Universidade Federal do Espírito Santo
Programa Institucional de Iniciação Científica
Relatório Final de Pesquisa
Ciências Exatas e da Terra

Interpretabilidade em modelos preditivos na área da saúde

Edital:	Edital PIIC 2020 /2021
Grande Área do Conhecimento (CNPq):	Ciências Exatas e da Terra
Área do Conhecimento (CNPq):	Probabilidade e Estatística
Título do Projeto:	Ciência de Dados e Aprendizado Estatístico Aplicados à Saúde
Título do Subprojeto:	Interpretabilidade em modelos preditivos na área da saúde
Professora Orientadora:	Agatha Sacramento Rodrigues
Estudante:	Ornella Scardua Ferreira

Resumo

Na abordagem preditivista em ajuste de modelos, o interesse consiste na construção de uma regra para predizer novas observações. Em uma classe de modelos candidatos, é escolhido aquele que apresenta o melhor desempenho preditivo. O modelo vencedor pode ser um modelo não-interpretável, ou seja, um modelo cujas decisões não podem ser explicadas (modelo caixa-preta). No entanto, destaca-se a importância em entender os fatores que levam a certas previsões ou comportamentos, além de ser mais fácil para os humanos confiarem em um sistema que explique suas decisões. Nesse cenário, métodos de interpretabilidade se fazem necessários, uma vez que podem ser aplicados a qualquer modelo preditivo previamente ajustado. Com esse estudo se objetiva estudar métodos de interpretabilidade globais e individuais, comparando-os em uma aplicação da área da saúde e discutindo suas vantagens e desvantagens. Os métodos de interpretabilidade serão aplicados a um conjunto de dados real do Departamento de Obstetrícia da Faculdade de Medicina da Universidade de São Paulo (FM-USP), para os quais ferramentas para analisá-los serão concedidas por meio de documentação computacional bem formulada, disponível no GitHub, e por uma plataforma web via aplicativo Shiny. Com esse projeto, espera-se discutir a utilização de métodos de interpretabilidade na área da saúde, assentindo que o melhor modelo para um problema em questão seja utilizado, ainda que este seja um modelo caixa-preta.

Palavras-chave: Aplicações na área da saúde. Aplicativo Shiny. Métodos de interpretabilidade individuais. Métodos de interpretabilidade globais. Modelos caixa-preta.

1 Introdução

Em *Machine Learning* (ML, em português Aprendizado de Máquina), as análises são embasadas por modelos estatísticos e algoritmos computacionais, muita vezes sofisticados, cujos resultados, predições ou decisões, decorrem do aprendizado direto sobre os dados. ML pode ser supervisionado ou não-supervisionado (Morettin & Singer, 2020), e consideramos aqui ML supervisionado, o qual engloba modelos para estudar o valor de uma variável resposta (*output*, *label* ou desfecho) a partir de cováriaveis (*input*, *features*, variáveis explicativas ou preditoras). Em ML supervisionado, Breiman (2001) argumenta que uma distinção entre duas culturas precisa ser feita. A primeira cultura, chamada de modelo explicativo (ou modelo inferencial), foca na interpretação dos parâmetros envolvidos

do modelo e testa hipóteses para entender a relação entre as covariáveis e a variável resposta. Na segunda cultura, chamada de *algorithmic modeling culture* por Breiman, o principal objetivo é a construção de um modelo (regra) para predizer novas observações (Izbicki & Santos, 2020).

Em particular, quando o objetivo é predição, as escolhas no processo de modelagem são guiadas por medidas de desempenho preditivo e acurácia do modelo (capacidade em acertar ou errar uma predição dentro de um limiar aceitável (James et al., 2013). Nesse caso, o modelo escolhido é aquele com melhor desempenho preditivo e acurácia, e esse "melhor" pode ser um modelo não-interpretável, ou seja, um modelo em que as decisões não podem ser explicadas, visto que seu funcionamento interno não pode ser facilmente acessado. No entanto, mesmo com o intuito de predição, muitos pesquisadores e médicos têm o interesse em entender as variáveis no modelo e discutir o porquê delas e suas relações nesse modelo preditivo. Para facilitar o aprendizado sobre os fatores mais importantes de certas previsões ou comportamentos, a interpretabilidade e as explicações das decisões são cruciais (Molnar, 2019). Em razão disso, foram propostos métodos de interpretabilidade, individuais e globais, que podem ser aplicados a qualquer modelo preditivo ajustado. Se as decisões de um modelo de ML podem ser explicadas, as seguintes questões podem ser checadas mais facilmente (Doshi-Velez & Kim, 2017): 1) justiça (*fairness*) – garantir que as previsões sejam imparciais, não viesadas e que não discriminem grupos sub-representados; e 2) confiança - é mais fácil para os humanos confiarem em um sistema que explica suas decisões.

Nesse sentido, métodos de interpretabilidade em uma aplicação real da área da medicina obstétrica serão discutidos neste projeto. Nesta aplicação, ao predizer, no momento do diagnóstico de diabetes gestacional, se uma gestante fará o uso de insulina em algum momento antes do parto, deseja-se entender o motivo por que ocorreu tal desfecho. Essa aplicação é resultante de pesquisas realizadas no Departamento de Obstetrícia da Faculdade de Medicina da Universidade de São Paulo (FM-USP), em particular pelo ambulatório de diabetes gestacional.

Esse subprojeto é ligado ao projeto de pesquisa "Ciência de Dados e Aprendizado Estatístico Aplicados à Saúde" (número de registro 10225/2020 e certificado no CNPq), coordenado pela professora orientadora e que se objetiva em aplicar e desenvolver metodologias na área de Ciência de Dados para resolver problemas e inovação na área da saúde.

Esse relatório está organizado como segue: na Seção 2 estão os objetivos geral e específicos e na Seção 3 é apresentado o embasamento teórico. A Seção 4 é dedicada aos modelos de ML e à descrição dos métodos de interpretabilidade propostos e na Seção 5 são apresentados os resultados dos ajustes dos modelos e das aplicações dos métodos de interpretabilidade sobre os dados reais. Por fim, na Seção 6 estão as considerações finais.

2 Objetivos

O objetivo geral deste projeto é avaliar alguns métodos de interpretabilidade sobre um modelo preditivo caixa-preta em uma aplicação real da área da saúde.

São os objetivos específicos:

1. avaliar os modelos preditivos interpretáveis e não-intepretáveis, discutindo suas aplicações em problemas da área da saúde;
2. comparar os métodos de interpretabilidade, identificando suas vantagens e desvantagens sobre um modelo preditivo previamente ajustado;
3. discutir a *fairness* e confiança a partir das decisões do modelo preditivo;

4. documentar os pacotes e códigos computacionais dos métodos de interpretabilidade para a aplicabilidade dos métodos estudados;
5. desenvolver uma plataforma *web* por meio de um aplicativo Shiny para melhor acessibilidade dos métodos de interpretabilidade para os usuários; e
6. apresentar e discutir os resultados para os grupos de pesquisa do Departamento de Obstetrícia da FM-USP.

