**UNIVERSIDADE FEDERAL DE MINAS GERAIS**
**Instituto de Ciências Exatas**
**Programa de Pós-Graduação em Ciência da Computação**

Tiago Amador Coelho

# Regularização de modelos para predição precoce: Um estudo na predição de complicações na UTI

Belo Horizonte
2023

Tiago Amador Coelho

# Regularização de modelos para predição precoce: Um estudo na predição de complicações na UTI

**Versão Final**

Tese apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de Minas Gerais, como requisito parcial à obtenção do título de Doutor em Ciência da Computação.

Orientador: Adriano Alonso Veloso

Belo Horizonte
2023

Coelho, Tiago Amador

C672r Regularização de modelos para predição precoce: um estudo na predição de complicações na UTI [recurso eletrônico] :/ Tiago Amador Coelho–2022.
1 recurso online (63 f. il, color.) : pdf.

Orientador: Adriano Alonso Veloso.

Tese (Doutorado) - Universidade Federal de Minas Gerais, Instituto de Ciências Exatas, Departamento de Ciências da Computação.
Referências: f.55-63

1. Computação – Teses. 2. Aprendizado de máquina – Teses. 3. Ciência de dados - Medicina – Teses. 4. Explicação do modelo – Teses. I. Veloso, Adriano Alonso. II. Universidade Federal de Minas Gerais, Instituto de Ciências Exatas, Departamento de Computação. III.Título.

CDU 519.6*82(043)

Ficha catalográfica elaborada pela bibliotecária Belkiz Inez Rezende Costa
CRB 6/1510 Universidade Federal de Minas Gerais - ICEx


UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS EXATAS
DEPARTAMENTO DE CIÊNCIA DA COMPUTAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

**FOLHA DE APROVAÇÃO**

# REGULARIZAÇÃO DE MODELOS PARA PREDIÇÃO PRECOCE: UM ESTUDO NA PREDIÇÃO DE COMPLICAÇÕES NA UTI

TIAGO AMADOR COELHO

Tese defendida e aprovada pela banca examinadora constituída pelos Senhores(a):

Prof. Adriano Alonso Veloso - Orientador
Departamento de Ciência da Computação - UFMG

Prof. Renato Vimieiro
Departamento de Ciência da Computação - UFMG

Prof. Saulo Fernandes Saturnino
Faculdade de Medicina - UFMG

Prof. Wagner Meira Júnior
Departamento de Ciência da Computação - UFMG

Prof. Leandro Balby Marinho
Departamento de Sistemas e Computação - UFCG

Profa. Soraia Raupp Musse
Faculdade de Informática - PUCRS

Belo Horizonte, 14 de junho de 2022.

Documento assinado eletronicamente por **Adriano Alonso Veloso**, **Professor do Magistério Superior**, em 19/10/2022, às 21:44, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Renato Vimieiro**, **Professor do Magistério Superior**, em 24/10/2022, às 11:28, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Wagner Meira Junior**, **Professor do Magistério Superior**, em 01/11/2022, às 10:37, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Leandro Balby Marinho**, **Usuário Externo**, em 03/11/2022, às 09:55, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Saulo Fernandes Saturnino**, **Professor do Magistério Superior**, em 03/11/2022, às 19:13, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Soraia Raupp Musse**, **Usuária Externa**, em 16/11/2022, às 11:03, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

A autenticidade deste documento pode ser conferida no site https://sei.ufmg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **1843681** e o código CRC **B94A9BD2**.

*Dedico este trabalho a todos que diretamente ou indiretamente, o tornaram possível.*

# Agradecimentos

Inicialmente agradeço a Deus por sempre estar ao meu lado em todos os momentos.

A minha esposa Patrícia e a minha filha Elisa por estarem ao meu lado todos os dias me dando carinho e força para não desistir.

Aos meus pais Salustriano e Gilca por todo o ensinamento e suporte que me deram ao longo da minha vida.

Ao meu irmão Rodrigo pelas conversas e apoio mesmo estando distante.

A minha família que sempre estiveram torcendo.

Aos profs Adriano Veloso, Saulo Saturnino e Nivio Ziviani pelos ensinamentos e contribuições que fizeram para a construção deste trabalho. Em especial ao prof Adriano, agradeço sempre por ter me dado o seu voto de confiança para a construção deste trabalho.

Aos amigos e colegas da UEFS que me deram a oportunidade da licença para a pós-graduação.

Aos amigos que fiz nos laboratórios SPEED e LIA, pelos momentos de descontração, estudos e trabalhos.

A Sônia, que é uma verdadeira mãe para todos os alunos da PGCC, por sempre estar disposta a fazer tudo o que é possível para ajudar.

A todos, o meu sincero muito obrigado.

# Resumo

Modelos de predição estão se mostrando importantes para a tomada de decisão na UTI, porém eles geralmente apresentam o problema da caixa preta porque não fornecem a informação da lógica envolvidas nas previsões específicas do paciente. Já existem técnicas capazes de analisar os modelos e gerar explicações valiosas sobre o seu funcionamento. Porém, uma vez que o modelo é gerado, é importante garantir que ele continuará a mesma lógica de predição originalmente pretendida. Sabendo que complicações podem ocorrer a qualquer momento durante a permanência do paciente na UTI, construímos nossos modelos de aprendizado de máquina utilizando atributos obtidas a partir dos dados administrativos, resultados laboratoriais e sinais vitais do paciente, disponíveis na primeira hora após a sua admissão na UTI. Para construir modelos que continuem a funcionar como originalmente projetados, primeiro propomos medir (i) como as explicações fornecidas variam para diferentes entradas (ou seja, robustez) e (ii) como as explicações fornecidas mudam com modelos construídos a partir de diferentes subpopulações de pacientes (isto é, estabilidade). Posteriormente, empregamos essas medidas como termos de regularização que são acoplados a um procedimento de seleção de atributos de modo que o modelo final forneça previsões com explicações mais robustas e estáveis. Os experimentos foram conduzidos em um conjunto de dados contendo 6.000 internações na UTI de 5474 pacientes. Os resultados obtidos em uma coorte de validação externa de 1069 pacientes com 1086 internações em UTI mostraram que a seleção de atributos com base na robustez levou a ganhos em termos de poder preditivo que variaram de 6,8% a 9,4%, enquanto a seleção de atributos com base na estabilidade levou a ganhos que variaram de 7,2% a 11,5%, dependendo da complicação. Nossos resultados são de importância prática, pois nossos modelos preveem complicações com grande antecipação, facilitando intervenções oportunas e protetoras.

**Palavras-chave:** Aprendizado de máquina. Ciência de Dados Médicos. Explicabilidade do modelo. Regularização.

# Abstract

Predictive models are proving to be important for decision-making in the ICU, but they often present the black box problem because they do not provide the information from the logic involved in patient-specific predictions. There are already techniques capable of analyzing the models and generating valuable explanations about their functioning. However, once the model is generated, it is important to ensure that it will continue with the same prediction logic originally intended. Knowing that complications can occur at any time during a patient's ICU stay, we built our machine learning models using features obtained from the patient's administrative data, laboratory results, and vital signs, available within the first hour after admission to the ICU.This enables our models to provide great anticipation because complications can occur at any moment during ICU stay. To build models that continue to work as originally designed we first propose to measure (i) how the provided explanations vary for different inputs (that is, robustness), and (ii) how the provided explanations change with models built from different patient sub-populations (that is, stability). Second, we employ these measures as regularization terms that are coupled with a feature selection procedure such that the final model provides predictions with more robust and stable explanations. Experiments were conducted on a data set containing 6000 ICU admissions of 5474 patients. Results obtained on an external validation cohort of 1069 patients with 1086 ICU admissions showed that selecting features based on robustness led to gains in terms of predictive power that varied from 6.8% to 9.4%, whereas selecting features based on stability led to gains that varied from 7.2% to 11.5%, depending on the target complication. Our results are of practical importance as our models predict complications with great anticipation, thus facilitating timely and protective interventions.

**Keywords:** Machine Learning. Medical Data Science. Model Explainability. Regularization

# Lista de Figuras

# Lista de Tabelas

# Sumário

# Capítulo 1

# Introdução

As Unidade de Terapia Intensivas (UTIs) fornecem instalações, recursos e pessoal especializado para o manejo abrangente de pacientes que apresentam ou estão em risco de desenvolver disfunções orgânicas com risco de vida [48]. Tratar o distúrbio fisiológico evidente é apenas o primeiro passo no cuidado dos pacientes da UTI, pois eles estão sujeitos a muitas complicações decorrentes da terapia avançada [79]. Complicações importantes, como infecções, delirium, miopatias, neuropatias e distúrbios nutricionais, têm se mostrado consistentemente associadas a prejuízos na qualidade de vida em longo prazo [56] e morte [84]. Assim, todos os esforços devem ser empreendidos para prevenir essas complicações.

Como as UTIs são ambientes ricos em dados onde vários sinais são monitorados continuamente, modelos de aprendizado de máquina já foram desenvolvidos para identificar pacientes em risco de complicações [4, 31, 34, 38, 39, 49, 71]. Embora esses modelos geralmente sejam excelentes na captura de relações complexas entre sinais [20, 53], eles sofrem do problema da caixa-preta [22], ou seja, o mecanismo pelo qual os sinais são combinados a fim de converter as entradas nas saídas é opaco [77]. Modelos complicados de caixa-preta levantam algumas preocupações, pois as decisões médicas na UTI podem ter consequências de vida ou morte [78]. Assim, o mecanismo de predição de um modelo de risco deve estar de acordo com o conhecimento clínico real antes que o modelo seja colocado em uso [54, 17, 33].

Algoritmos que visam explicar modelos de caixa-preta estão sendo cada vez mais usados para dar sentido ao mecanismo de previsão [1, 66]. Esses algoritmos fornecem uma compreensão mecanicista do funcionamento do modelo, revelando como os recursos estão relacionados em conjunto para formar a previsão final [50]. Um tipo importante de explicação é determinar a importância do recurso para uma predição particular [45, 62]. No caso da modelagem de dados da UTI, a importância de um sinal é calculada levando-se em consideração muitas formas de interação envolvendo o sinal. Este tipo de explicação pode mitigar o problema com os modelos caixa-preta, uma vez que se pode projetar ou selecionar um modelo que associe importâncias apropriadas aos sinais, fornecendo assim uma razão confiável para as previsões do modelo [2, 77].

Uma vez que o modelo é colocado em uso, no entanto, uma grande preocupação é

se ele continuará funcionando como originalmente planejado [59]. Confiar nas explicações fornecidas significa não apenas confiar no mecanismo de predição do modelo, mas também nos dados que ele foi construído [81]. Os dados da UTI são razoavelmente complexos e o modelo resultante pode ter falhas em potencial simplesmente porque os dados de treinamento estão incompletos, no sentido de que não esgotaram todas as maneiras pelas quais os sinais podem interagir. Nesse caso, mesmo pequenas variações no conjunto de treinamento podem levar a explicações completamente diferentes.

## 1.1 Motivação

Embora alguns modelos possam ser excelentes para a predição de complicações, muitos deles são uma verdadeira caixa-preta para os clínicos, que não entendem o seu funcionamento, implicando em uma desconfiança ou em uma falta de confiança no modelo. Por outro lado, as implicações de confiar demais em um modelo de risco de UTI podem ser desastrosas, pois um modelo falho (mas confiável) pode tranquilizar falsamente um clínico, possivelmente levando a uma mudança sistemática na prática clínica [5, 61]

Dentro deste contexto, o trabalho pretende investigar se as explicações dadas pelos modelos são robustas e estáveis para diferentes subpopulações, afim de verificar se entradas similares não geraram explicações substancialmente diferentes, como também se pequenas variações nos dados de treinamento não geram a explicações diferentes, para posteriormente criar um termo de regularização para que as explicações das predições geradas pelo modelo final sejam robustas e estáveis.

## 1.2 Contribuições

Esta tese visa alcançar as seguintes contribuições principais:

- criação e avaliação de um termo de regularização com base na estabilidade das explicações do modelo preditor
- criação e avaliação de um termo de regularização com base na robustez das explicações do modelo preditor
- Avaliação dos modelos de um modelo em um cenário real

## 1.3 Objetivo Geral

O objetivo geral deste trabalho é construir um modelo que realize predições precoces de maneira eficaz utilizando uma parcial dos dados. Um exemplo da aplicação deste modelo seria em uma UTI onde temos que predizer as complicações o mais cedo possível, utilizando apenas os dados da primeira hora dos pacientes, onde muitos eventos podem acontecer após a primeira hora até o momento do desfecho (alta/óbito) do paciente.

### 1.3.1 Objetivos Específicos

Especificamente, pretende-se:

- Medir como as explicações fornecidas variam dadas as diferentes entradas (robustez);
- Medir como as explicações fornecidas mudam com modelos construídos a partir de diferentes subpopulações (estabilidade);
- Criar um termo de regularização para que o modelo final forneça previsões com explicações mais robustas e estáveis.

## 1.4 Organização

O trabalho está organizado da seguinte forma: no Capítulo 2 há uma discussão dos conceitos e trabalhos relacionados com o trabalho afim de prover uma base teórica. No Capítulo 3 e apresentada a base de dados utilizada no trabalho. O Capítulo 4 aborda o método proposto para o problema, explicando como foi a sua elaboração e o seu funcionamento. O Capítulo 5 apresenta os experimentos realizados e os resultados obtidos. Por fim, no Capítulo 6 é apresentada a conclusão do trabalho, bem como são sugeridos trabalhos futuros.

# Capítulo 2

# Conceitos e Trabalhos Relacionados

Esse trabalho foca na construção modelos de aprendizado de máquina para prever complicações graves usando elementos administrativos e clínicos coletados imediatamente após a admissão do paciente na unidade de terapia intensiva (UTI), bem como a medição da qualidade da explicabilidade dos modelos gerados através de sua robustez e estabilidade. Revisaremos os trabalhos relacionados: Aprendizado de Máquina, LightGBM, SHAP, Teste de Mantel e trabalhos relacionados que são componentes chaves deste trabalho.

## 2.1 Conceitos

### 2.1.1 Aprendizado de Máquina

Nas últimas décadas uma área da Inteligência Artificial (IA) conhecida como Aprendizado de Máquina (AM) vem chamado muita atenção da comunidade científica por sua capacidade de aprendizado de padrões complexos a partir da análise dos dados, permitindo assim realizar predições com alta taxa de precisão.

Com o seu destaque, o Aprendizado de Máquina passou a ser amplamente utilizado em diversos problemas como *Knowledge Discovery in Databases* (KDD) [25], sistemas de recomendações [74], reconhecimento de caractere [12], predição de doenças [8], entre outros.

Afim de categorizar o Aprendizado de Máquina, as formas de aprendizado podem ser divididas da seguinte maneira: Aprendizado Supervisionado e Aprendizado Não Supervisionado.

### 2.1.2 Aprendizado Supervisionado

É utilizado em problemas de predição ou classificação, onde os dados estão classificados e rotulados. O termo "supervisionado"se dá justamente pela necessidade de conhecimento prévio dos rótulos, para que durante o aprendizado seja avaliado a capacidade de predizer/classificar o valor de saída para novos dados.

Segundo [29] a predição/classificação é definido como uma tarefa no qual a partir da observações históricas de casos anteriores, pretende-se prever/classificar o valor para um novo caso.

Formalmente podemos definir como: Dado um conjunto de treinamento na forma $(X_i, Y_i)$, no qual $X_{i\in\{1,..,N\}} = \{x_1, ..., x_k\}$, onde $X$ contem $N$ instâncias e $k$ atributos, e $Y_i \in IR$. O algoritmo deverá ser capaz de aprender a partir dos dados de treinamento uma correlação entre os atributos, e ao ser apresentado um novo dado de entrada $X_a$ ele consiga prever o valor de $Y_a$.

### 2.1.3 Performance e Interpretabilidade

Aprendizado de máquina vem sendo utilizado para resolver problemas de tomada de decisão, como em hospitais. No entanto, modelos que tem um alto grau de complexidade (baseadas em redes neurais profundas) são verdadeiras caixas-pretas, tendo uma alta performance e uma baixa interpretabilidade de como os resultados foram obtidos, enquanto modelos clássicos são altamente interpretáveis e tem uma baixa performance. Essa limitação "performance X interpretabilidade"veem limitando o seu uso, pois em vários problemas onde é crítico o bem entendimento das contribuições individuais dos atributos para o resultado do modelo, não se pode utilizar modelos de alta performance.

### 2.1.4 LightGBM

*Gradient Boosting Decision Trees* (GBDT) é um comitê de *Decision Trees* no qual é realizado um treinamento sequencial onde em cada iteração, a árvore será construída baseada nos erros das *trees* anteriormente construídas (*Negative Gradients*), gerando um modelo final com uma grande capacidade de predição.

O *Light Gradient Boosting Decision Trees* (LightGBM) é um algoritmo da família dos *GBDT* desenvolvido por [41] que contém duas novas técnicas que são o *Gradient-based One-Side Sampling* e o *Exclusive Feature Bundling*. Ainda em seu trabalho, [41] afirma que um dos maiores problemas do *GBDT* é a criação das *Decision Trees*, pois encontrar os melhores pontos para a divisão das árvores é a parte que mais demanda tempo, já que precisam verificar todas as instâncias de dados para estimar o ganho de informação de todos os pontos de divisão possíveis, o que consome muito tempo.

Para resolver esse problema foi proposto o *Gradient-based One-Side Sampling* (GOSS) que foca em pegar as instâncias que contem gradientes altos e pega uma amostra das instâncias com gradientes baixos, assim são utilizados uma quantidade reduzida de instâncias para a divisão das árvores. O pseudocódigo GOSS pode ser visto em Algorithm 1, o pseudocódigo foi retirado e adaptado de [41].

