MINISTÉRIO DA EDUCAÇÃO
Secretaria de Educação Profissional e Tecnológica
Instituto Federal de Educação, Ciência e Tecnologia do Piauí
Campus Teresina Central

Disciplina: Tópicos Especiais em Computação

Período / Turma: 2025.1 / ADS-5

Análise Comparativa de Algoritmos de Classificação em Bases de Dados Padrão com Python

Ana Beatriz Farias, Gabriel de Almeida e Mariana Andrade

1 INTRODUÇÃO

O avanço da inteligência artificial e do aprendizado de máquina tem permitido a criação de modelos computacionais capazes de aprender padrões a partir de dados. No contexto da classificação supervisionada, diferentes algoritmos podem ser aplicados para prever categorias ou classes com base em atributos fornecidos. Este trabalho tem como objetivo aplicar e comparar três algoritmos clássicos de classificação: Árvore de Decisão, Naive Bayes e K-Nearest Neighbors (KNN). Foram utilizadas seis bases de dados amplamente conhecidas na literatura: a base Iris, base Breast Cancer Wisconsin (Diagnostic), base Oxford Parkinson's Disease Detection Dataset, base Heart Disease, Optical Recognition of Handwritten Digits e Wine.

A comparação dos modelos é realizada por meio de métricas de desempenho amplamente utilizadas em problemas de classificação: acurácia, precisão, revocação (recall), F1-score e matriz de confusão, todas calculadas utilizando a biblioteca scikit-learn em Python.

- Acurácia representa a proporção de previsões corretas em relação ao total de exemplos avaliados, indicando o desempenho geral do modelo.

- Precisão mede a proporção de verdadeiros positivos entre todas as instâncias classificadas como positivas pelo modelo, refletindo a exatidão das previsões positivas.

- Revocação (Recall) indica a capacidade do modelo de identificar corretamente todas as instâncias positivas reais, ou seja, a proporção de verdadeiros positivos em relação ao total de positivos existentes.

- F1-score é a média harmônica entre precisão e revocação, fornecendo uma única métrica que equilibra os dois aspectos, especialmente útil quando há desbalanceamento entre classes.

- A matriz de confusão apresenta a distribuição dos acertos e erros do modelo, detalhando as classificações verdadeiras versus as preditas para cada classe, facilitando a análise dos tipos de erros cometidos.

Essas métricas fornecem uma avaliação abrangente do desempenho dos algoritmos aplicados, possibilitando uma comparação criteriosa entre eles para as bases de dados analisadas.

2 METODOLOGIA

A metodologia adotada envolve as seguintes etapas:

1. Coleta e preparação dos dados: foram utilizadas seis bases do repositório UCI Iris (com três classes de flores), Breast Cancer (com diagnóstico benigno ou maligno), Wine (com três classes de cultivares de vinho) e Optical Recognition of Handwritten Digits (com dez classes de dígitos de 0 a 9), Heart Disease (valores inteiros de 0 a 4, sendo que 0 indica a não presença da doença cardíaca), Parkinson Disease (classes 0 e 1, que indicam a ausência ou presença da DP). As variáveis categóricas foram codificadas com LabelEncoder e os dados foram divididos entre treino e teste (70/30 ou 80/20, dependendo do experimento).

2. Pré-processamento: para o algoritmo KNN, foi realizada a padronização dos atributos com StandardScaler, visto que esse modelo é sensível à escala dos dados.

3. Modelagem: três algoritmos foram implementados:

    - Árvore de Decisão (ID3) com profundidade máxima limitada a 3, utilizando critério de entropia.

    - Naive Bayes (GaussianNB), baseado na suposição de normalidade dos atributos.

    - K-Nearest Neighbors (KNN) com k=3.

4. Avaliação dos modelos: as previsões foram comparadas com os rótulos reais do conjunto de teste utilizando métricas de desempenho e visualizações, como relatórios de classificação e matrizes de confusão com seaborn.

