INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DE SÃO PAULO

Felipe Domingues Bonfim	SP3071227
Gabriela Dias Dutra	SP3030041
Henrique Oliveira De Souza	SP3060292
Lucas Nogueira De Souza	SP3072703

Projeto de Introdução a Ciência de Dados - Relatório CRISP-DM

SÃO PAULO

2023

INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DE SÃO PAULO

Felipe Domingues Bonfim	SP3071227
Gabriela Dias Dutra	SP3030041
Henrique Oliveira De Souza	SP3060292
Lucas Nogueira De Souza	SP3072703

Projeto de Introdução a Ciência de Dados - Relatório CRISP-DM

Atividade apresentada ao curso de Análise e Desenvolvimento de Sistemas do Instituto Federal de Educação, Ciência e Tecnologia de São Paulo.

Orientadora Prof.ª Dra. Josceli M. Tenorio

SÃO PAULO

2023

Sumário

Introdução **3**
Descrição do Problema 3
Análise de dados **6**
Variáveis utilizadas **15**
Tratamentos gerais realizados na base de dados **16**
Avaliação de modelos **18**
Regressão logística 18
Árvore de decisão 20
Florestas aleatórias 23
Conclusão **24**

Introdução

Este relatório tem como objetivo aplicar técnicas de estatística para investigar as múltiplas variáveis que podem determinar se um parto será vaginal ou cesárea, explorar a relação dos dados por meio de representações gráficas aplicar técnicas de ciência de dados para criar e avaliar três modelos estatísticos utilizando dados de partos no Brasil, no ano de 2021. As informações de nascimento utilizadas foram extraídas das bases de dados do Sistema de Informação sobre Nascidos Vivos (Sinasc) disponibilizadas pelo Ministério da Saúde através do portal DataSUS.

O Sistema de Informações sobre Nascidos Vivos (Sinasc), foi implantado oficialmente a partir de 1990, com o objetivo de coletar dados sobre os nascimentos informados em todo território nacional e fornecer dados sobre natalidade para todos os níveis do Sistema de Saúde.

Descrição do Problema

O parto vaginal é um processo natural e fisiológico. Contudo, em certas circunstâncias, uma cesariana (CS) pode ser necessária para proteger a saúde da mulher e do bebê. Nessas circunstâncias, a não adesão da CS contribui para o aumento da mortalidade e morbidade materna e perinatal.

A problemática gira em torno das indicações desnecessárias do procedimento em mulheres que naturalmente não correm risco nenhum de saúde, apenas pelo fato de ser mais lucrativo para quem as indica.

No ano de 2018, em matéria publicada pelo Senado Federal, especialistas já indicavam uma epidemia de cesáreas no Brasil.

Segundo dados do Ministério da Saúde, baseado no relatório da revista científica The Lancet, o Brasil tem a segunda maior taxa de cesáreas do mundo, sendo 57,7% dos partos realizados, o recomendado pela Organização Mundial da Saúde é menos de 15%. A falta de autonomia da gestante em escolher a melhor opção, a naturalização da CS e a intervenção médica contribuem para este cenário.

O uso indiscriminado do procedimento cirúrgico traz danos não somente à gestante como ao feto e a sociedade como um todo, por se tratar de uma questão de saúde pública.

Figura 1 - Ranking de cesarianas no mundo (Fonte: The Lancet, 2018)

Sobre as opções de parto e a escolha das mulheres, a Fiocruz aponta que 70% das mulheres em um primeiro momento desejam ter o parto vaginal (PV), porém no decorrer da gravidez e do pré natal são desencorajadas por seus familiares e médicos. Comparado com os dados do gráfico apresentado abaixo 10% das mulheres relataram que o motivo da opção pela CS foi por indicação médica e 14% não queria sentir dor ou ser mais conveniente. Outros estudos também indicam como motivo a questão estética e a praticidade da cirurgia (MULLER,2015).

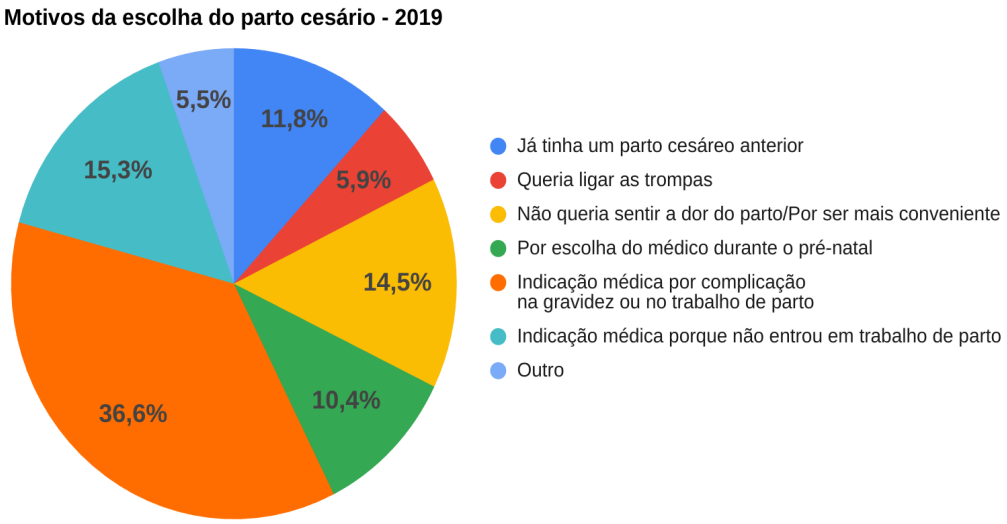


Figura 2 - Motivos da escolha do parto cesáreo (Fonte: PNS, 2019. Adaptado pelos autores)

Outro fator importante de se mencionar é a questão socioeconômica das parturientes, que em relação ao parto vaginal as mulheres de baixa renda e menor nível de escolaridade representam 62,3% da porcentagem total.

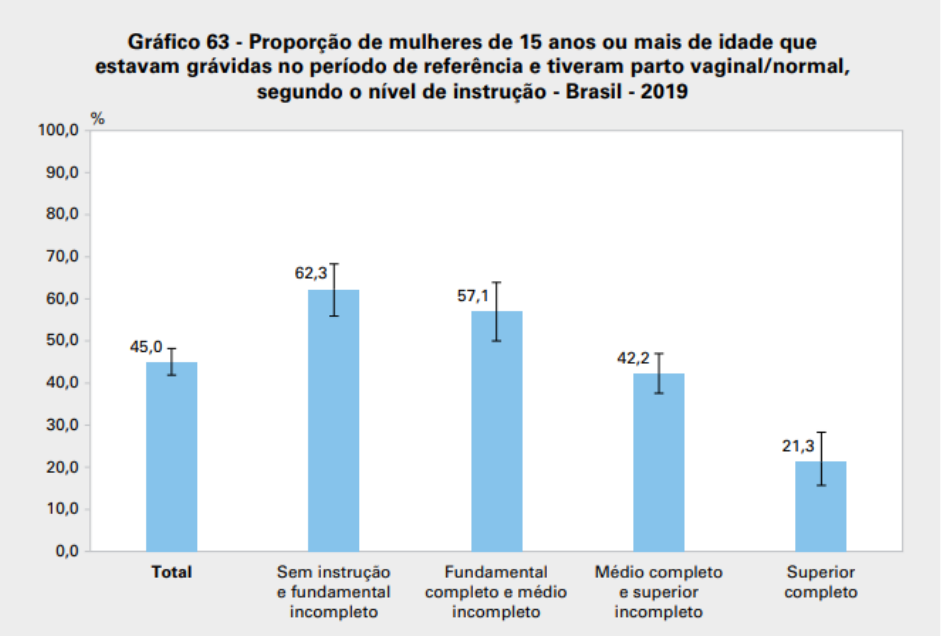


Figura 3 - Proporção de mulheres de 15 anos ou mais de idade que estavam grávidas no período de referência e tiveram parto vaginal/normal segundo o nível de instrução (Fonte: IBGE)

Tendo em vista as estatísticas supracitadas conclui-se que é necessário uma maior exploração da temática, divulgação de informações para que as gestantes tenham o poder de escolha respeitado e consciente. O engajamento dos órgãos públicos e privados na resolução da questão.

Análise de dados

A análise de dados destaca-se como uma etapa fundamental para desvendar padrões e características presentes na base de dados do Sistema de Informação sobre Nascidos Vivos (SINASC) referente ao ano de 2021. O foco central desta análise recai sobre a variável "*PARTO*", um indicador crucial que distingue entre partos vaginais e cesarianas. A importância desta investigação reside na capacidade de oferecer uma compreensão aprofundada dos fatores associados a diferentes métodos de parto, indo além de aspectos demográficos e abrangendo variáveis como escolaridade, idade materna e outras. Ao explorar a distribuição da variável "*PARTO*", busca-se não apenas identificar as tendências predominantes, mas também fornecer uma visão abrangente das dinâmicas de saúde materna.

Na Figura 4, podemos visualizar a distribuição da idade das mães separadas por tipo de parto.

