Trabalho Prático Árvore de Decisão
Relatório

Dario M. V. P. Mousinho[1], **Layse Gomes Castro**[1]

[1]Universidade Federal Fluminense (UFF)

1. Introdução

Esse trabalho tem como objetivo aplicar técnicas de Aprendizado de Máquina supervisionado para resolver problemas de classificação utilizando o algoritmo de Árvores de Decisão. Foram utilizados dois conjuntos de dados: o *Iris* [Fisher 1936] e o *Breast Cancer Wisconsin (Diagnostic)* [Wolberg and Street 1993]. Ambos os datasets estão disponíveis no repositório da Universidade da Califórnia em Irvine (UCI Machine Learning Repository).

As Árvores de Decisão permitem a interpretação (caixa branca), realizam partições sucessivas no espaço de atributos com base em critérios de impureza, como Gini ou Entropia. A principal vantagem desse tipo de modelo está na simplicidade e na facilidade de interpretar os resultados [Mitchell and Mitchell 1997].

A proposta metodológica consiste em realizar a análise exploratória dos dados, treinar modelos de árvore com e sem ajuste de hiperparâmetros, avaliar o desempenho por meio de métricas clássicas de classificação e, por fim, comparar os resultados obtidos nos dois contextos.

2. Iris Dataset

2.1. Contextualização

O gênero Iris compreende diversas espécies de plantas com flores reconhecidas por sua simetria e coloração marcante. Entre elas, destacam-se três espécies em particular: Iris setosa, Iris versicolor e Iris virginica. Essas espécies têm sido amplamente utilizadas como modelo em estudos de aprendizado de máquina, notadamente no conjunto de dados introduzido por Ronald Fisher, um marco na história da classificação estatística multivariada [Unwin and Kleinman 2021].

As espécies Iris setosa, Iris versicolor e Iris virginica diferenciam-se, entre outros aspectos, pelo padrão de disposição e morfologia de suas folhas. A Iris setosa, encontrada em regiões de clima frio como o Alasca, o Canadá e partes da Ásia, apresenta folhas estreitas e rígidas, com uma base frequentemente pigmentada. A Iris versicolor, nativa de áreas úmidas da América do Norte, possui folhas em forma de espada que se organizam em leques verticais a partir de rizomas densos. Por sua vez, a Iris virginica, que habita zonas alagadiças do leste dos Estados Unidos, exibe folhas longas, brilhantes e levemente arqueadas, adaptadas a ambientes com alta umidade [Anderson 1928].

As distinções morfológicas entre essas três espécies possibilitam seu uso em estudos de classificação, sendo essenciais para a validação de métodos estatísticos e computacionais aplicados à análise de dados biológicos.

2.2. Objetivos

O objetivo desse trabalho é construir um modelo capaz de classificar corretamente uma amostra de flor entre três espécies do gênero *Iris*: *Setosa*, *Versicolor* e *Virginica*. Foi utilizada a classificação com base em quatro atributos morfológicos:

1. Comprimento da sépala;
2. Largura da sépala;
3. Comprimento da pétala;
4. Largura da pétala.

Dessa forma, o atributo alvo deste estudo é a variável "classe", de natureza categórica, a qual indica a espécie correspondente a cada amostra. O propósito foi treinar um modelo capaz de, a partir do que foi aprendido nos dados rotulados, prever corretamente a classe de novos espécimes.

