

Ciência de Dados (Big Data Processing and Analytics)

Big Data Analytics – Mineração e Análise de Dados



Professor curador
Prof. Dr. Rogério de Oliveira





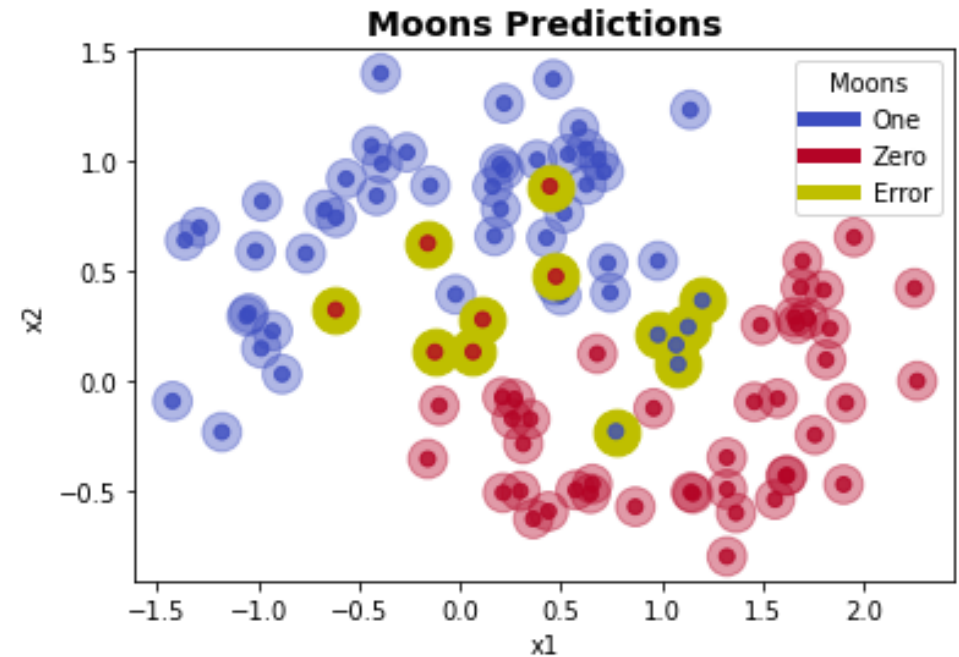
TRILHA 3

Classificação e Métricas

Parte A

Estimadores do Scikit-Learn

1. Import do estimador
2. Preparação dos dados X e y do estimador (variáveis preditoras e objetivo)
3. Instanciação e configuração do estimador
4. Ajuste do modelo (treinamento ou estimativa dos parâmetros)
5. Predição
6. Obtenção de Métricas e Avaliação do Modelo



Estimadores do Scikit-Learn

Import dos Estimadores

```
from sklearn.linear_model import LogisticRegression
```

Instanciando e Definindo os Parâmetros

```
[ ] clf = LogisticRegression(max_iter=1000)  
print(clf)
```

Ajustando os Parâmetros do Modelo

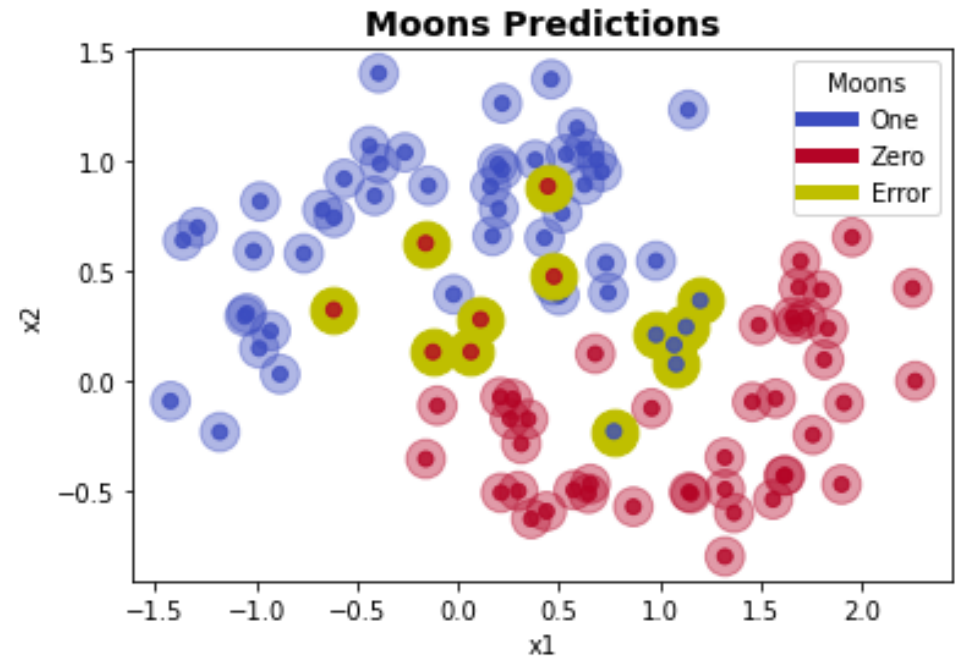
```
[ ] X = df[['x1', 'x2']]  
y = df.y  
  
clf.fit(X,y)  
  
print( clf.coef_, clf.intercept_)
```