**Algorithm 1:** *Gradient-based One-Side Sampling*

```
input: I: dados de treinamento
input: d: iterações
input: a: taxa de amostragem de dados com gradiente alto
input: b: taxa de amostragem de dados com gradiente baixo
input: loss: função de perda
input: L: modelo classificador
models ← {};
fact ← (1−a)/b
topN ← a × len(I);
randN ← b × len(I);
for i = 0 to d do
    preds ← models.predict(I);
    g ← loss(I, preds);
    w ← {1,1,...};
    sorted ← GetSortedIndices(abs(g));
    topSet ← sorted[1:topN];
    randSet ← RandomPick(sorted[topN:len(I)], randN);
    usedSet ← topSet + randSet;
    w[randSet] × = fact // atribui o peso fact as amostras com gradientes baixos;
    newModel ← L(I[usedSet], - g[usedSet], w[usedSet]);
    models.append(newModel);
end
```

Outro problema abordado em [41] foi a alta dimensionalidade dos dados, que interfere na diretamente da performance. Para atacar esse problema foi elaborado o método *Exclusive Feature Bundling* (EFB), no qual o método combina muitos atributos em um novo atributo, assim reduzindo a dimensionalidade dos dados sem que haja grande perda de informação.

O método EFB é dividido em duas etapas, a primeira *Greedy Bundling* no qual é construído um grafo ponderado onde os pesos das arestas correspondem aos conflitos existentes entre os atributos, posteriormente os atributos são ordenados de forma decrescente com relação ao seu grau no grafo e por fim verifica cada um dos atributos atribuindo ele a um *bundle* com baixo conflito ou criando um novo *bundle*. O pseudocódigo do *Greedy Bundling* pode ser visto em Algorithm 2, retirado e adaptado de [41].

**Algorithm 2:** *Greedy Bundling*

```
Input: F: atributos
Input: K: número máximo de conflitos
Construa o grafo G
searchOrder ← G.sortByDegree();
bundles ← {};
bundlesConflicts ← {};
for i in searchOrder do
    needNew ← True;
    for j = 1 to len(bundles) do
        cnt ← ConflictCnt(bundles[j],F[i])
        if cnt + bundlesConflict[i] ≤ K then
            bundles[j].add(F[i])
            needNew ← False
            break
        end
    end
    if needNew then
        Add F[i] como um novo bundle em bundles
    end
    Output: bundles
end
```

A segunda etapa do método EFB, chamado de *Merge Exclusive Features*, é reduzir a complexidade do treinamento, para isso é realizado uma junção de dois ou mais atributos de um mesmo *bundle*, onde o ponto principal é garantir que os valores dos atributos originais possam ser identificados nos *bundles*. O algoritmo desta etapa pode ser visto em Algorithm 3, retirado e adaptado de [41].

**Algorithm 3:** *Merge Exclusive Features*

```
Input: numData: número de dados
Input: F: um bundle com atributos exclusivos
binRanges ← {0};
totalBin ← 0;
for f in F do
    totalBin += f.numBin
    binRanges.append(totalBin)
end
newBin ← new Bin(numData)
for i = 1 to numData do
    newBin[i] ← 0
    for j = 1 to len(F) do
        if F[j].bin[i] ≠ 0 then
            newBin[i] ← F[j].bin[i] + binRanges[j]
        end
    end
end
Output: newBin
Output: binRanges
```

Sobre a performance e a eficiência do *LightGBM*, [44], [41] e [9] demonstram a sua superioridade sobre o algoritmo *XGBoost* e outros algoritmos de classificação em vários aspectos como: maior velocidade de treinamento e maior eficiência, maior precisão dos

resultados, menor consumo de memória e capacidade de processamento de grandes bases de dados.

### 2.1.5 SHAP - Shapley Additive exPlanations

*Shapley Additive exPlanations* (SHAP) foi proposto por [46], é empregado para interpretar os resultados de qualquer modelo de aprendizado de máquina e analisar a importância individual dos atributos. Ele se baseia na teoria do jogo Valores de Shapley [69] e em explicações locais [63], para que possa oferecer meios para estimar a contribuição de cada atributo.

O SHAP está sendo utilizado nos últimos anos na área de aprendizado de máquina para explicar os resultados dos modelos, principalmente daqueles que são uma caixa preta. [58] utilizou o SHAP para explicar o modelo de predição de acidentes de trânsito, [28] em seu trabalho sobre predição precoce de mortalidade de pacientes idosos, utilizou a ferramenta para evidenciar quais atributos influenciavam mais o modelo gerado.

Em seu livro, [52] define que o objetivo do SHAP é explicar a previsão de uma instância $x$ calculando a contribuição de cada atributo para a previsão, onde são calculados os valores de Shapley a partir da teoria dos jogos de coalizão e, cada valor dos atributos de uma instância atuam como jogadores em uma coalizão.

Os valores de Shapley ($\phi$) são determinados através do retreinamento do modelo em todos os subconjuntos de atributos $S \subseteq N$, onde $N$ é o conjunto com $n$ atributos. É atribuído um valor de importância a cada atributo que representa o efeito na previsão do modelo ao incluir esse atributo [46] por meio da equação 2.1:

$$\phi_i = \sum_{S \subseteq N\{i\}} \frac{|S|!(n - |S| - 1)!}{n!}[v(S \cup \{i\}) - v(S)] \tag{2.1}$$

Assim, uma função linear de características binários $g$ é definido com:

$$g\left(z'\right) = \phi_0 + \sum_{j=1}^{M} \phi_j z'_j \tag{2.2}$$

onde $g$ é a explicação do modelo, $z' \in \{0,1\}^M$ é o vetor de coalizão, $M$ o tamanho máximo da coalizão e $\phi_j \in \mathbb{R}$ é o valor para um atributo $j$ [46].

A Figura 2.1, retirada do presente trabalho, exemplifica como o SHAP classifica os atributos pela soma das magnitudes dos valores SHAP de todas amostras, evidenciando a distribuição dos impactos de cada atributo na saída do modelo, revelando o atributo mais influente é a "Highest Creatinine1h/Highest Fi O21h".

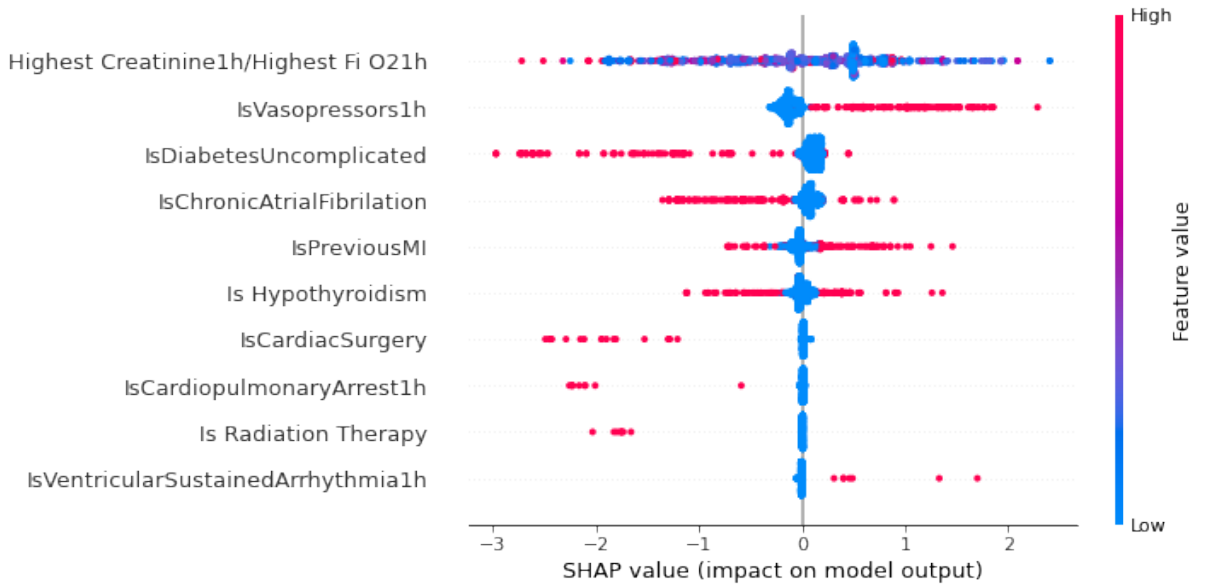


Figura 2.1: Classificação dos atributos de acordo com a soma das magnitudes - SHAP

### 2.1.6 AUROC - *Area Under the ROC Curve*

[24] e [43] definem a curva *Receiver Operating Characteristics* (ROC) como um gráfico (espaço ROC) que mostra a taxa de verdadeiros positivos[1] no eixo vertical e a taxa de falsos positivos[2] no eixo horizontal, à medida que o *threshold* de classificação varia. [43] ainda afirma que a área sob a curva ROC (*Area Under the ROC Curve* (AUROC)) será equivalente a estatística $U$ de Wilcoxon-Mann-Whitney.

Classificadores probabilísticos produzem naturalmente uma pontuação (*score*) ou um valor probabilístico que represente o grau em que uma instância é membro de uma classe. Esses valores são usados pelos classificadores, juntamente com um *threshold* para produzir um classificador binário, onde se a saída do classificador estiver acima de um determinado *threshold*, o classificador produzirá um A, caso contrário um B.

A exemplo, Figura 2.2 demonstra uma curva ROC em um conjunto de teste de 10 instâncias. As instâncias utilizadas são demonstradas na Tabela 2.1, que ilustra a classe da cada instância e o *score* obtido por cada uma delas ao ser processada pelo classificador.

[1] $sensibilidade = TP/(TP + FN)$
[2] $FP/(FP + TN)$

Tabela 2.1: Instâncias geradas por um classificador probabilístico

| Inst # | Classe | Score |
|---|---|---|
| 1 | p | 0.95 |
| 2 | n | 0.86 |
| 3 | p | 0.70 |
| 4 | p | 0.55 |
| 5 | n | 0.51 |
| 6 | n | 0.49 |
| 7 | p | 0.40 |
| 8 | p | 0.35 |
| 9 | n | 0.27 |
| 10 | n | 0.15 |

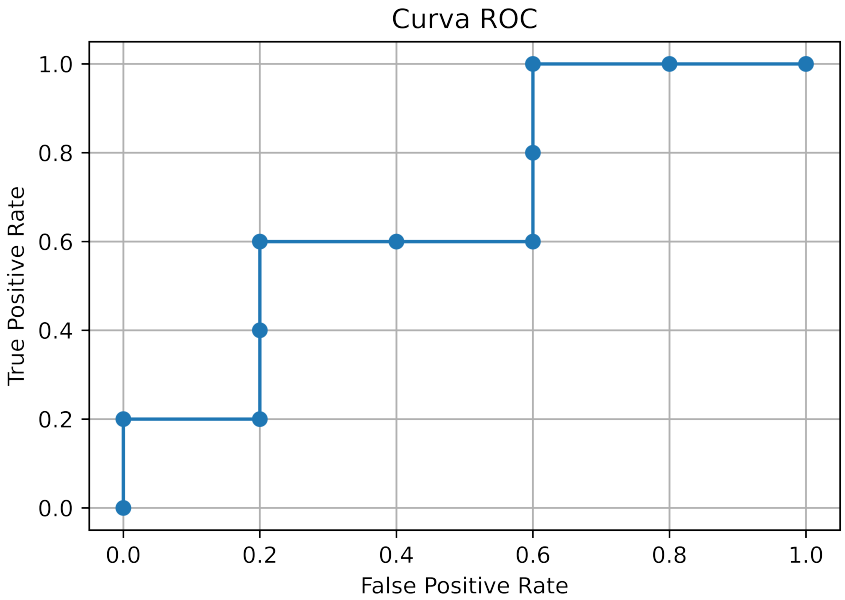


Figura 2.2: Curva ROC gerada através da Tabela 2.1

Para gerar a curva ROC do exemplo da Figura 2.2 segue os seguintes passos:

1. O passo no sentido para cima é dado pelo valor $\frac{1}{\#InstanciasPositivas}$, 0,2 neste exemplo;
2. O passo no sentido para direita é dado pelo valor $\frac{1}{\#InstanciasNegativas}$, 0,2 neste exemplo;
3. Ordena de forma decrescente pelo score a saída do classificador (Tabela 2.1);
4. Partindo do ponto $(0, 0)$ e, para cada instância da Tabela 2.1, se a classe for p (positiva) faça o passo para cima, caso contrário faça o passo será para a direita.

Toda curva ROC é gerada através de um número finito de instância produzindo um degrau no gráfico (espaço ROC) e a medida que o número de instâncias aumenta, mais contínua vai ficando a curva.

Para se calcular a AUROC, basta calcular a área do retângulo formado a cada passo para a direita, no exemplo acima teremos: $((0,2*0,2)+(0,2*0,6)+(0,2*0,6)+(0,2*1,0)+)0,2*1,0))$ totalizando um valor de 0,68.

### 2.1.7 Teste de Mantel

[47] em seu estudo projetou um método para calcular o coeficiente de correlação entre dois agrupamentos de doenças em estudos epidemiológicos, o qual gera uma estimativa de associação entre uma doença e um fator de risco. Muito frequentemente, os dados para pesquisa em saúde não são obtidos de forma adequada, por meio de amostragem probabilística de uma população-alvo. Embora não seja um problema calcular o coeficiente de correlação entre dois conjuntos, testar sua significância não pode ser feito usando a abordagem usual, já que é necessário assumir a independência dos dados.

O teste de Mantel consiste em um processo de comparação entre duas matrizes simétricas de distância. O teste pode ser descrito da seguinte forma: seja o conjunto de dados $(xi, yi)$ com $i = 1, ..., n$ elementos. São calculas as matrizes simétricas $n$ x $n$ de distância $D^x_{ij} = dist(x_i, x_j)$ e $D^y_{ij} = dist(y_i, y_j)$. O teste de Mantel é dado pela soma dos produtos entre as matrizes de distâncias $D^x_{ij}$ e $D^y_{ij}$, de acordo com a seguinte equação:

$$r = \sum_{i=1}^{n-1} \sum_{j=1+1}^{n} D^x_{ij} \times D^y_{ij} \tag{2.3}$$

em que $r$ é o coeficiente de Mantel. Estendendo o teste de Mantel, em seus estudos [73] apresenta o teste parcial de Mantel, em que uma terceira matriz de distância é adicionada. Este teste é baseado no coeficiente de correlação entre as matrizes de distâncias $D^x_{ij}$ e $D^y_{ij}$ enquanto controla o efeito da terceira matriz de distância $D^z_{ij}$. O trabalho de [21] apresenta uma análise do teste de Mantel junto com a análise de correlação de Pearson e quando ambos os métodos são aplicáveis.

É notável a diversidade de áreas com que o teste de Mantel e suas extensões são aplicadas. No trabalho de [19] apresenta uma revisão do uso do teste de Mantel e suas extensões, de sua aplicação associada a modelos teóricos em genética populacional. Extensões do teste de Mantel também tem sido usadas na área de aprendizado de máquina, como no trabalho [75], o qual propõe um novo método baseado em rede neural profunda para a previsão do tempo de sobrevivência global de pacientes. Foram analisadas imagens de ressonância magnética do celebro feitas no processo pré-operatório de pacientes que possuem tumor cerebral para orientar a previsão do tempo de sobrevivência.

### 2.1.8 Regularização

Durante o processo de treinamento, os algoritmos de Aprendizado de Máquina, principalmente os algoritmos mais complexos, realizam ajustes no modelo que está sendo construído seja capaz de realizar a predição/classificação. Neste processo, pode ocorrer o *overfitting* ou o *underfitting* do modelo.[85, 37]

Segundo [37] o *underfitting* significa que durante o processo de aprendizagem, o modelo gerado não conseguiu encontrar nenhuma relação entre os atributos, resultado em um classificador que não consegue realizar a classificação correta do seu próprio conjunto de treinamento.

Já o *overfitting*, segundo [85], o modelo gerado tem um desempenho excelente em sua base de treinamento, porém na base de treinamento ele não consegue ter o mesmo desempenho. Isso ocorre porque durante o treinamento o modelo aprendeu tão bem as regras e relações existente entre os atributos dos dados de treinamento, que ao receber novos dados as regras e relações aprendidas não tem validade, sendo assim o modelo não tem a capacidade de generalização.

Uma alternativa de se evitar o *overfitting* é aumentar o tamanho da base de dados para o treinamento do modelo, mas nem sempre é possível. Uma outra alternativa é o uso de técnicas de regularização ou de reguladores.[67]

A técnica de regularização é utilizada para penalizar os coeficientes de valores altos da função objetivo, fazendo com que o modelo fique menos sensível a ruídos, aumentando a sua capacidade de generalização [60]. Dentre várias técnicas de regularização, destacamos a L1 e L2.

A regularização L1 ou *Lasso* (equação 2.4) atribui uma penalização igual ao valor absoluto dos coeficientes, em destaque. Note que a função *Loss* pode ter valor 0, fazendo com que a regularização L1 realize uma *"feature selection"* quando há uma grande quantidade de atributos.

$$Loss = Error(Y - \widehat{Y}) +\lambda \sum_{1}^{n} |w_i| \tag{2.4}$$

A regularização L2 ou *Ridge* (equação 2.5), também atribui uma penalização aos coeficientes, como visto na regularização L1, porém a penalização é igual ao quadrado da magnitude dos coeficientes.

$$Loss = Error(Y - \widehat{Y}) +\lambda \sum_{1}^{n} w_i^2 \tag{2.5}$$

## 2.2 Trabalhos Relacionados

A identificação prévia de pacientes com risco de complicação na UTI ou fora dela tem sido motivo de preocupação e vem sendo abordados em vários estudos [68] [35] [23] [32] [76]. Mais especificamente, alguns estudos que avaliam os riscos de complicações nos pacientes vêm obtendo bons resultados quando utilizam as técnicas de Aprendizado de Máquina.

