UNIVERSIDADE ESTADUAL DE CAMPINAS
INSTITUTO DE MATEMÁTICA, ESTATÍSTICA E COMPUTAÇÃO CIENTÍFICA
DEPARTAMENTO DE MATEMÁTICA APLICADA

João Victor Ribeiro dos Santos

Aplicação de métodos interpretáveis de aprendizado de máquina para predições na área de saúde

Campinas
13/07/2022

João Victor Ribeiro dos Santos

Monografia apresentada ao Instituto de Matemática, Estatística e Computação Científica da Universidade Estadual de Campinas como parte dos requisitos para obtenção de créditos na disciplina Projeto Supervisionado, sob a orientação do(a) Prof. Cristiano Torezzan.

Resumo

Modelos de aprendizado de máquina têm impulsionando mudanças de paradigma em processos decisórios nas mais variadas áreas. Na área de saúde, por exemplo, é crescente o interesse por métodos de apoio à decisão baseados em dados para auxiliar os profissionais de saúde em predições ou diagnósticos. Nesse contexto, é de grande interesse que os resultados obtidos pelos algoritmos sejam auditáveis e passíveis de interpretação. Neste projeto estudamos alguns dos principais modelos interpretáveis de aprendizado de máquina e realizamos a interpretação dos resultados de modelos de aprendizado de máquina para a predição de via de parto, com base em dados públicos disponibilizados pelo Sistema de Informação sobre Nascidos Vivos - SINASC. Com os resultados obtidos pudemos compreender o funcionamento de tais métodos a importância da interpretabilidade para a área de aprendizado de máquina.

Conteúdo

1 Introdução

Aprendizado de máquina é um segmento do campo de Inteligência Artificial que reúne métodos matemáticos e algoritmos computacionais que permitem ao computador analisar um determinado conjunto de dados e a partir deles construir uma espécie de programa, usualmente nomeados como modelo, que é capaz de realizar previsões.

Nessa tarefa de aprender a partir dos dados os computadores possuem grande vantagem em velocidade, diferente de nós, conseguem manipular e calcular com exatidão um quantidade enorme de operações em muito pouco tempo e para tudo isso precisam apenas de energia elétrica. Dessa forma o uso e estudo sobre o aprendizado de máquina vem crescendo exponencialmente e abrangendo cada vez mais ramos da atuação humana.

Entretanto, quanto mais importante a decisão do sistema de aprendizado, mais necessário se faz conseguir interpretar como sua previsão é alcançada, tanto para prevenir erros quanto encontrar possíveis aperfeiçoamentos relevantes. A área responsável por interpretar a tomada de decisão da maquina na Inteligência Artificial é formalmente denominada *Interpretabilidade* ou *Explainable Artificial Intelligence(XIA)* (Stuart Russell [2013]).

As técnicas de interpretabilidade podem ser divididas em dois seguimentos, as que observam parâmetros particulares de como o modelo final foi construido, chamados métodos intrínsecos, e as técnicas que explicam o modelo estudando como se da a previsão do modelo em relação as informações que ele recebe, chamados métodos *model-agnostic*.

Neste trabalho estudamos uma referência clássica da literatura dessa área, o livro Molnar [2022], e aplicamos algumas técnicas do tipo *model-agnostic* de interpretabilidade, *Partial Dependence Plot*, *Permutation Importance* e SHAP para interpretabilidade de alguns modelos de aprendizado de máquina para problemas de predição na área da saúde.

2 Aprendizado de Máquina

Antes de abordarmos diretamente sobre o objetivo deste projeto vamos descrever alguns conceitos básicos de aprendizado de máquina. Diferente da "programação comum", onde são passadas instruções detalhadas sobre o que o algoritmo deve retornar

ao receber um dado, no aprendizado de máquina ou Machine Learning (ML) uma "programação indireta" ocorre, onde o algoritmo irá, a partir da analise de uma base de dados, procurar uma solução $\widehat{f}$ que consiga realizar previsões que se aproximem das observadas nos dados.

Uma base de dados é uma matriz, é composta por um conjunto de amostras dispostas nas linhas e um conjunto de atributos, ou variáveis, representados na forma de colunas. Por exemplo, em um banco de dados para cadastro de pacientes, características como "Idade", "Sexo"e "Peso"são atributos e um caso especifico, como um paciente chamado com 21 anos, do sexo Masculino, com 82kg pode ser uma amostra.

Para o treinamento do algoritmo de aprendizado a base de dados é separada entre a coluna com o atributo alvo y, o que a máquina tentara prever, e o restante da base de dados se torna a base de treino X. Usualmente cada uma dessas partes são separadas em outras duas, que seriam as bases X e y de treinamento e de teste, para que se possa analisar a assertividade das previsões quanto a casos desconhecidos. O resultado do treinamento é chamado de modelo f, ele é responsável por realizar as predições $f(x_m)$, onde x_m é a m-ésima amostra de X.

Existem vários algoritmos de aprendizado, cada um tendo sua vantagem ou desvantagem para diferentes objetivos ou estruturas das bases de dados. Os algoritmos pode ser classificado em categorias generalizadas de aprendizado (supervisionado, não supervisionado e por reforço), neste trabalho utilizaremos o supervisionado, onde o algoritmo tem acesso aos dados do atributo alvo durante o treinamento.

3 Interpretabilidade

Diferente do aprendizado humano, no aprendizado de máquina o computador não tem compreensão racional sobre a tarefa que esta realizando. Neste caso, uma função matemática é definida para treinar um algoritmo com base em um conjunto de dados. Nestes casos, o modelo irá carregar vieses inerentes à base de treinamento e às escolhas da função de mérito escolhida. A presença desses vieses influencia diretamente a confiança do modelo.

Devido ao aumento da uso da inteligência artificial, se faz necessário o enten-

dimento das decisões dessas máquinas. Entender como essas predições são geradas pode auxiliar na missão de aperfeiçoar o modelo, melhorando a confiabilidade do mesmo.

De acordo com Molnar [2022], interpretabilidade no aprendizado de máquina visa buscar explicações para o comportamento do modelo de aprendizado na sua tomada de decisão. Quanto maior a interpretabilidade, mais fácil é para entender por que certas predições foram feitas.

