Machine Learning aplicado para previsão do tempo de permanência de pacientes em UTIs

Caio César Silva

Disciplina: Monografia em Sistemas de Informação II

Orientador: Adriano Alonso Veloso

Departamento de Ciência da Computação – Universidade Federal de Minas Gerais – Belo Horizonte, Brasil

caiocs2019@ufmg.br

Abstract. *This monograph addresses the application of Machine Learning techniques to solve a critical challenge in the context of medicine and hospital management: the overcrowding of Intensive Care Units (ICUs). Overcrowding in ICUs results in adverse consequences such as increased mortality, greater risk of hospital-acquired infections, delayed care, overload on the medical team and increased hospital costs. The central objective of this research is to develop a predictive Machine Learning model capable of predicting the length of stay of a patient in the ICU based on data collected in the first hour of hospitalization, comparing with a model that uses SAPS-3, with the aim of assisting in the management of these units.*

Resumo. *Esta monografia aborda a aplicação de técnicas de Machine Learning para solucionar um desafio crítico no contexto da medicina e da gestão hospitalar: a superlotação das Unidades de Terapia Intensiva (UTIs). A superlotação nas UTIs resulta em consequências adversas como aumento da mortalidade, maior risco de infecções hospitalares, retardo no atendimento, sobrecarga da equipe médica e aumento dos custos hospitalares. O objetivo central desta pesquisa é desenvolver um modelo preditivo de Machine Learning capaz de prever o tempo de permanência de um paciente na UTI com base nos dados coletados na primeira hora de internação, comparando com um modelo que usa o SAPS-3, com o objetivo de auxiliar na gestão destas unidades.*

1. Introdução

A presente monografia aborda a solução de um desafio crítico na medicina e gestão hospitalar: a superlotação das Unidades de Terapia Intensiva (UTIs). As UTIs são essenciais para o atendimento de pacientes graves, necessitando de monitoramento constante. Porém, a superlotação impacta negativamente a qualidade do atendimento, aumenta os custos hospitalares e sobrecarrega os profissionais de saúde, resultando em um aumento da mortalidade, maior risco de infecções hospitalares e retardos no atendimento. Para abordar esse problema, esta pesquisa propõe o uso de Machine Learning para prever o tempo de permanência de um paciente na UTI com base em dados fisiológicos e laboratoriais coletados na primeira hora de internação, comparando

com um modelo que usa o SAPS-3, um escore de mortalidade que é calculado em um tempo maior, que possivelmente também pode ser usado para este propósito.

A pesquisa utiliza uma base de dados fornecida pelo orientador Adriano Veloso, com informações de pacientes internados em três UTIs do Hospital Life Center em Belo Horizonte (MG), entre janeiro de 2017 e dezembro de 2018. Os dados incluem identificação, demografia, diagnósticos, CID-10, evoluções na UTI e no hospital, comorbidades, capacidade funcional, razões para internação, complicações no primeiro dia de internação, dados fisiológicos e laboratoriais (1h), uso de suporte e complicações na UTI. Os objetivos do trabalho são desenvolver um modelo de Machine Learning para prever o tempo de permanência na UTI, realizar limpeza e preparação dos dados, análise exploratória e avaliar a eficácia dos modelos propostos.

2. Referêncial Teórico e Trabalhos Correlatos

Para a realização deste projeto foi estudada uma vasta literatura científica da área, contando com vários artigos que se tratavam do uso de Machine Learning, não só para predição do tempo de permanência do paciente em UTIs, mas também predição de outros fatores, como complicações de saúde. Mesmo que estas não sejam exatamente o que este projeto esteja buscando, muito da metodologia e dos desafios encontrados nestes outros artigos serviram de inspiração e aprendizado.

O primeiro estudo, que se destaca dos que foram analisados, trata da construção de modelos de aprendizado de máquina para prever complicações severas usando dados administrativos e clínicos coletados logo após a admissão do paciente na unidade de terapia intensiva. Neste estudo foram construídos modelos que levavam em conta a medição de como as explicações (das complicações do paciente) fornecidas variam para diferentes pacientes, ou seja, a robustez da solução, assim como as explicações providas mudam com modelos construídos a partir de diferentes sub-populações de pacientes, ou seja, a estabilidade do modelo. O estudo mostrou que a escolha dos dados mais importantes para a decisão do modelo, quando definidos usando o critério de robustez, levaram a ganhos de potencial preditivo variando de 6.8% a 9.4%. Assim como a seleção de features para o modelo baseado no critério da estabilidade, levaram a ganhos de 7.2% a 11.5%. [Amador et al., 2022]

Outro estudo analisado compara métodos de regressão para a modelagem do tempo de permanência de pacientes em unidades de terapia intensiva. Neste estudo foram utilizados dados de 32,667 admissões em UTI não planejadas. Foram utilizados oito modelos de regressão:

1. Regressão de mínimos quadrados ordinários no tempo de permanência.
2. Regressão de mínimos quadrados ordinários no tempo de permanência truncado a 30 dias.
3. Regressão de mínimos quadrados ordinários no tempo de permanência transformado pelo logaritmo.
4. Modelo linear generalizado com distribuição Gaussiana e função de ligação logarítmica.
5. Regressão de Poisson.
6. Regressão binomial negativa.

7. Regressão Gamma com função de ligação logarítmica.
8. Modelo APACHE IV original e recalibrado, tanto para todos os pacientes quanto separadamente para sobreviventes e não sobreviventes.

