



escola
britânica de
artes criativas
& tecnologia

Profissão: Analista de dados



APRENDIZADO DE MÁQUINA – CLASSIFICAÇÃO



GUIA DA AULA 1



Classifique

● **Motivação**

● **Árvore de
decisão**

● **Pacote
Scikit-Learn**



Acompanhe aqui
os temas que
serão tratados
na videoaula



Motivação

Dado a largura e o comprimento das pétalas e sépalas de uma flor do gênero *iris*, qual é a sua espécie:



Iris Versicolor



Iris Setosa



Iris Virginica



Queremos um conjunto de regras que representem essa relação. Uma possível solução seria o conjunto de condições if-else como no exemplo a seguir:

```

def f(petal_length: float, petal_width: float, sepal_length: float, sepal_width: float) ->
str:
    if sepal_width > 5.0:
        if petal_width > 2.0:
            return 'versicolor'
        else:
            return 'virginica'
        else:
            return 'setosa'
  
```



Este conjunto de regras pode ser representado graficamente por uma árvore de decisão, no qual as folhas representam as classes do atributo categórico ou variável resposta a ser predito e os nós, as regras de decisão.

Qual o melhor conjunto de regras (atributos e valores de corte) para esse conjunto de dados?



Árvore de decisão

A árvore de decisão é uma abordagem estatística que busca encontrar a relação entre um atributo categórico alvo y (variável resposta) e um conjunto de atributos preditores x_i através de um conjunto de regras simples que, quando combinadas, formam uma complexa classificação. De maneira geral, utilizamos métodos exaustivos (força bruta) para definir a quantidade de nós necessário para classificar as classes do atributo alvo. O racional por trás da construção de um nó vem do uso do conceito de **impureza** do nó.



Para cada atributo x_i :

para cada* valor w_i entre $\min(x_i)$ e $\max(x_i)$:

calcule a impureza do nó

Selecione x_i e w_i com a menor impureza

Crie um nó com x_i e w_i

Repita



As métricas de impureza mais utilizadas são **Gini** e **Entropia**. Uma das grandes vantagens das árvores de decisão é a sua capacidade de explicação da relação entre a variável resposta e os atributos preditores, uma vez que é possível visualizá-la. Outra vantagem é que a técnica dispensa o tratamento dos atributos preditores (normalização, padronização, *one-hot encoding* etc.) pois estes não são comparados entre si.



Pacote Scikit-Learn

Pacote Python para ciência de dados e *machine learning*. Possui diversos modelos para aprendizado supervisionado, não supervisionado, etc. além de métodos auxiliares.

A documentação pode ser encontrada no link <https://scikit-learn.org/stable/>.

Para a árvore de decisão, temos:

```
In [ ]: from sklearn.tree import DecisionTreeClassifier
```

```
In [ ]: model = DecisionTreeClassifier()
```

