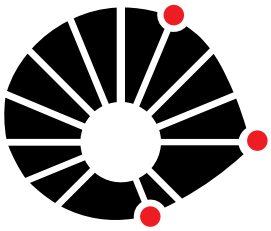

UNICAMP

UNIVERSIDADE ESTADUAL DE CAMPINAS

Instituto de Matemática, Estatística e Computação Científica

LIVIA OHANA DA ROCHA CARVALHO ROSA

Um estudo de recomendação de UTI para gestantes de alto risco usando técnicas de aprendizagem de máquina

Campinas

2020

Livia Ohana da Rocha Carvalho Rosa

Um estudo de recomendação de UTI para gestantes de alto risco usando técnicas de aprendizagem de máquina

Dissertação apresentada ao Instituto de Matemática, Estatística e Computação Científica da Universidade Estadual de Campinas como parte dos requisitos exigidos para a obtenção do título de Mestra em Matemática Aplicada e Computacional.

Orientador: Estevão Esmi Laureano

Este exemplar corresponde à versão final da Dissertação defendida pela aluna Livia Ohana da Rocha Carvalho Rosa e orientada pelo Prof. Dr. Estevão Esmi Laureano.

Campinas

2020

Ficha catalográfica
Universidade Estadual de Campinas
Biblioteca do Instituto de Matemática, Estatística e Computação Científica
Ana Regina Machado - CRB 8/5467

R71e

Rosa, Livia Ohana da Rocha Carvalho, 1988-
Um estudo de recomendação de UTI para gestantes de alto risco utilizando técnicas de aprendizagem de máquina / Livia Ohana da Rocha Carvalho Rosa. – Campinas, SP : [s.n.], 2020.

Orientador: Estevão Esmi Laureano.
Dissertação (mestrado profissional) – Universidade Estadual de Campinas, Instituto de Matemática, Estatística e Computação Científica.

1. Aprendizado de máquina. 2. Gestantes. 3. Unidades de terapia intensiva. 4. Gravidez de alto risco. 5. Algoritmos de computador. I. Esmi, Estevão, 1982-. II. Universidade Estadual de Campinas. Instituto de Matemática, Estatística e Computação Científica. III. Título.

Informações para Biblioteca Digital

Título em outro idioma: An study of ICU recommendation to high-risk pregnant women using machine learning techniques
Palavras-chave em inglês:
Machine learning
Pregnant women
Intensive care units
Pregnancy, High-risk
Computer algorithms
Área de concentração: Matemática Aplicada e Computacional
Titulação: Mestra em Matemática Aplicada e Computacional
Banca examinadora:
Estevão Esmi Laureano [Orientador]
João Batista Florindo
Samira El Maerrawi Tebecherane Haddad
Data de defesa: 16-12-2020
Programa de Pós-Graduação: Matemática Aplicada e Computacional

Identificação e informações acadêmicas do(a) aluno(a)
- ORCID do autor: https://orcid.org/0000-0002-9275-3012
- Currículo Lattes do autor: http://lattes.cnpq.br/1861131561464120

Dissertação de Mestrado Profissional defendida em 16 de dezembro de 2020 e aprovada pela banca examinadora composta pelos Profs. Drs.

Prof(a). Dr(a). ESTEVÃO ESMI LAUREANO

Prof(a). Dr(a). JOÃO BATISTA FLORINDO

Prof(a). Dr(a). SAMIRA EL MAERRAWI TEBECHERANE HADDAD

A Ata da Defesa, assinada pelos membros da Comissão Examinadora, consta no SIGA/Sistema de Fluxo de Dissertação/Tese e na Secretaria de Pós-Graduação do Instituto de Matemática, Estatística e Computação Científica.

Aos meus pais, Áurea e Robson, que são minha força e suporte.
Ao meu companheiro de vida, Lucas, que me ajudou em cada passo desta jornada e, incansavelmente, disse que tudo daria certo.
À minha sogra, Teresinha, que me ajudou nos cuidados do Arthur para que eu pudesse concluir este trabalho.
Ao meu filho, Arthur, que é a razão de tudo fazer sentido.

Agradecimentos

Ao o meu orientar Prof. Dr. Estevão Esmi, por me guiar e proporcionar a possibilidade de trabalhar com o que realmente importa.

Ao Prof. Dr. Rodolfo Pacagnella, sem seus ensinamentos e experiência, este trabalho não seria possível.

Ao Ms. Fabiano Soares, seus artigos foram essenciais para a conclusão deste trabalho.

Aos meus colegas da pós-graduação, sem eles a jornada seria muito mais difícil.

Aos pesquisadores e cientistas que contruíram a estrada para que eu pudesse seguir.

E finalmente, à UNICAMP, ao IMECC e ao CAISM que me ofereceram a oportunidade de estudar.

"Não é o crítico que importa, nem aquele que aponta onde foi que o homem tropeçou ou como o autor das façanhas poderia ter feito melhor. O crédito pertence ao homem que está por inteiro na arena da vida, cujo rosto está manchado de poeira, suor e sangue. Que luta bravamente, que erra, que decepciona, porque não há esforço sem erros e decepções. Mas que, na verdade, se empenha em seus feitos, que conhece o entusiasmo, as grandes paixões, que se entrega a uma causa digna e que, na melhor das hipóteses, conhece no final o triunfo da grande conquista e, na pior, se fracassar, ao menos fracassa ousando grandemente. (Theodore Roosevelt)

Resumo

O objetivo desta dissertação é desenvolver um modelo para recomendar internação em UTI de gestantes de alto risco. Para tanto, foram utilizadas técnicas de aprendizagem de máquina e o algoritmo escolhido, devido aos seus melhores resultados, foi o *XGBoost*, cujos parâmetros foram ajustados via *grid search*. Obteve-se resultados de acurácia superiores a 83% e sensibilidade superior a 51%, dentro do grupo onde foi recomendada a internação estão 100% dos óbitos e 66% dos casos de *near miss*. Também foi considerada a inclusão de interpretações das predições do modelo em casos particulares.

Palavras-chave: Aprendizagem de máquina, XGBoost, UTI, gestantes, gestação, *near miss*, óbito materno.

Abstract

The objective of this dissertation is to develop a model to recommend ICU admission to high-risk pregnant women. Machine learning techniques were used and XGBoost was the chosen algorithm, due to its better results, with the parameters adjusted via grid search. Accuracy results obtained were greater than 83% and sensitivity greater than 51%, within the group where hospitalization was recommended it was observed 100% of deaths and 66% of near miss cases. The inclusion of interpretations of the model's predictions in particular cases was also considered.

Keywords: Machine Learning, XGBoost, ICU, pregnant women, pregnancy.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

MSS	*Maternity Severity Score*
SMO	*Severe Maternal Outcome*
CEP	Comitê de Ética em Pesquisa
CONEP	Comissão Nacional de Ética em Pesquisa
CPAV	Condições Potencialmente Ameaçadoras da Vida
UTI	Unidade de Terapia Intensiva
MM	Morte Materna
MMG	Morbidade Materna Grave
NMM	*Near Miss* Materno
OMS	Organização Mundial da Saúde
ONU	Organização das Nações Unidas
REDE	Rede Brasileira de Vigilância de Morbidade Materna Grave
WHO	World Health Organization
ACC	Acurácia
SENS	Sensibilidade
ESPEC	Especificidade
PR	Prevalência Real
PE	Prevalência Estimada
VPP	Valor Preditivo Positivo
VPN	Valor Preditivo Negativo
RVP	Razão de Verissimilhança Positiva
RVN	Razão de Verossimilhança Negativa
XGBOOST	*Extreme Gradient Boosting*
LIME	*Local Interpretable Model-Agnostic Explanations*

Sumário

1 Introdução

1.1 Justificativas e motivações

A morte materna (MM), mais que um evento isolado da perda da vida de uma pessoa, tem um impacto na família e comunidade nas quais esta mulher está inserida. Devido a sua importância, é considerada como um indicador socioeconômico de um país (WHO, 1990). Segundo (SOARES et al., 2016), a internação em UTI nos casos onde a gravidade é considerada, pode reduzir em até 80% a chance de óbito, ou seja, ferramentas que otimizem o uso de leitos de UTI podem ser instrumentos eficientes para gestão de políticas em saúde.

O uso de técnicas de aprendizagem de máquina em saúde é relativamente recente, no entanto, extremamente promissor. Entende-se aprendizagem de máquina como a busca por algoritmos computacionais que melhorem a capacidade de realizar uma tarefa através da experiência (JORDAN; MITCHELL, 2015).

A escolha de estrutura, critério de adaptação e algorítimo de otimização são centrais no aprendizado de máquina (BOCCATO; ATTUX, 2019). A primeira escolha está relacionada à estrutura do modelo que será utilizado para realizar a tarefa, assim, é preciso considerar o mapeamento da entrada-saída gerado pelo modelo e aspectos como linearidade e causalidade, por exemplo. O critério de adaptação expressa matematicamente o que se deseja atingir durante o processo de aprendizado, a partir dessa escolha se pode medir o grau de sucesso do modelo. A definição de critério dá origem a um problema de otimização, pois, o objetivo do sistema será encontrar um conjunto de parâmetros que otimize uma determinada função que representa o objetivo do modelo. Por fim, é necessário escolher o algorítmo de otimização que será responsável por escolher os parâmetros ótimos do sistema.

A aprendizagem de máquina possui paradigmas de aprendizagem, esta segmentação baseia-se no conhecimento se a resposta desejada é conhecida ou não. Desta forma temos os seguintes paradigmas de aprendizado.

- Supervisionado: Para cada dado (ou padrão) de treinamento existe uma resposta desejada que é conhecida. Neste caso, se diz que os dados são rotulados.
- Não-Supervisionado: Para cada dado (ou padrão) de treinamento não existe uma resposta desejada que é conhecida. Neste caso, se diz que os dados são não-rotulados.
- Semi-Supervisionado: Para alguns dados (ou padrões) de treinamento existe uma

resposta desejada que é conhecida e para outros não. São utilizados tanto dados rotulados quanto não-rotulados.

- Reforço: Para cada dado (ou padrão) de treinamento não existe uma resposta desejada que é conhecida, mas se tem acesso a informação da qualidade da saída gerada pelo modelo na forma de um sinal de recompensa ou punição.

No caso deste trabalho, o modelo segue o paradigma supervisionado. Os dados são rotulados, pois, para cada gestante sabemos se houve ou não internação em UTI.

Alguns exemplos de trabalhos na área da saúde utilizando aprendizagem de máquina, inclui (PAN et al., 2017) que avalia o valor preditivo positivo de algoritmos de aprendizado de máquina para uma avaliação comparativa do risco de parto adverso entre gestantes como forma de melhorar a alocação de serviços sociais. E ainda, em (SILVA et al., 2017) é usado o algoritmo de Naive Bayes para identificar relações entre dados de nascimento e de morte de crianças com menos de um ano. Para (FILHO, 2015), existem três áreas auspiciosas para o uso de *big data* (assim como as técnicas de aprendizagem de máquina) em saúde: medicina de precisão (*precision medicine*), prontuários eletrônicos do paciente, e internet das coisas (*internet of things*).

A medicina de precisão é um termo usado para descrever o tratamento individualizado que engloba o uso de novos diagnósticos e terapias, direcionados às necessidades de um paciente com base em suas próprias características genéticas, biomarcadores, fenótipo e características psicossociais (JAMESON; LONGO, 2015). Neste contexto o uso de aprendizado de máquina pode ajudar em políticas de atendimento que direcionem corretamente os pacientes, tais práticas podem funcionar como uma triagem para que os casos relevantes não sejam desconsiderados pelo uso de procedimentos que se baseiam apenas em médias populacionais.

Desta forma, a correta utilização dos caros e raros leitos de UTI (caros em relação aos atendimentos de saúde básica) podem ser disponibilizados de forma a atender melhor e com mais eficiência.

1.2 Problema de internação em UTI de gestantes de alto risco

O reconhecimento de gravidade e fatores de risco em gestantes é um problema que antecede a internação em UTI. Considerando todas as mulheres gestantes, apenas uma parte dessas mulheres irá desenvolver alguma patologia durante a gestação, e um percentual ainda menor irá desenvolver uma condição ainda mais grave capaz de por em risco a sua vida. Estas condições são chamadas condições potencialmente ameaçadoras da vida (CPAV). A Fig. 1 apresenta um quadro de condições potencialmente ameaçadoras da

vida dividida em quatro categorias: complicações hemorrágicas, complicações hipertensivas, outras complicações (onde temos as infecções) e os indicadores de manejo de gravidade.

Evidências motram que as principais causas associadas a morte materna estão relacionadas a complicações hemorrágicas, hipertensivas e infecções. Estima-se que 80% das mortes maternas poderia ser evitada caso houvesse uma maior agilidade no reconhecimento da gravidade da doença com posterior encaminhamento ao serviçõe de cuidado adequados (SOARES et al., 2016).

COMPLICAÇÕES HEMORRÁGICAS	
Descolamento prematuro de placenta	Hemorragia pós-parto
Placenta prévia / acreta/increta/percreta	Atonia
Prenhez ectópica	Retenção placentária
Rotura uterina	Lacerações de trajeto
Hemorragia grave por aborto	Coagulopatia
COMPLICAÇÕES HIPERTENSIVAS	
Pré-eclampsia grave	Hipertensão grave
Eclâmpsia	HELLP síndrome
OUTRAS COMPLICAÇÕES	
Edema pulmonar	Choque
Convulsões	Insuficiência respiratória aguda
Sepse grave	Acidose
Endometrite pós-parto	Cardiopatia
Endometrite pós-aborto	AVC
Foco urinário	Distúrbios de coagulação
Foco pulmonar	Tromboembolismo
Trombocitopenia < 100 mil	Cetoacidose diabética
Crise tireotóxica	Icterícia / disfunção hepática
	Meningite
	Insuficiência Renal Aguda
INDICADORES DE MANEJO DE GRAVIDADE	
Transfusão de hemoderivados	Intubação não relacionada à anestesia
Acesso venoso central	Retorno à sala cirúrgica
Admissão em UTI	Intervenção cirúrgica maior (histerectomia, laparotomia)
Hospitalização prolongada (>7dias)	Uso de sulfato de magnésio

Figura 1 – Condições potencialmente ameaçadoras da vida (WHO, 1990).

A base da Rede Brasileira de Vigilância de Morbilidade Materna Grave, ou apenas REDE, será utilizada neste estudo, esta base comtempla dados de 27 unidades de referência em diferentes regiões geográficas do país. Segundo (SOARES et al., 2016), 1.5% das mulheres chegam a óbito ao final de um gestação, no entanto existem contextos em que este indicador é consideravelmente maior. Para facilitar o estudo desses casos, criou-se o conceito de morbidade materna grave (MMG) que engloba os casos de CPAV,

near miss e óbito materno. Entende-se *near miss* como os casos onde não houve o óbito, mas, as mulheres sobreviveram à um evento grave durante a gestação. Os critérios para identificação do *near miss* são mostrados na Fig. 2.

A sobrevivência da mulher e a presença de qualquer uma das condições abaixo classifica o caso como *near-miss* materno

Sistema disfuncional	Quadro Clínico	Quadro Laboratorial	Condições de Manejo
Cardiovascular	Choque Falência cardíaca	pH<7.1 Lactato >5 mEq/ml	Uso contínuo de drogas vasoativas Ressuscitação cárdio-pulmonar
Respiratório	Cianose aguda Respiração ofegante Respiração >40bpm, <6bpm	Sat. O_2 <90% por >60 minutos PaO_2 / FiO_2<200 mmHg	Intubação e ventilação não relacionada à anestesia
Renal	Oligúria não responsiva a reposição volêmica e diurético	Creatinina ≥300µmol/l ou ≥3,5 mg/dL	Diálise para insuficiência renal aguda
Coag/hemat.	Distúrbio de coagulação	Trombocitopenia aguda grave <50,000 plaquetas/ml	Transfusão de >5 concentrados de hemácias
Hepático	Icterícia na presença de pré-eclâmpsia	Bilirrubina>100 µmol/l ou > 6,0 mg/dL	
Neurológico	Perda de consciência ≥12 h Perda de consciência E falta de batimentos cardíacos Acidente Vascular Cerebral Mal epilético e Paralisia		
Endócrino		Perda de consciência E a presença de glicose e ácidos cetônicos na urina	
Uterino			Histerectomia seguida de hemorragia ou infecção

Figura 2 – Critérios de identificação de *near miss* materno (WHO, 1990).

O MMG foi proposto como um indicador da qualidade dos serviços maternos disponibilizados à população (CECATTI et al., 2009). Trabalhar com este indicador mostrou-se viável, pois, com a maior volume de dados, pode-se estudar de forma mais adequada condições clínicas que se aplicam não apenas as mulheres que sobreviveram a quadros graves, mas também das que vieram a óbito.

As mortes maternas e o *near miss* possuem em comum a falência de um orgão ou função (CECATTI et al., 2009). Estas duas condições são referidas como Desfecho Materno Grave (do inglês, *Severe Maternal Outcome (SMO)*).

1.3 Objetivo

O objetivo principal deste estudo é apresentar um modelo baseado em aprendizagem supervisionada que, levando em conta as internações de gestantes em centros de

referência, seja capaz de generalizar e treinar um algoritmo que responda de maneira apropriada a recomendação de internação em UTI e esteja aderente os critérios de gravidade e manejo estabelecidos. De modo que, em última análise, as internações reduzam o risco de morte materna.

O objetivo secundário é apontar explicações considerando as predições individuais, visando explicar localmente as recomendações do modelo e garantir que os profissionais de saúde tenham maior embasamento para tomar decisões.

2 Base de dados

2.1 Rede Brasileira de Vigilância de Morbidade Materna Grave

O estudo original envolveu 27 centros obstétricos de referência no Brasil em diferentes regiões geográficas entre os anos de 2009 e 2010. Os pesquisadores realizaram vigilância prospectiva e coleta de dados para a identificação dos casos de morbidade materna grave assim como a avaliação dos fatores e desfechos associados (CECATTI et al., 2009).

Na Fig. 3 é apresentado um esquema mostrando a quantidade de mulheres consideradas no estudo com morbidade materna grave (MMG), *Severe Maternal Outcome (SMO)* e as admitidas em UTI.