[87] descreve o uso de técnicas aprendizado de máquina, especificamente o XGBoost, para a predição de complicações após cirurgia cardíaca pediátrica. Os autores criaram dois modelos de predição com o XGBoost, um sem otimização e outro otimizado, no qual ambos tiveram resultados superiores (AUROC de 0,82) quando comparados aos sistemas de pontuação RACHS-1, Aristotle, STS-EACTS e STS (AUROC de 0,75).

[30] demonstra que modelos supervisionados de aprendizado de máquina podem prever melhor as complicações pós-operatórias após artroplastia total do ombro (TSA) do que os índices de comorbidade. Neste trabalho os autores utilizaram atributos como idade, índice de massa corporal (IMC), tempo operatório, tabagismo, comorbidades, diagnóstico, hematócrito e albumina pré-operatórios para prever complicações como transfusão, longo período de estadia ($>$ 3 dias), infecção de sítio cirúrgico, o retorno à sala de cirurgia, embolia pulmonar e readmissão.

A pesquisa de [83] faz uso de uma *Long Short-Term Memory* (LSTM), um tipo de *Deep Neural Networks* (DNN), para realizar a predição da mortalidade de pacientes em UTIs. Nos resultados, foi mostrado que o escore SOFA teve uma preditividade moderada (AUC de 0,72), o modelo utilizando regressão logística mostrou bom desempenho (AUC de 0,82), enquanto o modelo desenvolvido por eles teve o melhor desempenho, obtendo um AUC de 0,88.

[64] criaram, treinaram e avaliaram mais de 100 modelos de XGBoost para prever se o paciente evoluiria para um estado crítico após ser diagnosticado com o COVID-19. Os resultados apresentados no trabalho mostram um alto desempenho preditivo dos modelos criados (AUCROC 0,861), além disso, a análise de interpretabilidade identificou idade avançada, pneumonia, IMC mais alto, diabetes, sexo masculino, falta de ar, doença cardiovascular, ausência de tosse, etnia não hispânica e temperatura corporal elevada como os fatores preditivos mais importantes para estado crítico.

Em [51], os autores desenvolveram um algoritmo de aprendizados de máquina para monitoramento em tempo real, no qual utilizando 5 minutos de dados fisiológicos, o algoritmo consegue predizer com 30 minutos de antecedência um evento de hipotensão em pacientes. Nos experimentos foram utilizados dados de 400 pacientes e aproximadamente 181000 horas de dados fisiológicos, onde o algoritmo o algoritmo demonstrou 94% de precisão, 85% de sensibilidade e 96% de especificidade na previsão de hipotensão dentro

de 30 minutos que antecedem os eventos.

No trabalho de [3], foi proposto um modelo de rede neural híbrida para predizer o risco de mortalidade, em UTIs neonatais, com janelas de risco de 3, 7 e 14 dias, usando apenas os atributos: peso ao nascer, idade gestacional, sexo, frequência cardíaca e frequência respiratória. O melhor resultado se deu com o modelo que utilizou a janela de risco de mortalidade em 3 dias, conseguindo superar os estados da arte encontrados na literatura, obtendo uma AUROC de 0,9336 com desvio padrão de 0,0337 no 5 *folds cross validation*. Além disso, o modelo proposto é capaz de atualizar continuamente a avaliação do risco de mortalidade, permitindo a análise das tendências de saúde e respostas ao tratamento.

O foco do estudo realizado por [57] foi a identificação precoce de pacientes críticos que iriam necessitar de ventilação mecânica prolongada (VMP) e a realização da traqueostomia. No estudo realizado sobre a base de dados MIMIC III, foi utilizado o algoritmo *Gradient-boosted decision tree* para realizar a predição, alcançando uma AUROC de 0.820 para a predição da VMP e uma AUROC de 0.830 para a predição da realização da traqueostomia.

[27] diz que reinternações de pacientes em unidades de terapia intensiva estão associados a um aumento da mortalidade, morbidade e também dos custos. Ainda no trabalho, é afirmado no trabalho que os modelos atuais de predição de reinternações em UTI tem um valor preditivo moderado e que utilizam de diversos atributos fisiológicos que podem ser avaliados em momentos distintos da internação do paciente. Assim, ele propõe uma abordagem combinando a modelagem *fuzzy* com a seleção de atributos através de árvore de busca, mostrando que um é possível a criação de um modelo com um maior valor preditivo pode ser alcançado utilizando menos atributos fisiológicos, nas quais esses atributos podem ser avaliados nas 24 horas anteriores à alta.

O trabalho de [11] utilizou de aprendizado de máquina para identificar exames desnecessários em pacientes com sangramentos gastrointestinais, visando a redução de custos na UTI. Para alcançar o objetivo foi utilizada a técnica de modelagem *fuzzy* em uma base de dados reais com 746 pacientes para a criação de um modelo preditivo. A acurácia do modelo criado foi superior a 80%, demonstrando que poderia haver uma redução média de 50% no total de exames laboratoriais.

Já em [55] faz uso de *Deep Convolutional Neural Networks* (DCNN) para o diagnóstico da doença de Alzheimer através da análise das imagens de ressonância magnética. Para tal, foi desenvolvido pelos autores um método chamado de *Swap Test*, que produz mapas de calor que retratam as áreas do cérebro que são indicativos da doença de Alzheimer, fornecendo interpretabilidade ao modelo desenvolvido, deixando em um formato que é melhor compreensível para os médicos.

Em um outro trabalho de diagnóstico da doença de Alzheimer, [18] treinou um modelo usando redes neurais profundas utilizando 2109 imagens para treinamento e 40 imagens para teste, conseguindo predizer a doença com 75 meses de antecedência do

diagnóstico final, com uma AUROC de 0,98, uma confiança de 95%, uma taxa de 82% de especificidade e com 100% de sensibilidade.

As técnicas de aprendizado de máquina também veem sendo usadas para a predição da necessidade de um paciente precisar ser internado na UTI. Esse problema é extremamente importante pois devido às restrições de recursos, os profissionais de saúde da linha de frente podem não conseguir fornecer o monitoramento e a avaliação frequentes necessários para todos os pacientes com alto risco de deterioração clínica, ocasionando um maior risco a vida do paciente.[10] atacou o problema utilizando algoritmos de *random forest*, conseguindo resultados alcançando valores de 76,2% de acurácia e 79,9% de AUC para pacientes internados com COVID-19.

[16] afirma que o monitoramento extensivo em unidades de terapia intensiva (UTIs) gera grandes quantidades de dados que contêm inúmeras tendências que são difíceis para os médicos avaliarem sistematicamente, e que as abordagens atuais descartam informações importantes destes dados. Para resolver esse problema [16] criou um modelo baseado no aprendizado profundo (*Deep Learning*) que é capaz de utilizar todos os dados coletados, sem a necessidade de realizar seleção de atributos ou quaisquer outro tratamento nos dados, para predizer o risco de morte do paciente. Em seus testes, foram obtidos utilizando as primeiras 24 horas dos dados de cada paciente, conseguindo um AUC superior a 0.8, enquanto os modelos SAPS II e OASIS obtiveram AUCs de 0,72 e 0,76 respectivamente.

[42] apresentou um estudo onde técnicas de aprendizado de máquinas como *Logistic Regression*, *Decision Tree*, *Random Forest*, *Support Vector Machine* e *Adaptive Boosting* foram utilizadas para identificar precocemente ou predizer doenças do coração, câncer de mama e diabetes. Os resultados do estudo evidenciaram o quão assertivos as técnicas são em predizer essas doenças, obtendo uma acurácia de 87,1% para doenças do coração utilizando a regressão logística, 85,71% para diabetes com o *Support Vector Machine* (SVM) e 98,57% utilizando o classificador AdaBoost para a detecção do câncer de mama.

O aprendizado de máquina foi utilizado por [26] para predizer se pacientes, que estavam infectados pelo COVID-19, iram evoluir para um estado crítico, necessitando de uma UTI. A base de dados foi fornecida pelo hospital Beneficência Portuguesa de São Paulo e continha dados de 1040 pacientes de COVID-19, no qual 53,3% eram homens, com média de idade de 51,7 anos e 63,8% dos pacientes eram brancos. Todos os modelos criados por eles tiveram um resultado de AUROC superior a 0,91, com sensibilidade de 0,92 e especificidade de 0,82.

[15] propôs um modelo baseada em uma rede neural convolucional para que através de imagens de raio X pudesse predizer os pacientes que estavam infectados com a COVID-19. Para o treinamento do modelo foram usadas 1384 imagens de pacientes com idades de 18 a 63 anos e para teste foram usadas 350 imagens. O modelo nos testes chegou a obter 89,7% de acurácia e obteve 94,0% de AUC.

[72] em seu estudo afirma que durante a pandemia de COVID-19, a tomografia

computadorizada (TC) era uma alternativa ao teste RT-PRC para o diagnóstico da doença, porém a segmentação de imagens de TC é demorada e apresenta vários desafios, como as altas disparidades de textura, tamanho e localização das infecções. Para atacar esses problema ele propôs um *framework* baseado em uma DNN para realizar a segmentação das imagens. Como resultado o método proposto obteve um *dice score* de 80,3% e um *IoU score* de 68,77%, sem um resultado superior aos algoritmos em estado da arte.

[65] utiliza técnicas de aprendizado de máquina para realizar análise de texturas de imagens de ressonância magnética para criar um modelo preditivo para o diagnóstico de lesões adrenais. O modelo preditivo criado obteve uma acurácia de 80%, já um radiologista especialista obteve uma acurácia de 73%.

[40] realizou um estudo no qual descreve sobre o futuro e as possibilidades do uso de aprendizado de máquina no contexto da cirurgia robótica. Ele analisa que o procedimento cirúrgico pode ser decomposto em episódios, que para cada episódio pode ser aprendido como uma habilidade e essas habilidade podem ser incorporadas no robô cirúrgico para que ele possa tomar as decisões apropriadas no momento apropriado. Em suas conclusões ele aponta que ao longo do tempo os robôs cirurgiões podem ganhar mais autonomia, resultando em sistemas semi-autônomos ou até mesmo autônomos.

O gerenciamento de leitos e de recursos é um grande problema na área de saúde, baseado nisso [14] utilizando registros médicos eletrônicos de pacientes cardíacos do hospital *King Abdulaziz Cardiac Center* desenvolveu quatro modelos de aprendizado de máquina ( *Random Forest*, Redes Bayesianas, SVM e Redes Neurais ). Nos experimentos foram utilizados dados de 16414 internações, onde haviam 12769 pacientes, dentre eles 68,2% eram homens. Dentre os modelos desenvolvidos, o *Random Forest* foi o obteve o melhor resultado com sensibilidade de 0,8, acurácia de 0,8 e AUROC de 0,94.

[86] projetou vários modelos preditores de taquicardia utilizando aprendizado de máquina, para criar um score de risco de pacientes na UTI. Normalmente pacientes que tiveram episódios de taquicardia tiveram aumento do suporte vasopressor, houve um aumento no tempo de permanência na UTI e também aumento da mortalidade. Dentre os modelos projetados, o que obteve o melhor resultado foi o *Random Forest* com acurácia de 0,847 e AUROC de 0,921.

Por fim, o problema da caixa preta vem sendo considerado um aspecto importante quando é utilizado aprendizado de máquina para criação de modelos de predição clínica [13]. A falta de interpretabilidade dos modelos de aprendizado de máquina está sendo cada vez mais associada à confiança clínica insuficiente nos modelos e à pouca aprovação dos médicos [80]. [70] empregaram um método de aprendizado profundo que produziu um modelo interpretável para uma previsão precoce de sepse.

# Capítulo 3

# Materiais e Métodos

Neste capítulo serão apresentados os materiais e métodos utilizados na pesquisa, incluindo os dados e algoritmos.

## 3.1 Dados

Para o desenvolvimento do modelo, foram coletados os dados dos prontuários eletrônicos de pacientes internados em três diferentes UTIs de um hospital localizado na cidade de Belo Horizonte, entre os períodos de 31 de julho de 2016 a 31 de dezembro de 2018. Os dados coletados foram anonimizados e abrangem informações demográficas, diagnósticos, valores laboratoriais, dentre outros valores que são atualizados frequentemente pelos equipamentos de monitoramentos das UTIs. Ao todo, a coorte compreende 6000 admissões e 5474 pacientes.

Com o intuito de validar do modelo, foram coletados os dados dos prontuários eletrônicos de pacientes internados nas mesmas três UTIs, mas do período de 01 de janeiro de 2019 a 30 de abril de 2019. A coorte contém 1086 internações na UTI e 1069 pacientes distintos. O objetivo principal da validação é demonstrar que os modelos fornecem explicações com alta robustez e estabilidade.

A Tabela 3.1 mostra as características dos pacientes usando os dados de sua primeira admissão na UTI.

Já a Tabela 3.2 mostra as características dos pacientes agrupados pelas complicações.

Para a adesão ao modelo SAPS III, tanto as coortes de desenvolvimento quanto de validação são compostas por pacientes maiores de 16 anos com internação em UTI superior a 24 horas.

Tabela 3.1: Características dos pacientes da UTI

| | |
|---|---|
| Idade - anos | 65 (54 − 80) |
| Sexo | |
| Feminino | 3636 (51,32%) |
| Masculino | 3450 (48,68%) |
| Comorbidades | |
| Aids | 39 (0,5%) |
| Hipertensão | 4075 (58,1%) |
| Diabetes | 1115 (15,9%) |
| Arritmia Cardíaca | 331 (4,7%) |
| Demência | 173 (2,4%) |
| Obesidade Mórbida | 432 (6,1%) |
| Terapia de Câncer | 302 (4,3%) |
| Categoria de admissão | |
| Médica | 3939 (56,2%) |
| Cirurgia programada | 2430 (34,6%) |
| Cirurgia não programada | 656 (9,3%) |
| Tipo de cirurgia | |
| Cardíaca | 54 (0,9%) |
| Trauma | 45 (0,6%) |
| Neuro cirurgia | 25 (0,3%) |
| Tempo de internação antes da UTI - dias | 2.64 (0 − 1) |
| Tempo de internação na UTI - dias | 3.89 (1 − 4) |
| Número de admissões na UTI | |
| 1 | 5962 (85,0%) |
| 2 | 392 (5,5%) |
| 3 | 53 (0,7%) |
| $\geq 4$ | 20 (0,2%) |

### 3.1.1 Atributos e Rótulos

Os atributos usados para construir os modelos são uma mistura de informações estáticas extraídas de dados que incluem dados demográficos e diagnósticos (obtidos antes da admissão na UTI), resultados laboratoriais diários (obtidos antes da admissão na UTI) e informações dinâmicas como os sinais vitais coletados de equipamentos da UTI (durante a primeira hora após a admissão na UTI). Resumidamente, nossos dados são um subconjunto do modelo SAPS III, para que possamos vincular nosso modelo à prá-

Tabela 3.2: Características dos pacientes da UTI agrupados pelas complicações

| | **Delirium** | **VAP** | **CLABSI** | **Mortalidade** |
|---|---|---|---|---|
| **# Instâncias** | 236 | 41 | 13 | 477 |
| **Idade - Anos** | 72 (63 - 86) | 66 (58 - 80) | 55 (43 - 71) | 71,82 (63 - 84) |
| **Sexo** | | | | |
| **Masculino** | 114 | 23 | 5 | 238 |
| **Feminino** | 122 | 18 | 8 | 239 |
| **Alta** | 213 | 28 | 8 | - |
| **Óbito** | 23 | 13 | 5 | - |
| **Tempo de internação antes da UTI - dias** | 2,97 (0 - 1,25) | 3,9 (0 - 1) | 24,38 (0 - 14) | 7,44 (0 - 8) |
| **Tempo de internação na UTI - dias** | 8 (3 - 11) | 21 (12 - 27) | 24,07 (11 - 19) | 8,91 (2 - 11) |
| **Número de admissões na UTI** | | | | |
| 1 | 206 | 41 | 13 | 477 |
| 2 | 15 | 0 | 0 | 0 |
| 3 | 0 | 0 | 0 | 0 |
| $\geq 4$ | 0 | 0 | 0 | 0 |

tica clínica atual. A Figura 3.1 demonstra alguns subconjunto de atributos usados para aprendizado dos modelos.

A Figura 3.2 mostra a extração e rotulagem de recursos. O foco do trabalho foi a previsão precoce, ou seja, as previsões são realizadas durante a primeira hora de admissão na UTI. Assim, empregamos apenas os sinais vitais medidos durante a primeira hora após a admissão na UTI. Esses sinais foram agregados (*down-sampling*), resultando em valores mínimos e máximos para cada sinal. Já os desfechos consistem em possíveis complicações ocorridas em qualquer momento da internação na UTI.

A Tabela 3.3 mostra as complicações consideradas neste estudo, que são explicadas a seguir.

