

# CIÊNCIA DE DADOS PROJETO

GRUPO 4  
[Link GitHub](#)

# Exploração de dados

Foram realizadas várias análises para compreender a estrutura e o conteúdo do conjunto de dados e prepará-los para etapas seguintes.

Análise preliminar do conjunto de dados

- **Contagem de Dados e Características:** Verificou-se o número total de colunas (características) e linhas (dados) no conjunto de dados.
- **Listagem das Características:** Listaram-se todas as colunas presentes no conjunto de dados.

Identificação dos Tipos de Dados

- Identificaram-se os tipos de dados (por exemplo, numéricos, categóricos) de cada coluna utilizando 'data.dtypes'.

Análise de Valores Ausentes

- **Contagem de Valores Ausentes:** Contou-se o número de '?' em cada coluna.
- **Substituição de Valores Ausentes por NaN:** Garantiu-se que os valores ausentes fossem representados como NaN no conjunto de dados.

# Exploração de dados

Foram realizadas várias análises para compreender a estrutura e o conteúdo do conjunto de dados e prepará-los para etapas seguintes.

## Análise Estatística e Criação de Tabela Resumo

- **Realizou-se uma análise estatística para cada característica do conjunto de dados. Isso incluiu:** o cálculo de medidas como média, desvio padrão, máximo, mínimo, contagem de valores ausentes e a moda (valor mais frequente).
- **Criou-se uma tabela que contém os seguintes elementos para cada característica:** o nome da característica, o tipo (numérico ou categórico), a média ou moda (dependendo do tipo de dados), o desvio padrão (expresso como uma percentagem da média), os valores máximo e mínimo para as características numéricas, e a contagem e percentagem de valores ausentes.