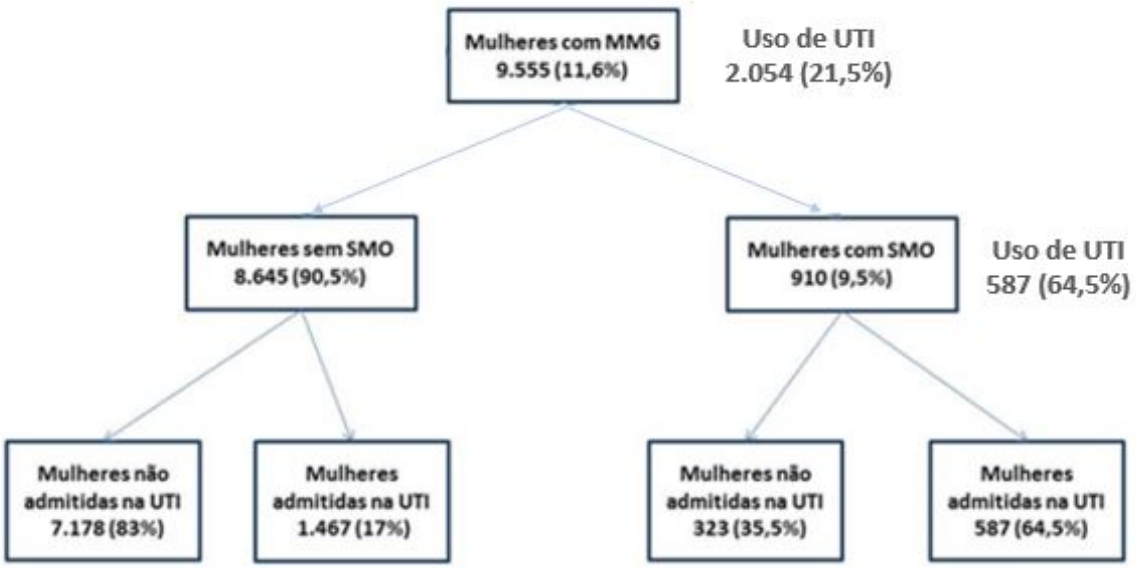

Figura 3 – Esquema para ilustrar a segmentação das mulheres pertinentes ao estudo (Fonte: Própria).

O centro de coordenação do estudo original obteve a aprovação do comitê de ética em pesquisa com seres humanos (CEP) e do Conselho Nacional de Ética em Pesquisa (CONEP). A aprovação dos centros colaboradores neste estudo foi confirmada após o projeto ter sido aprovado por seus respectivos comitês de ética (SOARES et al., 2016). O termo de consentimento livre e esclarecido individual não foi necessário. Os detalhes do método do estudo original podem ser verificados de forma integral em outra publicação (CECATTI et al., 2009).

2.2 Atributos utilizados

Neste estudo foram selecionados da base da REDE mais de 100 atributos e a classe de interesse que é a internação em UTI para todas as gestantes com MMG. A base contém informações de condições observadas após a internação em UTI, além de informações do bebê, o que poderia prejudicar a confiabilidade do modelo proposto. Para contornar esse problema, foi feita uma seleção por especialistas dos atributos relevantes. No Anexo A estão listados os atributos considerados pertinentes a este estudo.

3 Metodologia

3.1 Modelos de classificação

Os modelos de classificação são modelos de aprendizado supervisionado, ou seja, são algoritmos guiados a partir de amostras fornecidas previamente para produzir um modelo geral, capaz de fazer previsões sobre amostras futuras. Em outras palavras, o objetivo de um algoritmo supervisionado de classificação é construir um modelo da distribuição de rótulos de classe sob o espaço de atributos das amostras. O classificador resultante é então usado para atribuir rótulos de classe às amostras de teste em que os valores das variáveis preditoras são conhecidos, mas o valor do rótulo da classe é desconhecido (KOTSIANTIS, 2007).

Matematicamente, o problema de classificação é apresentado da seguinte forma por (ESMI; VALLE, 2014): *Dado o conjunto* $\{(x_i, y_i) \in X \times Y : i = 1, \ldots, n\}$ *em que* X *é um universo arbitrário e* $Y = \{\ell_1, \ldots, \ell_L\}$ *é um conjunto finito deseja-se determinar* $f : X \to Y$ *tal que* $f(x_i) = y_i, \forall i = 1, \ldots, n$. A função $f : X \to Y$ apresentada é chamada *classificador*, X é o *espaço de características ou espaço de atributos* e $Y = \{\ell_1, \ldots, \ell_L\}$ é referido como o *conjunto de rótulos de classe*.

Neste trabalho, os modelos de classificação estudados são modelos baseados em árvores. O problema de classificação para a internação de gestantes de alto risco em UTI será avaliado usando os modelos de Floresta Aleatória, *Gradient Boosting Machine* e *XGBoost*.

3.2 Avaliando modelos de classificação

3.2.1 Testes Diagnósticos

Neste trabalho alguns conceitos aplicados à testes diagnósticos típicos em medicina serão extrapolados para testes que usam modelos preditivos. Estes testes também auxiliam na tomada de decisão médica, no entanto, é recomendado que o profissional de saúde continue sendo responsável pela avaliação do teste e tomada de decisão a respeito do manejo no paciente.

Segundo (BEAGLEHOLE et al., 1993) o propósito de um teste diagnóstico é ajudar a confirmar possíveis diagnósticos provenientes, por exemplo, de atributos demográficos e sintomas apresentados pelo paciente. O desempenho de um teste diagnóstico depende da ausência de desvios (ou ausência de viés) e ainda da reprodutibilidade. Do

ponto de vista epidemiológico para medir a reprodutibilidade deve-se considerar que **o mesmo teste aplicado ao mesmo paciente reproduza os mesmos resultados**. Para avaliar a qualidade de um teste diagnóstico é necessário considerar dois aspectos: *reprodutibilidade* e *acurácia*.

Reprodutibilidade é a capacidade de um teste em produzir resultados consistentes (tipicamente os mesmos resultados) quando os testes forem realizados de maneira independente sob as mesmas condições. Por exemplo, dois especialistas leêm o mesmo exame de forma independente e chegam no mesmo diagnóstico, note que os dois especialistas podem estar igualmente certos ou igualmente errados.

A Figura 4 ilustra a relação entre reprodutibilidade e acurácia. De maneira geral, um teste com baixa reprodutibilidade e alta acurácia terá pouca utilidade, pois, não será considerado um teste confiável. Da mesma forma que uma alta reprodutibilidade e uma baixa acurária também invalidam o uso do teste pois os resultados podem estar errados. A reprodutibilidade e a acurácia devem ser adequadamente mensuradas, pois, só assim podemos validar a qualidade da informação produzida pelo teste.

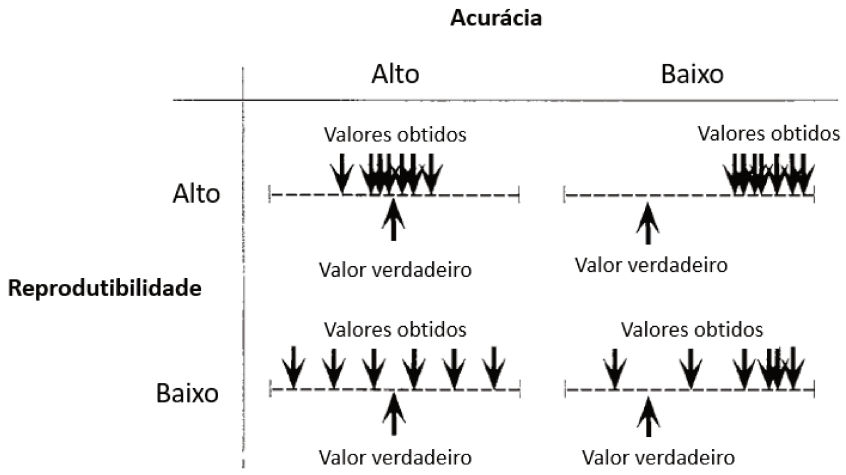

Figura 4 – Relação entre o valor verdadeiro de uma medida quantitativa e o valor obtido pelo estudo considerando baixa e alta validade e reprodutibilidade. Adaptado de (BEAGLEHOLE et al., 1993).

Com base nestes conceitos, o modelo de classificação foi avaliado. Fixou-se um conjunto de teste com as mesmas pacientes de forma que todos os modelos propostos pudessem ser avaliados de forma equivalente. A variância foi usada para avaliar a reprodutibilidade do teste, pois, uma alta acurácia com baixa variância indica que o modelo consistentemente prediz o mesmo diagnóstico para a paciente avaliada, neste caso, recomendação ou não recomendação de internação em UTI.

Validação de testes diagnósticos

A validação ou acurácia de um teste diagnóstico do ponto de vista médico, refere-se a quanto um teste é útil para diagnosticar um evento (acurácia simultânea) ou predizê-lo (acurácia preditiva). Para fazer esta validação, compara-se os resultados obtidos no teste com a internação em UTI indicada pelos especialistas. Como os dados foram obtidos a partir de diversos centro de referência nacionais, é razoável assumir tal premissa.

Os indicadores usados como referência neste trabalho para medir a qualidade das predições foram:

- Acurácia (ACC): probabilidade do indivíduo ser corretamente classificado de acordo com o padrão-ouro estabelecido;
- Sensibilidade (SENS): probabilidade de um teste positivo nos indivíduos verdadeiramente positivos, ou seja, diagnosticar corretamente as gestantes que necessitem de UTI;
- Especificidade (ESPEC): probabilidade de um teste negativo nos indivíduos verdadeiramente negativos, ou seja, diagnosticar corretamente as gestantes que não necessitem de UTI;
- Valor preditivo positivo (VPP): probabilidade do indivíduo ser verdadeiramente positivo quando o teste é positivo, ou seja, indica a probabilidade da gestante verdadeiramente necessitar de UTI caso teste recomende UTI.
- Valor preditivo negativo (VPN): probabilidade do indivíduo ser verdadeiramente negativo quando o teste é negativo, ou seja, indica a probabilidade da gestante verdadeiramente não necessitar de UTI caso teste não recomende UTI.
- Razão de verossimilhança positiva (RVP): probabilidade de um teste positivo em indivíduos verdadeiramente positivos dividida pela probabilidade de um teste positivo em indivíduos verdadeiramente negativos. Varia de $[1, \infty)$, o valor igual a 1 indica que a probabilidade do teste ser positivo para indivíduos verdadeiramente positivos e verdadeiramente negativos é mesma, o que torna o teste inútil, então o ideal é que esse valor seja o maior possível.
- Prevalência estimada (PE): proporção de indivíduos indicados pelo teste como positivos entre a população total, ou seja, indica a proporção de internações em UTI recomendadas pelo teste.
- Prevalência real (PR): proporção de indivíduos verdadeiramente positivos entre a população total, ou seja, indica a proporção real de internação em UTI.

- F1 *Score* (F1): média harmônica entre o valor preditivo positivo e a sensibilidade, quanto mais perto de 1, melhor.

Todos os indicadores mencionados acima são extraídos da matriz de confusão (Fig. 5) gerada a partir dos resultados dicotômicos obtidos pelos classificadores.

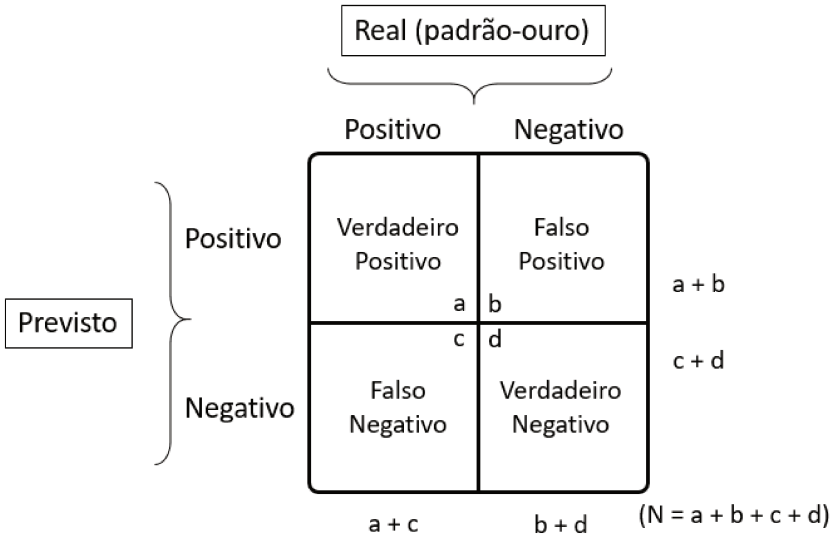

Figura 5 – Matriz de confusão - relação entre os resultados esperados e obtidos do teste diagnóstico.

Adaptando os conceitos para o contexto do estudo, entende- se a classe *positivo* da matriz de confusão como internação em UTI (real) ou recomendação de internação em UTI (previsto), já a classe *negativo* se entende como a não internação em UTI (real) ou a não recomendação de internação em UTI (previsto). Seguem as fórmulas para o cálculo dos indicadores listados:

$$ACC = \frac{(a+d)}{N}$$

$$SENS = \frac{a}{(a+c)}$$

$$ESPEC = \frac{d}{(b+d)}$$

$$VPP = \frac{a}{(a+b)}$$

$$VPN = \frac{d}{(c+d)}$$

$$RVP = \frac{a/(a+c)}{b/(b+d)}$$

$$PE = \frac{(a+b)}{N}$$

$$PR = \frac{(a+c)}{N}$$

$$F1 = 2 \times \frac{(VPP \times SENS)}{(VPP + SENS)}$$

3.3 Modelos baseados em árvores

3.3.1 Classificadores estruturados em árvores binárias

Todas as definições referentes à caracterização de árvores presentes neste trabalho foram extraídas de (BREIMAN et al., 1984).

Classificadores estruturados em árvores binárias, ou apenas classificadores estruturados em árvores, são construídos por repetidas divisões de subconjuntos de X em dois subconjuntos descendentes, começando pelo próprio X. Este processo, para uma hipotética árvore de 6 classes está descrito na Figura 6.

O algoritmo CART (*Classification And Regression Trees*) proposto por Leo Breiman (BREIMAN et al., 1984) em meados de 1980 e consiste em um modelo não paramétrico (isto é, não assume premissas a respeito da distribuição das variáveis) que pode ser usado tanto em modelos de classificação quanto de regressão, este algoritmo é capaz de trabalhar com variáveis disretas, contínuas, categóricas e dados faltantes o que popularizou o uso de árvores como preditores nas últimas décadas.

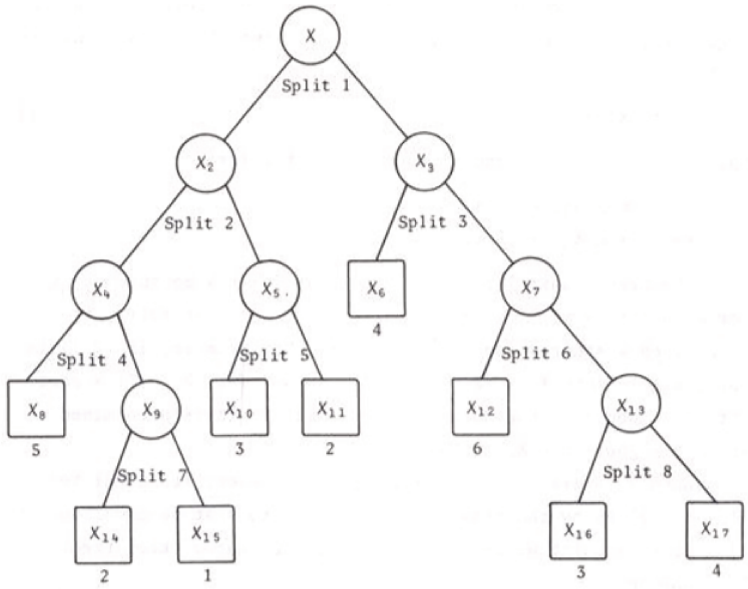

Figura 6 – Exemplo de um classificador com 6 classes estruturado em uma árvore binária. (Fonte: (BREIMAN et al., 1984)).

A estrutura que será apresentada a seguir comtempla as árvores de decisão e árvores de regressão.

Na Figura 6, X_2 e X_3 são disjuntos com $X = X_2 \mathrm{U} X_3$. De maneira análoga, X_4 e X_5 são disjuntos com $X_2 = X_4 \mathrm{U} X_5$ e $X_3 = X_6 \mathrm{U} X_7$. Os subconjuntos X_6, X_8, X_{10},

X_{11}, X_{12}, X_{14}, X_{15}, X_{16} e X_{17} onde não há novas divisões são chamados de subconjuntos terminais. Ainda na Figura 6 os subconjutos terminais são indicados por retângulos e os não terminais são indicados por círculos.

Os subconjuntos terminais formam uma partição de X. Cada subconjunto terminal é designado por um rótulo de classe. Pode haver dois ou mais subconjuntos terminais com o mesmo rótulo de classe. A partição A_j, onde j são os rótulos da j-ésima classe, é obtida juntando todos os subconjuntos terminais correspondentes a esta classe. Assim, para este exemplo em que $j = 1, \ldots, 6$ temos,

$$\begin{aligned} A_1 &= X_{15} \\ A_3 &= X_{10} \cup X_{16} \\ A_5 &= X_8 \\ A_2 &= X_{11} \cup X_{14} \\ A_4 &= X_6 \cup X_{17} \\ A_6 &= X_{12}. \end{aligned} \tag{3.1}$$

Os *splits* são formados por condições nas coordenadas de $\mathbf{x} = (x_1, x_2, \ldots)$. Por exemplo, o *split 1* de X em X_2 e X_3, poderia ser descrito como:

$$X_2 = \{\mathbf{x}; x_4 \leqslant 7\}, \quad X_3 = \{\mathbf{x}; x_4 > 7\}. \tag{3.2}$$

O *split 3* de X_3 em X_6 e X_7, poderia ser descrito por exemplo como:

$$\begin{aligned} X_6 &= \{\mathbf{x} \in X_3; x_3 + x_5 \leqslant -2\} \\ X_7 &= \{\mathbf{x} \in X_3; x_3 + x_5 > -2\} \end{aligned}. \tag{3.3}$$

Desta forma, um classificador dado por uma árvore prevê a classe j para uma observação $\mathbf{x} \in X$. Assim, se $\mathbf{x} \in A_j$, a partir da definição do primeiro *split*, é determinado se $\mathbf{x}$ pertence a X_2 ou X_3. Por exemplo, se (3.2) for usado, $\mathbf{x}$ pertence à X_2, se $x_4 \leqslant 7$ e a X_3 se $x_4 > 7$. Se $\mathbf{x}$ pertence à X_3, então a partir da definição do *split 3*, é determinado se $\mathbf{x}$ pertence à X_6 ou X_7. Deste modo a árvore é avaliada e $\mathbf{x}$ pertence à um subconjunto terminal, sua classe prevista é fornecida pelo rótulo de classe anexado a esse subconjunto terminal. Neste ponto, mudamos a terminologia para a teoria das árvores. A partir de agora, um subconjunto X_t será chamado de nó t da árvore e $X = X_1$ será o nó raiz. Os subconjuntos terminais se tornam nós terminais, ou ainda, folhas e os subconjuntos não terminais serão nós não terminais.

Toda a construção de classificadores estruturados em árvores envolve 3 decisões:

- A seleção dos *splits*.
- A decisão de quando um nó deve se tornar uma folha ou não.

- A designação de cada folha a uma classe.

Dada a estrutura básica de uma árvore binária, estas decisões dependem de qual algoritmo de treinamento é utilizado.

3.3.2 Árvores de Decisão

O método de árvores de decisão utiliza regras para dividir os dados. Neste trabalho serão abordados apenas árvores de decisão que utilizam regras binárias para fazer a partição ou *split* dos dados (BREIMAN, 2001). Em modelos baseados em árvores o primeiro nó é chamado raiz, a partir desse nó são geradas sucessivas ramificações que são ligadas à outros nós, chamados nós internos, as ligações acontecem até atingirem um nó terminal ou folha. Um exemplo genérico da estrutura de uma árvore de decisão é mostrado na Figura 6.

