Classificando frutas usando Aprendizado de máquina

Árvores de decisão

Henrique Benatti - 2025

1 - Introdução

1.1 - Objetivo

O objetivo deste trabalho é realizar a construção e análise de um código de aprendizado de máquina treinado para identificar frutas com base em algumas características fornecidas pela base de dados. Para esse problema utilizou-se um algoritmo de árvore de decisão que é uma abordagem de aprendizado supervisionado usado em estatística, mineração de dados e aprendizado de máquina, para resolver problemas de classificação e regressão como um modelo preditivo que tira conclusões sobre um conjunto de dados ou observações.

Neste trabalho, pela sua natureza expositiva e didática, foi utilizado um conjunto pequeno de dados no formato xlsx, o que corresponde a uma planilha de excel 8x5.

A análise consiste na introdução dos fundamentos teóricos do aprendizado de máquina e do algoritmo de árvore de decisão, no estudo das diferentes partes do código e de sua estrutura, e por fim na interpretação dos resultados obtidos.

1.2 - Conceitos teóricos fundamentais

Com o avanço dos problemas científicos, sociais e tecnológicos, junto de uma crescente complexidade relacionada a esses sistemas, surge a necessidade do desenvolvimento de maneiras mais eficientes para processar um volume cada vez maior de dados, essa necessidade é evidenciada nos dados coletados em experimentos como os do LHC onde para se obter boas chances de encontrar novos fenômenos e partículas pode ser necessário gerar até 300 gigabytes de dados por segundo. Problemas semelhantes surgem em outras áreas como em diagnósticos médicos, processamento de linguagem natural e astrofísica de altas energias.

É impossível para um pesquisador analisar um volume tão grande de dados sem o auxílio de ferramentas com sofisticação capazes de processar e analisar dados, para esse problemas surge uma alternativa de mitigação, que vem de um

subcampo da ciência da computação ou Inteligência artificial, conhecido como aprendizado de máquina.

1.2.1 Aprendizado de Máquina (Machine Learning)

O aprendizado de máquina é um subcampo da inteligência artificial, tem como objetivo fazer com que o computador imite a capacidade humana de aprender e realizar tarefas autônomas por meio da experiência e da exposição a um conjunto de dados.

Esse campo de estudo explora o desenvolvimento de algoritmos que podem aprender e melhorar sem serem explicitamente programados.

Dos diferentes tipos de algoritmos de aprendizado de máquina existentes usaremos o de árvore de decisão.

1.2.2 Árvores de Decisão

Árvore de decisão é o modelo preditivo que iremos utilizar, ele servirá para tirar conclusões a partir de um conjunto de dados. Esse tipo de algoritmo é bastante inteligível pois apresenta uma estrutura de fácil visualização, o que permite entender sua lógica de funcionamento.

2 - Análise do código

2.1 - Dados

Utilizamos uma pequena base de dados organizada em uma planilha de excel que contém quatro características associadas aos diferentes tipos de frutas.

Arredondada	Suculenta	Vermelha	Doce	Fruta
0	1	1	1	Morango
1	0	0	0	Limão
1	1	0	1	Pera
0	0	0	1	Banana
1	1	1	1	Cereja
1	1	1	0	Tomate
1	1	1	1	Maçã

Cada característica possui um peso, um valor que pode ser 0 ou 1, esse valor indica a característica que cada fruta não possui ou possui respectivamente, sendo

0 representando a inexistência e 1 representando a existência da característica em questão.

Esses dados são guardados em uma pasta que será organizada em um repositório para facilitar a construção do código e o acesso aos dados.

2.2 Organização do repositório

A organização do repositório é importante pois facilita o acesso do código às pastas que possuem os dados que iremos utilizar para treinar nosso modelo.

Organizamos o repositório da seguinte forma:

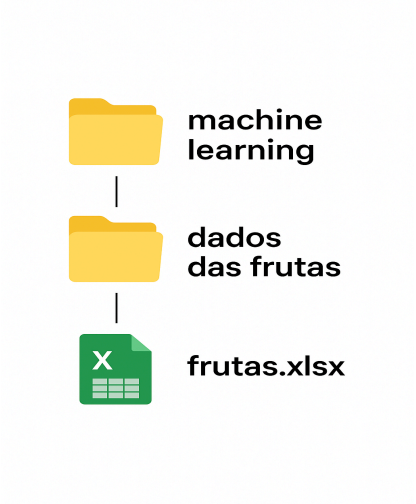


Figura 1: Repositório.

