UFC

UNIVERSIDADE FEDERAL DO CEARÁ
CAMPUS JARDINS DE ANITA - ITAPAJÉ

Análise de Dados e Reconhecimento de Padrões: Um estudo Com Iris Dataset e Qualidade do Vinho Dataset

Autores: Giovanna Dias Castro de Oliveira
Jônatas Fernandes Silva
Leandro Nascimento Adegas

Disciplina: Aprendizado de Máquina e Reconhecimento de Padrões

Itapajé, Maio, 2025

Conteúdo

1 Resumo

Este trabalho trata da análise de dois datasets, a saber, dataset Iris e um *dataset* de qualidade do vinho. Análise das features de cada dataset. Uso de dois classificadores para realizar a análise, vizinhos mais próximos (*(k-nearest neighbors, KNN*) e máquina de vetores de suporte (*suport vector machine, SVM*) e *Random Forest* (RF). A contribuição principal do estudo reside na justificativa que será apresentada em relação a escolha dos classificadores.

2 Introdução

O conjunto de dados Iris é um dos conjuntos de dados mais populares em ciência de dados e aprendizado de máquina. Ele contém informações sobre 150 flores de íris, divididas em três espécies diferentes: Iris setosa, Iris versicolor e Iris virginica. Este conjunto de dados é frequentemente usado como um exemplo para demonstrar técnicas de análise de dados e classificação. Em resumo, o conjunto de dados Iris é uma ferramenta valiosa para a comunidade de ciência de dados e aprendizado de máquina. Ele nos permite explorar técnicas de análise de dados e classificação em um conjunto de dados bem compreendido e interessante [1].

O *dataset* Qualidade do Vinho está dividido em dois tipos de vinho. Os dois conjuntos de dados estão relacionados com as variantes vermelha e branca do vinho português "Vinho Verde". Para mais detalhes, consulte a referência [2]

Para realizar a análise dos dois conjuntos de dados e extrair a maior quantidade de informações pertinentes possíveis foram utilizados dois classificadores, vizinhos mais próximos (*(k-nearest neighbors, KNN*) e máquina de vetores de suporte (*suport vector machine, SVM*) e *Random Forest* (RF).

Os conjuntos de dados foram disponibilizados pelo professor Dr. Júlio César Santos dos Anjos. O mesmo também disponibilizou as explicações necessárias para a realização desta tarefa.

A organização deste trabalho segue a seguinte ordem. A seção 3 contém informações a respeit das features dos dois *datasets* utilizados e alguns trabalhos relacionados. A seção 4 explica de maneira breve o funcionamento dos classificadores KNN e SVM e a justificativa por trás da escolha destes modelos de classificação. A seção 5 explica o ambiente e os métodos utilizados para as demonstrações empíricas deste trabalho. A seção 6 traz os resultados obtidos após a análise dos *datasets*. Por fim, a seção 6 traz a conclusão deste trabalho.

3 Background e Trabalhos Relacionados

3.1 Backgroud

3.1.1 Dataset Iris

O dataset Iris é composto por seis (06) features, a saber:

1. Id;
2. SepalLengthCm;
3. SepalWidthCm;
4. PetalLengthCm;
5. PetalWidthCm; e
6. Species

Feature Id A feature **Id** representa apenas a identificação de cada flor dentro do dataset, ou seja, ele é apenas um indicador para guiar o analista. Dessa forma, vem facilitar quanto a identificação de qual flor está sendo observada.

Feature SepalLengthCm A feature **SepalLengthCm** representa o comprimento em centímetros das sépalas que compõem cada flor. A saber, sépalas são folhas modificadas, geralmente de coloração verde, localizadas abaixo das pétalas e com função de proteger o botão floral enquanto ainda não abriu [3].

Feature SepalWidthCm A feature **SepalWidthCm** representa a largura em centímetros das sépalas que compõem cada flor.

Feature PetalLengthCm A feature **PetalLengthCm** representa o comprimento em centímetros das pétalas que compõem cada flor. A saber, pétalas são uma unidade da corola, pode possuir cores vivas e marcantes e sua função é atrair agentes polinizadores. Inserem-se logo após as sépalas. [3].

Feature PetalWidthCm A feature **PetalLengthCm** representa a largura em centímetros das pétalas que compõem cada flor.

Feature Species A feature **Species** representa a espécie que cada flor indicada por seu **Id** pertence.