Para o algoritmo CART (*Classification And Regression Trees*) proposto por Leo Breiman, os *splits* acabam quando todos os atributos, ou variáveis, forem incluídas no modelo ou quando todas as amostras forem classificadas corretamente (BREIMAN et al., 1984).

Em problemas de classificação com n de classes o critério padrão para realizar o *split* em um nó é o índice Gini, mostrado na Equação (3.4) onde a denota a classe e p_a é a proporção da classe a no respectivo nó.

$$G_{(\text{nó})} = 1 - \sum_{a=1}^{n} \left(p_a^2\right) \tag{3.4}$$

Um valor pequeno para G indica que aquele nó contém predominante amostras de uma classe a (CAO et al., 2012). O melhor *split* para um conjunto de dados pode ser calculado através do índice Gini, pois este representa o ganho de informação $\left(G_{(\text{nó pai})} - G_{(\text{nó filho})}\right)$ considerando todos os possíveis *splits*. A variável que apresentar o menor índice Gini é escolhida para criar o *split* neste nó.

Dado um conjunto de dados, para cada atributo i, o número total de possíveis *splits* é $K_i - 1$ onde K_i é o número de valores distintos que esse atributo assume considerando as amostras de treinamento pertencentes ao respectivo nó . O índice Gini é calculado para a todos os possíveis *splits* de cada atributo e o que apresentar o menor índice de Gini será escolhido para criar o *split* neste nó.

Nota-se que criar uma árvore de decisão com grandes conjuntos de dados onde todas, ou a maioria, das variáveis ou atributos são categóricos ou contínuos requer muito esforço computacional devido ao elevado número de possíveis combinações. No entanto,

após a construção do modelo as amostras de validação ou teste são estimadas de maneira muito rápida (SANTANA et al., 2020).

3.3.3 Árvores de Regressão

A estrutura de uma árvore regressão é semelhante à estrutura de árvores de decisão. No entanto, ao invés do índice Gini, o critério para definir o *split* em um nó é feito através da minimização do erro quadrático médio, dado na Equação (3.5), onde y_i é o valor observado nas amostras do conjunto de treino $\{y_i, \mathbf{x}_i\}_1^N$ e $\bar{y}$ é a média dos valores observados contidos neste nó.

$$\frac{1}{\mathrm{n}} \sum_{i=1}^{\mathrm{N}} (y_i - \bar{y})^2 \tag{3.5}$$

O treinamento de uma árvore de regressão para quando um critério de parada é atingido. No caso do CART, o critério de parada padrão implantado no *software* MATLAB é quando se tem 5 amostras em um nó terminal ou folha, ou ainda, se o erro quadrático médio é menor que $10^{-6} \times \|y_i - \bar{y}\|$ (MATHWORKS, 2017). Este processo tem por objetivo reduzir erros de generalização que geram sobreajuste (BREIMAN et al., 1984).

As árvores de regressão estimam o valor de saída associado a uma determinada amostra/observação, utilizando a média aritmética das amostras y_i do conjunto de treino $\{y_i, \mathbf{x}_i\}_1^N$ contidas no nó terminal/folha (BREIMAN et al., 1984).

Na Figura 7 é mostrada a aplicação de uma amostra desconhecida w sobre uma árvore de regressão genérica que foi construída considerando apenas dois atributos (x_1, x_2). Neste exemplo genérico o valor das folhas representa a média aritmética do valor observado y das amostras presentes em cada folha. A caminho em azul representa o caminho que a amostra desconhecida w percorreu na árvore onde $x_1 = 2$ e $x_2 = 3$, sendo o valor estimado para a amostra desconhecida nesta árvore de regressão foi de $80,7$.

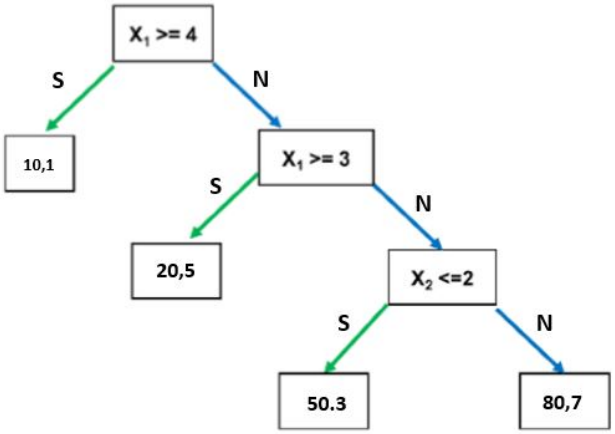

Figura 7 – Exemplo genérico de uma árvore de regressão com apenas 2 atributos (x_1, x_2).

3.3.4 *Bagging*

Em problemas de classificação, as árvores de decisão e regressão são preditores altamente interpretáveis, pois, os atributos e pontos de *split* escolhidos mostram com clareza as informações consideradas na atribuição da classe. No entanto, também são preditores facilmente suscetíveis a sobreajuste (ou, do inglês, *overfitting*) (BREIMAN, 1996). Para mitigar este problema, Breiman introduziu em 1996 (BREIMAN, 1996) o conceito de *bagging*.

Bagging é um método para gerar várias versões de um preditor e usá-los para obter um preditor agregado. A média das agregações pode considerar a pluralidade em um resultado numérico ao prever uma classe. Testes em conjuntos de dados reais e simulados usando classificação, árvores de regressão e seleção de subconjunto na regressão linear mostram que o *bagging* pode fornecer ganhos substanciais em precisão (BREIMAN, 1996). Ao perturbar o conjunto de aprendizagem, usando técnicas como o *bootstrap* pode causar mudanças no preditor construído melhorando a acurácia (BREIMAN, 1996).

Seja $\mathcal{L}$ um conjunto de treinamento cujos dados podem ser representados por $\{(y_n, \mathbf{x}_n), n = 1 \ldots, N\}$ onde os y_i são os rótulos de classe ou a resposta numérica associadas as amostras $\mathbf{x_i} \in X$ onde X é o espaço de atributos. Assuma que a partir deste conjunto de treinamento foi obtido um preditor $\varphi(\mathbf{x}, \mathcal{L})$, ou seja, dado $\mathbf{x}$ pode-se estimar y por $\varphi(\mathbf{x}, \mathcal{L})$. Agora suponha uma sequencia de conjuntos de treinamento $\{\mathcal{L}_k\}$ onde cada sequencia consiste em N observações independentes a partir da mesma distribuição subjacente a $\mathcal{L}$. O objetivo é criar um preditor melhor que $\varphi(\mathbf{x}, \mathcal{L})$ usando $\{\mathcal{L}_k\}$ e a sequencia correspondente de preditores $\{\varphi(\mathbf{x}, \mathcal{L}_k)\}$.

Se y é numérico, um procedimento usual para estimar y seria utilizar a média

de $\varphi(\mathbf{x}, \mathcal{L}_k)$ sobre k, ou seja, tomando $\varphi_A(\mathbf{x}) = E_{\mathcal{L}}\varphi(\mathbf{x}, \mathcal{L})$ onde $E_{\mathcal{L}}$ denota da esperança sobre $\mathcal{L}$ e φ_A representa a *agregação* (do inglês, *aggregation*) de preditores.

Se $\varphi(\mathbf{x}, \mathcal{L})$ prediz uma classe $j \in \{1, \ldots, J\}$, então um método de agregação de $\varphi(\mathbf{x}, \mathcal{L}_k)$ é pelo voto: $\varphi_A(x) = \underset{\cup N_j}{\arg\max} N_j$ onde $N_j = \#\{k; \varphi(\mathbf{x}, \mathcal{L}_k) = j\}$, isto é, N_j é a quantidade de classificadores que associam $\mathbf{x}$ a classe j.

Normalmente, temos um único conjunto de treinamento $\mathcal{L}$, sem as réplicas de $\mathcal{L}$. Ainda assim, uma imitação do processo que leva a φ_A pode ser feita obtendo os conjuntos de dados replicados $\left\{\mathcal{L}^{(B)}\right\}$. Cada um destes conjuntos possui N observações tiradas ao acaso, mas com reposição, de $\mathcal{L}$. Este processo de amostragem se chama *amostragem bootstrap* onde cada $(y_n, \mathbf{x}_n)$ pode aparecer repetidas vezes ou não aparecer em qualquer conjunto de dados $\mathcal{L}^{(B)}$.

Assim, $\left\{\mathcal{L}^{(B)}\right\}$ são os conjuntos de dados replicados extraídos da distribuição bootstrap e que se aproximam da distribuição subjacente de $\mathcal{L}$ (BREIMAN, 1996). Usando as amostras bootstrap $\left\{\mathcal{L}^{(B)}\right\}$ de $\mathcal{L}$ e gerando $\left\{\varphi\left(\mathbf{x}, \mathcal{L}^{(B)}\right)\right\}$, obtem-se o preditor ou classificador φ_A. Este procedimento é chamado de ***b**ootstrap **agg**regat**ing***, de onde vem o acrônimo *bagging*.

Neste trabalho foram aplicados 3 algoritmos estruturados em árvores para comparação de resultados: Floresta Aleatória, GBM e *XGBoost*. Todos eles usam o *bagging* como um técnica para reduzir erros de generalização (BREIMAN, 1996).

3.3.5 *Boosting*

Boosting é uma técnica geral para melhorar a acurácia de qualquer algoritmo de aprendizagem (FREUND; SCHAPIRE; ABE, 1999). Esta técnica tem suas raízes em uma estrutura chamada modelo de aprendizagem PAC (VALIANT, 1984), veja Kearns e Vazirani (KEARNS; VAZIRANI; VAZIRANI, 1994) para uma boa introdução a este modelo. Kearns e Valiant (KEARNS; VALIANT, 1988) (KEARNS, 1994) foram os primeiros a colocar a questão de que um algoritmo de aprendizagem "fraco" (do inglês, *weak learner*) que é apenas ligeiramente melhor do que a escolha aleatória. No modelo PAC, um *weak learner* pode ser "impulsionado" (do inglês, *boosted*) em um algoritmo de aprendizagem "forte", ou seja, com melhor acurácia.

O AdaBoost, introduzido em 1995 por Freund e Schapire (FREUND; SCHAPIRE, 1997), é um dos algoritmos baseados em *boosting* mais populares e baseia-se em reponderação dos dados com base nos resíduos. No entanto, neste trabalho será abordado o estudo do Gradiente Impulsionado, ou ainda, *Gradient Boosting*, introduzido por no início dos anos 2000 por Friedman (FRIEDMAN, 2002).

Gradient Boosting

Segundo Friedman (FRIEDMAN, 2002), o *Gradient Boosting* constrói modelos de regressão aditivos ajustando sequencialmente uma função parametrizada simples aos pseudo-resíduos atuais por mínimos quadrados em cada iteração. Os pseudo-resíduos são o gradiente da função de perda que está sendo minimizada, em relação aos valores do modelo em cada observação no conjunto de treinamento, avaliados na iteração atual.

No problema de estimativa de uma função, tem-se um modelo que consiste em uma saída ou variável resposta y e um conjunto de variáveis explicativas ou atributos $\mathbf{x} = \{x_1, \cdots, x_n\}$. Dado um conjunto de treinamento $\{y_i, \mathbf{x}_i\}_1^N$, o objetivo é encontrar uma função $F^*(\mathbf{x})$ que mapeia $\mathbf{x}$ em y, de modo que sobre a distribuição conjunta de todos os $(y, \mathbf{x})-$valores, o valor esperado de uma determinada função de perda $\Psi(y, F(\mathbf{x}))$ seja minimizado:

$$F^*(\mathbf{x}) = \arg\min_{F(\mathbf{x})} E_{y,\mathbf{x}} \Psi(y, F(\mathbf{x})) \tag{3.6}$$

O método de *Boosting* aproxima F^* por uma função F que possui uma expansão "aditiva" da forma:

$$F(\mathbf{x}) = \sum_{m=0}^{M} \beta_m h\left(\mathbf{x}; \mathbf{a}_m\right). \tag{3.7}$$

As funções $h(\mathbf{x}; \mathbf{a})$ são chamadas de preditores base, e usualmente, são funções bem simples com parâmetros $\mathbf{a} = \{a_1, a_2, \cdots\}$. No caso deste trabalho, as funções $h(\mathbf{x}; \mathbf{a})$ são árvores binárias e sua estrutura é explicada com mais detalhes na Seção 3.3.1. Os coeficientes da expansão $\{\beta_m\}_0^M$ e os parâmetros $\{\mathbf{a}_m\}_0^M$ são definidos a medida que o treinamento do modelo avança. Seja $F_0(\mathbf{x})$ o modelo inicial e M o número máximo de etapas do algoritmo. A cada etapa m o algoritmo obtém um modelo $F_m(\mathbf{x})$ dado por

$$F_m(\mathbf{x}) = F_{m-1}(\mathbf{x}) + \beta_m h\left(\mathbf{x}; \mathbf{a}_m\right) \tag{3.8}$$

onde β_m e a_m são determinados pelo seguinte problema de minimização:

$$(\beta_m, \mathbf{a}_m) = \arg\min_{\beta, \mathbf{a}} \sum_{i=1}^{N} \Psi\left(y_i, F_{m-1}\left(\mathbf{x}_i\right) + \beta h\left(\mathbf{x}_i; \mathbf{a}\right)\right) \tag{3.9}$$

Note que o termo $\beta h\left(\mathbf{x}_i; \mathbf{a}\right)$, dado em termos do preditor base h é adicionado ao modelo F_{m-1}. Assim o modelo F_m pode ser visto como um refinamento ou aprimoramento de F_{m-1}. O algoritmo resolve o problema de estimar os parâmetros $\{\beta_m\}_1^M$ e $\{\mathbf{a}_m\}_1^M$ resolvendo o problema (3.9) em dois passos. Aqui estamos supondo que a função de perda $\Psi(y, F(\mathbf{x}))$ é diferenciável com respeito ao segundo argumento.

O primeiro passo é obter $\mathbf{a}_m$, resolvendo o problema de minimização (3.10),

$$\mathbf{a}_m = \arg\min_{\mathbf{a},\rho} \sum_{i=1}^{N} \left[\tilde{y}_{im} - \rho h\left(\mathbf{x}_i; \mathbf{a}\right)\right]^2, \tag{3.10}$$

onde $\tilde{y}_{im}$ é o pseudo-resíduo atual dado por

$$\tilde{y}_{im} = -\left[\frac{\partial \Psi\left(y_i, F\left(\mathbf{x}_i\right)\right)}{\partial F\left(\mathbf{x}_i\right)}\right]_{F(\mathbf{x})=F_{m-1}(\mathbf{x})}. \tag{3.11}$$

Uma vez obtido o parâmetro a_m, o valor ótimo do coeficiente β_m é determinado através do seguinte problema de otimização:

$$\beta_m = \arg\min_{\beta} \sum_{i=1}^{N} \Psi\left(y_i, F_{m-1}\left(\mathbf{x}_i\right) + \beta h\left(\mathbf{x}_i; \mathbf{a}_m\right)\right). \tag{3.12}$$

Esta estratégia substitui um problema potencialmente difícil de otimização dado na Equação (3.9), por dois problemas mais simples de minimização descritos nas Equações (3.10) e (3.12). Note que β_m pode ser obtido resolvendo $\dfrac{\partial \psi}{\partial \beta} = 0$.

Gradient Tree Boosting

Gradient Tree Boosting, ou ainda, Árvores Impulsionadas por Gradiente, é uma abordagem do método de *Gradient Boosting* especializada para preditores base $h\left(\mathbf{x}_i; \mathbf{a}_m\right)$ dados por árvores de regressão com L nós terminais (sessão 3.3.3). A cada iteração m, a árvore de regressão particiona o espaço de atributos X em L regiões disjuntas $\{R_{lm}\}_{l=1}^{L}$ e prediz um valor $\bar{y}_{lm}$ para cada uma delas, conforme Equação (3.13), onde $\chi_{R_{lm}}(\mathbf{x})$ é a função indicadora ou característica do conjunto R_{lm} e $\bar{y}_{lm} = \text{media}_{\mathbf{x}_i \in R_{lm}}\left(\tilde{y}_{im}\right)$ é a média dos pseudo-resíduos dado como na(Equação (3.11) para cada região R_{lm}.

$$h\left(\mathbf{x}; \{R_{lm}, \bar{y}_{lm}\}_1^L\right) = \sum_{l=1}^{L} \bar{y}_{lm} \chi_{R_{lm}}(\mathbf{x}) \tag{3.13}$$

Os parâmetros deste preditor base, neste caso, árvores de regressão, são as variáveis onde ocorrerão os *splits* e, ainda, seus respectivos pontos de *split*. Isto definirá as regiões $\{R_{lm}\}_{l=1}^{L}$ correspondentes as partições definidas na m-ésima iteração. Estes parâmetros são induzidos da melhor maneira em uma estratégia de cima para baixo, considerando o problema da Equação (3.10). O peso associado ao nó terminal l da árvore de regressão é obtido resolvendo a Equação (3.14) para cada uma das regiões R_{lm} da árvore de regressão.

Dado que a árvore (3.13) estima um valor constante $\bar{y}_{lm}$ dentro de cada região R_{lm}, a solução para a Equação (3.12) se reduz a uma simples "locação" estimada baseada no critério de perda:

$$\gamma_{lm} = \arg\min_{\gamma} \sum_{\mathbf{x}_i \in R_{lm}} \Psi\left(y_i, F_{m-1}\left(\mathbf{x}_i\right) + \gamma\right) \tag{3.14}$$

A aproximação inicial $F_{m-1}(\mathbf{x})$ é recalculada separadamenta para cada região correspondente. Pode-se ainda considerar um parâmetro de "encolhimento" $0 < \nu \leqslant 1$ controla a taxa de aprendizagem do processo. Assim, definimos a árvore (3.15)

$$h\left(\mathbf{x}; \{R_{lm}, \bar{y}_{lm}\}_1^L\right) = \sum_{l=1}^{L} \nu\gamma_{lm}\chi_{R_{lm}}(\mathbf{x}) \tag{3.15}$$

e

$$F_m(\mathbf{x}) = F_{m-1}(\mathbf{x}) + h\left(\mathbf{x}; \{R_{lm}, \nu\gamma_{lm}\}_1^L\right). \tag{3.16}$$

onde $a_{l,m} = \{R_{lm}, \nu\gamma_{lm}\}$, $l = 1, \ldots, L$.

Note que, como β_m está incorporado ao peso $\gamma_{l,m}$, a adição do parâmetro ν pode ser vista como uma extensão do modelo (3.8) que pode ser obtido com $\nu = 1$. Empiricamente, verifica-se que pequenos valores para esta taxa como ($\nu \leqslant 0.1$), em geral, reduzem erros de generalização (FRIEDMAN, 2002).

Friedman apresentou vários algoritmos baseados neste modelo para funções de perda específicas. No caso da classificação, que é o que o objeto de interesse deste trabalho, a função Ψ utilizada é a da **probabilidade logarítmica negativa multinomial** para k classes.