O resultado final deste estudo foi que os modelos não tiveram uma performance preditiva boa, tendo os modelos de regressão explicando apenas 20% da variabilidade observada nos dados. Apesar disto, o estudo se destaca por uma série de fatores, como o uso de métodos avançados de modelagem, a grande escala de dados e termos cíclicos. [I. Verburg et al., 2014]

Também foi analisado um estudo no qual foi proposto um framework para prever o tempo de permanência dos pacientes na UTI utilizando diferentes técnicas de aprendizado de máquina. Este estudo também tem destaque por utilizar dados médicos gerais coletados na admissão do paciente, o que se assemelha bastante ao projeto deste relatório. Este estudo acabou tendo como conclusão dois modelos de aprendizado de máquina que foram bastante precisos em suas predições, no modelo que utilizava a técnica fuzzy, obteve uma precisão de 92% e no modelo que utilizava árvore de classificação, obteve uma precisão de 90%. Apesar de alguns desafios enfrentados durante a implementação, como trade-off entre viés e variância, ajuste excessivo e insuficiente, várias tentativas foram realizadas para superar os problemas, fornecendo um framework com resultado eficiente e estável ao longo de diversos experimentos. [Merhan A. Abd-Elrazek et al., 2021]

Houve também um estudo realizado utilizando dados de 109 UTIs do Brasil, de tipos variados em 38 hospitais, assim como validação externa em 93 UTIs médico-cirúrgicas de 55 hospitais. Este estudo visa desenvolver uma metodologia para previsão do tempo de permanência do paciente na unidade de terapia intensiva, assim como o risco de estadia prolongada. Com isso, foi desenvolvido um modelo utilizando técnica de Random Forest e Regressão Linear, que mostrou-se preciso em uma grande coorte multicêntrica de pacientes gerais de UTI. Este artigo também se destaca por sua análise que envolve uma seleção das melhores features para o modelo, que começou com 90 e terminou com 32, após a filtragem com base na importância. [Peres, Igor Tona et al., 2022]

Em um outro estudo, também foi proposto um modelo para predição do tempo de permanência, só que neste caso, é o tempo de permanência no hospital, e não somente na UTI, assim como a base só possuía pacientes que sofreram acidente vascular cerebral (AVC). Utilizando uma abordagem baseada em dados, foram analisados dados de 16,592 pacientes adultos com AVC isquêmico ou hemorrágico de 130 UTIs de 43 hospitais, entre 2011 e 2020. Modelos de aprendizado de máquina, incluindo Random Forests, foram aplicados e demonstraram eficácia na previsão de tempo de permanência prolongado e mortalidade. Fatores como condições pré-mórbidas, disfunção de múltiplos órgãos e aspectos neurológicos do AVC foram identificados como importantes para essas previsões. Concluiu-se que esses modelos podem ser úteis para planejamento de alocação de recursos e avaliação de desempenho em UTIs que tratam pacientes com AVC. [Kurtz, Pedro et al., 2022]

Vale citar que foram consultados estudos envolvendo o escore de gravidade SAPS-3 (Simplified Acute Physiology Score III) que é um sistema muito utilizado em unidades de terapia intensiva para prever a mortalidade do paciente. O uso deste modelo gera custo, portanto, como é detalhado posteriormente neste relatório, foi feito a

comparação do modelo que utiliza somente o SAPS-3 (como baseline), com um modelo que utiliza dos dados de primeira hora de admissão do paciente e com outro modelo que utiliza tanto do SAPS-3 como destes dados de primeira hora também. Portanto, vale uma análise acerca do que o SAPS-3 busca propor para os hospitais, assim como sua metodologia e eficácia do ponto de vista da literatura científica.

Com isso, também há um estudo que busca revisar as versões mais recentes dos sistemas de escore de gravidade: Acute Physiology and Chronic Health Evaluation (APACHE), Simplified Acute Physiology Score (SAPS) e Mortality Probability Model (MPM), comparando suas características e descrevendo suas forças e limitações. Os sistemas mais recentes (APACHE IV, MPM 0-III e SAPS-3) foram amplamente validados globalmente em diversos contextos, incluindo em pacientes gerais de UTI e subgrupos específicos como pacientes com condições graves, como câncer, problemas cardiovasculares, cirúrgicos, com lesão renal aguda necessitando de terapia de substituição renal e aqueles que necessitam de oxigenação por membrana extracorpórea. Os resultados são promissores, mostrando uma boa capacidade de discriminação. Porém, a calibração pode não ser tão precisa quanto a dos estudos originais, para melhorar a precisão dos escores, é recomendável fazer algumas adaptações. Em resumo, esses sistemas são valiosos para caracterizar a gravidade da doença dos pacientes, avaliar o desempenho das UTIs e iniciar melhorias na qualidade. Eles também são úteis para benchmarking, apesar disso, atualizações contínuas e personalizações regionais são necessárias para garantir a precisão ideal. [Salluh, Jorge Ibrain Figueira e Márcio Soares., 2014]

Com isso, a partir dos estudos analisados, pretende-se seguir este projeto através de variadas análises e metodologias. Será possível o uso de diferentes modelos e técnicas de análise para se verificar a construção de um modelo para a predição do tempo de permanência dos pacientes nas unidades de terapia intensiva.

3. Metodologia

Toda a implementação deste projeto foi realizada no Google Colab com Python, um serviço de uso de Jupyter Notebooks porém com a hospedagem fornecida pela Google. A base de dados foi fornecida pelo orientador Adriano, em um arquivo no formato Excel (xlsx), com uma grande variedade de dados.

3.1. Dados disponíveis

Os dados foram coletados pela Epimed, uma empresa que oferece sistemas para ajudar a melhorar o desempenho de hospitais, incluindo UTIs. A base possui dados de pacientes internados em três UTIs médicas e cirúrgicas mistas do Hospital Life Center em Belo Horizonte (MG), abrangendo o período de 01 de janeiro de 2017 até 31 de dezembro de 2018, contendo 7,085 registros ao todo. A base possui dados como:

- Identificação, Dados Demográficos, Internação na UTI, Diagnósticos, CID-10 e Evoluções na UTI e no Hospital
- Comorbidades e Capacidade Funcional
- Razões para internação na UTI
- Complicações no Primeiro Dia de Internação na UTI

- Dados Fisiológicos e Laboratoriais (1h)
- Uso de Suporte e Complicações na UTI

Com estes dados disponíveis foi possível realizar uma análise muito rica, mas antes, foi necessário realizar uma preparação para que as devidas análises sejam realizadas, e para que o modelo receba estes dados de entrada da melhor forma possível. Portanto, foi utilizado técnicas de limpeza, pré processamento e organização, para que o modelo tivesse o melhor resultado possível dentro das limitações do projeto.

