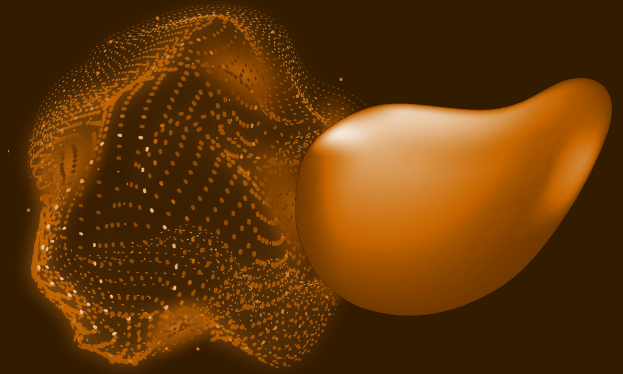


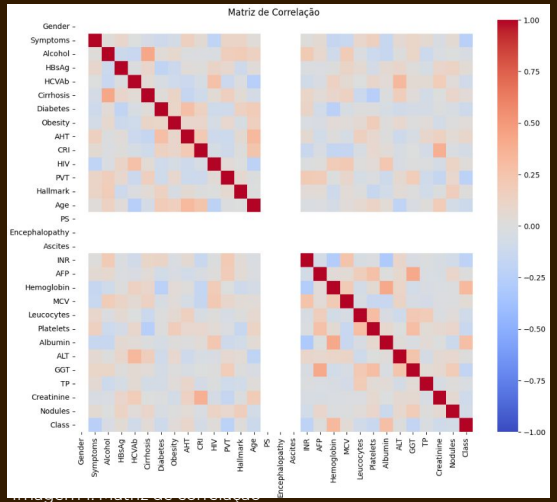
Data exploration and enrichment for supervised classification

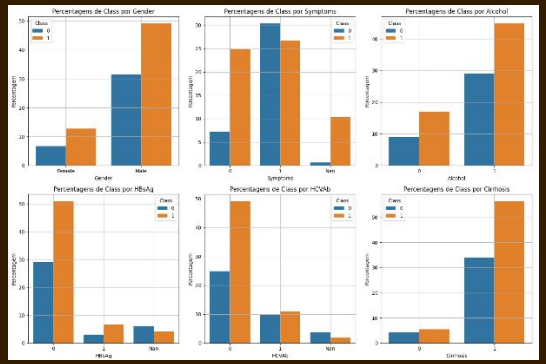
Elementos de IACD

Grupo 16

Daniela Leitão, Gonçalo Cruz, Leonor Ribeiro

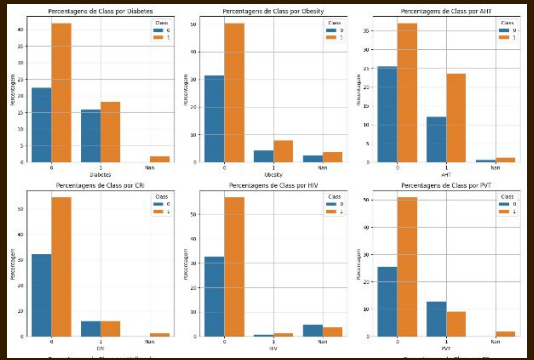
DATA EXPLORATION





Nota:
Para além da matriz de correlação e estes plot diffs, fizemos também histogramas para todas as categorias, testes de relevância, etc.

Imagens 2 e 3: alguns gráficos (plot_diffs()) percentuais de comparação entre cada categoria e a 'Class'



A exploração dos dados permitiu identificar relações entre variáveis, destacar padrões e possíveis áreas de melhoria.

DATA PREPROCESSING

Para melhorar a qualidade dos dados e o desempenho dos métodos de machine learning:

01 **Imputação de valores em falta**

Substituição dos valores em falta pela moda da coluna correspondente

02 **Remoção de colunas com base na correlação**

Identificação de pares de colunas com alta correlação e eliminação das que possuem menor correlação com a variável alvo (‘Class’).