- Delirium é definida como uma rápida alteração de consciência podendo durar horas a dias, caracterizada pela diminuição da consciência ambiental, atenção diminuída

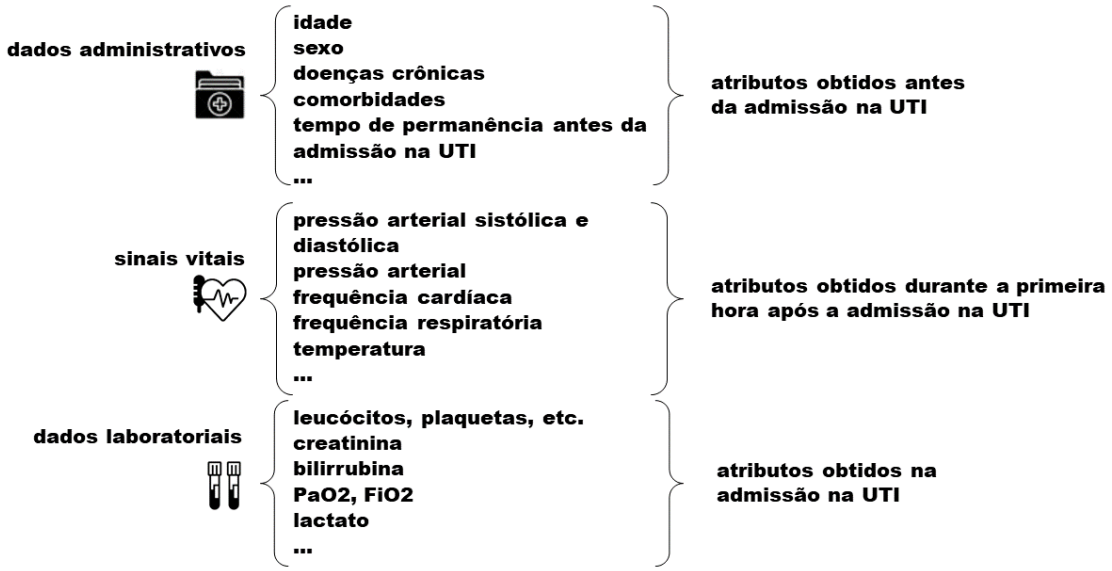


Figura 3.1: Subconjunto de atributos usados para o aprendizado dos modelos

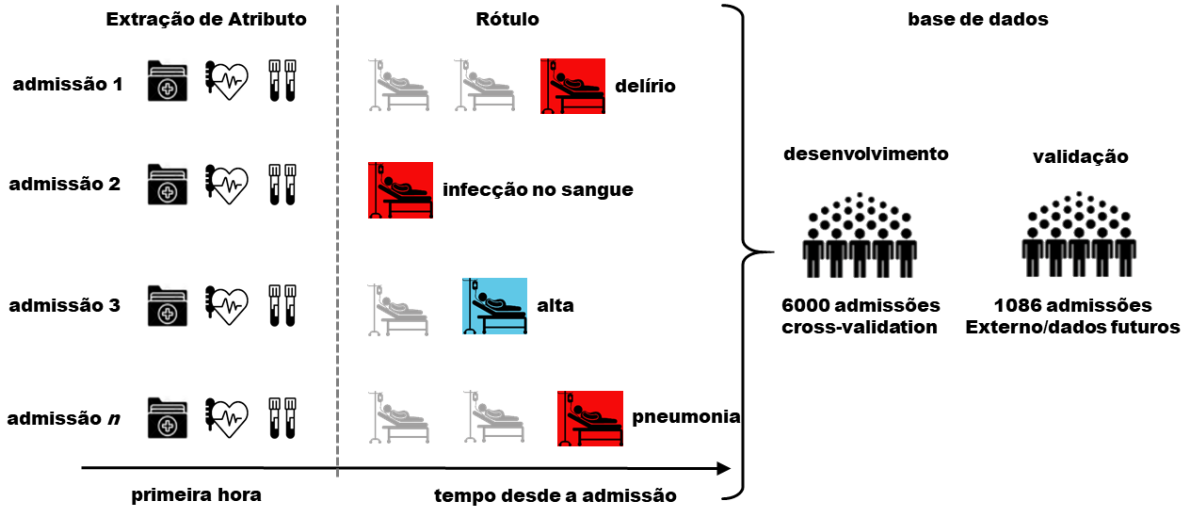


Figura 3.2: Configuração de previsão precoce: os atributos estão disponíveis na primeira hora após a admissão na UTI, mas os rótulos (complicações) podem ocorrer a qualquer momento durante a permanência na UTI. Os dados são divididos em conjuntos de desenvolvimento e validação.

e cognição alterada [6]. Essas características clínicas podem se manifestar como déficits de memória, desorientação, alucinações, níveis flutuantes de alerta e anormalidades motoras.

- Infecções no fluxo sanguíneo associadas à inserção de cateteres centrais (CLABSI):

Tabela 3.3: Complicações durante a internação na UTI. O número entre parênteses indica a fração de casos com óbito do paciente.

| | Desenvolvimento (n=6,000) | Validação (n=1,086) |
|---|---|---|
| Delirium | 147 (10%) | 89 (11%) |
| Pneumonia associada à ventilação mecânica (VAP) | 39 (31%) | 8 (50%) |
| Infecções no fluxo sanguíneo associadas à inserção de cateteres centrais (CLABSI) | 11 (27%) | 4 (50%) |
| Mortalidade | 414 | 63 |

Uma linha central é um cateter que é colocado na veia grande do paciente, geralmente no pescoço, tórax, braços ou virilha. O cateter central costuma ser usado para tirar sangue ou para administrar fluidos e medicamentos com mais facilidade. A linha pode ser deixada no local por várias semanas ou meses, se necessário. Às vezes, bactérias ou outros germes podem entrar na linha central do paciente e entrar em sua corrente sanguínea. Isso pode causar uma infecção que é chamada de infecção da corrente sanguínea associada à linha central.

- Pneumonia associada à ventilação mecânica (VAP) é um tipo de infecção pulmonar que ocorre em pacientes que usam máquinas de ventilação mecânica na UTI. A taxa de mortalidade por essa pneumonia varia de 24 a 50% e pode chegar a 76% em alguns casos específicos ou quando a infecção pulmonar é causada por patógenos de alto risco [7].

## 3.2 Desenvolvimento do Modelo

Aplicamos um algoritmo de *additive boosting* para prever a ocorrência de complicações específicas nas admissões à UTI. Especificamente, usamos o algoritmo LightGBM [41], que segue a técnica *gradient boosting* que se ajusta às *boosted decision trees* minimizando o *gradient error*. As árvores são adicionadas iterativamente ao conjunto e são adequadas para corrigir os erros de previsão feitos por árvores de decisão anteriores. O conjunto minimiza o *cross-entropy loss function* usando *gradient descent*. LightGBM fornece hiperparâmetros que devem ser ajustados, incluindo o número de árvores para compor o conjunto ($T$), a taxa de aprendizado ($\gamma$) e a profundidade máxima da árvore ($\theta$).

Amostramos o espaço do modelo selecionando aleatoriamente $k$ características do conjunto de atributos disponíveis com $2 \leq k \leq 25$, e para cada conjunto de atributos

construímos modelos usando combinações de $T$ (10, 50, 100 ), $\gamma$ (0,05, 0,1, 0,2) e $\theta$ (5, 10). Portanto, cada conjunto de atributos resultou em 18 modelos diferentes. Repetimos a exploração aleatória de conjuntos de atributos de modo que cerca de 1.000.000 de modelos foram produzidos para cada complicação.

### 3.2.1 Performance do Modelo

Avaliamos o desempenho da previsão em termos do AUROC. AUROC mede se o modelo é capaz de classificar exemplos corretamente. Leva valores entre 0,5 (previsões aleatórias) e 1 (todos os casos positivos são classificados acima dos casos negativos).

Para cada modelo avaliado, realizamos um *5-fold cross-validation* na coorte de desenvolvimento. Usamos a coorte de validação como um conjunto de dados independente no qual avaliamos os modelos construídos a partir da coorte de desenvolvimento.

### 3.2.2 Explicabilidade

Aplicamos o algoritmo SHAP a cada modelo para obter explicações sobre os atributos que conduzem as previsões específicas do paciente. SHAP é uma representação agnóstica de modelo de importância de atributos, onde o impacto de cada atributos em uma previsão particular é representado usando valores de *Shapley*. Dado o conjunto atual de valores de atributos, um valor de *Shapley* quantifica o quanto um único atributo no contexto de sua interação com outros atributos contribui para a diferença entre a previsão real e a média. Ou seja, a soma dos valores de *Shapley* para todos os atributos mais a previsão média é igual à previsão real.

É importante ressaltar que o valor de *Shapley* para um atributo não deve ser visto como seu efeito direto e isolado, mas como seu efeito composto ao interagir também com os outros atributos. Os valores de *Shapley* consideram todas as previsões possíveis para uma instância usando todas as combinações possíveis de entradas e, por causa dessa abordagem exaustiva, o SHAP pode garantir propriedades como consistência e precisão local [45]. Em resumo, o SHAP fornece um vetor de valores *Shapley* (aproximados) para cada entrada (também conhecido como vetor de explicação). Os vetores de explicação SHAP têm a mesma dimensão das entradas, e cada valor em um vetor de explicação indica a importância do recurso correspondente em uma previsão específica.

# Capítulo 4

# Regularização baseada na Robustez e Estabilidade da Explicação

Neste capítulo, definiremos os conceitos de robustez e estabilidade da explicação e apresentaremos abordagens para medi-los. Em seguida, discutiremos como empregar esses conceitos para regularização do modelo.

## 4.1 Robustez das Explicações

Robustez ($\alpha$) mede até que ponto, entradas semelhantes têm vetores de explicação semelhantes. Para medir a robustez das explicações, primeiro criamos uma matriz de similaridade a partir das entradas (pacientes). Em seguida, criamos uma matriz de similaridade paralela a partir dos vetores de explicação correspondentes. Para ambas as matrizes, a similaridade é dada em termos de distância euclidiana. Uma vez que as duas matrizes de similaridade paralelas são criadas, empregamos o coeficiente de Mantel $r$ [47], que fornece a auto correlação espacial entre duas matrizes de similaridade. O coeficiente de Mantel é um método não paramétrico que calcula a significância da correlação por meio de permutações das linhas e colunas de uma das matrizes de similaridade. Para o teste estatístico dentro do coeficiente de Mantel foi selecionado o coeficiente de correlação de Pearson, também chamado de coeficiente de correlação produto momento $r$. Sendo a faixa de $r$ entre $-1$ a $+1$, onde estar perto de $-1$ indica uma forte correlação negativa (ou seja, as explicações não são robustas) e $+1$ indica forte correlação positiva. Um valor de $r$ de 0 indica que não há correlação entre entradas e vetores de explicação. A Figura 4.1 demonstra a teoria da robustez das explicações.

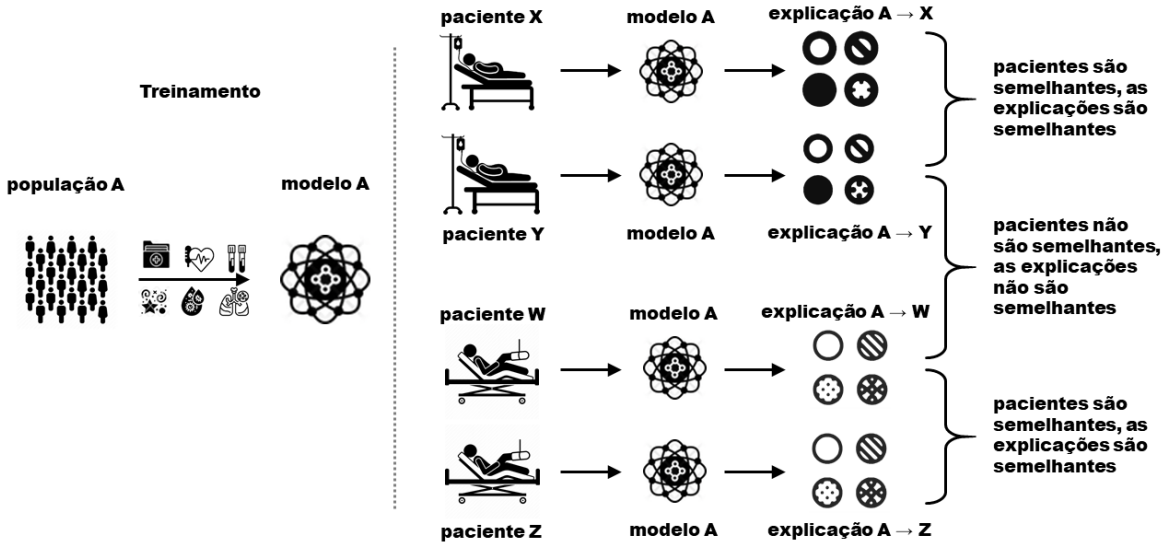


Figura 4.1: Entradas semelhantes (pacientes) são dadas ao modelo e, se as explicações correspondentes são semelhantes, a robustez do conjunto de recursos é alta.

## 4.2 Estabilidade das Explicações

A estabilidade ($\beta$) mede até que ponto a mesma entrada (pacientes) leva a vetores de explicação semelhantes quando é fornecida a modelos construídos a partir de conjuntos de treinamento com pequenas variações. Para medir a estabilidade das explicações, realizamos várias rodadas de *bootstrap* (amostramos os dados de treinamento uniformemente com substituição) e, em cada rodada, criamos uma matriz de similaridade a partir dos vetores de explicação obtidos. O processo de *bootstrapping* resulta em múltiplas matrizes de similaridade (ou seja, uma matriz por rodada), que são então comparadas com a matriz de similaridade dos vetores de explicação obtidos a partir do conjunto de treinamento original (*baseline*). Uma comparação entre as matrizes *bootstrapped* e a matriz *baseline* é novamente dada em termos do coeficiente de Mantel $r$. Como temos várias matrizes, simplesmente calculamos o valor médio de $r$. Valores médios altos de $r$ indicam que as explicações são estáveis com variações no conjunto de treinamento, enquanto valores médios baixos de $r$ indicam que as explicações variam muito com pequena variação no conjunto de treinamento. A Figura 4.2 demonstra a teoria da estabilidade das explicações.

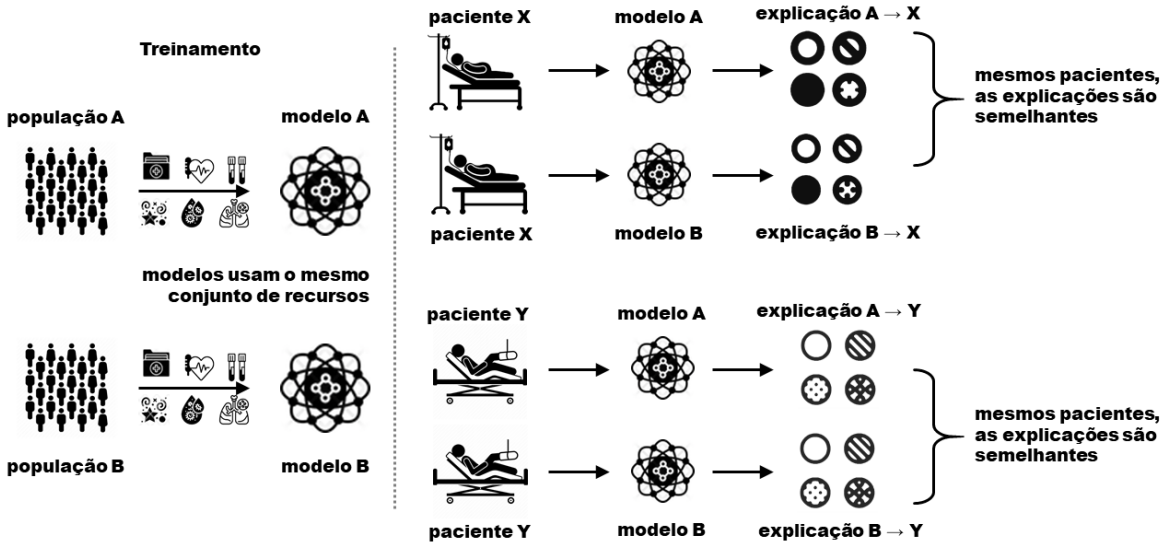


Figura 4.2: Os modelos A e B empregam o mesmo conjunto de atributos, mas são treinados em duas populações ligeiramente diferentes A e B. Então, as mesmas entradas (pacientes) são dadas aos modelos A e B, e se as explicações correspondentes forem semelhantes, então a estabilidade do conjunto de atributos é alta

## 4.3 Amostragem do Espaço do Modelo

Amostramos o espaço do modelo selecionando os atributos que compõem um modelo. Mais precisamente, começamos enumerando todos os modelos possíveis compostos de um único atributo. Em seguida, selecionamos o atributo dentro do modelo de melhor desempenho de acordo com um critério de utilidade específico e, em seguida, enumeramos todos os modelos possíveis compostos por dois atributos. O processo de enumeração do modelo continua incluindo um atributo após cada iteração, até que o modelo comece a degradar. Garantimos que nenhum atributo apareça mais de uma vez no mesmo modelo.

Durante a amostragem do espaço do modelo, queremos desencorajar o aprendizado de um modelo $f$ para o qual as explicações não são robustas nem estáveis, de modo a evitar o risco de confiar demais no mecanismo de previsão do modelo. Portanto, definimos a utilidade $U_\alpha(f)$ e $U_\beta(f)$ de um modelo $f$ como:

$$U_\alpha(f) = P + C \times \alpha, \text{ sendo } C \geq 0, \alpha > 0 \tag{4.1}$$

$$U_\beta(f) = P + C \times \beta, \text{ sendo } C \geq 0, \beta > 0 \tag{4.2}$$

em que $P$ é uma medida de desempenho de previsão (AUROC) e $C$ é o coeficiente de regularização. Valores mais baixos de $C$ encorajarão modelos de amostragem com de-

sempenho de previsão aparentemente alto, não importa quão robustas e estáveis sejam as explicações fornecidas. Valores maiores de $C$, por outro lado, podem levar a modelos com baixo desempenho de previsão. Valores apropriados de $C$ podem levar a modelos de alto desempenho com explicações robustas e estáveis e, portanto, esses modelos são mais propensos a funcionar como originalmente projetados. Desconsideramos qualquer modelo com robustez negativa ou estabilidade negativa.