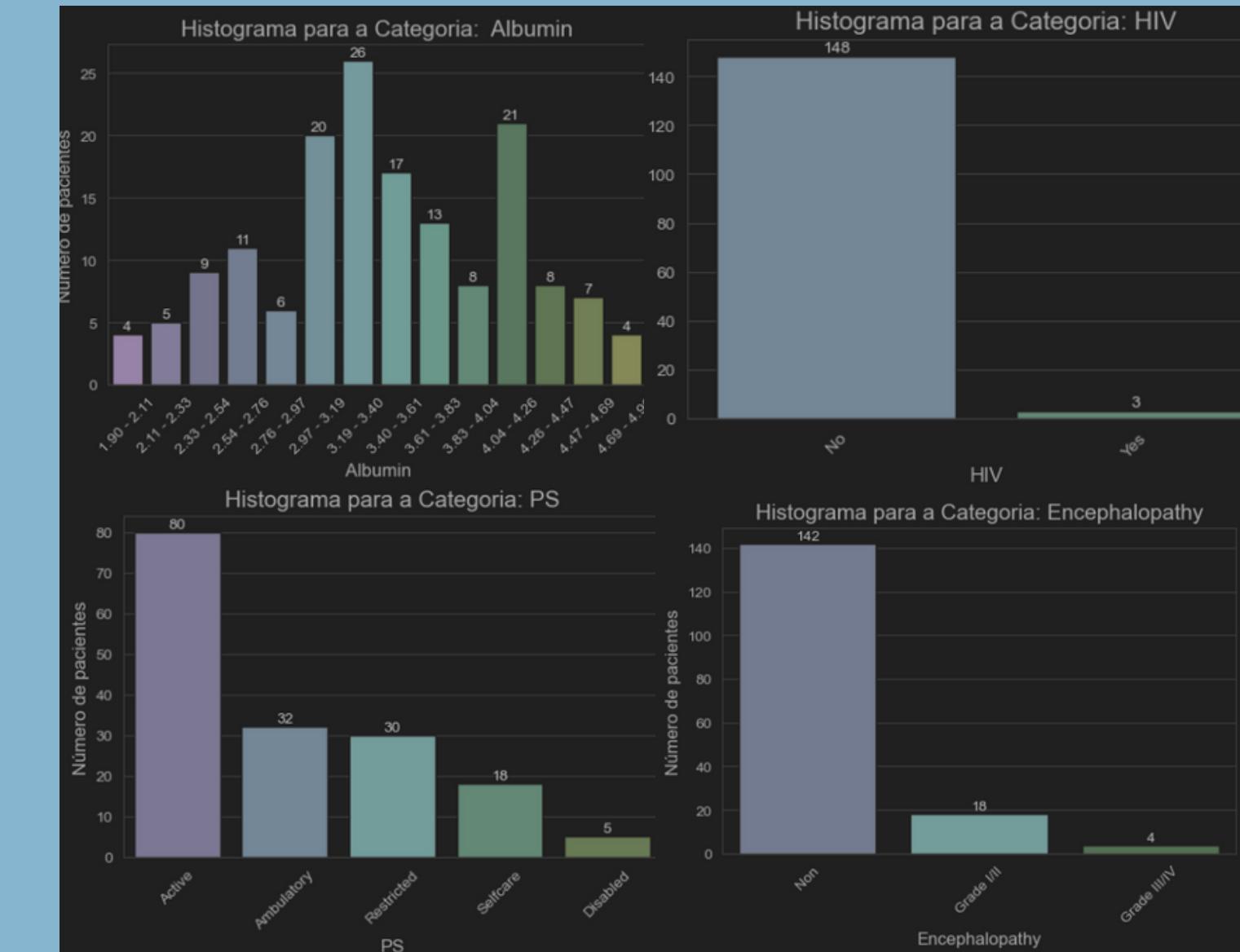
## Visualização com Histogramas

- **Criaram-se histogramas para todas as categorias** com o objetivo de compreender melhor a distribuição dos dados .

# Exploração de dados

	Tipo	Média/Moda	Desvio_Padrão (%)	Máximo	Mínimo	Nan	Nan (%)
<b>Gender</b>	object	Male	-	-	-	0	0.00%
<b>Symptoms</b>	object	Yes	-	-	-	18	10.91%
<b>Alcohol</b>	object	Yes	-	-	-	0	0.00%
<b>HBsAg</b>	object	No	-	-	-	17	10.30%
<b>HBeAg</b>	object	No	-	-	-	39	23.64%
<b>HBcAb</b>	object	No	-	-	-	24	14.55%
<b>HCVAb</b>	object	No	-	-	-	9	5.45%
<b>Cirrhosis</b>	object	Yes	-	-	-	0	0.00%
<b>Endemic</b>	object	No	-	-	-	39	23.64%
<b>INR</b>	float64	1.42	33.61%	4.82	0.84	4	2.42%
<b>AFP</b>	float64	19299.95	772.53%	1810346.00	1.20	8	4.85%
<b>Hemoglobin</b>	float64	12.88	16.66%	18.70	5.00	3	1.82%
<b>MCV</b>	float64	95.12	8.84%	119.60	69.50	3	1.82%
<b>Leucocytes</b>	float64	1473.96	197.37%	13000.00	2.20	3	1.82%
<b>Platelets</b>	float64	113206.44	94.62%	459000.00	1.71	3	1.82%
<b>Albumin</b>	float64	3.45	19.88%	4.90	1.90	6	3.64%
<b>Total_Bil</b>	float64	3.09	178.09%	40.50	0.30	5	3.03%
<b>ALT</b>	float64	67.09	85.76%	420.00	11.00	4	2.42%
<b>AST</b>	float64	96.38	90.77%	553.00	17.00	3	1.82%

Tabela Resumo



Histogramas

# Pré- processamento

## Transformação de dados

**Conversão de valores categóricos** (ex: Yes, No) **em numéricos** (ex: 1,0).

## Remoção de variáveis

**Remoção de variáveis com base na análise de uma matriz de correlação.** **Uma menor correlação de uma variável com a variável alvo: 'Class', leva à remoção desta.**