3.1.2 Qualidade do Vinho Dataset

Features Vinho Branco e Vermelho O Dataset Qualidade do Vinho é composto por doze (12) features, a saber:

1. fixed acidity: nível de acidez fixa;
2. volatile acidity: nível de acidez volátil;
3. citric acid: nível de ácido cítrico;
4. residual sugar: nível de de açucar residual;
5. chlorides: nível de cloreto;
6. free sulfur dioxide: nível de dióxido de enxofre livre;
7. total sulfur dioxide: nível total de dóxido de enxofre;
8. density: densidade (base h2o -¿ 1.0);
9. pH: O pH é uma escala numérica que determina o grau de acidez de uma solução aquosa, baseado na concentração de íons hidrônio (H3O+). Soluções ácidas possuem excesso de íons hidrônio e pH menor do que 7. Soluções básicas possuem excesso de íons hidroxila (OH-) e valores de pH superiores a 7. Soluções consideradas neutras têm igual concentração de íons H3O+ e íons OH-, e sua medida de pH é 7. [4];
10. sulphates: nível de sulfatos, a saber: Os sulfatos são compostos iônicos que contêm o ânion SO42-, que é chamado de ânion sulfato. [5];
11. alcohol: nível de álcool presente no vinho; e
12. quality (score between 0 and 10): índice de qualidade do vinho

3.2 Trabalhos Relacionados

3.2.1 Trabalhos Relacionados ao Dataset Qualidade do Vinho

Modeling wine preferences by data mining from physicochemical properties [6]. Este estudo apresenta uma análise detalhada de dois conjuntos de dados relacionados ao vinho "Vinho Verde" (variantes tinto e branco). O objetivo é prever a qualidade do vinho com base em 11 atributos físico-químicos, utilizando técnicas como regressão, árvores de decisão, redes neurais e máquinas de vetor de suporte (SVM). Os autores concluíram que os métodos baseados em árvore e SVM apresentaram desempenho superior na predição da qualidade dos vinhos. O dataset foi posteriormente disponibilizado publicamente e tornou-se referência em benchmarks de classificação.

Comparison of classification techniques on wine dataset [7]. Este trabalho compara diversas técnicas de classificação aplicadas ao dataset de qualidade do vinho. Os algoritmos avaliados incluíram K-Nearest Neighbors (KNN), Support Vector Machines (SVM), Naive Bayes e Árvores de Decisão. O foco foi medir a acurácia, tempo de execução e robustez de cada algoritmo. Os autores identificaram que KNN e SVM apresentaram resultados mais consistentes em termos de desempenho e precisão, especialmente após ajustes de parâmetros e normalização dos dados.

Machine Learning in Action [8]. Este livro apresenta uma abordagem prática para o aprendizado de máquina, explicando os fundamentos teóricos e demonstrando a aplicação de algoritmos como KNN e SVM em Python. O autor utiliza exemplos com datasets reais, como o Iris e o de vinhos, e fornece códigos completos para ajudar na implementação. É uma excelente referência para quem deseja entender, codificar e comparar classificadores na prática.

Scikit-learn: Machine Learning in Python [9]. Este artigo apresenta a biblioteca Scikit-learn, uma das mais populares para aprendizado de máquina em Python. A biblioteca fornece implementações eficientes de diversos algoritmos, incluindo KNN, SVM e Decision Trees. Os autores destacam a modularidade, a facilidade de uso e o suporte a avaliação de modelos. A Scikit-learn é amplamente utilizada para pesquisas aplicadas com os datasets Iris e Vinho, sendo um padrão de fato para experimentos com classificadores supervisionados.

4 Modelo de Solução

4.1 Classificador KNN

Este classificador é conhecido como vizinhos mais próximo, isto é, ele utiliza as características de um centróide como base para identificar qual ponto pertence a um certo grupo [10], Figura 1.

O KNN utiliza o seguinte algoritmo:

1. Guarda em memória **todo** o conjunto X, no formato adequado;
2. Para cada novo vetor de atributos $\mathbf{x}_{new}$ sem classificação, realiza uma busca em X pelo índice do vetor de atributos mais próximo de $\mathbf{x}_{new}$:

$$i^* = \arg \min_{i=1,\ldots,N} dist(\mathbf{x}_{new}, \mathbf{x}_i)$$

   em que $dist(\mathbf{x}_{new}, \mathbf{x}_i)$ é uma função que mede a distância entre os dois vetores $\mathbf{x}_{new}$ e $\mathbf{x}_i$;
3. Atribuir ao vetor $\mathbf{x}_{new}$ à mesma classe que $\mathbf{x}_{i*}$.