3.2. Preparação dos dados

Antes de realizar uma análise exploratória dos dados, foi necessário uma preparação destes para eliminar quaisquer ocorrências de dados incorretos ou desnecessários. Para essa preparação, foi a realizado a remoção de colunas com dados faltantes, para evitar problemas que eles pudessem causar no modelo, além disso, também foi eliminado todas a colunas que dizem respeito a algum tipo de código que faça mapeamento com alguma entidade de negócio, para evitar quaisquer problemas que envolvam regras de negócio das quais este projeto não teria acesso. Todos os dados que possivelmente se tratavam de dados binários também foram convertidos para o formato correto da linguagem Python, o formato booleano, o que permite o entendimento melhor pelos modelos e para a criação das visualizações e análises que vieram a ser necessárias. Além disso, todas as planilhas que se encontravam no arquivo Excel cujo a base estava, foram convertidas para dataframes da biblioteca Pandas, dessa forma possibilitando uma integração melhor com a linguagem Python.

Com isso, se possui todas as tratativas necessárias para se realizar uma análise rica, e também a criação de modelos de Aprendizado de Máquina, dessa forma, este estudo prossegue com a realização de uma Análise Exploratória dos Dados.

3.3. Análise exploratória dos dados

Começa-se a Análise Exploratória dos dados com uma análise da distribuição do target inicial, que é o tempo de permanência.

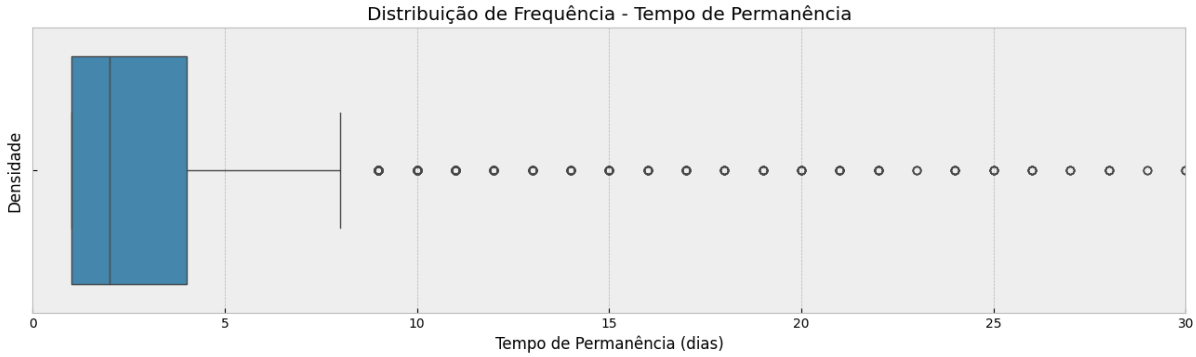

Figura 1. Distribuição do tempo de permanência

Conforme pode-se notar na Figura 1, o target tem uma distribuição muito concentrada entre 0 e 9 aproximadamente, acima disso há uma quantidade menor de exemplos, o que torna bem complexo realizar uma previsão do target sendo ele contínuo.

Com isso, segue-se para a realização de análises bivariadas, ou seja, uma analise da relação de cada variável individualmente com o target. Para isso, foi separado as variáveis categóricas entre dicotômicas e politômicas, porém não foi encontrado nenhuma relação que fosse muito evidente. O mesmo ocorre quando realiza-se uma análise bivariada utilizando como comparação as variáveis numéricas.

Concluindo esta análise, pode-se notar que o target é uma variável muito complexa para ser analisada individualmente com outras variáveis, o que faz muito sentido, pois por se tratar do tempo de permanência do paciente na unidade de terapia intensiva, não seria arriscado supor que ela depende de uma série de fatores que ocorram em conjunto. Ou seja, uma análise multivariada se faz pertinente para que seja observado quais os conjuntos de fatores mais importantes que levam a determinação dessa variável. Outro ponto que se pode concluir, a partir da distribuição do target, é o fato de que os pacientes que ficaram mais de nove dias são considerados outliers, dessa forma, eles teriam impacto no modelo.

A solução encontrada foi a alteração do target para categórico dicotômico, ou seja, ao invés de utilizar o modelo de aprendizado de máquina para encontrar o tempo de permanência do paciente na UTI, o foco muda para a verificação de se o modelo consegue prever se o paciente vai ficar mais de 9 dias ou não, que foi o valor “limiar” para o que se define como outlier na nossa base de dados. Isso ainda assim torna a pesquisa de extrema importância, pois, encontrar pacientes que podem ser considerados outliers quanto ao seu tempo de permanência ainda assim pode evitar muitos custos desnecessários por conta da administração da unidade, assim como auxiliar em outros pontos da gestão e da logística dela.

3.4. Desenvolvimento dos modelos

Para enriquecer a análise, foi desenvolvido e comparado o resultado de três modelos. Em todos os modelos foi previsto se o paciente ficou ou não mais de nove dias internado na UTI, o que muda entre esses modelos serão as features que eles terão. O primeiro modelo utiliza apenas as colunas que dizem respeito ao SAPS-3, dessa forma, tem-se um modelo "baseline" e ao final da implementação dos três modelos foi possível avaliar também o quanto o SAPS-3, apesar de gerar custos para a gestão da unidade, pode contribuir para o resultado final da predição. O segundo modelo tem todas as colunas dos dados fisiológicos e laboratoriais coletados na primeira hora de admissão do paciente na unidade, exceto o SAPS-3, assim, é avaliado se é possível construir um modelo que tenha sua eficácia próxima da capacidade de predição que o SAPS-3 pode fornecer, pois, apesar de ser um dado muito importante, o SAPS-3 não é uma métrica calculada na primeira hora, portanto não seria “justo” realizar uma comparação que tente superar o SAPS-3 em termos de predição. O terceiro modelo conta com o SAPS-3 e também com os demais dados fisiológicos e laboratoriais coletados na primeira hora, assim, é comparado com o modelo baseline e com o modelo sem SAPS-3, utilizando a medida de importância com as demais features do modelo.

Para o desenvolvimento dos modelos foi utilizado o LightGBM Classifier, que é um algoritmo de boosting de gradiente que constroi múltiplas árvores de decisão de forma iterativa, ajustando cada uma para corrigir os erros da anterior. Ele também fornece uma lista de quais features foram as mais importantes, o que permite analisar quais são as que tiveram maior impacto na sua decisão.

