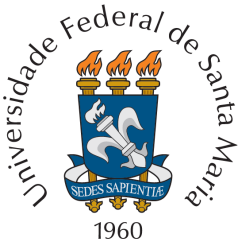


Universidade Federal de Santa Maria
ELC1098 - MINERAÇÃO DE DADOS

TRABALHO FINAL

Diego Ribeiro Chaves
Lucas Xavier Pairé

Orientado por
Prof. Dr. Joaquim V. C. Assunção

02 de dezembro de 2024

1 Objetivo

O presente trabalho tem como objetivo explorar dados acerca de pacientes internados nas unidades de tratamento intensivas de hospitais em busca de padrões entre pacientes que vieram a óbito em uma UTI, agrupando-os em *clusters* de características semelhantes para, posteriormente, analisar as variáveis mais importantes que diferenciam esses grupos.

2 Resumo

Nesse estudo, foi realizada uma análise para identificar fatores relacionados ao óbito de pacientes internados em UTIs, utilizando técnicas de clusterização e árvores de decisão. Os dados clínicos e laboratoriais dos pacientes, incluindo variáveis como idade, tempo de internação e resultados de exames, foram empregados na análise.

Inicialmente, foi utilizado o `KMeans` para agrupar os pacientes em três *clusters*, com base nas variáveis selecionadas. O número de *clusters* foi determinado pelo método do cotovelo, e os *clusters* foram visualizados por meio de PCA em um gráfico 2D, o que possibilitou a identificação de padrões nos dados.

Em seguida, um modelo de árvore de decisão foi treinado para prever os *clusters* dos pacientes a partir das mesmas variáveis. O desempenho do modelo foi avaliado, sendo observada uma acurácia de 98%, com bom desempenho na maioria das classes. Contudo, a classe minoritária apresentou maior dificuldade na previsão.

A análise da árvore de decisão revelou que os resultados dos exames laboratoriais foram identificados como os fatores mais influentes para a classificação dos pacientes, tendo um impacto significativo na distinção entre os grupos de pacientes, sendo determinantes no risco de óbito.

Este trabalho forneceu uma visão aprofundada sobre os fatores que influenciam os desfechos dos pacientes em UTIs, destacando a importância dos exames laboratoriais na previsão do risco de óbito. A aplicação de técnicas de aprendizado de máquina permitiu a construção de modelos preditivos eficientes, com potencial para auxiliar na tomada de decisões clínicas.

3 Metodologia

Para cumprir os objetivos desse trabalho, inicialmente os dados foram tratados, integrados, limpos e selecionados. Em seguida, foram normalizados utilizando o método `StandardScaler` da biblioteca `scikit-learn`. Na fase de mineração, aplicou-se o o algoritmo `KMeans` para identificar agrupamentos entre os pacientes. O número ideal de *clusters* foi determinado pelo método do cotovelo,

resultando em 3 *clusters* principais. Dois desses *clusters* apresentaram características mais compactas, enquanto o terceiro era mais disperso, indicando uma maior variabilidade nas características dos pacientes. Após isso, para entender as características distintivas de cada grupo, treinou-se uma árvore de decisão com os dados de *cluster* como variável-alvo, utilizando as demais variáveis como preditoras. O modelo foi avaliado em termos de acurácia e métricas de classificação. Os passos serão mais detalhados a seguir.

3.1 Coleta e pré-processamento

Os dados acerca dos pacientes encontravam-se, inicialmente, divididos em três *datasets*, sendo eles:

1. `B1_Exam.csv`, que continha informações acerca dos exames realizados pelos pacientes, e seus respectivos resultados. Esse *dataset* apresentava várias linhas duplicadas, pois o mesmo paciente realizava o mesmo exame mais de uma vez.

2. `B1_Prescription.csv`, que continha informações acerca dos medicamentos prescritos para os pacientes, bem como de fatos importantes constatados como alergias, intervenções e quantidade de drogas diferentes e drogas controladas.

3. `B1_Admission.csv`, que continha informações como a data de internação, data de nascimento e principalmente o motivo da saída do paciente da UTI, seja por alta, óbito ou outro motivo.