Nosso código será construído dentro da pasta ‘machine learning’.

2.3 Código

Para que possamos realizar uma análise pragmática e clara do código, foi feita a separação do mesmo em blocos de código menores, dessa forma podemos analisar cada bloco separadamente, não foi realizada a criação de nenhuma função o que permite a quebra do programa em blocos de código e sua análise em um fluxo linear. Utilizamos o Visual Studio Code como ambiente de codificação.

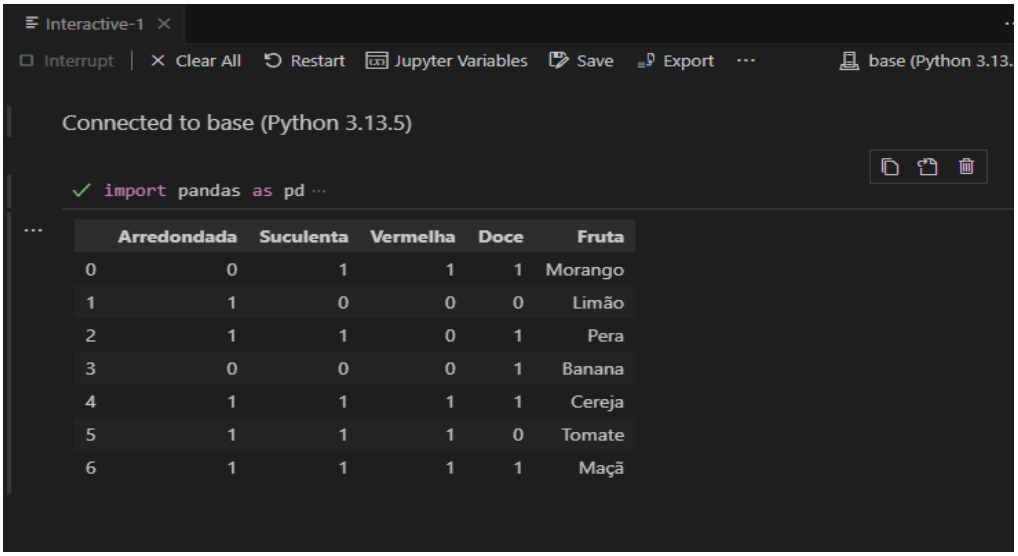
2.3.1 Bloco 1: Primeiros Comandos

Neste primeiro bloco importamos a biblioteca pandas utilizada neste primeiro pedaço de código para criar um data frame. Para a criação do data frame foi necessário especificar o formato dos dados utilizados e o caminho do arquivo em que os dados estão salvos, por causa da disposição das pastas no repositório fica mais fácil especificar o caminho até os dados.

```python
# %%
import pandas as pd
# escolhendo e observando os dados
df = pd.read_excel("data/dados_frutas.xlsx")
df
```

Figura 2: Escolha e observação dos dados utilizados.

Nosso código também está sendo executado em notebooks do ambiente mútuo criado pelo jupyter notebooks, isso facilita na observação e execução do programa.



Connected to base (Python 3.13.5)

```python
import pandas as pd
```

	Arredondada	Suculenta	Vermelha	Doce	Fruta
0	0	1	1	1	Morango
1	1	0	0	0	Limão
2	1	1	0	1	Pera
3	0	0	0	1	Banana
4	1	1	1	1	Cereja
5	1	1	1	0	Tomate
6	1	1	1	1	Maçã

Figura 3: Observando o data frame com os dados.

2.3.2 Bloco 2: Escolha do Algoritmo

Nesse segundo bloco escolhemos o tipo de algoritmo de aprendizado de máquina que iremos utilizar, esse é também o modelo que será treinado por meio dos nossos dados. Utilizamos um algoritmo de árvore de decisão, que importamos usando a biblioteca sklearn.

```python
# %%
# escolhendo o algoritmo que será usado
from sklearn import tree
arvore = tree.DecisionTreeClassifier()
```

Figura 4: Definindo o algoritmo utilizado, árvore de decisão.

2.3.3 Bloco 3: Separando Variáveis

Nessa etapa separamos as variáveis em ‘variável resposta’ que são as frutas que temos na nossa base de dados e que representam as possibilidades de respostas que o modelo poderá no dá com base nas características ‘variáveis independentes’, após o treinamento.

```
12  # %%
13  # Variaveis
14  y = df['Fruta'] #variavel resposta
15  caracteristicas = ['Arredondada','Suculenta','Vermelha','Doce']
16  x = df[caracteristicas]
17  print(y)
18  print(x)
19
```

Figura 5: separação das variáveis.