Também foi utilizado uma técnica de otimização dos modelos, o GridSearch, que realiza uma busca exaustiva testando várias combinações de hiperparâmetros até encontrar uma configuração que proporciona o melhor desempenho para o modelo. Ele funciona ao observar uma série de valores para cada parâmetro, e em seguida, treinar e validar o modelo em cada combinação destes valores. Com isso ele encontra os melhores parâmetros, permitindo o modelo capturar padrões nos dados mais efetivamente, e evitando problemas de ajuste excessivo ou insuficiente.

Como os dados estão desbalanceados, ou seja, há uma proporção muito maior de pacientes que ficaram menos de 9 dias comparado aos que ficaram mais que isso, foi necessário utilizar de uma técnica chamada Undersampling juntamente com K-means. Essa técnica envolve reduzir a quantidade de dados da classe majoritária (no caso, os pacientes que ficaram menos de 9 dias) usando um algoritmo de clusterização (K-means) para selecionar as representações que irão manter a classe majoritária mais diversificada. Com isso, foi possível manter um equilíbrio entre as classes quanto a sua quantidade de dados, ajudando o modelo a ter uma representação das classes mais equilibrada e variada, o que pode melhorar sua precisão e robustez ao realizar a predição.

Dessa forma, tem-se todos os preparativos para o desenvolvimento e implementação dos três modelos. A base de dados foi separada entre treino e teste, separando 30% dos dados para teste e o restante para treino. A técnica de Undersampling foi aplicada somente à base de treino, dessa forma a base de teste foi mantida intacta, o que permite uma validação mais fiel e concisa. Todo os procedimentos informados neste capítulo foram aplicados aos três modelos, o que também fez com que se possa comparar os modelos de maneira mais adequada.

3.5. Resultados dos modelos

Após a realização do treino de cada modelo e sua validação na base de teste, foi obtido os resultados para comparar os três modelos de forma adequada. Dessa forma, foi avaliado e comparado o resultado de cada modelo. Para a avaliação dos resultados também foram utilizadas técnicas que forneceram dados muito valiosos acerca da performance de cada um.

Uma das técnicas que foi utilizada para analisar os resultados dos modelos é a curva ROC (Receiver Operating Characteristic), que permite visualizar a performance de um modelo ao exibir sua taxa de verdadeiros positivos, contra a taxa de falsos positivos. A área sob a curva ROC (AUC-ROC) realiza a quantificação da performance em um intervalo de 0 a 1, onde os valores próximos de 1 indicam um modelo com uma boa capacidade de discriminação. Com isso, pode-se efetivamente avaliar se os modelos têm um equilíbrio adequado entre sensibilidade e especificidade.

Juntamente à curva ROC, foi utilizada outra ferramenta para avaliar os resultados dos modelos, que é a matriz de confusão. A matriz de confusão permite visualizar o desempenho do modelo ao mostrar os resultados das predições em uma tabela de contingência. Nela, é possível ver os valores verdadeiros positivos, verdadeiros negativos, falsos positivos e falsos negativos, o que fornece a análise de onde o modelo está acertando, e onde ele está errando.

Também foram utilizadas de métricas para avaliar os resultados dos modelos, no caso, quatro métricas muito importantes: acurácia, precisão, recall e F1 Score. A

acurácia mede a proporção total de predições corretas em relação ao número total de casos, o que mostra uma boa visão da performance geral do modelo. A precisão indica a porcentagem de predições positivas que são verdadeiramente positivas, ou seja, há a possibilidade de avaliar a necessidade de atuar na minimização de falsos positivos. O recall (sensibilidade) mede a proporção de verdadeiros positivos identificados corretamente pelo modelo, se diferenciando da precisão ao focar na quantidade de verdadeiros positivos encontrados, enquanto que a precisão foca na qualidade das predições positivas. O F1 Score equivale a média harmônica entre precisão e recall, mostrando o equilíbrio entre eles, o que é muito útil em casos onde há um desbalanceamento entre as classes. Com essas quatro métricas pode-se ter uma análise detalhada do desempenho de cada modelo.

3.6. Comparação dos resultados dos três modelos

Após o treino de cada modelo e a obtenção de suas métricas, além do uso das ferramentas citadas para avaliação de seus resultados, foi necessário comparar os resultados de cada modelo, o que é feito nesta seção.

Métricas dos modelos	Acurácia		Precisão		Recall		F1 Score	
	Treino	Teste	Treino	Teste	Treino	Teste	Treino	Teste
Modelo baseline (somente SAPS-3)	0.696	0.678	0.697	0.487	0.693	0.696	0.695	0.573
Modelo somente dados fisiológicos e laboratoriais de 1h	0.636	0.522	0.616	0.355	0.723	0.660	0.665	0.462
Modelo SAPS-3 + dados fisiológicos e laboratoriais de 1h	0.747	0.671	0.747	0.480	0.749	0.693	0.748	0.567

Tabela 1. Métricas dos modelos referente a treino e teste

Analisando as métricas obtidas na fase de testes, pode-se notar que o modelo baseline, que conta somente com o SAPS-3, apresenta a melhor acurácia, recall e F1 Score (em base de teste), mas uma precisão média, o que pode indicar que ele identifica bem os verdadeiros positivos, mas gera uma quantidade razoável de falsos positivos. O modelo que utiliza apenas os dados fisiológicos e laboratoriais coletados na primeira hora de admissão do paciente tem a menor acurácia, precisão e F1 Score, mas um recall razoável, o que mostra que embora ele capture uma boa proporção de verdadeiros positivos, não é muito preciso e acurado. Agora, o modelo que combina o SAPS-3 com os dados fisiológicos e laboratoriais apresenta uma acurácia, precisão, F1 Score e recall ligeiramente menores que o baseline. O que pode-se tirar disso é que a inclusão dos dados fisiológicos e laboratoriais não adicionam uma melhoria significativa quando já se usa o SAPS-3 no modelo, apesar disso, a combinação ainda é competitiva quando se fala de performance geral.

Um ponto que vale ser observado, é o fato de que este padrão não se repete na base de treino, onde é possível ver que as métricas são ligeiramente superiores, e além disso, o modelo que conta com os dados fisiológicos e laboratoriais, juntamente com o SAPS-3 performa melhor que o baseline e o modelo sem o SAPS-3. Apesar disto, ao se avaliar na base de teste o comportamento muda, trazendo à tona a importância de se manter a base de teste intacta para realizar uma validação mais concisa.

