Mortalidade em Unidades de Terapia Intensiva: uma Abordagem para Predição Explorando Aprendizado de Máquina

Alexandre Renato Rodrigues de Souza[1,5], **Fabrício Neitzke Ferreira**[2], **Rodrigo Blanke Lambrecht**[3,4], **Leonardo Costa Reichow**[4], **Rogério da Costa Albandes**[5], **Adenauer Correa Yamin**[3,5]

[1]Instituto Federal de Educação, Ciência e Tecnologia do Rio Grande do Sul (IFRS)
[2]Instituto Federal de Educação, Ciência e Tecnologia Sul-Rio-Grandense (IFSUL)
[3]Universidade Católica de Pelotas (UCPEL)
[4]Lifemed Ind. de Equip. e Artigos Médicos e Hospitalares S.A. (LIFEMED)
[5]Universidade Federal de Pelotas (UFPEL)

Abstract. *The main objective of this research is the conception of an approach to predict mortality in ICUs. A cohort of 17,734 patients was used, from the MIMIC-III Database, considering 12 input predictor variables and 7 Machine Learning methods. The best performance was achieved by the Gradient Boosting Classifier (GBC) method, which obtained a 0.53 of F1 score and 0.85 for AUC. The approach conceived enables the generation of robust models capable of detecting hidden patterns, dealing with large amounts of data, and having greater power of discrimination in classifications. The results are promising and, in some cases, superior to those obtained by other proposals identified in the literature review.*

Resumo. *Esta pesquisa tem por objetivo central a concepção de uma abordagem para predição de mortalidade em UTIs. Foi empregada uma coorte de 17.734 pacientes, provenientes do Banco de Dados MIMIC-III, sendo consideradas 12 variáveis preditoras de entrada e 7 métodos de Aprendizagem de Máquina. A melhor performance foi alcançada pelo método Gradient Boosting Classifier (GBC), que atingiu 0,53 de F1 score e 0,85 de AUC. A abordagem concebida viabiliza a geração de modelos robustos, capazes de detectar padrões ocultos, lidar com grandes quantidades de dados e ter maior poder de discriminação nas classificações. Os resultados são promissores e, em alguns casos, superiores a outras propostas identificadas na revisão de literatura.*

1. Introdução

As Unidades de Terapia Intensiva (UTIs) dos hospitais tem atraído relevantes esforços de pesquisa, pois os pacientes requerem monitoramento contínuo de seus parâmetros fisiológicos devido à gravidade de sua condição de saúde e apresentarem elevado risco de rápida deterioração clínica [Purushotham and at al. 2018].

A mortalidade é um desfecho primário de interesse no tratamento intensivo, pois as taxas de mortalidade nas UTIs são as mais altas entre as unidades hospitalares (em torno de 10 a 29%, dependendo da idade e da doença), e a identificação precoce de pacientes em risco é fundamental para melhorar os desfechos do tratamento [Harutyunyan et al. 2019].

Oportuno ter presente, que escores clínicos de alerta precoce, tais como MPM, EWS, NEWS, MEWS, SOFA, qSOFA, SAPS e APACHE, classificam os pacientes de acordo com o risco de morte. A maioria desses escores escolhe um pequeno número de preditores explicativos e usa modelos matemáticos simples para prever o desfecho clínico através de suposições de relação linear e aditiva. Embora essas pontuações ainda sejam amplamente utilizadas no ambiente hospitalar, diversos estudos mostram que os modelos customizados para avaliação da mortalidade utilizando Aprendizado de Máquina têm um desempenho superior a esses sistemas tradicionais de pontuações [Purushotham and at al. 2018].

Em 2012, o Laboratório *PhysioNet* do *Massachusetts Institute of Technology (MIT)* propôs um desafio para incentivar o desenvolvimento de novas técnicas de Aprendizado de Máquina para identificar o risco de mortalidade hospitalar de pacientes internados em UTIs. Os métodos desenvolvidos pelos vencedores do desafio superaram a performance de previsão de alguns desses sistemas de pontuação de risco [Johnson et al. 2012].

Este desafio promoveu junto à comunidade científica internacional um aumento de interesse por esse tema, bem como por conjuntos de dados abertos de saúde, tais como o MIMIC, que estão gradualmente se tornando disponíveis para fins de pesquisa, o que tem viabilizado um número crescente de trabalhos sobre previsão de mortalidade em UTIs.

Considerando este cenário, o presente trabalho tem por objetivo geral investigar o uso de métodos de classificação de Aprendizado de Máquina no desenvolvimento de um modelo de predição de mortalidade em UTIs para auxiliar os médicos na tomada de decisão. Isto é feito analisando diferentes dados clínicos de pacientes coletados nas primeiras 24 e 48 horas após a internação.

Por sua vez, os objetivos específicos deste estudo são: (i) considerar os principais desafios no uso de técnicas de Aprendizado de Máquina para previsão de mortalidade hospitalar de pacientes internados em UTIs; (ii) propor uma abordagem para a previsão de mortalidade para estes pacientes; e, (iii) avaliar e comparar o desempenho de diferentes métodos de Aprendizado de Máquina, empregando um banco de dados construído a partir de informações provenientes de UTIs.

A próxima Seção discute Trabalhos Relacionados à área de estudo deste artigo, os quais foram selecionados a partir de uma Revisão Sistemática de Literatura realizada no decorrer da pesquisa, a qual é foco de outro artigo.