3 CARACTERÍSTICAS DAS BASES DE DADOS

Base de Dados	Nº de Instâncias	Nº de Atributos (Features)
Heart Disease	303	13
Iris	150	4
Wine	178	13
Digits	5.620	64
Parkinson’s Disease	197	22
Breast Cancer	569	30

4 ALGORITMOS

4.1 Árvore de decisão

Base de dados	Acurácia	Precisão	Revocação	F1-Score
Breast Cancer	0,96	0,97	0,95	0,96
Iris	1,00	1,00	1,00	1,00
Wine	0,91	0,93	0,87	0,88
Digits	0,55	0,42	0,55	0,46
Heart Disease	0,53	0,24	0,26	0,24
Parkinson's Disease	0,86	0,90	0,90	0,90
Média por Algoritmo	**0,80**	**0,74**	**0,76**	**0,74**

4.2 Classificador Naive Bayes

Base de dados	Acurácia	Precisão	Revocação	F1-Score
Breast Cancer	0,94	0,94	0,93	0,94
Iris	0,98	0,98	0,97	0,97
Wine	1,00	1,00	1,00	1,00
Digits	0,76	0,85	0,77	0,77
Heart Disease	0,43	0,21	0,31	0,21
Parkinson's Disease	0,74	0,91	0,72	0,81
Média por Algoritmo	**0,81**	**0,82**	**0,78**	**0,78**

4.3 K-Nearest Neighbors (KNN)

Base de dados	Acurácia	Precisão	Revocação	F1-Score
Breast Cancer	0,96	0,96	0,95	0,96
Iris	1,00	1,00	1,00	1,00
Wine	0,96	0,96	0,96	0,96
Digits	0,98	0,98	0,98	0,98

Heart Disease	0,58	0,31	0,31	0,31
Parkinson's Disease	0,89	0,89	0,97	0,93
Média por Algoritmo	**0,90**	**0,85**	**0,86**	**0,86**

5. DISCUSSÃO DOS RESULTADOS

Com base nas três tabelas apresentadas para os algoritmos: Árvore de Decisão, Naive Bayes, K-Nearest Neighbors (KNN):

Melhor Algoritmo por Base de Dados:

Base de Dados	Melhor Algoritmo	Justificativa
Breast Cancer	KNN	Métricas todas ≥ 0,95.
Iris	Todos empataram	Todos os algoritmos tiveram boas métricas .
Wine	Naive Bayes	Todos os algoritmos tiveram boas métricas .
Digits	KNN	Obteve todas as métricas em torno de 0,98, a árvore teve desempenho fraco.
Heart Disease	KNN	Apesar da baixa performance geral, teve métricas superiores aos demais.
Parkinson's Disease	KNN	F1-Score 0,93 contra 0,90 (Árvore) e 0,81 (Naive Bayes).

Base de dados com melhor desempenho geral:

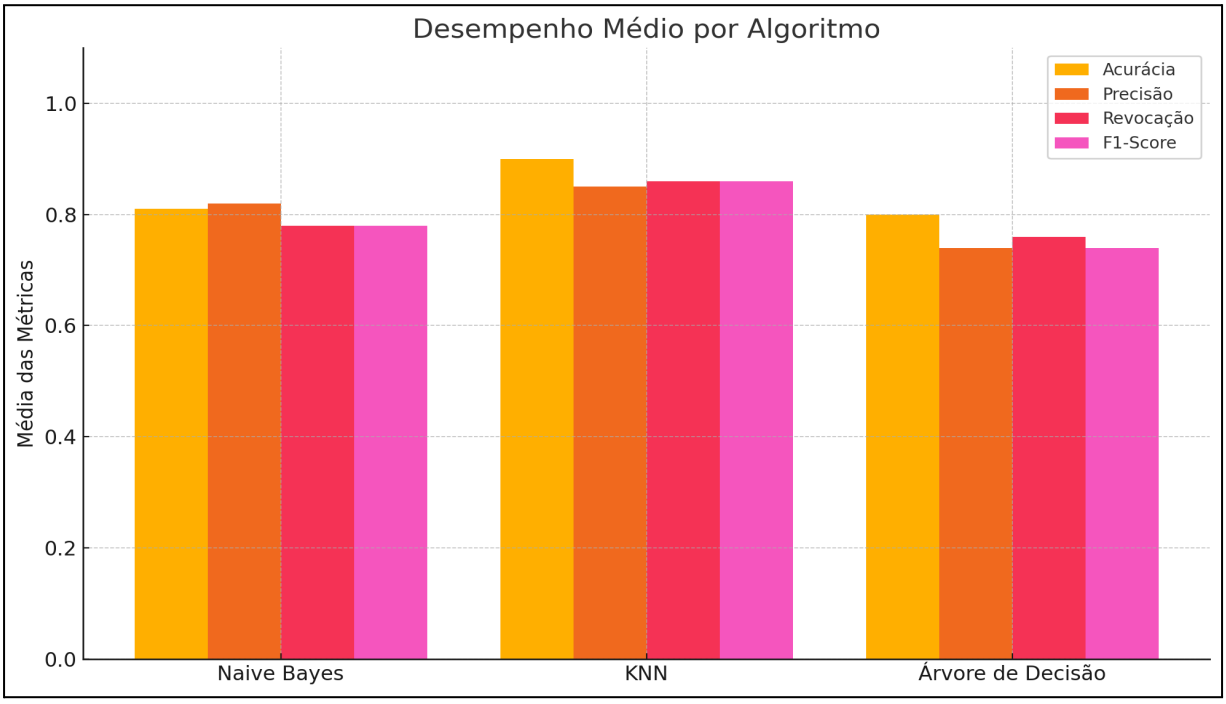
Iris	Todos os algoritmos com métricas = 1
Wine	Naive Bayes atingiu 1.00 em todas as