Motivado por Breiman (BREIMAN, 1996), uma pequena modificação foi feita no algoritmo do *Gradient Boosting* para incorporar a aleatoriedade como parte integrante do procedimento. Especificamente, em cada iteração, uma subamostra dos dados de treinamento é retirada aleatoriamente (sem substituição) do conjunto de dados de treinamento completo. Esta subamostra selecionada aleatoriamente é então usada, em vez da amostra completa, para testar o preditor base (tipicamente, regressões linear e árvores binárias de classificação) e, então, calcular a atualização do modelo para a iteração atual. Esta modificação foi descrita por Friedman em (FRIEDMAN, 2002) e chama-se *Stochastic Gradient Boosting*.

3.3.6 Floresta Aleatória

Florestas aleatórias são preditores gerados a partir de uma combinação de árvores, de modo que cada árvore depende dos valores de um vetor de parâmetros amostrado de forma independente e com a mesma distribuição para todas as árvores da floresta

(BREIMAN, 2001). O algoritmo de Floresta Aleatória se baseia na aplicação de *bagging* em árvores de decisão, no entanto, além de reamostrar o conjunto de aprendizagem (observações) para cada uma das ávores, também são reamostradas as variáveis (atributos) considerados na construção de cada árvore.

Geralmente, aproximadamente dois terços das amostras do conjunto de treino são usadas para construir uma árvore, estas amostras são denominadas *dentro do saco* (ou, do inglês, *in bag samples*) e o restante das amostras é denominada *fora do saco* (ou, do inglês, *out of bag samples*). As amostras *fora do saco* são usadas para avaliar a performance do modelo.

Além do processo de aleatorização das amostras usando *Bagging*, há também a aleatorização da seleção das variáveis que serão consideradas na definição de cada nó durante a estruturação de cada árvore. O número de variáveis é tipicamente chamado de *mtry*. A utilização destes dois processos de aleatorização e o fato das árvores serem construídas de maneira independente umas das outras, torna o modelo mais robusto e menos suscetível à sobreajuste (do inglês, *overfitting*) (BREIMAN, 2001).

Breiman define Floresta Aleatória como um classificador que consiste em um conjunto de classificadores estruturados por árvores $\{h(\mathbf{x}, \theta_k)\}$ onde $\{\theta_k, k = 1, \ldots\}$ são vetores de parâmetros distribuídos de forma independente e idêntica de forma que cada árvore lança um voto unitário e a classe mais popular é atribuida à amostra $\mathbf{x}$ (BREIMAN, 2001).

Do ponto de vista computacional, o algorítimo pode ser paralelizado, distribuindo a construção das ávores em cada um dos núcleos dos processadores, reduzindo o tempo necessário para treinar o modelo (BREIMAN, 2001).

3.3.7 *Gradient Boosting Machine - GBM*

No início dos anos 2000, Friedman introduziu o algoritmo *Gradient Boosting Machine* (FRIEDMAN, 2001) aplicando os conceitos abordados em (FRIEDMAN, 2002) para problemas de regressão e classificação em duas ou mais classes. Para o caso particular em que os preditores base são árvores de regressão, o método do gradiente impulsionado produz procedimentos competitivos e altamente robustos (FRIEDMAN, 2001).

Seja F^* uma função que minimiza o valor esperado de uma função de perda específica $L(y, F(\mathbf{x}))$, sobre todos os valores do conjunto de treino $\{y_i, \mathbf{x}_i\}_1^N$, isto é,

$$F^* = \arg\min_F E_{y,\mathbf{x}} L(y, F(\mathbf{x})) = \arg\min_F E_{\mathbf{x}} \left[E_y (L(y, F(\mathbf{x}))) \mid \mathbf{x} \right] \tag{3.17}$$

As funções frequentemente empregadas para regressão incluem o erro quadrático $(y - F)^2$ e o erro absoluto $|y - F|$ quando $y \in \mathbb{R}$. Para o problema de classificação binário,

quando $y \in \{-\mathbf{1}, 1\}$, tipicamente utiliza-se o log da probabilidade binomial negativa $\log\left(1 + e^{-2yF}\right)$.

Neste trabalho, $F(\mathbf{x})$ é obtido a partir de expansões "aditivas", conforme Equação (3.18), onde $h(\mathbf{x}; \mathbf{a}_m)$ são pequenas árvores de regressão ponderadas pelo parâmetro de *encolhimento* ν, para mais detalhes vide Seções 3.3.1 e 3.3.5, ou ainda, (BREIMAN et al., 1984):

$$F\left(\mathbf{x}; \{\beta_m, \mathbf{a}_m\}_1^M\right) = \sum_{m=1}^{M} \nu \beta_m h\left(\mathbf{x}; \mathbf{a}_m\right) \tag{3.18}$$

O algoritmo GBM pode ser aplicado em vários problemas e não apenas problemas de classificação. Neste trabalho, tem-se por objetivo um modelo de classificação binária, então, será onde o algoritmo será apresentado em detalhes.

Seja $L(y, F)$ a função de perda, dada pela probabilidade logarítmica negativa binominal onde y é a classe observada de $\mathbf{x}$ e y pode corresponde à classe -1 ou 1. Mais precisamente, temos

$$L(y, F) = \log(1 + \exp(-2yF)), \quad y \in \{-1, 1\}, \tag{3.19}$$

onde

$$F(\mathbf{x}) = \frac{1}{2} \log\left[\frac{\Pr(y = 1 \mid \mathbf{x})}{\Pr(y = -1 \mid \mathbf{x})}\right] \tag{3.20}$$

O pseudo-resíduo $\tilde{y}_i$ é o gradiente de L com respeito ao segundo argumento em $F_{m-1}(x)$, isto é,

$$\tilde{y}_i = -\left[\frac{\partial L\left(y_i, F\left(\mathbf{x}\right)\right)}{\partial F\left(\mathbf{x}\right)}\right]_{F(\mathbf{x}) = F_{m-1}(\mathbf{x}_i)} = 2y_i / \left(1 + \exp\left(2y_i F_{m-1}\left(\mathbf{x}_i\right)\right)\right) \tag{3.21}$$

Tomando as Equações (3.19) e (3.18), obtem-se a Equação (3.22) que minimiza o problema de otimização proposto na Equação (3.17)

$$\left(\rho_m, \mathbf{a}_m\right) = \arg\min_{\rho} \sum_{i=1}^{N} \log\left(1 + \exp\left(-2y_i\left(F_{m-1}\left(\mathbf{x}_i\right) + h\left(\mathbf{x}_i; \mathbf{a}_m\right)\right)\right)\right). \tag{3.22}$$

Com árvores de regressão representadas pela Equação (3.13) da Seção 3.3.5, para cada folha ou nó terminal R_{ljm} obtem-se um valor de γ_{ljm}, conforme Equação (3.23).

$$\gamma_{ljm} = \arg\min_{\gamma} \sum_{\mathbf{x}_i \in R_{ljm}} \log\left(1 + \exp\left(-2y_i\left(F_{m-1}\left(\mathbf{x}_i\right) + \gamma\right)\right)\right) \tag{3.23}$$

Como não há uma solução fechada para a Equação (3.23), é usada uma aproximação de Newton-Raphson de passo único onde a Equação (3.23) se torna:

$$\gamma_{ljm} = \sum_{\mathbf{x}_i \in R_{jm}} \tilde{y}_i / \sum_{\mathbf{x}_i \in R_{ljm}} |\tilde{y}_i| \left(2 - |\tilde{y}_i|\right) \tag{3.24}$$

onde $\tilde{y}_i$ é dado pela Equação (3.21).

A aproximação final $F_M(\mathbf{x})$ é representada pela probabilidade logarítmica sendo, então, necessário transformar a estimativa em probabilidade através das Equações (3.25) e (3.26):

$$p_+(\mathbf{x}) = \widehat{\Pr}(y = 1 \mid \mathbf{x}) = 1/\left(1 + e^{-2F_M(\mathbf{x})}\right) \tag{3.25}$$

$$p_-(\mathbf{x}) = \widehat{\Pr}(y = -1 \mid \mathbf{x}) = 1/\left(1 + e^{2F_M(\mathbf{x})}\right) \tag{3.26}$$

Este cálculo é necessário pois trata-se de um problema de classificação em duas classes, ou seja, precisamos estimar se $\mathbf{x}$ pertence à classe 1 ou -1. Para isto, usamos a Equação (3.27) onde $\hat{y}(\mathbf{x})$ é a classe estimada de $\mathbf{x}$,

$$\hat{y}(\mathbf{x}) = 2 \cdot 1\left[p_+(\mathbf{x}) > p_-(\mathbf{x})\right] - 1. \tag{3.27}$$

3.3.8 *Extreme Gradient Boosting - XGBoost*

O *Extreme Gradient Boosting (XGBoost)* é um algoritmo que pode ser usado em problemas de classificação e regressão, sendo considerado um aprimoramento do algoritmo *Gradient Boosting* introduzido por Friedman (FRIEDMAN, 2001). Suas modificações são descritas por Chen e Guestrin em (CHEN; GUESTRIN, 2016).

O *XGBoost* é recente na literatura e utiliza os princípios de árvores de regressão (vide Seção 3.3.3), *bagging* (vide Seção 3.3.4) e *tree boosting* (vide Seção 3.3.5). A junção destas diferentes técnicas, aliada a um algoritmo escalável e técnicas de regularização produziu um algoritmo de aprendizagem de máquina robusto e largamente utilizado (CHEN; GUESTRIN, 2016).

Seja $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}\ (|\mathcal{D}| = n, \mathbf{x}_i \in \mathbb{R}^m, y_i \in \mathbb{R})$ um conjunto de dados com n amostras e m atributos ou variáveis, $\mathcal{F}$ é o conjunto das árvores de regressão em $\mathbb{R}^m$ com T folhas, onde K é o número total de árvores e f é uma função no espaço $\mathcal{F}$ dado por

$$\mathcal{F} = \left\{f(\mathbf{x}) = w_{q(\mathbf{x})}\right\}\left(q : \mathbb{R}^m \to T, w \in \mathbb{R}^T\right) \tag{3.28}$$

mais conhecido como CART (BREIMAN et al., 1984). Por conveniência de notação, vamos denotar $f_k(\mathbf{x}) = h\left(\mathbf{x}, \{(R_{jk}, w_{jk})\}_{j=1}^{T}\right), \forall k$. Além disso, para cada árvore f_k tem-se a função $q_k : \mathcal{R}^m \rightarrow 1, \ldots, T$ dada por $q_k(\mathbf{x}) = j$ se $\mathbf{x} \in R_{jk}$ e o vetor $w_k = (w_{1k}, w_{2k}, \ldots, w_{Tk})^T \in \mathbb{R}^T$, q representa a estrutura de uma árvore de regressão que atribui uma folha j a uma amostra $\mathbf{x}$. Note que, cada f_k corresponde a uma árvore de regressão independente q cujos pesos das suas folhas são representados por w. Diferentemente das árvores de decisão, árvores de regressão contém um peso w contínuo em cada uma das folhas. Assim, o peso w atribuído à uma amostra desconhecida $\mathbf{x}$ é a soma dos w obtidos em cada uma das folhas para todas as K árvores de regressão ponderados por um *fator de encolhimento* $\nu \in (0, 1]$, esta técnica foi introduzida por Friedman (FRIEDMAN, 2002) (vide Seção 3.3.5 e 3.3.7) e tem como objetivo prevenir o sobreajuste (CHEN; GUESTRIN, 2016). Desta forma, a predição de um conjunto de árvores $\hat{y}_i$ é dada por

$$\hat{y}_i = \phi\left(\mathbf{x}_i\right) = \sum_{k=1}^{K} \nu f_k\left(\mathbf{x}_i\right), \quad f_k \in \mathcal{F}. \tag{3.29}$$

Afim de obter um modelo treinado, é necessário minimizar a seguinte função:

$$\begin{aligned} \mathcal{L}(\phi) &= \sum_i \Psi\left(\hat{y}_i, y_i\right) + \sum_k \Omega\left(f_k\right) \text{ onde} \\ \Omega(f_k) &= \gamma T + \frac{1}{2}\lambda \|w_k\|^2 \end{aligned} \tag{3.30}$$

Ψ é a função de perda diferenciável que mede a diferença entre a predição $\hat{y}_i$ e o valor observado y_i, Ω é a função de regularização que tem por objetivo penalizar a complexidade das árvores de regressão.

O modelo descrito pela Equação (3.30) não pode ser otimizado usando os métodos tradicionais de otimização no espaço Euclidiano, pois, possui funções como parâmetros (CHEN; GUESTRIN, 2016). Sendo assim, o modelo será treinado de forma "aditiva". Considere a função objetivo modificada na Equação (3.31), onde $\hat{y}_i^{(t)}$ é a predição da i-ésima amostra do conjunto de dados na t-ésima iteração. Como a minimização é obtida de forma iterativa, é adiconado f_t. Precisamente, é preciso resolver o seguinte problema de minimização $f_k = \arg\min_{f_k \in \mathcal{F}} \mathcal{L}^{(t)}$ onde

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} \Psi\left(y_i, \hat{y}_i^{(t-1)} + f_t\left(\mathbf{x}_i\right)\right) + \Omega\left(f_t\right) \tag{3.31}$$

Pode-se empregar uma aproximação de segunda ordem proposta por Friedman em (FRIEDMAN et al., 2000) para otimizar a função objetivo da Equação (3.31). Seja $g_i = \partial_{\hat{y}(t-1)} \Psi\left(y_i, \hat{y}_i^{(t-1)}\right)$ e $h_i = \partial^2_{\hat{y}(t-1)} \Psi\left(y_i, \hat{y}_i^{(t-1)}\right)$ são, respectivamente, as derivadas parciais de primeira e segunda ordem da função de perda Ψ em relação à $\hat{y}^{(t-1)}$, obtem-se

a seguinte aproximação

$$\mathcal{L}^{(t)} \simeq \sum_{i=1}^{n} \left[\Psi\left(y_i, \hat{y_i}^{(t-1)}\right) + g_i f_t\left(\mathbf{x}_i\right) + \frac{1}{2} h_i f_t^2\left(\mathbf{x}_i\right) \right] + \Omega\left(f_t\right). \tag{3.32}$$

Removendo os termos constantes $\Psi(y_i, \hat{y_i}^{(t-1)})$, chega-se em uma função objetivo simplificada no passo t:

$$\tilde{\mathcal{L}}^{(t)} = \sum_{i=1}^{n} \left[g_i f_t\left(\mathbf{x}_i\right) + \frac{1}{2} h_i f_t^2\left(\mathbf{x}_i\right) \right] + \Omega\left(f_t\right). \tag{3.33}$$

O objetivo aqui é determinar a árvore $f_k \in \mathcal{F}$ que minimiza a função objetivo $\mathcal{L}^{(t)}$ dada como na Equação (3.33). Seja $I_j = \{i \mid q\left(\mathbf{x}_i\right) = j\}$ o conjunto de amostras ou observações da folha j é possível reescrever a Equação (3.33) expandindo Ω, conforme Equação (3.34)

$$\begin{aligned} \tilde{\mathcal{L}}^{(t)} &= \sum_{i=1}^{n} \left[g_i f_t\left(\mathbf{x}_i\right) + \frac{1}{2} h_i f_t^2\left(\mathbf{x}_i\right) \right] + \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_{jt}^2 \\ &= \sum_{j=1}^{T} \left[\left(\sum_{i \in I_j} g_i \right) w_{jt} + \frac{1}{2} \left(\sum_{i \in I_j} h_i + \lambda \right) w_{jt}^2 \right] + \gamma T. \end{aligned} \tag{3.34}$$

Desta forma, para uma árvore específica f_t associada a $q(\mathbf{x})$ é possível calcular o peso ótimo w_{jt}^* da folha j segundo a Equação (3.35),

$$w_{jt}^* = -\frac{\sum_{i \in I_j} g_i}{\sum_{i \in I_j} h_i + \lambda}. \tag{3.35}$$

Assim, obtem-se o valor ótimo da função objetivo.

$$\tilde{\mathcal{L}}^{(t)} = -\frac{1}{2} \sum_{j=1}^{T} \frac{\left(\sum_{i \in I_j} g_i\right)^2}{\sum_{i \in I_j} h_i + \lambda} + \gamma T. \tag{3.36}$$

O valor obtido pela Equação (3.36) é usada para medir a qualidade da árvore de regressão f_t. Este valor funciona como o índice Gini usado para construir árvores de decisão.

Sabendo que é normalmente inviável avaliar todas as possíveis árvores de regressão, um algorítmo guloso é usado começando em apenas uma folha e iterativamente, adicionando novos *splits* à árvore. Assumindo que I_L e I_R são o conjuntos com amostras

contidas nos nós da esquerda e da direita após o *split* tal que $I = I_L \cup I_R$, tem-se que a redução na função de perda é dada pela seguinte equação

$$\mathcal{L}_{split} = \frac{1}{2}\left[\frac{\left(\sum_{i\in I_L} g_i\right)^2}{\sum_{i\in I_L} h_i + \lambda} + \frac{\left(\sum_{i\in I_R} g_i\right)^2}{\sum_{i\in I_R} h_i + \lambda} - \frac{\left(\sum_{i\in I} g_i\right)^2}{\sum_{i\in I} h_i + \lambda}\right] - \gamma. \tag{3.37}$$

A fórmula mostrada na Equação (3.37) é usada para avaliar os candidatos a *split* de um nó da árvore de regressão. Para mais detalhes da construção da estrutura da árvore, o leitor pode consultar (CHEN; GUESTRIN, 2016).

Além da função regularizadora Ω, o *XGBoost* utiliza outras técnicas adicionais para prevenir sobreajuste como a inserção do parâmetro de *encolhimento* ν (do inglês, *shrinkage*) introduzido inicialmente por Friedman em (FRIEDMAN, 2002) (vide Seção 3.3.5 e 3.3.7) e a seleção de atributos para a construção das árvores de regressão, esta técnica já é utilizada no algoritmo de Floresta Aleatória (vide Seção 3.3.6).

Busca em grade para seleção de hiperparâmetros

Os hiperparâmetros do modelo *XGBoost* foram ajustados via busca em grade ou *grid search*. E, para cada combinação desses parâmetros, a validação cruzada foi feita usando $5 - Folds$. Os hiperparâmetros selecionados para o ajuste foram:

- subsample (subamostragem): proporção das amostras de treinamento. Configurá-lo como 0.5 significa que o XGBoost coletaria aleatoriamente metade das amostras de treinamento antes de montar árvores e isso evitaria sobreajustes (*overfitting*). A subamostragem ocorrerá uma vez em cada iteração, ié, a cada nova árvore inserida no treinamento. O valor obtido para este hiperparâmetro pela análise de *grid search* foi 1, ou seja, 100% das amostras do treino foram utilizadas no treinamento. Os valores testados foram: $0,5; 0,8; 1$.

- colsample_bytree: é a taxa de subamostra de atributos ao construir cada árvore. O valor obtido para este hiperparâmetro pela análise de *grid search* foi 0.5, ou seja, 50% dos atributos foram considerados na contrução de cada árvore, lembrando que a subamostragem ocorre uma vez para cada árvore construída. Os valores testados foram: $0,5; 0,8; 0,9; 1$.

- max_depth: Profundidade máxima de uma árvore. Aumentar esse valor tornará o modelo mais complexo e mais propenso a superajustar (*overfitting*). O valor obtido para este hiperparâmetro pela análise de *grid search* foi 6. Os valores testados foram: $3; 6; 9$.