Predição

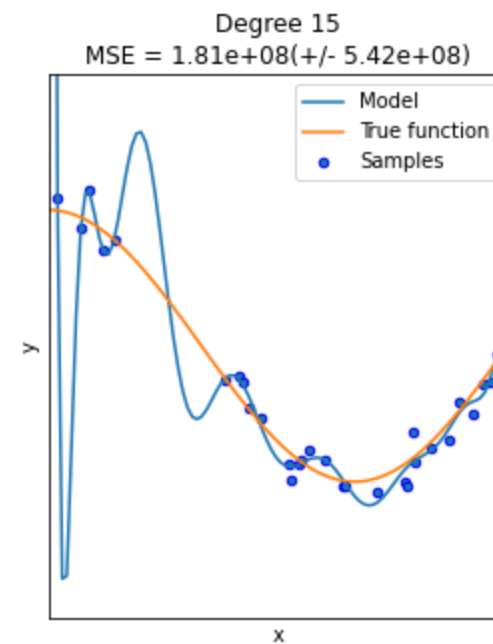
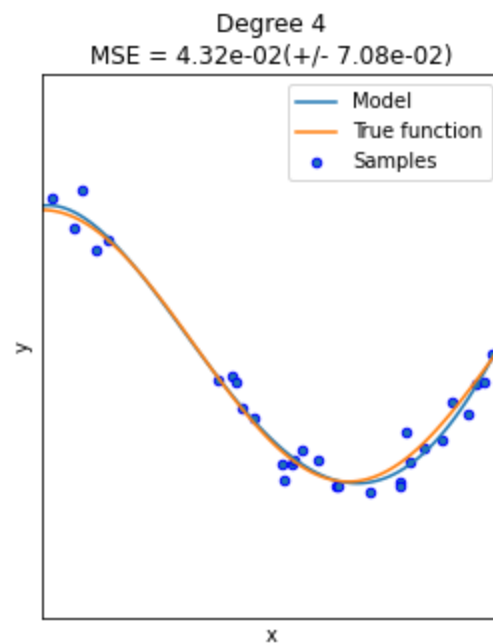
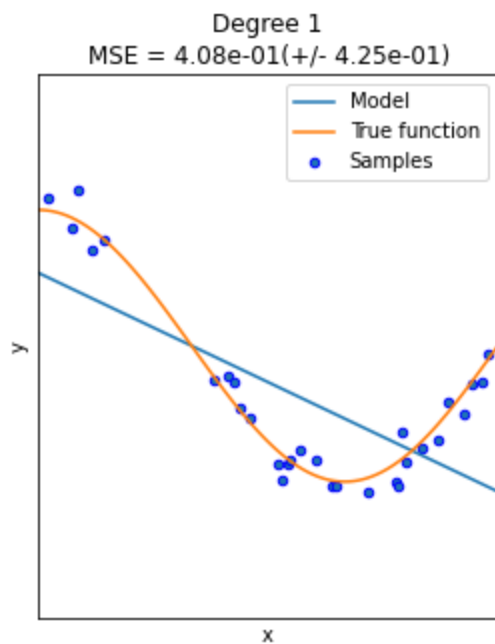
```
[ ] clf.predict(X)
```

