CURSO DE DATA SCIENCE

Prof. Marcelo Bianchi

Atividade da aula do dia 30/03/2021 - Decision Tree Algorithm

Integrantes do Grupo: Daniel Moreira, Lia Morimoto, Raphael Bezerra, Thainan Abreu

Instruções da Atividade

Essa atividade deverá ser feita em grupo de 4 pessoas.

Data de Entrega: 06/04/2021

Siga as instruções e poste o arquivo word com as respostas na plataforma.

1) Execute o código no jupyter e responda às seguintes perguntas:

   a) Do que se trata o Dataset ?

   >> O dataset trata-se de um cenário hipotético criado pela IBM para explorar questões de desligamentos de funcionários. Esses desligamentos podem ser causados por motivos pessoais ou profissionais e criam vacâncias de cargos, o que pode interferir nos processos da empresa.

   b) Quais são as suas conclusões sobre cada um dos itens abaixo do dataset:

   BusinessTravel :
   >> A taxa de desligamento entre os trabalhadores aumenta entre aqueles com viagens mais frequentes. Possivelmente pelo desgaste ou rotina que os mantém afastados de casa.

   Department :
   >> A taxa de desligamento é mais baixa no setor de Pesquisa e Desenvolvimento.

   EducationField :
   >> A taxa de desligamento é mais alta entre os trabalhadores de Recursos Humanos e os de grau Técnico do que em quaisquer outras áreas.

   Gender :
   >> A taxa de desligamento é maior entre os Homens.

JobRole :
>>Dado um cargo, a taxa de desligamento é mais alta quando se trata de Técnicos de Laboratório, Representantes de Vendas e Recursos Humanos.

MaritalStatus :
>> A taxa de desligamento é mais alta entre os trabalhadores Solteiros

OverTime :
>> A taxa de desligamento cresce à medida em que a carga horária também aumenta.

2) Descreva cada um dos parâmetros do algoritmo decision tree

- `criterion`: medida de qualidade da divisão do algoritmo, pode ser “gini” ou “entropy”
- `splitter`: estratégia que o algoritmo usará para dividir o nó, pode ser dividido da “melhor” forma ou de forma aleatória
- `max_depth`: o limite de profundidade da árvore de decisão, ou seja quantas camadas abaixo da inicial (camada 0) a árvore pode descer.
- `min_samples_split`: número de amostras mínimas nas quais um nó pode se dividir
- `min_samples_leaf`: menor número de amostras para que um nó seja considerado uma folha.
- `min_weight_fraction_leaf`:
- `max_features`: quantas features serão considerada enquanto procura o melhor resultado para o nó
- `max_leaf_nodes`: quantidade máxima de folhas formadas
- `min_impurity_decrease`: um nó apenas será dividido se reduzir a entropia em um valor menor ou igual a esse.
- `min_impurity_split`: quantidade mínima de entropia para que um nó possa se dividir, se estiver abaixo disso o nó será considerado uma folha.

3) Explique o que é CONFUSION MATRIX ?

>> Confusion Matrix, Matriz de Confusão, é uma matriz com as frequências (absolutas ou relativas) de classificação para cada classe do modelo e mostra se a árvore separa as classes bem e corretamente usando as seguintes métricas:

Positivo Verdadeiro (TP - True Positive), Positivo Falso (FP - False Positive), Negativo Verdadeiro (TN - True Negative) e Negativo Falso (FN - False Negative).

******	Treinamento +	Treinamento -
Teste +	TP	FP
Teste -	FN	TN

4) Explique o que é Classification Report

```
_____________________________________________________
CLASSIFICATION REPORT:
                0      1  accuracy  macro avg  weighted avg
precision    1.00   1.00      1.00       1.00          1.00
recall       1.00   1.00      1.00       1.00          1.00
f1-score     1.00   1.00      1.00       1.00          1.00
support    853.00 176.00      1.00    1029.00       1029.00
_____________________________________________________
```

>> Assim como a função de mesmo nome da biblioteca scikit-learn, o classification Report é um texto gerado com as principais métricas de classificação sumarizadas.

5) Descreva cada item do Classification Report

Precision :
>> Precisão se refere à proporção dos positivos identificados corretamente(TP)
Precision = TP/(TP+FP)

Recall:
>> Recall (Sensibilidade) é a proporção de positivos corretamente identificados em relação ao total de positivos.
Recall = TP/(TP+FN)

F1-score:
>> F1-Score é uma média ponderada entre a Precision e a Recall
F1-Score = 2 * (Precision * Recall) / (Precision + Recall)

Support:

>>Suporte é o número total de ocorrências de uma classe em questão

Support = TP + FP ou Support = TN + FN