A partir disso, iniciou-se um processo para garantir a padronização e integração dos dados, reduzindo a complexidade e facilitando as análises subsequentes. O processo foi composto por várias etapas descritas abaixo:

3.1.1 Agregação dos exames

Os valores dos exames foram agrupados por paciente, hospital e tipo de exame, calculando a média de cada exame para cada paciente. O resultado foi salvo em um novo arquivo chamado `Agg_Exams.csv`.

3.1.2 Agregação de Prescrições

Os dados de prescrições foram reduzidos a variáveis essenciais, removendo colunas irrelevantes. Valores booleanos foram agregados com a função "any"para indicar se ao menos uma prescrição continha determinado atributo. Variáveis numéricas foram agregadas utilizando médias, somas ou proporções. Os resultados foram salvos em `Agg_Prescription.csv`.

3.1.3 Processamento das internações

A idade dos pacientes foi calculada com base na data de internação e nascimento, enquanto o tempo de permanência foi computado pela diferença entre as datas de admissão e alta. Colunas irrelevantes foram excluídas e os motivos de alta foram padronizados em categorias mais simples, como "OBITO", "ALTA", "TRANSFERENCIA", e "PERMANENCIA". Os resultados foram salvos em `Remap_Admission.csv`.

3.1.4 Integração dos dados

Os datasets de internação, exames e prescrições foram integrados por paciente e hospital, mantendo apenas registros completos e eliminando duplicatas. Esse dataset combinado foi salvo como `Dataset_Merged.csv`.

3.1.5 Transformação dos Exames

Os valores médios dos exames foram reorganizados em formato de tabela pivô, com cada tipo de exame representando uma coluna, e as médias preenchendo os valores. Essa tabela foi salva como `Dataset_Pivot_Exams.csv`.

3.1.6 Filtragem e limpeza

Pacientes sem informações sobre sexo foram excluídos, e colunas irrelevantes, como identificadores sem significado ou colunas com dados nulos, foram removidas. O dataset resultante foi salvo como `Dataset_Pivot_Sparse.csv`

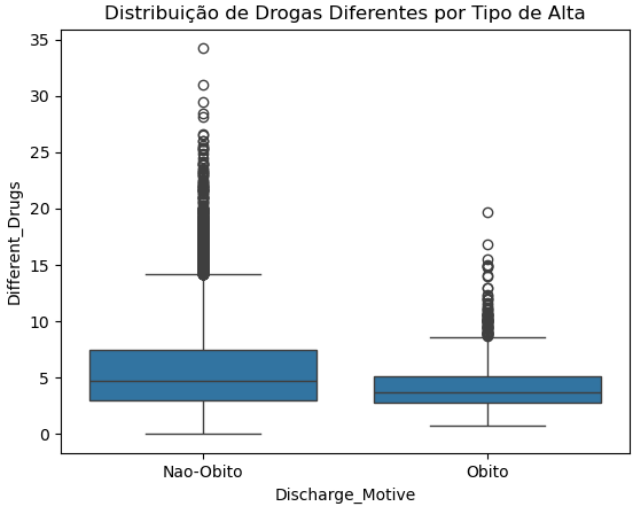
3.1.7 Preparação para a análise binária

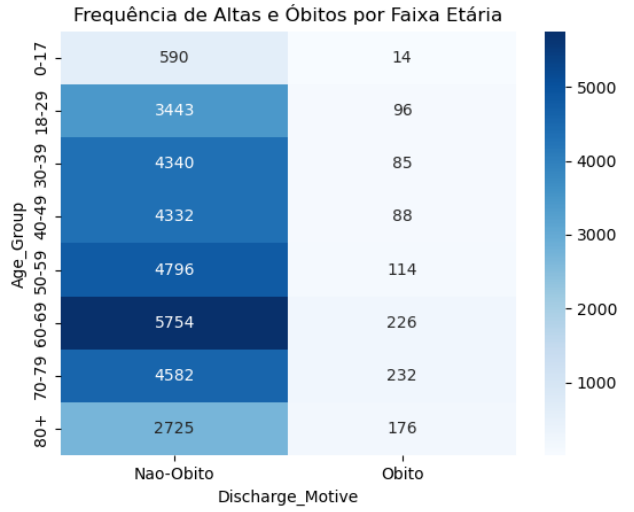
O motivo de alta foi binarizado em duas categorias: "Óbito"e "Não-Óbito". A idade foi categorizada em grupos etários, e valores ausentes em colunas de exames foram preenchidos com a média do respectivo grupo etário e sexo ou, na ausência desta, pela média geral dos demais pacientes que realizaram o mesmo exame. O resultado final foi salvo em `Dataset_Final_Binary.csv`.