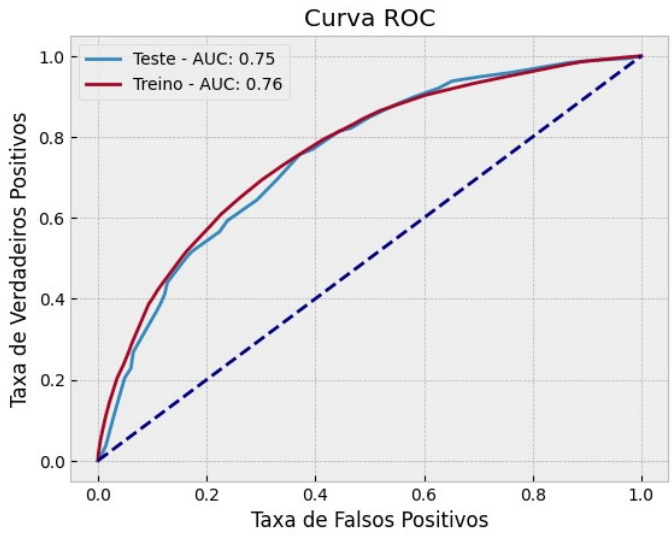

Figura 2. Gráfico da curva ROC do modelo baseline (somente SAPS-3), para as bases de treino e teste

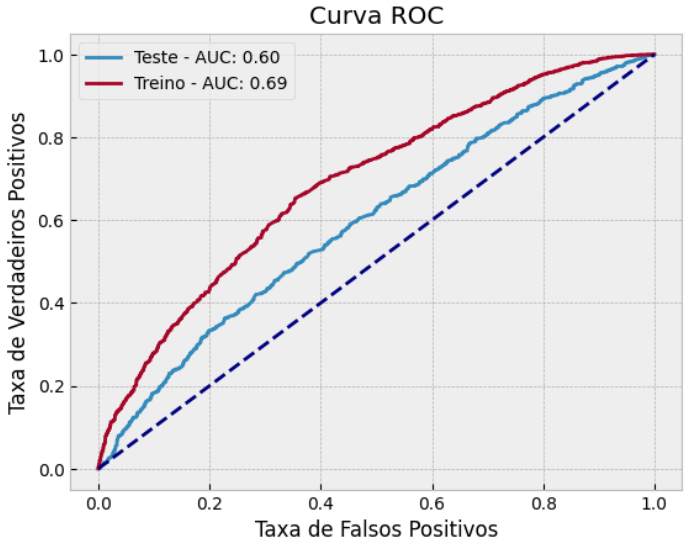

Figura 3. Gráfico da curva ROC do modelo somente com os dados fisiológicos e laboratoriais coletados na primeira hora (sem SAPS-3), para as bases de treino e teste

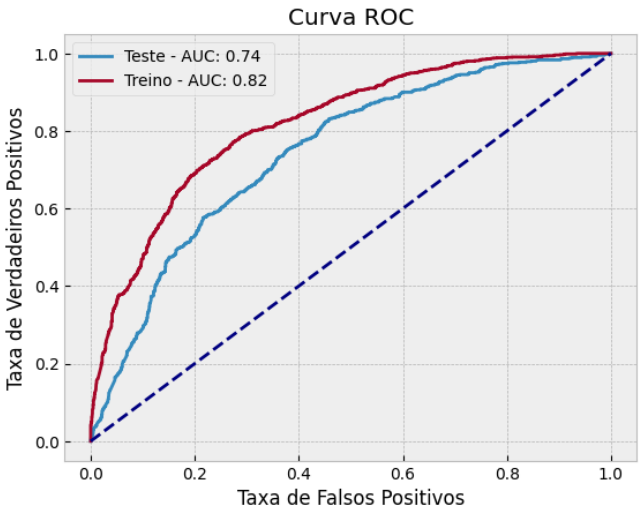

Figura 4. Gráfico da curva ROC do modelo com os dados fisiológicos e laboratoriais coletados na primeira hora e também com o SAPS-3, para as bases de treino e teste

Observando a Figura 2, Figura 3 e a Figura 4 pode-se observar a curva ROC, assim como a área sob a curva (AUC) para os diferentes modelos. O modelo baseline apresenta a maior AUC, indicando que ele tem uma capacidade muito boa em distinguir entre as classes. O modelo que utiliza apenas os dados fisiológicos e laboratoriais coletados na primeira hora tem a menor AUC, o que sugere que ele tem uma performance moderada e uma capacidade menor de discriminar entre as classes. O modelo que combina o SAPS-3 e os dados fisiológicos e laboratoriais tem uma AUC muito próxima do modelo baseline, mas ainda inferior. Todos estes pontos reforçam que a adição dos dados fisiológicos e laboratoriais juntamente com a informação do SAPS-3 não traz uma melhoria significativa para a capacidade discriminativa do modelo. O mesmo padrão é observado na fase de treino, onde o modelo somente com o SAPS-3 apresenta o melhor valor para a métrica AUC, tendo logo em seguida o modelo com os dados fisiológicos e laboratoriais com o SAPS-3, e então o modelo sem o SAPS-3, por último.