métricas, e também métricas próximas a 1 nos outros dois algoritmos.

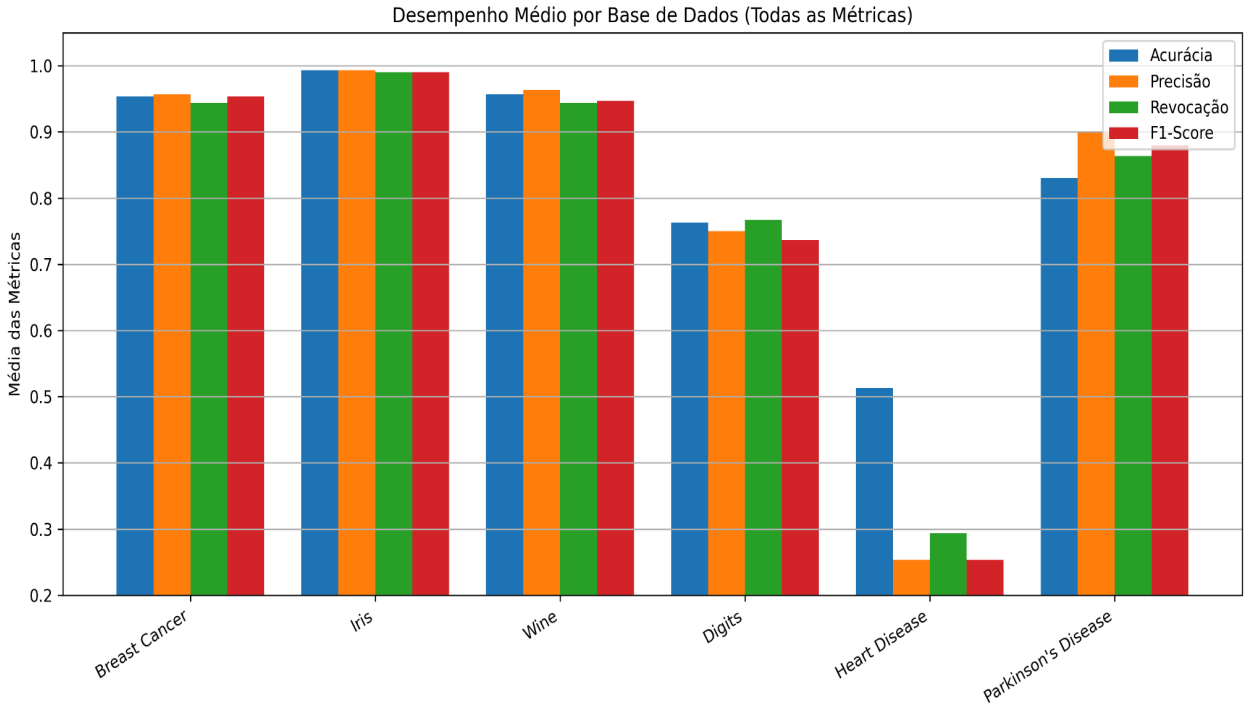
Com base nesses dados:

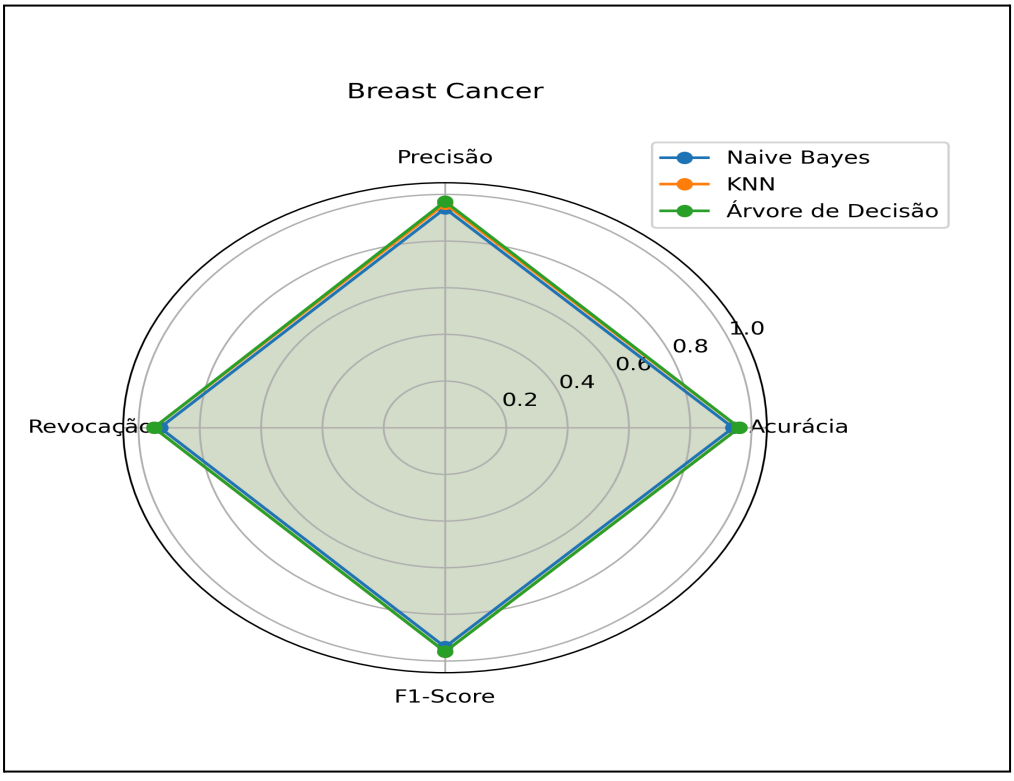
KNN se destacou como o algoritmo mais robusto, mantendo bom desempenho em praticamente todas as bases, especialmente nas mais desafiadoras(Heart Disease, Parkinson).

