Data exploration and enrichment for supervised classification

Elementos de Inteligência Artificial e Ciência de Dados

22 DE MAIO, 2024

Vinicius Abrunhosa
Marco Dinis Sousa
Bárbara Neto

ESPECIFICAÇÃO DO TRABALHO A REALIZAR

1) Análise e exploração de dados
2) Tratamento de dados
3) Criação modelos supervised learning com sklearn
4) Experimentar diferentes modelos e seleção de maior eficiência
5) Analisar os resultados
6) Incorporar uma interface streamlit

O problema de "machine learning" é classificar se um paciente diagnosticado com carcinoma hepatocelular vai ou não sobreviver passado 1 ano do diagnóstico, baseado em dados clínicos anteriores.

ANÁLISE DE DADOS

- Ver qual o tipo de dados em cada coluna
- Verificar features com dados em falta
- Analisar correlações entre as features e classificação final

TRATAMENTO DE DADOS

Tendo como base a análise:

- Descartar colunas com pouca relação com a classificação final
- Completar espaços em falta
- Discretizar dados contínuos
- Codificação de dados discretos

MODELAÇÃO

- Decision Tree
- Random Forest
- Logistic Regression
- SVM
- KNN

- número de splits: 5
- número de repeats: 10
- random state: 0

Treino: 20% dos dados

OBJETIVO:

Calcular precisão, accuracy, recall, f1-score, support e matriz de dispersão

Evitar ao máximo falsos negativos - devido à natureza dos dados tratados e o objetivo médico do trabalho

INTERFACE

Usámos o StreamLit para fazer a nossa interface.

1- Escolher o modelo

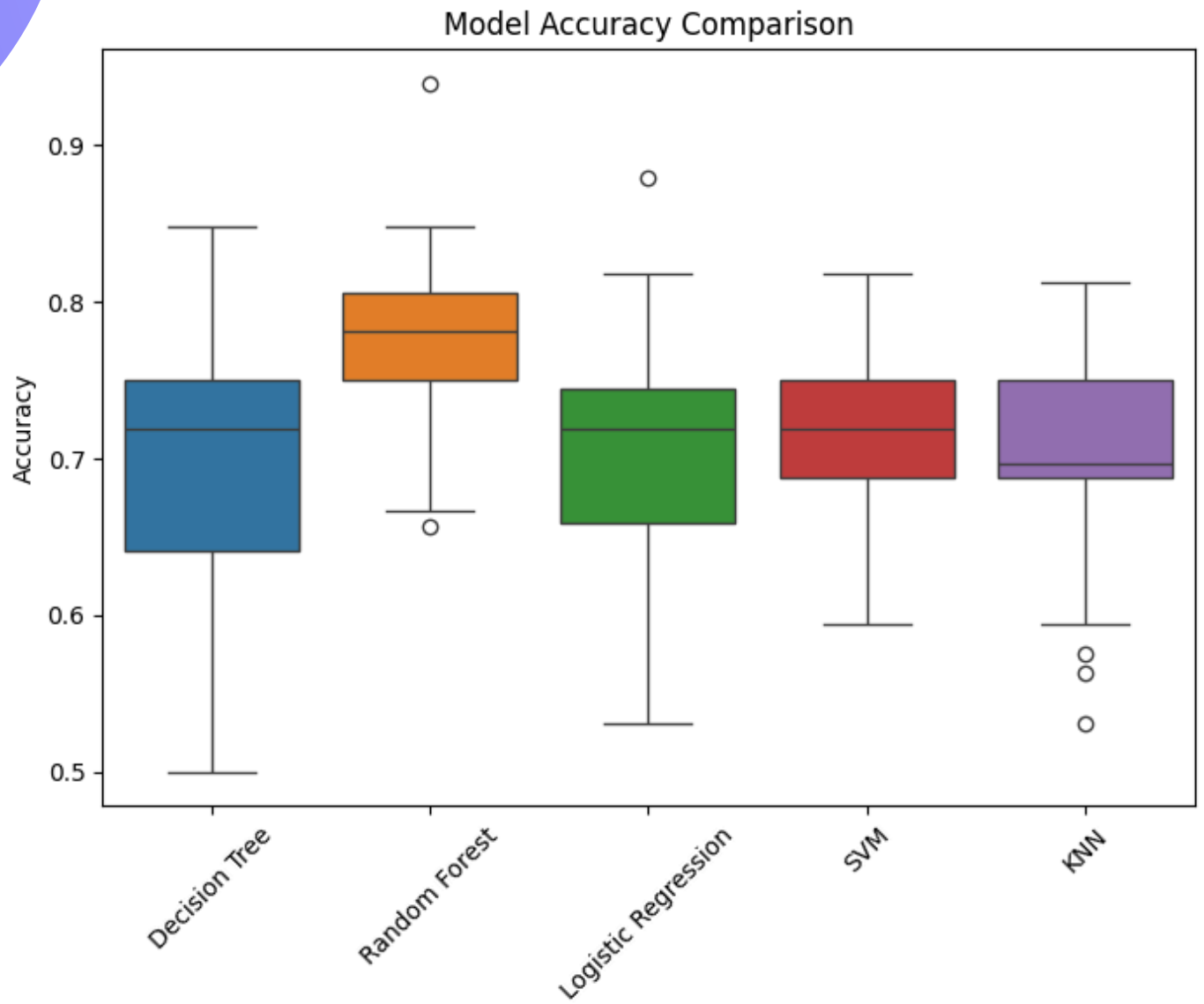
2- Introduzir os dados

3- Submeter!

