



**UNIVERSIDADE ESTADUAL DE FEIRA DE SANTANA**

Autorizada pelo Decreto Federal nº 77.496 de 27/04/76  
Recredenciamento pelo Decreto nº 17.228 de 25/11/2016



**PRÓ-REITORIA DE PESQUISA E PÓS-GRADUAÇÃO**  
COORDENAÇÃO DE INICIAÇÃO CIENTÍFICA

## **XXIX SEMINÁRIO DE INICIAÇÃO CIENTÍFICA DA UEFS** **SEMANA NACIONAL DE CIÊNCIA E TECNOLOGIA - 2025**

### **Utilização de Aprendizado de Máquina na UTI para Predição de Mortalidade de Pacientes**

**Filipe Carvalho Matos Galvão<sup>1</sup>; Tiago Amador Coelho<sup>2</sup>**

1. Filipe Carvalho Matos Galvão, PROBIC/UEFS, ENGENHARIA DE COMPUTAÇÃO, e-mail: filipeglv7@gmail.com

2. Tiago Amador Coelho, DEPARTAMENTO DE CIÊNCIAS EXATAS, e-mail: tiago@uefs.br

**PALAVRAS-CHAVE:** Aprendizado de Máquina; MIMIC-III; Predição de Mortalidade

### **INTRODUÇÃO**

Este relatório técnico final descreve o desenvolvimento e a avaliação de modelos de aprendizado de máquina para predição de mortalidade em pacientes de UTI, utilizando o banco de dados **MIMIC-III**. O projeto, conduzido no âmbito do programa de Iniciação Científica e Tecnológica da UEFS, buscou aplicar técnicas avançadas para extrair insights preditivos de um complexo conjunto de dados de saúde. O objetivo principal é desenvolver um modelo que possa auxiliar na tomada de decisões clínicas rápidas e precisas, identificando pacientes de alto risco de forma precoce.

O estudo empregou uma metodologia rigorosa, que incluiu o pré-processamento e a sumarização de uma vasta quantidade de dados, o balanceamento de classes com a técnica **SMOTE** (*Synthetic Minority Over-sampling Technique*), e a otimização de hiperparâmetros usando a biblioteca *Optuna*. Foram utilizados dois algoritmos de destaque em aprendizado de máquina, *Random Forest* e *XGBoost*, conhecidos por sua robustez e alta performance. Os resultados mostraram que o modelo foi capaz de atingir um desempenho satisfatório, com a constatação de que as informações mais críticas para a predição de mortalidade já estão contidas nas primeiras 36 horas de internação.

### **MATERIAL E MÉTODOS OU METODOLOGIA (ou equivalente)**

O projeto usou a base de dados **MIMIC-III**. Foram selecionadas seis das 26 tabelas disponíveis: *ADMISSIONS*, *PATIENTS*, *DIAGNOSES ICD*, *CHARTEVENTS*, *D\_ITEMS* e *ICUSTAYS* [Johnson et al. 2016, PhysioNet 2016]. A tabela *chartevents*, a maior do banco de dados (34,5 GB), foi otimizada devido a limitações de hardware. O objetivo inicial era usar dados das primeiras 24 horas de internação, mas testes também foram realizados com 36 e 72 horas para obter melhores resultados.

As principais etapas metodológicas incluíram:

- **Otimização da tabela *chartevents*:** Foram excluídos procedimentos com menos de 100.000 aparições. A tabela foi transformada para que cada paciente tivesse uma linha por hora, contendo valores agregados como máximo, médio, mínimo (para registros numéricos) e mais recorrente ou último registrado (para registros categóricos).
- **Normalização Z-Score:** Aplicada para padronizar as características e evitar que escalas muito diferentes afetassem o desempenho do modelo.
- **Sumarização do quadro do paciente:** Dados sociais, étnicos, geográficos e hospitalares de internações anteriores à UTI foram condensados em uma única entrada para aumentar a precisão do modelo.
- **Análise Exploratória dos Dados:** Análise da distribuição de idade, relação entre diagnóstico e tempo de morte, tipo de admissão e a correlação entre mortes e média de dias de internação.
- **Seleção das *features*:** A seleção de *features*, ou variáveis, foi uma etapa crucial para garantir a confiabilidade dos dados e a precisão do modelo. A abordagem inicial focou em incluir apenas as variáveis presentes em pelo menos 100 mil registros na tabela.
- **Seleção dos modelos:** Foram escolhidos os algoritmos de árvore de decisão *Random Forest* e *XGBoost*.
- **Pré-processamento e Divisão dos Dados:** Dados ausentes foram preenchidos com a média e o conjunto de dados foi dividido em 80% para treino e 20% para teste.
- **Balanceamento de Classes:** A técnica **SMOTE** (*Synthetic Minority Over-sampling Technique*) foi utilizada para lidar com o desequilíbrio entre a classe minoritária (pacientes falecidos) e a majoritária (sobreviventes).
- **Otimização de Hiperparâmetros:** A biblioteca *Optuna* foi usada para encontrar a combinação ideal de hiperparâmetros, maximizando métricas como Acurácia, ROC AUC e F1-Score.

