403]
mother_age_25-29	0,5531	0,065	8,480	0,000	[0,425, 0,681]
mother_age_30-34	0,7891	0,067	11,789	0,000	[0,658, 0,920]
mother_age_35-39	1,0595	0,071	14,970	0,000	[0,921, 1,198]
mother_age_40-44	1,3872	0,088	15,703	0,000	[1,214, 1,560]
mother_age_45+	1,4788	0,221	6,697	0,000	[1,046, 1,912]
mother_9-12gr_nodiploma	0,2093	0,080	2,624	0,009	[0,053, 0,366]
mother_highschool	0,1684	0,073	2,321	0,020	[0,026, 0,311]
mother_college-credit-but-degree	0,2109	0,073	2,871	0,004	[0,067, 0,355]

Table 4: Results of the selected model for logistic regression (Python output).

(continued)

mother_associatedegree	0,1549	0,078	1,991	0,046	[0,002, 0,307]
mother_bachelordegree	0,0636	0,074	0,861	0,389	[-0,081, 0,208]
mother_masterdegree	-0,0624	0,078	-0,803	0,422	[-0,215, 0,090]
mother_doc-or-profdegree	-0,1344	0,091	-1,476	0,140	[-0,313, 0,044]
father_age_20-44	0,0870	0,084	1,035	0,301	[-0,078, 0,252]
father_age_45-49	0,2650	0,106	2,499	0,012	[0,057, 0,473]
father_age_50+	0,2495	0,123	2,026	0,043	[0,008, 0,491]
live_birth_order_2	-0,5439	0,055	-9,810	0,000	[-0,653, -0,435]
live_birth_order_3	-1,0182	0,062	-16,517	0,000	[-1,139, -0,897]
live_birth_order_4+	-1,3996	0,066	-21,122	0,000	[-1,529, -1,270]
interval_last_live_birth_pluraldelivery	1,4379	0,112	12,857	0,000	[1,219, 1,657]
interval_last_live_birth_<2y	-0,7036	0,062	-11,304	0,000	[-0,826, -0,582]
interval_last_live_birth_2-3y	-0,7187	0,061	-11,774	0,000	[-0,838, -0,599]
interval_last_live_birth_3-5y	-0,5943	0,058	-10,168	0,000	[-0,709, -0,480]
interval_last_live_birth_5-7y	-0,3086	0,061	-5,024	0,000	[-0,429, -0,188]
mother_BMI_under	-0,2065	0,064	-3,236	0,001	[-0,331, -0,081]
mother_BMI_over	0,2742	0,026	10,752	0,000	[0,224, 0,324]
mother_BMI_obesityI	0,6111	0,031	19,772	0,000	[0,551, 0,672]
mother_BMI_obesityII	0,8929	0,041	22,017	0,000	[0,813, 0,972]
mother_BMI_obesityIII	1,1768	0,048	24,593	0,000	[1,083, 1,271]
weight_gain_11-20p	0,1348	0,043	3,130	0,002	[0,050, 0,219]
weight_gain_21-30p	0,2436	0,041	6,007	0,000	[0,164, 0,323]
weight_gain_31-40p	0,3027	0,042	7,175	0,000	[0,220, 0,385]
weight_gain_41-98p	0,5819	0,043	13,657	0,000	[0,498, 0,665]
num_prev_cesarean_1	2,7994	0,046	61,220	0,000	[2,710, 2,889]
num_prev_cesarean_2+	4,6676	0,106	43,986	0,000	[4,460, 4,876]
attendant_osteopath	-0,0088	0,035	-0,254	0,000	[-0,077, 0,059]
attendant_nurse-midwife	-3,6125	0,097	-37,209	0,000	[-3,803, -3,422]
attendant_other	-2,0801	0,127	-16,390	0,000	[-2,329, -1,831]
constant	1,7760	0,125	14,242	0,000	[1,532, 2,020]

Source: from the authors (2024).

Table 5: Odds ratios in the selected model for logistic regression (Python output).