As explicações geradas pelas técnicas de interpretabilidade devem ser capaz de informar as principais causas do comportamento do modelo, não necessariamente o entendimento exato de todas as decisões. Um modelo confiável e assertivo em suas predições se torna também uma fonte de informação, além do dados em si, e a interpretabilidade possibilita extrair esse conhecimento.

Podemos resumir as técnicas de interpretabilidade em dois níveis, global e local. Nas técnicas globais, o objetivo é compreender como o modelo realiza as predições num escopo geral, elas descrevem o comportamento esperado do modelo. Técnicas locais, por sua vez, buscam explicar como o modelo realiza predições num caso ou grupo especifico, são os métodos que descrevem as predições para um conjunto de interesse.

Além desta, existe outra ramificação dos métodos de interpretação, intrínsecos e model-agnostic. Os intrínsecos, se referem aos métodos que consistem em avaliar informações fundamentais do modelo a ser analisado, devido a isto apenas algoritmos de aprendizado mais simples podem ser interpretados desta forma, Decision Tree e Linear Regression são alguns exemplos de modelos fundamentalmente interpretáveis. Técnicas model-agnostic consistem em avaliar o modelo depois do treinamento, lidando apenas com as predições retornadas pelo modelo e não com a estrutura do modelo em si, logo, estes métodos conseguem ser aplicados para qualquer tipo de modelo. A seguir iremos apresentar três métodos *model-agnostic*: Partial Dependence Plot, *Permutation Importance* e SHAP que foram utilizados neste estudo.

3.1 Permutation Importance

O método de permutação de atributos consiste em medir a importância de um atributo de acordo com o aumento do erro gerado quando se utiliza uma base de dados com os dados de tal atributo embaralhados. Dessa forma, quão maior o aumento

do erro gerado pelos dados adulterados em comparação com os dados honestos, maior a importância durante o calculo da predição do modelo, em contra partida quando à base com um atributo adulterado gera pouca diferença no erro do modelo, menor a importância daquela variável durante a decisão do modelo é menor.

O método *Permutation Importance* foi proposto por Fisher et al. [2018] e pode ser resumido pelos passos apresentados a seguir Molnar [2022].

Sejam $\widehat{f}(X)$ o modelo treinado, X a base de dados sem o atributo alvo(sendo x uma amostra), y o atributo alvo e $L(y, \widehat{f}(x))$ uma função pra mensurar o erro:

1. Calcular o erro $e_{original} = L(y, \widehat{f}(X))$
2. Para cada variável j entre todas as as variáveis da matriz X:

   Gerar uma matriz adulterada, X_{perm}, a partir da permutação dos dados da coluna j, o que quebra a relação entre tal atributo e o resultado correto y.
3. Calcular erro com base na matriz permutada $e_{perm} = L(y, \widehat{f}(X_{perm}))$
4. Calcular importância de cada atributo como a divisão $I_j = e_{perm}/e_{original}$ ou como a diferença $I_j = e_{perm} - e_{original}$.

Importante ressaltar que quando permutamos uma variável do banco de dados não quebramos somente a relação de tal variável com a resposta, mas também a com as outras variáveis, durante a decisão do modelo. Logo, quando calculamos a importância de permutação para muitos atributos, as correlações entre eles também vão ser consideradas no calculo da importância para cada variável. Por exemplo, considere um modelo confiável que tem como tarefa prever se uma pessoa tera complicações graves se contrair uma certa doença, suponha que já se sabe que o peso e idade, conjuntamente, tem um grande influência no diagnostico médico da doença. Neste caso, quando calcularmos a importância de permutação individualmente para peso ou idade no modelo em análise, o resultado vai considerar, além da influência individual da variável, também a importância da relação das variáveis.

Uma das vantagens de tal método é que ele provem uma compreensão global sobre o comportamento do modelo, pois a importância resultante será a média do erro

excedente, como o erro é dependente da base de dados utilizada no método, logo tal resultado tem influência de todos os casos do banco de dados. Além disso, a permutação de atributos pode poupar muito tempo em relação a outras técnicas existentes, como precisamos apenas do modelo a ser avaliado e um banco de dados, para calcular os erros em sequencia a importância, não é necessário um novo treinamento ou reconstrução do modelo.

Em contra partida, devido ao calculo do erro, é necessário ter conhecimento sobre o que seria a resposta correta(y) para comparar com a predição($\widehat{f}(x)$). Outro ponto negativo é que ao embaralhar os dados de um atributo, com outras variáveis correlacionadas, podemos gerar informações incondizentes com a realidade e uma importância segregada entre as variáveis associadas. Considerando o modelo do ultimo exemplo, utilizado para tentar diagnosticar uma doença, saiba que na predição ele utiliza também as informações do tipo sanguíneo do pai e do paciente, quando realizamos a permutação, podemos ter novas amostras onde os tipos sanguíneos do pai e filho não condizem biologicamente e além disso a importância do tipo sanguíneo possivelmente seria partilhada entre as duas variáveis, logo é importante se atentar a como interpretar a permutação desses tipos de variáveis. Podemos pensar na importância de permutação de variáveis correlacionadas, também como erro aumentou utilizando dados que não serão observados em casos reais.

3.2 Partial Dependence Plot

Partial Dependence Plot, ou PDP, é um procedimento muito útil no entendimento do comportamento do modelo. Sua função é gerar gráfico para a visualização de como se da a média das previsões ($\widehat{f}(x)$) para cada valor das variáveis selecionada (Friedman [2001]).

A dependência parcial $\widehat{f_s}$, também conhecida como por método de Monte Carlo, é definida por:

$$\widehat{f_s}(x_s) = \frac{1}{n}\sum_{i=1}^{n} \widehat{f}(x_s, x_c^{(i)})$$

Onde S é conjunto de variáveis de interesse e $\widehat{f}(x)$ diz respeito ao modelo.

Nesta formula x_S representa um possível valor para as variáveis de S e $x_C^{(i)}$ contem o valor do resto dos atributos que o modelo $\widehat{f}$ precisa, ou seja $x_S \cup x_C = x$. O n representa o número total amostras do banco de dados.

Ou seja, o PDP substitui todos os valores originais das variáveis contidas em S por x_S , calculas a média de todos essas predições realizadas a partir desse banco de dados modificado e gera um gráfico dessas médias para todos os x_S possíveis.Com o esse método podemos visualizar o comportamento médio das predições modelo f para todos os valores de x_S. Devido as limitações dimensionais de exibição, usualmente o conjunto S contem uma ou duas variáveis.