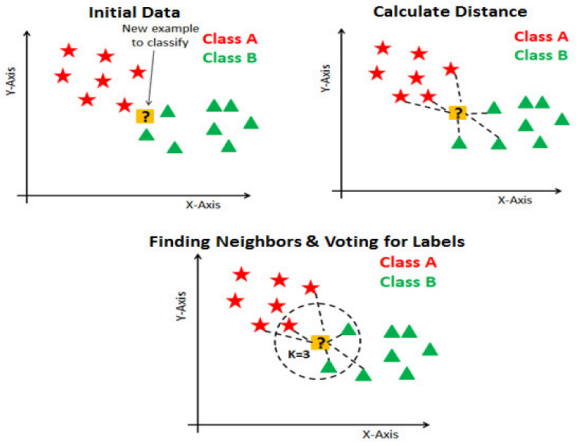

Figura 1: Classificação KNN
Source: Machine Learning, Prof. Dr. Júlio César, [10]

4.2 Classificador SVM

A técnica de classificação SVM é uma técnica mais moderna. Semelhante aos *perceptrons*. O SVM surge como uma melhoria do *perceptron*. O SVM seleciona um hiperplano, não somente separa os pontos em duas classes, ele maximiza as margens. A equação (1), exemplifica o a fórmula que o SVM usa para isso [11].

$$SVM = w \cdot x + b = 0 \tag{1}$$

Observando a equação (1), nota-se que, é uma equação de 1º grau, ou seja, seu resultado é uma reta, como demonstrado na Figura 2.

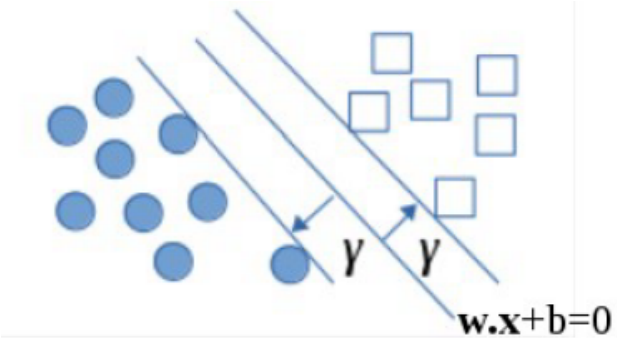

Figura 2: SVM seleciona maximizando a distância $\gamma entreos hiperplanos$
Source: Algoritmos e técnicas de classificação, Prof. Dr. Júlio César, [11]

Para este trabalho ele foi utilizado de forma estentida no (SVM estendido), veja [12].

4.3 Random Forest

O classificador RF, também conhecido como floresta aleatória, é um modelo composto por várias árvores de decisão. Cada árvore é treinada individualmente e no fim o algoritmo faz uma soma das decisões, tomando-as como base para a classificação.

5 Metodologia

Nesta seção será apresentada os requisitos necessários para a avaliação do modelo de solução.

5.1 Ambiente

Os códigos e testes foram respectivamente, escritos e realizados no *google colab*. A linguagem de programação utilizada foi Python em sua versão mais atualizada (*latest*). Os datasets utilizados foram o dataset Iris, disponível em: `https://www.kaggle.com/datasets/uciml/iris`, e o dataset Qualidade do vinho, disponível em: `https://www.kaggle.com/datasets/uciml/red-wine-quality-cortez-et-al-2009`. Os testes foram realizados 10 (dez) vezes, para garantir uma melhor qualidade dos resultados.

Os códigos utilizados estão disponíveis em: `https://colab.research.google.com/drive/1Tn0eRmwhVgR`

5.2 Preparação dos Dados

Para preparar os dados para a realização dos testes foram realizadas algumas etapas, descritas a seguir. Após carregar o dataset foi feito o mapeamento da feature *'species'* para nome. Depois verificou-se o gráfico da dispersão dos dados, em que sua dimensão é d^3, como mostra a Figura 3.