A Figura 4.3 apresenta uma visão geral do processo de regularização proposto. Conforme explicado anteriormente, os atributos obtidos dos dados do paciente, sinais vitais e resultados laboratoriais estão disponíveis na primeira hora após a admissão na UTI. Portanto, as previsões de rótulos (complicações) estão disponíveis na primeira hora após a admissão na UTI, mas podem ocorrer a qualquer momento durante a permanência na UTI. Em seguida, os atributos são selecionados iterativamente, maximizando a robustez ou a estabilidade do modelo até que o modelo de melhor desempenho seja selecionado. Por fim, o modelo final é utilizado para a identificação precoce de pacientes com risco de complicações.

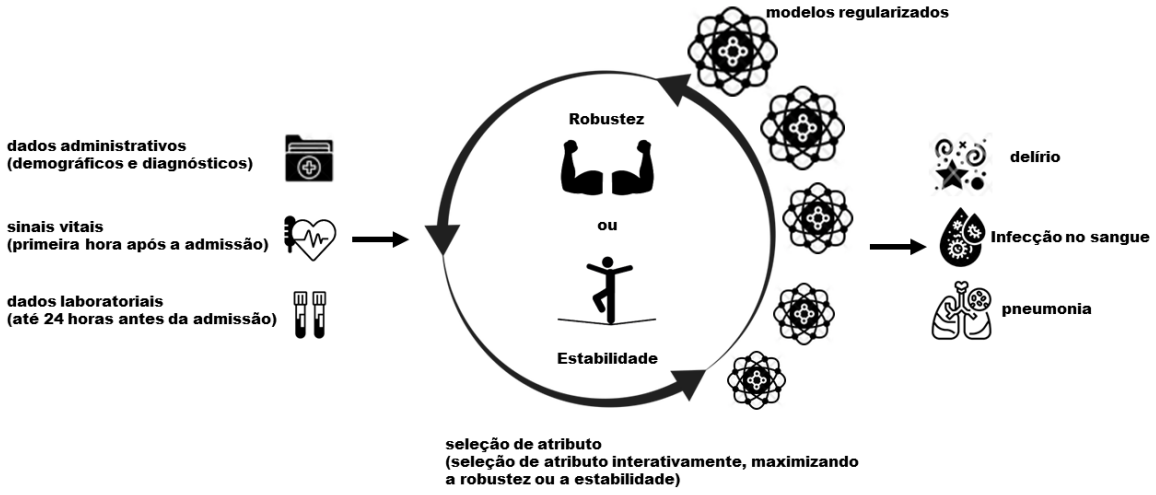


Figura 4.3: O processo de regularização proposto. Os atributos são selecionados de forma iterativa, maximizando a robustez ou a estabilidade do modelo. O modelo final é usado para prever as complicações.

# Capítulo 5

# Experimentos

Neste Capítulo serão apresentados os resultados para a identificação precoce de pacientes na UTI com riscos de complicações. Em particular, os experimentos têm também como objetivos responder as seguintes perguntas:

**1:** Existe um *trade-off* entre a estabilidade, robustez da explicação e o desempenho da predição?

**2:** Podemos melhorar o desempenho dos modelos usando estabilidade e robustez da explicação como termos de regularização?

## 5.1 Base de Dados

A Tabela 5.1 mostra as características dos pacientes nos conjuntos de dados de desenvolvimento e validação usando os dados de sua primeira admissão na UTI. No conjunto de dados de desenvolvimento, a idade média dos pacientes é de 65 anos (intervalo interquartil - IQR $54 - 80$) e 3.074 (51,2%) mulheres.

## 5.2 Configuração dos Experimentos

Realizamos o *cross-validation* com cinco *folds* utilizando todo o conjunto de dados de desenvolvimento, ou seja, os dados são organizados em cinco *folds* (cada uma contendo 1200 instâncias), onde em cada execução, quatro *folds* são usadas como conjunto de treinamento ($n = 4800$), e o *folds* restante é usado como conjunto de teste ($n = 1200$). O resultado apresentado do AUC é a média das cinco execuções. A robustez e a estabilidade da explicação também são calculadas pelas médias das *folds*. Todo o processo foi executado

Tabela 5.1: Características dos pacientes da UTI dos dados de desenvolvimento e validação

| | Desenvolvimento ($n = 6,000$) | Validação ($n = 1,086$) |
|---|---|---|
| Idade - anos | 65 (54 − 80) | 65 (53 − 80) |
| Sexo | | |
| Feminino | 3,074 (51,2%) | 562 (51,7%) |
| Masculino | 2,926 (48,8%) | 524 (48,3%) |
| Comorbidades | | |
| Aids | 32 (0,005%) | 7 (0,007%) |
| Hipertensão | 3543 (59,0%) | 532 (52,1%) |
| Diabetes | 919 (15,3%) | 196 (19,5%) |
| Arritmia Cardíaca | 304 (5,1%) | 27 (2,6%) |
| Demência | 528 (8,8%) | 145 (14,2%) |
| Obesidade Mórbida | 373 (6,2%) | 59 (5,6%) |
| Terapia de Câncer | 212 (3,5%) | 90 (8,6%) |
| Categoria de admissão | | |
| Médica | 3275 (54,6%) | 664 (61,1%) |
| Cirurgia programada | 2158 (36,0%) | 272 (25,0%) |
| Cirurgia não programada | 553 (9,4%) | 103 (9,5%) |
| Tipo de cirurgia | | |
| Cardíaca | 53 (0,9%) | 1 (0,1%) |
| Trauma | 42 (0,7%) | 3 (0,3%) |
| Neuro cirurgia | 24 (0,4%) | 1 (0,1%) |
| Tempo de internação antes da UTI - dias | 2,76 (0 − 1) | 2,00 (0 − 1) |
| Tempo de internação na UTI - dias | 3,87 (1 − 4) | 4,08 (2 − 5) |
| Número de admissões na UTI | | |
| 1 | 5034 (92,1%) | 928 (96,5%) |
| 2 | 361 (6,6%) | 31 (3,2%) |
| 3 | 51 (0,9%) | 2 (0,02%) |
| $\geq 4$ | 19 (0,4%) | 1 (0,01%) |

separadamente para cada rótulo: Delirium, Infecções no fluxo sanguíneo associadas à inserção de cateteres centrais (CLABSI), Pneumonia associada à ventilação mecânica (VAP) e Mortalidade. Também avaliamos o desempenho dos modelos em um conjunto de dados de validação separado ($n = 1086$) como um conjunto de dados independente (dados futuros), no qual os modelos são construídos usando todo o conjunto de dados de desenvolvimento ($n = 6000$).

## 5.3 Avaliação do Modelo

Avaliamos a performance dos preditores utilizando a *Area Under the Receiver Operating Characteristic Curve* (AUROC). Ela mede se o modelo é capas de classificar as instâncias corretamente, obtendo valores entre $0,5$ para predições aleatórias e 1 caso o preditor tenha uma boa capacidade de separabilidade entre classes.

## 5.4 Resultados e Discussão

Para responder a primeira pergunta, amostramos o espaço do modelo (cerca de um milhão de modelos para cada complicação), para que possamos compreender a relação entre a robustez da explicação, estabilidade e o desempenho preditivo, a Figura 5.1 mostra essa relação entre os modelos treinados para prever delirium. Claramente, o desempenho preditivo aumenta tanto com a robustez quanto com a estabilidade da explicação, e os modelos de melhor desempenho são aqueles localizados no canto superior direito. Ao comparar os dois *heatmaps*, encontramos diferenças nos valores AUROC quando a estabilidade da explicação varia de 0,3 a 0,5 e a robustez da explicação varia de 0,8 a 0,9. Ainda assim, o desempenho da predição atinge os valores mais altos quando a robustez e a estabilidade da explicação são maiores. Especificamente, os valores AUROC chegam a 0,88 com validação cruzada usando o conjunto de dados de desenvolvimento e a 0,85 no conjunto de dados de validação externa.

Tendências semelhantes são observadas ao analisar modelos treinados para prever as outras complicações direcionadas como VAP (os resultados são mostrados na Figura 5.2), CLABSI (os resultados são mostrados na Figura 5.3), e mortalidade (os resultados são mostrados na Figura 5.4). Em todos os casos, o desempenho preditivo aumenta tanto com a robustez da explicação quanto com a estabilidade da explicação. Em resumo, os modelos que VAP têm valores AUROC de até 0,92 no *cross-validation* e nos dados de validação. Os modelos que preveem a CLABSI têm valores AUROC tão altos quanto 0,88 no *cross-validation* e 0,85 nos dados de validação. Finalmente, os modelos de previsão de mortalidade têm valores AUROC de até 0,84 no *cross-validation* e 0,83 nos dados de validação.

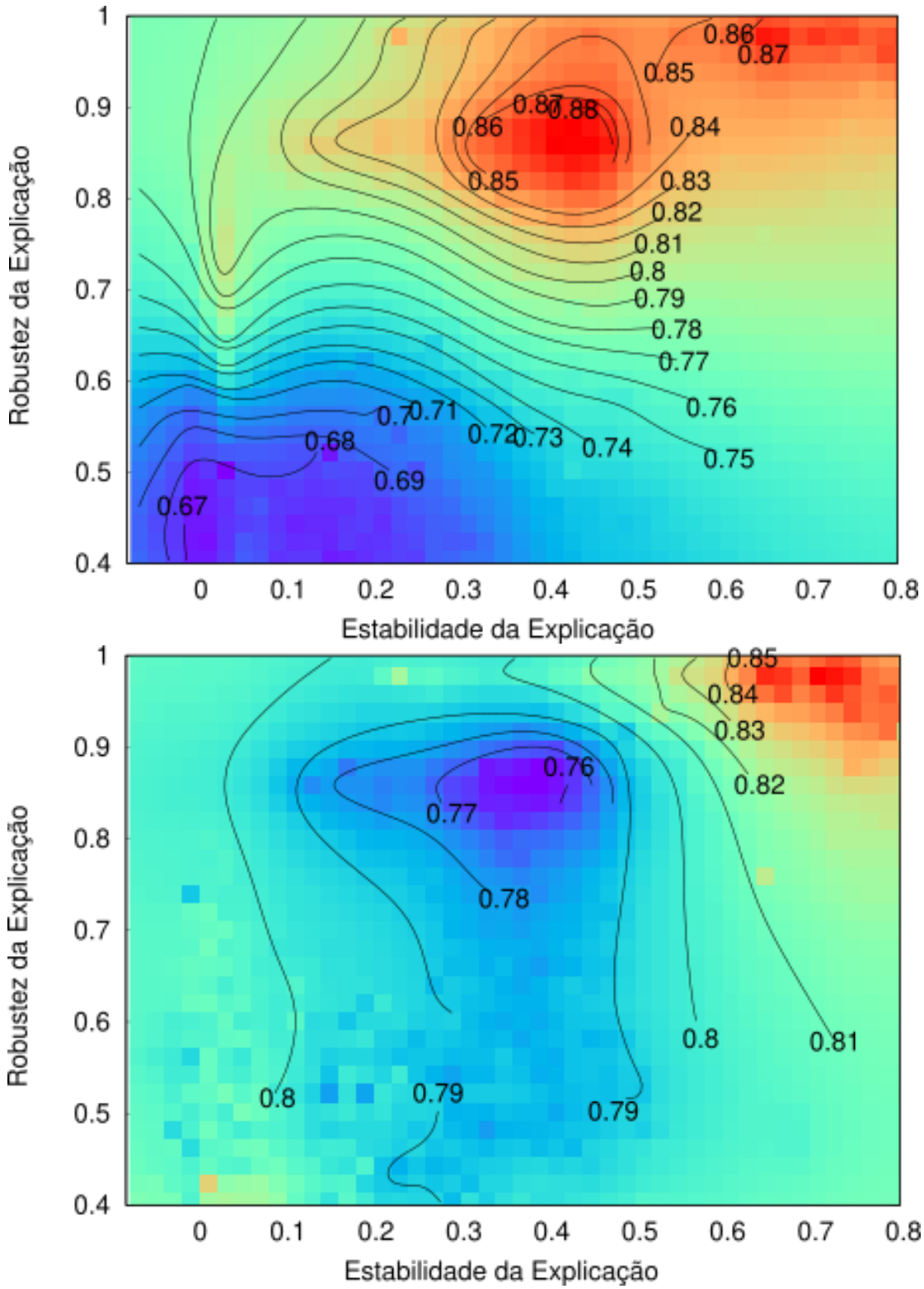


Figura 5.1: *Heatmap* dos modelos que preveem Delirium. A cor indica a distribuição dos valores AUROC para aos modelos: o azul está associado a valores baixos do AUROC enquanto o vermelho está associado com valores mais altos. Superior − dados de treinamento usando o *cross-validation*. Inferior − dados de validação.

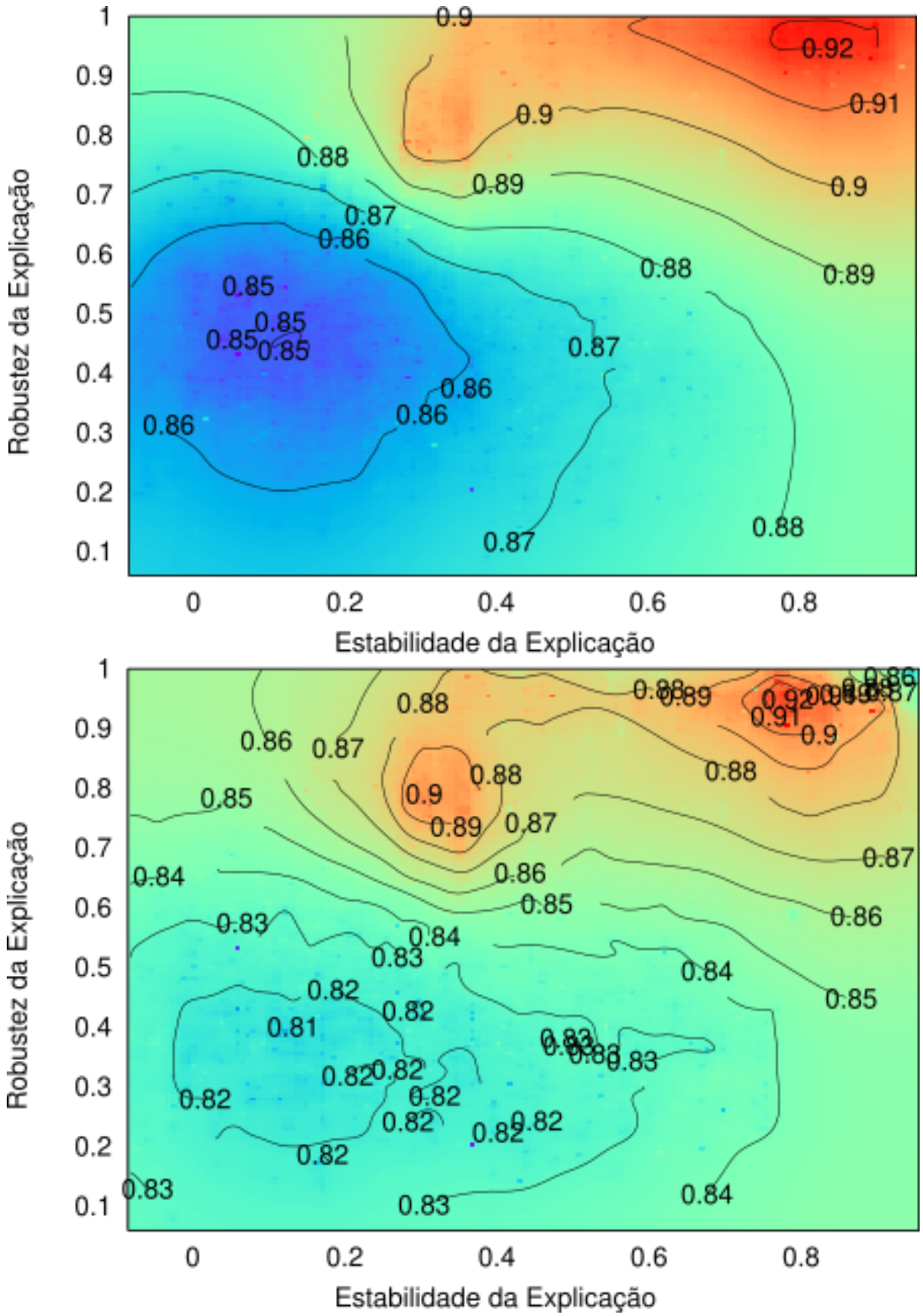


Figura 5.2: *Heatmap* dos modelos que preveem VAP. A cor indica a distribuição dos valores AUROC para aos modelos: o azul está associado a valores baixos do AUROC enquanto o vermelho está associado com valores mais altos. Superior − dados de treinamento usando o *cross-validation*. Inferior − dados de validação.