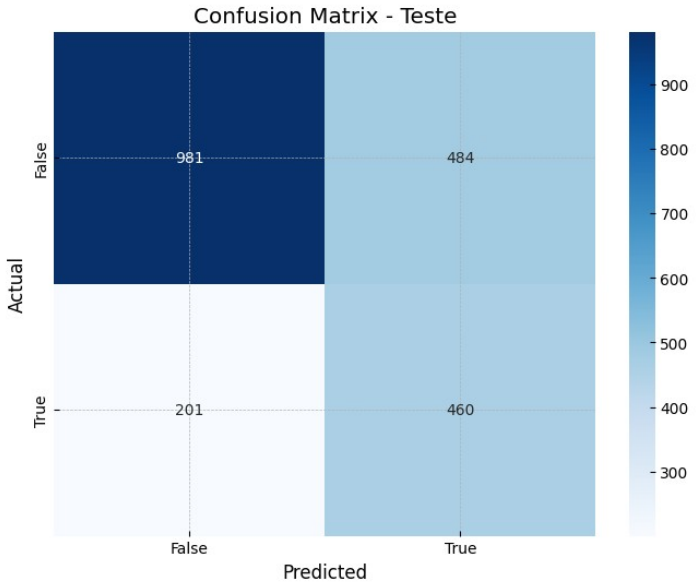

Figura 5. Matriz de confusão para o modelo baseline na base de teste

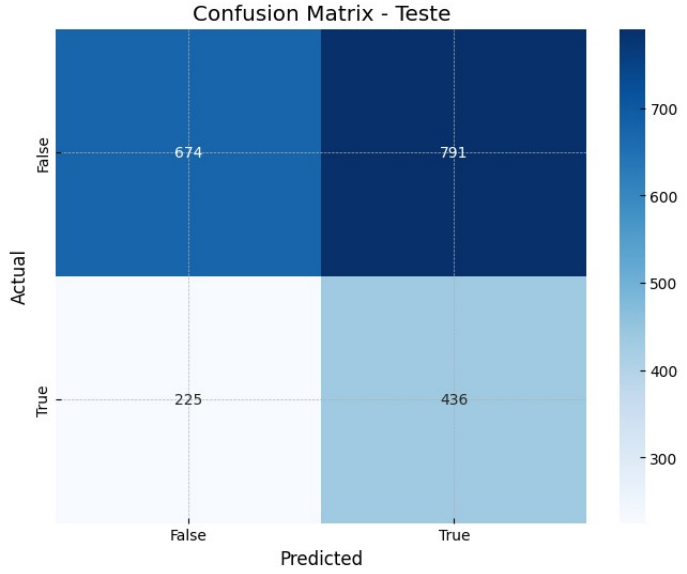

Figura 6. Matriz de confusão para o modelo sem o SAPS-3 na base de teste

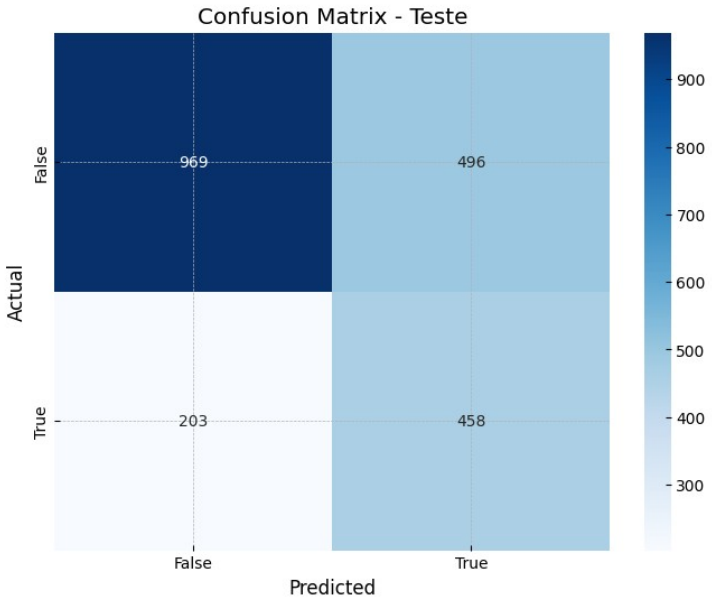

Figura 7. Matriz de confusão para o modelo com os dados fisiológicos e laboratoriais, juntamente com o SAPS-3, na base de teste

É possível observar na Figura 5, Figura 6 e na Figura 7 as matrizes de confusão para os três modelos na base de teste, nelas é possível obter insights bem claros sobre suas performances. No modelo baseline, pode-se observar que há uma boa capacidade de identificação dos casos negativos, ou seja, os pacientes que ficaram menos de 9 dias, mas há uma boa proporção de falsos positivos. O modelo que utiliza apenas os dados fisiológicos e laboratoriais (sem o SAPS-3), mostra uma taxa maior de falsos positivos, além de uma capacidade moderada para identificar casos positivos. Agora, o modelo que combina o SAPS-3 com os dados fisiológicos e laboratoriais mostrou uma melhoria em relação ao modelo sem o SAPS-3, mas ainda tem dificuldades nos casos dos falsos positivos.