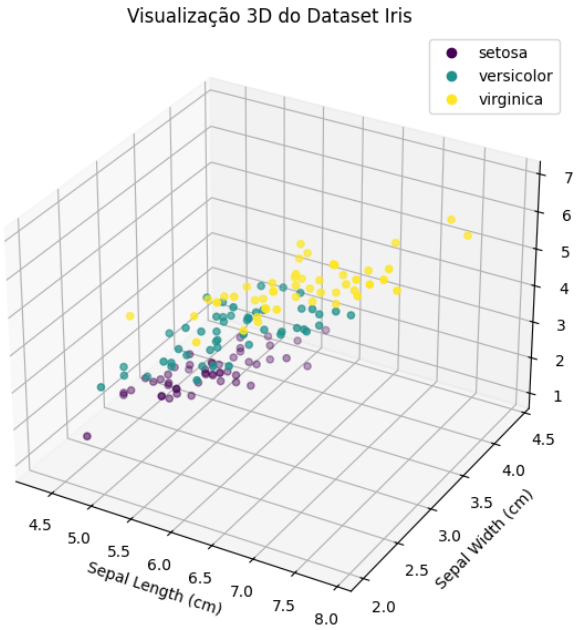

Figura 3: Dispersão dos Dados Dataset Iris.

Após esta etapa, foi realizada a redução de d^3 para d^2, utilizando da técnica de análise dos componentes principais (*Principal Components Analysis*, PCA).

5.3 Classificação

Concluída a etapa de preparação dos dados, começou-se a classificação. Como já citado anteriormente, foram utilizados três classificadores, KNN, SVM e RF.

Para iniciar a classificação primeiramente o dataset foi dividido em **X_train, X_test, y_train, y_test)**. Onde 70% dos dados foram utilizados para treinamento e 30% para validação, este método é conhecido como *data splitting* [13]. E logo após isto, o KNN, SVM e o RF foram treinados.

Em seguida foi realizada a avaliação do modelo por meio das seguintes métricas: acurácia, precisão, recall e f1-score. Em seguida foi feita a validação cruzada para os dois classificadores Por fim foi gerada a matriz de confusão para os dois classificadores e o gráfico de fronteira.

6 Resultados

6.1 Iris Dataset

Para o conjunto de dados Iris foram registrados os seguintes resultados.:

A Figura 4 mostra que a espécie 'Setosa' está bem mais definida do que as outras, sendo assim, temos que classificar as outras espécies para separarmos as duas espécies, para isso, usamos os classificadores KNN e SVM.

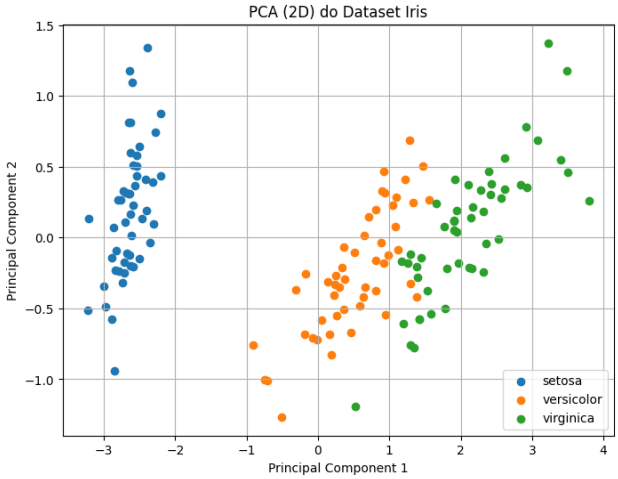

Figura 4: PCA.

6.1.1 Métricas de Avaliação

As métricas de avaliação demonstradas nas Tabelas 1, 2 e 3, nos mostram que os modelos KNN e RF de classificação escolhidos acertaram todas as classificações feitas no *dataset*, já o SVM estendido também acertou quase todas, ou seja, o modelo pode ser classificado como ótimo.

Acurácia	Precisão	Recall	F1-Score
100%	100%	100%	100%

Tabela 1: Métricas de Avaliação KNN.

Acurácia	Precisão	Recall	F1-Score
91%	100%	100%	100%

Tabela 2: Métricas de Avaliação SVM.

Acurácia	Precisão	Recall	F1-Score
100%	100%	100%	100%

Tabela 3: Métricas de Avaliação RF.

6.1.2 Validação Cruzada

A Tabela 4 demonstra que, apesar dos números ótimos demonstrados na Tabela 1 e na Tabela 2, a acurácia real esperada nos modelos é um pouco menor. Sendo a acurácia esperada para SVM igual a 0.98, para KNN igual a 0.97 e para RF 0.96.

Acurácia Média SVM	Acurácia Média KNN	Acurácia Média RF
0.98	0.97	0.96

Tabela 4: Validação Cruzada.