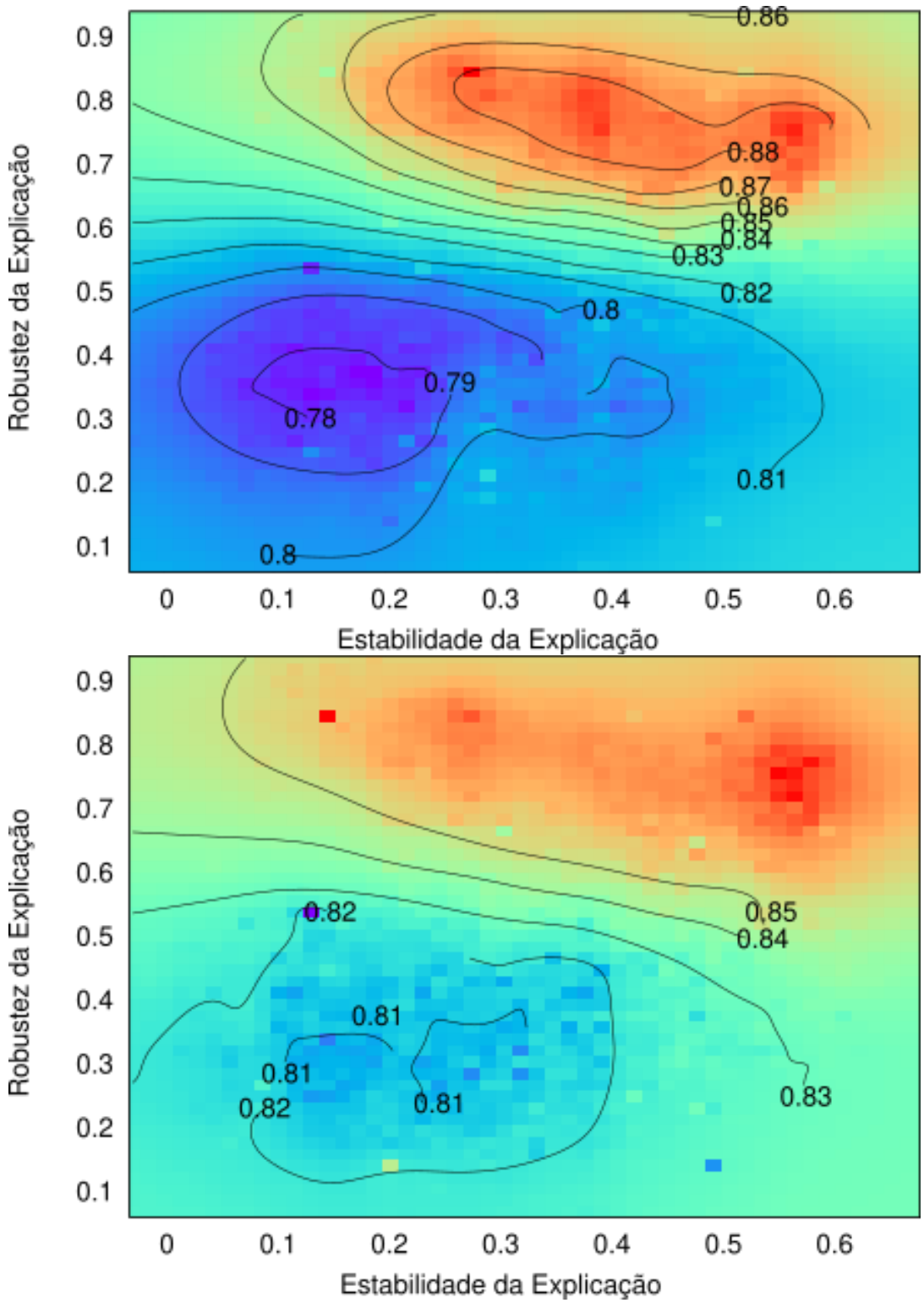


Figura 5.3: *Heatmap* dos modelos que preveem CLABSI. A cor indica a distribuição dos valores AUROC para aos modelos: o azul está associado a valores baixos do AUROC enquanto o vermelho está associado com valores mais altos. Superior − dados de treinamento usando o *cross-validation*. Inferior − dados de validação.

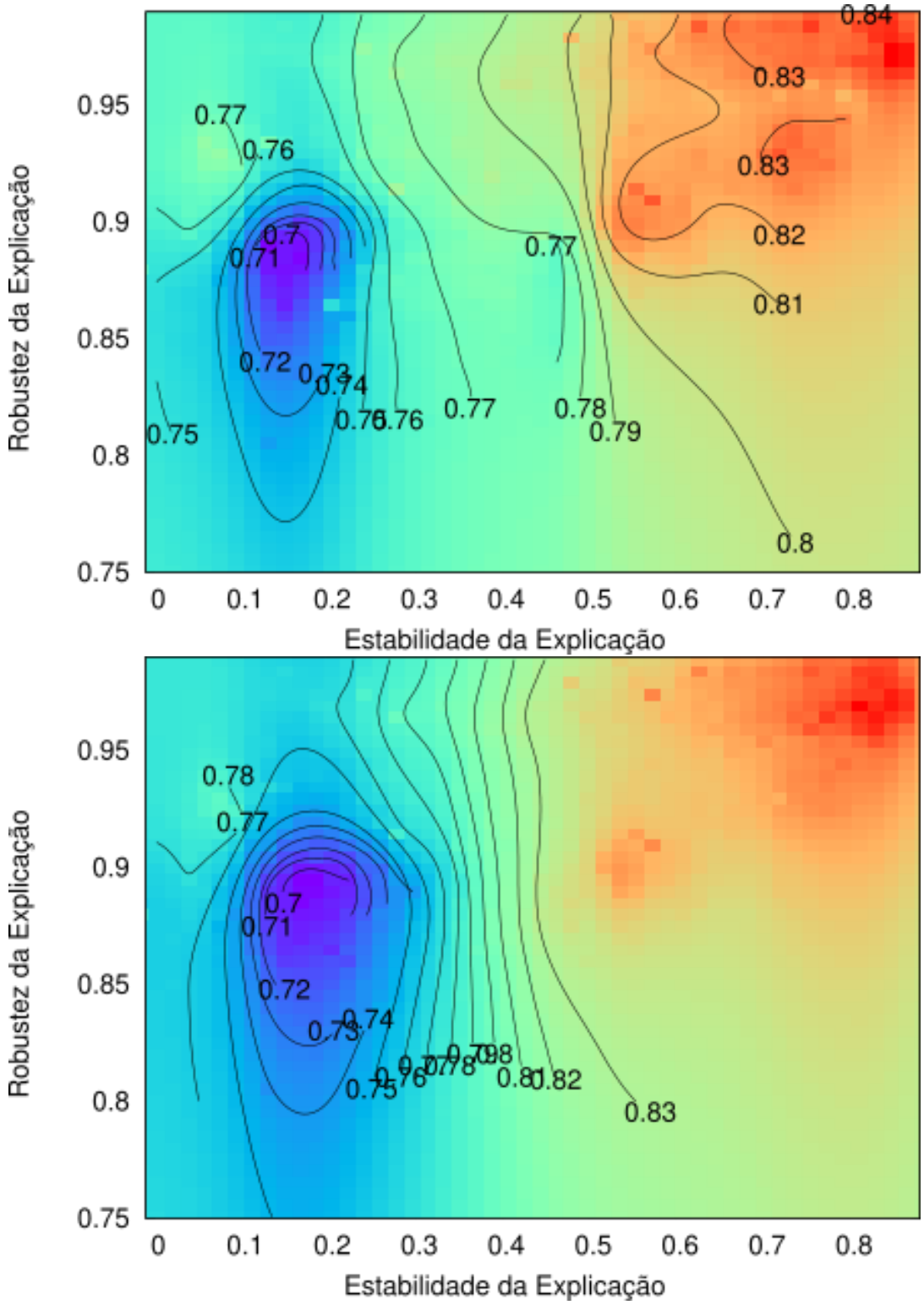


Figura 5.4: *Heatmap* dos modelos que preveem Mortalidade. A cor indica a distribuição dos valores AUROC para aos modelos: o azul está associado a valores baixos do AUROC enquanto o vermelho está associado com valores mais altos. Superior − dados de treinamento usando o *cross-validation*. Inferior − dados de validação.

Em uma analise mais aprofundada, nota-se que durante o processo de seleção de atributos (*feature selection*) para compor o modelo de predição, o algoritmo seleciona

atributos que o leva a encontrar ótimos locais, não conseguindo sair destas regiões. Porém quando é utilizado regularizadores baseados na explicação ou robustez, o algoritmo consegue selecionar atributos que trazem um maior ganho ao modelo, fazendo que ele saia dos ótimos locais, buscando melhores soluções para que consiga performar melhor. Essa análise é válida em todos os experimentos que foram realizados, como podem ser observadas nas Figuras 5.5, 5.6 e 5.7.

A Figura 5.5 demonstra os atributos selecionados, não utilizando de regularizadores, para a criação do modelo. Esses atributos foram ordenados pelos seus respectivos valores de Shapley.

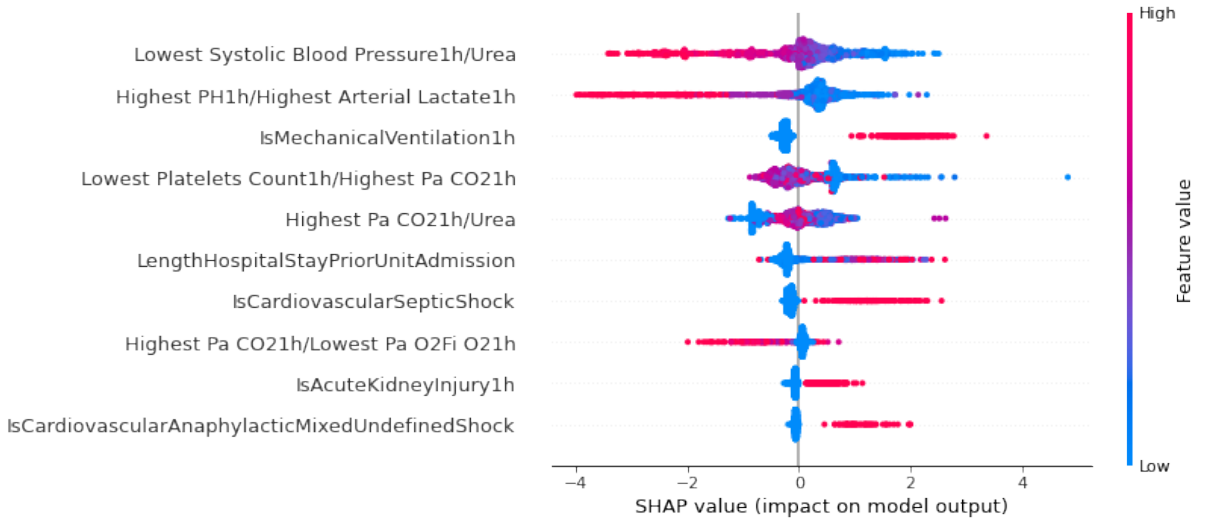


Figura 5.5: Classificação dos atributos de acordo com a soma das magnitudes - SHAP para a predição da mortalidade

Na figura 5.6 durante o processo de criação do modelo de predição, foi feito o uso da regularizador baseado na explicação. Pode-se notar que há alguns atributos que foram selecionados em ambos os modelos, porém há diferenças entre os atributos que os modelos selecionaram.

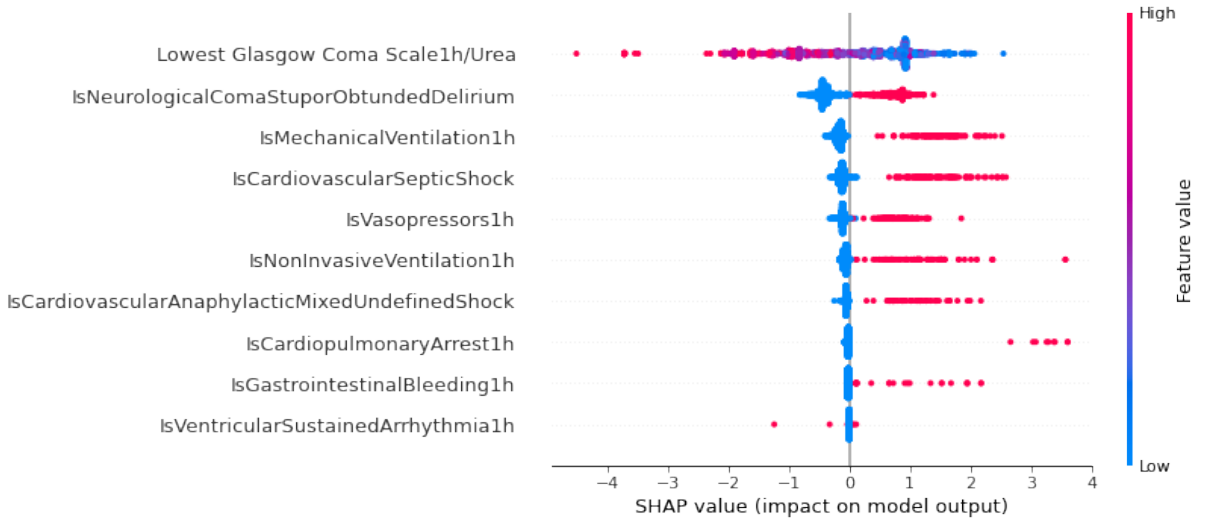


Figura 5.6: Classificação dos atributos de acordo com a soma das magnitudes - SHAP para a predição da mortalidade, utilizando a regularização baseada na estabilidade

Por fim, a Figura 5.7 mostra que *Highest Pa O21h/Highest Arterial Lactate1h* foi o atributo com o maior valor Shapley para o modelo criado utilizando o regularizador baseado na robustez e como mostrado nas figuras anteriores, esse modelo tem alguns atributos que são compartilhados com o modelo que não utiliza regularizador e com o modelo que utiliza o regularizador baseado na estabilidade.

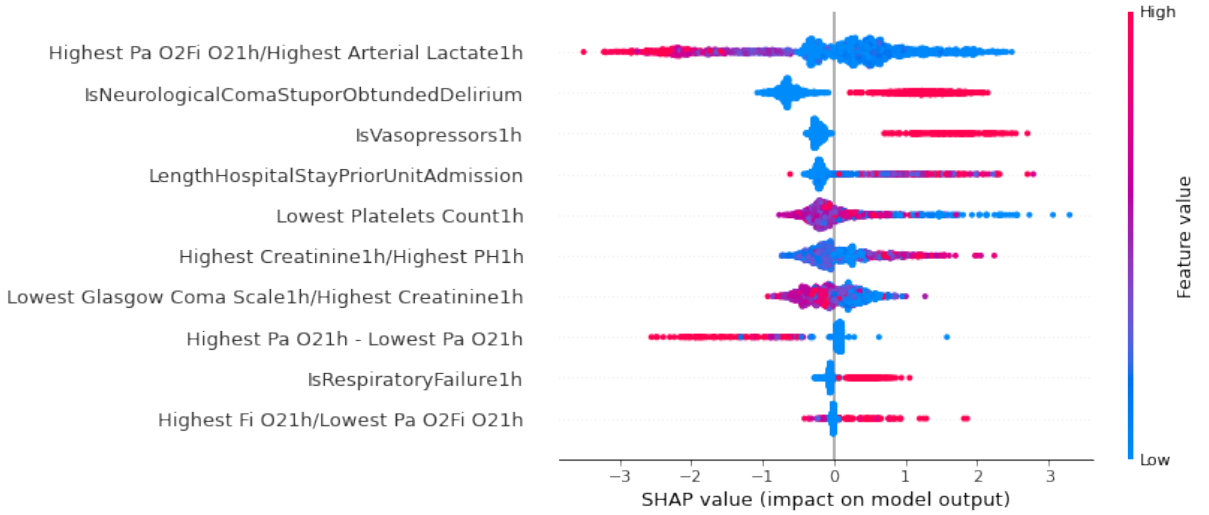


Figura 5.7: Classificação dos atributos de acordo com a soma das magnitudes - SHAP para a predição da mortalidade, utilizando a regularização baseada na robustez

Para responder a segunda pergunta, amostramos o espaço do modelo conforme descrito na Seção 4.3. Especificamente, selecionamos iterativamente os recursos para compor o modelo. A cada iteração, o recurso selecionado é aquele que fornece utilidade máxima. Variando o coeficiente de regularização $C$, controlamos a importância que os termos de regularização terão no processo de seleção de características. Para $C = 0$, a utilidade

é dada apenas em termos de uma medida de desempenho preditiva (ou seja, AUROC), ou seja, os recursos são incorporados ao modelo simplesmente maximizando AUROC. À medida que $C$ aumenta, os recursos selecionados tendem a aumentar a estabilidade da explicação e a robustez do modelo resultante. Variamos o coeficiente de regularização $C$ de 0 a 0,5 e relatamos os resultados em termos de AUROC. Além disso, consideramos o método introduzido por [36] a fim de fornecer uma comparação de *baseline*, no qual utiliza de uma LightGBM para realizar a predição precoce de falha no sistema circulatório.

A Tabela 5.2 mostra os números de desempenho dos modelos que preveem o Delirium. Nela, facilmente pode ser observado que a seleção de recursos baseados apenas em AUROC não leva aos melhores modelos, mas os termos de regularização propostos desempenham um papel importante na seleção de recursos. Os melhores resultados, foram obtidos com $C = 0,3$. É interessante ressaltar que o desempenho nos conjuntos de dados de desenvolvimento e validação se aproximam à medida que $C$ aumenta, sugerindo que a estabilidade e a robustez da explicação são úteis para a generalização do modelo. Além disso, o desempenho obtido com $C = 0,3$ supera em muito o desempenho de previsão do *baseline*.