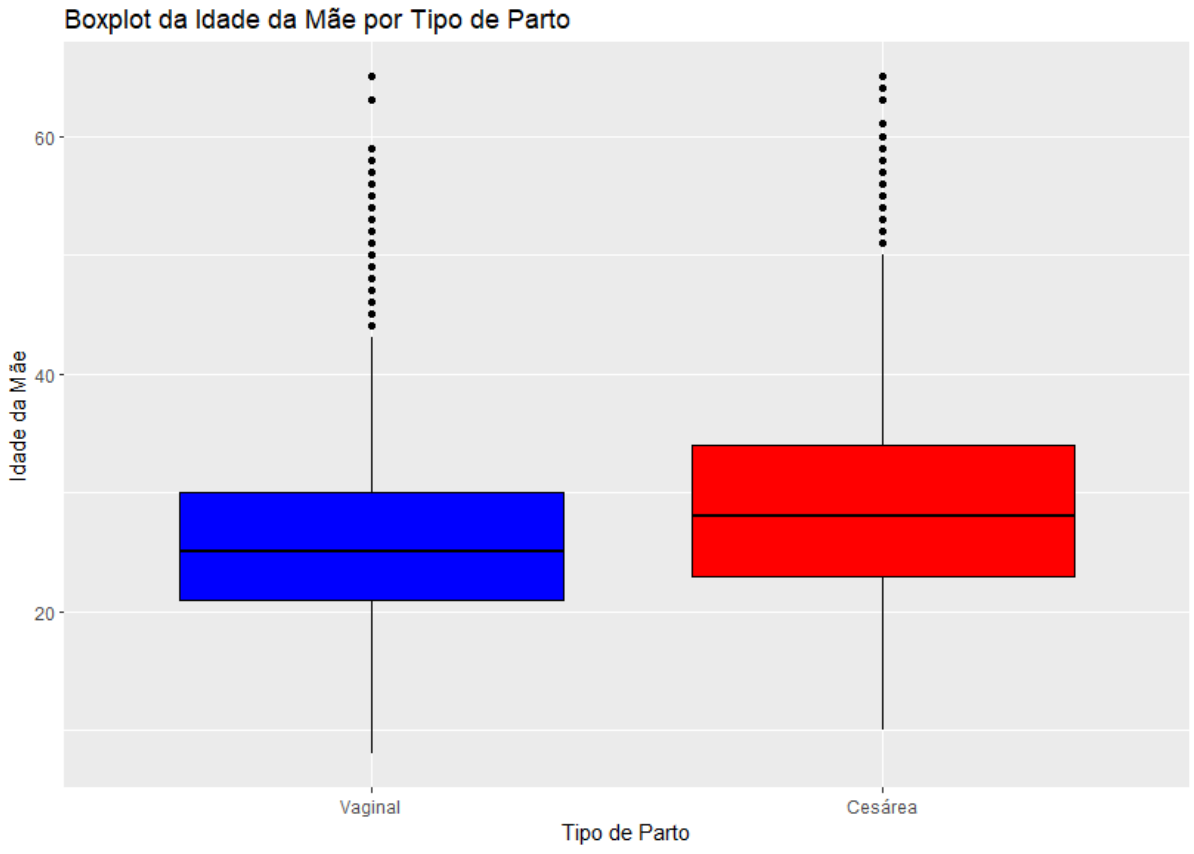


Figura 4 - Boxplot de idade da mãe por tipo de parto (Fonte: Os autores)

A Figura 5 apresenta um gráfico de barras separando os tipos de parto por local de nascimento.

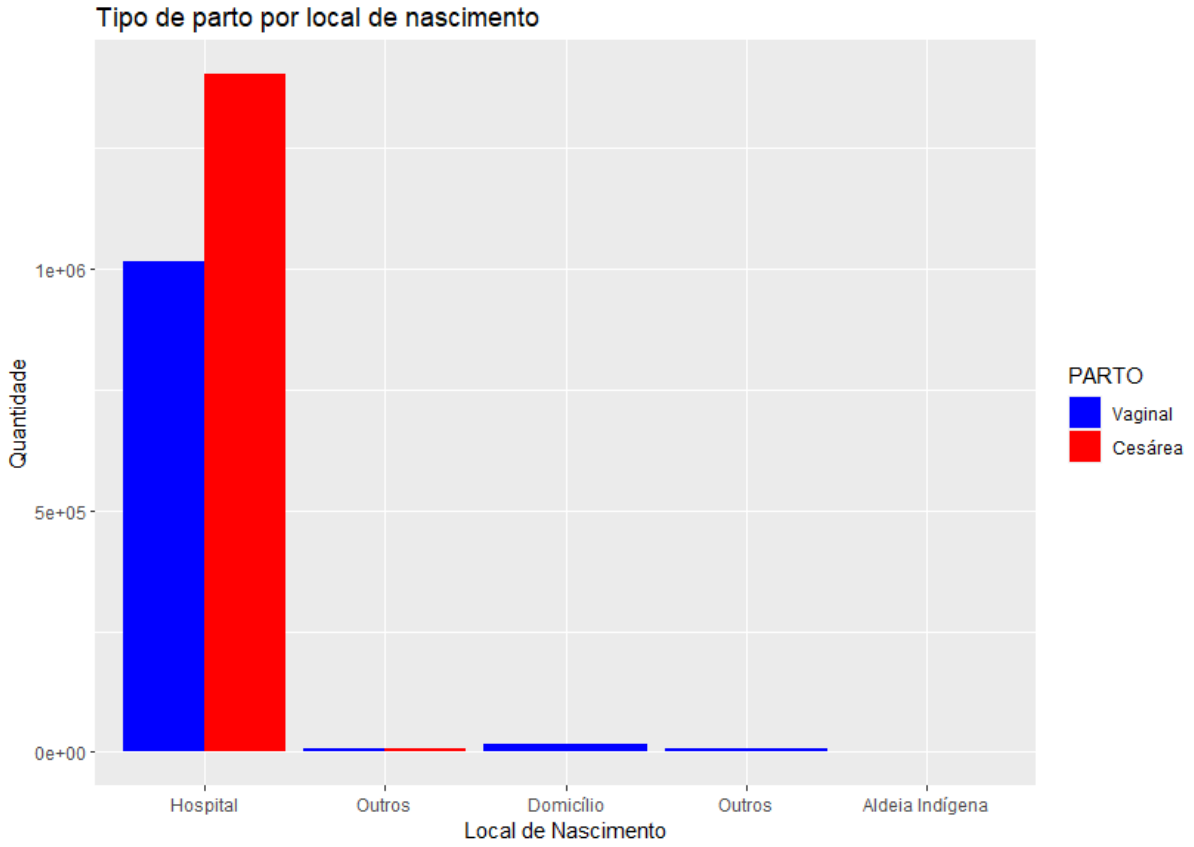


Figura 5 - Tipo de parto por local de nascimento (Fonte: Os autores)

A Figura 6 apresenta um gráfico de barras representando os tipos de parto por situação conjugal da mãe.

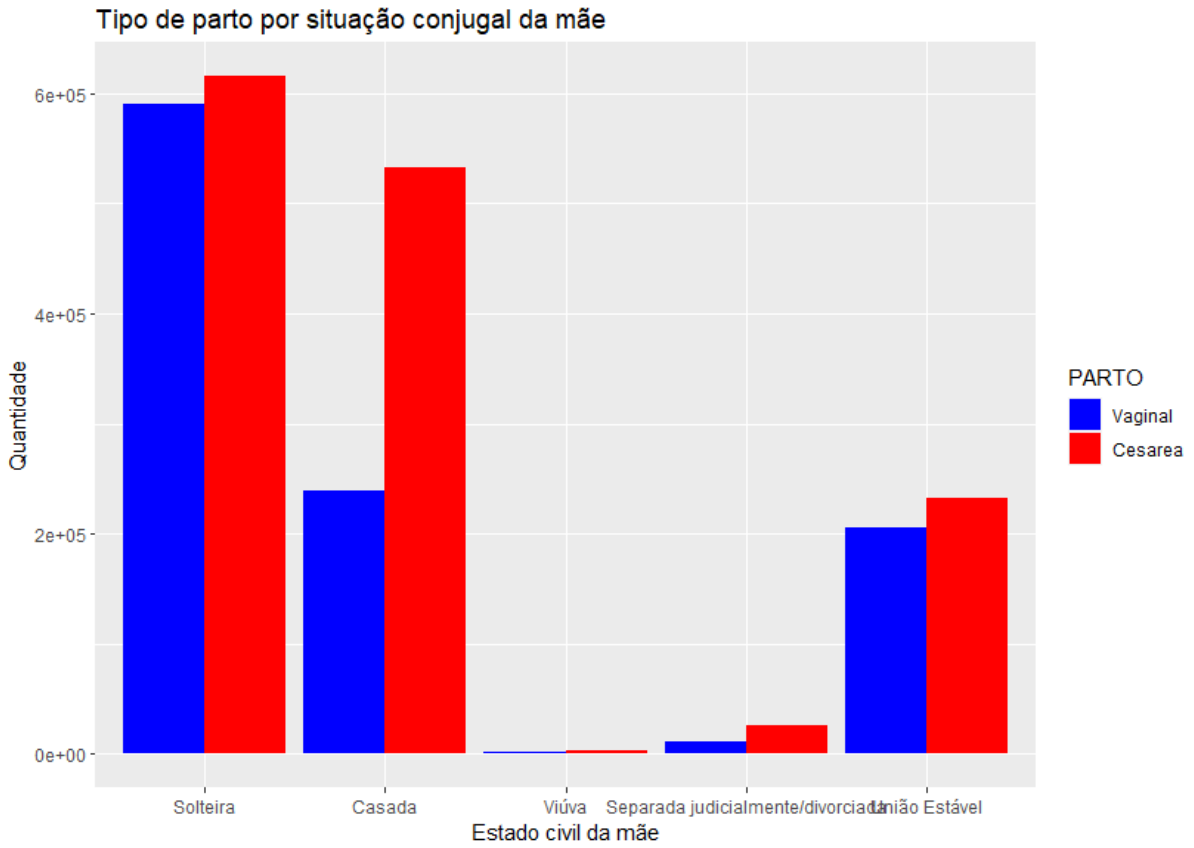


Figura 6 - Tipo de parto por situação conjugal da mãe (Fonte: Os autores)