Justamente por retornar o comportamento das predições em forma de um simples gráfico, existe a vantagem desse resultado ser intuitivo e de fácil compreensão. O PDP também é de fácil implementação, por precisar essencialmente do modelo e um banco de dados, mas note que ele não utiliza o valor correto y, ele depende somente das respostas que o modelo entrega.

Semelhantemente ao *Permutation Importance*, o caso de variáveis correlacionadas também pode ser um problema. Como os valores da base de dados são substituídos por x_S durante o calculo das predições, podemos construir uma amostra que não condiz com a realidade e utilizar esse resultado irreal no calculo da média de predições. Logo, quanto mais variáveis independentes para o modelo, com melhor precisão o PDP irá demonstrar o comportamento do modelo para os dados que ele realmente pode receber.

3.3 SHAP

O método SHAP (SHapley Additive exPlanations) é um método de interpretabilidade que pertence ao grupo dos *Additive Feature Attribuition Methods*, tal classe consiste de métodos que, para compreender uma previsão do modelo $f(x)$ de maneira local, constroem um novo modelo de interpretação $g(z)$ com o intuito de realizar uma aproximação da previsão do modelo $f(x)$. através da analise do modelo auxiliar g podemos estudar como foi realizada a predição do modelo principal (Lundberg and Lee [2017]).

Additive Feature Attribuition Methods possuem um modelo de interpretação que é uma função linear de variáveis binarias:

$$g(z') = \phi_0 + \sum_{i=1}^{M} \phi_i z'_i,$$

onde $z' \in 0, 1^M$, M é o número de atributos do modelo de interpretação, e $\phi_i \in \mathbb{R}$.

Nessa formula, ϕ_i representa a influencia que cada atributo i teve sobre a predição e z' é um vetor binário de controle sobre quais efeitos serão somados. O ϕ_0, no método SHAP, é igual ao valor esperado do modelo $f(X)$, ou a esperança ($E(f(x))$), e os efeitos ϕ_i são somados a partir desse resultado inicial, ao somar todos temos $g(z) \approx f(x)$ para uma amostra x especifica. No SHAP o calculo do ϕ_i tem origem da teoria *Shapley Value*.

Nomeado em homenagem ao Lloyd Shapley, que introduziu o assunto em 1951, *Shapley Value* vem como uma ideia de solução na teoria de jogo cooperativo (*coalitional game theory*). De acordo com Lundberg and Lee [2017], este método mede a importância de um atributo para modelos lineares na presença de correlações entre as variáveis. A ideia é que as variáveis de uma amostra são "jogadores"e a "pontuação final", respectiva a atuação de todos os jogadores é a predição do modelo é uma pontuação proveniente das ações dos jogadores, que são as variáveis daquela amostra. Uma maneira de imaginar é que o jogo é um cabo de guerra e cada variável pode atuar aplicando "força"em um dos lados da "corda", aumentando ou diminuindo o valor da predição final.

Resumidamente, o shapley value da variável i, na predição de um x especifico, corresponde á diferença que teríamos no resultado do modelo caso ele não tivesse acesso á essa variável, os cálculos necessário para alcançar esse resultado não serão abordados neste trabalho devido a sua complexidade, entretanto eles podem ser encontrados no artigo Lundberg and Lee [2017].

Podemos visualizar as importância distribuídas pelo método SHAP como um gráfico de barras onde cada barra representa a influência do atributo i naquela previsão. A soma dessas influencias resulta no valor da resposta do modelo. Além disse pode se utilizar os Shapley values para uma analise global, basta realizar o método SHAP para os n casos da sua base de dados, então você terá como resultado uma matriz $\Phi \in \mathbb{R}^{(n \times j)}$ onde cada elemento $\phi_{m,i}$, tal que $m \in \{1, ..., n\}$ e $i \in \{1, ..., j\}$, equivale ao *Shapley value*

encontrado para cada variável i e amostra m, assim, tirando a media de todos os valores da coluna i da matriz, conseguimos calcular a importância média desta variável. Com isto podemos construir um gráfico para visualizar a importância média de todos os atributos. método de interpretabilidade, baseado na teoria *Shapley value*, a partir destes valores de importância, é possível criar análises locais e blogais sobre o comportamento do modelo

4 Estudo de caso

Nesta seção apresentamos alguns resultados obtidos por meio de um estudo de caso baseado em dados reais oriundos de uma base pública. O objetivo da seção é aplicar os métodos de interpretabilidade apresentados anteriormente para investigar a importância dos atributos (variáveis) utilizadas na predição. . Utilizaremos na implementação dos algoritmos a linguagem Python [b].

4.1 Descrição do conjunto de dados

Em nosso estudo utilizamos o Sistema de Informação sobre Nascidos Vivos SINASC [2010], disponibilizado de forma aberta pelo SUE e baixado por meio do pacote PySUS (Coelho [2016]). Esse sistema de dados disponibiliza bancos de dados sobre os nascimentos ocorridos no brasil divididos por estados, aqui utilizaremos a Base de dados de São paulo, foram selecionadas as seguintes variáveis dos casos ocorridos entre 2016 e 2020:

- **IDADEMAE**: Idade da mãe
- **IDADEPAI**: Idade do pai
- **QTDGESTANT**: Número de Gestações Anteriores
- **QTDPARTNOR**: Número de partos normais
- **QTDPARTCES**: Número de partos cesariana
- **QTDFILVIVO**: Número de filho vivo
- **QTDFILMORT**: Número de filho morto

- **ESCMAE2010**:Escolaridade da mãe (sem escolaridade, fundamental 1, fundamental 2, ensino médio, superior incompleto, superior completo)
- **ESTCIVMAE**: Estado civil da mãe (solteira, casada, viúva, divorciada ou união estável)
- **RACACORMAE**: Raça da mãe (branca, preta, amarela, parda ou indígena)
- **PARTO**: Tipo de parto (normal ou cesário)

Após esta seleção, as amostras com valores faltantes foram removidas para não comprometer o funcionamento do modelo, além disso, as variáveis categóricas foram formatadas para variáveis binarias, onde cada categoria da variável terá seu próprio atributo e um valor 1 para este atributo representa que aquela amostra pertence a esta categoria. Neste ponto nossa base de dados já possui 1.072.629 casos de parto cesário e 457.184 casos de parto normal, para melhorar a qualidade do modelo, iremos igualar a quantidade de cada caso.