3 Embasamento Teórico

O embasamento teórico das técnicas e modelos de *Machine Learning* é baseado nos trabalhos de Izbicki & Santos (2020), James et al. (2013) e Morettin & Singer (2020). Amorim (2019) e Molnar (2019) foram quem fundamentam a parte teórica e prática acerca dos métodos de interpretabilidade individuais (para cada observação) e globais (quando a interpretação é feita em termos de média) considerados neste projeto. Vamos descrever os métodos de interpretabilidade considerados de forma resumida nas subseções subsequentes.

3.1 Gráfico de Dependência Parcial

O gráfico de dependência parcial é a representação gráfica do efeito marginal que uma ou até duas covariáveis têm sobre o resultado predito de um modelo de ML (Friedman, 2001). A partir de um modelo ajustado $f(\mathbf{X},\mathbf{C})$, em que $\mathbf{X}$ é a matriz de covariáveis de interesse e $\mathbf{C}$ é a matriz das covariáveis restantes, o algoritmo inerente a este gráfico compreende os seguintes passos:

1. fixe valores para $\mathbf{X}$;
2. para cada valor x_i, $i = 1,\ldots,M$, fixado:
   2.1 substitua o valor observado de $\mathbf{X}$ por x_i em cada uma das n observações da amostra e calcule o valor predito do modelo, i.e., $\hat{f}(x_i,c)$;
   2.2 calcule a média das n predições:
$$\bar{f}_{x_i}(x_i) = \frac{1}{n}\sum_{j=1}^{n}\hat{f}(x_i,c_j); \tag{1}$$
3. construa o gráfico dos valores x_i contra as médias das predições $\bar{f}(x_i)$.

A função de dependência parcial, dada na Equação (1), calcula o efeito médio das covariáveis de $\mathbf{X}$ ao marginalizar a distribuição das predições sobre as covariáveis para as quais não se tem interesse em explicar, pois, dessa forma, a função se torna dependente apenas das covariáveis de $\mathbf{X}$. A grande vantagem desse gráfico está em sua interpretação, que é intuitiva e causal. No entanto, essa interpretação só é coerente quando os preditores $\mathbf{X}$ e $\mathbf{C}$ não são correlacionados. Em caso de se haver correlação, a média pode ser influenciada por valores que não fazem sentido para todas as observações, podendo causar interpretações pouco prováveis ou mesmo irreais.

3.2 Gráfico da Esperança Condicional Individual

Enquanto gráficos de dependência parcial expressam a relação média de uma variável com o resultado predito, em gráficos da esperança condicional individual o grau de dependência entre variável e predição é considerado individualmente, isto é, para cada observação. De outro modo, a curva de um gráfico de dependência parcial é exatamente a média das curvas de um gráfico da esperança condicional individual. Assim, o algoritmo desse gráfico se restringe aos passos 1. e 2.1 do algoritmo do gráfico de dependência parcial, em que cada umas das

n curvas é formada pelo ponto $\left(x_i, \hat{f}(x_i, c_j)\right), i = 1,\ldots,M,\ j = 1,\ldots,n$. Sendo possível quantificar as relações separadamente, a interpretação é ainda mais direta e as relações heterogêneas de covariáveis são mais facilmente identificáveis. Apesar disso, o problema decorrente de uma possível correlação entre a variável sob investigação e as demais é o mesmo dos gráficos de dependência parcial.

3.3 Gráfico de Efeitos Locais Acumulados

O gráfico de efeitos locais acumulados tem o mesmo objetivo do gráfico de dependência parcial: calcular o efeito médio das covariáveis sobre a predição de um modelo. No entanto, esse método atua sobre a distribuição condicional das covariáveis, acumulando em uma grade a média das diferenças entre predições condicionadas a pequenas variações nos valores de um dado x da matriz de covariáveis $\mathbf{X}$. Por essa razão, o gráfico de efeitos locais acumulados é uma alternativa ao gráfico de dependência parcial, pois o efeito estimado para as covariáveis de $\mathbf{X}$ não é interferido por valores de variáveis correlacionadas. A partir de uma matriz $\mathbf{X}$ de interesse, a construção do gráfico de efeitos locais acumulados se dá pelo seguinte algoritmo:

1. divida $\mathbf{X}$ em M intervalos;
2. calcule os efeitos locais para cada uma das m_i observações dentro do i-ésimo intervalo: $\hat{f}_{d_{ij}} = (x_i^+ - c_j) - \hat{f}(x_i^- - c_j),\ j = 1,\ldots,m_i$, com x_i^+ e x_i^- correspondendo, nesta ordem, aos limites superior e inferior do intervalo i;
3. calcule a média acumulada para cada valor x de $\mathbf{X}$: $\bar{f}_a(x) = \sum_{i=1}^{k(x)} \frac{1}{m_i} \sum_{j=1}^{m_i} \hat{f}_{d_{ij}}$, em que $k(x)$ é o índice do intervalo que x faz parte; e
4. calcule o valor centralizado de $\bar{f}_a(x)$ para todas as n observações: $\bar{f}_c(x) = \bar{f}_a(x) - \frac{1}{n}\sum_{i=1}^{n} \hat{f}_a(x_i)$.

No passo 2., os efeitos são considerados locais porque a diferença entre predições significa o quanto o resultado predito é influenciado por variações em torno de x. No passo 3., o efeito médio do preditor $\mathbf{X}$ é calculado ao somar a diferença média das predições de cada intervalo. No passo 4., os resultados centrados em zero permitem interpretar o valor de um ponto em uma curva como a diferença para o valor predito médio.

3.4 Interação das Covariáveis

O método de interação das covariáveis avalia o peso da interação de dada covariável com as demais, conhecida como interação bidirecional, e da interação entre todos os pares possíveis de covariáveis, refletindo a interação total. Uma maneira de estimar a força da interação é medir o quanto da variação da predição depende da interação das covariáveis. Uma medida possível é a estatística H, introduzida por Friedman et al. (2008). A variância explicada de uma interação é determinada pela diferença entre a função de dependência parcial observada e a função de dependência parcial sem interação. Em particular, se não há interação entre duas covariáveis, digamos j e k, a função de dependência parcial definida na Equação (1) pode ser decomposta em $DP_{jk}(x_j, x_k) = DP_j(x_j) + DP_k(x_k)$, em que $DP_j(x_j)$ e $DP_k(x_k)$ são as respectivas funções de dependência parcial das covariáveis j e k. Agora, se a covariável j não interagir com nenhuma outra, a função de predição $\hat{f}$ é definida como $\hat{f}(x) = DP_j(x_j) + DP_{-j}(x_{-j})$, com $DP_j(x_j)$ sendo a função de dependência parcial que depende apenas de j e $DP_{-j}(x_{-j})$ sendo a função de dependência parcial de todas as covariáveis, exceto j. Portanto, a estatística H usada para medir a força de uma interação bidirecional é $H_j^2 = \frac{\sum_{i=1}^{n}\left[DP_{jk}\left(x_j^{(i)}, x_k^{(i)}\right) - DP_j\left(x_j^{(i)}\right) - DP_k\left(x_k^{(i)}\right)\right]^2}{\sum_{i=1}^{n} DP_{jk}^2\left(x_j^{(i)}, x_k^{(i)}\right)}$. Ao passo que quando a interação é total, a estatística é $H_j^2 = \frac{\sum_{i=1}^{n}\left[\hat{f}\left(x^{(i)}\right) - DP_j\left(x_j^{(i)}\right) - DP_{-j}\left(x_{-j}^{(i)}\right)\right]^2}{\sum_{i=1}^{n} \hat{f}^2\left(x^{(i)}\right)}$. Se H resultar em zero, significa que não há qualquer