A Figura 7 representa os tipos de parto separados pelo nível de escolaridade da mãe. O valor do nível da escolaridade é separado em períodos de anos estudados pela mesma.

a

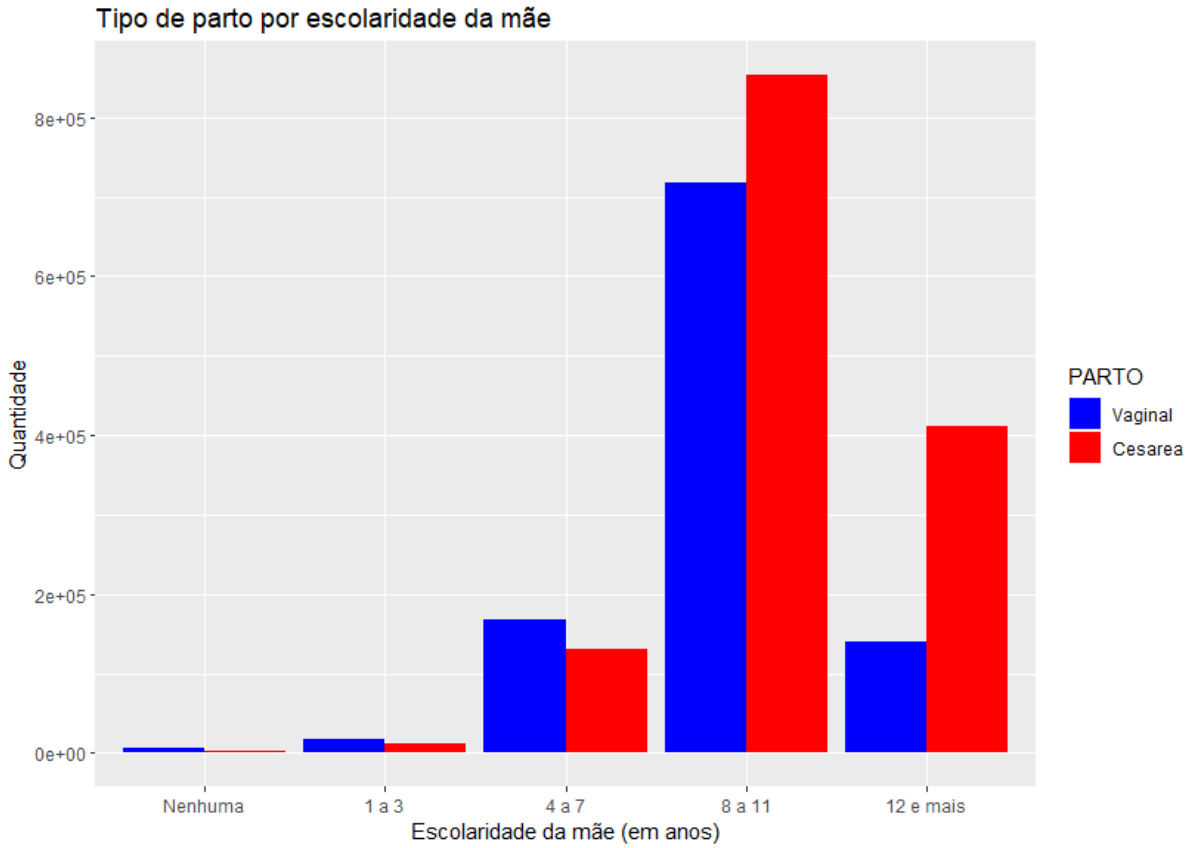


Figura 7 - Tipo de parto por escolaridade da mãe (Fonte: Os autores)

A Figura 8 representa os tipos de parto considerando se a mãe já passou por um processo de parto vaginal anteriormente.

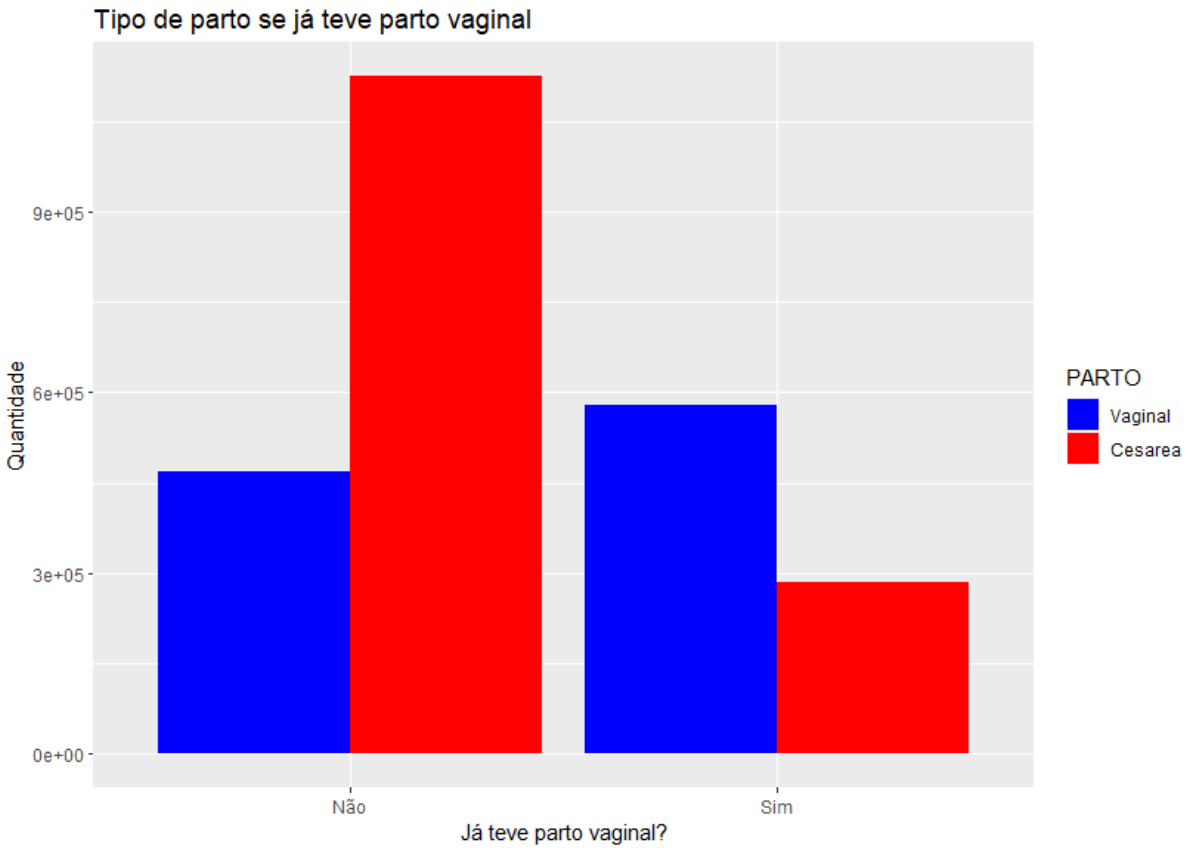


Figura 8 - Tipo de parto levando considerando se já houve parto vaginal previamente(Fonte: Os autores)

A Figura 9 representa os tipos de parto considerando se a mãe já passou por um processo de parto cesárea anteriormente.

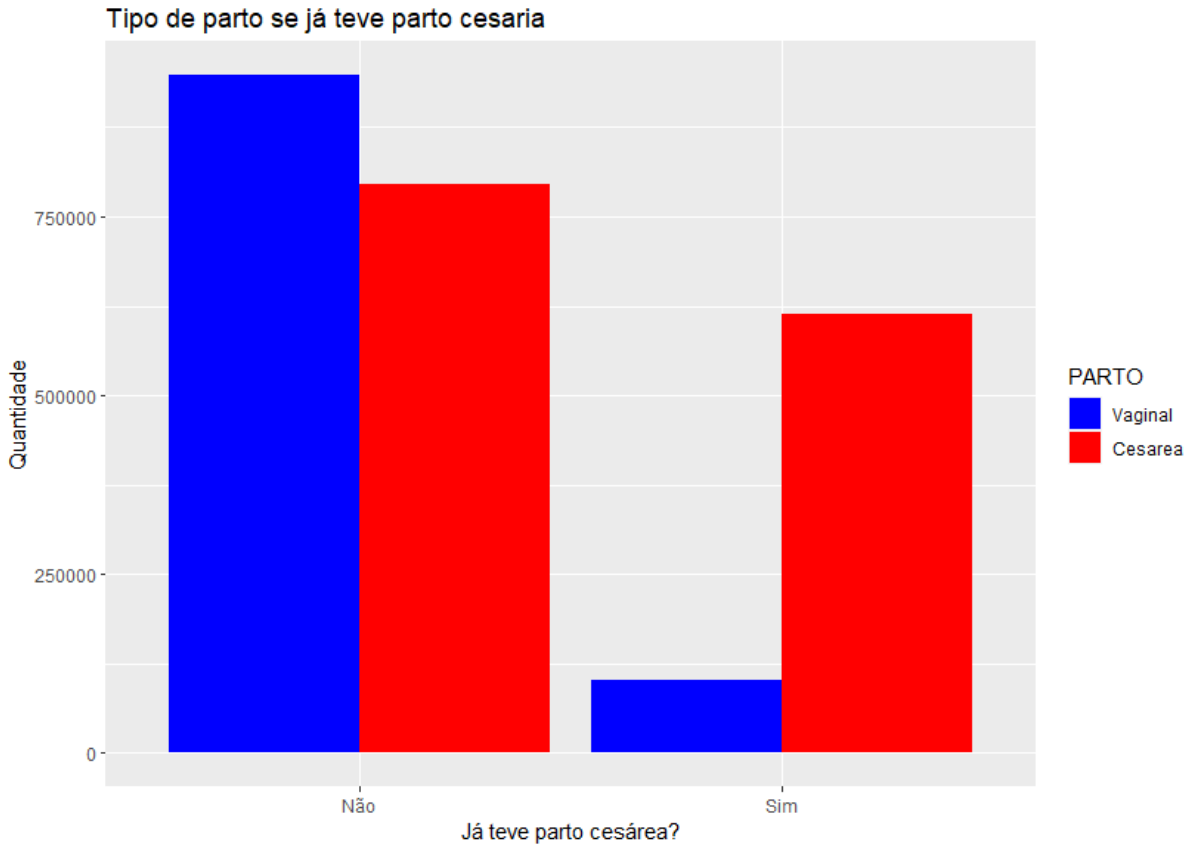


Figura 9 - Tipo de parto levando considerando se já houve parto cesárea previamente(Fonte: Os autores)

Com a junção das bases de dados do Sistema de Informação sobre Nascidos Vivos – Sinasc e da base de dados de municípios do Brasil, é possível visualizarmos a separação dos tipos de parto por cada estado do Brasil, conforme mostra a Figura 10.