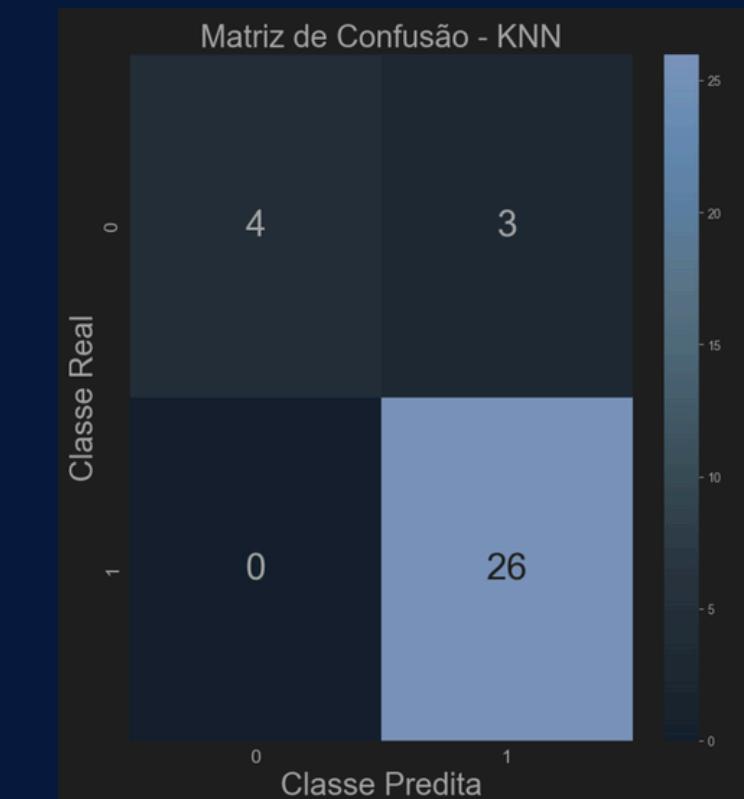
## Imputação de valores ausentes

### Substituição dos valores em falta pela:

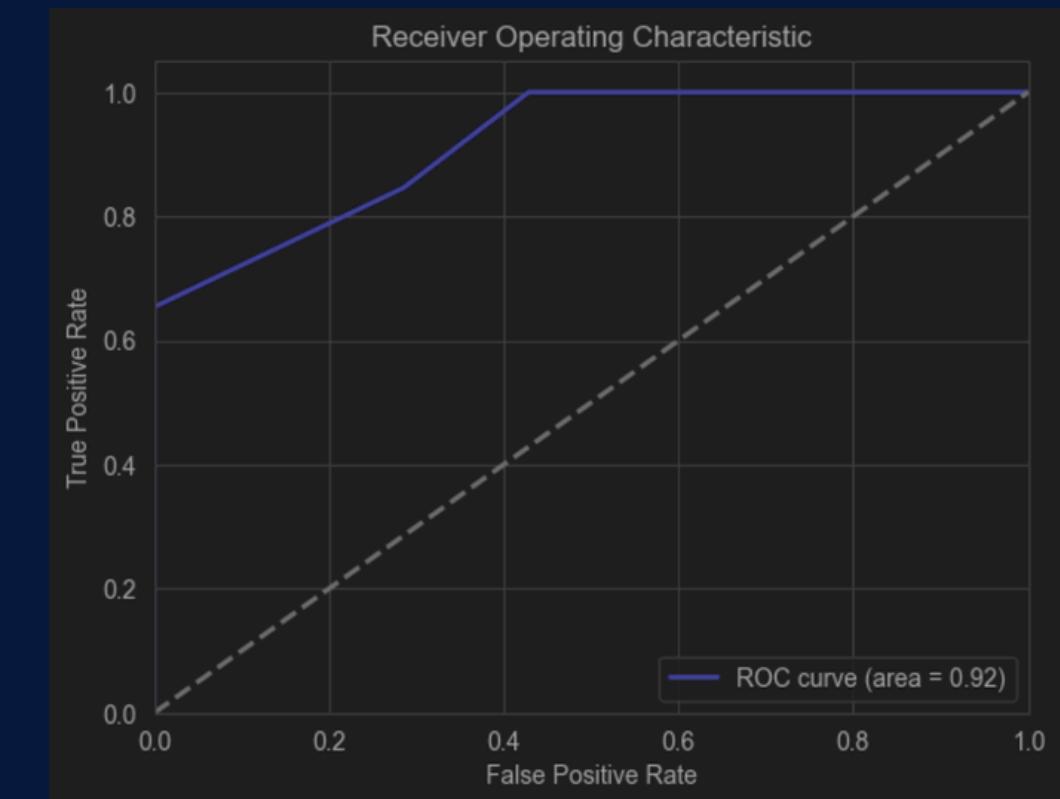
- Média da coluna (em colunas numéricas)
- Moda da coluna (em colunas categóricas)