- eta (taxa de *encolhimento* ou aprendizagem ν): é o peso atribuído a cada árvore do processo de forma a tornar o aprendizado mais conservador. O valor obtido para este hiperparâmetro pela análise de *grid search* foi 0.1. Os valores testados foram: $0,1; 0,4; 0,8; 0,9$.

- n_rounds: este hiperparâmetro controla a quantidade máxima de árvores no treinamento. Os valores testados foram: $100; 500; 1000$.

Os demais hiperparâmetros de treinamento considerados foram os valores padrão da biblioteca implementada em R (CRAN, 2020) (XGBOOST DEVELOPERS, 2020).

3.4 Interpretação dos Resultados

3.4.1 *Local Interpretable Model-Agnostic Explanations (LIME)*

A aprendizagem de máquina está no cerne de muitas inovações tecnológicas nos dias de hoje o que torna cada vez mais necessário modelos que expliquem as predições dos mais diversos algoritmos. Ao usar um algoritmo de aprendizagem de máquina para diagnóstico médico ou detecção de terrorismo, por exemplo, as predições não podem ser postas em prática sob uma fé cega, pois as conseqüências podem ser catastróficas (RIBEIRO; SINGH; GUESTRIN, 2016). O objetivo primordial de um modelo preditivo é encontrar bons resultados medidos através da acurácia, no entanto, um segundo objetivo deve ser interpretar o modelo e entender como ele funciona. Com os modelos se tornando cada vez mais complexos, eles tambem se tornam cada vez mais difíceis de serem interpretados (KUHN; JOHNSON, 2013).

O *LIME* é um algoritmo que foi introduzido por Ribeiro, Singh e Guestrin (RIBEIRO; SINGH; GUESTRIN, 2016) que pode explicar as predições de qualquer classificador ou regressor através de aproximações locais e usando um modelo interpretável . Este método é *model-agnostic*, isto é, ele pode ser aplicado nos mais diversos tipos de modelos "caixa-preta", isto é, modelos com baixa interpretabilidade, tais como, floresta aleatória, redes neurais, etc. Tais modelos são aplicados nos mais variados modelos como, por exemplo, classificação de textos complexos e de imagens.

Outro ponto relevante é que o *LIME* é capaz de explicar predições individualmente, o que para as aplicações médicas é especialmente relevante. Na Fig. 8 o exemplo ilustra o caso de um paciente que foi diagnosticado com GRIPE pelo modelo, neste caso o médico pode avaliar o que foi considerado pelo classificador para tomar a decisão. Isto pode ajudar o médico a refazer a avaliação se o mesmo julgar necessário.

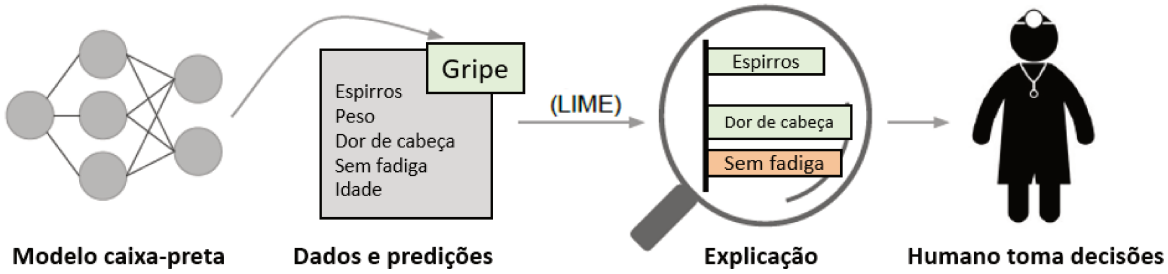

Figura 8 – Explicando predições individuais. Neste exemplo o modelo classifica o paciente com GRIPE e o *lime* destaca os sintomas que contribuem para a classificação GRIPE (em verde: espirros e dor de cabeça) e os que não contribuem (em vermelho: sem fadiga). Adaptado de (RIBEIRO; SINGH; GUESTRIN, 2016).

Com frequência, quando um classificador é considerado não confiável é feito o que se pode chamar de engenharia de atributos (do inglês, *feature engineering*). Isto pode significar desenvolvimentos como redução de dimensionalidade (FILHO, 2015) que são feitos, em geral, por especialistas em aprendizagem de máquina.

Explicações podem ajudar no processo e possibilitar que não especialistas em aprendizado de máquina possam selecionar atributos não relevantes e melhorar modelos. Foram realizados experimentos como mostrado na Figura 9 envolvendo indivíduos não especialistas em aprendizado de máquina e diversos classificadores, a conclusão foi que mesmo indivíduos não especialistas, depois de terem acesso as explicações, tipicamente escolheram classificadores que melhor generalizam os dados do mundo real.

Na Figura 9 a explicação sugere um mau classificador, pois, o critério que está senso utilizado para definir se há um *husky* ou um lobo na imagem é a predominantemente a neve e não as características do próprio animal (RIBEIRO; SINGH; GUESTRIN, 2016). Após as explicações serem apresentadas aos usuários a quantidade de indivíduos que confiavam no modelo ruim foi de 37% para 11%.

(a) Husky classificado como lobo (b) Explicação

Figura 9 – Exemplo de explicação gerada pelo *lime*. Fonte: (RIBEIRO; SINGH; GUESTRIN, 2016).

Um critério essencial para a explicação é que ela precisa ser interpretável, ou seja, prover informações quantitativas que auxiliem no entendimento de como os atributos geraram a resposta do classificador. Para que a explicação seja confiável é preciso que a mesma seja localmente fiel, isto é, a explicação precisa corresponder a predição do modelo na vizinhança da amostra que está sendo predita e com esta medição pode-se avaliar a qualidade da explicação.

Confiabilidade local não implica confiabilidade global (vide Figura 10), ou seja, atributos importantes localmente podem não ser importantes globalmente. Encontrar explicações globalmente confiáveis para modelos complexos (chamados "caixa-preta") permanece um desafio. Na Figura 10 o fundo azul e rosa representa o modelo "caixa-preta" e a complexa função de decisão que não pode ser bem aproximada por um modelo linear. A cruz vermelha em destaque é a amostra que está sendo explicada, o *LIME* gera perturbações na amostra e faz predições usando função de decisão complexa. O peso destas perturbações é dado pela proximidade da amostra que está sendo explicada. A linha pontilhada é a explicação aprendida. Neste caso, um modelo linear ajustado localmente em torno da amostra de interesse.

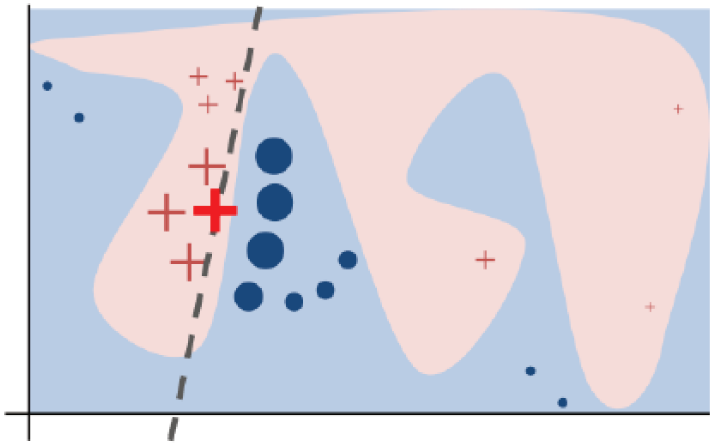

Figura 10 – Exemplo ilustrativo para apresentar intuitivamente o conceito do *LIME*. Fonte: (RIBEIRO; SINGH; GUESTRIN, 2016).

3.5 *Software* e computador utilizados

Os classificadores foram desenvolvidos no ambiente RStudio (3.6). Utilizou-se um computador processador Intel(R) Core(TM) i7-5500U CPU @ 2.40 GHz/8 Gb (RAM), Sistema Operacional Windows10 Pro x64.

4 Resultados das classificações para recomendação em UTI de gestantes de alto risco

4.1 Análise exploratória de dados

O primeiro passo para qualquer projeto em ciência de dados é explorar os dados (BRUCE; BRUCE, 2019). A análise exploratória de dados é uma área relativamente nova da estatística. A estatística clássica buscava tirar conclusões de grandes populações com base em pequenas amostras usando complexos procedimentos de inferência (BRUCE; BRUCE, 2019). Em 1962, *John W. Tukey* propôs uma nova disciplina científica chamada *Análise de Dados* que inclui a inferência como um dos seus componentes, mas, não o único (TUKEY, 1977).

Neste trabalho nosso objeto de estudo são as internações em UTI registradas na base da REDE em pacientes que tiveram MMG. Destas pacientes cerca de 22% foram internadas em UTI, tal qual podemos observar na Figura 3. Por se tratar de uma base proveniente de outras pesquisas, boa parte dos dados já estavam previamente tratados. Assim, o objetivo da análise exploratória foi entender melhor o perfil das pacientes e, ainda, remover e/ou tratar variáveis com muitos dados faltantes.

Fazendo a análise bivariada das variáveis sociais que compõem o modelo verifica-se que não há diferenças significativas entre a população com e sem internação em UTI. No entanto, pode-se descrever o perfil da população do estudo mesmo que as distribuições das variáveis sociais sejam muito parecidas tanto para a população internada quanto para a população não internada.

As 9555 mulheres desse estudo têm idades entre 10 e 49 anos, 67% tem menos de 29 anos, 31% se declararam brancas e 24% negras, 68% são analfabetas ou não completaram o ensino fundamental, 69% não fizeram pré-natal, 98% foram atendidas pela rede pública de saúde, 45% se declararam casadas, 45% chegaram sem qualquer encaminhamento, 48% estavam tendo seu primeiro filho, 75% nunca passaram por uma cesárea e 22% já sofreram algum tipo de aborto.

Das variáveis utilizadas no estudo, apenas as variáveis P10 (IMC), p5cat (cor), p6cat (escolaridade), p7cat (estado civil), p15cat (número de gestações), p16cat (número de partos), p17cat (número de abortos), p18cat (números de cesáreas), p22cat (pré-natal), p24cat (idade gestacional na admissão), p27cat (idade gestacional na resolução), P29 (aborto espontâneo/induzido), P30 (aborto seguro/inseguro), P52 (outras condições), P20cat (parto uterino) e p28cat (tipo de parto) possuíam dados faltantes.

As variáveis P29 (aborto espontâneo/induzido), P30 (aborto seguro/inseguro), P52 (outras condições) e P20cat (parto uterino) foram removidas por terem um percentual alto de valores faltantes. Nos demais casos, os dados faltantes foram substituídos pela moda da variável, o uso da moda, comparado à média e mediana, produziu os melhores resultados nos testes computacionais.

Na Figura 11 temos um gráfico de *Boxplot* da variável IMC mostrando que não há diferenças significativas na população internada versus não internada.

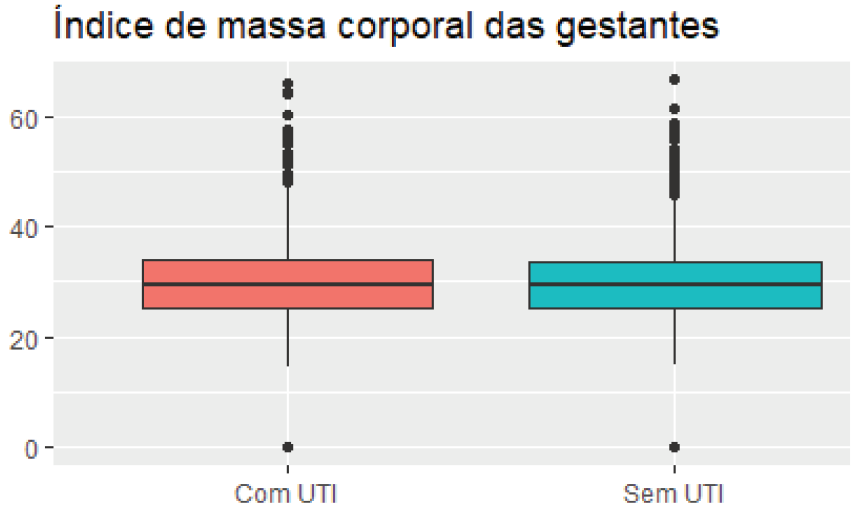

Figura 11 – Gráfico dos valores do índice de massa corporal das pacientes do estudo

A seleção de atributos ou variáveis foi feita considerando critérios clínicos dos especialistas envolvidos no estudo para evitar que houvesse inclusão de variáveis com informações após a internação em UTI e, ainda, foi utilizada a correlação entre as variáveis e com a variável resposta p55_3 (internação em UTI).

A Figura 12 mostra a correlação linear entre as variáveis específicas dos critérios de *near miss* clínico, ou seja, qualquer uma dessas variáveis já classifica a paciente como *near miss*. Quanto mais fina e escura a elipse é, mais positivamente correlacionadas são as variáveis, sendo as variáveis $p57\_1$ (cianose) e $p57\_3$ (FR>40 ou FR<6) as mais correlacionadas entre si, mas não suficientemente significativa. Não há correlação forte positiva ou negativa com a variável resposta $p55\_3$ (internação em UTI).

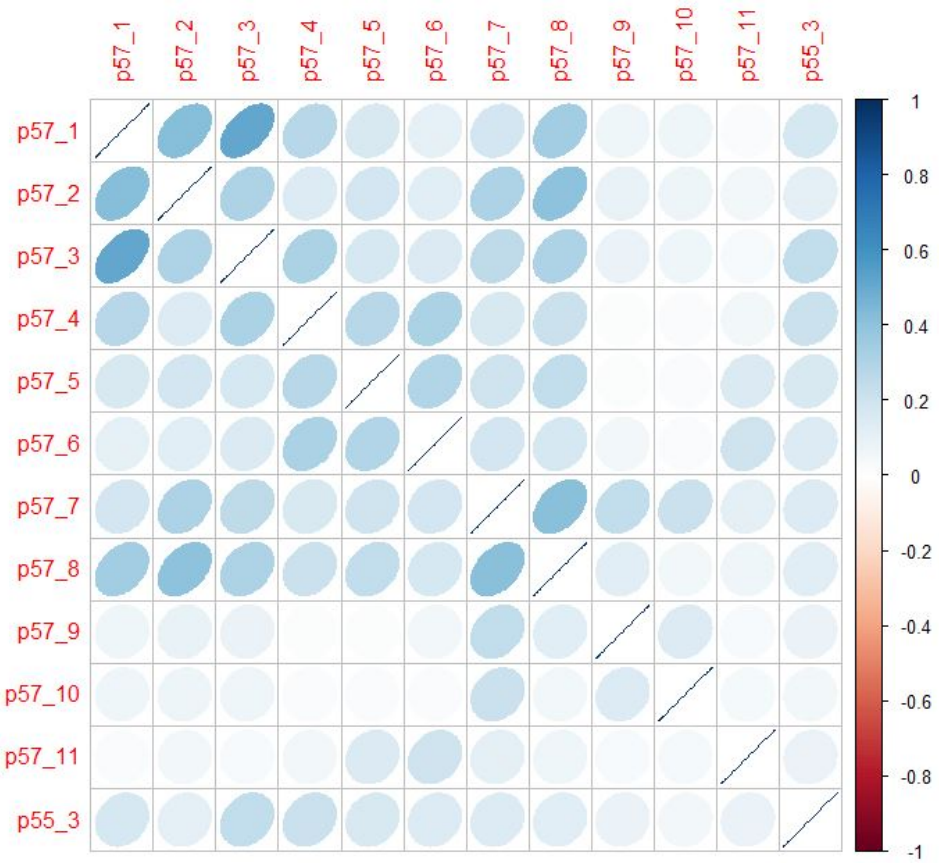

Figura 12 – Correlação entre as variáveis com informações de *near miss* clínicas e a variável resposta com informação de internação em UTI

A Figura 13 mostra a correlação linear entre as variáveis específicas dos critérios de *near miss* laboratorial, ou seja, qualquer uma dessas variáveis já classifica a paciente como *near miss*. Quanto mais fina e escura a elipse é, mais positivamente correlacionadas são as variáveis, sendo as variáveis $p59\_1$ (saturação de oxigênio $< 90\%$ por > 60 minutos) e $p59\_2$ (PaO2 FiO2 < 200) as mais correlacionadas entre si, mas não suficientemente significativas. Não há correlação forte positiva ou negativa com a variável resposta $p55\_3$ (internação em UTI).

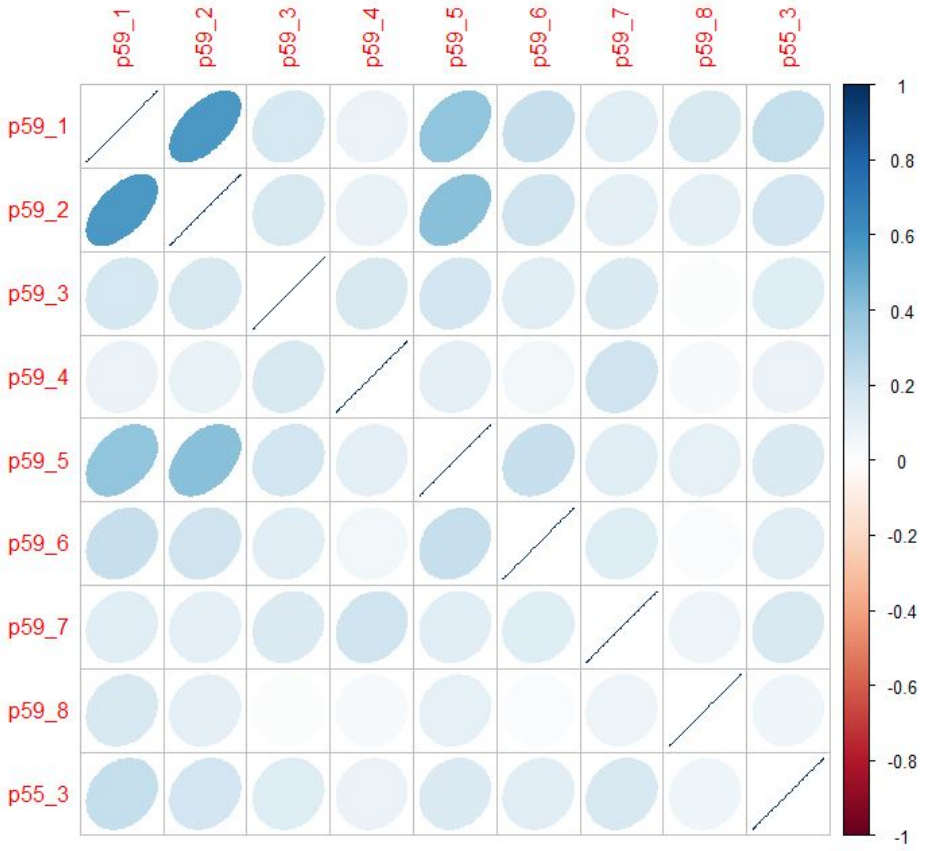

Figura 13 – Correlação entre as variáveis com informações de *near miss* laboratoriais e a variável resposta com informação de internação em UTI

A Figura 14 mostra a correlação linear entre as variáveis específicas dos critérios de *near miss* de manejo, ou seja, qualquer uma dessas variáveis já classifica a paciente como *near miss*. Quanto mais fina e escura a elipse é, mais positivamente correlacionadas são as variáveis, sendo as variáveis $p61\_1$ (uso de droga vasoativa contínua) e $p61\_4$ (intubação e ventilação por tempo >=60 minutos não relacionado a anestesia) as mais correlacionadas entre si, mas não significativamente. Não há correlação forte positiva ou negativa com a variável resposta $p55\_3$ (internação em UTI).

