---------mariana--------------catarina------------------liliana----

SLIDE 1

Introdução:

Neste trabalho foi nos proposto analisar um conjunto de dados sobre pacientes com HCC. A análise visa identificar as variáveis que influenciam mais a sobrevivência dos pacientes e qual modelo que oferece os melhores resultados.

SLIDE 2:

Iniciamos o nosso trabalho com uma análise exploratória dos dados.

Utilizamos métodos como shape, head, tail, info, duplicated, nunique e describe, e assim identificamos que o conjunto de dados tem 165 linhas e 50 colunas, não existem linhas repetidas, exitem colunas numéricas e colunas categóricas e ainda percebemos que existiam valores ausentes.

Seguidamente, a partir de histogramas das colunas numéricas observamos que a maioria dos pacientes são idosos, com idade média acima de 60 anos. Além disso, há uma prevalência maior de pacientes do sexo masculino. Muitos pacientes apresentam cirrose e sintomas associados, como varizes e esplenomegalia,. Também identificamos um consumo significativo de álcool e tabaco entre os pacientes.

Ao analisar as taxas de sobrevivência, verificamos que aproximadamente 62% dos indivíduos sobrevivem.

A partir de boxplots verificamos a existência de outliers que são valores que se afastam significativamente dos outros dados e podem influenciar negativamente o desempenho dos modelos. Explicando como se pode analisar este tipo de gráficos, a linha horizontal dentro da caixa representa a mediana e divide o primeiro quartil e o terceiro quartil. As linhas horizontais acima e abaixo representam o alcance iterquartílico, ou seja, o IQR, que inclui 50% dos dados. Os pontos fora do IQR são considerados outliers.

SLIDE 3

Preprocessamento de Dados:

Para garantir a qualidade dos dados, realizamos um preprocessamento abrangente. O preprocessamento inclui várias etapas:

Para o tratamento de valores Ausentes, utilizamos técnicas de imputação, como a média para substituir os valores em falta das colunas numéricas e com a moda para as colunas categóricas. Convertemos as colunas categóricas usando LabelEncoder.

Removemos ainda os outliers.Como se pode ver, nestes boxplots que já não estam presentes.

Analisando agora este gráfico, que é denominado de gráfico de radar: podemos ver que ele só inclui variáveis numéricas , comparando duas classes, representadas como 0 , ou seja, vive, e 1 no que caso de morrer.

Em relação a ALP é de destacar que a Classe 1 tem níveis médios, mais altos em comparação com a Classe 0 . Já para a variável AFP, a Classe 1 mostra níveis elevados em relação à Classe 0, o que pode ser um marcador significativo. Para INR, a Classe 1 tende a ter valores ligeiramente mais altos. Hemoglobina: Classe 0 tem níveis ligeiramente mais altos em comparação com a Classe 1.

SLIDE 4:

A percentagem de pacientes que sobrevive e que morre é respetivamente 38.2%, o que corresponde a 63 indivíduos, e 61,8% morre, que corresponde a 102 indivíduos.

Dos pacientes que sobrevivem 19.5% são homens e 17.2% são mulheres, relativamente aos que morrem, 30,5% são homens e 32,8% sao mulheres.

O eixo x representa os intervalos de idade.O eixo y mostra a contagem de pessoas em cada intervalo.A maior barra representa o intervalo de 50 a 75 anos, indicando que a maioria das pessoas do conjunto de dados está nessa faixa etária.

As áreas onde as duas curvas se sobrepõem indicam idades onde tanto sobreviventes quanto falecidos estão presentes. As áreas onde uma curva está mais alta que a outra, indicam idades onde há uma maior concentração de indivíduos daquela classe específica.

Analisando este último gráfico:

Para mulheres, a sobrevivência está mais distribuída entre diferentes faixas etárias, enquanto os falecimentos são mais concentrados em idades avançadas (70-80 anos).

Para homens, existe uma maior concentração de sobreviventes e falecidos nas faixas etárias de 60-70 e 70-80 anos, com um número relativamente menor de falecidos em idades mais jovens.

SLIDE 5:

A seguir fizemos uma matriz correlação para entender como as variáveis se relacionam entre si.

Quanto mais vermelho, mais forte é a correlcao positiva, mais proxima de 1. Quanto mais azul menos forte é a correlacao negativa, mais proxima de −1.

Assim , concluimos que variáveis como AST, ALT, GGT, Bilirrubina Total e Albumina mostraram correlações significativas entre si.