2. Trabalhos Relacionados Explorando Aprendizado de Máquina para Predição de Mortalidade em UTIs

Durante os esforços de Revisão Sistemática de Literatura associada à pesquisa em desenvolvimento, foram identificados diversos trabalhos relacionados à previsão de mortalidade em UTIs. Foram priorizadas aqueles que contemplassem os seguintes aspectos: (i) uso de uma das versões do MIMIC como base de dados, de forma a considerar coortes de pacientes de UTIs originados de um único centro; (ii) emprego de no máximo vinte variáveis brutas retiradas do banco de dados; e, (iii) uso da métrica AUC (*Area Under ROC Curve*) como forma de avaliar a performance dos modelos propostos, tornando possível assim uma comparação entre os diversos trabalhos, os quais estão apresentados a seguir.

Oportuno ressaltar que a literatura tem apontado a AUC como métrica de

"número único" para avaliação de performance de métodos de Aprendizado de Máquina por apresentar várias propriedades desejáveis quando comparada à outras técnicas [Muralitharan and at al. 2021, Bradley 1997].

 # 1 - **[Baker et al. 2020]** propôs um modelo híbrido que combina CNN e BiLSTM para prever a mortalidade a partir de estatísticas que descrevem a variação da frequência cardíaca, pressão arterial, frequência respiratória, níveis de oxigênio no sangue e temperatura corporal. O modelo de melhor performance obteve uma AUC de 0,88. O trabalho teve como conclusão que o uso de uma rede híbrida CNN-BiLSTM é eficaz na determinação do risco de mortalidade para as janelas de 3, 7 e 14 dias de sinais vitais. Os resultados mostram que é possível implementar um sistema preciso para prever o risco de mortalidade de forma contínua e automática, reduzindo a carga de trabalho dos profissionais de saúde e melhorando os desfechos clínicos dos pacientes.

 # 2 - **[Alghatani et al. 2021]** desenvolveu modelos para previsão de tempo de permanência nas UTIs e predição de mortalidade baseados no banco de dados MIMIC-III. Foram aplicados seis métodos de Aprendizado de Máquina comumente usados para prever a mortalidade, utilizando 11 variáveis de entrada (dados demográficos e sinais vitais) em cada modelo. A melhor AUC alcançada no modelo de mortalidade foi de 0,78 usando o algoritmo *Random Forest*. A novidade nessa abordagem foi a construção de modelos para prever o tempo de permanência nas UTIs e mortalidade com razoável precisão com base em uma combinação de Aprendizado de Máquina e a abordagem de quantis que utiliza apenas os sinais vitais disponíveis no perfil do paciente. A técnica utilizada é baseada na engenharia de atributos dos sinais vitais, incluindo suas médias modificadas, desvios padrão e quantis das variáveis originais, o que forneceu um conjunto de dados mais apropriado para obter um melhor poder preditivo dos modelos.

 # 3 - **[Purushotham and at al. 2018]** apresentou os resultados de performance para várias tarefas de previsão clínica, como mortalidade, tempo de permanência e código ICD-9 usando modelos de *Deep Learning*, *ensembles* de modelos de Aprendizado de Máquina (algoritmos *Super Learner*), SAPS II e pontuações SOFA. ICD-9 é o sistema oficial de códigos para diagnósticos e procedimentos em utilização nos hospitais dos Estados Unidos. Foi empregado o conjunto MIMIC-III como fonte de dados. Os resultados mostraram que os modelos de aprendizado profundo superam consistentemente todas as outras abordagens, especialmente quando os dados brutos de séries temporais clínicas são usados como variáveis de entrada para os modelos. O método de aprendizado profundo MMDL (*Multimodal Deep Learning Model*) alcançou uma AUC de 0,87 utilizando 17 variáveis preditoras e 48 horas de dados.

 # 4 - **[Bhattacharya and et al. 2017]** propôs um novo algoritmo para predição de mortalidade em UTIs para resolver o problema de desequilíbrio entre classes. O método é baseado na transformação das variáveis preditoras para reduzir a correlação existente entre elas. A eficácia do algoritmo foi demonstrada em conjuntos de dados simulados e no MIMIC-II. Uma vantagem da proposta é o uso de apenas seis dados clínicos do paciente (pressão arterial média, frequência cardíaca, temperatura corporal, nível de sódio, nível de potássio e nível de magnésio) que podem ser facilmente obtidos em prontuários eletrônicos. Em comparação, outros métodos ou sistemas de pontuação usam medidas que podem não estar disponíveis para todos os pacientes, podendo assim exigir intervenção manual ou análise de notas clínicas. O modelo desenvolvido, chamado pelos autores de

CHISQ-NEW, obteve uma AUC de 0,87.

 # 5 - **[Pirracchio and at al. 2015]** desenvolveu um algoritmo *Super Learner* para previsão de mortalidade para pacientes em UTIs, comparando sua performance com escores tradicionais de pontuação. Foram avaliadas a calibração, discriminação e classificação de risco da mortalidade hospitalar prevista com base no *Super Learner* em comparação com SAPS-II, APACHE-II e SOFA. Como fonte de dados primária foram utilizados dados clínicos de 24.508 pacientes do banco de dados MIMIC-II. Foram produzidos dois conjuntos de previsões com base no *Super Learner*; o primeiro baseado nas 17 variáveis conforme aparece no escore SAPS-II (SL1), e o segundo, baseado nas variáveis originais sem transformações (SL2). O *Super Learner* teve uma AUC de 0,85 quando foi utilizado o SL1, e de 0,88 utilizando o SL2. Comparado com os escores de gravidade convencionais, o modelo proposto apresentou melhor desempenho para prever a mortalidade hospitalar em pacientes em UTIs.