Então, depois deste tratamento, terminamos com uma base de dados de 21 variáveis e 914.368 amostras.

No tratamento da base de dados foi utilizada a biblioteca Python [a].

4.2 Apresentação do problema de predição

O problema que será utilizado como base para o estudo consiste na previsão da via de parto, normal ou cesariana. Assim, o vetor alvo y, será o atributo "PARTO"e as demais variáveis da base serão denotadas por X. As amostras serão divididas nos conjuntos de X_{treino}, X_{teste}, y_{treino} e y_{teste} na proporção de 50%.

Em seguida as bases de dados de treino foram utilizadas para construir os modelos: *Logistic Regression* e *Decision Tree*. Na implementação dos modelos de aprendizado, foi utilizada a biblioteca *sckiti-learn*.

4.3 Resultados dos modelos de predição

Após o treinamento dos modelos, tivemos os seguintes resultados de precisão sobre previsão correta, estimadas a partir da base de treino. Para mensurar o erro foi

utilizaremos a proporção de acertos para predições realizadas e um método denominado matriz de confusão, este método gera um gráfico que mostra, para todas as possíveis combinações de previsão $f(x)$ e resposta correta y, mostra a quantidade de vezes que aquela combinação aconteceu na bateria de testes.

4.3.1 Regressão logística

O modelo de regressão logística apresentou uma precisão de 73% com seus erros distribuídos como apresentado na figura 1:

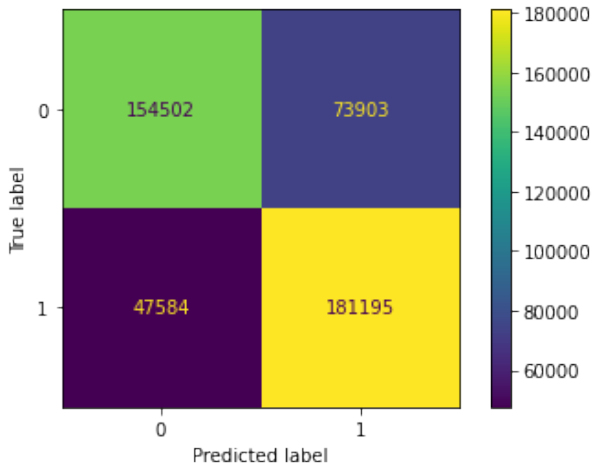

Figura 1: Gráfico da matriz de confusão do modelo de regressão logística, onde valor 1 corresponde à parto cesário e 0 ao parto normal, o eixo horizontal representa o valor predito e o eixo vertical diz respeito à resposta verdadeira

4.3.2 Decision Tree

O modelo de Decision Tree, ou arvore de decisão, apresentou uma precisão de 72% e gerou a matriz de confusão exibida na figura 2:

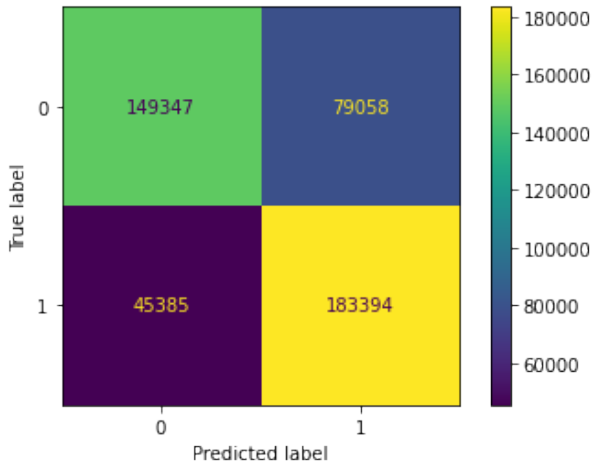

Figura 2: Gráfico da matriz de confusão do modelo da arvore de decisão, onde valor 1 corresponde à parto cesário e 0 ao parto normal, o eixo horizontal representa o valor predito e o eixo vertical diz respeito à resposta verdadeira

4.4 Modelos de interpretabilidade

Nesta seção iremos finalmente implementar as técnicas de interpretabilidade abordadas anteriormente, utilizaremos os modelos de regressão logística e arvore de decisão já construídos, além disso, será usada a base de teste para tal implementação.

4.4.1 Permutation Importance

Executando o *Permutation Importance* inicialmente no modelo da arvore de decisão tivemos o resultado apresentado na figura 3.

Vemos que, no modelo da arvore de decisão que criamos, a maior importância de permutação vem da variável "QTDPARTNOR", as próximas variáveis com mais influência possuem valor de importância consideravelmente menor, enquanto grande parte das variáveis não provoca uma diferença no erro considerável quando permutada, lembrando que a importância de permutação é calculada a partir das discrepâncias entre o erro gerado pela permutação e o erro padrão do modelo.

Agora veja, na figura 4, a tabela de importância de permutação do modelo de regressão logística.

Em comparação com o outro modelo, a regressão logística apresentou valores menores para importância das variáveis mais relevantes, porém a distribuição das im-

Weight	Feature
0.1622 ± 0.0017	QTDPARTNOR
0.0198 ± 0.0004	IDADEMAE
0.0157 ± 0.0003	ESTCIVMAE_casada
0.0129 ± 0.0004	RACACORMAE_branca
0.0129 ± 0.0004	QTDGESTANT
0.0057 ± 0.0002	QTDPARTCES
0.0013 ± 0.0004	RACACORMAE_parda
0 ± 0.0000	ESCMAE2010_fund1
0 ± 0.0000	RACACORMAE_indigena
0 ± 0.0000	RACACORMAE_amarela
0 ± 0.0000	ESTCIVMAE_viuva
0 ± 0.0000	ESTCIVMAE_un_estav
0 ± 0.0000	ESCMAE2010_sup_incomp
0 ± 0.0000	ESTCIVMAE_divorc
0 ± 0.0000	ESCMAE2010_sup_compl
0 ± 0.0000	ESCMAE2010_sem_escolar
0 ± 0.0000	ESCMAE2010_medio_compl
0 ± 0.0000	ESCMAE2010_fund2
0 ± 0.0000	ESTCIVMAE_solteira
0 ± 0.0000	RACACORMAE_preta
... 1 more ...	