3.2 Exploração dos dados

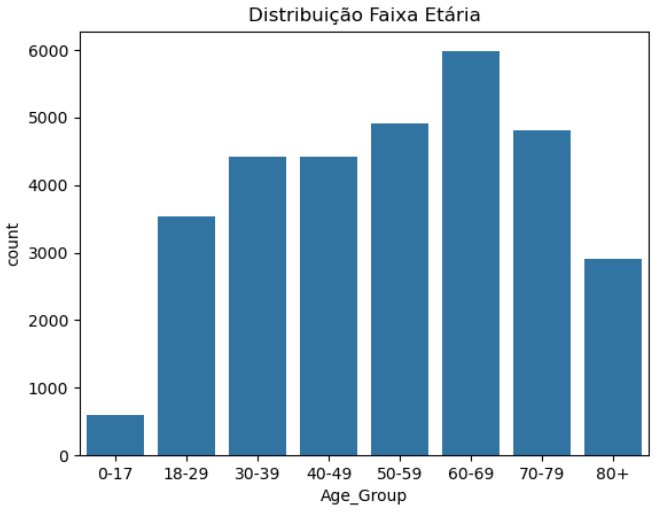
Para a etapa exploratória dos dados, inicialmente buscou-se encontrar relações mais simples entre as variáveis observadas e o motivo de alta binário (Óbito e Não-Óbito). Nesse sentido, obteve-se alguns resultados que podem ser visualizados a seguir.

Buscou-se verificar se existia uma relação entre a quantidade de drogas diferentes ingeridas pelo paciente, bem como da faixa etária com o motivo de alta. Obteve-se o seguinte resultado:

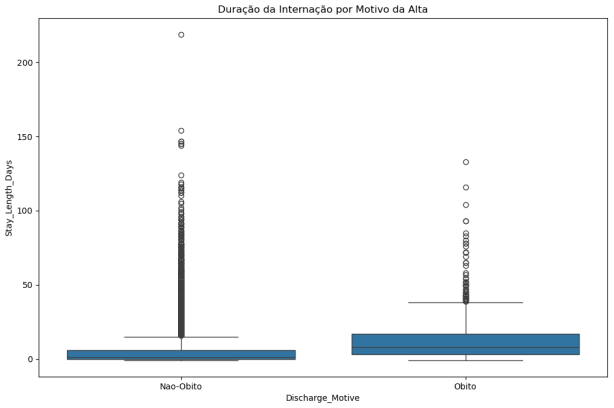




Foi possível constatar que os pacientes que vem a óbito, em geral, ingerem menores quantidades de drogas diferentes. Além disso, pacientes da faixa etária de 70 a 79 anos faleceram com maior frequência do que os demais, ainda que essa faixa não seja a mais frequente, como pode ser constatado abaixo:



Ainda, buscou-se uma possível correlação entre o tempo de internação e o motivo de alta, e constatou-se que em média, pacientes que vem a óbito, ficam internados por mais tempo:



Durante a análise inicial dos dados, não foi possível identificar correlações simples ou relações diretas entre as variáveis e o desfecho clínico (alta ou óbito). Essa ausência de padrões claros pode indicar a influência de múltiplos fatores interagindo de maneira complexa.

Diante disso, decidiu-se adotar abordagens mais profundas para explorar possíveis relações nos dados, utilizando técnicas como *clustering* e árvores de decisão. O objetivo foi identificar grupos de pacientes com características semelhantes e compreender melhor os fatores que podem contribuir para os desfechos clínicos.

As árvores de decisão permitem visualizar as variáveis mais relevantes e como elas se combinam para influenciar os resultados, enquanto o *clustering* agrupa os pacientes com base em suas semelhanças, oferecendo *insights* adicionais sobre padrões não triviais nos dados.