Podemos observar que as colunas das características foram separadas da coluna das frutas:

```
✓ # Variaveis ···

0    Morango
1      Limão
2       Pera
3     Banana
4     Cereja
5     Tomate
6       Maçã
Name: Fruta, dtype: object
   Arredondada  Suculenta  Vermelha  Doce
0            0          1         1     1
1            1          0         0     0
2            1          1         0     1
3            0          0         0     1
4            1          1         1     1
5            1          1         1     0
6            1          1         1     1
```

Figura 6: conjunto ‘variável resposta’ e ‘variáveis independentes’.

2.3.4 Bloco 4: Treinando o Modelo

Após a separação das variáveis podemos treinar nosso modelo, etapa a qual a máquina realmente “aprende”, o algoritmo irá identificar os padrões existentes no nosso conjunto de dados separando cada característica por meio do estabelecimento de condições, esse processo é o que define a estrutura da *Árvore de Decisão*, que veremos após a análise do código.

A última linha de código desse bloco faz uma predição com base no treinamento realizado.

```
Run Cell | Run Above | Debug Cell
# %%
# Ensinar a maáquina
arvore.fit(x,y)
arvore.predict([[1,1,0,0]])

```

Figura 7: treinando modelo e realizando uma classificação.

Para realizar a predição, nós atribuímos uma sequência aleatória de características (1,1,0,0) e com base no treinamento do modelo, foi realizada uma classificação que relaciona a fruta compatível com as características dadas.

```
array(['Limão'], dtype=object)

Press Enter to execute.
```

Figura 8: resultado da predição com base nas características (1,1,0,0).

O resultado obtido da classificação foi ‘limão’, apesar das características dadas não serem exatamente as do limão, o algoritmo classifica com base na fruta mais provável. Esses detalhes serão observados e corrigidos em trabalhos de análises de algoritmos posteriores.

2.3.5 Bloco Final: Observando a Árvore e Probabilidade da Predição

Essas são as últimas linhas de código, nas quais iremos plotar o fluxograma da nossa árvore de decisão e entender mais claramente como o algoritmo funciona, além disso iremos também ver a probabilidade da classificação feita anteriormente.

```
Run Cell | Run Above | Debug Cell
# %%
# Visualizando a arvore
import matplotlib.pyplot as plt
plt.figure(dpi=400)
tree.plot_tree(arvore,
               feature_names = caracteristicas,
               class_names = arvore.classes_,
               filled = True )

# Probabilidades
proba = arvore.predict_proba([[1,1,0,0]]) [0]
pd.Series(proba, index = arvore.classes_)

```

Figura 9: código para visualizar a árvore e calcular a probabilidade.

As primeiras linhas importam a biblioteca de visualização e dá algumas especificações básicas para plotar a imagem da árvore.

As últimas linhas do código calcula a probabilidade da fruta com base nas características fornecidas para realizar a classificação, o valor é dado entre 0 e 1, e coloca as frutas em série para que a probabilidade de cada uma possa ser observada, essa é uma boa estratégia quando temos mais de uma fruta possível como resultado da classificação, o que não é o caso para o exemplo que utilizamos.