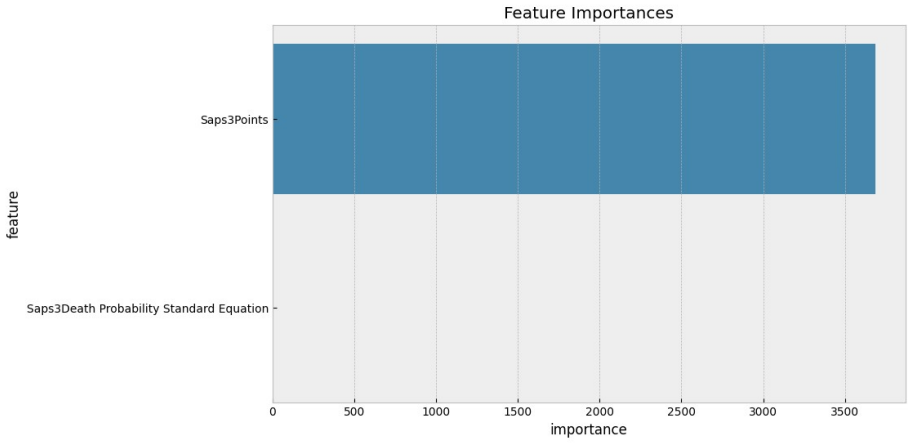

Figura 8. Gráfico de importância de cada feature para o modelo baseline

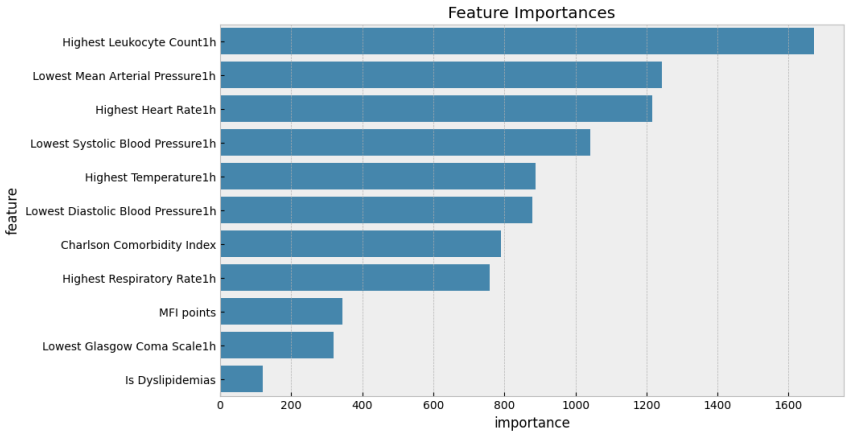

Figura 9. Gráfico de importância de cada feature para o modelo sem o SAPS-3

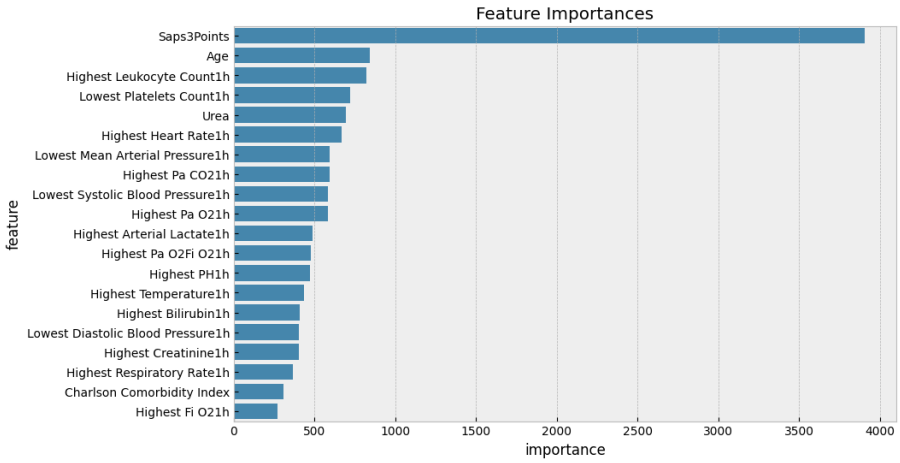

Figura 10. Gráfico de importância de cada feature para o modelo com o SAPS-3 e os dados fisiológicos e laboratoriais coletados na primeira hora de admissão do paciente na unidade

Também é muito importante avaliar a importância de cada feature nos modelos, o que permite verificar o que cada modelo mais levou em conta na hora de se realizar uma predição. Conforme pode-se ver na Figura 8, Figura 9 e Figura 10, nelas se encontram as features mais importantes para cada modelo, ordenadas com base na sua importância. Com isso, observa-se ao notar na Figura 8, que o modelo baseline só levou em conta uma das features referentes ao SAPS-3, que é de fato o escore que ele determina. Na Figura 9 há a importância das features para o modelo sem o SAPS-3, o que permite ver que as cinco features mais importantes foram:

1. Maior contagem de leucócitos
2. Pressão arterial média mais baixa
3. Frequência cardíaca mais alta
4. pressão arterial sistólica mais baixa
5. Temperatura mais alta

Já na Figura 10, onde estão as importâncias de cada feature para o modelo com o SAPS-3 e os dados fisiológicos e laboratoriais coletados na primeira hora, tem-se que as cinco features mais importantes para o modelo foram:

1. Escore SAPS-3
2. Idade
3. Maior contagem de leucócitos
4. Contagem de plaquetas mais baixa
5. Ureia

4. Conclusão

Com as métricas de cada modelo analisadas, segue-se para a conclusão deste trabalho. No modelo baseline, que usa apenas como feature o SAPS-3, foi obtido uma boa performance tanto na base de treino, como na base de teste. O modelo apresenta uma acurácia, precisão e recall bons, além de ter uma capacidade de discriminação boa, conforme foi visto através da métrica AUC. A matriz de confusão deste modelo reforça isso, ao mostrar que há um equilíbrio entre verdadeiros positivos e verdadeiros negativos, apesar de uma quantidade alta de falsos positivos. Já o modelo que usa somente dos dados fisiológicos e laboratoriais coletados na primeira hora, observa-se uma performance geral menor quando se comparado ao baseline. Sua acurácia é mais baixa, tendo um valor de 0.522, uma precisão de 0.355, um recall de 0.660 e um F1 Score de 0.462. Comparado ao baseline, que obteve uma acurácia no valor de 0.678, uma precisão de 0.487, recall no valor de 0.695 e um F1 Score de 0.573. Ainda neste modelo sem o SAPS-3, há uma taxa maior de falsos positivos, e uma AUC menor que a do modelo baseline, tendo um valor de 0.60, comparado ao modelo baseline que obteve 0.75. Isso mostra que, embora os dados coletados na primeira hora de admissão do paciente na UTI sejam importantes, eles não conseguem fornecer uma capacidade preditiva próxima ao que o SAPS-3 fornece. Agora, observando o modelo que combina o SAPS-3 juntamente com os dados coletados na primeira hora de admissão, é possível observar métricas muito próximas do modelo baseline, embora um pouco inferiores em um aspecto geral. A acurácia do modelo baseline teve um valor de 0.678, sua precisão 0.487, recall de 0.695, F1 Score de 0.573 e uma AUC de 0.75. Já este último modelo, que conta com o SAPS-3 e os dados fisiológicos e laboratoriais coletados na primeira hora, contou com uma acurácia de 0.671, precisão de 0.480, recall de 0.693, F1 Score de 0.567 e uma AUC de 0.74. Mostrando que a combinação dos dados não parece trazer uma melhoria significativa em relação ao modelo que conta somente com o SAPS-3.

Dessa forma, quando é analisado as features do modelos, pode-se observar que o SAPS-3 é de fato uma feature "dominante", até quando é combinado com outros dados. No modelo que conta com o SAPS-3 e as demais features, é visto que o SAPS-3, seguido da idade e da maior contagem de leucócitos são as features mais importantes, mas o SAPS-3 ainda permanece no topo. Portanto, quando é feito o questionamento de quais são os dados mais importantes para predição do tempo de permanência do paciente na UTI, pode-se considerar que o SAPS-3 é de fato o dado mais importante.

A combinação do SAPS-3 com os demais dados fisiológicos e laboratoriais não traz uma melhoria significativa na performance do modelo, na verdade, é plausível até dizer que é arriscado que estes dados de primeira hora podem até causar uma certa

"confusão" no modelo. Portanto, a escolha de um modelo mais simples, baseado no escore SAPS-3, pode ser preferido para a previsão do tempo de permanência dos pacientes.