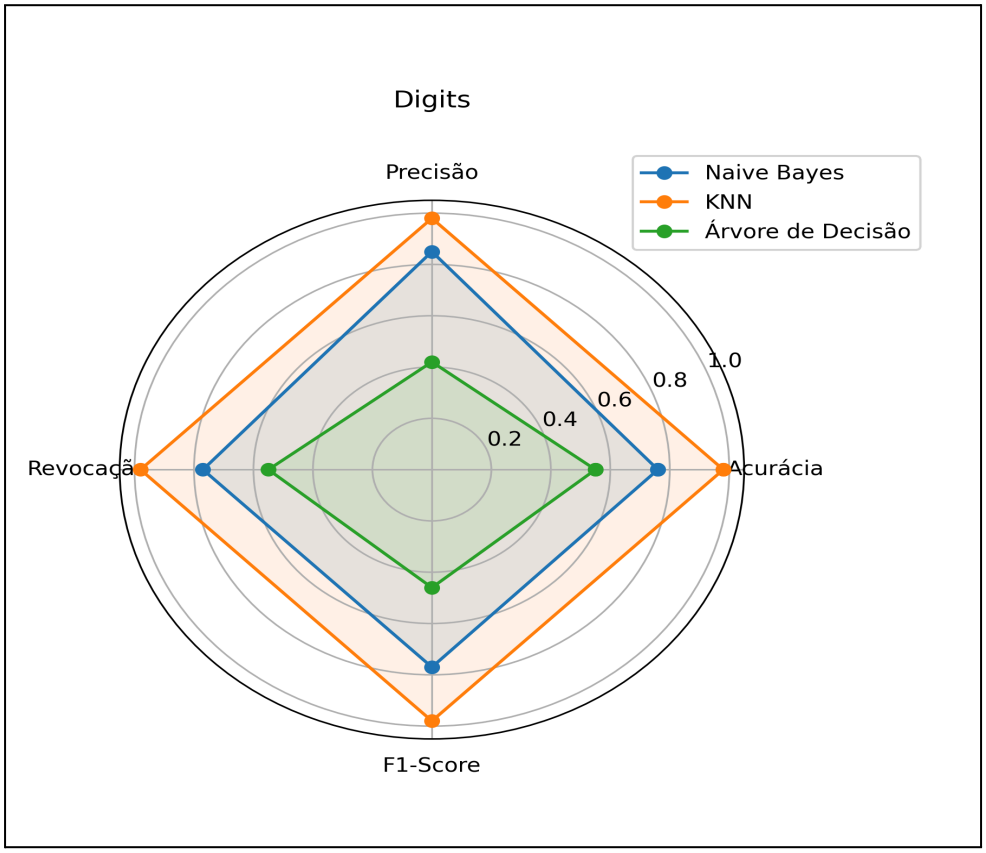
Naive Bayes foi o melhor em Wine, alcançando métricas perfeitas, mas teve grande queda em Heart Disease.

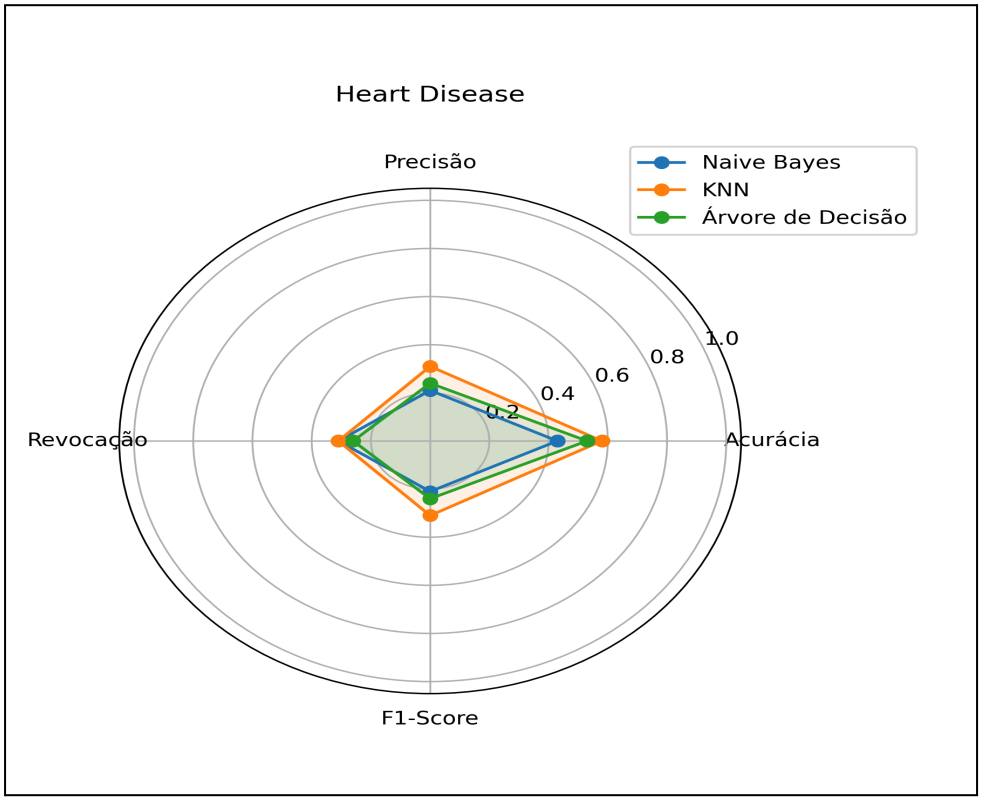
Árvore de Decisão teve desempenho instável, com bons resultados em Iris, Breast Cancer e Parkinson, mas fraco em Digits e Heart Disease. As bases Iris, Wine e Breast Cancer são mais fáceis de classificar, dado o alto desempenho de todos os modelos. A base Heart Disease é a mais desafiadora, com desempenho insatisfatório mesmo com o melhor algoritmo.