A aplicação retornará se o paciente sobrevive ou não passado 1 ano desde o diagnóstico.

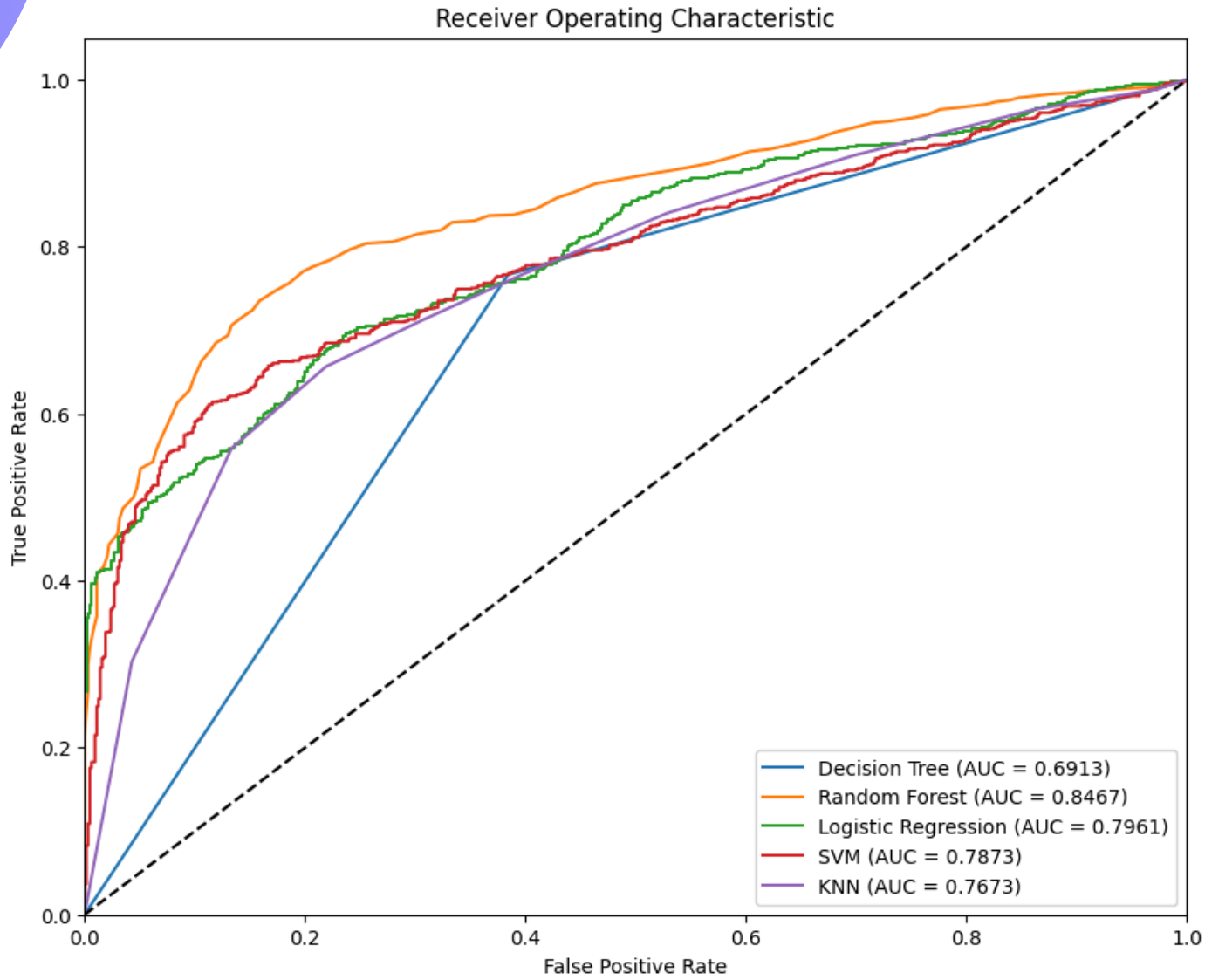
Na última linha da interface aparece o resultado - “Lives” ou “Dies”.