2.3. Metodologia

2.3.1. Obtenção e Preparação dos Dados

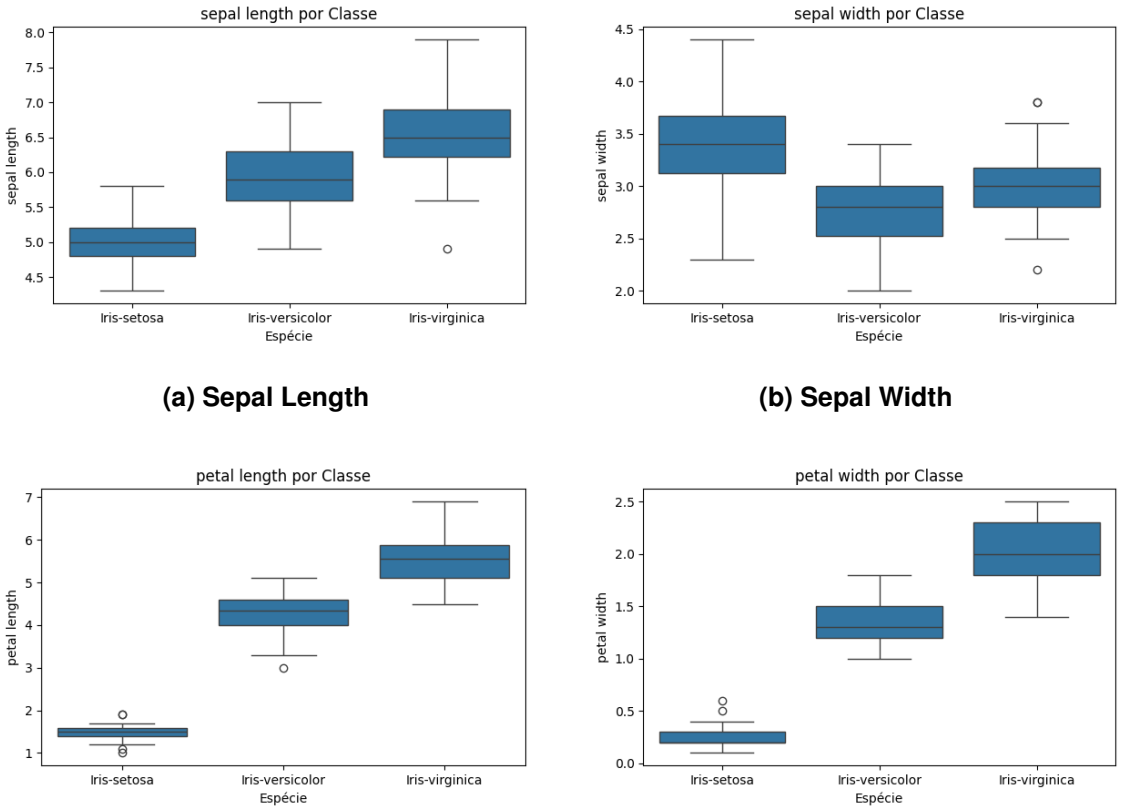
- O conjunto de dados Iris foi obtido a partir da biblioteca `ucimlrepo`, que acessa diretamente o repositório UCI Machine Learning Repository [Fisher 1936] .
- Os dados foram divididos em conjuntos de treino (70%) e teste (30%), com estratificação da variável alvo para manter a proporção das classes.

2.3.2. Análise Exploratória dos Dados

- O conjunto de dados Iris [Fisher 1936] contém 150 amostras, distribuídas igualmente entre as três espécies. Cada instância é caracterizada por quatro atributos contínuos. A análise exploratória revelou que a classe *Setosa* apresenta uma separação clara em relação às demais, enquanto *Versicolor* e *Virginica* possuem maior sobreposição nos atributos (figura 1).
- Foram utilizados gráficos de dispersão (*pairplot*), boxplots para cada atributo por classe, e a matriz de correlação para avaliar a relação entre os atributos. Identificou-se alta correlação entre o comprimento e a largura das pétalas, indicando possível redundância de informação (figura 2).

2.3.3. Árvore de Decisão: Treinamento Inicial

- Foi utilizado o algoritmo de árvore de decisão `DecisionTreeClassifier` da biblioteca `scikit_learn` [Pedregosa et al. 2011] como modelo preditivo.
- O modelo foi treinado com os hiperparâmetros padrão, utilizando o conjunto de treino.
- A avaliação inicial foi feita sobre o conjunto de teste, por meio das métricas: acurácia, matriz de confusão e relatório de classificação.



(a) Sepal Length **(b) Sepal Width**

(c) Petal Length **(d) Petal Width**

Figura 1. Boxplots dos atributos do conjunto Iris por classe. Observa-se boa separação entre as classes, especialmente nos atributos relacionados às pétalas.

2.3.4. Cross Validation

- Para estimar a generalização do modelo, foi aplicado o k-fold cross-validation, no qual foi utilizado k = 10.
- O procedimento envolveu o particionamento do conjunto completo em subconjuntos de treino e teste rotativos, a fim de obter uma avaliação mais robusta do desempenho do classificador.