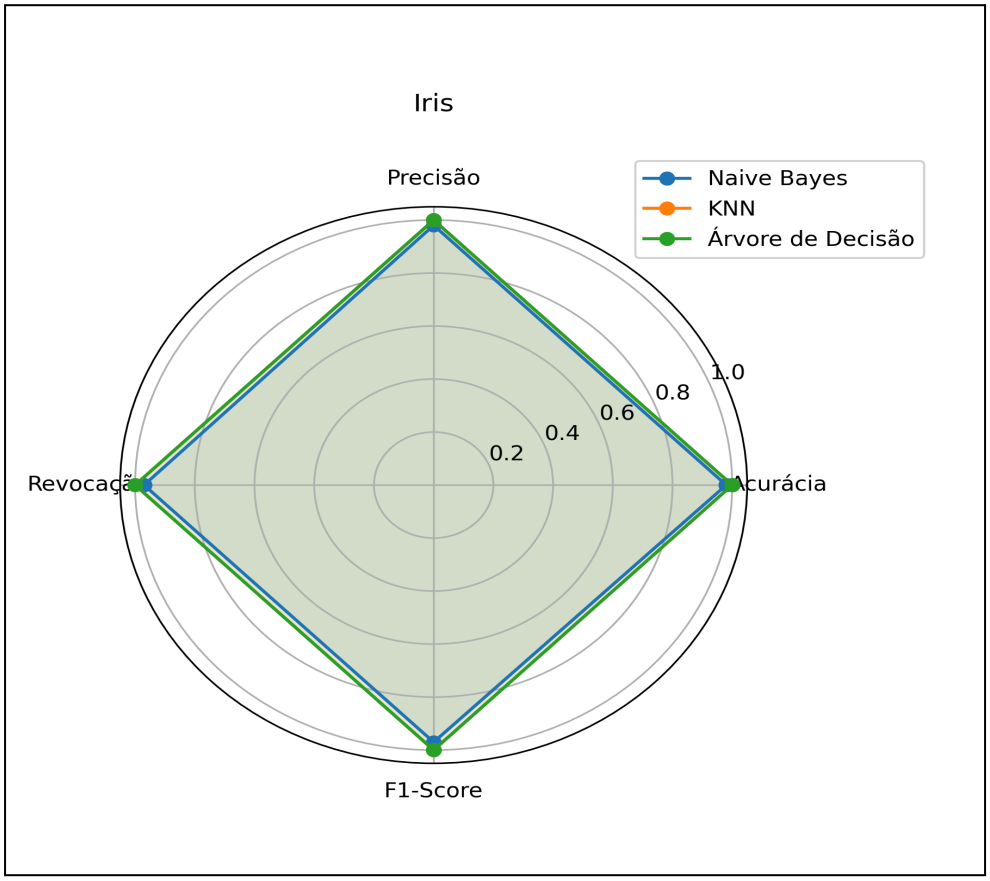


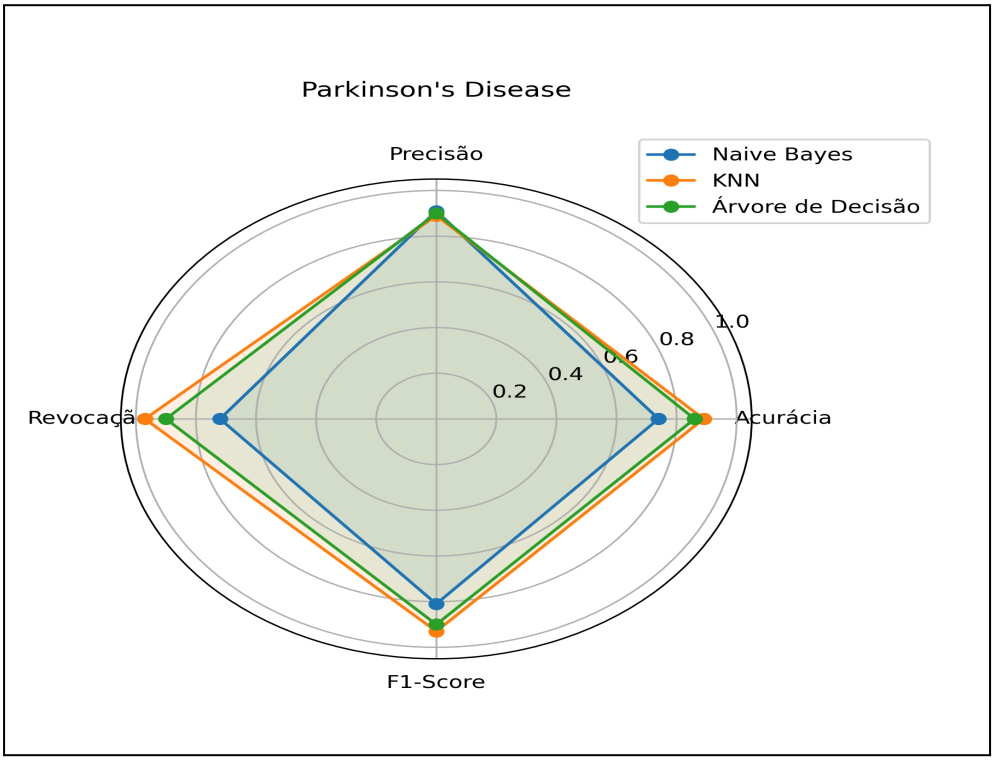

Desempenho Médio por Base de Dados (Todas as Métricas)
Média das Métricas
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
Acurácia
Precisão
Revocação
F1-Score
Breast Cancer
Iris
Wine
Digits
Heart Disease
Parkinson's Disease