Figura 3: Importância de permutação para as variáveis mais influentes no modelo da arvore de decisão

portância esta bem mais equilibrada neste caso. Agora a maior relevância vem da variável "QTDGESTANT", a variável "QTDPARTNOR"recebeu uma das maiores importâncias juntamente com "ESTCIVMAE_casada"e "RACACORMAE_branca".

Na implementação deste método, foi utilizada a biblioteca eli5.

4.4.2 Partial Dependence Plot

Como já abordamos anteriormente, o *Partial Dependence Plot* de um modelo sobre a variável i constrói um gráfico onde cada ponto (x, y) da curva representa a média do resultados das previsões (y) quando é passada uma base de treino com todas os valores da variável i são substituídos por x.

Executamos o PDP para algumas das variáveis classificadas como relevantes na implementação do método *Permutation Importance*, no modelo da arvore de decisão analisamos a variável "QTDPARTNOR"e no modelo de regressão logística as variaveis "IDADEMAE", "QTDPARTNOR"e "QTDPARTCES"foram analisadas.

O comportamento do modelo em relação a variavel "QTDPARTNOR"está apresentado na figura 5. Vemos que o modelo tem uma associação negativa com a variável, além disso a partir de um valor maior igual a 1, para quantidade de parto normal anterior,

Weight	Feature
0.0520 ± 0.0008	QTDGESTANT
0.0453 ± 0.0010	QTDPARTNOR
0.0424 ± 0.0006	ESTCIVMAE_casada
0.0322 ± 0.0005	RACACORMAE_branca
0.0279 ± 0.0002	QTDPARTCES
0.0201 ± 0.0004	ESCMAE2010_sup_compl
0.0153 ± 0.0004	ESTCIVMAE_solteira
0.0140 ± 0.0006	ESCMAE2010_medio_compl
0.0137 ± 0.0007	IDADEMAE
0.0102 ± 0.0004	RACACORMAE_parda
0.0034 ± 0.0001	IDADEPAI
0.0030 ± 0.0002	ESCMAE2010_fund1
0.0023 ± 0.0001	ESTCIVMAE_un_estav
0.0023 ± 0.0001	ESCMAE2010_sup_incomp
0.0021 ± 0.0002	ESTCIVMAE_divorc
0.0004 ± 0.0001	RACACORMAE_amarela
0.0002 ± 0.0001	ESTCIVMAE_viuva
0.0002 ± 0.0001	RACACORMAE_indigena
0.0001 ± 0.0000	ESCMAE2010_sem_escolar
-0.0000 ± 0.0001	RACACORMAE_preta
... 1 more ...	

Figura 4: Importância de permutação para as variáveis mais influentes no modelo de regressão logística

a curva da dependência se estabiliza em aproximadamente -40%. Ou seja, neste modelo, ao possuir quantidade de parto normal anterior maior igual a 1 representa uma diminuição de 40% na chance da gestação atual ser cesário.

Agora, em relação ao modelo de regressão linear, temos os *Parcial Dependence Plot* das variáveis "IDADEMAE", "QTDPARTNOR"e "QTDPARTCES"apresentadas nas figuras 6, 7 e 8 respectivamente. No PDP da idade da mãe vemos que o modelo apresenta um comportamento aparentemente linear em relação a idade da mãe. Para a quantidade de partos normais e cesariana tem observa se um comportamento semelhantemente oposto, a partir de 6 partos as curvas das duas variáveis se estabilizam em 50%, porém enquanto a maior quantidade de partos cesários resulta um comportamento mais próximos da classificação de parto cesário a quantidade de partos normais resulta o comportamento oposto.

Durante a implementação do método *Partial Dependence Plot*, foi utilizada a biblioteca PDPbox.

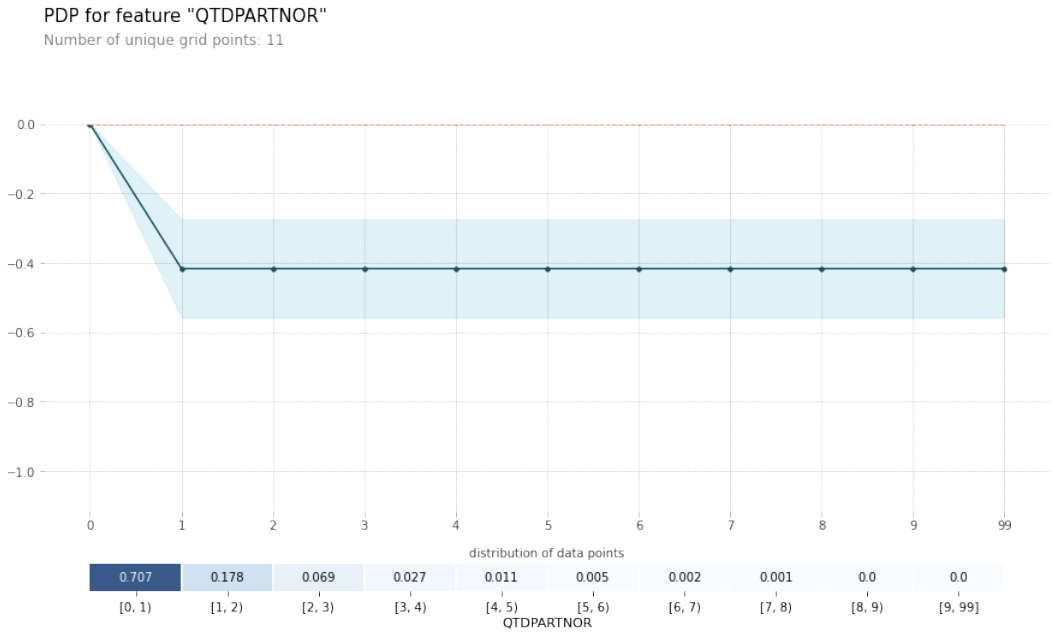

Figura 5: *Partial Dependence Plot* do modelo da arvore de decisão em relação a variável "QTDPARTNOR"

