Inteligência Artificial e Big Data
Aula 07

Prof. Me Daniel Vieira

SENAI

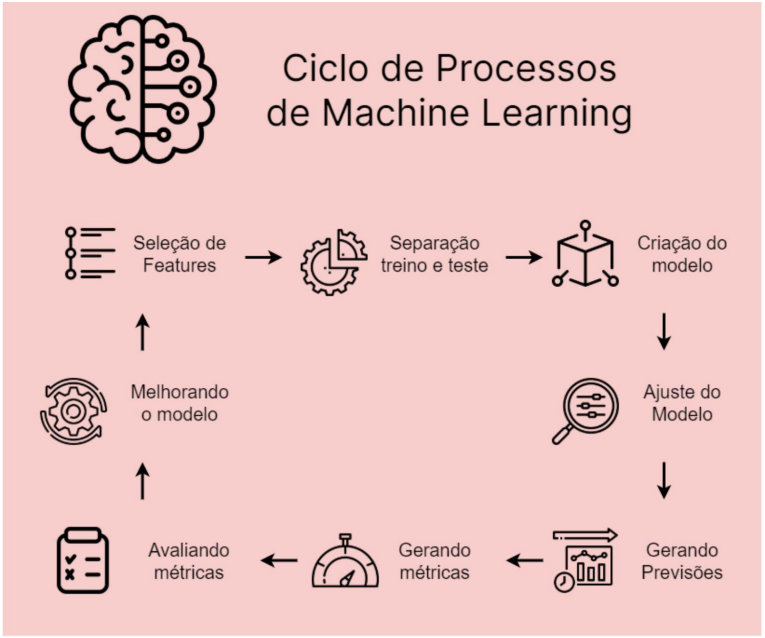
Agenda

1- Estudo de caso Imóveis
2- Métricas
3- Árvore de decisão
4 - Random Forest

Estudo de caso

Você foi contratado por uma empresa para criar uma modelo de aprendizado de máquina para prever o preço dos imóveis com base na quantidade de andares, cômodos.

Etapas de Machine Learning



Métricas

R2 score - coeficiente de determinação, indica o quão próximo a reta da previsão está do valor real do conjunto de dados

```
#importar a biblioteca para calcular a métrica r2_score
from sklearn.metrics import r2_score

r2_lr = r2_score(y_teste, previsao_lr)
r2_lr
```

$$R^2 = 1 - \frac{SS_{RES}}{SS_{TOT}} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \overline{y})^2}$$

Métricas

MSE (Mean Squared Error ou Erro Quadrático Médio): média da diferença elevada ao quadrado entre o valor real e o previsto. (penalidade sobre o erro)

```
from sklearn.metrics import mean_squared_error

y_true = [[0.5, 1],[-1, 1],[7, -6]]
y_pred = [[0, 2],[-1, 2],[8, -5]]

mean_squared_error(y_true, y_pred, squared=True)
```

O erro quadrático médio, MSE (da sigla em inglês Mean Squared Error), é comumente usado para verificar a acurácia de modelos e dá um maior peso aos maiores erros, já que, ao ser calculado, cada erro é elevado ao quadrado individualmente e, após isso, a média desses erros quadráticos é calculada.

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

Métricas

MAE (Mean Absolute Error ou Erro Absoluto Médio): média da diferença absoluta (módulo) entre o valor real e o previsto.

```
from sklearn.metrics import mean_absolute_error

y_true = [3, -0.5, 2, 7]
y_pred = [2.5, 0.0, 2, 8]
mean_absolute_error(y_true, y_pred)

from sklearn.metrics import mean_absolute_percentage_error

y_true = [3, -0.5, 2, 7]
y_pred = [2.5, 0.0, 2, 8]

mean_absolute_percentage_error(y_true, y_pred)
```

O erro médio absoluto, MAE (da sigla em inglês Mean Absolute Error), é calculado a partir da média dos erros absolutos, ou seja, utilizamos o módulo de cada erro para evitar a subestimação, isso porque, o valor é menos afetado por pontos especialmente extremos (outliers).

$$MAE = \frac{1}{n}\sum_{i=1}^{n} | \mathrm{Y}i - \hat{\mathrm{Y}}i |$$

Métricas

Root Mean Squared Error - RMSE

A Raiz Quadrada do Erro Médio é bem similar à métrica MSE, de modo que ela é calculada pela raiz quadrada dele. O RMSE determina a distância de um ponto de dados em relação a linha ajustada, medida ao longo de uma linha vertical. Desse modo, ela informa a concentração dos dados em torno da linha de ajuste.