Observamos também uma correlação moderada entre a idade dos pacientes e a classe de sobrevivência, sugerindo que a idade pode influenciar a probabilidade de sobrevivência. Identificamos que a Bilirrubina Direta e a Bilirrubina Total fornecem informações redundantes devido à alta correlação entre elas, indicando que uma das duas pode ser removida para simplificar os modelos.

Além disso, destacamos variáveis como ALP, AFP, Hemoglobina, PS , Metástase e Albumina como importantes para a análise.

SLIDE 6:

Seguidamente, este dois gráficos representam a matriz correlacao das variáveis que consideramos mais relevantes e o seu heatmap, assim conseguimos concluir que Dir_Bil e Total_Bil fornecem informações redundantes devido à sua alta correlação e que ALP, AFP, Hemoglobina, PS, Metástase e Albumina são cruciais para a análise preditiva da sobrevivência do paciente.

SLIDE 7:

Utilizamos vários algoritmos para construir nossos modelos de previsão. Vou passar a detalhar cada um:

- Decision Tree: Cria uma árvore de decisão baseada nas variáveis dos dados. Fácil de interpretar, mas pode sofrer de overfitting, especialmente com dados pequenos.
- K-Nearest Neighbors (KNN): Classifica pacientes com base nos k vizinhos mais próximos.
- Random Forest: Constrói múltiplas árvores de decisão e combina seus resultados para melhorar a precisão e reduzir o overfitting. Robusto e eficaz em classificação.
- Gradient Boosting: Melhora a previsão corrigindo erros sequencialmente. Poderoso, mas sensível a outliers e requer ajuste cuidadoso dos hiperparâmetros.
- Multi-Layer Perceptron (MLP): Rede neural que aprende representações complexas. Captura relações não lineares, mas exige mais dados e tempo de treino.
- Regressão Logística: Modela a probabilidade de sobrevivência com base numaa combinação linear das variáveis. Fácil de interpretar e eficiente para problemas binários.
- Stacking Classifier: Combina previsões de vários modelos base para melhorar a precisão. Utiliza um meta-modelo para aprender a melhor combinação dos modelos base.
- Support Vector Classifier (SVC): Encontra o hiperplano que separa as classes. Eficaz em espaços de alta dimensão e ajustável para diferentes margens de erro.

Análise do gráfico dos clusters

Para gerar este gráfico analisamos a matriz de correlacao e verificamos as variaveis que tinham uma maior correlação entre si e com a variável Class relativa e o resultado foram as variáveis dir bill e total bill. Com a análise deste gráfico observamos precisamente essa correlação e a classes 0 e 1 correspondem a não sobreviventes e sobreviventes respetivamente.

SLIDE 8 :

Para avaliar os modelos, utilizamos a validação cruzada e a análise da curva ROC, que nos permite avaliar a sensibilidade e especificidade dos modelos.

- Random Forest: Melhor desempenho com precisão de 0.74, destacando-se pela robustez e eficácia na classificação dos pacientes.

- Gradient Boosting: Precisão de 0.70, sendo a segunda melhor opção, embora menos robusto que o Random Forest.

- KNN: Promissor, mas apresentou alguns falsos negativos.

- MLP: Necessita de ajustes devido ao maior número de falsos negativos. Requer mais dados e ajustes .

- Decision Tree: Simples de interpretar, com desempenho moderado. Útil para compreensão inicial.

- Regressão Logística e SVC: Priorizaram a precisão, sendo rápidos.

SLIDE 9:

Neste slide, apresentamos as matrizes de confusão para todos os modelos, resumindo o desempenho através da comparação das previsões com os valores reais. As matrizes incluem quatro componentes principais: verdadeiros positivos, falsos positivos, falsos negativos e verdadeiros negativos.

Com base nesta análise, calculamos métricas de avaliação como acurácia, precisão, recall e F1-Score. Utilizamos validação cruzada para estimar a capacidade de generalização dos modelos.

Conclusões:

- Random Forest: Melhor desempenho para prever a sobrevivência dos pacientes.
- Gradient Boosting: Segunda melhor opção.

- SVC e MLP: Necessitam de melhorias.

O dendrograma hierárquico mostra a proximidade e estrutura dos clusters de modelos, indicando que KNN e Random Forest compartilham características e padrões de previsão comuns.

SLIDE 10 :

Com base na nossa análise, o Random Forest foi identificado como o modelo mais robusto e preciso para a classificação de sobrevivência dos pacientes com Carcinoma Hepatocelular. Gradient Boosting e KNN também mostraram bons desempenhos, enquanto modelos como MLP e Decision Tree precisam de ajustes adicionais.