Decidiu-se focar os esforços nos pacientes que vieram a óbito, dada a relevância desses casos para os hospitais.

3.3 Análise de Dados através de *Clustering* e Árvore de Decisão

3.3.1 Pré-processamento dos Dados

Para explorar os fatores que podem estar relacionados ao desfecho de óbito dos pacientes que foram internados em unidades de terapia intensiva (UTI), aplicou-se técnicas de *clustering* e árvores de decisão com o intuito de descobrir padrões ocultos nos dados, como a segmentação de pacientes com características similares e a identificação de variáveis preditivas para o desfecho de óbito.

O *dataset* foi inicialmente carregado e segmentado para incluir apenas os pacientes que vieram a óbito. Além disso, foi selecionado um conjunto de colunas com variáveis numéricas e categóricas que podem ser relevantes para o *clustering*, como variáveis clínicas (e.g., idade, score de prescrição) e resultados de exames laboratoriais. Por último, removeu-se os valores ausentes, dado que a presença dos mesmos poderia prejudicar a análise.

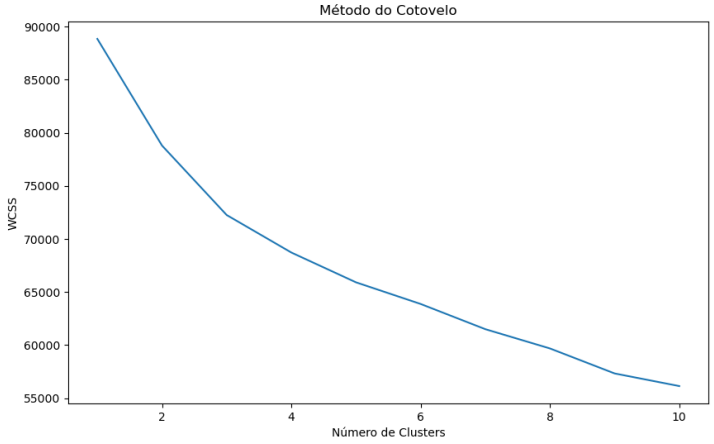
3.3.2 Escalonamento dos Dados

Antes de aplicar qualquer algoritmo de *machine learning*, foi necessário realizar o escalonamento das variáveis. Isso é importante porque muitos algoritmos, como o `KMeans`, são sensíveis à escala das variáveis. Para isso, utilizou-se o `StandardScaler` para normalizar os dados, transformando as variáveis para terem média zero e desvio padrão igual a um. Esse processo garante que todas as variáveis contribuam igualmente para os cálculos de distância.

3.3.3 *Clustering* com `KMeans`

O primeiro passo para entender os padrões dos dados foi aplicar o algoritmo `KMeans` de *clustering*. O objetivo do *clustering* é segmentar os pacientes em grupos homogêneos com base nas suas características, sem ter conhecimento prévio sobre quais grupos existem (sem supervisão).

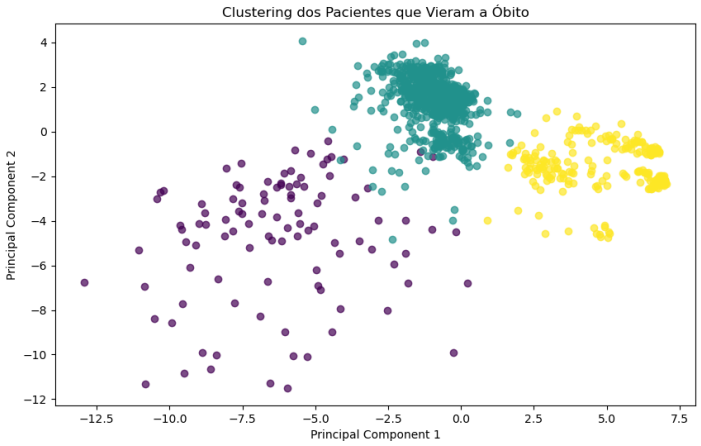
A escolha do número de clusters no `KMeans` foi feita utilizando o método do cotovelo, amplamente utilizado para determinar o número ideal de *clusters*, conforme descrito por Jain et al. [1999]. Esse método calcula a soma dos quadrados dentro do *cluster* (*WCSS - Within-Cluster Sum of Squares*) para diferentes números de *cluster* e escolhe o número de *cluster* onde a redução do WCSS começa a ser menos acentuada. O gráfico resultante ajuda a visualizar o ponto em que a adição de mais *clusters* não melhora significativamente a qualidade do agrupamento, indicando o número ideal de *clusters*. Nesse caso, o número ideal de clusters foi escolhido como 3, que foi visualmente indicado pelo ponto de inflexão no gráfico.