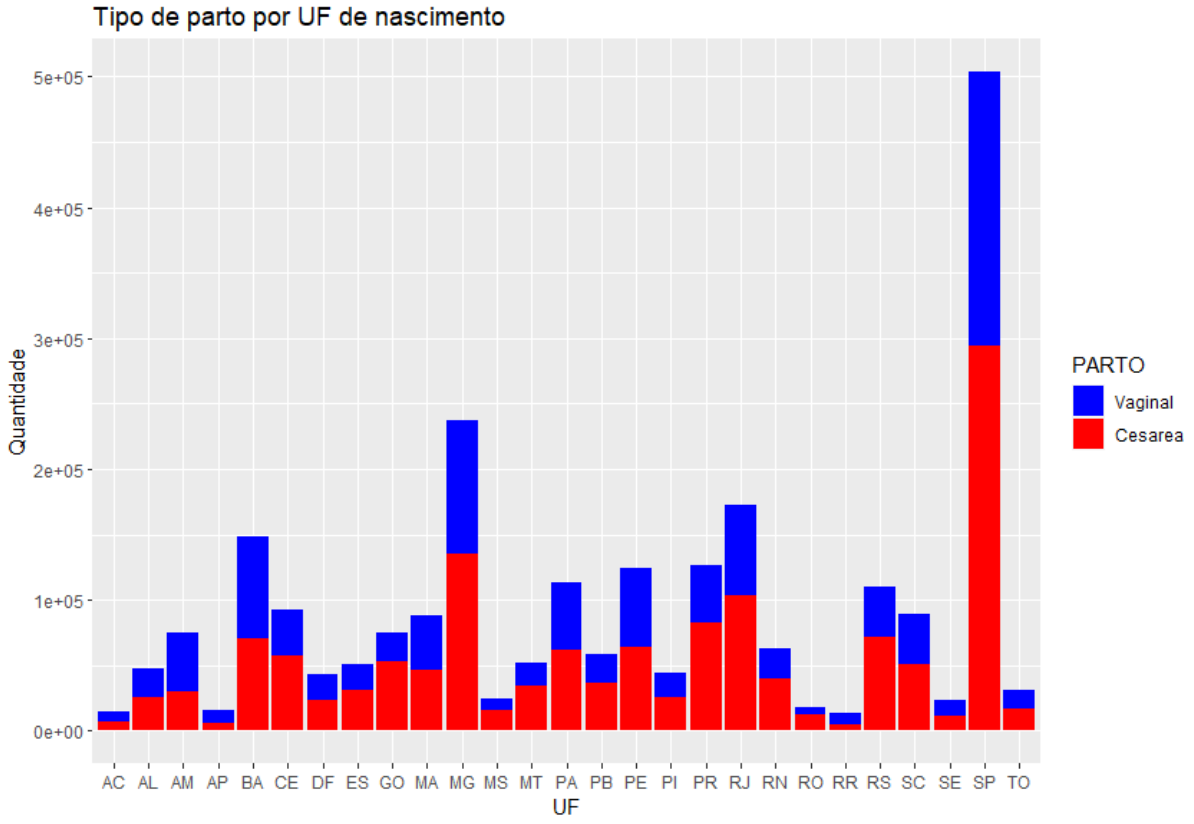


Figura 10 - Tipo de parto separado por estado (Fonte: Os autores)

A Figura 11 exibe as correlações entre as variáveis independentes e a variável dependente (*PARTO 1*, que corresponde a partos normais). A coluna de *Correlation*, que vai de -1 a 1 aponta o grau de correlação entre as variáveis.

Exemplificando, devido ao fato da variável dependente ser dicotômica (considerou-se parto normal como verdadeiro e parto cesáreo como falso), as correlações para *PARTO 2* consistem no inverso daquelas encontradas para *PARTO 1*.

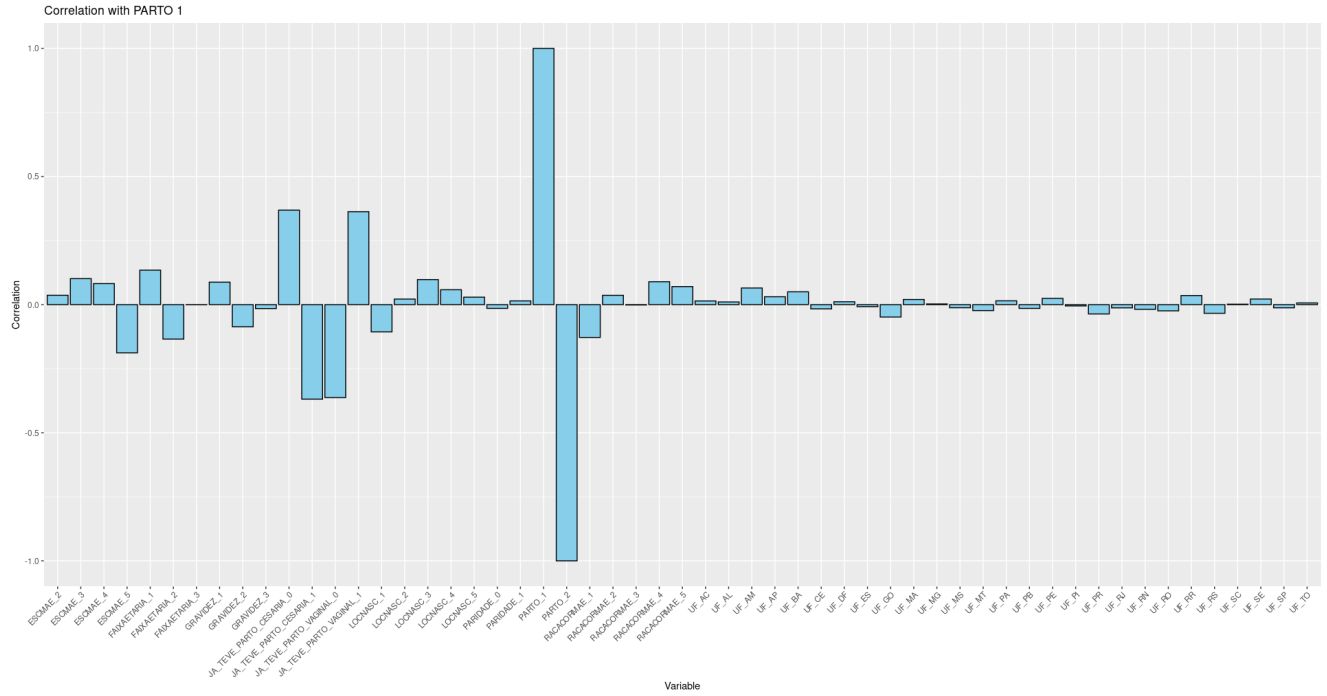


Figura 11 - Correlações entre variáveis (Fonte: Os autores)

Com essa análise exploratória, nota-se que, em 2021, o parto cesárea, também conhecido como cesariana, foi o mais ocorrido de forma geral no Brasil. Os partos vaginais ocorrem mais comumente entre mulheres mais novas do que as que fazem cesariana.

Evidencia-se também que as mães casadas optam muito mais pelos partos cesárea do que pelos partos normais. Também é observado que as mães com menos tempo de estudo tendem a optar pelos partos normais, e quanto maior o tempo de estudo, mais comuns são as cesarianas.

Se observa também que as mães que já passaram por um processo de cesariana tendem a optar pelo mesmo procedimento na segunda ou próxima ocorrência de gravidez, e o mesmo comportamento ocorre para o parto normal, porém, a tendência de optar pelo mesmo procedimento é maior para o parto cesárea.

Por fim, nota-se que a região Norte do Brasil é a que mais opta pelo parto normal, seguido do Nordeste. De resto, as outras regiões brasileiras concentram mais partos cesárea.

Variáveis utilizadas

PARTO - Tipo de parto. 1 - Vaginal. 2 - Cesáreo. Essa informação é originária da base de dados de nascidos vivos.

ESCMAE - Escolaridade da mãe, em anos de estudo concluídos: 1 – Nenhuma; 2 – 1 a 3 anos; 3 – 4 a 7 anos; 4 – 8 a 11 anos; 5 – 12 e mais; 9 – Ignorado. Essa informação é originária da base de dados de nascidos vivos.