Score do Modelo

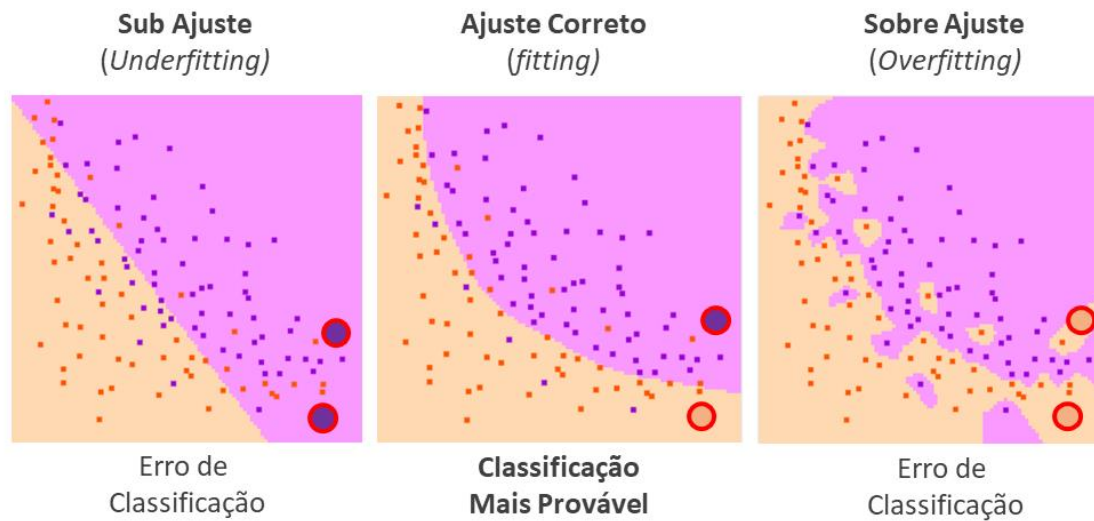
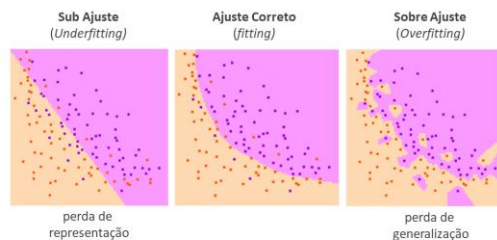
```
[ ] clf.score(X,y)
```