03 **Remoção de colunas com base na frequência de 'Nan' e na variância**

Identificação de colunas com mais valores em falta ('Nan') e mais baixa variância

Colunas removidas: Grams_day, Spleno, PHT, Dir_Bil, AST, Sat, Ferritin, Iron, Packs_year, Varices, Smoking, HBeAg, Endemic, HBcAb e Hemochro.

DATA MODELING

```
Melhor número de vizinhos para KNN: 16
Classification Report for KNN:
              precision    recall  f1-score   support

           0       0.83      0.45      0.59        11
           1       0.76      0.95      0.84        20

    accuracy                           0.77        31
   macro avg       0.80      0.70      0.72        31
weighted avg       0.79      0.77      0.75        31

Confusion Matrix for KNN:
[[ 5  6]
 [ 1 19]]
Acurácia média com validação cruzada para KNN: 0.5966666666666667
```

Imagem 4: Resultados do algoritmo KNN

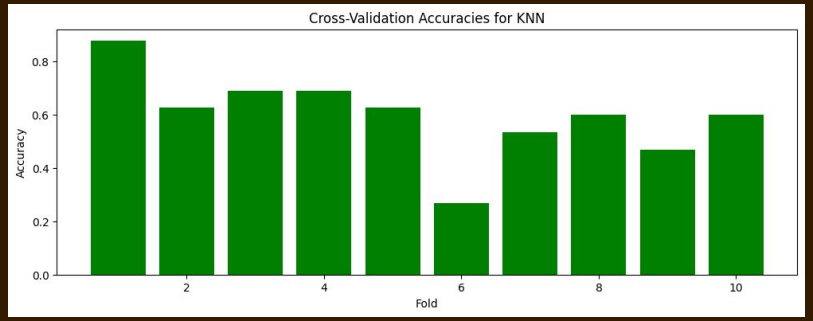


Imagem 5: Gráfico de barras para exatidão da validação cruzada

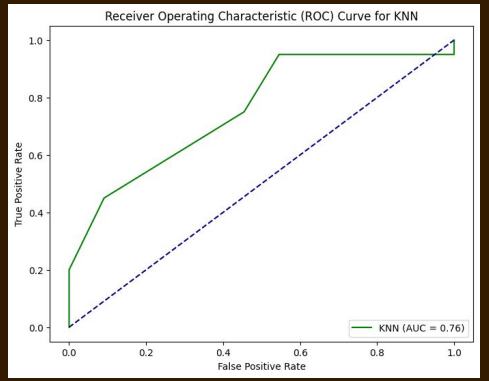


Imagem 6: Curva ROC

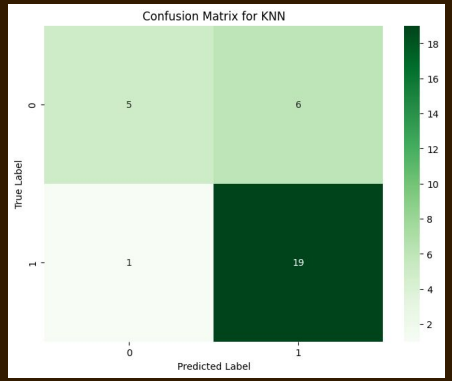


Imagem 7: Matriz de confusão

DATA MODELING

```
Classification Report for Decision Tree:
              precision    recall  f1-score   support

           0       0.56      0.45      0.50        11
           1       0.73      0.80      0.76        20

    accuracy                           0.68        31
   macro avg       0.64      0.63      0.63        31
weighted avg       0.67      0.68      0.67        31

Confusion Matrix for Decision Tree:
[[ 5  6]
 [ 4 16]]
Acurácia média com validação cruzada: 0.6266666666666667
```