2.3.5. Otimização de Hiperparâmetros

- Realizou-se uma busca em grade (Grid Search) para encontrar os melhores hiperparâmetros.
- Foram testadas diferentes combinações de profundidade máxima da árvore (max-depth) e critérios de impureza (“gini” e “entropy”).
- Também foi realizado o cross validation durante a execução do Grid Search, utilizamos k=10 novamente.

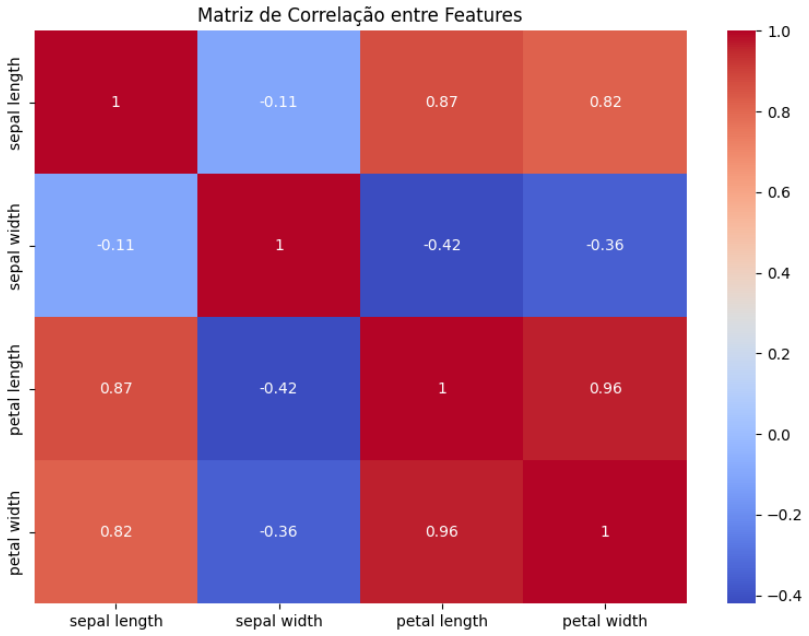


Figura 2. Correlação entre os atributos presentes no dataset Iris.

2.4. Métricas e Resultados

O desempenho do modelo de árvore de decisão foi avaliado com base no conjunto de teste, que corresponde a 30% das amostras do conjunto de dados Iris. A acurácia obtida nesse teste foi de 93,3%, o que indica uma boa capacidade preditiva do modelo mesmo antes da otimização.

O relatório de classificação apontou métricas perfeitas para a classe *Iris-setosa*, com precisão, recall e f1-score iguais a 1,00. No entanto, observou-se um recall inferior para a classe *Iris-versicolor* (0,80), o que revela que algumas amostras foram incorretamente classificadas como *Iris-virginica*. Esse comportamento é reforçado pela matriz de confusão, que mostra três erros especificamente entre essas duas classes, enquanto *Iris-setosa* não foi confundida com nenhuma outra espécie.

Com o objetivo de melhorar a performance do modelo, foi realizada uma busca em grade (*GridSearchCV*) para otimização de hiperparâmetros. Nesse processo, restringiu-se a busca à manipulação de dois parâmetros principais:

- **max_depth**: profundidade máxima da árvore, com valores testados entre 2 e 5;
- **criterion**: critério de impureza para divisão dos nós, sendo testados *gini* e *entropy*.

Durante a busca, utilizou-se novamente cross validation com 10 divisões (*k=10*), de forma a garantir a robustez estatística dos resultados. Os resultados das combinações testadas estão apresentados na Tabela 1.

A combinação que obteve o melhor desempenho (ranking 1) foi a árvore de decisão com profundidade máxima igual a 3 e critério de impureza *gini*. A acurácia média alcançada por essa configuração foi de 96,2% na validação cruzada, com um desvio padrão de 0,047, indicando boa estabilidade do modelo. A Figura 3 mostra a representação gráfica da árvore treinada com os hiperparâmetros ideais.

max_depth	critério	Acurácia Média	Desvio Padrão	Ranking
3	gini	0.9618	0.0469	1
3	entropy	0.9618	0.0469	1
2	gini	0.9518	0.0658	2
2	entropy	0.9518	0.0658	2
4	gini	0.9427	0.0765	3
4	entropy	0.9427	0.0765	3
–	entropy	0.9427	0.0765	3
–	gini	0.9418	0.0890	4
5	gini	0.9327	0.0872	4
5	entropy	0.9327	0.0749	4