# Modelação e Avaliação de dados - KNN

Matrizes de confusão



Curvas ROC



Relatórios de classificação

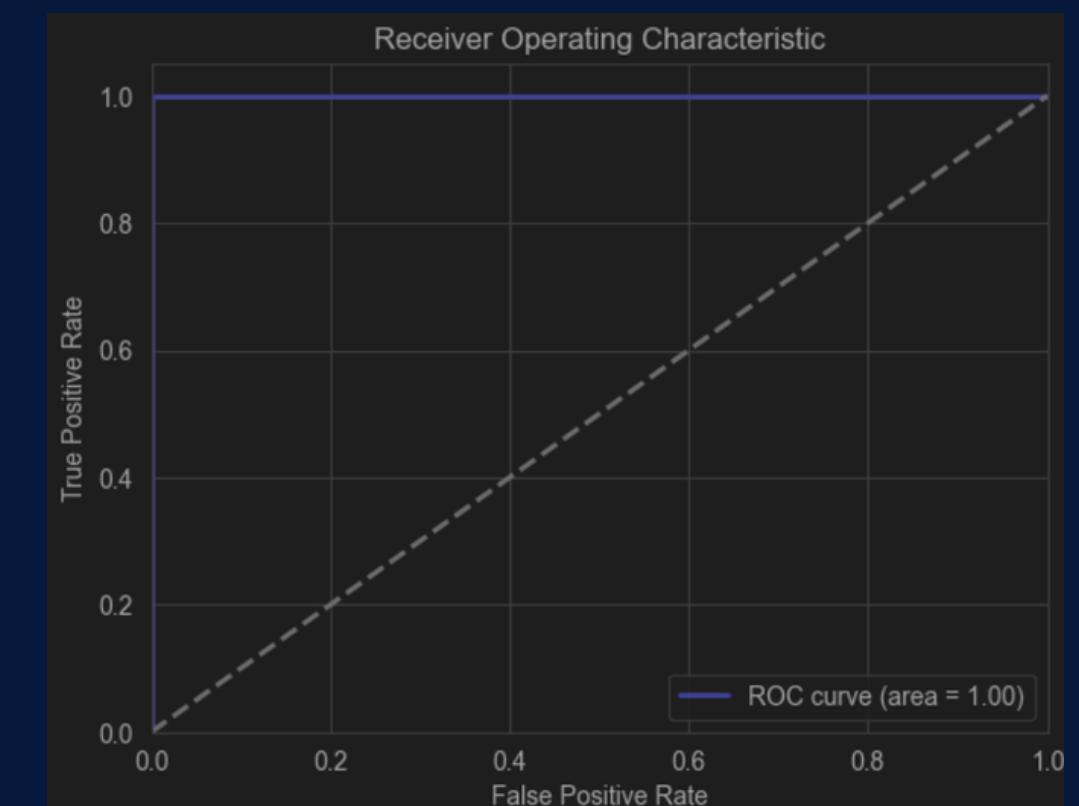
	precision	recall	f1-score	support
False	1.00	0.57	0.73	7
True	0.90	1.00	0.95	26
accuracy			0.91	33
macro avg	0.95	0.79	0.84	33
weighted avg	0.92	0.91	0.90	33

# Modelação e Avaliação de dados - DT

Matrizes de confusão

Curvas ROC

Relatórios de classificação



Relatório de Classificação:

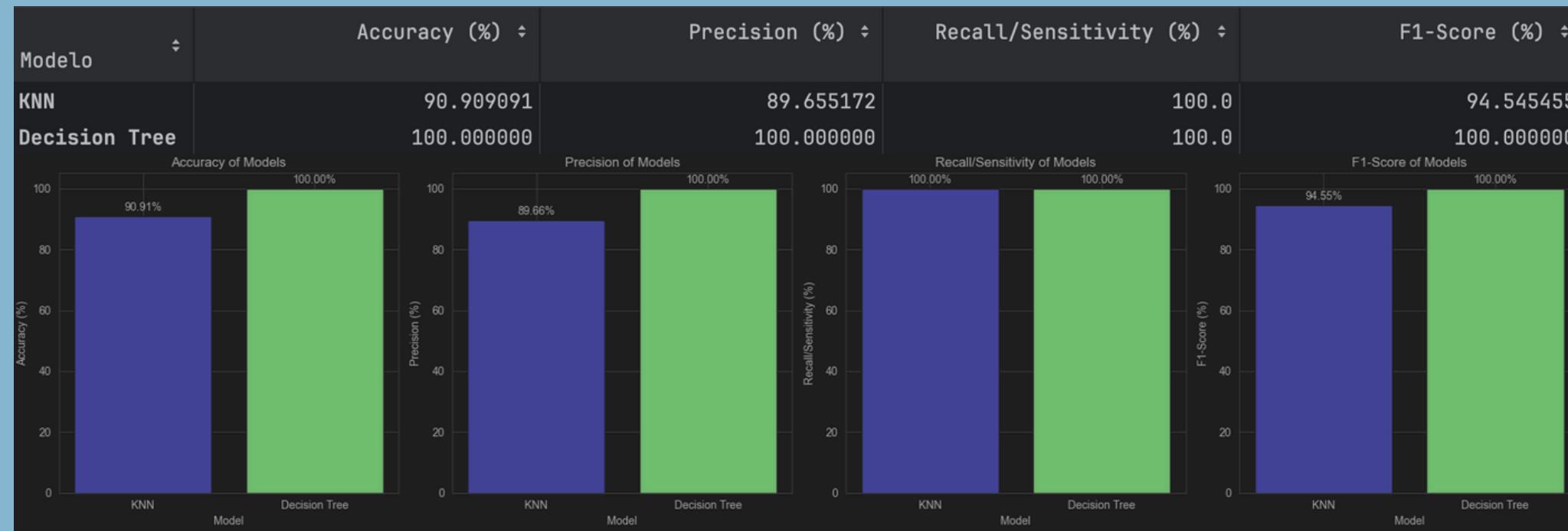
	precision	recall	f1-score	support
False	1.00	1.00	1.00	7
True	1.00	1.00	1.00	26
accuracy			1.00	33
macro avg	1.00	1.00	1.00	33
weighted avg	1.00	1.00	1.00	33