RESULTADOS



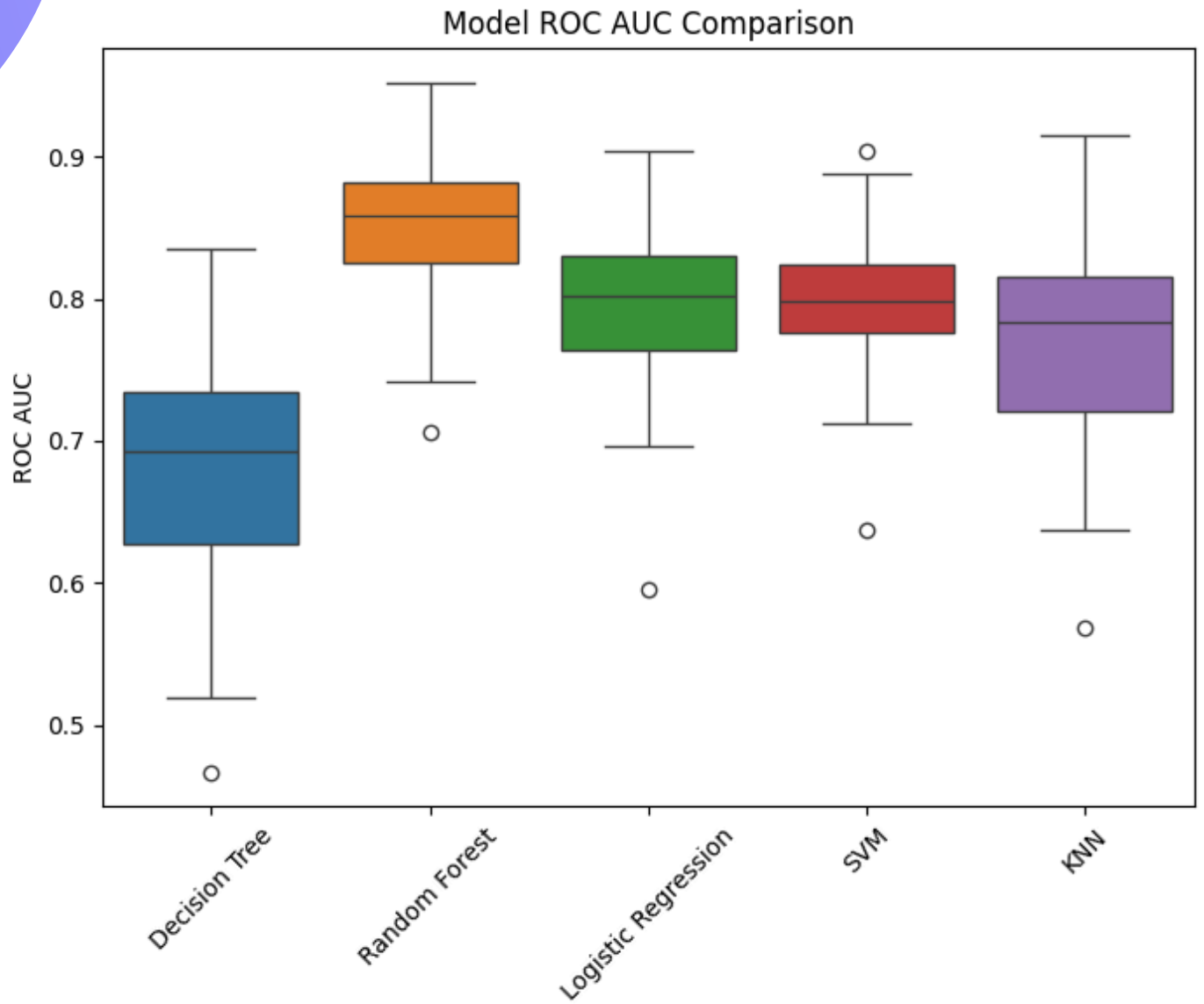
Representação os quartis e outliers da accuracy de cada modelo depois de vários testes.

RESULTADOS



Representação das curvas ROC para cada um dos modelos.

RESULTADOS



Representação os quartis e outliers da área por baixo da curva ROC de cada modelo depois de vários testes.

CONCLUSÕES

O modelo **random forest** mostrou ter os melhores resultados.

Na precisão e acurácia, para além de ter os valores mais elevados é o que tem a menor amplitude interquartil.

Na quantidade de falsos negativos é o que apresenta os menores quando comparado aos verdadeiros positivos.

É também o modelo com maior ROC AUC.

Tendo em conta os parâmetros avaliados o segundo melhor modelo é o SVM.

BIBLIOGRAFIA

- Chat-GPT
- Bing chat
- Documentação da biblioteca “scikit-learn”
- Notebooks fornecidos pelos professores

BIBLIOTECAS USADAS:

- scikit-learn
- pandas
- numpy
- matplotlib
- seaborn
- graphviz
- streamlit