Tabela 1. Resultados do GridSearchCV para DecisionTreeClassifier

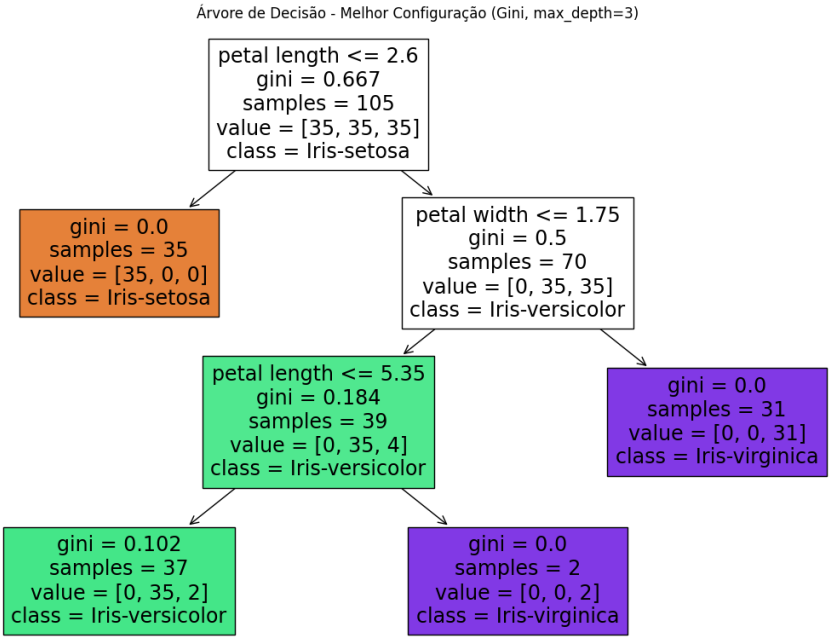


Figura 3. Árvore de Decisão treinada com os melhores hiperparâmetros (max_depth = 3, criterion = gini).

3. Breast Cancer Wisconsin Dataset

3.1. Contextualização

O diagnóstico do câncer de mama por meio de biópsia aspirativa por agulha fina (FNA do inglês *fine needle aspiration*) representa uma alternativa menos invasiva em relação à biópsia cirúrgica tradicional. No entanto, a avaliação morfológica das células extraídas por FNA depende fortemente da experiência do profissional, sendo, portanto, suscetível a variações subjetivas [Amedee and Dhurandhar 2001]. Com o intuito de aumentar a precisão e a objetividade desse processo diagnóstico, foi desenvolvido um sistema computacional que utiliza técnicas de processamento de imagem e aprendizado de máquina

para analisar características dos núcleos celulares em imagens microscópicas do câncer [Street et al. 1993].

As amostras presentes no conjunto de dados *Breast Cancer Wisconsin (Diagnostic)* foram obtidas a partir da digitalização de imagens de lâminas contendo células tumorais coradas, extraídas por FNA. Utilizando modelos de contorno ativos, conhecidos como *"snakes"*, a partir disso foi possível delimitar com precisão os núcleos celulares das células tumorais. Com esses contornos, foram extraídas 10 características morfológicas, sendo elas: raio, perímetro, área, suavidade, concavidade, simetria, textura e dimensão fractal [Wolberg and Street 1993].

O conjunto contém 569 amostras rotuladas como benignas ou malignas, sendo a variável "diagnóstico" o atributo alvo de interesse. Essas informações estruturadas em um espaço de características numéricas viabilizam a aplicação de algoritmos de classificação, com o objetivo de prever, de forma automatizada e precisa, a natureza de novos casos a partir do padrão aprendido nos dados rotulados. [Street et al. 1993]

3.2. Objetivos

O segundo objetivo deste trabalho consiste na utilização do conjunto de dados *Breast Cancer Wisconsin (Diagnostic)* para a construção de um modelo de árvore de decisão capaz de realizar a predição das classes *"benign"* (benigno) e *"malignant"* (maligno).