RACACORMAE - Tipo de raça e cor da mãe: 1– Branca; 2– Preta; 3– Amarela; 4– Parda; 5– Indígena. Essa informação é originária da base de dados de nascidos vivos.

LOCNASC - Local de nascimento: 1 – Hospital; 2 – Outros estabelecimentos de saúde; 3 – Domicílio; 4 – Outros; 5- Aldeia Indígena. Essa informação é originária da base de dados de nascidos vivos.
IDADE - Idade da mãe. Essa informação é originária da base de dados de nascidos vivos.

FAIXAETARIA - Faixa etária da mãe: 1 - Menor de idade (19 anos ou menos), Adulto (entre 20 e 60 anos) e Idoso (maior que 60 anos). Essa informação foi extraída a partir do campo ***IDADEMAE***, o critério de separação é baseado na definição das Nações Unidas (NU) que idosos são pessoas com mais de 60 anos e na definição da Organização Mundial de Saúde (OMS) que a idade adulta começa a partir dos 19 anos.

CODMUNNASC - Código de município do nascimento. Essa informação é originária da base de dados de nascidos vivos.

UF - Código de UF onde a criança nasceu. Essa informação foi extraída fazendo uma junção com a base de Municípios, correlacionando o código de município do nascimento (***CODMUNNASC***) da base de nascidos vivos com com o valor equivalente na base de Municípios (***CODMUNIC***), extraindo por fim o código da Unidade Federativa (***uf_code***). Essa informação é originária da base de dados de nascidos vivos.

PARIDADE - Define se é a primeira gravidez ou se teve mais de uma. 1 – Multípara; 0- Nulípara. Essa informação é originária da base de dados de nascidos vivos.

GRAVIDEZ - Tipo de gravidez: 1– Única; 2– Dupla; 3– Tripla ou mais; 9– Ignorado.

QTDPARTCES - Número de partos cesáreos anteriores pela gestante. Informação originária da base de dados de nascidos vivos.

QTDPARTNOR - Número de partos vaginais anteriores pela gestante. Informação originária da base de dados de nascidos vivos.

JA_TEVE_PARTO_CESARIA - Variável informando se a gestante já fez parto cesáreo no passado. Extraído do dado ***QTDPARTCES*.**

JA_TEVE_PARTO_VAGINAL - Variável informando se a gestante já fez parto cesáreo no passado. Extraído do dado ***QTDPARTNOR* .**

ESTCIVMAE - estado civil da mãe. Informação originária da base de dados de nascidos vivos.

Tratamentos gerais realizados na base de dados

Devido ao grande volume da base de nascidos vivos, é inevitável que alguns registros possuam colunas com valores nulos ou não informados. Para propósitos deste relatório, resolvemos ocultar todos os registros onde as seguintes váriaveis não estavam presentes: *PARTO, ESCMAE, RACACORMAE, IDADEMAE, GRAVIDEZ, UF, QTDPARTCES, QTDPARTNOR*. Considerando que a entrada inicial de dados de nascidos vivos era de 2.677.101 registros, e após a remoção consiste de 2.473.568 registros, podemos concluir que foram perdidos aproximadamente 8% do volume da base durante o processo.

Além disso, como é possível observar, algumas variáveis categóricas estão sendo representadas numericamente, elas foram convertidas para factor, a fim de facilitar as operações que serão realizadas durante a análise exploratória e treinamento durante o relatório. Nesse caso, as seguintes váriaveis passaram por esse tratamento: *PARTO, ESCMAE, RACACORMAE, LOCNASC, FAIXAETARIA, GRAVIDEZ, PARIDADE, ESTCIVMAE, UF, JA_TEVE_PARTO_CESARIA E JA_TEVE_PARTO_VAGINAL*.

Finalizando, os dados resultantes dessas mudanças foram divididos em dois *datasets* diferentes com uma proporção de 70/30, o primeiro para treinamento dos modelos e o segundo para testar os modelos após o treinamento. Isso quer dizer que 1.731.497 registros foram utilizados no treinamento dos modelos, e que 742.071 registros foram utilizados para testar os resultados.

Avaliação de modelos

Nesse relatório foram escolhidos, para propósito de análise, três modelos estatísticos diferentes. A ideia é avaliar a performance de cada modelo de maneira geral durante a classificação do tipo de parto baseado nas variáveis extraídas, fazendo uma comparação de performance nos dados de treinamento e nos dados de teste, além de uma avaliação das características individuais de cada modelo.

Entre as informações que serão igualmente analisadas, podemos destacar:

Acurácia - Acertos em relação ao total de itens, é a capacidade do modelo de medir aquilo que se propõe a medir.

Precisão - Capacidade de fornecer os mesmos resultados quando repetido, ou seja, indica o quão próximo os resultados são uns dos outros.

Sensibilidade - Taxa de valores verdadeiramente positivos. A padronização neste relatório é que partos vaginais são considerados positivos, portanto, a sensibilidade é dada pela proporção entre quantos partos vaginais os modelos previram corretamente e o total de partos vaginais.

Especificidade - Taxa de valores verdadeiramente negativos. Seguindo a padronização de que partos vaginais são considerado positivos, e considerando a dicotomia do tipo de parto, neste relatório partos cesáreos são considerados negativos; portanto, a especificidade é dada pela proporção entre quantos partos cesáreos os modelos previram corretamente e o total de partos cesáreos.

Curva ROC e **AUC**: A Curva ROC é uma ferramenta gráfica que representa o desempenho de um modelo de classificação em diferentes pontos de corte para a probabilidade de predição. Ela exibe a taxa de sensibilidade em relação à taxa de especificidade para. Quanto mais próxima a curva estiver do canto superior esquerdo, melhor é o desempenho do modelo. AUC (Area Under Curve ou Área Sob a curva) é uma medida numérica associada à Curva ROC, representando a área sob a curva. Quanto maior a AUC, melhor o desempenho do modelo em discriminar entre classes. Uma AUC de 1 indica um modelo perfeito, enquanto uma AUC de 0,5 sugere um desempenho equivalente ao acaso.

Essas informações serão extraídas das matrizes de confusão geradas em conjunto com as predições pelos modelos. A matriz de confusão separa as decisões tomadas pelo classificador em verdadeiros positivos, falso positivos, falso negativos e verdadeiros negativos.

Regressão logística

A regressão logística é um modelo estatístico que calcula a probabilidade de um determinado evento dicotômico baseado na combinação de diferentes variáveis independentes. O modelo calcula o peso de cada variável independente por meio de função de ativação e com base nisso classifica a variável dependente entre 0 e 1.

Avaliação individual do modelo

Após a criação do modelo de regressão logística, é possível extrair os coeficientes sendo levados em consideração e calcular a importância de cada valor dentro do modelo. Os valores relevantes que podem ser extraídos são:

Estimativa: Coeficiente utilizado para estimar a probabilidade baseado no valor da variável. Nesta análise, valores acima de 0 indicam que a variável aumenta as chances de que o parto tenha sido cesáreo e valores abaixo de 0 indicam que a variável reduz as chances de que o parto tenha sido cesáreo.

Erro padrão: Medida que expressa variabilidade associada a uma estimativa, que no caso de regressão logística é quão consistentes ou dispersos são os valores observados em torno do valor estimado para o coeficiente.. Quanto menor o erro padrão de uma variável, maior a sua significância estatística.

p-value: Usado para avaliar se um resultado observado é estatisticamente significativo ou resultado de acaso. Quanto menor o p-value, maior a possibilidade que a variável tenha significância estatística e efeito na determinação do tipo de parto.

z-value: Essa informação é calculada dividindo o coeficiente estimado pelo seu erro padrão, é uma maneira mais simples de enxergar a significância estatística de uma variável. Quanto maior o valor, mais significância estatística.

Com base nisso, o modelo foi treinado com os dados de treinamento designados, onde então extraímos as variáveis com maior z-value modular e interpretamos como preditores mais consistentes do tipo de parto, eles são:

Variável	Estimativa	Erro padrão	z-value	p-value
JA_TEVE_PARTO_CESARIA1	1.655.200	0.005876	281.688	< 2e-16
JA_TEVE_PARTO_VAGINAL1	-1.560.995	0.006057	-257.699	< 2e-16
FAIXAETARIA2	0.715740	0.005445	131.450	< 2e-16
GRAVIDEZ2	1.831692	0.016374	111.865	< 2e-16
RACACORMAE4	-0.229057	0.004624	-49.533	< 2e-16

Algumas das conclusões que o modelo chega sobre o *dataset*, com base nesses dados, são:

- Em casos onde uma mulher já teve um parto cesárea anteriormente, as chances são mais altas que o parto dela seja cesárea.
- Em casos onde uma mulher já teve um parto vaginal anteriormente, as chances são mais altas que o parto dela seja cesárea.
- Em casos onde a faixa etária da gestante seja entre 20 e 60 anos, as chances são mais altas que o parto dela sejá cesárea.
- Em casos onde a gravidez é dupla as chances do parto ser cesárea aumentam.
- A probabilidade do parto ser vaginal é maior caso a gestante seja parda.

Esse relatório será acompanhado da tabela completa de coeficientes da da regressão logística com seus respectivos indicadores.