interação entre covariáveis. Por outro lado, se o resultado for 1, quer dizer que toda a variância da predição pode ser explicada pela soma das funções de dependência parcial. Ao usar todos as n observações, a estatística de interação H requer muita carga computacional, e uma das formas de contornar essa problemática é selecionar uma parte de n. No entanto, isso aumenta a variabilidade das estimativas de dependência parcial, tornando os resultados da estatística inconsistentes. Além disso, H é uma medida que funciona bem sobre variáveis independentes. Do contrário, as interpretações vão decorrer de interações inverossímeis.

3.5 Importância da Covariável por Permutação

A ideia geral da teoria que há por detrás da importância da covariável por permutação, introduzida por Breiman (2001), diz que a importância de uma covariável em um modelo é quantificada pelo aumento do erro em sua predição após permutar essa variável. Se permutando os valores da covariável o erro da predição aumenta, então essa variável é importante para o modelo. Caso contrário, isto é, se o erro da predição não se altera, a variável é considerada insignificante. Assim como a teoria, o algoritmo que mede a importância da covariável por permutação é bastante simples. Dados como entrada o modelo treinado f, a matriz de preditores $\mathbf{X}$, o vetor da variável resposta y e uma função de perda $L(y, f(x))$, o algoritmo proposto por Fisher et al. (2018) é dado por:

1. estime o erro do modelo f, e_{mod}, por meio da função de perda $L(y, f(\mathbf{X}))$ (como exemplos, o erro quadrático médio em problemas de regressão e a taxa de erro de classificação em problemas de classificação);
2. para cada j-ésima variável, $j = 1, \ldots, p$:
   2.1 gere a matriz de variáveis permutadas $\mathbf{X_p}$ ao permutar a variável j nos dados da matriz original $\mathbf{X}$;
   2.2 estime o erro do modelo com base na matriz de variáveis permutadas $\mathbf{X_p}$, i.e., $e_p = L(y, f(\mathbf{X_p}))$;
   2.3 calcule a importância da covariável j por permutação: $I_j = e_p / e_{mod}$ ou $I_j = e_p - e_{mod}$; e
3. ordene as importâncias I de forma decrescente.

Um grande ganho ao avaliar a importância de uma covariável por permutação é não precisar reciclar o modelo, um processo custoso e demorado computacionalmente. Ao permutar as covariáveis, ganha-se, no mínimo, tempo para identificar quais delas são realmente relevantes. Apesar disso, a importância da covariável por permutação é calculada com base em uma estimativa de erro do modelo, o que não é interessante se o objetivo maior é saber o quanto de variância pode ser explicada por cada covariável (ao contrário de querer examinar se o desempenho do modelo diminui permutando os valores de uma covariável, por exemplo). Em Molnar (2019), levanta-se a questão sobre em qual amostra deve-se calcular a importância da covariável, se na amostra de treino ou na de teste. E ao que parece, por enquanto, fica a critério de cada usuário. Se o objetivo é saber o quanto o modelo é dependente de cada variável, opta-se pelos dados de treinamento. Mas se o desejo está em medir a contribuição de cada variável sobre as decisões acertadas do modelo em dados novos, é sugerido escolher os dados de teste.

3.6 Modelo Interpretável Substituto Global

Um modelo substituto global é um método interpretável treinado para aproximar as predições de um modelo caixa-preta. Dessa forma, é possível tirar conclusões sobre um modelo caixa-preta interpretando globalmente o modelo substituto, que usa a predição do modelo não-interpretável como sua variável resposta. De outro modo, ao usar um modelo substituto global, o objetivo está em aproximar a função de predição do modelo caixa-preta f à função de predição do modelo substituto g, com a diferença de que g deve ser interpretável. A escolha do modelo substituto independe do modelo caixa-preta que está sendo usado, uma vez que não é necessário conhecer

o funcionamento interno do modelo complexo. O algoritmo para obter um modelo interpretável substituto global consiste nos seguintes passos:

1. selecione o conjunto ou um subconjunto de dados $\mathbf{X}$ que foi usado para treinar o modelo caixa-preta f;
2. para este conjunto ou subconjunto de dados, obtenha as predições de f;
3. ajuste um modelo interpretável g com os dados selecionados em 1. e as predições obtidas em 2.;
4. avalie o quão bem o modelo simples g replica as predições do modelo complexo f; e
5. interprete g e tire as conclusões acerca de f.

Uma medida recomendada e bastante usada no passo 4. é a R^2 (Montgomery et al., 2021), que pode ser interpretada como o percentual de variância que é explicada pelo modelo substituto global. Se $R^2 \approx 1$, significa que o modelo substituto g se comporta de maneira similar ao modelo caixa-preta f e possibilita generalizar as interpretações de g para f. Do contrário, se $R^2 \approx 0$, é inviável interpretar o modelo caixa-preta a partir do modelo substituto simples. A principal vantagem desse método é poder escolher qualquer modelo que seja interpretável, sendo possível optar por aquele em que as interpretações são mais familiares para o usuário. Por outro lado, como o modelo substituto não tem acesso aos resultados reais, as interpretações feitas não são baseadas nos dados, mas se referem exclusivamente ao modelo complexo.