Após determinar o número de clusters, o `KMeans` foi aplicado aos dados escalonados para dividir os pacientes em 3 grupos. Essa operação resultou em uma coluna adicional no dataset, representando o *cluster* ao qual cada paciente pertence. Os *clusters* podem ser analisados para identificar padrões de características comuns.

Como o `KMeans` trabalha em um espaço multidimensional, a visualização direta dos *clusters* seria difícil. Para facilitar a análise, utilizou-se a técnica de Análise de Componentes Principais (PCA), que reduz a dimensionalidade dos dados para duas variáveis principais, permitindo a visualização dos clusters em um gráfico 2D.

A plotagem dos *clusters* foi realizada da seguinte forma, permitindo a análise visual da separação entre eles:



3.3.4 Árvore de Decisão

A segunda técnica aplicada foi a árvore de decisão, um modelo supervisionado que permite identificar quais variáveis são mais relevantes para predizer a classe de um paciente (no caso, o *cluster*).

Primeiramente, os dados foram divididos em dois conjuntos: treinamento e teste, utilizando a proporção 80/20. Após a divisão, conjunto de treinamento resultante foi utilizado para treinar a árvore de decisão, enquanto o conjunto de teste foi utilizado para avaliar o desempenho do modelo. A medição do desempenho do modelo foi realizado através da acurácia e das métricas de classificação (precisão, *recall*, f1-score):

Classe	Precisão	*Recall*	*F1-Score*	Suporte
0	0.91	0.87	0.89	23
1	0.98	0.99	0.99	133
2	1.00	0.98	0.99	51

Tabela 1: Desempenho da árvore de decisão

Interpretando as métricas, a acurácia do modelo é 0.9758, o que significa que o modelo classificou corretamente cerca de 98% das amostras no conjunto de teste. Esse é um excelente desempenho global, considerando que o modelo foi capaz de prever corretamente a classe de quase todos os pacientes.

A precisão mede a capacidade do modelo de classificar corretamente os casos positivos em cada classe. Em termos de precisão por classe:

1. A classe 2 (com 51 pacientes) teve precisão de 1.00, o que significa que todas as previsões feitas para esta classe foram corretas.

2. Já a classe 1 (com 133 pacientes) também teve uma precisão alta de 0.98, indicando que as previsões para essa classe foram em grande parte corretas.

3. Entretanto, a classe 0 (com 23 pacientes) teve uma precisão de 0.91, o que ainda é um bom valor, mas indica que houve uma pequena quantidade de previsões incorretas para esta classe.

O *recall*, por outro lado, mede a capacidade do modelo de identificar corretamente todos os casos positivos de cada classe:

1. A classe 1 teve o melhor *recall* (0.99), ou seja, o modelo conseguiu identificar quase todos os pacientes dessa classe de forma correta.

2. A classe 2 teve um *recall* de 0.98, o que também indica um bom desempenho em identificar pacientes dessa classe.

3. A classe 0 teve um *recall* um pouco mais baixo (0.87), sugerindo que o modelo teve dificuldades em identificar todos os pacientes dessa classe corretamente.

O *F1-Score* é uma média harmônica entre precisão e *recall*, equilibrando essas duas métricas. Um *F1-Score* alto indica que o modelo está bem equilibrado entre as duas métricas, nesse sentido a classe 2 teve o melhor *F1-Score* (0.99), seguida pela classe 1 com 0.99 também. A classe 0 teve um *F1-Score* de 0.89, que, embora menor, ainda indica um bom desempenho.

A média macro (*Macro Average*) considera a média aritmética das métricas de cada classe, tratando todas as classes igualmente. O *macro avg* para precisão, *recall* e *F1-Score* são 0.96, 0.95, e 0.95, respectivamente. Isso indica que, em média, o modelo teve um bom desempenho geral, mas o modelo tem uma performance levemente inferior em identificar a classe 0.