Avaliação geral do modelo

Utilizando a regressão logística para prever os valores nos dados de treinamento, podemos extrair a seguinte matriz de confusão e métricas:

Real

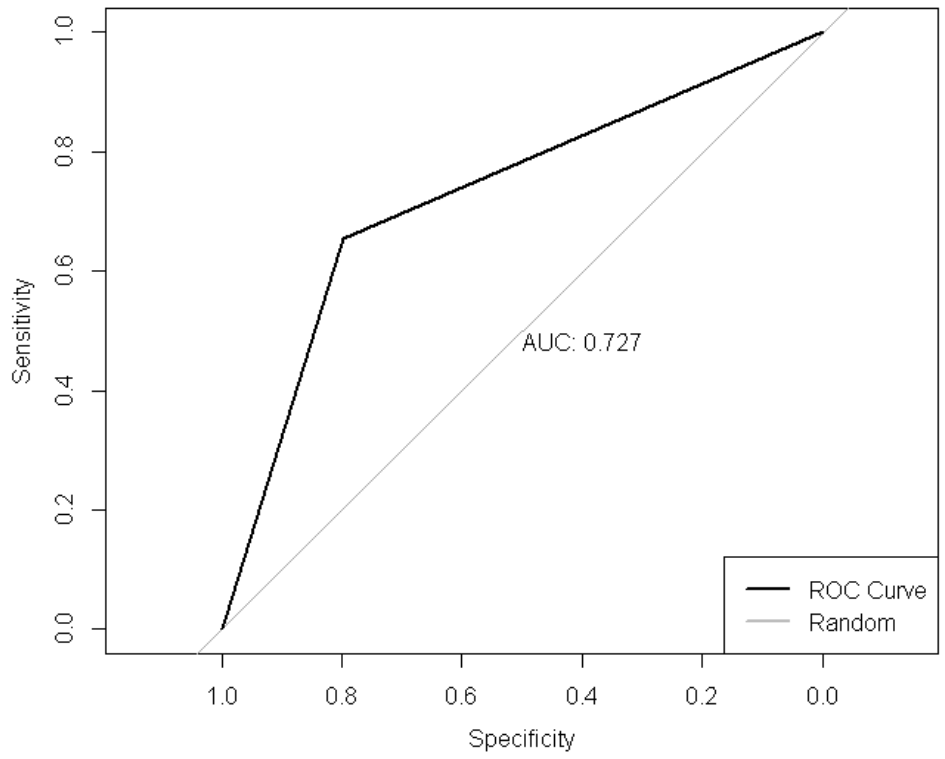
		Positivo (Vaginal)	Negativo (Cesáreo)
Predito	Positivo (Vaginal)	590074	342534
	Negativo (Cesáreo)	148706	650183

Acurácia: 0.7163. Isso quer dizer que o modelo consegue prever o tipo de parto com sucesso nos dados de treinamento em 71.63% dos casos.

Sensibilidade: 0.7987. Esse dado indica que o modelo conseguiu prever partos vaginais corretamente em 79,87% dos casos.

Especificidade: 0.6550. Indica que o modelo conseguiu prever partos cesáreas corretamente em 65,50% dos casos

AUC: 0.727. Uma área de cobertura sob a curva ROC indica que o modelo tem uma performance significativamente melhor que a aleatoriedade. Segue representação da curva ROC:



Seguindo com os dados designados para teste, extraímos a seguinte matriz de confusão e métricas:

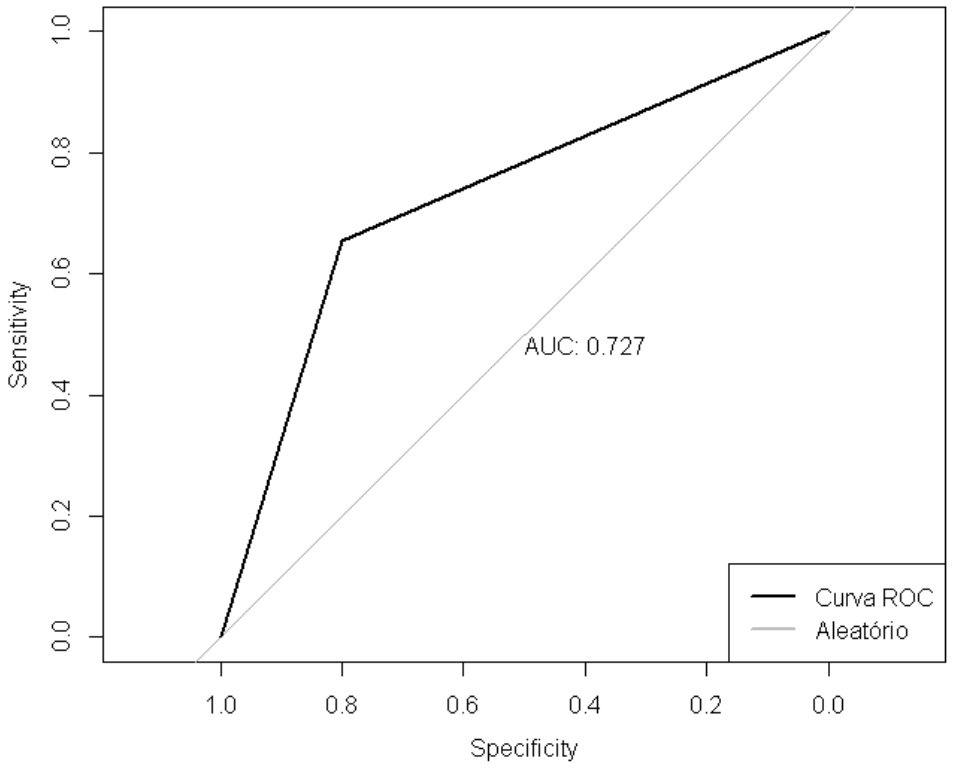
		Real	
		Positivo (Vaginal)	Negativo (Cesáreo)
Predito	Positivo (Vaginal)	253470	146513
	Negativo (Cesáreo)	63673	278415

Acurácia: 0.7168. Isso quer dizer que o modelo consegue prever o tipo de parto com sucesso nos dados de treinamento em 71.68% dos casos.

Sensibilidade: 0.7992. Esse dado indica que o modelo conseguiu prever partos vaginais corretamente em 79,92% dos casos.

Especificidade: 0.6552. Indica que o modelo conseguiu prever partos cesáreas corretamente em 65,52% dos casos

AUC: 0.727. Uma área de cobertura sob a curva ROC indica que o modelo tem uma performance significativamente melhor que a aleatoriedade seguindo dados de teste. Segue representação da curva ROC:



Seguindo as informações apresentadas, é possível observar que o modelo de regressão logística apresentou uma capacidade superior de classificar partos vaginais corretamente em relação à classificação de partos cesárea, dado seu valor de sensibilidade em relação ao valor de especificidade. Além disso, é possível observar que o modelo apresentou uma boa precisão, garantindo valores próximos de acurácia, sensibilidade, especificidade e AUC tanto nas predições da base de treinamento quanto na base de testes, sendo ligeiramente melhor na predição utilizando os dados de teste.

O valor de AUC apresentado tanto nas predições de treinamento quanto de teste indica que o modelo é uma opção consideravelmente melhor que a aleatoriedade em diversos limiares de sensibilidade/especificidade. Esse valor de AUC, em conjunto com a acurácia de 71,63% e 71,68% (dados de treinamento e teste, respectivamente), indicam que o modelo possui confiabilidade, mas não é um preditor perfeito.

Considerando a tabela de coeficientes que acompanha esse relatório, é possível observar que algumas váriaveis também possuem um erro padrão e p-value relativamente alto. Sendo assim, é possível que parte das dificuldades do modelo de classificar com acurácia seja devido ao overfitting de dados.

Portanto, a aplicação desse modelo precisa levar em consideração a sua maior dificuldade em classificar partos cesáreos, além das possibilidades de melhora considerando seus coeficientes atuais.

Árvore de decisão

A árvore de decisão é um modelo que utiliza dados de treinamento para segmentar o espaço do preditor em regiões não sobrepostas conhecidas como nós da árvore. Em cenários de classificação, o valor previsto para cada nó é a resposta mais comum no nó. Resumidamente, cada nó interno da árvore representa uma condição sobre os dados e cada ramo representa o resultado dessa condição. As folhas representam a decisão final, que no caso é se o parto sendo predito será vaginal ou cesárea.

Avaliação individual do modelo

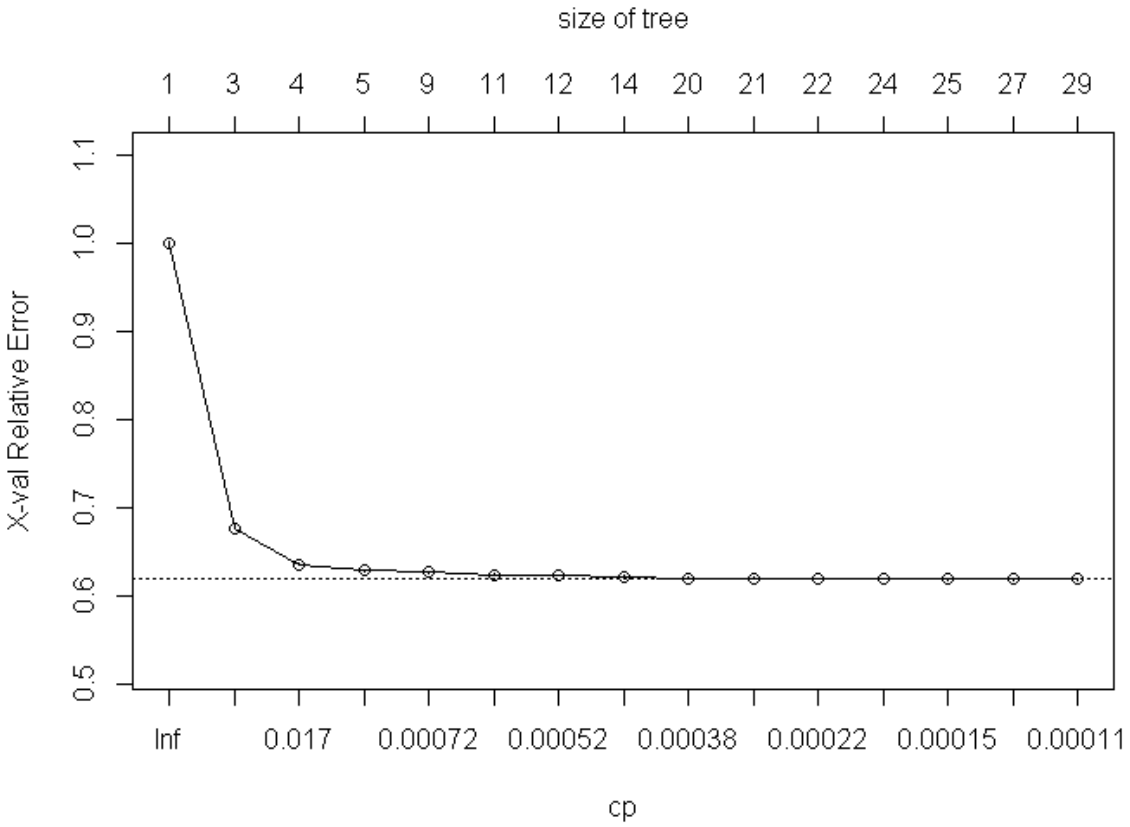
Durante a criação de uma árvore de decisão, o parâmetro de complexidade é utilizado para definir em quantos nós a árvore será dividida e ajustar seu poder preditivo. Essa divisão afeta os seguintes valores:

Erro relativo no Nó: indica o ganho de informação com cada divisão, quanto mais próximo de 0, mais a divisão foi efetiva na predição do modelo. Ou seja, melhor dividida essa árvore está.

Erro de validação cruzada: representa a taxa de erro quando a árvore é validada em um conjunto de dados separado, quanto mais próximo de zero melhor. Indica uma boa generalização do modelo.

Desvio padrão de validação cruzada: mede a variabilidade associada a estimativa do erro cruzado, quanto menor o valor, mais confiança na estimativa de validação cruzada.

Com base nessas informações, foi possível gerar uma projeção de erro relativo por fator de complexidade, o resultado foi esse:



Baseado nessa projeção, decidimos seguir então com um fator complexidade de 0.0001, com um erro relativo de 0.61857, erro de validação cruzada de 0.61926 e desvio padrão de validação cruzada de 0.00078533. O resultado foi uma árvore com 28 divisões que pode ser vista aqui.

Nessa complexidade, o erro na classificação do nó inicial é de 0.42667, isso indica que 42,66% dos valores são inicialmente classificados errados. Multiplicando esse valor pela taxa de erro relativo do nó, podemos adquirir a taxa de erro de resubstituição e chegar ao valor 0,2639, indicando que 26,39% dos valores de treinamento foram classificados incorretamente, importante notar na avaliação geral que esse dado está consistente com a acurácia em treinamento do modelo.

Os valores informados fornecem a informação que a árvore possui margem para melhorias que vão além do aumento da complexidade, já que é possível observar uma melhoria minúscula nas taxas de erro conforme a complexidade aumenta. Isso sugere que o modelo possa estar passando por overfitting de dados ou que ele carece de variáveis melhores.

Avaliação geral do modelo

Utilizando essa árvore de decisão para classificar os dados de treinamento, podemos extrair a seguinte matriz de confusão e métricas:

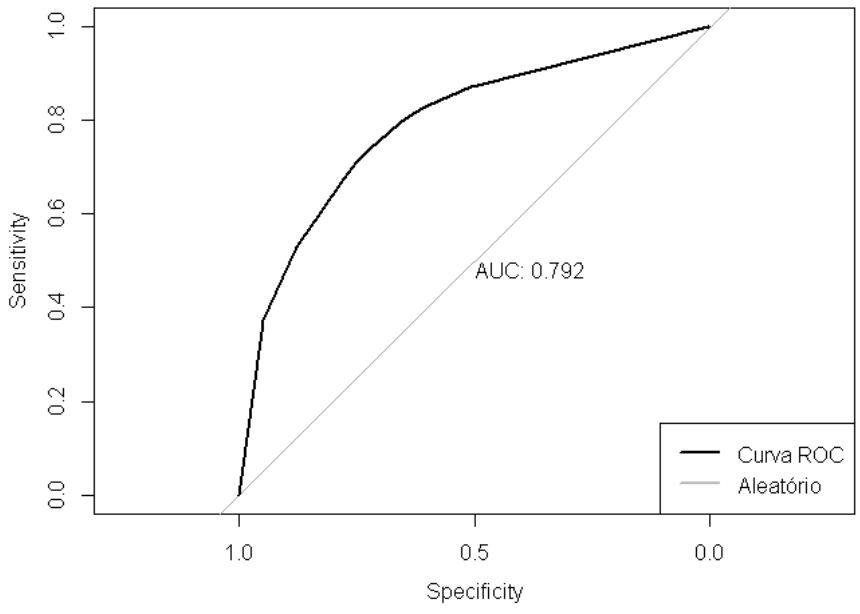
		Real	
		Positivo (Vaginal)	Negativo (Cesáreo)
Predito	Positivo (Vaginal)	483622	201828
	Negativo (Cesáreo)	255158	790889

Acurácia: 0.7361. Isso quer dizer que o modelo consegue prever o tipo de parto com sucesso nos dados de treinamento em 73.61% dos casos.

Sensibilidade: 0.6564. Esse dado indica que o modelo conseguiu prever partos vaginais corretamente em 65,64% dos casos.

Especificidade: 0.7967. Indica que o modelo conseguiu prever partos cesáreas corretamente em 79,67% dos casos

AUC: 0.792. Uma área de cobertura sob a curva ROC indica que o modelo tem uma performance significativamente melhor que a aleatoriedade seguindo dados de teste. Segue representação da curva ROC:



Já com os dados designados de teste, temos a seguinte matriz de confusão e seus dados:

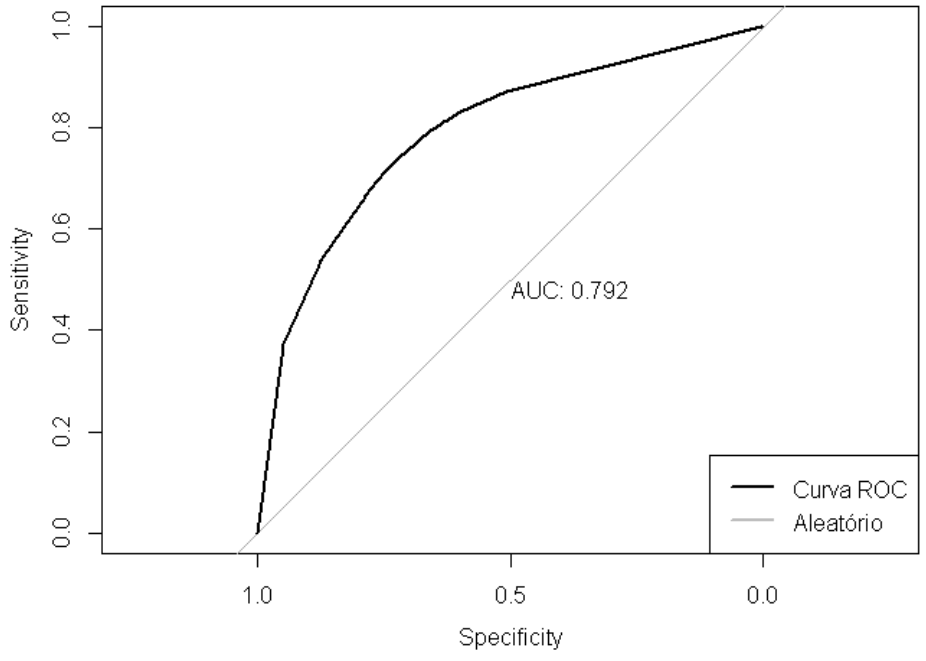
		Real	
		Positivo (Vaginal)	Negativo (Cesáreo)
Predito	Positivo (Vaginal)	480368	209236
	Negativo (Cesáreo)	258412	783481

Acurácia: 0.7299. Isso quer dizer que o modelo consegue prever o tipo de parto com sucesso nos dados de treinamento em 72.99% dos casos.

Sensibilidade: 0.6502. Esse dado indica que o modelo conseguiu prever partos vaginais corretamente em 65,02% dos casos.

Especificidade: 0.7892. Indica que o modelo conseguiu prever partos cesáreas corretamente em 78,92% dos casos

AUC: 0.792. Uma área de cobertura sob a curva ROC indica que o modelo tem uma performance significativamente melhor que a aleatoriedade seguindo dados de teste. Segue representação da curva ROC:



Seguindo os dados levantados, é possível observar que o modelo de árvore de decisão apresentou uma capacidade superior de classificar partos cesáreos corretamente em relação à classificação de partos vaginais, dado seu valor de especificidade maior em relação ao seu valor de sensibilidade. Além disso, é possível observar que o modelo apresentou uma boa precisão, garantindo valores próximos de acurácia, sensibilidade, especificidade e AUC tanto nas predições da base de treinamento quanto na base de testes, sendo ligeiramente melhor na predição utilizando os dados de treinamento.

O valor de AUC apresentado tanto nas predições de treinamento quanto de teste indica que o modelo é uma opção significativamente melhor que a aleatoriedade em diversos limiares de sensibilidade/especificidade. Esse valor de AUC, em conjunto com a acurácia de 73,61% e 72,99% (dados de treinamento e teste, respectivamente), indicam que o modelo possui confiabilidade, mas possivelmente tem margens para aprimoramento. Além disso, sua aplicação precisa levar em consideração a sua maior dificuldade em classificar partos vaginais.