3.7 Modelo Interpretável Substituto Local

Ribeiro et al. (2016) propõem uma implementação do modelo interpretável substituto local (*Local Interpretable Model-Agnostic Explanations* - LIME, em inglês) para aproximar as predições do modelo caixa-preta. A ideia é a mesma do método do substituto global visto na seção anterior, mas ao invés de treinar um modelo cuja interpretação é verdadeira globalmente, o LIME se concentra em explicar as predições individualmente. De forma geral, o LIME funciona como um modelo interpretável simples que se aproxima bem do modelo caixa-preta nas proximidades de uma observação x de interesse, gerando, portanto, uma interpretação que seja verdadeira apenas em torno da observação que se quer explicar. Sendo assim, as predições feitas pelo modelo caixa-preta podem facilmente ser interpretadas de forma individual pelo modelo interpretável escolhido. O algoritmo associado a essa técnica tem como sequência de passos:

1. para cada predição do modelo caixa-preta f a ser explicada, permute as observações n vezes;
2. obtenha a predição de cada observação permutada por meio do modelo f;
3. pondere as observações permutadas segundo sua proximidade com a observação original ao calcular uma medida de distância e dissimilaridade;
4. selecione as m variáveis consideradas mais importantes para as predições feitas por f e use-as para explicar os dados permutados;
5. ajuste um modelo interpretável g ao conjunto de dados permutados, em que as m variáveis selecionadas em 4. sejam as covariáveis e as predições do modelo caixa-preta representem a variável resposta; e
6. interprete localmente um ponto x do modelo caixa-preta com base nas estimativas de g geradas pela minimização de uma função de perda L somada a uma medida de complexidade Ω:

$$\xi(x) = arg\min_{g \in G} L(f, g, \pi_x) + \Omega(g), \tag{2}$$

em que π_x define o tamanho da vizinhança em torno de x e G é o conjunto de modelos interpretáveis possíveis. É possível ver na Equação (2) que as explicações produzidas pelo estimador LIME vão depender dos valores escolhidos para seus parâmetros, em especial o tamanho da vizinhança. Entretanto, precisar pré-definir esses

valores é considerado um dos principais pontos negativos desse método, uma vez que é necessário testar diferentes configurações até que as interpretações façam sentido para um determinado contexto. Outra grande fraqueza do LIME é apontada por Alvarez-Melis & Jaakkola (2018), que mostram em cenários simulados que as explicações são instáveis mesmo para dois pontos de dado muito próximos. Já Amorim (2019) discute a dificuldade em se encontrar um modelo interpretável que explique bem localmente as predições de modelos muito complexos, principalmente de regressão e que tenham muitas covariáveis. Mas ainda assim, o LIME ainda é um dos poucos métodos de interpretabilidade aplicável a dados de natureza tabular, textual e de imagem.

3.8 Valores Shapley

Desenvolvido por Shapley (2016), o valor Shapley é um conceito da Teoria dos Jogos que descreve como distribuir de forma justa uma premiação aos jogadores de uma coalizão de acordo com a contribuição individual de cada um no resultado final de um jogo. Essa contribuição, o valor Shapley, é medida pela média da diferença entre todas as contribuições marginais de todas as coalizões possíveis que contém e não contém determinado jogador. Em *Machine Learning*, podemos traduzir o "jogo" como o modelo preditivo, os "jogadores" como as covariáveis e o "prêmio" como a predição. Por sua vez, o valor Shapley é definido pela média ponderada (por uma constante normalizadora) da soma de todas as diferenças possíveis entre os modelos treinados com $\left(f_{S\cup\{j\}}\right)$ e sem (f_S) a j-ésima covariável sob investigação. Formalmente, $\phi_j = \sum_{S\subseteq N\setminus\{j\}} \frac{|S|!(|N|-|S|-1)}{N!}\left[f_{S\cup\{j\}}\left(x_{S\cup\{j\}}\right) - f_S(x_S)\right]$, em que N é o conjunto de todas as covariáveis e x_S é o vetor com todos os valores de entrada das covariáveis do subconjunto S. Uma vez que calcular as contribuições exige todos os subconjuntos possíveis de covariáveis, o aumento exponencial do tempo impede que o valor exato dessas contribuições seja calculado. Para contornar esse problema, Štrumbelj & Kononenko (2014) propuseram um procedimento de amostragem via método de Monte-Carlo para estimar os valores Shapley ao perturbar os valores de entrada de uma observação x para a qual há o interesse em explicar. Dados o modelo não-interpretável f e a matriz de covariáveis $\mathbf{X}$, o algoritmo consiste em:

1. selecione uma observação x de interesse e uma covariável j, $j = 1,\ldots,p$;
2. para cada $m = 1,\ldots,M$ iteração:
   - 2.1 selecione uma permutação aleatória do conjunto de todas as permutações possíveis entre as covariáveis, i.e., $O \in S(\{1,\ldots,p\})$;
   - 2.2 selecione uma observação aleatória z que pertença à matriz de dados $\mathbf{X}$ e gere uma nova amostra, i.e., $z_0 = \left(z_{(1)},\ldots,z_{(j)},\ldots,z_{(p)}\right)$;
   - 2.3 crie uma nova observação com a covariável j ao substituir os valores de x pelos valores de z após a operação j: $x_{+j} = \left(x_{(1)},\ldots,x_{(j-1)},x_{(j)},z_{(j+1)},\ldots,z_{(p)}\right)$;
   - 2.4 crie uma nova observação sem a covariável j repetindo o passo 2.3, mas fazendo a substituição a partir da operação j: $x_{-j} = \left(x_{(1)},\cdots,x_{(j-1)},z_{(j)},z_{(j+1)},\cdots,z_{(p)}\right)$;
   - 2.5 calcule a contribuição marginal de j ao fazer a diferença entre as predições do modelo complexo com e sem a covariável j: $\phi_j^m = \hat{f}(x_{+j}) - \hat{f}(x_{-}j)$;
3. calcule a média das M diferenças e encontre o valor Shapley para x: $\phi(x) = \frac{1}{M}\sum_{m=1}^{M}\phi_j^m$.

O valor Shapley encontrado no passo 3. é interpretado como o quanto a inclusão da j-ésima covariável contribuiu, positiva ou negativamente, para a predição de uma observação x em particular em relação à predição média feita para todos os dados. Além de distribuir justamente as contribuições, a teoria sólida por detrás dos acontecimentos dos valores Shapley faz com ele seja considerado um dos métodos de interpretabilidade mais completos em termos de explicação. As desvantagens geralmente esbarram no tempo computacional, que é muito grande devido a todas

as permutações possíveis de 2^p subconjuntos de covariáveis. Uma implementação rápida e eficiente dos valores Shapley é o método Explicações Aditivas Shapley, popularmente conhecido como SHAP (*SHapley Additive exPlanations*, em inglês). Introduzido por Lundberg & Lee (2017), SHAP é uma abordagem unificada de vários métodos para interpretar a previsão de qualquer observação x como a soma das contribuições individuais de cada covariável de um modelo não-interpretável. A diferença para os valores Shapley é que são usados métodos auxiliares que possibilitam flexibizar seu algoritmo para cada tipo de modelo (linear, baseado em árvores, de aprendizagem profunda etc). Para a leitura da teoria do SHAP em detalhes é recomendado o livro de Molnar (2019).

4 Metodologia

Apesar da popularidade dos modelos de *Machine Learning*, muitos deles permanecem sendo não-interpretáveis (modelos caixa-preta). Em outras palavras, são sistemas fechados e complexos que inviabilizam o entendimento da função que é modelada pelos seus algoritmos. Na prática, isso significa que não conseguimos entender a relação entre as covariáveis e a variável resposta de forma direta, como acontece em um modelo de regressão logística, por exemplo. Na Figura 2.7, p. 25 de James et al. (2013), está muito claro que um modelo muito flexível (isto é, mais complexo) tem a interpretabilidade comprometida.