A média ponderada (*Weighted Average*) leva em consideração o número de amostras de cada classe, ajustando as métricas conforme o número de exemplos em cada classe. A *weighted avg* para precisão, *recall* e *F1-Score* são 0.98, 0.98, e 0.98, respectivamente, o que reflete a boa performance global do modelo, considerando que as classes 1 e 2 dominam em termos de número de amostras.

Concluindo a análise do desempenho, O modelo teve um desempenho excelente na classificação dos pacientes, com uma acurácia de 98% e uma boa performance em termos de precisão, *recall* e *F1-Score*, principalmente nas classes 1 e 2.

No entanto, a classe 0 (representando um subconjunto menor de pacientes) apresentou um desempenho um pouco inferior, com uma precisão de 0.91 e um *recall* de 0.87. Isso sugere que, embora o modelo esteja funcionando muito bem

em identificar pacientes das classes majoritárias (1 e 2), ele tem dificuldades em classificar corretamente alguns pacientes da classe minoritária (0), o que é comum em problemas com classes desbalanceadas. A precisão para essa classe ainda é boa, mas a diferença entre precisão e *recall* indica que o modelo pode estar falhando em identificar todos os pacientes dessa classe de forma eficaz.

Em resumo, o modelo de árvore de decisão teve uma performance robusta e pode ser aprimorado com técnicas adicionais de balanceamento de classes (como *SMOTE* ou *undersampling*) para melhorar a capacidade de classificação da classe minoritária, ou então explorar o uso de algoritmos mais avançados como *Random Forests* ou *XGBoost* para melhorar a performance geral.

A árvore de decisão então foi visualizada para interpretar como as variáveis influenciam a classificação dos pacientes em diferentes *clusters*. Devido ao alto grau de detalhes da árvore, optou-se por colocá-la em modo paisagem:

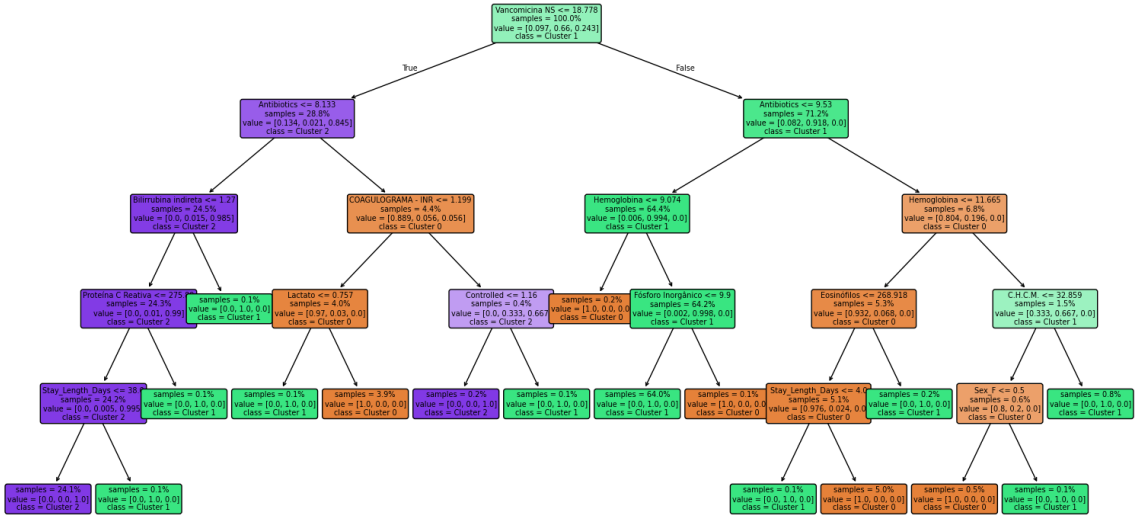


Figura 2: Visualização da árvore de decisão

A árvore de decisão revelou quais variáveis foram mais importantes para classificar os pacientes e como essas variáveis estavam sendo usadas para tomar decisões ao longo da árvore. Em geral, os resultados dos exames foram os principais fatores classificadores.