Tabela 5.2: Modelos Preditores de Delirium − Performance preditiva com a variação de valores de $C$.

| | Estabilidade | | Robustez | |
|---|---|---|---|---|
| $C$ | Desenvolvimento | Validação | Desenvolvimento | Validação |
| 0,0 | 0,79 ↓ (0,01) | 0,77 ↓ (0,01) | 0,79 ↓ (0,01) | 0,77 ↓ (0,01) |
| 0,1 | 0,82 ↑ (0,02) | 0,82 ↑ (0,04) | 0,80 ↕ (0,00) | 0,79 ↑ (0,01) |
| 0,2 | 0,84 ↑ (0,04) | 0,84 ↑ (0,06) | 0,82 ↑ (0,02) | 0,80 ↑ (0,02) |
| 0,3 | 0,86 ↑ (0,06) | 0,85 ↑ (0,07) | 0,84 ↑ (0,04) | 0,82 ↑ (0,04) |
| 0,4 | 0,82 ↑ (0,02) | 0,83 ↑ (0,05) | 0,83 ↑ (0,03) | 0,81 ↑ (0,03) |
| 0,5 | 0,82 ↑ (0,02) | 0,80 ↑ (0,02) | 0,81 ↑ (0,01) | 0,80 ↑ (0,02) |
| *baseline* | 0,80 | 0,78 | 0,80 | 0,78 |

As Tabelas 5.3, 5.4 e 5.5 mostram a mesma análise para modelos treinados para prever VAP, CLABSI e Mortalidade, respectivamente. A mesma tendência é observada na previsão dessas complicações, onde são obtidos utilizando valores moderados de $C$. Valores mais altos podem selecionar recursos que melhoram a estabilidade e a robustez da explicação, mas são fracos em termos de aumento de AUROC.

Claramente, os termos de regularização propostos são altamente eficazes na seleção de recursos que produzem modelos com alta generalização para todas as complicações consideradas. Finalmente, é importante notar que a estabilidade e a robustez da explicação são regularizadores altamente eficazes, mesmo quando usados em árvores de aumento de gradiente, que empregam outros tipos de regularização, como L1 e L2.

A fim de compreender melhor o impacto do uso de estabilidade e robustez da explicação como regularizadores durante a seleção de recursos, plotamos um *heatmap*

Tabela 5.3: Modelos Preditores de VAP − Performance preditiva com a variação de valores de $C$.

| | Estabilidade | | Robustez | |
|---|---|---|---|---|
| $C$ | Desenvolvimento | Validação | Desenvolvimento | Validação |
| 0,0 | 0,86 ↓ (0,02) | 0,84 ↓ (0,02) | 0,86 ↓ (0,02) | ↓ 0,84 (0,02) |
| 0,1 | 0,88 ↕ (0,00) | 0,86 ↕ (0,00) | 0,86 ↓ (0,02) | ↓ 0,84 (0,02) |
| 0,2 | 0,90 ↑ (0,02) | 0,90 ↑ (0,04) | 0,88 ↕ (0,00) | ↑ 0,87 (0,01) |
| 0,3 | 0,92 ↑ (0,04) | 0,90 ↑ (0,04) | 0,90 ↑ (0,02) | ↑ 0,88 (0,02) |
| 0,4 | 0,90 ↑ (0,02) | 0,90 ↑ (0,04) | 0,89 ↑ (0,01) | ↑ 0,88 (0,02) |
| 0,5 | 0,88 ↕ (0,00) | 0,88 ↑ (0,02) | 0,88 ↕ (0,00) | ↕ 0,86 (0,00) |
| *baseline* | 0,88 | 0,86 | 0,88 | 0,86 |

Tabela 5.4: Modelos Preditores de CLABSI − Performance preditiva com a variação de valores de $C$.

| | Estabilidade | | Robustez | |
|---|---|---|---|---|
| $C$ | Desenvolvimento | Validação | Desenvolvimento | Validação |
| 0,0 | 0,82 ↓ (0,02) | 0,79 ↓ (3) | 0,82 ↓ (0,02) | 0,79 ↓ (0,03) |
| 0,1 | 0,83 ↓ (0,01) | 0,81 ↓ (0,01) | 0,84 ↕ (0,00) | 0,82 ↕ (0,00) |
| 0,2 | 0,86 ↑ (0,02) | 0,84 ↑ (0,02) | 0,86 ↑ (0,02) | 0,84 ↑ (0,02) |
| 0,3 | 0,88 ↑ (0,04) | 0,86 ↑ (0,04) | 0,88 ↑ (0,04) | 0,85 ↑ (0,03) |
| 0,4 | 0,87 ↑ (0,03) | 0,85 ↑ (0,03) | 0,88 ↑ (0,04) | 0,85 ↑ (0,03) |
| 0,5 | 0,85 ↑ (0,01) | 0,85 ↑ (0,03) | 0,85 ↑ (0,01) | 0,83 ↑ (0,01) |
| *baseline* | 0,84 | 0,82 | 0,84 | 0,82 |

Tabela 5.5: Modelos Preditores de Mortalidade − Performance preditiva com a variação de valores de $C$.

| | Estabilidade | | Robustez | |
|---|---|---|---|---|
| $C$ | Desenvolvimento | Validação | Desenvolvimento | Validação |
| 0,0 | 0,81 ↓ (0,01) | 0,77 ↓ (0,03) | 0,81 ↓ (0,01) | 0,77 ↓ (0,03) |
| 0,1 | 0,82 ↕ (0,00) | 0,80 ↕ (0,00) | 0,82 ↕ (0,00) | 0,80 ↕ (0,00) |
| 0,2 | 0,83 ↑ (0,01) | 0,82 ↑ (0,02) | 0,84 ↑ (0,02) | 0,82 ↑ (0,02) |
| 0,3 | 0,86 ↑ (0,04) | 0,85 ↑ (0,05) | 0,86 ↑ (0,04) | 0,84 ↑ (0,04) |
| 0,4 | 0,84 ↑ (0,02) | 0,82 ↑ (0,02) | 0,84 ↑ (0,02) | 0,84 ↑ (0,04) |
| 0,5 | 0,81 ↓ (0,01) | 0,80 ↕ (0,00) | 0,80 ↓ (0,02) | 0,78 ↓ (0,02) |
| *baseline* | 0,82 | 0,80 | 0,82 | 0,80 |

mostrando os limites de decisão para modelos que preveem mortalidade. Basicamente, representamos cada ponto (ou seja, um paciente) usando os valores de recursos correspondentes e, em seguida, usamos t-SNE [82] para visualizar os dados em duas dimensões. A Figura 5.8 (superior) mostra o limite de decisão para o melhor modelo obtido com $C = 0$, enquanto a Figura 5.8 (inferior) mostra os melhores modelos obtidos com $C = 0,3$. Os pontos em vermelho correspondem a pacientes que morreram durante a internação na UTI. Curiosamente, os recursos selecionados usando $C = 0,3$ produziram um modelo que é muito mais homogêneo no sentido de que melhora a separabilidade dos diferentes

resultados.

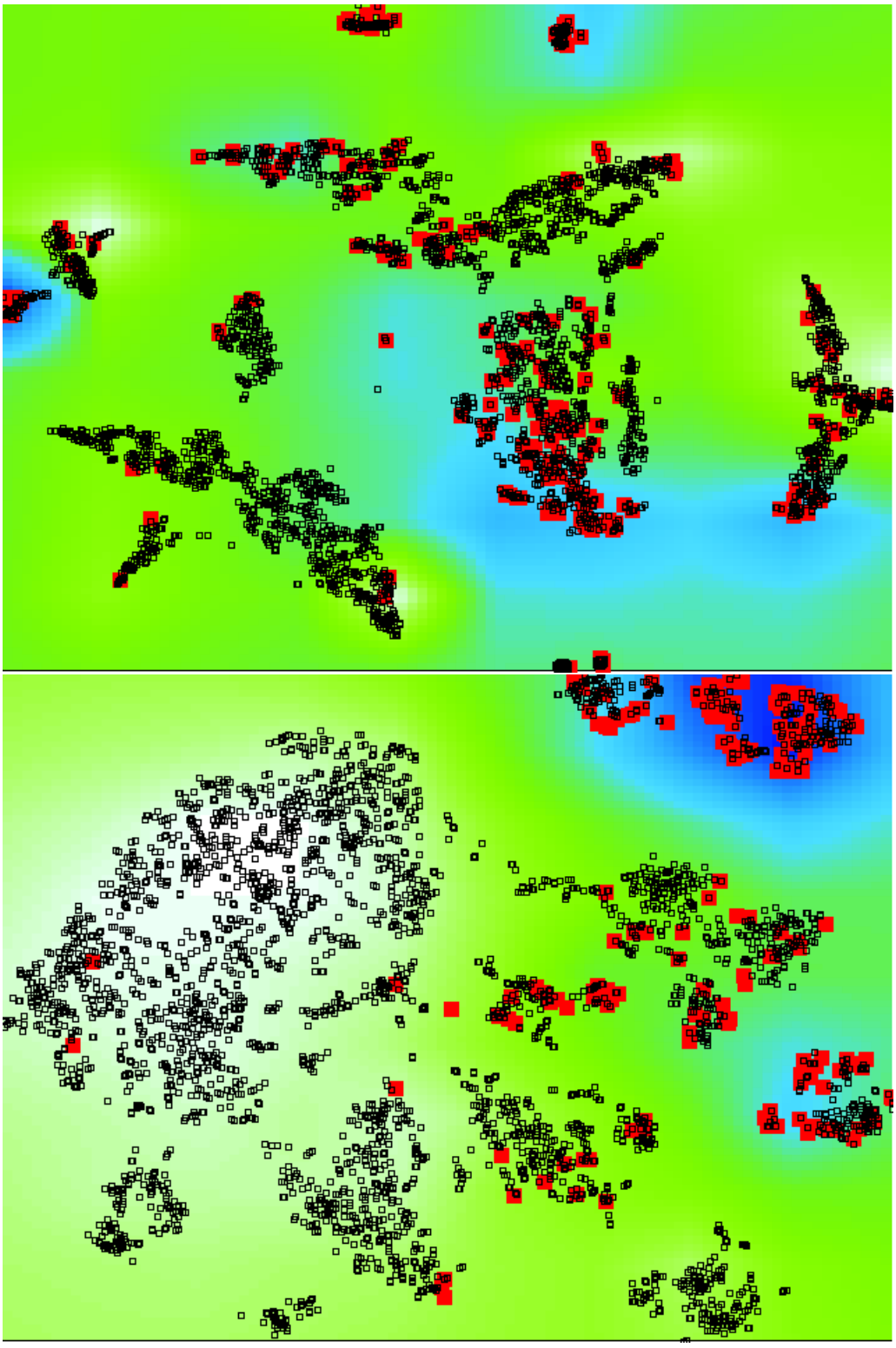

Figura 5.8: Espaço de decisão dos modelos de previsão da mortalidade. Superior − A utilidade do modelo é fornecida exclusivamente pela AUROC. Inferior − A utilidade do modelo é dada pela Estabilidade da Explicação (Equação 4.2.)

# Capítulo 6

# Conclusão e Trabalhos Futuros

Este trabalho apresentou um algoritmo para predição de riscos de complicações e mortes na primeira hora após a admissão na UTI, no qual foram utilizados técnicas de aprendizado de máquina. Neste capítulo serão resumidos os resultados já obtidos e apresentar os trabalhos futuros para a continuação

## 6.1 Principais Resultados

Essa tese tem como objetivo o estudo e a criação de modelos de aprendizado de máquina capazes de predizer precocemente se complicações ocorrerão durante o período em que o paciente está na UTI, para isso são utilizados atributos obtidas a partir dos dados administrativos, resultados laboratoriais e sinais vitais do paciente coletados na primeira hora da sua admissão.

Vem se tornando comum o uso de modelos de predição para auxiliar na tomada de decisão na UTI, um problema comum a alguns preditores é conhecido como problema da caixa preta, que não fornecem a informação da lógica envolvida nas predições dos pacientes. Atualmente já existem técnicas como o SHAP, que consegue analisar o modelo, executando-o com os dados de teste, gerando explicações do seu funcionamento.

Um outro problema é que uma vez o modelo gerado, sejamos capazes de garantir o seu funcionamento mantendo a mesma lógica de predição originalmente pretendida. Assim, para construir modelos que continuem a funcionar como originalmente projetados, primeiro propomos medir (i) como as explicações fornecidas variam para diferentes entradas (ou seja, robustez) e (ii) como as explicações fornecidas mudam com modelos construídos a partir de diferentes subpopulações de pacientes (isto é, estabilidade). Posteriormente, empregamos essas medidas como termos de regularização que são acoplados a um procedimento de seleção de atributos de modo que o modelo final forneça previsões com explicações mais robustas e estáveis.

Os experimentos foram conduzidos em um conjunto de dados contendo 6.000 in-

ternações na UTI de 5474 pacientes. Os resultados obtidos em uma coorte de validação externa de 1069 pacientes com 1086 internações em UTI mostraram que a seleção de atributos com base na robustez levou a ganhos em termos de poder preditivo que variaram de 6,8% a 9,4%, enquanto a seleção de atributos com base na estabilidade levou a ganhos que variaram de 7,2% a 11,5%, dependendo da complicação.

Nossos resultados são de importância prática, pois nossos modelos preveem complicações com grande antecipação, facilitando intervenções oportunas e protetoras.

## 6.2 Trabalhos Futuros

- Sabendo que o tempo de permanência no hospital é um dos principais indicadores, sendo utilizado na gerência para a obtenção do melhor uso dos recursos hospitalares, pretendemos desenvolver um modelo de predição em tempo real do tempo de permanência de um paciente na UTI e/ou no hospital;
- Pretende-se expandir o trabalho para predizer outras complicações durante a estadia do paciente na UTI.
- Um outro ponto importante que queremos abordar futuramente é a predição dinâmica da mortalidade na UTI, assim será possível avaliar o risco do paciente a cada intervenção que for realizada.
- Pretendemos desenvolver modelos que indiquem os recursos que serão necessários para o tratamento de grupos específicos de pacientes e ao longo do processo melhorar o desempenho do modelo com base nos métodos de regularização, enquanto incluímos novos recursos provenientes da evolução dos pacientes na UTI.

## 6.3 Publicações

- AMADOR, T., SATURNINO, S., VELOSO, A., ZIVIANI, N.. Early identification of ICU patients at risk of complications: Regularization based on robustness and stability of explanations. ARTIFICIAL INTELLIGENCE IN MEDICINE, v. 128, p. 102283, 2022.

- ALBUQUERQUE, A. ; COELHO, T. A. ; FERREIRA, R. ; VELOSO, A. ; ZIVIANI, N. . Learning to Rank with Deep Autoencoder Features. In: IJCNN - IEEE International Joint Conference on Neural Networks, 2018, Rio de Janeiro. IJCNN - IEEE International Joint Conference on Neural Networks 2018, 2018.

# Referências


[1] A. Adadi and M. Berrada. Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access*, 6:52138–52160, 2018.

[2] T. Alves, A. Laender, A. Veloso, and N. Ziviani. Dynamic prediction of ICU mortality risk using domain adaptation. In *Proc. of Big Data*, pages 1328–1336, 2018.

[3] Stephanie Baker, Wei Xiang, and Ian Atkinson. Hybridized neural networks for non-invasive and continuous mortality risk assessment in neonates. *Computers in Biology and Medicine*, 134:104521, July 2021.

[4] S. Barbieri, J. Kemp, O. Perez-Concha, S. Kotwal, M. Gallagher, A. Ritchie, and L. Jorm. Benchmarking deep learning architectures for predicting readmission to the ICU and describing patients-at-risk. *Scientific reports*, 10(1):1–10, 2020.

[5] A. Beam and I. Kohane. Big data and machine learning in health care. *JAMA*, 319(13):1317–1318, 2018.

[6] A Burns, A Gallagley, and J Byrne. Delirium. *Journal of Neurology, Neurosurgery & Psychiatry*, 75(3):362–367, 2004.

[7] Jean Chastre and Jean-Yves Fagon. Ventilator-associated pneumonia. *American journal of respiratory and critical care medicine*, 165(7):867–903, 2002.

[8] Min Chen, Yixue Hao, Kai Hwang, Lu Wang, and Lin Wang. Disease prediction by machine learning over big data from healthcare communities. *Ieee Access*, 5:8869–8879, 2017.

[9] Zhenyu Chen, Jiahao Li, and Yuzhi Xu. Machine learning-assisted high-throughput semi-empirical search of ofet molecular materials. *arXiv preprint arXiv:2107.02613*, 2021.

[10] Fu-Yuan Cheng, Himanshu Joshi, Pranai Tandon, Robert Freeman, David L Reich, Madhu Mazumdar, Roopa Kohli-Seth, Matthew A. Levin, Prem Timsina, and Arash Kia. Using machine learning to predict ICU transfer in hospitalized COVID-19 patients. *Journal of Clinical Medicine*, 9(6):1668, June 2020.

[11] F. Cismondi, L.A. Celi, A.S. Fialho, S.M. Vieira, S.R. Reti, J.M.C. Sousa, and S.N. Finkelstein. Reducing unnecessary lab testing in the ICU with artificial intelligence. *International Journal of Medical Informatics*, 82(5):345–358, May 2013.

[12] Adam Coates, Blake Carpenter, Carl Case, Sanjeev Satheesh, Bipin Suresh, Tao Wang, David J Wu, and Andrew Y Ng. Text detection and character recognition in scene images with unsupervised feature learning. In *2011 International Conference on Document Analysis and Recognition*, pages 440–445. IEEE, 2011.

[13] Harry Freitas Da Cruz, Boris Pfahringer, Tom Martensen, Frederic Schneider, Alexander Meyer, Erwin P. Böttinger, and Matthieu-P. Schapranow. Using interpretability approaches to update "black-box"clinical prediction models: an external validation study in nephrology. *Artificial Intelligence in Medicine*, 111:101982, 2021.