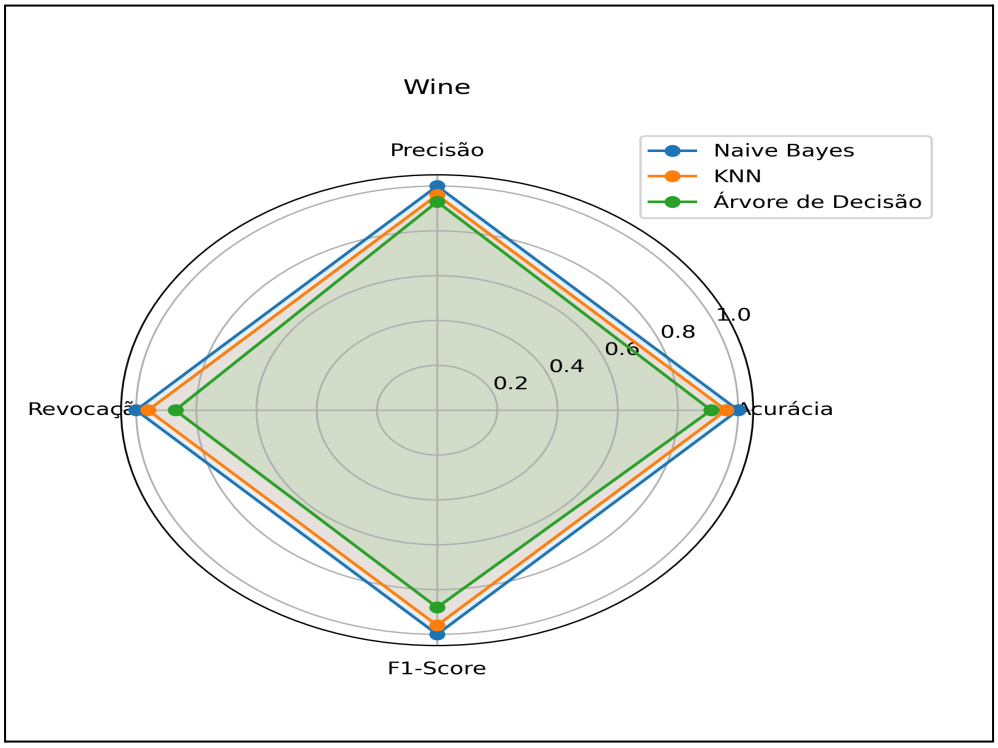

Breast Cancer
Naive Bayes
KNN
Árvore de Decisão
Precisão
Acurácia
F1-Score
Revocação
0.2
0.4
0.6
0.8
1.0


Digits
Precisão
Naive Bayes
KNN
Árvore de Decisão
1.0
0.8
0.6
0.4
0.2
Revocação
Acurácia
F1-Score


Heart Disease
Precisão
Naive Bayes
KNN
Árvore de Decisão
1.0
0.8
0.6
0.4
0.2
Revocação
Acurácia
F1-Score


Iris
Precisão
Naive Bayes
KNN
Árvore de Decisão
1.0
0.8
0.6
0.4
0.2
Revocaçã
curácia
F1-Score


Parkinson's Disease
Precisão
Naive Bayes
KNN
Árvore de Decisão
1.0
0.8
0.6
0.4
0.2
Revocaçã
Acurácia
F1-Score



6 CONCLUSÃO

A análise comparativa entre os algoritmos revelou que o desempenho dos modelos está fortemente associado às características das bases de dados, como número de instâncias, quantidade de atributos (features) e complexidade do problema.

O algoritmo K-Nearest Neighbors (KNN) apresentou, de forma geral, os melhores resultados médios em todas as métricas avaliadas (acurácia, precisão, revocação e F1-Score), com destaque para a base Digits, onde alcançou F1-Score de 0,98, indicando sua eficácia em bases com grande número de instâncias e alta dimensionalidade (64 atributos). Além disso, obteve excelente desempenho nas bases Iris, Wine e Parkinson's Disease, demonstrando robustez em diferentes contextos.

O classificador Naive Bayes mostrou desempenho satisfatório em bases mais simples ou bem comportadas, como Iris e Wine (F1 = 0,97 e 1,00, respectivamente), mas seu desempenho caiu em bases com atributos mais complexos e interdependentes, como Parkinson's (F1 = 0,81) e Heart Disease (F1 = 0,21). Isso é esperado, dado que o Naive Bayes assume independência entre os atributos, o que raramente é o caso em bases reais com muitos atributos correlacionados.

Já a Árvore de Decisão teve desempenho mais variável, indo de um F1-Score perfeito (1,00) na base Iris, até um valor bastante baixo (0,24) na base Heart Disease. Isso sugere

que esse algoritmo é mais sensível ao volume de dados e à complexidade da base, podendo sofrer com overfitting em bases pequenas e com ruído.

Os resultados mostram que não há um algoritmo universalmente melhor: o desempenho depende da natureza dos dados. O KNN se destacou pela consistência, mas Naive Bayes e Árvore de Decisão podem ser vantajosos em situações específicas, como bases simples ou quando a interpretabilidade é uma prioridade.

Portanto, a escolha do classificador ideal deve considerar não apenas as métricas, mas também o contexto da base de dados, os requisitos computacionais e os objetivos da aplicação.