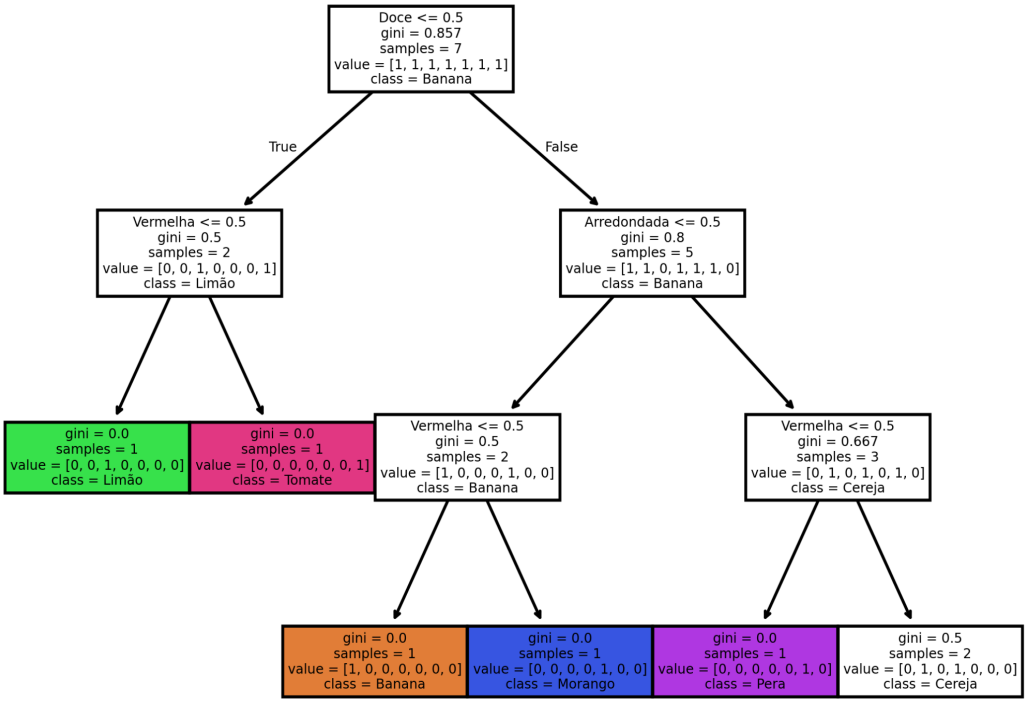


Figura 10: Estrutura da árvore de decisão.

Podemos observar por meio da estrutura da árvore a lógica de funcionamento do algoritmo, é possível perceber que o algoritmo escolhe uma característica como ponto de partida e usa uma condição para ramificar para outras características e assim sucessivamente, esse processo separa e encontra as frutas por meio dessa estrutura condicional. Como os valores de atributos das frutas é 0 ou 1, a condição estabelece que caso o valor seja menor ou igual a 0,5 então a característica tem valor igual a 0 e a condição é falsa e ramifica para outro atributo que estabelece outra condição caso o valor seja maior ou igual a 0,5 então a característica tem valor igual a 1 e a condição é verdadeira, o que igualmente cra outra ramificação.

Essa é a maneira a qual o modelo de árvore de decisão funciona, ela é particularmente fácil de compreender por conta da sua estrutura visual.

As probabilidades associadas a cada fruta, referentes a compatibilidade com as características especificadas na predição que fizemos, varia de entre 0 e 1. Podemos visualizá-las:

Figura 11: Probabilidades de 0 a 1 de cada fruta ser a compatível.

Isso implica que entre todas as frutas o limão é a fruta compatível.

3 - Resultados

Como o objetivo da análise foi introduzir conceitos e códigos básicos para começar a construção de sistemas baseados em aprendizado de máquina, omitimos alguns detalhes que são importantes em sistemas mais sofisticados, além de não cobrir uma variedade maior de resultados que o leitor poderá explorar variando as características das frutas ou aumentando a base de dados, no entanto todas as etapas do código foram executadas corretamente, desde a visualização dos dados até o treinamento do modelo e a verificação de um exemplo de classificação.

Particularmente sobre a classificação realizada alguns pontos são importante ressaltar, o modelo nos deu como resultado a fruta ‘limao’, com base nos dados ela era a mais provável e as probabilidade atribuída a ela foi de 100%, vale observar que esta probabilidade é dada em relação às outras frutas, então apesar de as características não terem sido exatamente as do limão o que impede uma probabilidade de 100%, em relação às outras frutas o limão era a escolha óbvia, dessa forma o algoritmo foi eficiente.

Existem métodos mais adequados para avaliar os resultados dos modelos, no entanto não serão explorados no presente trabalho mas sim em análises e projetos posteriores, aqui é importante perceber e compreender o funcionamento da estrutura do algoritmo e do código.

4- Conclusão

Foi realizada a análise de um código de machine learning que utilizou um modelo de árvore de decisão para classificar frutas, realizamos uma pequena introdução e comentamos brevemente sobre alguns conceitos importantes, o código foi separado em blocos para facilitar a compreensão da estrutura do programa, também visualizamos a estrutura do algoritmo e discutimos seu funcionamento. É importante perceber que não cobrimos boas práticas de codificação, aprofundamento teórico e nem outras técnicas essenciais para se construir programas de aprendizado de máquina, mas criamos um guia prático para as primeiras linhas de código que irá servir como base para códigos futuros.

Referências

- https://pt.wikipedia.org/wiki/Aprendizado_de_máquina

- https://www.ibm.com/br-pt/think/topics/machine-learning

- Aurélien Géron, Mãos à obra: Aprendizado de Máquina com Scikit-Learn, Keras e TensorFlow.