 # 6 - **[Harutyunyan et al. 2019]** apresentou quatro tarefas de predição clínica usando dados derivados do MIMIC-III. Essas tarefas abrangem uma série de problemas clínicos, incluindo modelagem de risco de mortalidade, previsão de tempo de hospitalização, detecção de deterioração fisiológica e classificação de fenótipo. Foram propostos modelos lineares e de redes neurais para todas as quatro tarefas e avaliado o efeito da supervisão profunda, treinamento multitarefa e modificações de arquiteturas específicas de dados no desempenho dos modelos neurais. O trabalho identificou que os modelos baseados em LSTM superam significativamente os modelos lineares e apresentou as vantagens de usar *channel-wise* LSTMs para prever várias tarefas usando um único modelo neural. A maior AUC alcançada pelo trabalho para previsão de mortalidade foi de 0,87.

 # 7 - **[Awad et al. 2020]** investigou como a mortalidade hospitalar precoce pode ser prevista para pacientes de UTIs. Os resultados mostraram que o poder de discriminação dos métodos de classificação de Aprendizado de Máquina após 6h da admissão superou os principais sistemas de pontuação utilizados em medicina intensiva (APACHE, SAPS e SOFA) após 48h de admissão. O classificador com melhor desempenho foi RF (AUC de 0,90), seguido por BN e depois PART em diferentes configurações experimentais. Os autores chegaram à conclusão que: (i) há uma melhora acentuada no desempenho na 6ª hora de admissão nas UTIs; (ii) a porcentagem de valores ausentes no conjunto de dados reduz drasticamente na 6ª hora de admissão nas UTIs e continua a diminuir gradualmente até a 48ª hora. O trabalho alerta para o problema dos *missing values* das variáveis coletadas, a fim de enfatizar a importância de coletar certas medidas desde o início da internação; já que isso influenciará no desempenho preditivo dos modelos de predição de mortalidade.

 # 8 - **[Este trabalho]**, cuja caracterização acontece na continuidade do artigo, aparece na Tabela 1 com o intuito de promover uma comparação de suas características ante a literatura. Uma discussão desta comparação é feita na Seção 4.

A Tabela 1 apresenta uma comparação entre os trabalhos relacionados, contemplando os aspectos centrais considerados na concepção da abordagem proposta por este estudo: (i) identificação do trabalho; (ii) número de variáveis clínicas; (iii) descrição das variáveis clínicas; (iv) versão do banco de dados MIMIC; (v) janela de tempo de medição das variáveis clínicas; (vi) métodos de predição avaliados (quando um trabalho avaliou

mais de um método, o que obteve a melhor performance está destacado em azul); (vii) métricas empregadas para medir a performance do modelo; (viii) melhor performance obtida pelo trabalho considerando a métrica AUC; e, (ix) coorte de pacientes ou internações.

3. Abordagem Proposta para Predição de Mortalidade

Esta seção descreve a abordagem proposta para lidar com a previsão de mortalidade de paciente internados em UTIs, conforme a Figura 1, considerando algumas decisões de projeto que orientaram sua construção.

3.1. Discussão do Problema de Pesquisa

Os sistemas de pontuação tradicionais para predição de mortalidade têm como objetivo identificar a deterioração do estado clínico do paciente. No entanto, esses sistemas de pontuação levam em consideração somente os dados de saúde de um determinado instante de tempo, sem considerar a tendência de variação dos mesmos no decorrer da internação [Churpek and at al. 2016].

A implantação crescente dos prontuários eletrônicos nos hospitais tem viabilizado o registro de dados históricos dos pacientes, cujos sinais vitais coletados ao longo do tempo podem ser interpretados como séries temporais, criando assim a oportunidade para a aplicação de técnicas computacionais que processam esses dados e permitem produzir previsões da evolução do estado clínico dos pacientes.

O objetivo do modelo desenvolvido por este trabalho é identificar o risco de morte durante a internação de pacientes em UTIs considerando como base os dados do MIMIC-III. Foram comparadas as performances de vários métodos de Aprendizado de Máquina, utilizando para isso dados de uma mesma coorte de pacientes coletados em janelas de tempo de 24 e 48 horas após a admissão nas UTIs. As janelas de tempo foram estabelecidas através de uma avaliação empírica inicial, que mostrou que com esses intervalos de tempo já é possível uma estimativa de risco de mortalidade suficientemente precisa. Como foram incluídos somente pacientes internados por pelo menos há 48 horas, o desfecho ocorreu após este intervalo de tempo de internação.

3.2. Banco de Dados e População Estudada

A coorte de pacientes relevantes para o desenvolvimento do modelo foi extraída do *Medical Information Mart for Intensive Care* (MIMIC-III) baseado no código fornecido por [Harutyunyan et al. 2019]. O MIMIC-III [Johnson and at al. 2016] é um repositório de acesso público criado a partir de dados clínicos de pacientes do *Beth Israel Deaconess Medical Center* (BIDMC), que é um hospital universitário da *Harvard Medical School*.