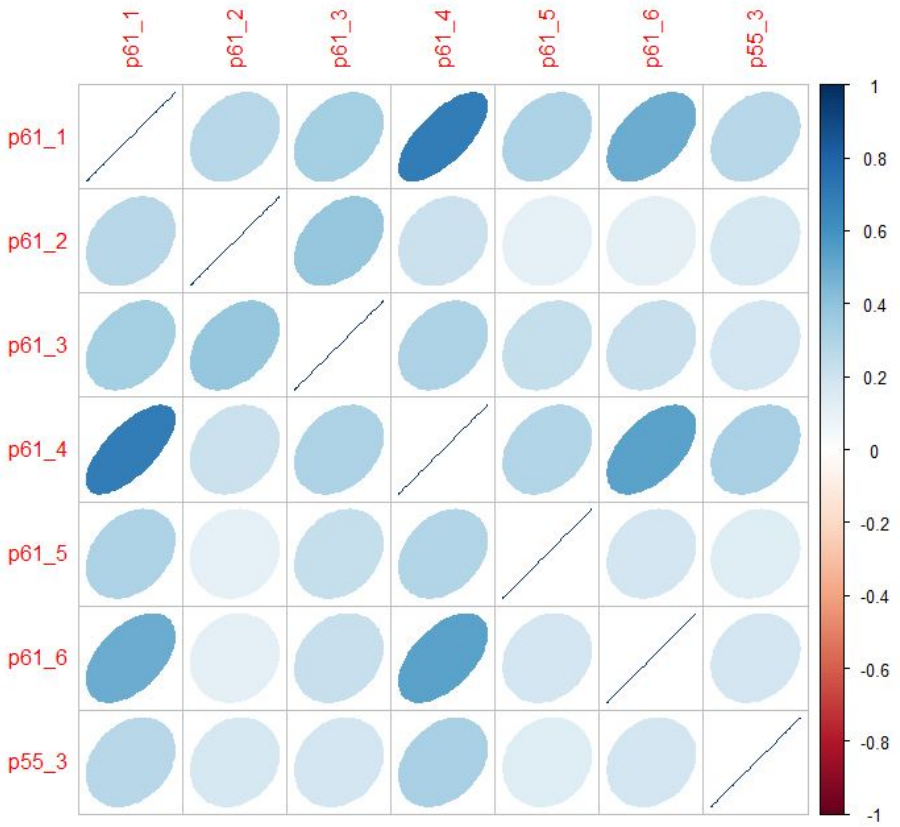

Figura 14 – Correlação entre as variáveis com informações de *near miss* de manejo e a variável resposta com informação de internação em UTI

Nas figuras 15 e 16 há alguns exemplos das distribuições das variáveis categóricas entre internação e não internação em UTI. Não há mudanças significativas no percentual de cada categoria da variável quando analisa-se a internação. Em alguns casos parece haver alguma distinção (como no caso das variáveis cor ou acompanhamento da gestação), no entanto, nesses casos verifica-se um volume pouco representativo ou ausência de dados.

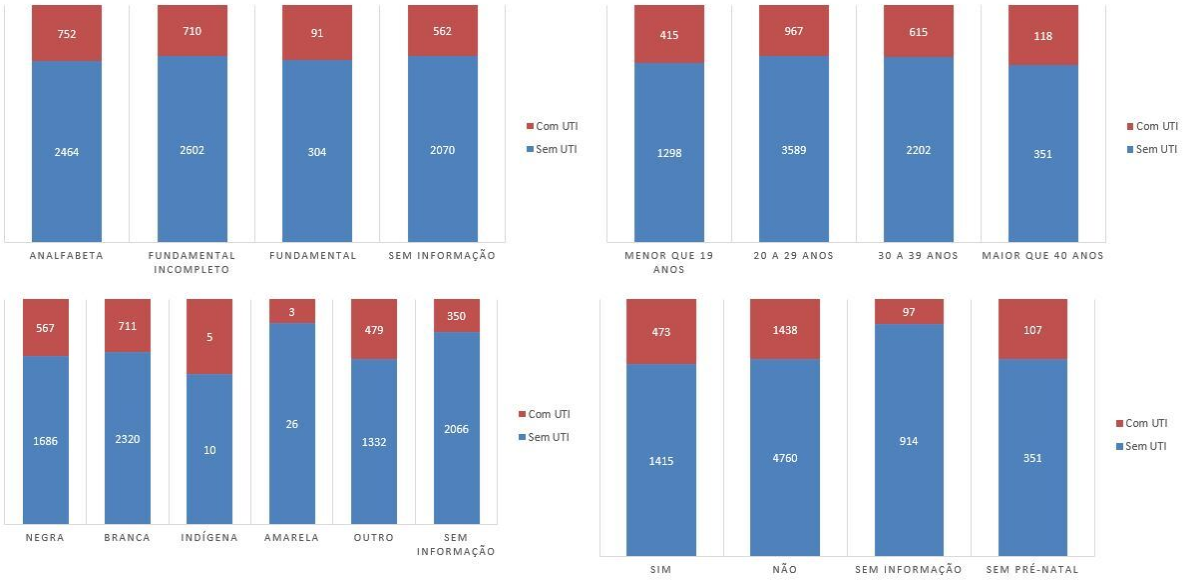

Figura 15 – Exemplos da análise das distribuições entre a variável resposta e as demais variáveis preditoras. Neste caso, as variáveis de cima para baixo e da direita para a esquerda são: escolaridade, idade, cor e acompanhamento na gestação.

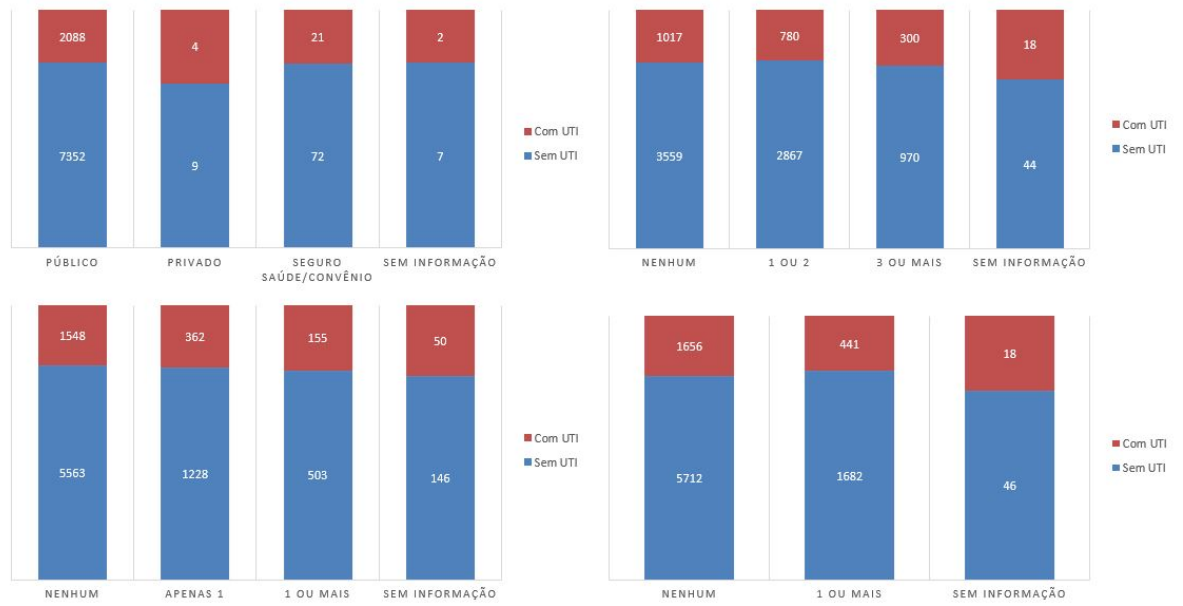

Figura 16 – Exemplos da análise das distribuições entre a variável resposta e as demais variáveis preditoras. Neste caso, as variáveis de cima para baixo e da direita para a esquerda são: cobertura, número de partos, número de cesáreas e número de abortos.

Por fim, com a análise exploratória de dados foi possível tratar os dados faltantes, removendo variáveis com muitos dados faltantes ou usando a moda para substituir estes

dados em algumas amostras, a moda foi escolhida após testes computacionais considerando a média e a mediana, esta técnica auxiliou na performance dos modelos testados. Utilizando a análise bivariada foi possível observar que as distribuições das variáveis preditoras de caráter demográfico em relação à variável resposta, em sua maioria, se comportam da mesma forma tanto para população com internação em UTI quanto para a população sem internação em UTI. A análise de correlação também mostrou que não há correlação direta entre as variáveis preditoras e a variável resposta.

4.2 Resultados e análises

4.2.1 Resultados das classificações

Neste capítulo são apresentados os resultados gerados pelos algoritmos de classificação: Floresta Aleatória, *Gradient Boosting Machine (GBM)* e *XGBoost*. Os dados utilizados foram da base REDE (Rede Brasileira de Vigilância de Morbidade Materna Grave), onde o objetivo dos algoritmos foi classificar as pacientes em duas classes, internação em UTI e não internação em UTI. Outros modelos foram testados como *naive Bayes*, SVM e redes neurais, no entanto, nos testes preliminares os modelos de árvore se mostraram mais promissores.

De modo a garantir a validade das medições geradas, o conjunto de teste foi fixado, ou seja, foram avaliadas as mesmas 1436 mulheres em todas as rodadas, isto é uma prática comum em medicina para garantir a reprodutibilidade e validação dos testes diagnósticos. Os conjuntos de treino e validação foram escolhidos aleatóriamente a cada nova repetição. Os dados restantes foram divididos em 90% treino e 10% validação. Desta forma, pode-se avaliar a consistência dos diagnósticos apresentados.

Cada um dos três modelos gerou 10 matrizes de confusão como resultado das repetições. Na Tabela 1 é apresentado a média e variância dos testes diagnósticos considerando as 10 repetições de cada modelo.

Modelo / Métrica	**Floresta Aleatória**	***GBM***	***XGBoost***
Acurácia (%)	82,69 ± 0,11	82,95 ± 0,14	83,46 ± 0,12
Sensibilidade (%)	37,40 ± 0,85	41,67 ± 0,56	51,61 ± 0,96
Especificidade (%)	95,83 ± 0,07	94,93 ± 0,25	92,70 ± 0,12
F1	0,49 ± 0,00	0,52 ± 0,00	0,58 ± 0,00
Valor Preditivo Positivo (%)	72,26 ± 2,21	70,52 ± 3,97	67,25 ± 1,31
Valor Preditivo Negativo (%)	84,06 ± 0,04	84,87 ± 0,03	86,84 ± 0,06
Razão de Verossimilhança Positiva	9,01 ± 0,46	8,29 ± 0,65	7,09 ± 0,14
Prevalência Estimada (%)	11,64 ± 0,07	13,30 ± 0,21	17,26 ± 0,12

Tabela 1 – Resultados dos testes diagnósticos para os modelos Floresta Aleatória, *GBM* e *XGBoost*.

A acurácia, mostrada na Figura 17, indica o total de acertos do modelo, ou seja, mostra a probabilidade de uma mulher ser corretamente diagnosticada. A escolha do *XGBoost* foi motivada não apenas pela acurácia, mas também pelo ganho na sensibilidade, sendo consistentemente superior aos demais modelos. A sensibilidade é um indicador importante, falsos negativos são mais indesejados que falsos positivos, pois, a não internação implica em maior risco à gestante do que uma internação desnecessária.

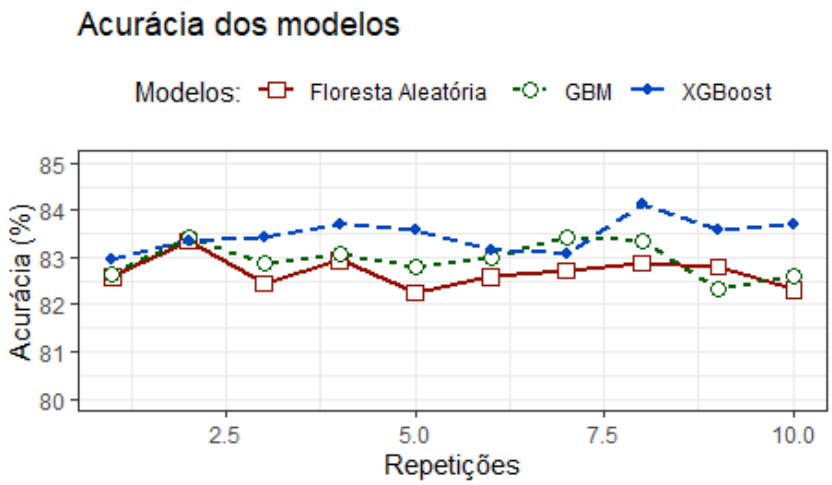

Figura 17 – Gráfico dos valores de acurácia para os quatro modelos utilizados.

A especificidade (Figura 18) é o indicador que mostra a probabilidade de um resultado negativo ser verdadeiramente negativo, a especificidade alta indica poucos falso positivos e que o algoritmo geralmente acerta casos de não internação em UTI. Neste indicador o modelo de Floresta Aleatória se destacou em relação aos outros modelos.

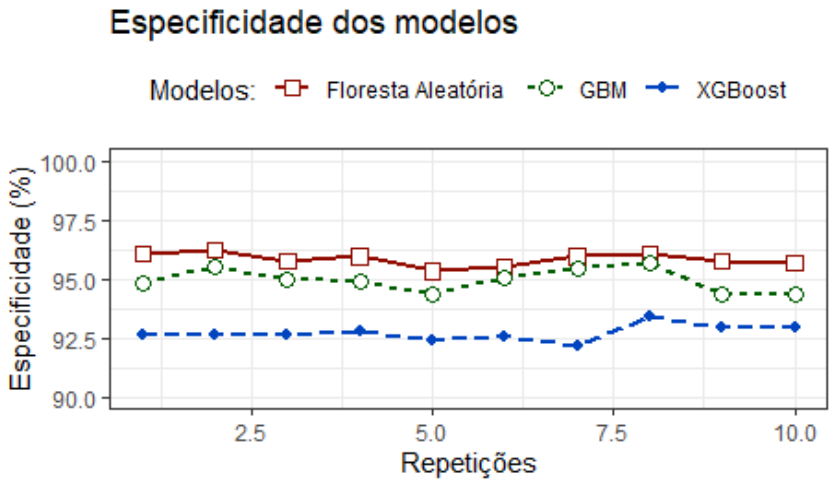

Figura 18 – Gráfico dos valores de especificidade para os quatro modelos utilizados.

A prevalência real de internação em UTI de gestamtes com MMG na base da REDE é de 22% e é constante para todos os modelos em todas as repetições para garantir a reprodutibilidade do teste. Lembrando que este percentual é sobre o dados de teste. Já a prevalência estimada (Figura 19) mostra o percentual de mulheres que os modelos recomendaram internação usando como ponto de corte uma probabilidade de internação estimada pelos modelos maior ou igual a 50% . Lembrando que leitos de UTI são recursos caros e não podem ser considerados disponíveis em abundância. No entanto, nenhum dos modelos gerou resultados superiores ao da prevalência real indicando que o volume de recomendações não seriam um problema. Assim, seria possível realocar os leitos para, por exemplo, atendimentos emergenciais ou transferências com urgência.

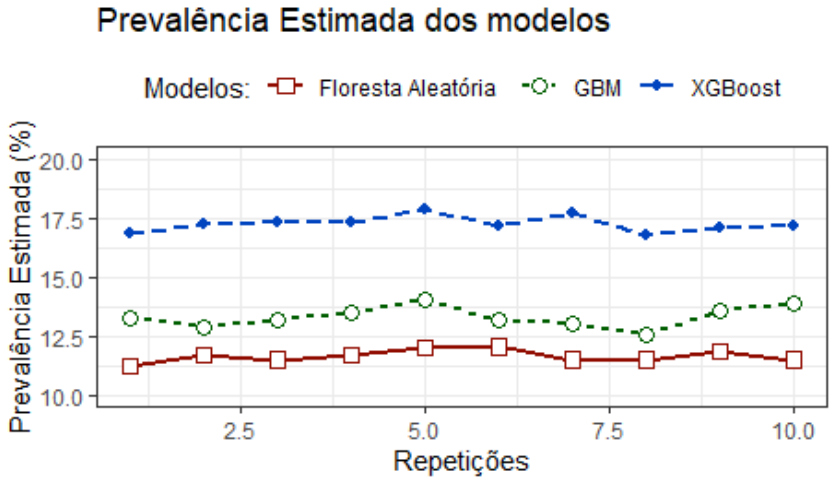

Figura 19 – Gráfico dos valores de prevalência estimada para os quatro modelos utilizados.

A precisão ou valor preditivo positivo (Fig. 20) é a probabilidade do indivíduo ser verdadeiramente positivo quando o teste é positivo. Este teste é especialmente relevante,

pois, quanto maior mais provável da paciente com recomendação de internação ser um caso com real necessidade de internação em UTI. O modelo Floresta Aleatória performou melhor para este indicador.

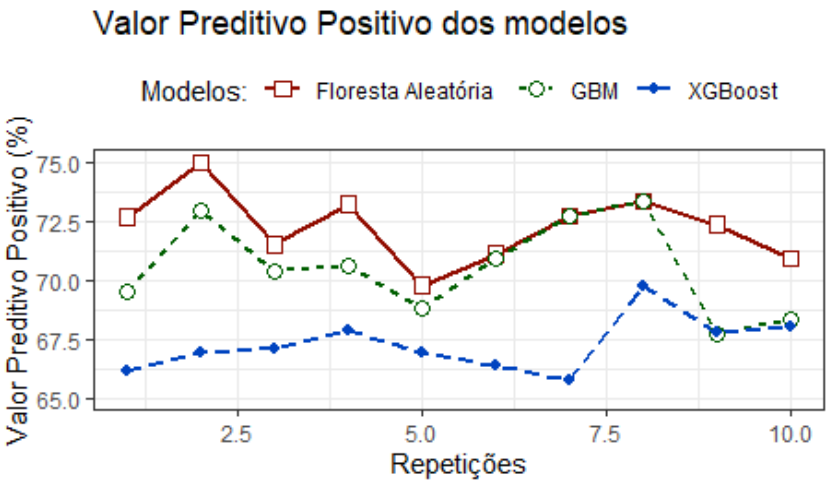

Figura 20 – Gráfico dos valores do valor preditivo positivo (predição) para os quatro modelos utilizados.

O VPN ou valor preditivo negativo (Figura 21) é a probabilidade do indivíduo ser verdadeiramente negativo quando o teste é negativo. O modelo *XGBoost* performou melhor para este indicador.