# Interpretação dos resultados



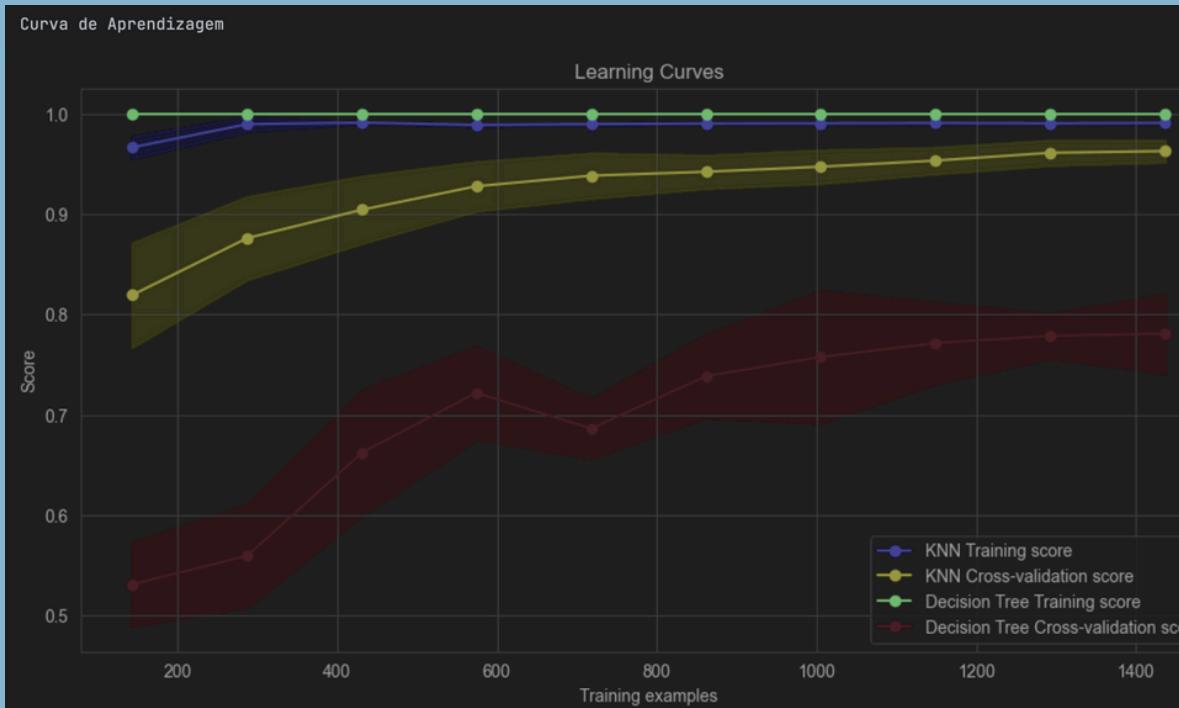
- Analisando estes dados, verificamos que a Árvore de Decisão tem uma precisão perfeita, ou seja, classificou corretamente todas as amostras do conjunto de teste, enquanto o K-Nearest Neighbors (KNN) apresenta uma precisão elevada.
- A precisão da Árvore de Decisão indica que todos os exemplos classificados como positivos são realmente positivos. No caso do KNN, ocorreram alguns falsos positivos.
- Quanto à sensibilidade, ambas as técnicas atingiram um valor perfeito, ou seja, todos os exemplos positivos reais foram identificados corretamente por ambos os modelos.

# Interpretação dos resultados



- Por fim, o F1-Score depende diretamente da combinação entre a precisão e a sensibilidade. Portanto, devido à menor precisão do modelo KNN, o seu F1-Score é inferior.
- O desempenho perfeito da Árvore de Decisão pode ser um sinal de overfitting, onde o modelo aprende muito bem os dados de treino, mas pode não generalizar bem para novos dados.

# Interpretação dos resultados



- Na primeira figura, a Curva de Aprendizagem mostra como os modelos se comportam à medida que temos mais dados para analisar. As linhas representam o desempenho de cada modelo e as sombras indicam o quanto variável esse desempenho pode ser.
- As linhas verdes e azuis mostram que os modelos têm um bom desempenho mesmo com poucos dados. Como as sombras são pequenas, isso significa que o desempenho é consistente e confiável.
- A curva amarela aproxima-se gradualmente do seu melhor desempenho à medida que temos mais dados. Também notamos que o modelo torna-se mais estável com mais dados. Já a curva vermelha melhora gradualmente, mas não atinge o mesmo nível de desempenho e tem uma sombra maior, o que indica que o modelo é menos estável e mais sensível a mudanças nos dados.
- Na curva ROC, o KNN (azul) tem uma pontuação de 0.92, o que significa que é excelente a distinguir corretamente entre classes positivas e negativas. A Árvore de Decisão tem um desempenho perfeito, o que significa que não comete erros ao distinguir entre classes.
- Ambos os modelos têm uma área sob a curva (AUC) elevada, o que sugere que são robustos para a tarefa de classificação. No entanto, como mencionado anteriormente, a perfeição do modelo Árvore de Decisão pode ser um indício de que se está a ajustar demasiado os dados (overfitting). Assim, é aconselhável optar pelo KNN, especialmente com conjuntos de dados pequenos ou pouco diversificados.