No entanto, compreender os motivos por que ocorreram certas previsões é adequado para se fazer inferência (Breiman, 2001) e muito importante para avaliar a justiça (*fairness*) e a confiança de um modelo (Doshi-Velez & Kim, 2017). Por esse motivo, serão considerados os métodos de interpretabilidade descritos na Seção 3 e empregá-los via pacote `iml`[1] do `R` (www.r-project.org), versão `4.1.0`, linguagem computacional usada neste projeto, sob a IDE `RStudio`. Já os modelos preditivos[2] considerados neste trabalho são apresentados na coluna "modelo" da Tabela 1 associadamente à função (coluna "função") e ao pacote (coluna "pacote") com os quais foram ajustados. Nesta tabela, também pode ser vista a performance correspondente a cada um dos modelos sob a perspectiva de algumas métricas de desempenho: acurácia, sensibilidade, especificidade, valor preditivo positivo (vpp) e valor preditivo negativo (vpn). O modelo escolhido é aquele com maior acurácia e mais detalhes sobre essas métricas podem ser vistos em Morettin & Singer (2020).

5 Resultados e Discussão

Analisamos a base de dados com 10 variáveis e 404 observações, em que todas são gestantes diagnosticadas com diabetes gestacional que realizaram o pré-natal entre os anos de 2012 a 2015 no Hospital das Clínicas da Universidade de São Paulo. Para o ajuste dos modelo preditivos considerados neste trabalho, usamos como covariáveis idade (média: 32,7; desvio-padrão: 6,1), número de gestações (média: 2,8; desvio-padrão: 1,7), IMC categórico (até normal: 21,3%; sobrepeso: 31,2%; obeso: 47,5%), histórico de diabetes na família (sim: 63,36%), antecedente de macrossomia (sim: 8,67%), antecedente de diabetes gestacional (sim: 12,13%), indicador de tabagista (sim: 8,42%), indicador de hipertensão (sim: 27,73%) e valor do exame de glicemia de jejum (média: 98,3; desvio-padrão: 6,5), e como variável resposta se a gestante fez o uso de insulina em algum momento precedente ao instante do parto (sim: 33%). A análise exploratória completa dos dados pode ser vista em: http://rpubs.com/ornscar/desc_diabetes.

[1] Com a exceção do método SHAP, para o qual foi usado o pacote `fastshap`.

[2] Devido à limitação de páginas deste relatório, para detalhes dos modelos preditivos aqui usados indicamos as referências Amorim (2019), Izbicki & Santos (2020), James et al. (2013) e Morettin & Singer (2020).

Como resultado dos ajustes dos modelos, as medidas de desempenho conjuntamente à função e ao pacote associados a cada um dos modelos preditivos candidatos estão na Tabela 1 que segue.

modelo	acurácia	sens.	espec.	vpp	vpn	função	pacote
Análise Discriminante Linear	0,712	0,277	0,953	0,769	0,704	`lda`	`MASS`
Análise Discriminante Quadrática	0,722	0,388	0,907	0,700	0,728	`qda`	`MASS`
Árvores de Classificação	0,730	0,393	0,895	0,650	0,750	`rpart`	`rpart`
Bagging	0,700	0,393	0,850	0,565	0,740	`train`	`caret`
Florestas Aleatórias	0,710	0,272	0,925	0,642	0,720	`train`	`caret`
K-vizinhos mais Próximos	0,710	0,181	0,970	0,750	0,706	`knn`	`class`
Regressão Logística	0,700	0,545	0,776	0,545	0,776	`glm`	`R Base`
Support Vector Machines Gaussiano	0,683	0,388	0,846	0,583	0,714	`svm`	`e1071`
Support Vector Machines Linear	0,653	0,138	0,938	0,555	0,663	`svm`	`e1071`
Support Vector Machines Polinomial	0,732	0,361	0,938	0,764	0,726	`svm`	`e1071`
XGBoost	0,762	0,757	0,754	0,609	0,866	`boost_tree`	`parsnip`

Tabela 1: Desempenho e funções e pacotes com os quais os modelos foram ajustados.

Tendo em vista a maior acurácia, o modelo escolhido para serem feitas as aplicações das métricas de interpretabilidade foi o *XGBoost*[3], para o qual os hiperparâmetros considerados foram: `trees` $= 1000$, `mtry` $= 7$, `min_n` $= 5$, `tree_depth` $= 9$, `learn_rate` $= 0,00209$, `loss_reduction` $= 0,00000461$ e `sample_size` $= 0,640$. Detalhes sobre os hiperparâmetros podem ser vistos em Izbicki & Santos (2020) e Morettin & Singer (2020).

Com o modelo preditivo previamente ajustado, inicialmente, queríamos saber qual a covariável mais importante para predizer se uma gestante diagnosticada com diabetes gestacional precisará fazer o uso de insulina em algum momento antes do parto. Para isso, usamos o método da importância da covariável por permutação, em que a importância da covariável foi medida pela variação na taxa de erro de classificação (TEC) ao permutá-la. Na Figura 1 fica bastante evidente que a covariável mais influente nas decisões do modelo foi o valor do exame de glicemia de jejum, que obteve o maior erro de classificação em relação às demais. Neste caso, podemos ver que o aumento na TEC foi de aproximadamente 1,34, com variabilidade entre 1,23 e 1,38.

Em virtude disso, selecionamos o valor do exame de glicemia de jejum como a covárivel a ser analisada nos gráficos de dependência parcial (PDP), da esperança condicional individual (ICE) e de efeitos locais acumulados (ALE) expostos na Figura 2. Ao analisar o gráfico de PDP (2a), podemos ver que quanto maior o valor do exame de glicemia de jejum, maior a probabilidade da necessidade do uso de insulina. Essa crescente torna-se ainda mais rápida para valores de exame superiores a 100. Ao observar o gráfico 2(b), no qual cada gestante é representada por uma curva em preto, constatamos que essa premissa parece ser verdadeira para todas as gestantes. O gráfico da esperança condicional individual mostra que o efeito do valor de glicemia de jejum parece seguir o padrão médio (em destaque pela curva em amarelo): o aumento da probabilidade é considerável em valores de glicemia de jejum maiores do que 100 (mesmo em gestantes que obtiveram valores abaixo de 100 e probabilidade predita acima de 50%). No entanto, curiosamente, valores acima de 105 não acarretaram em probabilidades maiores em ambos os gráficos. Provavelmente, isso se deve à pouca quantidade de dados com valores grandes de glicemia de jejum, dificultando o aprendizado do modelo em termos de predição e fazendo com que as estimativas de dependência parcial sejam pouco confiáveis nesse intervalo de valores.

Além disso, uma possível relação de causalidade entre as covariáveis também pode ter ocasionado interpretações

[3] O *XGBoost* é uma aplicação do método de *Gradient Boosting*, cuja função é minimizar uma função de perda específica $L(y, f(x))$ por meio do algoritmo de gradiente descendente (Friedman, 2001).