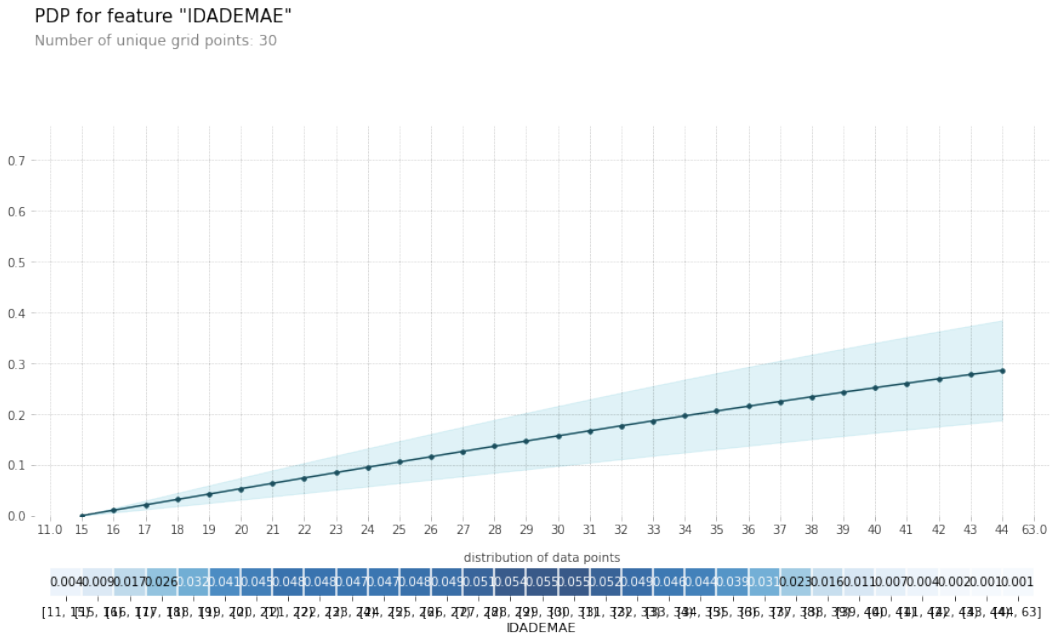

Figura 6: *Partial Dependence Plot* do modelo de regressão logística em relação a variavel "IDADEMAE"

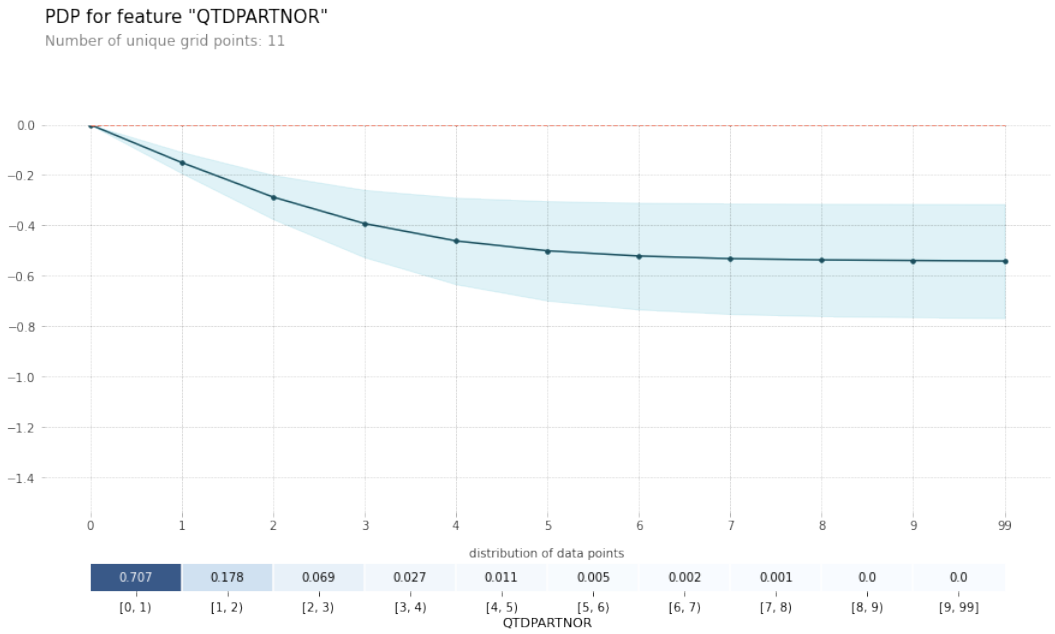

Figura 7: *Partial Dependence Plot* do modelo de regressão logística em relação a variavel "QTDPARTNOR"

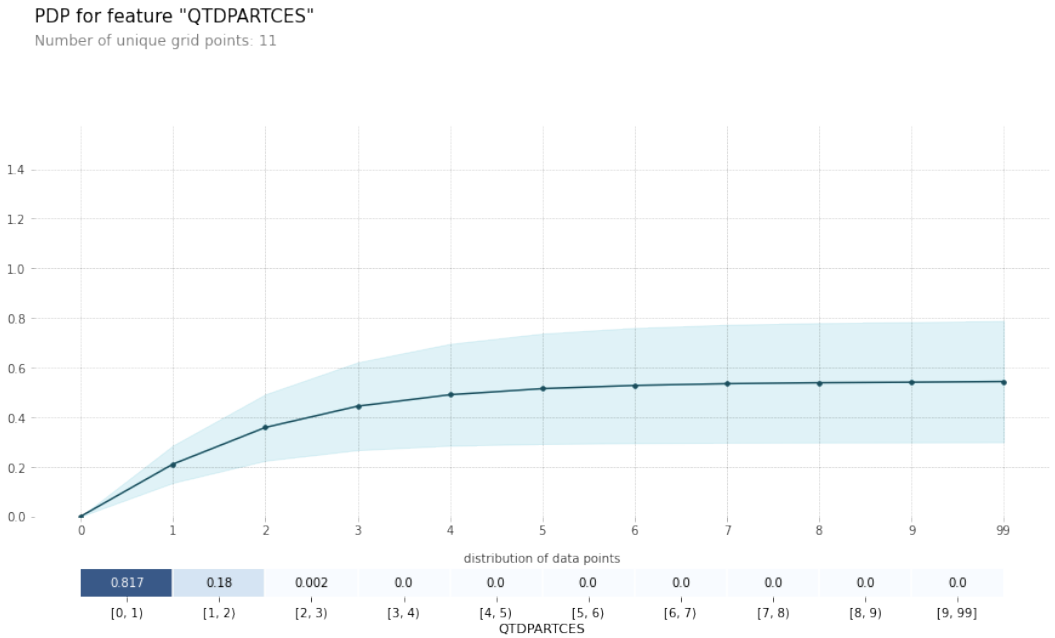

Figura 8: *Partial Dependence Plot* do modelo de regressão logística em relação a variavel "QTDPARTCES"

4.4.3 SHAP

Agora iremos implementar o ultimo método de interpretação abordado, o SHAP. Nesta implementação foi utilizado apenas o modelo de regressão logística.