Para o desenvolvimento do modelo foram incluídos apenas registros de pacientes com 18 anos ou mais e que permaneceram internados em UTIs por um período mínimo de 48 horas. Isso resultou em uma coorte de 17.734 pacientes e 1.456.610 observações. Desses pacientes, 15.328 sobreviveram e 2.406 morreram, resultando em uma taxa de mortalidade de 13,57%.

3.3. Análise Exploratória dos Dados e Seleção de Variáveis

As variáveis clínicas extraídas do MIMIC-III para desenvolvimento da abordagem proposta possuem registro do momento em que foram coletadas e contém informações demográficas dos pacientes, resultados de exames de laboratório, sinais vitais e a escala

Tabela 1. Comparação dos Trabalhos Relacionados

#	Variáveis Clínicas n	Variáveis Clínicas Descrição	MIMIC	Janela	Métodos de Predição	Métricas de Performance	AUC	Coorte
1	9	BT, DBP, HR, idade, MAP, RR, SBP, sexo, SpO2	III	24h	rede neural híbrida (CNN-BiLSTM)	AUC, AUPRC, acurácia, especificidade, sensibilidade	0,88	51.279 pacientes
2	11	altura, BT, DBP, glicose, HR, idade, peso, RR, SBP, sexo, SpO2	III	-	**RF**, LR, LDA, kNN, SVM, XGB	acurácia, sensibilidade, especificidade, NPV, PPV, AUC, curva ROC	0,78	44.626 internações
3	17	AIDS, bicabornato, bilirrubinas, BT, débito urinário, GCS, HR, idade, neoplasia hematológica, neoplasia metastática, PaO2/FiO2, potássio, SBP, sódio, tipo de admissão hospitalar, uréia, WBC	III	24/48h	super learner, **MMDL**	AUC, AUPRC	0,87	35.627 internações
4	6	BT, HR, magnésio, MAP, potássio, sódio	II	-	**CHISQ-NEW**, SVM, RF, LR, LDA, QDA, *Adaboost*	AUC	0,87	4.000 pacientes
5	17	bicabornato, bilirrubinas, BT, débito urinário, GCS, HR, idade, neoplasia hematológica, neoplasia metastática, PaO2/FiO2, potássio, SBP, sódio, tipo de admissão hospitalar, tipo de admissão hospitalar, uréia, WBC	II	-	*super learner*	AUC, curva ROC	0,88	24.508 pacientes
6	17	altura, BT, DBP, enchimento capilar, FiO2, GCS (pontuação total), GCS (resposta motora), GCS (resposta pupilar), GCS (resposta verbal), glicose, HR, MAP, peso, pH, RR, SBP, SpO2	III	-	LR, *channel-wise LSTM*, *multitask standard LSTM*, *standart LSTM*, *deep supervision*, ***multitask channel-wise LSTM***	AUC, AUPRC	0,87	21.139 internações
7	8	BT, creatinina, GCS, HR, idade, PaO2, RR, SBP	II	48h	**RF**, NB, PART, DT, SVM, JRip	AUC	0,90	11.722 pacientes
8	12	BT, DBP, GCS, glicose, HR, idade, MAP, pH, RR, SBP, sexo, SpO2	III	24/48h	LR, kNN, DT, RF, GNB, MLP, *Adaboost*, **GBC**	sensibilidade, especificidade, *F1 score*, AUC	0,85	17.734 pacientes

Descrição das variáveis: **HR** - frequência cardíaca; **SBP** - pressão arterial sistólica; **DBP** - pressão arterial diastólica; **MAP** - pressão arterial média; **RR** - frequência respiratória; **SpO2** -saturação periférica de oxigênio; **BT** - temperatura corporal; **GCS** - escala de coma de *Glasgow*; **FiO2** - fração inspirada de oxigênio; **PaO2** - pressão arterial de oxigênio; **WBC** - contagem de células brancas; **AIDS** - Síndrome da Imunodeficiência Humana; **pH** - potencial hidrogeniônico.

Métodos de predição: **CNN** - *Convolutional Neural Networks*; **LSTM** - *Long Short-Term Memory*; **BiLSTM** - *bidirectional Long Short-Term Memory*; **RF** - *Random Forest*; **LR** - *Logistic Regression*; **kNN** - *k-Nearest Neighbors*; **SVM** - *Support Vector Machines*; **XGB** - *eXtreme Gradient Boosting*; **MMDL** - *Multimodal Deep Learning Model*; **LDA** - *Linear Discriminant Analysis*; **QDA** - *Quadratic Discriminant Analysis*; **AdaBoost** - *Adaptive Boosting*; **NB** - *Naive Bayes*; **DT** - *Decision Tree*; **PART** - *partial Decision Tree*; **JRip** - *Repeated Incremental Pruning to Produce Error Reduction (RIPPER)*; **RNN** - *Recurrent Neural Network*; **GNB** - *Gaussian Naive Bayes*; **MLP** - *Multilayer Perceptron*; **GBC** - *Gradient Boosting Classifier*.

Métricas de performance: **ROC** - *Receiver Operating Characteristic*; **AUC** – *Area Under the Curve*; **AUPRC** - *Area Under Precision-Recall Curve*; **NPV** - *Negative Predictive Value*; **PPV** - *Positive Predictive Value*.