O projeto utilizou um limiar de probabilidade de 0.3 para o modelo *XGBoost* e 0.57 para o *Random Forest*, priorizando o *Recall* (sensibilidade) para identificar corretamente pacientes em risco, mesmo que isso resultasse em mais falsos positivos. Em cenários como predição de infarto ou mortalidade hospitalar, o recall é frequentemente priorizado em relação à precisão [Chen et al. 2023]. O documento discute que o custo de um falso negativo (não identificar um paciente de alto risco) é muito mais grave do que o de um falso positivo, que pode gerar custos e ansiedade, mas geralmente é tolerável.

## RESULTADOS E/OU DISCUSSÃO (ou Análise e discussão dos resultados)

Os resultados do projeto demonstraram que a **sumarização dos sintomas** durante a preparação dos dados foi crucial para o bom desempenho dos modelos. Os modelos treinados com dados das primeiras 24 horas sem essa sumarização obtiveram resultados insatisfatórios e são visualizáveis na Tabela 1.

O estudo comparou o desempenho dos modelos com dados de 24, 36 e 72 horas de internação. Os resultados mais notáveis são:

- Com dados de **24 horas**, os modelos obtiveram resultados satisfatórios, com o *XGBoost* apresentando uma Área sob a Curva (AUC-ROC) de 0.92, superando o *Random Forest* (0.91).
- Com dados de **36 horas**, o desempenho aumentou, com o **XGBoost** alcançando uma AUC-ROC de 0.94 e um *Recall* de 0.77, mostrando sua superioridade em relação ao Random Forest.
- Com dados de **72 horas**, os resultados foram os mesmos obtidos com 36 horas, indicando que as informações mais críticas para a predição da mortalidade estão concentradas nas primeiras 36 horas de internação.

| Modelo                   | Acurácia | AUC-ROC | Precisão | Recall | F1-Score |
|--------------------------|----------|---------|----------|--------|----------|
| Random Forest 24 pre sum | 0.80     | 0.84    | 0.32     | 0.68   | 0.43     |
| XGBoost 24 pre sum       | 0.83     | 0.86    | 0.37     | 0.68   | 0.48     |
| Random Forest 24         | 0.88     | 0.91    | 0.44     | 0.67   | 0.53     |
| XGBoost 24               | 0.89     | 0.92    | 0.47     | 0.72   | 0.57     |
| Random Forest 36         | 0.90     | 0.92    | 0.49     | 0.69   | 0.57     |
| XGBoost 36               | 0.91     | 0.94    | 0.53     | 0.77   | 0.63     |
| Random Forest 72         | 0.90     | 0.92    | 0.49     | 0.69   | 0.57     |
| XGBoost 72               | 0.91     | 0.94    | 0.53     | 0.77   | 0.63     |

**Tabela 1. Comparativo de desempenho dos modelos na predição de mortalidade.**

**CONSIDERAÇÕES FINAIS (ou Conclusão)**

O projeto demonstrou que é viável aplicar técnicas de aprendizado de máquina para prever a mortalidade em pacientes de UTI usando o banco de dados MIMIC-III. A otimização dos dados, o uso de SMOTE e a normalização Z-score foram fundamentais para a robustez do modelo. A escolha estratégica de um limiar de probabilidade não padrão, que priorizou a identificação de pacientes de alto risco, reforçou a importância clínica do modelo.

A conclusão mais significativa do projeto é que os dados das primeiras **36 horas** de internação já contêm as informações mais críticas para a predição de mortalidade, permitindo a tomada de decisões clínicas mais rápidas e eficientes sem a necessidade de um período de observação mais longo.

**REFERÊNCIAS**

Chen, X. et al. (2023). Machine learning-based prediction of in-hospital mortality for acute myocardial infarction patients. *Applied Sciences*, 15(16):9192.

Johnson, A., Pollard, T., Shen, L. *et al.* MIMIC-III, a freely accessible critical care database. *Sci Data* 3, 160035 (2016). <https://doi.org/10.1038/sdata.2016.35>

PhysioNet (2016). MIMIC-III clinical database. <https://physionet.org/content/mimiciii/1.4/>. Acesso em: 14 set. 2025.