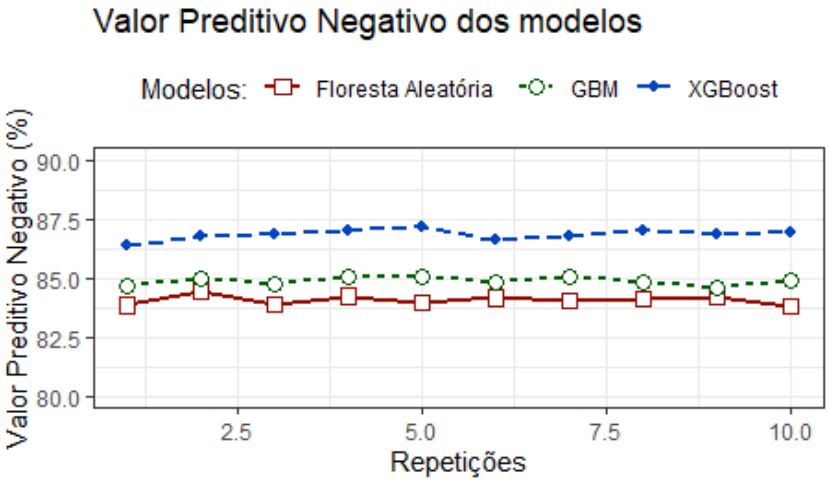

Figura 21 – Gráfico dos valores do valor preditivo negativo para os quatro modelos utilizados.

Para a razão de verossimilhança positiva (Figura 22) o modelo Floresta Aleatória performou melhor. Lembrando que o pior modelo possível teria RVP igual a 1, então quanto maior o indicador, melhor.

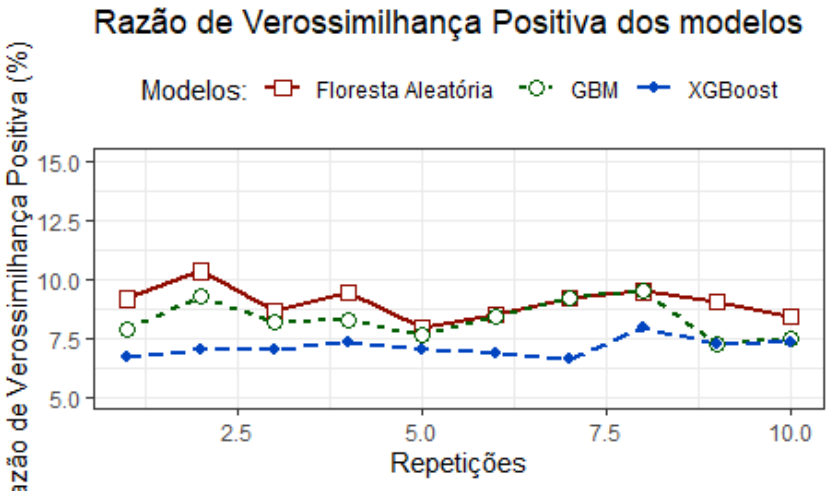

Figura 22 – Gráfico dos valores da razão de verossimilhança positiva para os quatro modelos utilizados.

A sensibilidade (Fig. 23) ou *recall* é o indicador que mostra a probabilidade de um resultado positivo ser verdadeiramente positivo, ou ainda, a taxa de verdadeiro positivo que corresponde a proporção de amostras da classe positiva (onde houve recomendação de classificação) corretamente classificadas.

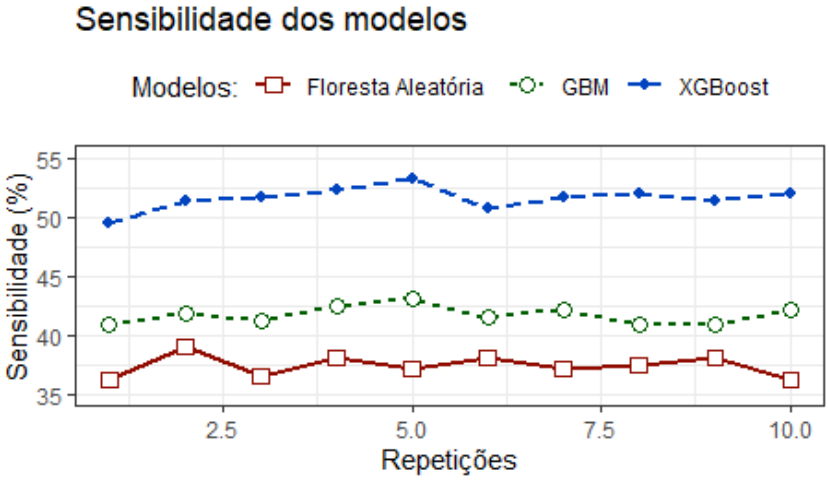

Figura 23 – Gráfico dos valores de sensibilidade (*recall*) para os quatro modelos utilizados.

Usando os valores de *recall* e precisão, foi calculado o valor de F_1 para os modelos (Tabela 1). Neste caso, valores próximos a 1 são esperados (DEMŠAR, 2006), o modelo *XGBoost* obteve o melhor resultado com AUC de 0,867.

Entendendo que falsos negativos podem ser preocupantes, especialmente em estudos na área da saúde, e ainda, avaliando testes estatísticos apresentados na próxima seção, o modelo *XGBoost* é o modelo recomendado para eventuais implementações práticas. Sendo também, por esta razão, o modelo escolhido para a realização das análises a seguir.

A Tabela 24 mostra a distribuição das pacientes do conjunto de teste em 10 faixas por ordem de propensão à UTI onde cada faixa contém 10% das pacientes. As faixas 8, 9 e 10 concentram 80% do total de pacientes com *near miss* e a faixa 10 concentra 100% do total de óbitos o que mostra que o modelo corretamente considera a estratificação da gravidade na recomendação de internação, o que produz melhores resultados, reduzindo a probabilidade de óbito (SOARES et al., 2016).

Vale ressaltar que, a base da REDE utilizada neste estudo contém pacientes consideradas de risco que foram classificadas em 3 grupos: CPAV, *near miss* e óbito. O modelo não contém a informação de óbito e/ou *near miss* como atributo, pois, estas informações não poderiam ser obtidas *a priori* no caso de uma internação real.

Faixas	Gestantes	% Gestantes	UTI	% UTI	Near Miss	% Near Miss	Óbito	% Óbito
1	143	10%	0	0%	0	0%	0	0%
2	143	10%	1	0%	2	2%	0	0%
3	143	10%	8	2%	5	4%	0	0%
4	143	10%	9	3%	1	1%	0	0%
5	144	10%	17	5%	4	3%	0	0%
6	144	10%	26	8%	4	3%	0	0%
7	144	10%	31	10%	9	8%	0	0%
8	144	10%	51	16%	11	9%	0	0%
9	144	10%	64	20%	27	23%	0	0%
10	144	10%	116	36%	54	46%	21	100%

Figura 24 – Faixas geradas a partir dos *scores* das pacientes internadas em UTI de modo que cada faixa concentre 10% da quantidade total de pacientes do teste. A faixa 1 é a faixa com menor probabilidade de internação e a faixa 10 é a faixa com maior probabilidade de internação. Esta tabela foi gerada a partir dos dados do grupo de teste.

Considerando a baixa variância entre as rodadas, podemos considerar que os resultados do modelo *XGBoost* são consistentes, corroborando que o modelo faz predições que consideram a gravidade da paciente. Considerar a gravidade é a maneira onde comprovadamente a internação em UTI é efetiva para evitar a morte materna (SOARES et al., 2016).

4.2.2 Teste de Wilcoxon e Kolmogorov-Smirnov

O teste dos postos sinalizados de Wilcoxon (WILCOXON, 1992) é uma alternativa não paramétrica ao teste t pareado, que classifica as diferenças de desempenho de dois classificadores para cada conjunto de dados (DEMŠAR, 2006). Do ponto de vista estatístico, o teste de Wilcoxon é mais seguro, pois não assume distribuições normais. Além disso, os outliers (desempenho excepcionalmente bom / ruim em alguns conjuntos de dados) têm menos efeito no Wilcoxon do que no teste t. Para realização deste tipo de teste é necessário que as amostras estejam pareadas, ou seja, o mesmo conjunto de teste

seja aplicado em todos os classificadores, assim, no nosso contexto, o teste irá avaliar os *scores* dos diferentes classificadores nas mesmas pacientes.

Com este teste é possível comparar classificadores em combinações de dois a dois, Neste estudo temos 3 classificadores: *XGBoost*, Floresta Aleatória e *GBM (Gradient Boosting Machine)*. Para cada combinação entre dois classificadores, tem-se a hipótese nula de que a mediana das diferenças entre os classificadores é igual a zero, e ainda, a hipótese alternativa implica que a mediana das diferenças entre os classificadores é diferente de zero. Para aceitar a hipótese nula o p-valor do teste precisa ser maior que $0,05$, caso contrário, a hipótese aceita é a alternativa.

A comparação entre os pares de classificadores encontra-se na Tabela 2. A partir deste resultados, pode-se rejeitar a hipótese nula na comparação entre os classificadores *GBM* e *XGBoost* e, ainda, entre os classificadores Floresta Aleatória e *XGBoost*, ou seja, a mediana das diferenças é diferente de zero e, portanto, há diferenças significativas entre estes classificadores. Já em relação aos classificadores *GBM* e Floresta Aleatória como o p-valor é maior que $0,05$ pode-se assumir que a mediana das diferenças é igual a zero e que, portanto, não há diferenças significativas entre estes classificadores. Sendo assim, por comparação, pode-se concluir que o classificador *XGBoost* é significativamente diferente aos demais, e ainda, usando os resultados dos testes diagnósticos (Tabela 1), pode-se inferir que o *XGBoost* é o melhor classificador para este problema dentre os testados.

Classificadores	**p-valor**
GBM e XGBoost	6.576e-16
GBM e Floresta Aleatória	0.3462
Floresta Aleatória e XGBoost	1.745e-14

Tabela 2 – Resultados dos testes de *Wilcoxon* para os modelos Floresta Aleatória, *GBM* e *XGBoost*.

O teste de Kolmogorov–Smirnov também é um teste não paramétrico e tem por objetivo de comparar a distância entre duas distribuições acumuladas (SICSÚ, 2010). Podendo ser utilizado para a comparação da separação das distribuições de um evento binário, no caso deste trabalho pacientes não internadas em UTI = 0 e pacientes internadas em UTI = 1, sendo cada uma das informações uma distribuição acumulada (FDA), ordenadas pela probabilidade ou *score*. No caso do teste de Kolmogorov–Smirnov (ou apenas, teste de K-S), quanto maior for o resultado melhor, pois, isto indica que a separação entre os eventos é maior, indicando também uma maior acurácia.

Os três modelos testados obtiveram um K-S em torno de 22%. Segundo a literatura (SICSÚ, 2010), um K-S emtre 20% e 25% é considerado baixo, porém aceitável.

4.2.3 Importância dos Atributos

O *XGBoost* pode medir o quão boa é a estrutura de uma árvore conforme Equação (3.37) (vide Seção 3.3.8). O ideal seria enumerar todas as possíveis árvores e seus respectivos ganhos, no entanto, isto é computacionalmente impraticável. Na prática, este ganho é calculado em cada nível da árvore em questão. A importância mostrada na Figura 25 reflete o valor deste ganho refletido na função objetivo por cada um dos atributos.

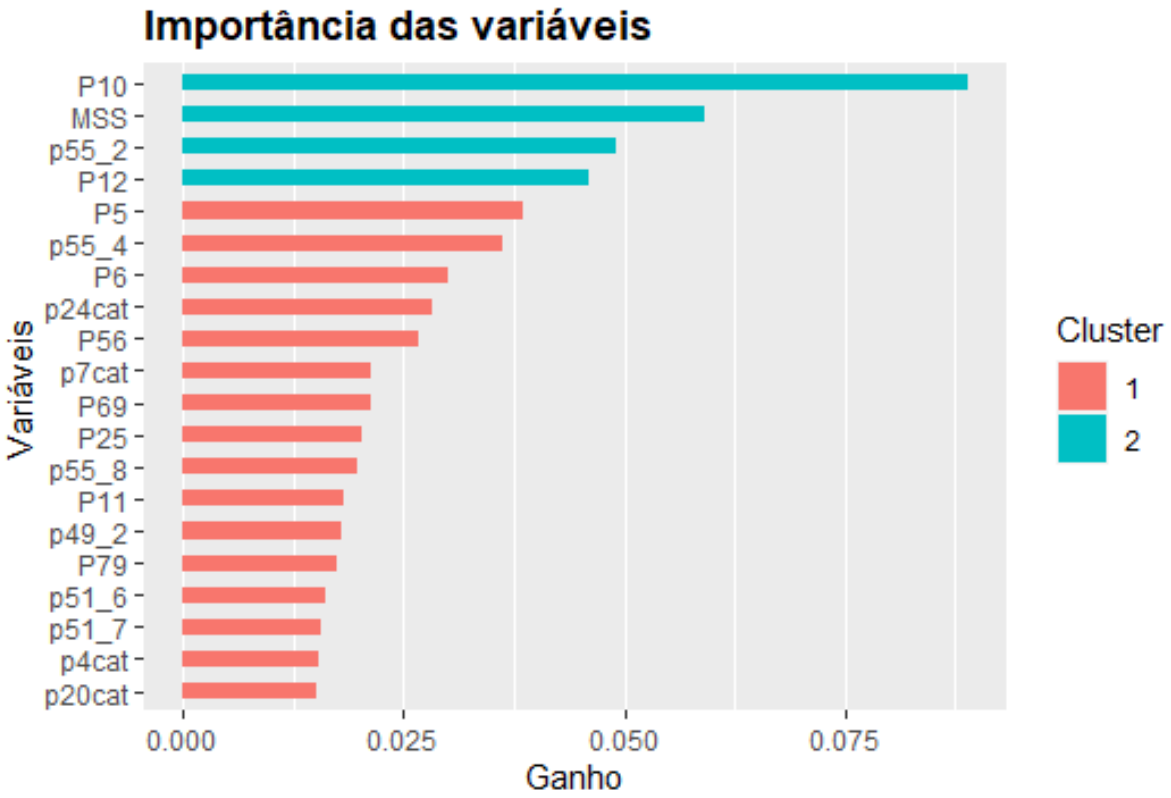

Figura 25 – Gráfico da importância dos atributos (*features*) relativo ao ganho para o modelo *XGBoost*.

Apesar de uma visão de quais atributos produzem ganhos de maneira geral, percebe-se que esta abordagem não gera saídas interpretáveis que ajudem a explicar o modelo globalmente ou para entradas específicas. Por exemplo, sabe-se que o atributo com o maior ganho para o modelo, ou seja, que mais ajuda a minimizar a função de perda é o P10 (IMC), mas não se sabe exatamente quais valores de IMC e quais relacionamentos com outros atributos levam a uma classificação de recomendação ou não de UTI.

Afim de buscar explicações mais consistentes do modelo usamos o algoritmo *LIME*.

4.2.4 Resultados das interpretações

Conforme demonstrado por (RIBEIRO; SINGH; GUESTRIN, 2016), pode-se gerar explicações de modelos caixa-preta a partir de aproximações locais usando funções de baixa complexidade e facilmente interpretáveis como árvores de decisão e modelos lineares.

A Figura 26 foi a saída gerada pelo algoritmo LIME para explicar explicar 3 casos de internação em UTI e 3 casos de não internação em UTI. A interpretação deve ser a mais simples possível e, neste caso, a explicação foi obtiva através de uma árvore de decisão. Em vermelho estão os valores das variáveis que reduzem a probabilidade de internação em UTI e em verde, os valores aumentam a probabilidade de internação. Lembrando que as explicações são locais, ou seja, podem explicar o caso, mas, não o modelo de forma geral.

Na Figura 26 a informação *Case* significa de fato a identificação de uma paciente em específico, o campo *Label* é a informação real se houve (*Label*=1) ou não (*Label*=0) a internação em UTI. O campo *Probability* indica o resultado da predição obtida pelo modelo caixa-preta que é, no nosso caso, o *XGBoost*. O campo *Explanation Fit* mostra o quão bem o modelo caixa-preta é explicado localmente (UC BUSINESS ANALYTICS, 2020).

Olhando para o *Case 1* da Figura 26 pode-se concluir que, segundo a avaliação do *LIME*, o modelo *XGBoost* aumentou a probabilidade de internação principalmente devido as variáveis: P10 (IMC), P13 (rede de atendimento), p27cat (idade gestacional na resolução por parto), P25 (tipo de trabalho de parto), P11 (acompanhamento da gestação), p43_1 (hipertensão arterial crônica), p28cat (tipo de parto), p4cat (idade), p15cat (número de gestações) e P12(internação espontânea/transferência). Além disso, por se tratar de uma explicação usando apenas uma árvore de decisão, pode-se indicar inclusive quais os dados em cada uma das variáveis foram responsáveis pelo aumento (*supports*) ou redução (*contradicts*) da probabilidade de internação.

Neste caso temos que a paciente *Case 1* foi recomendada para a UTI por ter IMC menor que 25.4, na sua informação de trabalho de parto consta aborto, permanece grávida ou sem informação. Além disso, esta paciente também tem hipertensão arterial crônica e a resolução do parto indica aborto, prenhez ectópica, permanece grávida ou não há informação no prontuário. As informações desta paciente que reduzem a sua probabilidade de ir para a UTI são a rede de atendimento, a idade gestacional na resolução é maior que 37 semanas, a paciente fez o pré-natal, tem menos de 29 anos, tem menos de 3 gestações e veio por encaminhamento interno ou de outra instituição.

A explicação dos demais casos segue de maneira análoga. Vale ressaltar que, apesar do LIME fornecer uma explicação, não necessariamente esta explicação representa uma explicação confiável do modelo caixa-preta que está sendo utilizado. Assim, cabe ao profissional de saúde que esteja avaliando o caso dar o parecer final.