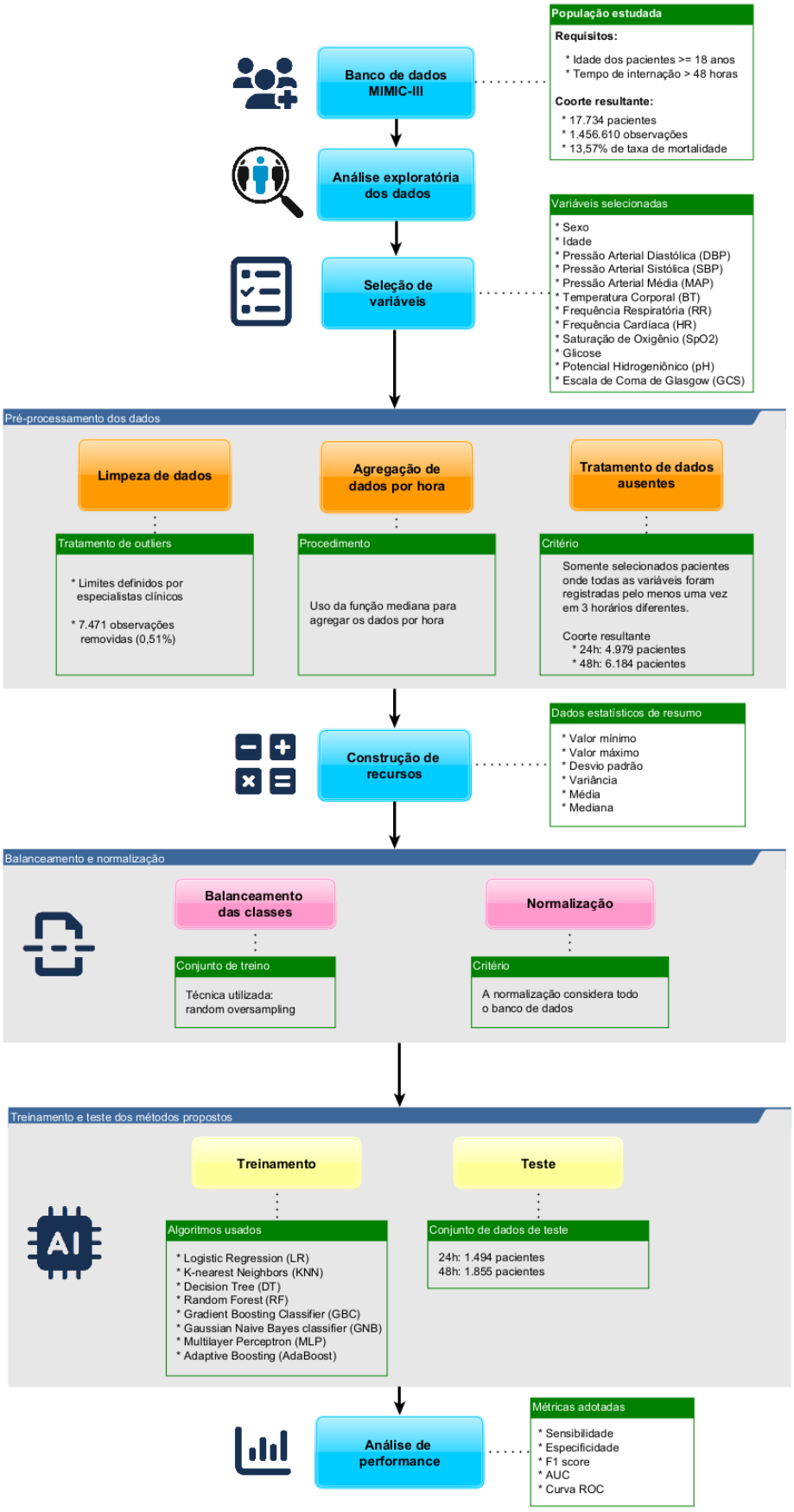


Figura 1. Organização da Abordagem Proposta (Fonte: os autores)

de coma de *Glasgow*, conforme listado na Tabela 2. Estas variáveis foram selecionadas através de uma análise de sensibilidade e considerando seu amplo emprego na literatura científica da área [Muralitharan and at al. 2021].

3.4. Pré-processamento dos Dados

A qualidade e a quantidade de informações úteis são importantes fatores que irão determinar o quanto um algoritmo de Aprendizagem de Máquina irá aprender. Com base nessa premissa, deve-se levar em consideração que os dados existentes nos prontuários eletrônicos apresentam diferentes formatos, dimensões e características, de forma que geralmente não estão prontos para serem inseridos diretamente nesses algoritmos. Dessa forma é indispensável uma boa preparação dos dados antes de alimentar os modelos, o que é realizado na etapa chamada de pré-processamento. Diante disso, a seguir são apresentadas as principais tarefas de preparação de dados utilizadas para o desenvolvimento do modelo de predição.