	[0,025, 0,975]	Odds Ratio
induction_labor	[0,490327, 0,536325]	0,512810
augmentation_labor	[0,334605, 0,370895]	0,352283
antibiotics	[2,027935, 2,219310]	2,121466
cephalic	[0,045981, 0,059270]	0,052204
risk_factors	[1,435904, 1,589779]	1,510884
mother_age_20-24	[1,164775, 1,496762]	1,320375
mother_age_25-29	[1,529962, 1,975629]	1,738573
mother_age_30-34	[1,930719, 2,509944]	2,201362
mother_age_35-39	[2,511373, 3,314364]	2,885066
mother_age_40-44	[3,367000, 4,760222]	4,003457
mother_age_45+	[2,846282, 6,763804]	4,387675
mother_9-12gr_nodiploma	[1,054394, 1,441401]	1,232804
mother_highschool	[1,026559, 1,364277]	1,183432
mother_college-credit-but-degree	[1,069189, 1,425985]	1,234767
mother_associatedegree	[1,002442, 1,359746]	1,167504
mother_bachelordegree	[0,922049, 1,231584]	1,065636
mother_masterdegree	[0,806773, 1,094016]	0,939480
mother_doc-or-profdegree	[0,731282, 1,045069]	0,874208
father_age_20-44	[0,925243, 1,286202]	1,090894
father_age_45-49	[1,058808, 1,604452]	1,303383
father_age_50+	[1,008174, 1,633665]	1,283362
live_birth_order_2	[0,520716, 0,647121]	0,580488
live_birth_order_3	[0,320135, 0,407640]	0,361248
live_birth_order_4+	[0,216656, 0,280913]	0,246701
interval_last_live_birth_pluraldelivery	[3,382769, 5,244078]	4,211829
interval_last_live_birth_<2y	[0,437954, 0,558985]	0,494783
interval_last_live_birth_2-3y	[0,432433, 0,549334]	0,487391
interval_last_live_birth_3-5y	[0,492230, 0,618967]	0,551973
interval_last_live_birth_5-7y	[0,651167, 0,828435]	0,734472
mother_BMI_under	[0,717851, 0,921801]	0,813459
mother_BMI_over	[1,251335, 1,382891]	1,315469
mother_BMI_obesityI	[1,734178, 1,957542]	1,842478
mother_BMI_obesityII	[2,255644, 2,644316]	2,442261
mother_BMI_obesityIII	[2,953547, 3,562921]	3,243957
weight_gain_11-20p	[1,051673, 1,245021]	1,144271
weight_gain_21-30p	[1,178356, 1,381399]	1,275845
weight_gain_31-40p	[1,246119, 1,470225]	1,353542
weight_gain_41-98p	[1,646000, 1,945187]	1,789351
num_prev_cesarean_1	[15,026456, 17,976401]	16,435377
num_prev_cesarean_2+	[86,452936, 131,048223]	106,440141
attendant_osteopath	[0,926007, 1,061051]	0,991232
attendant_nurse-midwife	[0,022308, 0,032639]	0,026983
attendant_other	[0,097414, 0,160202]	0,124923
constant	[4,625673, 7,541849]	5,906448

Source: from the authors (2024).

As maiores diferenças nas chances de partos cesáreos e vaginais são vistas nos atributos de número de cesáreas prévias e de posição cefálica. Ter passado por um parto cesáreo anteriormente aumenta mais de 15 vezes a chance de o parto atual ser cesáreo em relação às parturientes que não tiveram nenhuma cesárea anteriormente. Já se a quantidade de cesáreas prévias é de duas ou mais, a chance de cesárea aumenta em mais de impressionantes 105 vezes. Para as parturientes em que o feto está em posição cefálica, a probabilidade de o parto ser cesáreo é quase 95% menor do que para os casos em que o feto se apresenta em outras posições.

De forma geral, os resultados estão em linha com o indicado pela literatura para os Estados Unidos (Seção "Revisão da Literatura") e com o encontrado na análise de importância de atributos descrita na Subsubseção "Modelos Preditivos".

Apesar de não ser possível comparar resultados de predição deste modelo de regressão logística aos modelos preditivos apresentados na subsubseção anterior, pois foram utilizados subconjuntos diferentes dos dados, mesmo assim as métricas de avaliação da predição foram calculadas apenas a título de curiosidade. A performance preditiva pode ser observada na Tabela 6. Comparado aos modelos *default* (utilizando o total de atributos preditivos) e *default* com a inclusão de intercepto, o modelo escolhido foi o que alcançou menores acurácia, AUC e MCC, com performance ligeiramente inferior à dos demais modelos.

Table 6: Predictive performance of the logistic regression models.

	Acurácia	AUC	MCC
Default	0,7896	0,7895	0,5814
Default com intercepto	0,7901	0,7899	0,5823
Escolhido	0,7893	0,7891	0,5807

Source: from the authors (2024).

Considerações Finais

Este trabalho se propôs a construir modelos de predição de parto cesáreo e estudar os seus determinantes com base nos dados do NCHS de 2019. Para os modelos preditivos, foram utilizados os algoritmos que usualmente obtêm melhores performances para partos: RF e k-NN. Para a análise dos fatores determinantes no tipo de parto, foi usada a regressão logística. Os resultados preditivos e de análise de fatores determinantes obtidos se mostraram em linha com a literatura estudada.

Porém, a comparação entre as métricas de teste para os recortes da base de dados mostra tendências diferentes entre os algoritmos preditivos. Quanto menor a quantidade de atributos, melhores as métricas de predição para o k-NN; no caso da RF, as métricas pioram quando a quantidade de atributos diminui. O resultado provavelmente tem relação com a suscetibilidade do k-NN à alta dimensionalidade e a relativa imunidade da RF, dado que na floresta em geral cada árvore utiliza apenas uma amostra do total de atributos. Um aspecto interessante seria investigar se o resultado seria robusto a outras formas de seleção de variáveis, mais indicadas para algoritmos baseados em proximidade, como por exemplo, o *SelectKBest* do *scikit-learn* (Pedregosa *et al.*, 2011).