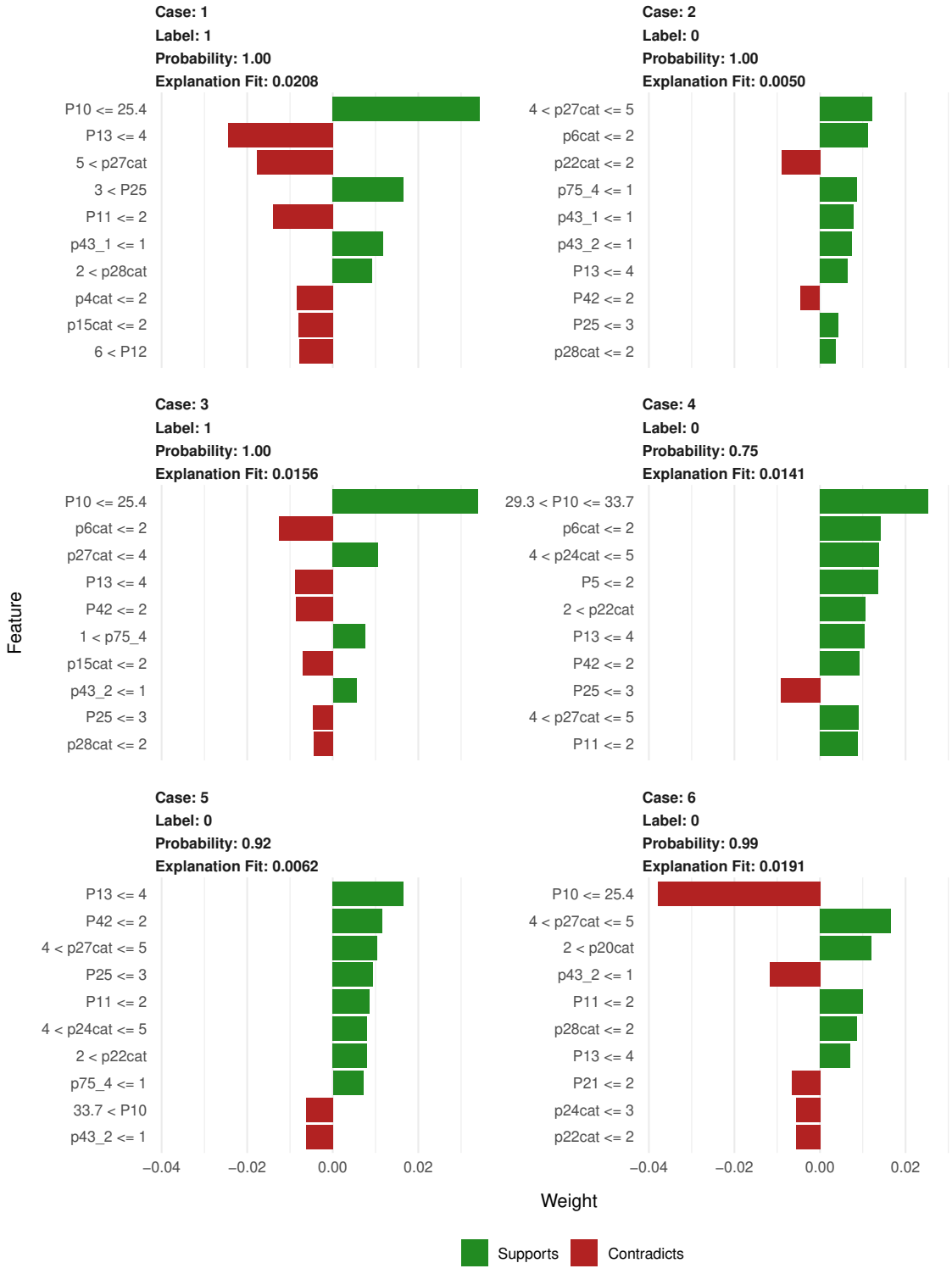

Figura 26 – Gráfico dos valores do peso das variáveis nas explicações de casos específicos do modelo *XGBoost* para os rótulos de classe 0 (não recomenda internação em UTI) e 1 (recomenda internação em UTI).

5 Conclusões

5.1 Considerações finais

Este estudo abordou o problema de recomendação de internação em UTI de gestantes com MMG sob a perspectiva do aprendizado de máquina. O algoritmo de aprendizagem supervisionada escolhido foi o *XGBoost* com parâmetros selecionados via *grid search*, obtendo valores de acurácia consistentemente em torno de 84% e sensibilidade ou *recall* em torno de 51%. Sendo a sensibilidade um teste diagnóstico muito relevante devido a gravidade de falsos negativos. A fixação do conjunto de teste e a baixa variância reforça a reprodutibilidade e a robustez que é outro aspecto importante ao se diagnosticar e predizer condições clínicas. Assim, como o as mesmas medições foram feitas para as mesmas 1436 mulheres com MMG em todos os modelos e todas as repetições, pode-se concluir que o modelo é consistente e pode ser usado para realizar testes diagnósticos, preferencialmente, com a supervisão de profissionais de saúde qualificados.

Distribuindo as pacientes do conjunto de teste em 10 faixas por ordem de propensão à UTI onde cada faixa contém 10% das pacientes, tem-se que as faixas 8, 9 e 10 concentram 80% do total de pacientes com *near miss* e a faixa 10 concentra 100% do total de óbitos o que mostra que o modelo corretamente considera a estratificação da gravidade. O tratamento em UTI possui maior eficácia em pacientes mais graves como já foi demonstrado em (SOARES et al., 2016).

5.2 Sugestões para trabalhos futuros

Usando uma abordagem simular a que foi apresentada nesta dissertação, outros modelos também podem ser gerados a partir da base da REDE. Por exemplo, os seguintes modelos poderiam ser investigados:

- Modelo 1: Predição de óbito materno;
- Modelo 2: Predição de óbito materno e *near miss*;
- Modelo 3: Recomendação de internação em UTI com o modelo 1 como entrada;
- Modelo 4: Recomendação de internação em UTI apenas com atributos laboratoriais;
- Modelo 5: Recomendação de internação em UTI apenas com atributos clínicos;
- Modelo 6: Recomendação de internação em UTI apenas com atributos sociais e de gestão.

A seleção de atributos também deve ser considerada devido ao grande número de atributos considerados neste estudo. Esta técnica poderia, entre outras coisas, melhorar a interpretabilidade do modelo preservando a acurácia. Além disso, é necessário investigar o modelo obtido em outras bases de dados do Brasil para verificar se o modelo é, de fato, capaz de generalizar e obter bons resultados na prática, validando os resultados com mais especialistas de área da saúde após uma eventual implementação.

Por fim, a base de REDE é uma fonte muito rica de informações para o desenvolvimento de modelos que auxiliem em políticas de atenção a saúde da mulher. Os primeiros passos ainda estão sendo dados para que estas informações gerem resultados palpáveis para a sociedade, mas o caminho já começou a ser trilhado.

Referências

BEAGLEHOLE, R.; BONITA, R.; KJELLSTRÖM, T. et al. *Basic epidemiology*. [S.l.]: World Health Organization Geneva, 1993. Citado 3 vezes nas páginas 10, 23 e 24.

BOCCATO, L.; ATTUX, R. *Tópicos em Sistemas Inteligentes II - Aprendizado de Máquina*. [S.l.]: DCA/FEEC/UNICAMP, 2019. Citado na página 16.

BREIMAN, L. . Random forests. *Machine learning*, Springer, v. 45, n. 1, p. 5–32, 2001. Citado 2 vezes nas páginas 29 e 36.

BREIMAN, L. Bagging predictors. *Machine learning*, Springer, v. 24, n. 2, p. 123–140, 1996. Citado 3 vezes nas páginas 31, 32 e 35.

BREIMAN, L.; FRIEDMAN, J. H.; OLSHEN, R. A.; STONE, C. J. Classification and regression trees. *Wadsworth International Group*, v. 8, p. 452–456, 1984. Citado 6 vezes nas páginas 10, 27, 29, 30, 37 e 39.

BRUCE, A.; BRUCE, P. *Estatística Prática para Cientistas de Dados*. [S.l.]: Alta Books, 2019. Citado na página 46.

CAO, D.-S.; HUANG, J.-H.; LIANG, Y.-Z.; XU, Q.-S.; ZHANG, L.-X. Tree-based ensemble methods and their applications in analytical chemistry. *TrAC Trends in Analytical Chemistry*, Elsevier, v. 40, p. 158–167, 2012. Citado na página 29.

CECATTI, J. G.; SOUZA, J. P.; PARPINELLI, M. A.; HADDAD, S. M.; CAMARGO, R. S.; PACAGNELLA, R. C.; SILVEIRA, C.; ZANARDI, D. T.; COSTA, M. L.; SILVA, J. L. P. e et al. Brazilian network for the surveillance of maternal potentially life threatening morbidity and maternal near-miss and a multidimensional evaluation of their long term consequences. *Reproductive Health*, BioMed Central, v. 6, n. 1, p. 15, 2009. Citado 2 vezes nas páginas 19 e 21.

CHEN, T.; GUESTRIN, C. Xgboost: A scalable tree boosting system. In: ACM. *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. [S.l.], 2016. p. 785–794. Citado 3 vezes nas páginas 38, 39 e 41.

CRAN. *Extreme Gradient Boosting*. 2020. Disponível em: <https://cran.r-project.org/web/packages/xgboost/xgboost.pdf>. Acesso em: 02 nov. 2020. Citado na página 42.

DEMŠAR, J. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine learning research*, v. 7, n. Jan, p. 1–30, 2006. Citado 2 vezes nas páginas 56 e 57.

ESMI, E.; VALLE, M. E. Uma visão geral de algumas abordagens para processamento de dados. *Biomatemática IMECC-UNICAMP*, v. 24, p. 91–108, 2014. Citado na página 23.

FILHO, A. D. P. C. Uso de big data em saúde no brasil: perspectivas para um futuro próximo. *Epidemiologia e Serviços de Saúde*, SciELO Public Health, v. 24, p. 325–332, 2015. Citado 2 vezes nas páginas 17 e 43.

FREUND, Y.; SCHAPIRE, R.; ABE, N. A short introduction to boosting. *Journal-Japanese Society For Artificial Intelligence*, JAPANESE SOC ARTIFICIAL INTELL, v. 14, n. 771-780, p. 1612, 1999. Citado na página 32.

FREUND, Y.; SCHAPIRE, R. E. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, Elsevier, v. 55, n. 1, p. 119–139, 1997. Citado na página 32.

FRIEDMAN, J. Stochastic gradient boosting. *Computational statistics and data analysis*, v. 38, n. 4, p. 367–378, 2002. Citado 6 vezes nas páginas 32, 33, 35, 36, 39 e 41.

FRIEDMAN, J.; HASTIE, T.; TIBSHIRANI, R. et al. Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). *The annals of statistics*, Institute of Mathematical Statistics, v. 28, n. 2, p. 337–407, 2000. Citado na página 39.

FRIEDMAN, J. H. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, JSTOR, p. 1189–1232, 2001. Citado 2 vezes nas páginas 36 e 38.

JAMESON, J. L.; LONGO, D. L. Precision medicine—personalized, problematic, and promising. *Obstetrical & Gynecological Survey*, LWW, v. 70, n. 10, p. 612–614, 2015. Citado na página 17.

JORDAN, M. I.; MITCHELL, T. M. Machine learning: Trends, perspectives, and prospects. *Science*, American Association for the Advancement of Science, v. 349, n. 6245, p. 255–260, 2015. Citado na página 16.

KEARNS, M. Cryptographic limitations on learning boolean formulae and finite automata. *Journal of the ACM (JACM)*, ACM New York, NY, USA, v. 41, n. 1, p. 67–95, 1994. Citado na página 32.

KEARNS, M.; VALIANT, L. Learning boolean formulae or finite automata is as hard as factoring. harvard university. *Center for Research in Computing Technology, Aiken Computation Laboratory*, 1988. Citado na página 32.

KEARNS, M. J.; VAZIRANI, U. V.; VAZIRANI, U. *An introduction to computational learning theory*. [S.l.]: MIT press, 1994. Citado na página 32.

KOTSIANTIS, S. Supervised machine learning: A review of classification techniques. *Informatica (03505596)*, v. 31, n. 3, 2007. Citado na página 23.

KUHN, M.; JOHNSON, K. *Applied predictive modeling*. [S.l.]: Springer, 2013. v. 26. Citado na página 42.

MATHWORKS, I. *MATLAB 2017b*. [S.l.]: The MathWorks, Inc. Natick, MA, USA, 2017. Citado na página 30.

PAN, I.; NOLAN, L. B.; BROWN, R. R.; KHAN, R.; BOOR, P. van der; HARRIS, D. G.; GHANI, R. Machine learning for social services: a study of prenatal case management in illinois. *American journal of public health*, American Public Health Association, v. 107, n. 6, p. 938–944, 2017. Citado na página 17.

RIBEIRO, M. T.; SINGH, S.; GUESTRIN, C. Why should i trust you?: Explaining the predictions of any classifier. In: ACM. *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. [S.l.], 2016. p. 1135–1144. Citado 6 vezes nas páginas 10, 42, 43, 44, 45 e 60.

SANTANA, F. B. d. et al. Floresta aleatória para desenvolvimento de modelos multivariados de classificação e regressão em química analítica. [sn], 2020. Citado na página 30.

SICSÚ, A. L. *Credit Scoring: desenvolvimento, implantação, acompanhamento*. [S.l.]: Blucher, 2010. Citado na página 58.

SILVA, C. L. da; ALVES, J. Q.; BRAGA, O. C.; JÚNIOR, J. W. P.; ANDRADE, L. O. M. de; OLIVEIRA, A. M. B. de. Usando o classificador naive bayes para geração de alertas de risco de óbito infantil. *Revista Electronica de Sistemas de Informaçao*, Faculdade Cenecista de Campo Largo-FACECLA, v. 16, n. 2, p. 1–15, 2017. Citado na página 17.

SOARES, F. M. et al. O papel da unidade de terapia intensiva na redução da mortalidade materna= the role of intensive care unit in reducing maternal mortality. [sn], 2016. Citado 5 vezes nas páginas 16, 18, 21, 57 e 62.

TUKEY, J. W. *Exploratory data analysis*. [S.l.]: Reading, Mass., 1977. v. 2. Citado na página 46.

UC BUSINESS ANALYTICS. *Visualizing ML Models with LIME*. 2020. Disponível em: <https://uc-r.github.io/lime>. Acesso em: 15 apr. 2020. Citado na página 60.

VALIANT, L. G. A theory of the learnable. *Communications of the ACM*, ACM New York, NY, USA, v. 27, n. 11, p. 1134–1142, 1984. Citado na página 32.

WHO, U. Unfpa, world bank. *Trends in maternal mortality*, v. 2008, n. 2010, p. 1, 1990. Citado 4 vezes nas páginas 10, 16, 18 e 19.

WILCOXON, F. Individual comparisons by ranking methods. In: *Breakthroughs in statistics*. [S.l.]: Springer, 1992. p. 196–202. Citado na página 57.

XGBOOST DEVELOPERS. *XGBoost Parameters*. 2020. Disponível em: <https://xgboost.readthedocs.io/en/latest/parameter.html>. Acesso em: 02 nov. 2020. Citado na página 42.

ANEXO A – Tabela de atributos

Tabela 3 – Tabela com os atributos selecionados para o modelo e suas respectivas descrições.

Atributo	**Descrição**
P10	IMC - Índice de Massa Corpórea
P11	acompanhamento da gestação
P12	informação se a internação foi espontânea, transf inter-hospitalar, etc
P13	informação se a paciente veio da rede pública, privada, sem pre natal, etc
P14	informação se a paciente veio da rede pública, privada, etc
P21	cirurgia no útero
P25	tipo de trabalho de parto
P29	aborto espontâneo ou induzido
P30	aborto seguro ou inseguro
P36	condição do nacimento - vivo/natimorto
P42	outras condições
P45	se houve hemorragia
P47	tipo de hemorragia
P48	se houve hipertensão
P52	outras condições
P56	condições clínicas
P58	condições laboratoriais
P67	demora no atendimento médico
P68	demora na transferência
P69	demora na comunicação entre os serviços de saúde
P70	demora na oferta de sangue e hemoderivados
P71	demora na monitorização do caso no serviço de saúde
P72	demora por motivos pessoais
P73	demora no acesso do caso ao serviço de saúde
P78	demora no acesso do caso à um profissional
P79	demora no acesso do caso à um profissional
P80	demora no acesso do caso à um profissional
p31_1	aborto - dilatação e curetagem
p31_2	aborto - ocitocina
p31_3	aborto - vácuo aspiração

Continua na próxima página

Tabela 3 – continuação da página anterior

Atributo	Descrição
p31_4	aborto - prostaglandinas
p31_5	aborto - outros
p31_6	aborto - nenhum
p43_1	hipertensão arterial crônica
p43_2	obesidade
p43_3	baixo peso
p43_4	diabetes mellitus
p43_5	tabagismo
p43_6	doenças cardíacas
p43_7	doenças respiratórias
p43_8	doenças renais
p43_9	anemia falciforme / talassemia
p43_10	HIV/AIDS
p43_11	tireoidopatias
p43_12	doenças neurológicas /epilepsia
p43_13	colagenoses
p43_14	neoplasias
p43_16	drogadição
p46_1	descolamento prematuro da placenta
p46_2	placenta prévia/acreta/increta/percreta
p46_4	rotura uterina
p46_5	hemorragia grave por aborto
p46_6	hemorragia pós parto
p46_9	outra hemorragia grave
p49_1	pré- eclâmpsia grave
p49_2	eclâmpsia
p49_3	hipertensão grave
p49_4	HELLP síndrome
p49_5	fígado gorduroso na gestação
p51_1	edema pulmonar
p51_2	convulsões
p51_3	trombocitopenia < 100mil
p51_4	crise tireotóxica
p51_5	choque
p51_6	insuficiência respiratória aguda

Continua na próxima página

Tabela 3 – continuação da página anterior

Atributo	Descrição
p51_7	acidose
p51_8	cardiopatia
p51_9	AVC
p51_10	distúrbios de coagulação
p51_11	CIVD
p51_12	tromboembolismo
p51_13	cetoacidose diabética
p51_14	icterícia/disfunção hepática
p51_15	meningite
p51_16	sepse grave
p51_17	insuficiência renal aguda
p55_1	transfusão de homoderivados
p55_2	acesso venoso central
p55_4	hospitalização prolongada (>7dias)
p55_5	intubação não relacionada à anestesia
p55_6	retorno à sala cirúrgica
p55_7	histerectomia/laparotomia
p55_8	uso de sulfato de magnésio
p55_10	outro procedimento cirúrgico maior
p57_1	cianose
p57_2	gasping
p57_3	FR>40 ou <6
p57_4	choque
p57_5	oligúria não responsiva à flúidos ou diuréticos
p57_6	distúrbio de coagulação
p57_7	perda de consciência durante 12h ou mais
p57_8	ausência de consciência E ausência de pulso-batimento cardíaco
p57_9	acidente Vascular Cerebral
p57_10	convulsão não controlada- Paralisia total
p57_11	icterícia na presença de pré eclâmpsia
p59_1	saturação de oxigênio < 90% por > 60 minutos
p59_2	PaO2 FiO2 < 200
p59_3	creatinina >= 300mmol/l ou >= 3.5 mg/dl
p59_4	bilirrubina >= 100 mmol/l ou >= 6.0 mg/dl
p59_5	pH < 7.1

Continua na próxima página

Tabela 3 – continuação da página anterior

Atributo	Descrição
p59_6	lactato > 5
p59_7	trombocitopenia Aguda (< 50 000 plaquetas)
p61_1	uso de droga vasoativa contínua
p61_2	histerectomia puerperal por infecção ou hemorragia
p61_3	transfusão de >= 5 unidades de concentrado de hemácias
p61_4	intubação e ventilação por tempo >=60 minutos não relacionado a anestesia
p61_5	diálise para insuficiência renal aguda
p61_6	ressuscitação cardio pulmonar (RCP)
p75_1	demora na procura ao serviço de saúde
p75_2	dificuldade geográfica ao acesso ao serviço de saúde
p75_3	recusa ao tratamento
p75_4	pré natal ausente ou inadequado
p75_5	qualquer outra demora
p4cat	idade
p5	cor
p6	escolaridade
p7cat	estado civil
p15cat	número de gestações
p16cat	número de partos
p17cat	número de abortos
p18cat	número de cesáreas
p20cat	parto uterino
p22cat	pré-natal
p24cat	idade gestacional na admissão
p27cat	idade gestacional na resolução do parto
p28cat	tipo de parto
p15cat2	gestações
MSS	maternal severity score