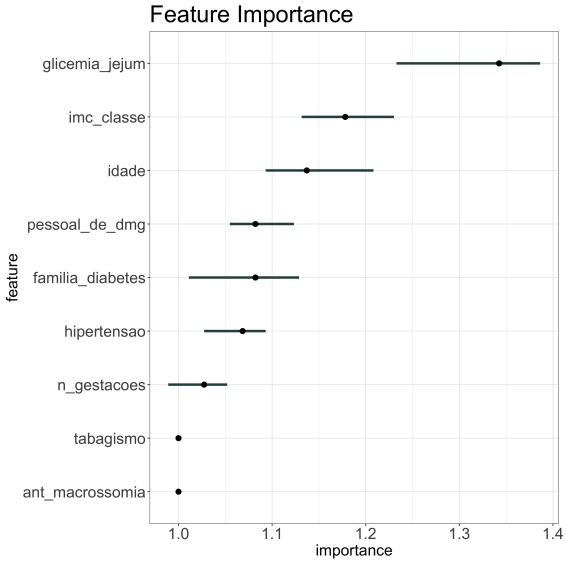

Figura 1: Gráfico da importância da covariável por permutação.

viesadas. Nessa situação, o gráfico de efeitos locais acumulados parece ser adequado, visto que ele desconsidera áreas com poucas informações e bloqueia o efeito de covariáveis correlecionadas. Apesar disso, podemos perceber no gráfico 2(c) que o comportamento da curva ALE é bastante semelhante às anteriores. Assim, mais uma vez é apresentado que uma gestante tem probabilidade aumentada em fazer o uso de insulina à medida que o resultado do exame de glicemia de jejum vai superando o valor de 100. Por esse motivo, podemos dizer que o gráfico de efeitos locais acumulados já seria o suficiente para resumir satisfatoriamente bem as relações entre a covariável glicemia de jejum e a probabilidade de uma gestante fazer o uso de insulina.

Em um gráfico de dependência parcial também é possível avaliar o efeito de até duas covariáveis de interesse. Para exemplificação, consideramos na Figura 3 as covariáveis glicemia de jejum e idade. É facil notar que gestantes com idade entre 16 e 47 anos, idades mínima e máxima do banco de dados, e que obtiveram uma pontuação de até pouco mais de 100 no exame de glicemia de jejum têm probabilidade zero, próxima a zero ou muito baixa de precisar usar insulina antes do parto - ainda que sob a perspectiva de gestantes com idade superior a 38 anos, essa probabilidade alcance cerca de 20%. Em gestantes com pouco mais de 30 anos e exames com valores entre 103 e 107, o risco de se fazer o uso de insulina é de aproximadamente 25% a 50%. O caso mais crítico parece estar na faixa de idade superior a 30 anos e de glicemia de jejum acima de 107, quando os percentuais são maiores do que 70%.

Outra maneira de fazer interpretações gerais é usar um modelo interpretável substituto global de modo que seu comportamento seja similar ao modelo caixa-preta *XGBoost* que foi usado para fazer as previsões deste trabalho. Utilizamos uma árvore de decisão como o modelo substituto para o qual a R^2 resultou em aproximadamente 70%. O resultado dessa aplicação é mostrado na Figura 4. Podemos identificar que a maior chance de uma gestante fazer o uso de insulina em algum momento antes do parto é quando ela obtém um resultado de exame de glicemia de jejum acima de 101, é classificada com sobrepeso ou como obesa e tem mais de 30 anos. Por outro lado, se o valor do exame é menor do que 101 e não há qualquer histórico de diabetes gestacional, o cenário se torna bastante favorável para a não necessidade do uso de insulina.

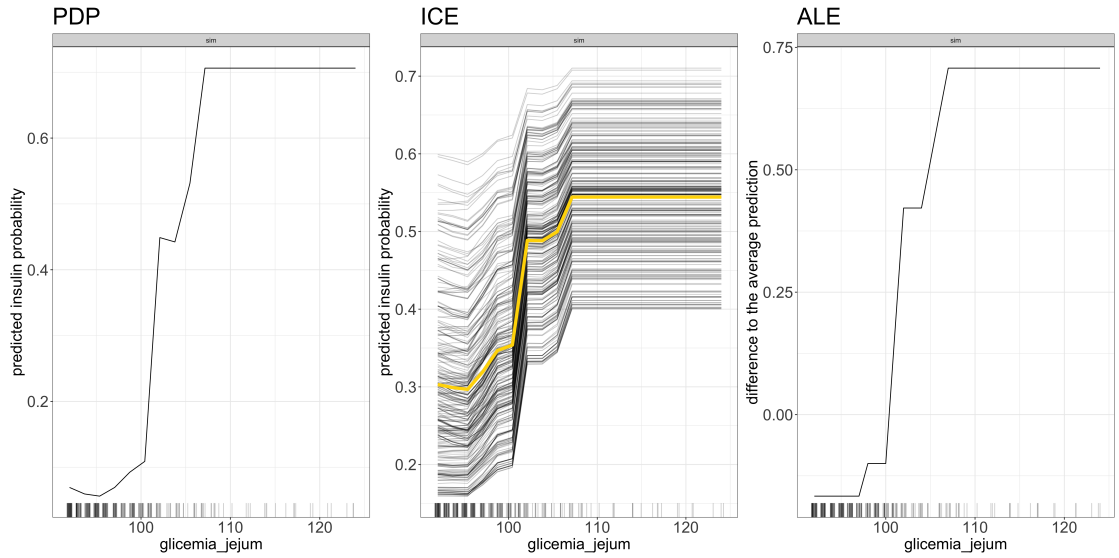

Figura 2: Gráficos (a) de dependência parcial, (b) da esperança condicional individual e (c) de efeitos locais acumulados para o modelo preditivo fazer o uso de insulina segundo o valor do exame de glicemia de jejum.

Para além disso, utilizamos o método da interação das covariáveis para explorar qual foi o impacto das interações totais e bidirecionais entre covariáveis sobre os resultados preditos do modelo. O resultado pode ser visto na Figura 5. Na Figura 5(a) é possível observar que, avaliadas individualmente, a interação do valor do exame de glicemia de jejum, da idade e do IMC categorizado com as demais covariáveis é capaz de explicar mais de 40% da variabilidade das predições, sugerindo que o efeito da interação entre cada uma delas com as outras são relativamente fortes. Em destaque nesse gráfico está o valor do exame de glicemia de jejum, com o maior efeito, e indicador de tabagista e antecedência de macrossomia, que não interagem com nenhuma outra covariável. Além do mais, uma vez que o valor do exame de glicemia de jejum foi a covariável que obteve a maior força de interação, na Figura 5(b) é mostrada a interação bidirecional dela com as restantes. E então, podemos perceber que o resultado do exame de glicemia de jejum tem interações mais robustas com as covariáveis idade, IMC categorizado e antecedência de diabetes gestacional, nesta ordem.