6.1.3 Matriz de Confusão e Fronteira de Decisão

As matrizes mostradas nas Figuras 5 e 6 nos relatam que nenhum erro de classificação ocorreu, ou seja, os dois classificadores tem 100% de acerto. Cada classe (setosa, versicolor, virginica) foi perfeitamente classificada por todos os modelos nos dados de teste.

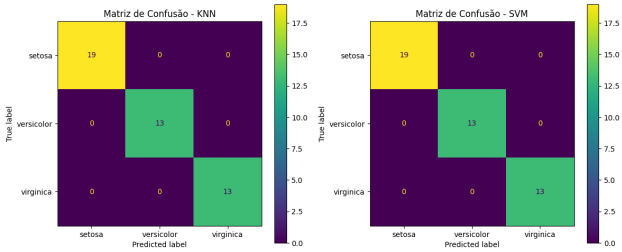

Figura 5: Matriz de Confusão KNN e SVM.

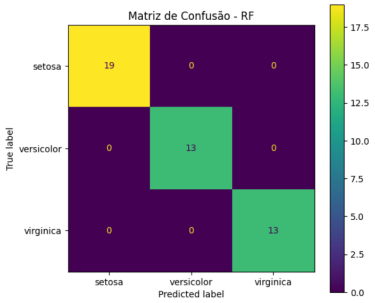

Figura 6: Matriz de Confusão RF.

Devido a classificação do KNN basear-se nos vizinhos mais próximos, as fronteiras mostradas na Figura 7 são não lineares e com muitos "recortes". As regiões são mais "quebradas", especialmente na área entre `versicolor` e `virginica`, o que pode indicar sensibilidade a ruídos ou outliers.

Já no SVM, a fronteira mostrada na Figura 7 é mais suave e elíptica, indicando uma separação mais generalizada das classes.

6.2 Qualidade do Vinho Dataset

Para o dataset qualidade do vinho foram usados 03 (três) classificadores, KNN, SVM estendido e RF.

Na Figura 8, é representada a redução de dimensionalidade já citada, de d^3 para d^2. Podemos notar que as classes estão bem misturadas, sendo de díficil separação.

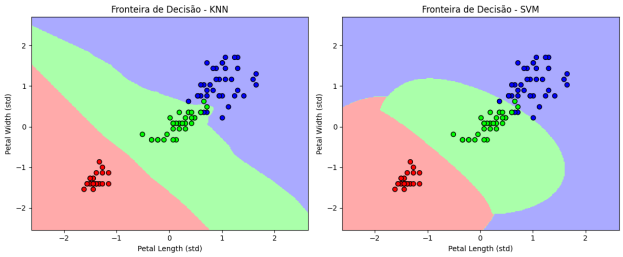

Figura 7: Fronteira de Decisão KNN e SVM

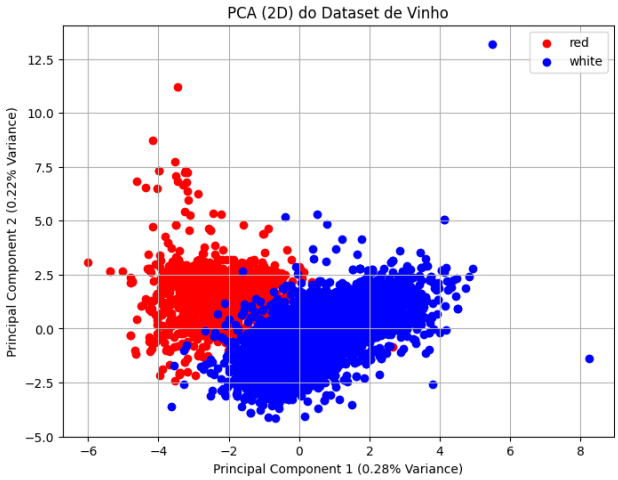

Figura 8: PCA.

6.2.1 Matriz de Confusão

Na Figura 9, é possível identificar que, RF apresenta o melhor desempenho neste cenário, com maior número de predições corretas e menos erros. KNN teve o pior desempenho, com mais erros tanto em falsos positivos quanto em falsos negativos. SVM está no meio-termo, melhor que KNN, mas inferior ao RF.