Na implementação do método SHAP foi utilizada a biblioteca *shap*. Geramos os gráficos resultantes dos valores Shapley para duas amostras especificas na figura 9 e 10. Observe que mesmo para valores iguais da mesma variável, o shapley value não necessariamente será o mesmo, o que faz parte do esperado sendo que o SHAP realiza uma analise sobre um caso especifico, logo podem se encontrar influências diferentes sobre a predição a depender da amostra predita.

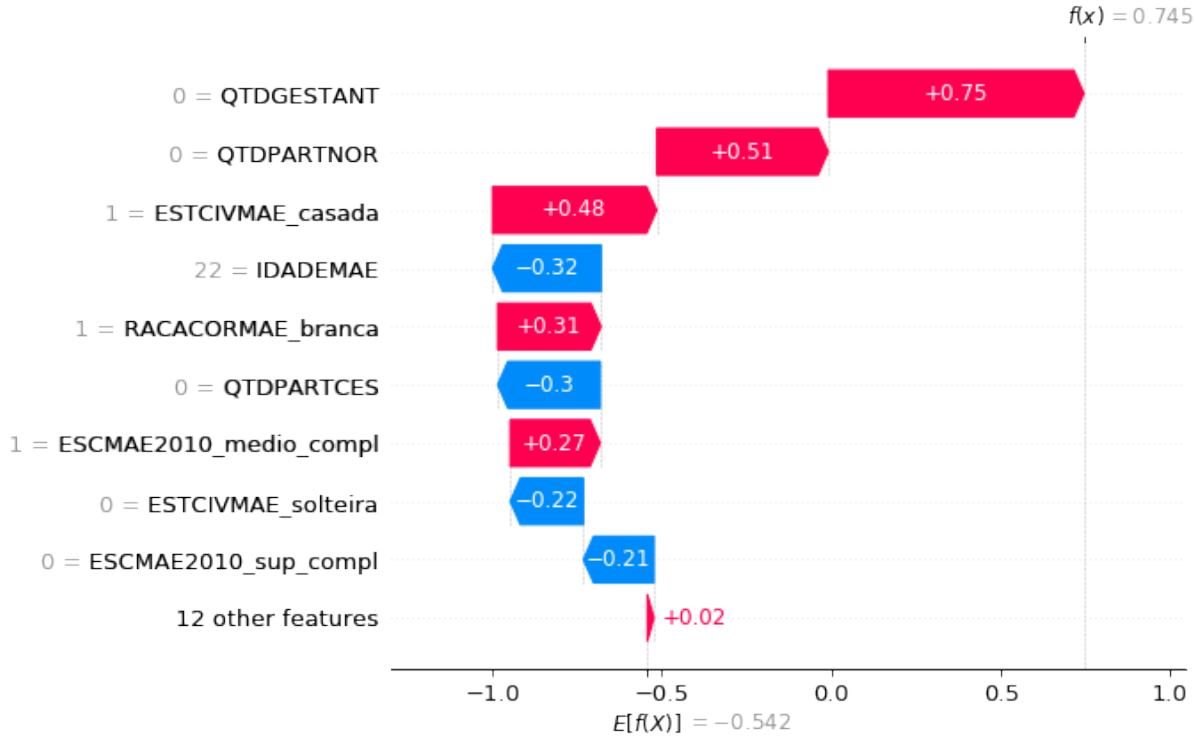

Figura 9: Gráfico dos *SHAP values* para o caso de uma mulher de 22 anos, casada, branca, de ensino médio completo e que nunca teve filhos

Podemos tentar utilizar os Shapley values para uma interpretação global sobrepondo os shap values para varios casos e vizualizar quais são os valores SHAP para cada valor possivel das variaveis. A figura 11 representa esta abordagem e nela vemos que as variaveis com maior influência shap, num caso geral, são "QTDPARTCES", "QTDPARTNOR", "QTDGESTANT".