Imagem 8: Resultados do algoritmo Decision Tree

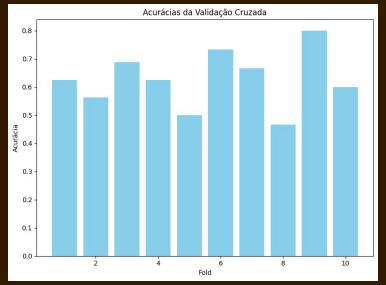


Imagem 9: Gráfico de barras para exatidão da validação cruzada

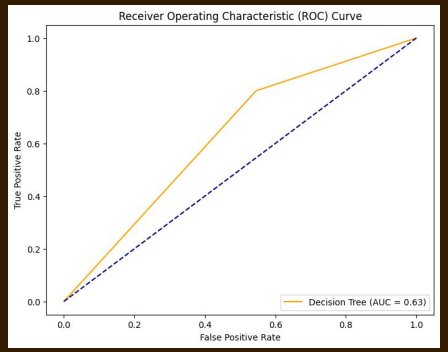


Imagem 10: Curva ROC

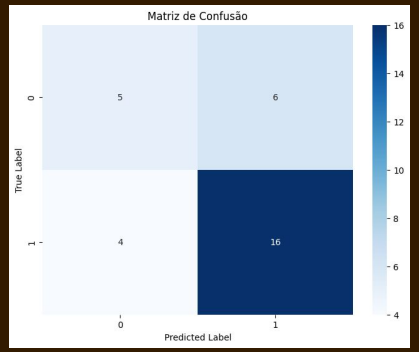


Imagem 11: Matriz de confusão

DATA MODELING

```
Classification Report for Random Forest:
              precision    recall  f1-score   support

           0       0.60      0.55      0.57        11
           1       0.76      0.80      0.78        20

    accuracy                           0.71        31
   macro avg       0.68      0.67      0.68        31
weighted avg       0.70      0.71      0.71        31

Confusion Matrix for Random Forest:
[[ 6  5]
 [ 4 16]]
Acurácia média com validação cruzada (Random Forest): 0.6775
```

Imagem 12: Resultados do algoritmo Random Forest

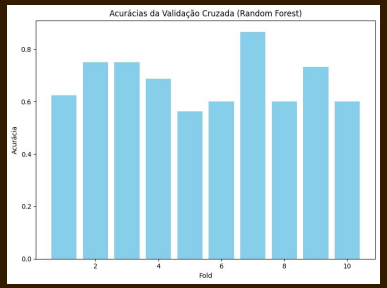


Imagem 13: Gráfico de barras para exatidão da validação cruzada

×

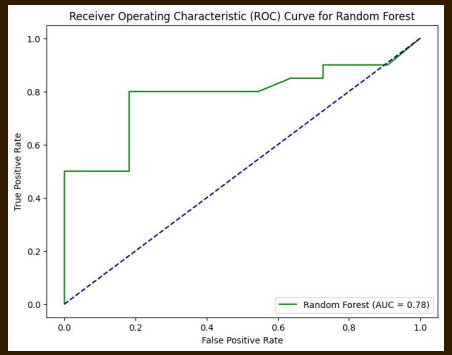


Imagem 14: Curva ROC

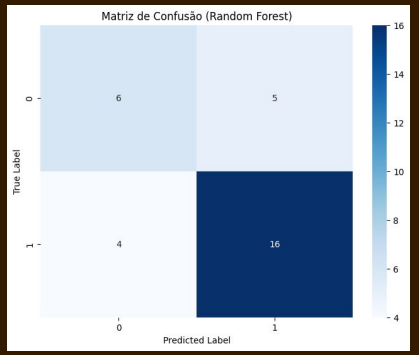


Imagem 15: Matriz de confusão

DATA EVALUATION

Comparison of Classification Metrics:

	Metric	Decision Tree	KNN
0	precision	0.666341	0.786022
1	recall	0.677419	0.774194
2	f1-score	0.668971	0.753532
3	AUC	0.627273	0.756818