Tabela 2. Variáveis Clínicas Selecionadas

Variável	Sigla	Tipo	Categoria	Unidade	Observações	Incidência
Sexo	-	dado demográfico	estática	-	1.456.610	100%
Idade	-	dado demográfico	estática	anos	1.456.610	100%
Pressão Arterial Diastólica	DBP	sinal vital	dinâmica	mmHg	1.013.725	70%
Pressão Arterial Sistólica	SBP	sinal vital	dinâmica	mmHg	1.014.126	70%
Pressão Arterial Média	MAP	sinal vital	dinâmica	mmHg	1.006.905	69%
Temperatura Corporal	BT	sinal vital	dinâmica	°C	320.929	22%
Frequência Respiratória	RR	sinal vital	dinâmica	RPM	1.064.918	73%
Frequência Cardíaca	HR	sinal vital	dinâmica	BPM	1.055.868	72%
Saturação de Oxigênio	SpO2	sinal vital	dinâmica	%	1.063.682	73%
Glicose	-	exame de laboratório	dinâmica	mg/dL	246.892	17%
Potencial Hidrogeniônico	pH	exame de laboratório	dinâmica	-	122.296	8%
Escala de Coma de *Glasgow*	GCS	pontuação	dinâmica	-	158.842	11%

3.4.1. Limpeza de Dados

Os dados extraídos do MIMIC-III possuem valores errôneos devido a ruídos, registros incorretos, erros tipográficos e imputação de informações ou unidades inconsistentes [Purushotham and at al. 2018]. Para tratar desses valores discrepantes presentes no banco de dados foram utilizadas as especificações constantes no repositório de código-fonte de [Harutyunyan et al. 2019], as quais foram definidas por especialistas clínicos com base em seu conhecimento de intervalos de medidas válidas.

No modelo proposto, cada variável numérica está associada a limites superior e inferior para detectar valores inutilizáveis (*outliers*). O valor observado será excluído se estiver fora desses limites. Ao aplicar essas regras para gerar a coorte do modelo foram removidas 7.471 observações (0,51%) classificadas como valores atípicos extremos.

3.4.2. Agregação dos Dados por Hora e Tratamento de Dados Ausentes

Os dados do MIMIC-III possuem o registro do momento de coleta (data/hora) para cada medição realizada. No entanto, a maioria das medições são amostradas de forma irregular e infrequente, de forma que as séries temporais brutas de cada variável são bastante

esparsas. A Tabela 2 apresenta a quantidade de observações com dados não nulos por variável. Conforme pode-se observar, existem poucas medições das variáveis Potencial Hidrogeniônico, Escala de Coma de *Glasgow* e Glicose, com incidências em 8%, 11% e 17% das observações, respectivamente.

Os modelos de Aprendizado de Máquina aplicados a séries temporais têm uma performance de classificação melhor quando recebem como entrada dados com representações de tempo discretizadas [Faceli et al. 2021]. Dessa forma, com o objetivo de obter uma representação mais densa dos dados fisiológicos de forma a oportunizar uma melhor inferência pelos algoritmos, as observações de cada série temporal foram agregadas em intervalos de hora em hora através do cálculo da mediana.

Como estratégia para evitar a geração de viés no modelo, os dados ausentes não foram substituídos por valores estimados. A abordagem utilizada foi incluir apenas pacientes com um alto percentual de informações completas, sendo então selecionados pacientes onde todas as variáveis foram registradas pelo menos uma vez em três horários diferentes dentro da janela de tempo de medição. Esta imposição resultou em um conjunto de dados com 6.184 pacientes na janela de tempo de 48h e 4.979 pacientes na janela de 24h.

3.5. Construção de Recursos

A construção de recursos aborda o problema de encontrar a transformação de variáveis que possuam a maior quantidade de informações úteis. Como as séries temporais presentes no banco de dados utilizado nesse trabalho contém um grande número de informações ausentes, o modelo proposto calcula os dados estatísticos resumidos (valor mínimo, valor máximo, desvio padrão, variância, média e mediana) de cada uma das variáveis dentro da janela de tempo estipulada (24 ou 48 horas). Esta estratégia reduz a complexidade do modelo, pois utiliza informações mais relevantes como entrada para a tarefa de previsão.

A identificação dos valores mínimo e máximo de cada série tem como objetivo mostrar os eventos extremos durante a internação. O desvio padrão e variância foram utilizados para quantificar a variabilidade dos eventos. A média e a mediana foram calculadas para fornecer uma representação precisa do evento médio para cada variável, sendo que o uso de ambas ajuda a reduzir o risco de distorção do resultado originado por alguma distribuição atípica dos dados.

3.6. Balanceamento das Classes

Conforme apresentado na Seção 3.2, o banco de dados utilizado no desenvolvimento desse modelo possui uma incidência de mortalidade de 13,57%, de forma que há uma distribuição desigual entre as classes. A modelagem preditiva com classes desequilibradas representa um desafio para o Aprendizado de Máquina, pois a maioria dos métodos usados para classificação foram projetados com base na suposição de um número igual de exemplos para cada classe [Bhattacharya and et al. 2017].

Neste trabalho foram experimentadas três diferentes estratégias de balanceamento entre as classes: (i) *random oversampling*: sobreamostragem aleatória da classe minoritária (método escolhido); (ii) *random downsampling*: seleciona e remove aleatoriamente amostras da classe majoritária; e, (iii) *SMOTE (Synthetic Minority Oversampling Technique)*: geração de exemplos sintéticos de treinamento da classe minoritária.