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}}$$

```
from sklearn.metrics import mean_squared_error
 y_true = [3, -0.5, 2, 7]
 y_pred = [2.5, 0.0, 2, 8]
mean_squared_error(y_true, y_pred)
```

Métricas

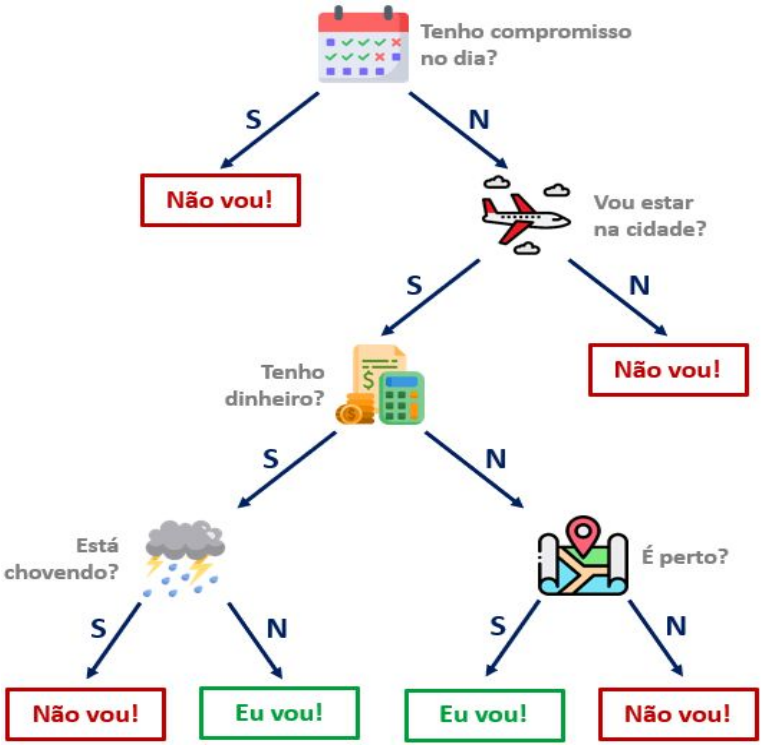
Mean Absolute Percentage Error - MAPE
O Mean Absolute Percentage Error é bem similar ao Mean Absolute Error - MAE, com a diferença de que ele mede a precisão como uma porcentagem e pode ser calculado como a porcentagem do MAE para cada amostra. O MAPE é muito utilizado em problemas de regressão pois traz uma interpretação bem intuitiva quanto ao erro relativo.

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{|y_i - \hat{y}_i|}{y_i}.100\%$$

```
from sklearn.metrics import mean_absolute_percentage_error
y_true = [3, -0.5, 2, 7]
y_pred = [2.5, 0.0, 2, 8]
mean_absolute_percentage_error(y_true, y_pred)
```

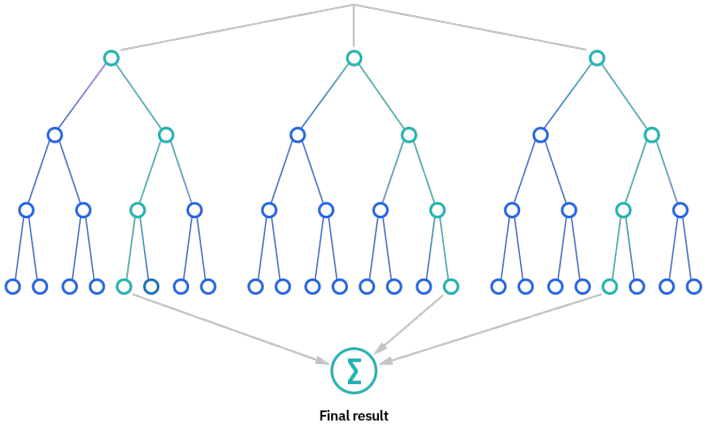
Modelo Árvore de decisão

É um algoritmo de aprendizado supervisionado que cria uma estrutura na forma de árvore para tomar decisões. Cada nó interno representa um teste em uma característica, cada ramo representa um resultado possível desse teste e cada folha representa uma classe ou valor de saída



Modelo Random Forest

É uma técnica de aprendizado conjunto que utiliza várias árvores de decisão. Cada árvore é treinada em uma amostra aleatória dos dados e a classificação final é determinada por votação ou média das previsões das árvores



Vantagens
Redução de ocorrência de overfitting
Flexibilidade
Lida com valores faltantes
Boa acurácia
Grandes volumes de dados

Desvantagens
Modelo complexo
Demanda mais poder computacional

Modelo Random Forest

Um algoritmo bastante conhecido e que utiliza essa abordagem é o Random Forest, que segue as seguintes etapas:

1. Escolher aleatoriamente amostras a partir do conjunto de treinamento;
2. Construir a árvore de decisão associada a esse conjunto de dados;
3. Repetir os passos 1 e 2 para uma quantidade de árvores escolhida pela pessoa cientista de dados; e
4. Definir o resultado da previsão como sendo a predição mais frequente para problemas de classificação ou média dos resultados obtidos para problemas de regressão.

Obrigado!

Prof. Me Daniel Vieira

Email: danielvieira2006@gmail.com

Linkedin: Daniel Vieira

Instagram: Prof daniel.vieira95