Imagem 16: Comparação dos métodos de classificação

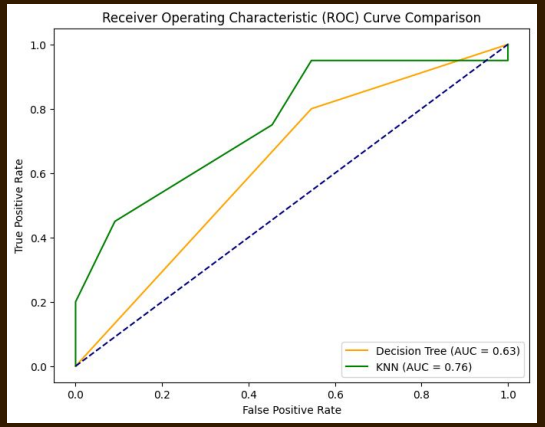


Imagem 17: Curva ROC com KNN e Decision Tree

DATA EVALUATION

Comparison of Classification Metrics:

	Metric	Decision Tree	KNN	Random Forest
0	precision	0.666341	0.786022	0.704455
1	recall	0.677419	0.774194	0.709677
2	f1-score	0.668971	0.753532	0.706305
3	AUC	0.627273	0.756818	0.784091

Imagem 18: Comparação dos métodos de classificação

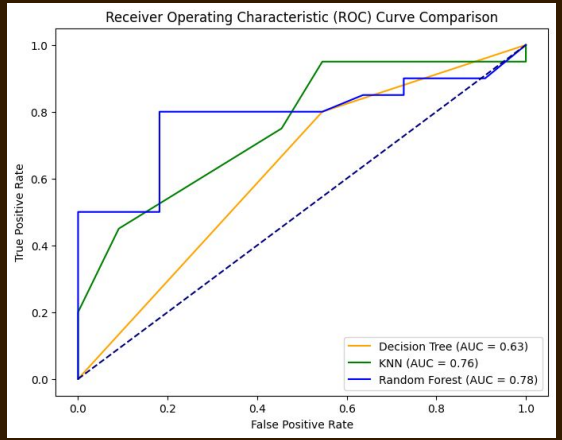


Imagem 19: Curva ROC com KNN , Decision Tree e Random Forest

INTERPRETATION OF RESULTS

	Decision Tree	KNN	Random Forest
×	• Precisão: 0.67 • Recall: 0.68 • F1-score: 0.67	• Precisão: 0.79 • Recall: 0.77 • F1-score: 0.75	• Precisão: 0.70 • Recall: 0.71 • F1-score: 0.71

O modelo Decision Tree tem um desempenho moderado, com precisão, recall e F1-score entre 0.67-0.68, ou seja, identifica corretamente cerca de 67-68% dos casos positivos.

O modelo KNN tem um desempenho relativamente bom, com precisão, recall e F1-score entre 0.75-0.79. Consegue identificar mais corretamente casos positivos e negativos. No entanto, possui um recall mais baixo em comparação com a sua precisão, o que pode indicar que pode perder alguns casos positivos.

O modelo Random Forest tem um desempenho semelhante ao KNN, com precisão, recall e F1-score entre 0.70-0.71. É equilibrado na identificação de casos positivos e negativos, e também possui um recall relativamente mais alto em comparação com a precisão, indicando que é melhor a capturar casos positivos do que o KNN.

INTERPRETATION OF RESULTS

Comparação dos algoritmos

- Precisão: KNN > Random Forest > Decision Tree
- Recall: KNN > Random Forest > Decision Tree
- F1-score: KNN > Random Forest > Decision Tree
- AUC: Random Forest > KNN > Decision Tree

De uma forma geral, podemos dizer que em termos de precisão, recall e F1-score, os algoritmos KNN e Random Forest superam a Decision Tree. No entanto, o Random Forest tem maior pontuação AUC, indicando um melhor desempenho geral para previsões de classificação.

Numa análise futura, poderíamos recorrer à escolha de um conjunto diferente de categorias/colunas para eliminar durante a etapa DATA PREPROCESSING, na tentativa de potenciar um melhor desempenho dos algoritmos usados, e tentar usar ainda outros algoritmos de classificação.

Assim, o método de classificação Random Forest parece ser o algoritmo mais eficaz para o nosso conjunto de dados, pois oferece um bom equilíbrio entre precisão, recall e desempenho preditivo geral.