3.7. Normalização

Muitos métodos de Aprendizado de Máquina exigem que as variáveis selecionadas estejam na mesma escala para um desempenho ideal [Faceli et al. 2021]. O método chamado "min-max" foi escolhido, através do qual os valores foram transformados para uma escala mínimo-zero e máximo-um.

4. Análise de Performance da Abordagem Proposta

Conforme apresentado na Tabela 3, foram empregados os indicadores sensibilidade, especificidade, *F1 score* e AUC para comparar diferentes métodos para previsão de mortalidade, considerando janelas de tempo com 24 e 48 horas de aquisição de dados. O realce na cor azul nas colunas *F1 score* e AUC destaca o método com as pontuações mais elevadas em cada uma das métricas.

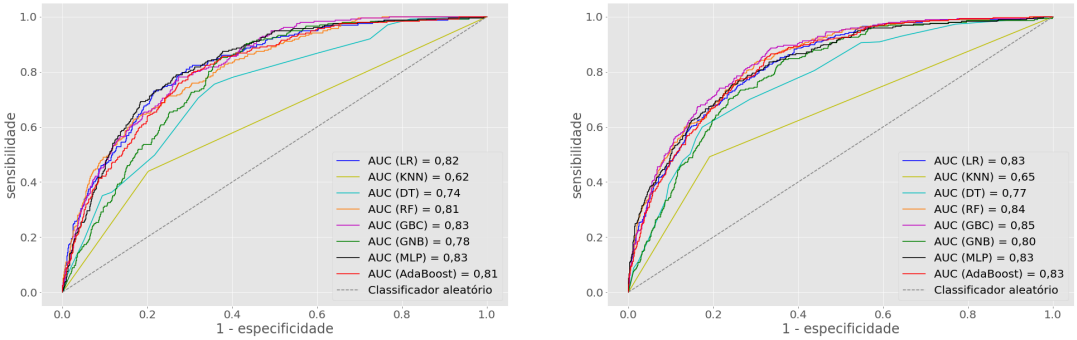
A sensibilidade, também chamada de Taxa de Verdadeiros Positivos (TPR), indica a capacidade do método de classificação em prever corretamente os pacientes que morreram (classe positiva). A especificidade indica a capacidade do método de classificação em prever corretamente os pacientes que sobreviveram (classe negativa). *F1 score* é uma média harmônica calculada com base na precisão e na sensibilidade [Faceli et al. 2021]. A AUC é uma grandeza escalar entre 0 e 1 que representa a área abaixo da curva ROC e mede a qualidade das previsões do modelo independentemente do ponto de operação do classificador [Bradley 1997].

Tabela 3. Performances de Classificação Usando Diferentes Métodos de Modelagem e Janelas de Tempo para Aquisição de Dados

Método	Janela de 24h				Janela de 48h			
	Sensib.	Espec.	F1 score	AUC	Sensib.	Espec.	F1 score	AUC
LR	0,75	0,75	0,49	0,82	0,74	0,75	0,48	0,83
kNN	0,44	0,80	0,35	0,62	0,49	0,81	0,39	0,65
DT	0,70	0,68	0,41	0,74	0,70	0,71	0,44	0,77
RF	0,65	0,80	0,48	0,81	0,69	0,79	0,49	0,84
GBC	0,59	0,84	0,49	**0,83**	0,68	0,83	**0,53**	**0,85**
GNB	0,68	0,71	0,43	0,78	0,74	0,70	0,45	0,80
MLP	0,66	0,83	**0,51**	**0,83**	0,68	0,80	0,50	0,83
AdaBoost	0,66	0,78	0,46	0,81	0,73	0,76	0,49	0,83

Na janela de tempo de 24 horas, o método MLP teve melhor performance considerando simultaneamente os indicadores *F1 score* e AUC, com valores de 0,51 e 0,83, respectivamente. O algoritmo GBC teve performance similar, mas com o valor de *F1 score* um pouco inferior. Considerando a Figura 2(a), que apresenta as curvas ROC dos modelos de 24 horas, pode-se observar que apesar dos métodos MLP, LR, RF, GBC e *Adaboost* apresentarem valores semelhantes em relação a métrica AUC, eles se comportam de maneira diferente dependendo do limiar de decisão a ser escolhido. Portanto a escolha do classificador mais adequado dependerá do ponto de equilíbrio desejado para a especificidade e sensibilidade.

Já na janela de tempo de 48 horas, o método GBC teve uma performance ligeiramente superior em ambos indicadores *F1 score* e AUC, com valores de 0,53 e 0,85, respectivamente. As curvas ROC dos modelos nesta janela de tempo estão apresentadas na Figura 2(b), onde também é possível constatar que apesar de vários dos algoritmos



(a) Janela de Tempo de 24h (b) Janela de Tempo de 48h

Figura 2. Curvas ROC dos Métodos de Predição Utilizados por Janela de Medição (Fonte: os autores)

apresentarem desempenhos semelhantes, a escolha depende do balanço pretendido entre especificidade e sensibilidade.