A classificação foi realizada com base em trinta atributos numéricos que representam 10 propriedades distintas e para cada uma delas foram computadas três estatísticas diferentes: a média, o erro padrão e o valor extremo (ou "pior"caso). As propriedades utilizadas para os trinta atributos foram:

1. Raio (média das distâncias do centro aos pontos do perímetro);
2. Textura (desvio padrão dos valores de intensidade em tons de cinza);
3. Perímetro;
4. Área;
5. Suavidade (variação local no comprimento dos raios);
6. Compacidade (razão entre o quadrado do perímetro e a área, subtraído de 1);
7. Concavidade (profundidade das regiões côncavas do contorno);
8. Pontos côncavos (número de segmentos côncavos ao longo do contorno);
9. Simetria,
10. Dimensão fractal (medida da complexidade da borda, aproximada pelo método da "linha costeira");

3.3. Metodologia

3.3.1. Obtenção e Preparação dos Dados

- O conjunto de dados Breast Cancer Wisconsin (Diagnostic) foi obtido a partir da biblioteca ucimlrepo, que acessa diretamente o repositório UCI Machine Learning Repository [Wolberg and Street 1993], contendo 569 amostras rotuladas como malignas (M) ou benignas (B).

- Os dados foram divididos em conjuntos de treino (70%) e teste (30%), com estratificação da variável alvo para manter a proporção original das classes.

3.3.2. Análise Exploratória dos Dados

- O conjunto possui 30 atributos numéricos derivados de 10 características morfológicas extraídas das imagens de biópsia, para cada uma das quais foram calculados três estatísticas: média, erro padrão e valor extremo.

- Devido à inviabilidade de se visualizar e interpretar todas as 30 variáveis individualmente, a análise foi direcionada apenas para os atributos com maior poder discriminativo, para isso, foi utilizada a matriz de correlação,

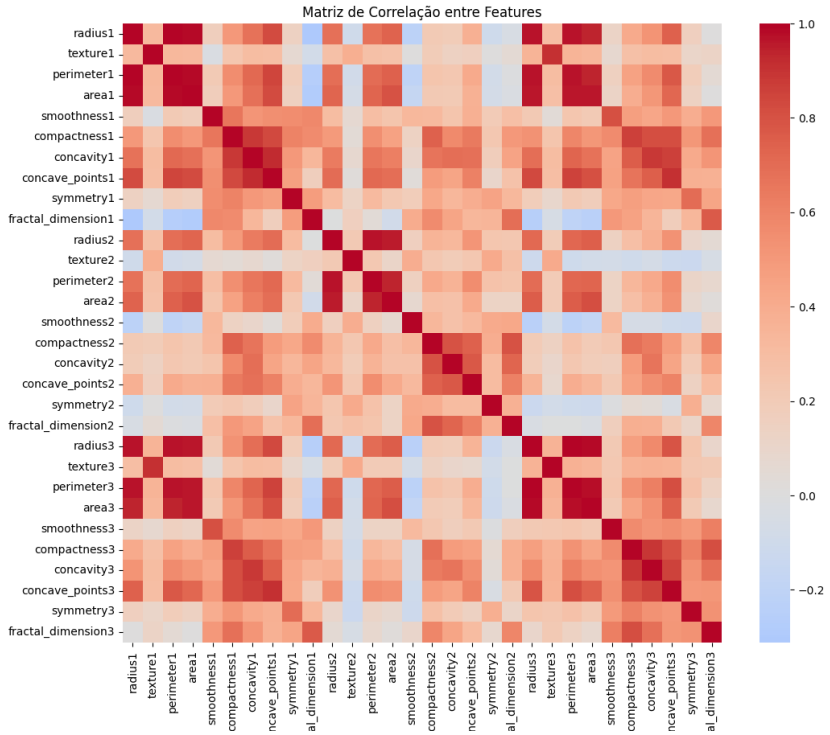


Figura 4. Matriz de correlação entre os 30 atributos do conjunto Breast Cancer Wisconsin.

- Foram gerados boxplots de atributos representativos, como `concave_points`, `perimeter` e `radius`, destacando a separação entre classes.