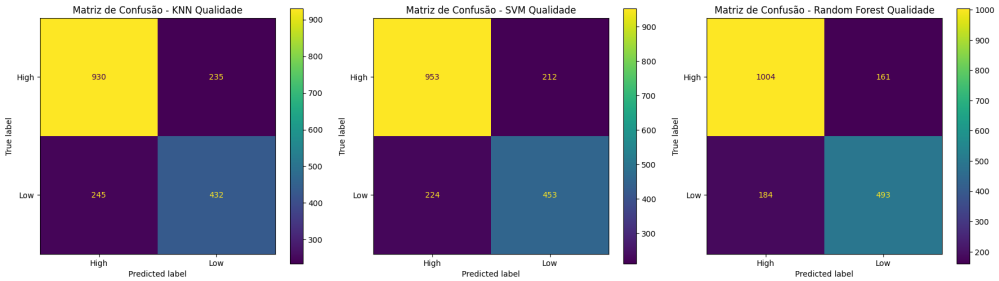

Figura 9: Matriz de Confusão KNN, SVM e RF.

7 Conclusão

A justificativa por trás da escolha dos classificadores KNN e SVM, se dá por conta da dimensão d^2. Como no PCA reduzimos a dimensão de d^3 para d^2, podemos trabalhar com distância euclidiana, logo podemos usar KNN e SVM. Isto para o primeiro conjunto de dados, o Iris. O SVM não usado

em sua forma usual, mas de forma estendida, isto é, ele é usado para classificação binária, mas no caso do conjunto de dados Iris, possuímos 03 (três) tipos de flores, então a forma estendida seria o idea, já que assim ele pode ser usado para multiclasses.

No dataset qualidade do vinho, além do uso do KNN e do SVM, foi utilizado o classificador RF. Porque o SVM já não era ideal neste caso, ele foi utilizado apenas para fins de comparação, mesmo utilizando de SVM estendido. O RF por outro lado, parecia ser ideal, por utilizar as árvores de decisão tanto nos *red wines* quanto nos *white wines*.

8 Agradecimentos

Agradecemos primeiramente a Deus pela inspiração e guia, ao professor Júlio Césas Santos dos Anjos pela orientação e dedicação, e aos valorosos colegas pela parceria e contribuições.

9 Referências

[1] "O conjunto de dados Íris é um dos conjuntos de dados mais populares em ciência de dados e aprendizado de máquina.." dio.me/articles/o-conjunto-de-dados-iris-e-um-dos-conjuntos-de-dados-mais-populares-em-ciencia-de dados-e-aprendizado-de-maquina. Online; Acesso em 2025-06-03.

[2] P. Cortez, A. Cerdeira, F. Almeida, T. Matos, and J. Reis, "Modeling wine preferences by data mining from physicochemical properties," *Viticulture Commission of the Vinho Verde region (CVRVV), 4050-501 Porto, Portugal*, 2009.

[3] "Partes da flor: verticilos florais, pedúnculo e pistilo." https://querobolsa.com.br/enem/biologia/flor-partes-da-planta. Online; Acesso em 2025-06-03.

[4] "O que é ph?." https://mundoeducacao.uol.com.br/quimica/voce-sabe-que-significa-ph-.htm. Online; Acesso em 2025-06-03.

[5] "Sulfatos." https://brasilescola.uol.com.br/quimica/sulfatos.htm. Online; Acesso em 2025-06-03.

[6] P. Cortez, A. Cerdeira, F. Almeida, T. Matos, and J. Reis, "Modeling wine preferences by data mining from physicochemical properties," *Decision Support Systems*, vol. 47, no. 4, pp. 547–553, 2009.

[7] M. A. Tayel, M. A. Elbehery, and W. M. Sheta, "Comparison of classification techniques on wine dataset," in *International Conference on Artificial Intelligence and Computer Science*, (Kuala Lumpur, Malaysia), pp. 89–94, 2014.

[8] P. Harrington, *Machine Learning in Action*. Shelter Island, NY: Manning Publications Co., 2012.

[9] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, *et al.*, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[10] J. C. S. dos Anjos, "Machine learning." Aula ou apresentação acadêmica, 2025.

[11] J. C. S. dos Anjos, "Algoritmos e técnicas de classificação." Aula ou apresentação acadêmica, 2025.

[12] "Classificação com máquina de vetores de suporte (svm)." https://analisemacro.com.br/econometria-e-machine-learning/classificacao-com-maquina-de-vetores-de-suporte-svm/. Online; em 2025 - 06 - 17.

[13] R. Izbick and T. M. dos Santos, *Aprendizado de Máquina: Uma Abordagem Estatística*. Livro Eletrônico, 2020.