Da mesma forma, poderiam ser testadas outras formas de amostragem alternativas à subamostragem aleatória realizada (*RandomUnderSampler*). Uma possibilidade seria o *CondensedNearestNeighbour*, também do *imbalanced-learn*, uma subamostragem baseada na comparação com o vizinho mais próximo (Lemaître *et al.*, 2017). Para a predição, poderiam ser

aplicados outros algoritmos além de k-NN e RF, como o AdaBoost, o SVM e o *Naive Bayes*. Esses algoritmos, como apresentado, são bastante utilizados na literatura de predição de cesárea.

O papel relevante do Brasil na discussão global sobre a utilização de partos cesáreos, por ser um dos países com as maiores proporções de cesarianas no mundo (Santos, 2011), aspira a aplicação desses modelos a dados brasileiros para uma comparação dos resultados obtidos e estudo de suas particularidades regionais.

Os códigos Python desenvolvidos neste trabalho estão disponíveis no GitHub do primeiro autor (endereço a ser informado posteriormente).

Referências

ABBAS, S. A.; REHMAN, A. U.; MAJEED, F.; MAJID, A.; MALIK, M. S. A.; KAZMI, Z. H.; ZAFAR, S. Performance analysis of classification algorithms on birth dataset. IEEE Access, v.8, p.102146–102154, 2020.

ALAM, M. S. B.; PATWARY, M. J. A.; HASSAN, M. Birth mode prediction using bagging ensemble classifier: A case study of Bangladesh. Em: 2021 International Conference on Information and Communication Technology for Sustainable Development (ICICT4SD). 2021. p.95–99.

BALDI, P.; BRUNAK, S.; CHAUVIN, Y.; ANDERSEN, C. A. F.; NIELSEN, H. Assessing the accuracy of prediction algorithms for classification: an overview. Bioinformatics, v.16, p.412–424, 2000.

DESYANI, T.; SAIFUDIN, A.; YULIANTI, Y. Feature selection based on Naive Bayes for caesarean section prediction. IOP Conference Series: Materials Science and Engineering, v.879, p.012091, 2020.

FACELI, K.; LORENA, A. C.; GAMA, J.; de ALMEIDA, T. A.; de LEON FERREIRA DE CARVALHO, A. C. P. Inteligência artificial: uma abordagem de aprendizado de máquina. 2a ed. Grupo Gen – LTC. 2021.

ISLAM, M. N.; MAHMUD, T.; KHAN, N. I.; MUSTAFINA, S. N.; ISLAM, A. K. M. N. Exploring machine learning algorithms to find the best features for predicting modes of childbirth. IEEE Access, v.9, p.1680–1692, 2021.

KAVITHA, D.; BALASUBRAMANIAN, T. A comparative study on mode of delivery and analyzing the risk factors of cesarean delivery using k-nearest neighbor, SVM and C5.0 classification techniques. Turkish Journal of Physiotherapy and Rehabilitation, v.32, n.2, p.1873–1878, 2021.

KOWSHER, M.; PROTTASHA, N. J.; TAHABILDER, A.; ISLAM, M. B. Machine learning based recommendation systems for the mode of childbirth. Em: Bhuiyan, T., Rahman, M. M., Ali, M. A. (Eds) Cyber Security and Computer Science. Cham: Springer International Publishing. 2020. p.295–306.

LEMAÎTRE, G.; NOGUEIRA, F.; ARIDAS, C. K. Imbalanced-learn: A Python toolbox to tackle the curse of imbalanced datasets in machine learning. Journal of Machine Learning Research, v.18, n.2017, p.1–5, 2017.

NCHS. Natality 2019. URL http://www.cdc.gov/nchs/data_access/VitalStatsOnline.htm.2020.

PAULA, G. A. *Modelos de regressão: com apoio computacional.* URL https://www.ime.usp.br/~giapaula/texto_2023.pdf. 2023.

PEDREGOSA, F.; VAROQUAUX, G.; GRAMFORT, A.; MICHEL, V.; THIRION, B.; GRISEL, O.; BLONDEL, M.; PRETTENHOFER, P.; WEISS, R.; DUBOURG, V.; VANDERPLAS, J.; PASSOS, A.; COURNAPEAU, D.; BRUCHER, M.; PERROT, M.; DUCHESNAY, E. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, v.12, n.2021, p.2825–2830, 2011.

RAHMAN, S.; KHAN, M. I.; SATU, M. S.; ABEDIN, M. Z. Risk Prediction with Machine Learning in Cesarean Section: Optimizing Healthcare Operational Decisions. Cham: Springer International Publishing. 2021. p.293–314.

SANTOS, T. T. Evidências de indução de demanda por parto cesáreo no Brasil. Dissertação de Mestrado em Economia, Universidade Federal de Minas Gerais. 2011.

SEABOLD, S.; PERKTOLD, J. statsmodels: Econometric and statistical modeling with Python. Em: 9th Python in Science Conference. 2010.

ULLAH, Z.; SALEEM, F.; JAMJOOM, M.; FAKIEH, B. Reliable prediction models based on enriched data for identifying the mode of childbirth by using machine learning methods: Development study. Journal of Medical Internet Research, v.23, n.6, e28856, 2021.

WHO. WHO statement on caesarean section rates. URL https://www.who.int/publications/i/item/WHO-RHR-15.02. 2015.