[14] Tahani A. Daghistani, Radwa Elshawi, Sherif Sakr, Amjad M. Ahmed, Abdullah Al-Thwayee, and Mouaz H. Al-Mallah. Predictors of in-hospital length of stay among cardiac patients: A machine learning approach. *International Journal of Cardiology*, 288:140–147, August 2019.

[15] Debabrata Dansana, Raghvendra Kumar, Aishik Bhattacharjee, D. Jude Hemanth, Deepak Gupta, Ashish Khanna, and Oscar Castillo. Early diagnosis of COVID-19-affected patients based on x-ray and computed tomography images using deep learning algorithm. *Soft Computing*, August 2020.

[16] Jacob Deasy, Pietro Liò, and Ari Ercole. Dynamic survival prediction in intensive care units from heterogeneous time series without the need for variable selection or curation. *Scientific Reports*, 10(1), December 2020.

[17] R. Delahanty, D. Kaufman, and S Jones. Development and evaluation of an automated machine learning algorithm for in-hospital mortality risk adjustment among critical care patients. *Crit Care Med*, 29:6, 2018.

[18] Yiming Ding, Jae Ho Sohn, Michael G. Kawczynski, Hari Trivedi, Roy Harnish, Nathaniel W. Jenkins, Dmytro Lituiev, Timothy P. Copeland, Mariam S. Aboian, Carina Mari Aparici, Spencer C. Behr, Robert R. Flavell, Shih-Ying Huang, Kelly A. Zalocusky, Lorenzo Nardo, Youngho Seo, Randall A. Hawkins, Miguel Hernandez Pampaloni, Dexter Hadley, and Benjamin L. Franc. Application of deep learning to predict standardized uptake value ratio and amyloid status on 18f-florbetapir pet using adni data. *Radiology*, 290(2):456–464, February 2019.

[19] José Alexandre F Diniz-Filho, Thannya N Soares, Jacqueline S Lima, Ricardo Dobrovolski, Victor Lemes Landeiro, Mariana Pires de Campos Telles, Thiago F Rangel, and Luis Mauricio Bini. Mantel test in population genetics. *Genetics and molecular biology*, 36:475–485, 2013.

[20] A. Docherty and N. Lone. Exploiting big data for critical care research. *Curr Opin Crit Care*, 21(5):467–72, 2015.

[21] Pierre Dutilleul, Jason D Stockwell, Dominic Frigon, and Pierre Legendre. The mantel test versus pearson's correlation analysis: Assessment of the differences for biological and environmental studies. *Journal of agricultural, biological, and environmental statistics*, pages 131–150, 2000.

[22] Editorial. Opening the black box of machine learning. *The Lancet Respiratory Medicine*, 6(11):801, 2018.

[23] Alexander C Fanaroff, Anita Y Chen, Laine E Thomas, Karen S Pieper, Kirk N Garratt, Eric D Peterson, L Kristin Newby, James A de Lemos, Mikhail N Kosiborod, Ezra A Amsterdam, et al. Risk score to predict need for intensive care in initially hemodynamically stable adults with non–st-segment–elevation myocardial infarction. *Journal of the American Heart Association*, 7(11):e008894, 2018.

[24] Tom Fawcett. An introduction to roc analysis. *Pattern Recognition Letters*, 27(8):861–874, 2006. ROC Analysis in Pattern Recognition.

[25] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. The kdd process for extracting useful knowledge from volumes of data. *Communications of the ACM*, 39(11):27–34, 1996.

[26] Fernando Timoteo Fernandes, Tiago Almeida de Oliveira, Cristiane Esteves Teixeira, Andre Filipe de Moraes Batista, Gabriel Dalla Costa, and Alexandre Dias Porto Chiavegatto Filho. A multipurpose machine learning approach to predict COVID-19 negative prognosis in são paulo, brazil. *Scientific Reports*, 11(1), February 2021.

[27] A.S. Fialho, F. Cismondi, S.M. Vieira, S.R. Reti, J.M.C. Sousa, and S.N. Finkelstein. Data mining using clinical physiology at discharge to predict ICU readmissions. *Expert Systems with Applications*, 39(18):13158–13165, December 2012.

[28] Ka Man Fong, Shek Yin Au, George Wing Yiu Ng, and Anne Kit Hung Leung. Interpretable machine learning model for mortality prediction in ICU: A multicenter study. October 2020.

[29] Ronaldo Goldschmidt and Emmanuel Passos. *Data Mining: Um Guia Prático*. Campus, 2005.

[30] Anirudh K Gowd, Avinesh Agarwalla, Nirav H Amin, Anthony A Romeo, Gregory P Nicholson, Nikhil N Verma, and Joseph N Liu. Construct validation of machine learning in the prediction of short-term postoperative complications following total shoulder arthroplasty. *Journal of shoulder and elbow surgery*, 28(12):e410–e421, 2019.

[31] G. Gutierrez. Artificial intelligence in the intensive care unit. *Critical care*, 24(1):101–101, 2020.

[32] Sascha Halvachizadeh, Larissa Baradaran, Paolo Cinelli, Roman Pfeifer, Kai Sprengel, and Hans-Christoph Pape. How to detect a polytrauma patient at risk of complications: a validation and database analysis of four published scales. *PloS one*, 15(1):e0228082, 2020.

[33] C. Hanson and B. Marshall. Artificial intelligence applications in the intensive care unit. *Crit Care Med*, 29(2):427–435, 2001.

[34] V. Huddar, B. Desiraju, V. Rajani, S. Bhattacharya, S. Roy, and C. Reddy. Predicting complications in critical care using heterogeneous clinical data. *IEE Access*, 4:7988–8001, 2016.

[35] Vijay Huddar, Bapu Koundinya Desiraju, Vaibhav Rajan, Sakyajit Bhattacharya, Shourya Roy, and Chandan K Reddy. Predicting complications in critical care using heterogeneous clinical data. *IEEE Access*, 4:7988–8001, 2016.

[36] Stephanie L. Hyland, Martin Faltys, Matthias Hüser, Xinrui Lyu, Thomas Gumbsch, Cristóbal Esteban, Christian Bock, Max Horn, Michael Moor, Bastian Rieck, Marc Zimmermann, Dean Bodenham, Karsten Borgwardt, Gunnar Rätsch, and Tobias M. Merz. Early prediction of circulatory failure in the intensive care unit using machine learning. *Nature Medicine*, 26(3):364–373, Mar 2020.

[37] H Jabbar and Rafiqul Zaman Khan. Methods to avoid over-fitting and under-fitting in supervised machine learning (comparative study). *Computer Science, Communication and Instrumentation Devices*, 70, 2015.

[38] A. Johnson, M. Ghassemi, S. Nemati, K. Niehaus, D. Clifton, and G. Clifford. Machine learning and decision support in critical care. In *Proc. of IEEE*, pages 444–466, 2016.

[39] T. Kamio, T. Van, and K. Masamune. Use of machine-learning approaches to predict clinical deterioration in critically ill patients: a systematic review. *Int J Med Res Health Sci*, 6(6):1–7, 2017.

[40] Yohannes Kassahun, Bingbin Yu, Abraham Temesgen Tibebu, Danail Stoyanov, Stamatia Giannarou, Jan Hendrik Metzen, and Emmanuel Vander Poorten. Surgical robotics beyond enhanced dexterity instrumentation: a survey of machine learning techniques and their role in intelligent and autonomous surgical actions. *International Journal of Computer Assisted Radiology and Surgery*, 11(4):553–568, October 2015.

[41] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. In *Advances in neural information processing systems*, pages 3146–3154, 2017.

[42] Pahulpreet Singh Kohli and Shriya Arora. Application of machine learning in disease prediction. In *2018 4th International Conference on Computing Communication and Automation (ICCCA)*, pages 1–4, 2018.

[43] Wojtek J. Krzanowski and David J. Hand. *ROC Curves for Continuous Data*. Chapman and Hall/CRC, 1st edition, 2009.

[44] JZ Li. Monthly housing rent forecast based on lightgbm (light gradient boosting) model. *International Journal of Intelligent Information and Management Science*, 7(6), 2018.

[45] S. Lundberg and S. Lee. A unified approach to interpreting model predictions. In *Proc. of Neurips*, pages 4765–4774, 2017.

[46] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in neural information processing systems*, pages 4765–4774, 2017.

[47] N. Mantel. The detection of disease clustering and a generalized regression approach. *Cancer Research*, 27(2):209–220, 1967.

[48] J. Marshall, L. Bosco, N. Adhikari, B. Connolly, J. Diaz, T. Dorman, R. Fowler, G. Meyfroidt, S. Nakagawa, P. Pelosi, J. Vincent, K. Vollman, and J. Zimmerman. What is an intensive care unit? a report of the task force of the world federation of societies of intensive and critical care medicine. *Journal of Critical Care*, 37:270–276, 2017.

[49] A. Meyer, D. Zverinski, B. Pfahringer, J. Kempfert, T. Kuehne, S. Sündermann, C. Stamm, T. Hofmann, V. Falk, and C. Eickhoff. Machine learning for real-time prediction of complications in critical care: a retrospective study. *The Lancet Respiratory Medicine*, 6(12):905–914, 2018.

[50] B. Mittelstadt, C. Russell, and S. Wachter. Explaining explanations in AI. In *Proc. of FAT**, pages 279–288, 2019.

[51] Mina Chookhachizadeh Moghadam, Ehsan Masoumi, Nader Bagherzadeh, Davinder Ramsingh, Guann-Pyng Li, and Zeev N Kain. A machine learning approach to predict hypotensive events in ICU settings. *bioRxiv*, October 2019.

[52] Christoph Molnar. *Interpretable Machine Learning*. Independently published, 2019. `https://christophm.github.io/interpretable-ml-book/`.

[53] T. Murdoch and A. Detsky. The inevitable application of big data to health care. *JAMA*, 309(13):1351–1352, 2013.

[54] E. Nigri, N. Ziviani, F. Cappabianco, A. Antunes, and A. Veloso. Explainable deep CNNs for mri-based diagnosis of alzheimer's disease. In *Proc. of IJCNN*, pages 1–8, 2020.

[55] Eduardo Nigri, Nivio Ziviani, Fabio Cappabianco, Augusto Antunes, and Adriano Veloso. Explainable deep cnns for mri-based diagnosis of alzheimer's disease, 2020.

[56] P. Pandharipande, T. Girard, J. Jackson, A. Morandi, J. Thompson, B. Pun, N. Brummel, C. Hughes, E. Vasilevskis, A. Shintani, K. Moons, S. Geevarghese, A. Canonico, R. Hopkins, G. Bernard, R. Dittus, and E.W. Ely for the BRAIN-ICU Study Investigators. Long-term cognitive impairment after critical illness. *N Eng J Med*, 369(14):1306–1316, 2013.

[57] Joshua Parreco, Antonio Hidalgo, Jonathan J. Parks, Robert Kozol, and Rishi Rattan. Using artificial intelligence to predict prolonged mechanical ventilation and tracheostomy placement. *Journal of Surgical Research*, 228:179–187, August 2018.

[58] Amir Bahador Parsa, Ali Movahedi, Homa Taghipour, Sybil Derrible, and Abolfazl Kouros Mohammadian. Toward safer highways, application of xgboost and shap for real-time accident detection and feature analysis. *Accident Analysis & Prevention*, 136:105405, 2020.

[59] R. Pirracchio, M. Cohen, I. Malenica, J. Cohen, A. Chambaz, M. Cannesson, C. Lee, M. Resche-Rigon, A. Hubbard, and ACTERREA Research Group. Big data and targeted machine learning in action to assist medical decision in the ICU. *Anaesth Crit Care Pain Med*, 38(4):377–384, 2019.

[60] Juncai Pu and Yong Chen. Data-driven forward-inverse problems and modulational instability for yajima-oikawa system using deep learning with parameter regularization. *arXiv preprint arXiv:2112.04062*, 2021.

[61] A. Rajkomar, J. Dean, and I. Kohane. Machine learning in medicine. *N Engl J Med*, 380(14):1347–1358, 2019.

[62] M. Ribeiro, S. Singh, and C. Guestrin. "why should I trust you?": Explaining the predictions of any classifier. In *Proc. of KDD*, pages 1135–1144, 2016.

[63] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?"explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

[64] Mike D Rinderknecht and Yannick Klopfenstein. Predicting critical state after covid-19 diagnosis: Model development using a large us electronic health record dataset. *npj Digital Medicine*, 4(1):1–14, 2021.

[65] Valeria Romeo, Simone Maurea, Renato Cuocolo, Mario Petretta, Pier Paolo Mainenti, Francesco Verde, Milena Coppola, Serena Dell'Aversana, and Arturo Brunetti. Characterization of adrenal lesions on unenhanced MRI using texture analysis: A machine-learning approach. *Journal of Magnetic Resonance Imaging*, 48(1):198–204, January 2018.

[66] W. Samek and K. Müller. Towards explainable artificial intelligence. In *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, volume 11700 of *Lecture Notes in Computer Science*, pages 5–22. Springer, 2019.

[67] Claudio Filipi Gonçalves Dos Santos and João Paulo Papa. Avoiding overfitting: A survey on regularization methods for convolutional neural networks. *ACM Computing Surveys (CSUR)*, 54(10s):1–25, 2022.

[68] Nirav Shah and Mark Hamilton. Clinical review: Can we predict which patients are at risk of complications following surgery? *Critical Care*, 17(3):1–8, 2013.

[69] Lloyd S Shapley. A value for n-person games. *Contributions to the Theory of Games*, 2(28):307–317, 1953.

[70] Supreeth P. Shashikumar, Christopher Josef, Ashish Sharma, and Shamim Nemati. Deepaise – an end-to-end development and deployment of a recurrent neural survival model for early prediction of sepsis, 2019.

[71] D. Shillan, J. Sterne, A. Champneys, and B. Gibbison. Use of machine learning to analyse routinely collected intensive care unit data: a systematic review. *Critical Care*, 23(1):284, 2019.

[72] Vivek Kumar Singh, Mohamed Abdel-Nasser, Nidhi Pandey, and Domenec Puig. LungINFseg: Segmenting COVID-19 infected regions in lung CT images based on a receptive-field-aware deep learning framework. *Diagnostics*, 11(2):158, January 2021.

[73] Peter E Smouse, Jeffrey C Long, and Robert R Sokal. Multiple regression and correlation extensions of the mantel test of matrix correspondence. *Systematic zoology*, 35(4):627–632, 1986.

[74] Nima Taghipour and Ahmad Kardan. A hybrid web recommender system based on q-learning. In *Proceedings of the 2008 ACM symposium on Applied computing*, pages 1164–1168, 2008.

[75] Zhenyu Tang, Yuyun Xu, Zhicheng Jiao, Junfeng Lu, Lei Jin, Abudumijiti Aibaidula, Jinsong Wu, Qian Wang, Han Zhang, and Dinggang Shen. Pre-operative overall survival time prediction for glioblastoma patients using deep learning on both imaging

phenotype and genotype. In *International Conference on Medical Image Computing and Computer Assisted Intervention*, pages 415–422. Springer, 2019.

[76] Julian Theis, William Galanter, Andrew Boyd, and Houshang Darabi. Improving the in-hospital mortality prediction of diabetes icu patients using a process mining/deep learning architecture. *IEEE Journal of Biomedical and Health Informatics*, 2021.

[77] H. Thorsen-Meyer, A. Nielsen, A. Nielsen, B. Kaas-Hansen, P. Toft, J. Schierbeck, T. Strom, P. Chmura, M. Heimann, L. Dybdahl, L. Spangsege, P. Hulsen, K. Belling, S. Brunak, and A. Perner. Dynamic and explainable machine learning prediction of mortality in patients in the intensive care unit: a retrospective study of high-frequency data in electronic patient records. *The Lancet Digital Health*, 2020.

[78] E. Tjoa and C. Guan. A survey on explainable artificial intelligence (XAI): towards medical XAI. *CoRR*, abs/1907.07374, 2019.

[79] K. To and L. Napolitano. Common complications in the critically ill patient. *Surgical Clinics North Amer.*, 92(6):1519–1557, 2012.

[80] Francisco Valente, Jorge Henriques, Simão Paredes, Teresa Rocha, Paulo de Carvalho, and João Morais. A new approach for interpretability and reliability in clinical risk prediction: Acute coronary syndrome scenario. *Artificial Intelligence in Medicine*, 117:102113, jul 2021.

[81] D. Valle, T. Pimentel, and A. Veloso. Assessing the reliability of visual explanations of deep models with adversarial perturbations. In *Proc. of IJCNN*, pages 1–8, 2020.

[82] L. van der Maaten and G. Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

[83] Bernhard Wernly, Behrooz Mamandipoor, Philipp Baldia, Christian Jung, and Venet Osmani. Machine learning predicts mortality in septic patients using only routinely available abg variables: a multi-centre evaluation. *International Journal of Medical Informatics*, 145:104312, 2021.

[84] C. Wollschlager and A. Conrad. Common complications in critically ill patients. *Disease-a-Month*, 34(5):225–293, 1988.

[85] Xue Ying. An overview of overfitting and its solutions. *Journal of Physics: Conference Series*, 1168(2):022022, feb 2019.

[86] Joo Heung Yoon, Lidan Mu, Lujie Chen, Artur Dubrawski, Marilyn Hravnak, Michael R. Pinsky, and Gilles Clermont. Predicting tachycardia as a surrogate for instability in the intensive care unit. *Journal of Clinical Monitoring and Computing*, 33(6):973–985, February 2019.

[87] Xian Zeng, Jiye An, Ru Lin, Cong Dong, Aiyu Zheng, Jianhua Li, Huilong Duan, Qiang Shu, and Haomin Li. Prediction of complications after paediatric cardiac surgery. *European Journal of Cardio-Thoracic Surgery*, 57(2):350–358, 2020.