Até aqui, os métodos de interpretabilidade foram apropriados para fazerem interpretações globais - ainda que o gráfico ICE também permita interpretação sob a ótica individual. Porém, como esses métodos geram interpretações baseadas em resultados médios das predições, nem sempre uma mesma explicação é válida para toda observação. Por isso, utilizamos um modelo de regressão logística ponderado como o modelo interpretável substituto local (LIME) e os valores Shapley para explicar o resultado previsto para uma gestante em particular (Figura 6). A partir do gráfico LIME, podemos dizer que o valor do exame de glicemia de jejum ser igual a 103 teve um efeito positivo na predição “sim”, assim como a idade ser igual a 43 anos e o IMC categorizado ser igual à obesa. Em contrapartida, não ter histórico de diabetes gestacional não parece favorecer o cenário em que a insulina deve ser receitada, visto que o efeito dessa covariável é negativo. Entretanto, devido à soma dos efeitos positivos serem maiores do que a soma dos efeitos negativos, é bastante provável que uma gestante com tais características precisará usar insulina antes do parto. As mesmas explicações conseguimos extrair do gráfico de valores Shapley, com a diferença que podemos avaliar a contribuição, se positiva ou negativa, de todas as covariáveis. Embora desejássemos explicar a previsão para uma gestante em específico, se lembrarmos dos resultados obtidos pelos

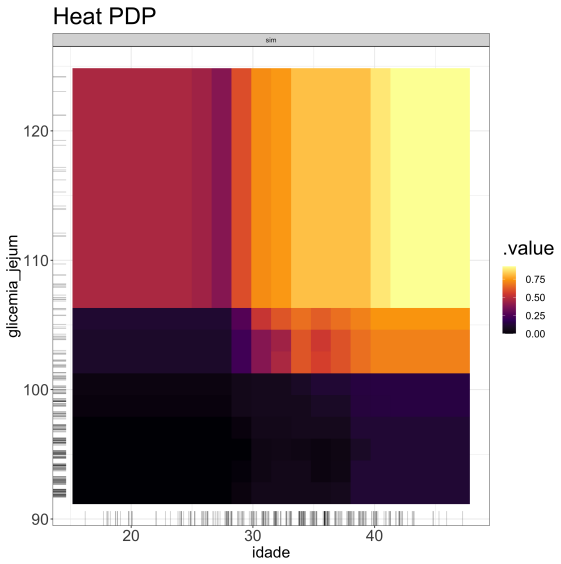

Figura 3: Gráfico de dependência parcial de calor para o modelo preditivo do uso de insulina segundo a dade e o valor do exame de glicemia de jejum.

métodos globais vistos anteriormente, essa parece ser uma análise bastante coesa com relação à interpretação média.

Uma alternativa deveras atrativa para fazer interpretações individuais é o gráfico de força SHAP. Em um gráfico de força SHAP é apresentado como as covariáveis contribuem para mover o valor da predição de uma observação para longe do valor médio de todas as predições, o qual é conhecido como valor base - em nosso problema, o valor base é a probabilidade média da classe "sim" da variável resposta insulina. Além disso, as covariáveis que têm mais impacto na predição $f(x)$ estão localizadas mais próximas da fronteira de divisão entre as covariáveis que impactaram positivamente (barras vermelhas) e negativamente (barras azuis), sendo o tamanho desse impacto quantificado pelo comprimento das barras. Vejamos a Figura 7 em que as predições a serem explicadas são de duas gestantes com diferentes características. A Figura 7(a) mostra que a combinação do valor de exame de glicemia de jejum ser relativamente baixo, a idade ser de 30 anos e ter um histórico de apenas duas gestações parece indicar uma probabilidade muito maior para o desfecho em que a prescrição de insulina não será necessária. Por outro lado, ao analisar o gráfico da Figura 7(b), vemos que as covariáveis glicemia de jejum, idade e IMC categorizado (1 = "sim") com seus respectivos valores atuam para aumentar a probabilidade da indispensabilidade em usar insulina em momentos precedentes ao parto. Mesma explicação feita pelos gráficos LIME e de valores Shapley, porém de um jeito muito mais interessante visualmente.

Um aplicativo Shiny para avaliar as medidas de interpretabilidade para todas as variáveis consideradas está disponível em https://ornscar.shinyapps.io/interpretabilidade_obstetricia.

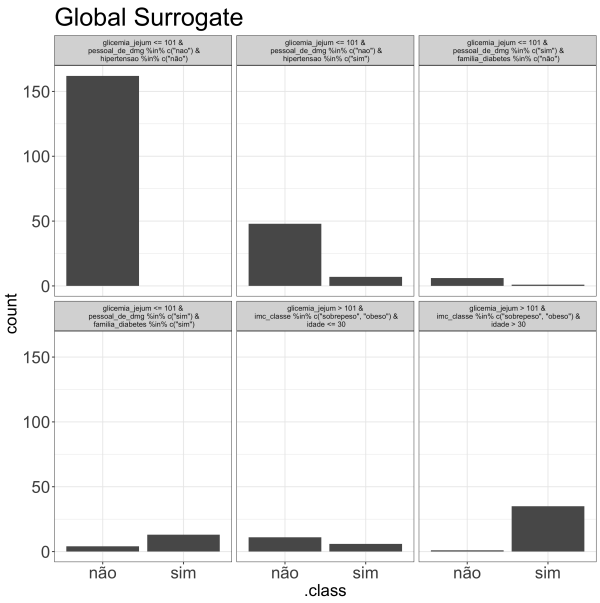

Figura 4: Gráfico de explicações do modelo interpretável substituto global.

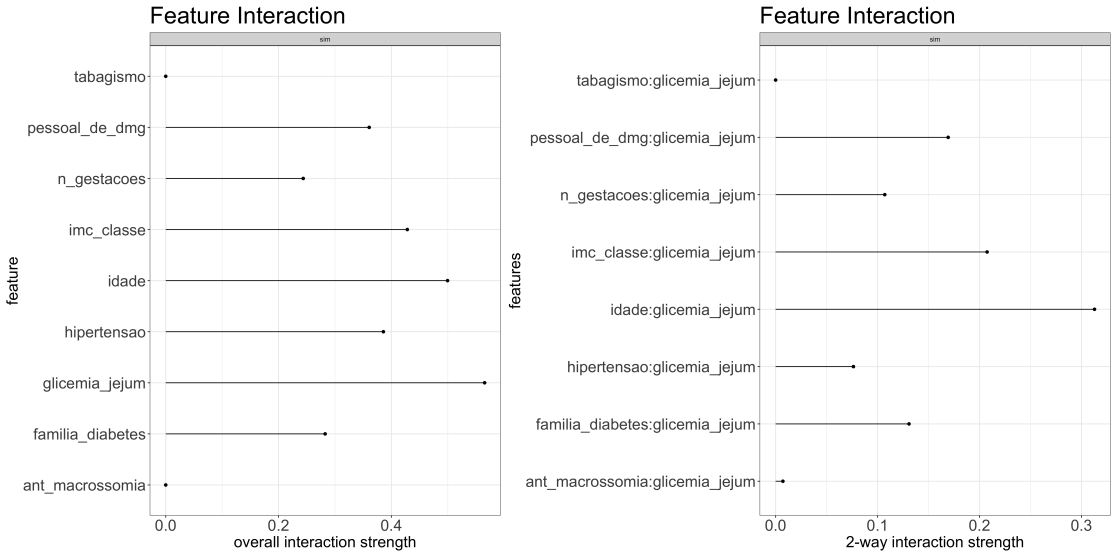

Figura 5: Gráficos (a) de interação total e (b) de interação bidirecional das covariáveis.