4 Conclusão

Nesse trabalho, buscou-se explorar e analisar dados de pacientes internados em UTIs, com o objetivo de entender os fatores associados ao desfecho de óbito. Para isso, utilizou-se técnicas de clusterização e árvores de decisão, abordagens poderosas para identificar padrões ocultos nos dados e prever categorias baseadas em variáveis clínicas e demográficas.

Primeiramente, aplicou-se o `KMeans` para realizar a clusterização dos pacientes. Utilizou-se a técnica do método do cotovelo para determinar o número ideal de *clusters*, o que levou a escolha de 3 *clusters*. Esses *clusters* foram visualizados em um gráfico 2D usando PCA, permitindo uma compreensão mais clara de como os pacientes foram agrupados com base nas características fornecidas. Embora não tenha sido possível encontrar relações simples de correlação diretamente visíveis, o uso do *clustering* ajudou a revelar padrões mais sutis nos dados.

Em seguida, para entender melhor a relação entre as variáveis e os *clusters*, aplicou-se uma árvore de decisão. O modelo foi treinado para prever os *clusters* a partir dos dados clínicos e exames dos pacientes. O desempenho do modelo foi feito usando métricas de classificação como precisão, *recall*, *F1-Score* e acurácia. O modelo obteve uma acurácia de 98%, com bom desempenho na maioria das classes, especialmente nas classes 1 e 2, mas com alguns desafios na classificação da classe minoritária (classe 0). Isso sugere que o modelo pode ser mais sensível a classes desbalanceadas, um problema comum em tarefas de classificação com dados assimétricos.

Por fim, as técnicas utilizadas, combinadas com a análise visual e estatística dos dados, proporcionaram uma compreensão valiosa sobre os fatores que influenciam os desfechos dos pacientes, especificamente os que foram a óbito. O bom desempenho do modelo de árvore de decisão, apesar de algumas dificuldades com a classe minoritária, confirma que os dados contêm informações relevantes para prever o risco de óbito.

Para trabalhos futuros, seria interessante explorar técnicas de balanceamento de classes para melhorar a performance do modelo, além de investigar outras abordagens de aprendizado de máquina mais avançadas, como *Random Forests* e *XGBoost*, que podem oferecer melhores resultados em termos de precisão e capacidade de generalização.

Esse estudo contribui para o entendimento dos fatores de risco associados ao

óbito em pacientes críticos, oferecendo uma base para o desenvolvimento de ferramentas preditivas em ambientes hospitalares, o que pode melhorar o cuidado e a tomada de decisões clínicas.

Os achados indicam que pacientes com resultados específicos em exames laboratoriais, idade avançada e tempo de internação prolongado têm maior probabilidade de vir a óbito. A aplicação do `KMeans` e das árvores de decisão possibilitou identificar padrões úteis que podem subsidiar melhorias nos cuidados clínicos.

> "Se vi mais longe foi por estar sobre os ombros de gigantes."– Isaac Newton

5 Referências

Referências

Aurélien Géron. *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*. O'Reilly Media, 2nd edition, 2019. ISBN 978-1492032649.

J. D. Hunter and the Matplotlib development team. Matplotlib: A 2d graphics environment, 2007. URL `https://matplotlib.org`. Acesso: 2024-11-30.

Anil K Jain, M Narasimha Murty, and Patrick J Flynn. Data clustering: A review. *ACM computing surveys (CSUR)*, 31(3):264–323, 1999.

Wes McKinney. *Python for Data Analysis: Data Wrangling with Pandas, NumPy, and Jupyter*. O'Reilly Media, 2nd edition, 2017. ISBN 978-1491957660.

The pandas development team. pandas: Powerful python data analysis toolkit, 2024. URL `https://pandas.pydata.org`. Acesso: 2024-11-30.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and É. Duchesnay. Scikit-learn: Machine learning in python, 2011. URL `https://scikit-learn.org`. Acesso: 2024-11-30.

Michael Waskom and the seaborn development team. seaborn: Statistical data visualization, 2024. URL `https://seaborn.pydata.org`. Acesso: 2024-11-30.

McKinney [2017] Géron [2019] Waskom and the seaborn development team [2024] Pedregosa et al. [2011] Hunter and the Matplotlib development team [2007] pandas development team [2024]