A comparação do desempenho da abordagem proposta com outros trabalhos da literatura será feita tendo como base a Tabela 1. A métrica AUC foi empregada para essa comparação por ser apontada pela literatura como a mais relevante para medição de desempenho quando da predição de mortalidade hospitalar [Muralitharan and at al. 2021]. Como referência será utilizado o resultado obtido pelo método GBC, desenvolvido pelo presente trabalho, o qual alcançou a melhor performance de classificação (AUC de 0,85). O modelo é baseado em 12 variáveis preditoras de entrada, sendo 7 sinais vitais (DBP, SBP, MAP, BT, RR, HR, SpO2), 2 dados demográficos (sexo e idade), 2 resultados de laboratório (glicose e pH) e um sistema de pontuação (GCS). Estes sinais vitais usualmente são registrados automaticamente por monitores multiparamétricos em ambientes de UTIs, simplificando assim a coleta destes dados fisiológicos.

A performance AUC de 0,85 alcançada é superior ao trabalho de [Alghatani et al. 2021], que obteve AUC de 0,78. Por sua vez, os trabalhos [Purushotham and at al. 2018] (AUC de 0,87), [Pirracchio and at al. 2015] (AUC de 0,88) e [Harutyunyan et al. 2019] (AUC de 0,87) apresentam melhor performance, entretanto exigem ao total 17 variáveis preditoras, dependem de até 6 exames laboratoriais ou requerem os valores individuais dos 4 critérios da escala de *Glasgow*, da qual muitas vezes é registrado somente o valor total. Estes aspectos podem dificultar sua implementação, pois os profissionais de saúde teriam de atualizá-los durante a internação do paciente e via de regra estes valores não são medidos regularmente.

Apesar da abordagem proposta não utilizar a menor quantidade de variáveis preditoras, ainda assim apresenta uma performance semelhante a trabalhos [Baker et al. 2020, Bhattacharya and et al. 2017, Awad et al. 2020] que contemplam outras estratégias para o esforço de predição, o que será objeto de estudo na continuidade da pesquisa.

Por fim, oportuno enfatizar que a performance obtida pela abordagem discutida neste artigo é superior aos escores tradicionais SAPS-II (AUC de 0,78) e SOFA (AUC de 0,71) [Pirracchio and at al. 2015], o que constitui um indicador bastante significativo para continuidade das pesquisas.

5. Considerações Finais

Este trabalho teve como contribuição central à concepção de uma abordagem explorando Aprendizado de Máquina para predição da mortalidade hospitalar utilizando dados coletados durante as primeiras 48 horas de internação nas UTIs. O MIMIC-III foi utilizado como banco de dados de informações clínicas coletadas do mundo real.

Comparando aos escores clínicos tradicionais, a abordagem proposta utiliza técnicas de mineração de dados e Aprendizado de Máquina que geram modelos mais sofisticados, robustos e capazes de detectar padrões ocultos, lidar com grandes quantidades de dados e ter maior poder de discriminação na classificação de mortalidade em UTIs.

Como perspectiva de continuidade da pesquisa, a expectativa é que esta abordagem tenha sua performance preditiva otimizada, na expectativa que seja possível empregar a mesma em um ambiente clínico real, como apoio à tomada de decisões considerando a celeridade inerente aos ambientes de cuidado intensivos.

6. Agradecimentos

O presente trabalho foi realizado com apoio do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001 e Instituto Federal do Rio Grande do Sul (IF-RS).

Referências

Alghatani, K., Ammar, N., Rezgui, A., and Shaban-Nejad, A. (2021). Predicting intensive care unit length of stay and mortality using patient vital signs: Machine learning model development and validation.

Awad, A., Bader-El-Den, M., McNicholas, J., Briggs, J., and El-Sonbaty, Y. (2020). Predicting hospital mortality for intensive care unit patients: Time-series analysis. *Health Informatics Journal*, 26(2).

Baker, S., Xiang, W., and Atkinson, I. (2020). Continuous and automatic mortality risk prediction using vital signs in the intensive care unit: a hybrid neural network approach.

Bhattacharya, S. and et al. (2017). ICU mortality prediction: A classification algorithm for imbalanced datasets.

Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30(7):1145–1159.

Churpek, M. M. and at al. (2016). Multicenter Comparison of Machine Learning Methods and Conventional Regression for Predicting Clinical Deterioration on the Wards. *Critical Care Medicine*, 44(2):368–374.

Faceli, K., Lorena, A. C., Gama, J., de Carvalho, A. C. P. d. L. F., and de Almeida, T. A. (2021). *Inteligência Artificial - Uma Abordagem de Aprendizado de Máquina*. 2a edição edition.

Harutyunyan, H., Khachatrian, H., Kale, D. C., Ver Steeg, G., and Galstyan, A. (2019). Multitask learning and benchmarking with clinical time series data. *Scientific Data*, 6(1).

Johnson, A. E. and at al. (2016). MIMIC-III, a freely accessible critical care database. *Scientific Data*.

Johnson, A. E. W., Dunkley, N., Mayaud, L., Tsanas, A., Kramer, A. A., and Clifford, D. (2012). Patient Specific Predictions in the Intensive Care Unit Using a Bayesian Ensemble. (Mimic):249–252.

Muralitharan, S. and at al. (2021). Machine learning–Based early warning systems for clinical deterioration: Systematic scoping review. *Journal of Medical Internet Research*, 23(2).

Pirracchio, R. and at al. (2015). Mortality prediction in intensive care units with the Super ICU Learner Algorithm (SICULA): A population-based study. *The Lancet Respiratory Medicine*, 3(1):42–52.

Purushotham, S. and at al. (2018). Benchmarking deep learning models on large healthcare datasets.