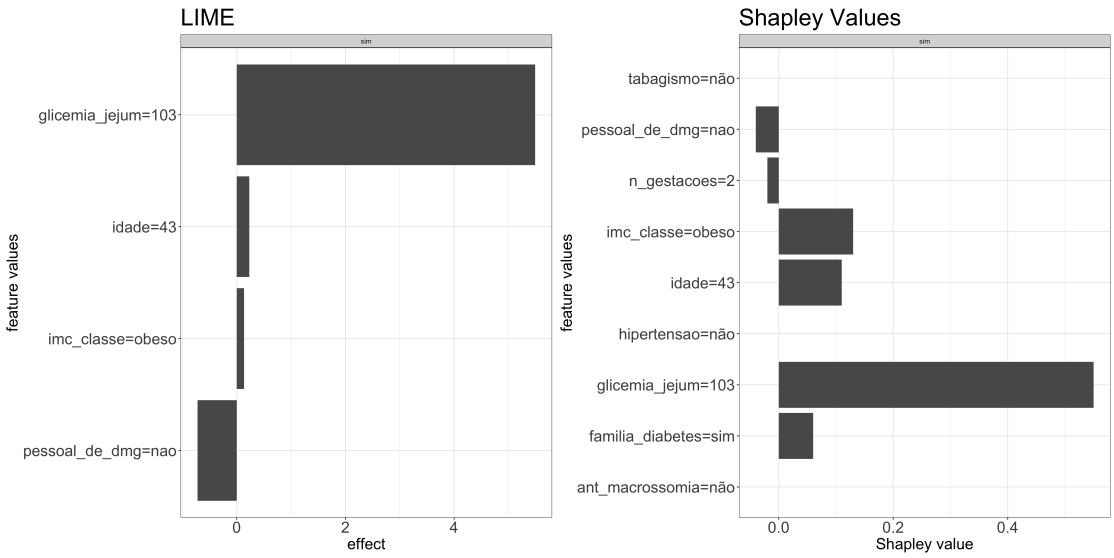

Figura 6: Gráficos de explicações (a) do modelo interpretável substituto local (LIME) e (b) dos valores Shapley para a gestante em particular. No gráfico LIME, as quatro covariáveis que aparecem foram as consideradas mais importantes para as predições feitas pelo modelo complexo *XGBoost* e o efeito do eixo-x é resultado do produto entre um peso β e os valores originais das covariáveis. Já no gráfico dos valores Shapley, o eixo-x representa o valor da contribuição de cada covariável.

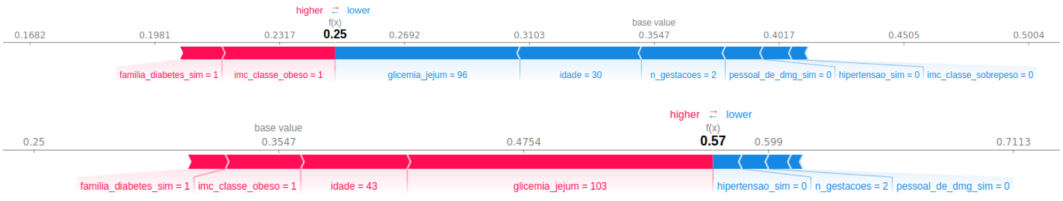

Figura 7: Gráficos de força SHAP (a) da primeira gestante e (b) da segunda gestante.

6 Conclusões

Neste projeto, usamos métodos de interpretabilidade para interpretar de maneira global e individual as decisões tomadas pelo modelo preditivo caixa-preta *XGBoost*, modelo que apresentou melhor desempenho preditivo dentre os modelos ajustados. Do ponto de vista estatístico e clínico, mostramos como esses métodos podem ser bastante úteis para entender facilmente a relação entre um dado fator e desfecho, sem abrir mão de um modelo com alto poder de predição. Poder dizer que um modelo é confiável e justo a ponto de usá-lo na tomada de decisão, em vez de simplesmente fazê-la com base em predições às cegas, é primordial, se não vital, para a área da saúde.

Esse projeto também disponibilizou um aplicativo web para acesso das medidas de interpretabilidade para todas as variáveis da aplicação (https://ornscar.shinyapps.io/interpretabilidade_obstetricia) e apresentado para o grupo de pesquisa do Departamento de Obstetrícia da USP. Além disso, todo código computacional utilizado nesse estudo pode ser acessado em https://github.com/ornscar/ic_saude_materna.

Referências Bibliográficas

Alvarez-Melis, D. & Jaakkola, T. S. (2018). On the robustness of interpretability methods. *arXiv preprint arXiv:1806.08049*.

Amorim, W. N. d. (2019). *Ciência de dados, poluição do ar e saúde*. PhD thesis, Universidade de São Paulo.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

Doshi-Velez, F. & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

Fisher, A., Rudin, C., & Dominici, F. (2018). Model class reliance: Variable importance measures for any machine learning model class, from the" rashomon" perspective. *arXiv preprint arXiv:1801.01489*, 68.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*.

Friedman, J. H., Popescu, B. E., et al. (2008). Predictive learning via rule ensembles. *Annals of Applied Statistics*, 2(3):916–954.

Izbicki, R. & Santos, T. M. d. (2020). *Aprendizado de máquina: uma abordagem estatística*. Câmara Brasileira do Livro, SP, Brasil, 2ª edition.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning with Applications in R*. Springer.

Lundberg, S. M. & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Proceedings of the 31st international conference on neural information processing systems*, pages 4768–4777.

Molnar, C. (2019). *Interpretable Machine Learning - A Guide for Making Black Box Models Explainable*.

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2021). *Introduction to linear regression analysis*. John Wiley & Sons.

Morettin, P. A. & Singer, J. M. (2020). *Introdução à Ciência de Dados - Fundamentos e Aplicações*.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "why should i trust you?"explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144.

Shapley, L. S. (2016). *17. A value for n-person games*. Princeton University Press.

Štrumbelj, E. & Kononenko, I. (2014). Explaining prediction models and individual predictions with feature contributions. *Knowledge and information systems*, 41(3):647–665.