Ao se realizar uma comparação do modelo baseline com o modelo que conta somente com os dados fisiológicos e laboratoriais coletados na primeira hora de admissão do paciente na UTI, fica bem claro, a partir das métricas, que o modelo sem o SAPS-3 não conseguiu atingir uma performance aceitável e próxima do modelo baseline. Com isso, é possível dizer que, devido as limitações dos dados coletados na primeira hora, assim como as limitações da nossa base de dados, não foi possível obter um resultado que chegasse próximo a capacidade preditiva que o escore SAPS-3 fornece. Isso se deve ao fato de que este projeto está lidando com um problema extremamente difícil, onde o tempo de permanência do paciente se deve a uma série de fatores que podem ocorrer de diversas maneiras, e com os dados coletados na primeira hora não foi possível observar com uma performance excelente se o paciente teria um tempo de permanência dentro ou fora do padrão esperado para a maioria dos pacientes de uma UTI. Além disso, o SAPS-3 não é um dado calculado na primeira hora de admissão, ele é coletado posteriormente, ou seja, em um momento onde já se há uma quantidade maior de dados disponíveis para se ter uma visão mais completa do estado do paciente, o que mostra que este projeto está tentando alcançar, com os dados de uma hora de admissão, uma capacidade preditiva próxima ou superior a que um dado que pode ser coletado com um dia de admissão, por exemplo, pode fornecer.

Apesar do modelo somente com o SAPS-3 contar com uma performance robusta e consistente, tanto na base de treino como teste, ainda não é possível afirmar que ele é adequado para uso em produção ou em testes, sendo necessário a realização de validações adicionais, e melhorias no modelo para que isto venha a ser possível. Em estudos futuros, é recomendado contar com técnicas de validação avançadas, como validação cruzada, dividindo os dados em múltiplos subconjuntos, e realizando o processo de treino e teste nestes conjuntos, ajudando a confirmar que o modelo não está super ajustando aos dados de treino. Também pode ser realizado estudos em ambientes controlados, como clínicas ou hospitais, onde há como ver o modelo performando em situações reais.

Para estudos futuros também é recomendado realizar uma série de melhorias no modelo, como a inclusão de novas features relevantes que ainda não foram exploradas, ou até a combinação de features, o que traria consigo uma necessidade de um conhecimento médico, pois iria ajudar a reduzir a complexidade do modelo. Com isso, talvez seria possível construir um modelo que tivesse a performance tão boa quanto o modelo baseline, sem a necessidade do SAPS-3. Outra possível melhoria seria contar com uma base de dados maior, onde não seria necessário o uso de técnicas como Undersampling, que apesar de serem muito úteis, trazem alguns custos a performance do modelo. Caso isto não seja possível, também seria possível contar com uso de outras técnicas para balanceamento de classes, como o Oversampling, que também poderia ajudar na capacidade preditiva do modelo.

5. Referências Bibliográficas

Amador, T., Saturnino, S., Veloso, A., & Ziviani, N. (2022). Early identification of ICU patients at risk of complications: Regularization based on robustness and stability of explanations. Artificial intelligence in medicine, 128, 102283.

Verburg, I.W., de Keizer, N.F., de Jonge, E., & Peek, N. (2014). Comparison of Regression Methods for Modeling Intensive Care Length of Stay. PLoS ONE, 9.

Abd-Elrazek, M.A., Eltahawi, A.A., Elaziz, M.E., & Abd-Elwhab, M.N. (2021). Predicting length of stay in hospitals intensive care unit using general admission features. Ain Shams Engineering Journal.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T. (2017). LightGBM: A Highly Efficient Gradient Boosting Decision Tree. Neural Information Processing Systems.

Peres, I.T., Hamacher, S., Oliveira, F.L., Bozza, F.A., & Salluh, J.I. (2022). Data-driven methodology to predict the ICU length of stay: A multicentre study of 99,492 admissions in 109 Brazilian units. Anaesthesia, critical care & pain medicine, 101142.

Kurtz, P., Peres, I.T., Soares, M., Salluh, J.I., & Bozza, F.A. (2022). Hospital Length of Stay and 30-Day Mortality Prediction in Stroke: A Machine Learning Analysis of 17,000 ICU Admissions in Brazil. Neurocritical Care, 37, 313 – 321.

Peres, I.T., Hamacher, S., Oliveira, F.L., Bozza, F.A., & Salluh, J.I. (2021). Prediction of intensive care units length of stay: a concise review. Revista Brasileira de Terapia Intensiva, 33, 183 – 187.

Verburg, I., Atashi, A., Eslami, S., Holman, R., Abu-Hanna, A., de Jonge, E., Peek, N., & de Keizer, N.F. (2017). Which Models Can I Use to Predict Adult ICU Length of Stay? A Systematic Review*. Critical Care Medicine, 45, e222–e231.

Lundberg, S.M., & Lee, S. (2017). A Unified Approach to Interpreting Model Predictions. Neural Information Processing Systems.

Metnitz, P., Moreno, R.P., Almeida, E., Jordan, B., Bauer, P., Campos, R.A., Iapichino, G., Edbrooke, D.L., Capuzzo, M., & le Gall, J. (2005). SAPS 3—From evaluation of the patient to evaluation of the intensive care unit. Part 1: Objectives, methods and cohort description. Intensive Care Medicine, 31, 1336 - 1344.

Moreno, R., Metnitz, P., Almeida, E., Jordan, B., Bauer, P., Campos, R.A., Iapichino, G., Edbrooke, D.L., Capuzzo, M., & le Gall, J. (2005). SAPS 3—From evaluation of the patient to evaluation of the intensive care unit. Part 2: Development of a prognostic model for hospital mortality at ICU admission. Intensive Care Medicine, 31, 1345 – 1355.

Kurtz, P., Bastos, L.S., Salluh, J.I., Bozza, F.A., & Soares, M. (2021). SAPS-3 performance for hospital mortality prediction in 30,571 patients with COVID-19 admitted to ICUs in Brazil. Intensive Care Medicine, 47, 1047 – 1049.

Salluh, J.I., & Soares, M. (2014). ICU severity of illness scores: APACHE, SAPS and MPM. Current Opinion in Critical Care, 20, 557–565.