Dilema Viés-Variância: Regressão



Dilema Viés-Variância: Classificação



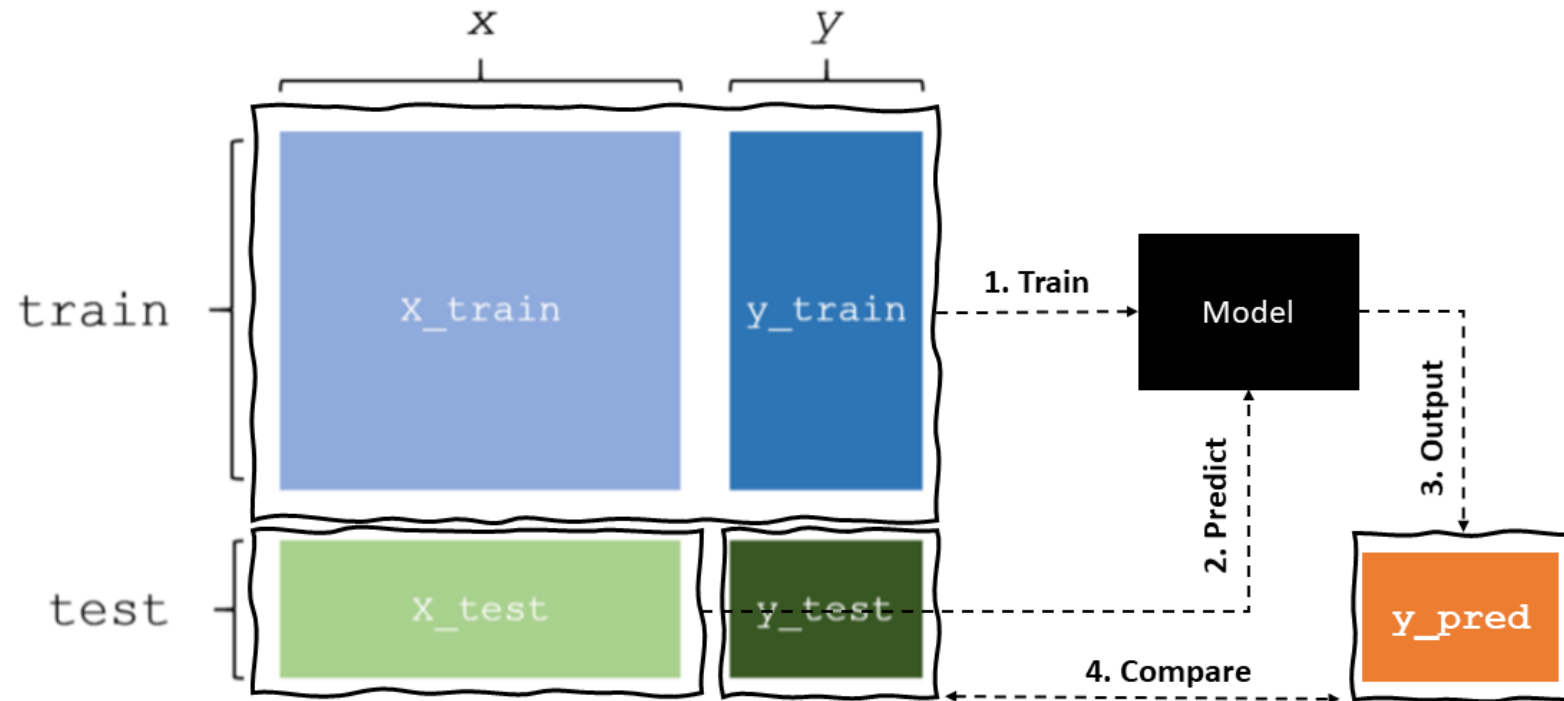


TRILHA 3

Classificação e Métricas

Parte B

Conjuntos de Treinamento e Teste



Matriz de Confusão

		Valor Previsto	
		Positivo	Negativo
Valor Real	Positivo	TP	FN
	Negativo	FP	TN

		y_pred	
		1	0
y_test	1	14	153
	0	11	920

Matriz de Confusão

		Valor Previsto	
		Positivo	Negativo
Valor Real	Positivo	TP	FN
	Negativo	FP	TN

		y_pred	
		1	0
y_test	1	14	153
	0	11	920

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

$$Accuracy = \frac{TotaldeAcertos}{TotaldeCasos}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

Classification Report

Matriz de Confusão:

```
[[ 14 153]
 [ 11 920]]
```

Score de Acuracidade:

0.8506375227686703

Classification Report:

	precision	recall	f1-score	support
1	0.56	0.08	0.15	167
0	0.86	0.99	0.92	931
accuracy			0.85	1098
macro avg	0.71	0.54	0.53	1098
weighted avg	0.81	0.85	0.80	1098

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

$$Accuracy = \frac{TotaldeAcertos}{TotaldeCasos}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

Cuidado!

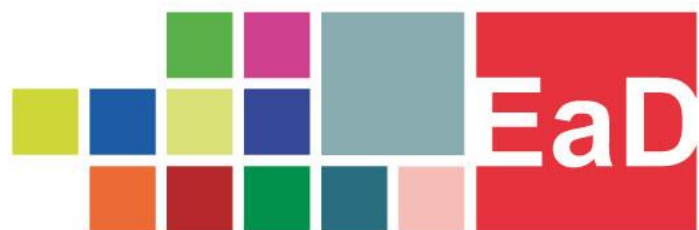
Real/Pred	Maligno	Benigno
Maligno	2	3
Benigno	0	95

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

$$Accuracy = \frac{TotaldeAcertos}{TotaldeCasos}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$



Universidade Presbiteriana
Mackenzie