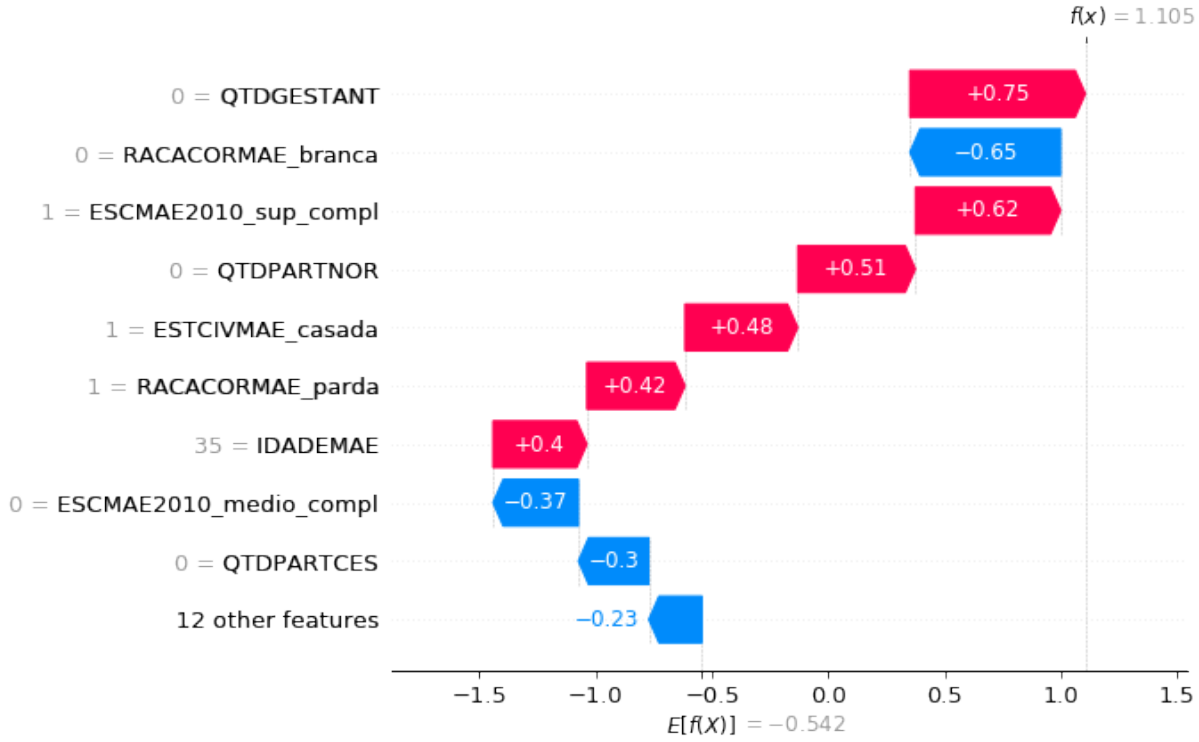

Figura 10: Gráfico dos *SHAP values* para o caso de uma mulher de 35 anos, casada, parda, de ensino superior completo e que nunca teve filhos

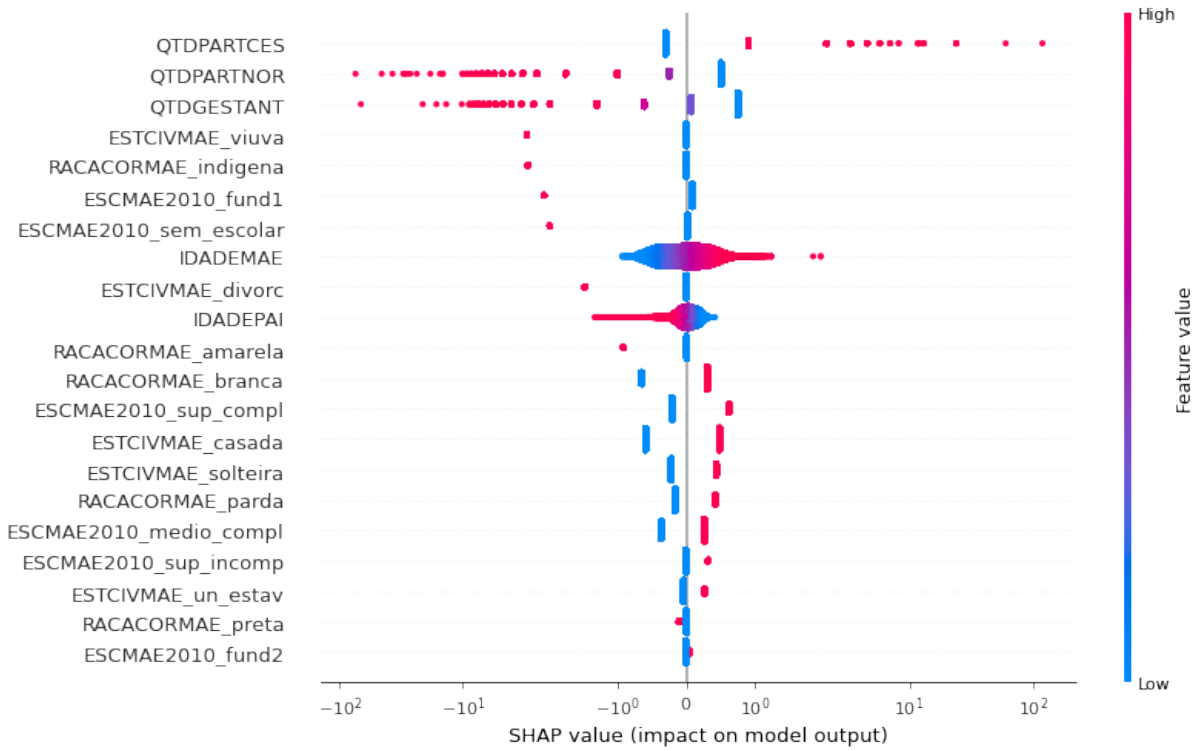

Figura 11: Gráfico dos *SHAP values* sobrepostos para cada atributo da base de teste

5 Considerações finais e perspectivas

Neste trabalhos estudamos o tema de interpretabilidade no aprendizado de máquina, com atenção especial em três métodos de interpretação: *Permutaion Importance*, *Partial Dependence Plot* e *SHAP*. A partir da base de dados SINASC [2010], construímos dois modelos genéricos e com base neles aplicamos os três métodos abordados. Os resultados obtidos permitiram compreender o funcionamento dos métodos de interpretabilidade e identificar algumas variáveis importantes para o modelo de predição. Por exemplo, foi possível visualizar a relação aproximadamente linear entre a idade da mãe e a predição do modelo.

Para futuros trabalhos construiremos modelos de aprendizado mais elaborados visando obter resultados mais robustos sobre interpretabilidade. Dentre os problemas que pretendemos abordar, ainda com base de dados do SINASC, estão a predição de peso ao nascer e predição de prematuridade.

Referências

Flavio Codeco Coelho. PySUS pacote python. https://pypi.org/project/PySUS/, 2016. Accessed: 2022-07-13.

eli5. eli5 library. https://eli5.readthedocs.io/en/latest/. Accessed: 2022-07-17.

Aaron Fisher, Cynthia Rudin, and Francesca Dominici. All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. 2018. doi: 10.48550/ARXIV.1801.01489. URL https://arxiv.org/abs/1801.01489.

Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189–1232, 2001. ISSN 00905364. URL http://www.jstor.org/stable/2699986.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

Christoph Molnar. *Interpretable Machine Learning*. 2 edition, 2022. URL https://christophm.github.io/interpretable-ml-book.

PDPbox. Pdpbox library. https://pypi.org/project/PDPbox/. Accessed: 2022-07-17.

Python. Python. https://www.python.org/, a. Accessed: 2022-07-17.

Python. Python. https://www.python.org/, b. Accessed: 2022-07-13.

SINASC. SINASC sistema de informação sobre nascidos vivos. http://sinasc.saude.gov.br/, 2010. Accessed: 2022-07-13.

Peter Norvig Stuart Russell. *Artificial Intelligence, A modern Approach*. 3 edition, 2013.