Floresta aleatória

Um modelo de floresta aleatória é construído a partir de um conjunto de árvores de decisão, tais quais as abordadas anteriormente, atingindo maior grau de precisão, com redução do viés e do overfitting (ou sobreajuste).

Avaliação geral do modelo

Igual aos dois modelos citados anteriormente, a floresta aleatória foi utilizada para classificar o tipo de parto tanto para os dados de treinamento quanto para os dados de teste. Com os dados de treinamento podemos extrair a seguinte matriz de confusão e informações:

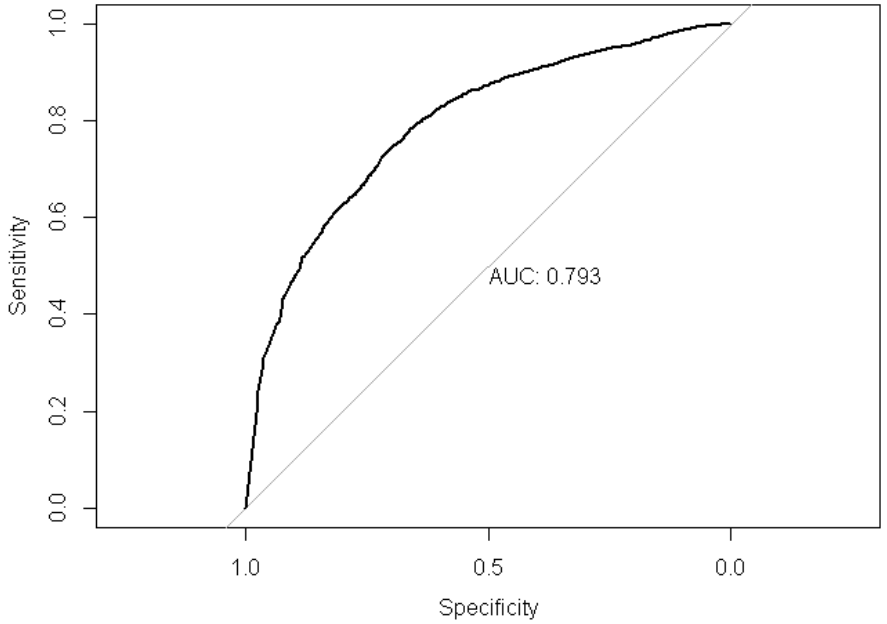
		Real	
		Positivo (Vaginal)	Negativo (Cesáreo)
Predito	Positivo (Vaginal)	480368	209236
	Negativo (Cesáreo)	258412	783481

Acurácia: 0.7299. Isso quer dizer que o modelo consegue prever o tipo de parto com sucesso nos dados de treinamento em 72.99% dos casos.

Sensibilidade: 0.6502. Esse dado indica que o modelo conseguiu prever partos vaginais corretamente em 65,02% dos casos.

Especificidade: 0.7892. Indica que o modelo conseguiu prever partos cesáreas corretamente em 78,92% dos casos.

AUC: 0.793. Uma área de cobertura sob a curva ROC indica que o modelo tem uma performance significativamente melhor que a aleatoriedade seguindo dados de teste. Segue representação da curva ROC:



Seguindo com as predições realizadas na base de teste, podemos extrair as seguintes informações:

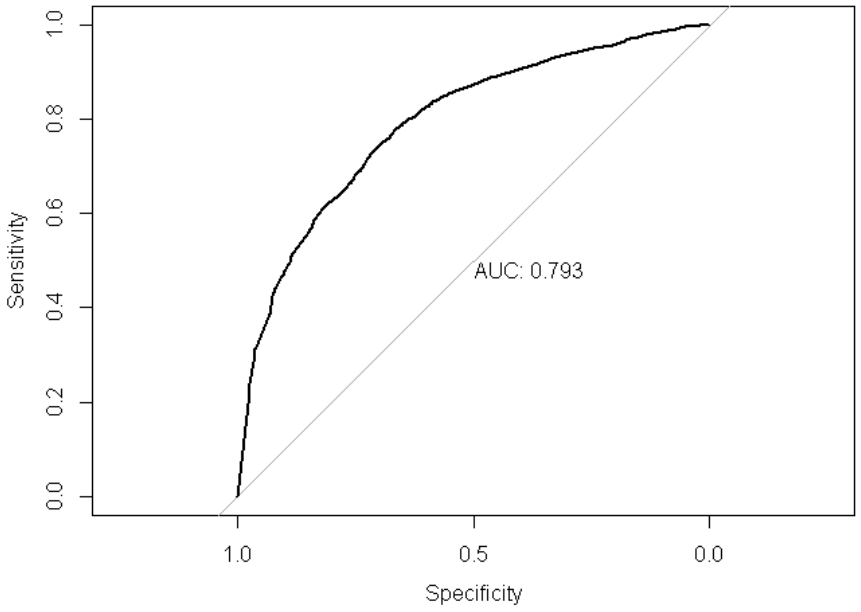
		Real	
		Positivo (Vaginal)	Negativo (Cesáreo)
Predito	Positivo (Vaginal)	206643	86980
	Negativo (Cesáreo)	109191	337948

Acurácia: 0.7356. Isso quer dizer que o modelo consegue prever o tipo de parto com sucesso nos dados de treinamento em 73.56% dos casos.

Sensibilidade: 0.6557. Esse dado indica que o modelo conseguiu prever partos vaginais corretamente em 65,57% dos casos.

Especificidade: 0.7967. Indica que o modelo conseguiu prever partos cesáreas corretamente em 79,67% dos casos

AUC: 0.793. Uma área de cobertura sob a curva ROC indica que o modelo tem uma performance significativamente melhor que a aleatoriedade seguindo dados de teste. Segue representação da curva ROC:



Avaliação individual do modelo

O modelo de floresta aleatória produziu um **F1 Score** de 0.6720663. Este indicador, que vai de 0 a 1, determinou que o considerando as duas classes (PARTO 1 e PARTO 2), utilizando uma média harmônica da capacidade do modelo capturar positivos (precisão) e de evitar falsos negativos (recall).

Também, o algoritmo fornece o **índice de Gini** para cada variável, denotando quantas observações seriam erroneamente com a remoção de cada variável.

	MeanDecreaseGini
ESCMAE	12101,004
RACACORMAE	2745,769
LOCNASC	2591,337
FAIXAETARIA	8210,387
UF	1246,36
GRAVIDEZ	2053,216
PARIDADE	11783,711
JA_TEVE_PARTO_CESARIA	74568,221

JA_TEVE_PARTO_VAGINAL	63788,666

Conclusão

Conforme visto nos dados levantados, todos os modelos apresentaram valores semelhantes de acurácia e AUC, indicando que todos tiveram aproximadamente o mesmo poder preditivo. É importante notar que, apesar disso, os modelos de árvore de decisão e random forest apresentaram maior poder preditivo na detecção de partos cesáreos, enquanto a regressão logística apresentou melhores resultados avaliando partos vaginais.

Todos os modelos apresentaram margem de melhoria, o levantamento de outras variáveis relevantes na determinação do tipo de parto poderia impactar positivamente o poder preditivo desses modelos. Ao mesmo tempo, seria cabível uma nova avaliação das variáveis já levantadas, buscando evidências de um overfitting das variáveis atuais nos conjuntos de dados utilizados.

Seguindo contra a ideia de declarar um modelo como superior, a aplicação dos modelos em conjunto pode ajudar no levantamento de valores de maneira mais útil. Ainda assim, a utilidade e valor desses modelos precisaria ser avaliada anteriormente por profissionais da saúde que poderiam se interessar pelas suas predições.

Referências

BOERMA, T. et al. **Global epidemiology of use of and disparities in caesarean sections**. The Lancet, v. 392, n. 10155, p. 1341–1348, out. 2018.

CARDOSO, J. E.; BARBOSA, R. H. S. **O desencontro entre desejo e realidade: a "indústria" da cesariana entre mulheres de camadas médias no Rio de Janeiro, Brasil.** Physis: Revista de Saúde Coletiva, v. 22, n. 1, p. 35–52, 2012.

Cesáreas respondem por 84% dos partos realizados por planos em 2019. Disponível em: <https://agenciabrasil.ebc.com.br/saude/noticia/2021-08/cesareas-respondem-por-84-dos-partos-realizados-por-planos-em-2019>.

CICLOS DE VIDA, B. **Pesquisa Nacional de Saúde 2019**. [s.l: s.n.]. Disponível em: <https://www.pns.icict.fiocruz.br/wp-content/uploads/2021/12/liv101846.pdf>.

Especialistas apontam epidemia de cesarianas no Brasil. Disponível em: <https://www12.senado.leg.br/noticias/especiais/especial-cidadania/especialistas-apontam-epidemia-de-cesarianas/especialistas-apontam-epidemia-de-cesarianas>.

MÜLLER, E.; RODRIGUES, L.; PIMENTEL, C. **O tabu do parto: Dilemas e interdições de um campo ainda em construção**. Civitas, v. 15, n. 2, p. 272–272, 4 set. 2015.

Taxas de cesarianas continuam aumentando em meio a crescentes desigualdades no acesso, afirma OMS - OPAS/OMS | Organização Pan-Americana da Saúde. Disponível em: <https://www.paho.org/pt/noticias/16-6-2021-taxas-cesarianas-continuam-aumentando-em-meio-crescentes-desigualdades-no-acesso>.