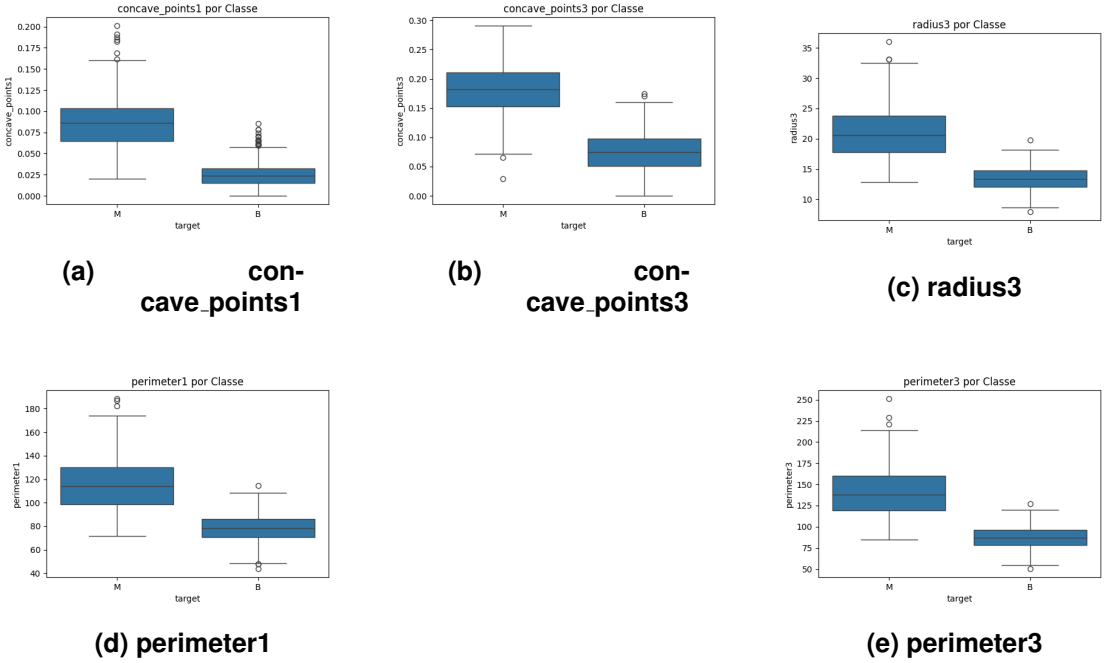


Figura 5. Boxplots de atributos selecionados por classe (M: Maligno, B: Benigno). Observa-se boa separação, principalmente em `concave_points` e `perimeter`.

3.3.3. Árvore de Decisão: Treinamento Inicial

- Foi utilizado o algoritmo `DecisionTreeClassifier` da biblioteca scikit-learn [Pedregosa et al. 2011] como modelo preditivo.
- O modelo foi treinado inicialmente com os hiperparâmetros padrão sobre o conjunto de treino.
- A avaliação inicial foi feita com base no conjunto de teste, utilizando acurácia, matriz de confusão e relatório de classificação.

3.3.4. Cross Validation

- Para estimar a capacidade de generalização do modelo, foi aplicado *cross validation* com k=10.
- O modelo foi avaliado por meio da acurácia média e do desvio padrão ao longo das 10 iterações, permitindo uma análise mais robusta do seu desempenho.

3.3.5. Otimização de Hiperparâmetros

- Foi realizada uma busca em grade (*GridSearchCV*) para ajuste dos hiperparâmetros.
- Foram testadas diferentes combinações de profundidade máxima da árvore (`max_depth`) e critério de impureza para divisão dos nós (`gini` e `entropy`).

- Também foi realizado o cross validation durante a execução do Grid Search, utilizamos k=10 novamente.

3.4. Métricas e Resultados

A avaliação inicial do modelo de árvore de decisão foi realizada utilizando um conjunto de teste correspondente a 30% das amostras. Os resultados obtidos mostraram bom desempenho, com acurácia satisfatória na distinção entre casos benignos e malignos.

Com o intuito de melhorar a performance do classificador, foi aplicada uma busca em grade (*GridSearchCV*) para otimização dos hiperparâmetros. A busca concentrou-se em duas variáveis principais:

- **max_depth**: profundidade máxima da árvore;
- **criterion**: critério de impureza para divisão dos nós, entre *gini* e *entropy*.

Durante a busca, foi utilizado cross validation com 10 divisões (*k=10*) para garantir estimativas mais confiáveis da acurácia média e da variabilidade do modelo em diferentes subconjuntos dos dados. A Tabela 2 apresenta os resultados obtidos.

max_depth	critério	Acurácia Média	Desvio Padrão	Ranking
4	entropy	0.9447	0.0292	1
3	entropy	0.9371	0.0345	2
5	gini	0.9321	0.0374	3
4	gini	0.9295	0.0336	4
–	entropy	0.9271	0.0136	5

Tabela 2. Resultados da validação cruzada com GridSearchCV para o conjunto Breast Cancer Wisconsin.

O melhor desempenho foi obtido com a configuração **max_depth = 4** e **criterion = entropy**, atingindo uma acurácia média de 94,5% e um desvio padrão de aproximadamente 2,9%. A Figura 6 mostra a representação gráfica da árvore de decisão gerada com esses hiperparâmetros.

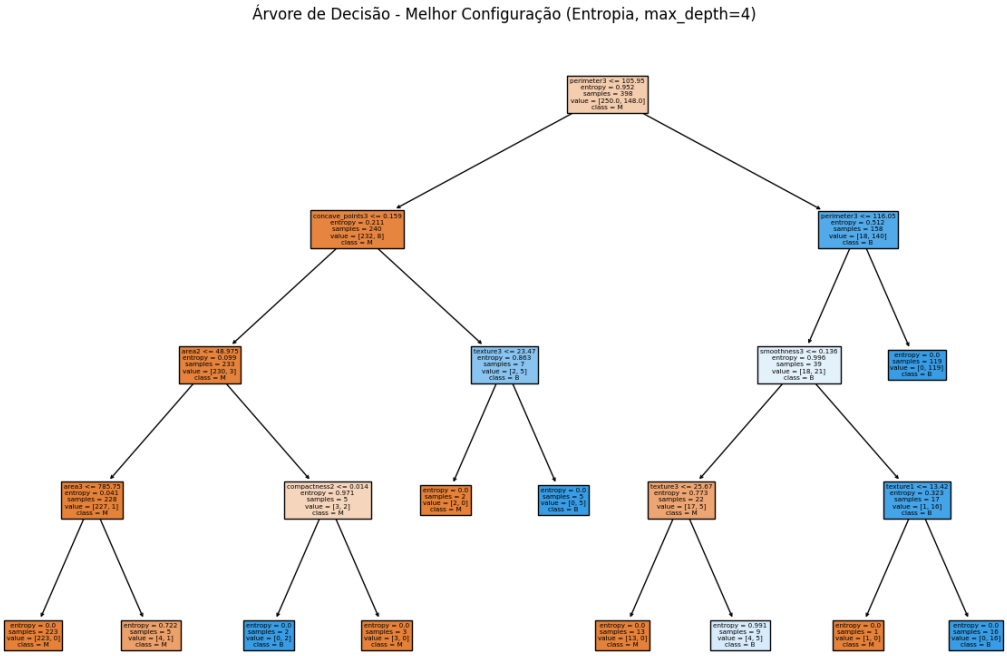


Figura 6. Árvore de decisão com os melhores hiperparâmetros (max_depth = 4, criterion = entropy) para o conjunto Breast Cancer.

Observando a árvore resultante, nota-se que os atributos relacionados a *concave_points* e *perimeter* foram os mais utilizados para as divisões internas, o que está de acordo com a análise exploratória inicial. Essas variáveis apresentaram maior separabilidade entre as classes, como pode ser observado nos boxplots incluídos na Seção de Análise Exploratória (Figuras 5), reforçando sua importância para a construção da árvore ao classificar.

Referências

Amedee, R. G. and Dhurandhar, N. R. (2001). Fine-needle aspiration biopsy. *The Laryngoscope*, 111(9):1551–1557.

Anderson, E. (1928). The problem of species in the northern blue flags, iris versicolor l. and iris virginica l. *Annals of the Missouri Botanical Garden*, 15(3):241–332.

Fisher, R. A. (1936). Iris. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C56C76.

Mitchell, T. M. and Mitchell, T. M. (1997). *Machine learning*, volume 1. McGraw-hill New York.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Street, W. N., Wolberg, W. H., and Mangasarian, O. L. (1993). Nuclear feature extraction for breast tumor diagnosis. In *Electronic imaging*.

Unwin, A. and Kleinman, K. (2021). The iris data set: In search of the source of virginica. *Significance*, 18(6):26–29.

Wolberg, William, M. O. S. N. and Street, W. (1993). Breast Cancer Wisconsin (Diagnostic). UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5DW2B.