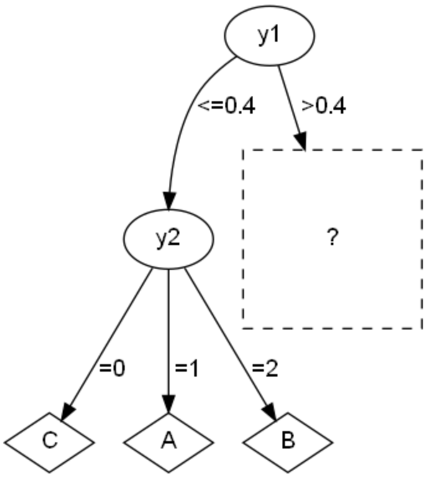
Aprendizagem 2023/24

Homework I - Decision Trees and Evaluation

I. Pen-and-paper [11v]

Consider the partially learnt decision tree from the dataset D. D is described by four input variables – one numeric with values in [0,1] and 3 categorical – and a target variable with three classes.

D	y_1	y_2	y_3	y_4	y_{out}
x_1	0.24	1	1	0	A
x_2	0.06	2	0	0	B
x_3	0.04	0	0	0	B
x_4	0.36	0	2	1	C
x_5	0.32	0	0	2	C
x_6	0.68	2	2	1	A
x_7	0.9	0	1	2	A
x_8	0.76	2	2	0	A
x_9	0.46	1	1	1	B
x_{10}	0.62	0	0	1	B
x_{11}	0.44	1	2	2	C
x_{12}	0.52	0	2	0	C



1) [5v] Complete the given decision tree using Information gain with Shannon entropy ($\log_2$). Consider that: i) a minimum of 4 observations is required to split an internal node, and ii) decisions by ascending alphabetic order should be placed in case of ties.

Homework I: Decision Trees and Evaluation

I. Pen-and-Paper

Exercício 1)

· Sendo $y_1 > 0{,}4$:

$\rightarrow IG(y_2) = H(y_{out}) - H(y_{out}|y_2) = -\left(\frac{3}{7}\cdot\log_2\left(\frac{3}{7}\right)+\frac{2}{7}\cdot\log_2\left(\frac{2}{7}\right) + \frac{2}{7}\cdot\log_2\left(\frac{2}{7}\right)\right) - \left[\frac{3}{7}\left(-\left(\frac{1}{3}\cdot\log_2\left(\frac{1}{3}\right)+\frac{1}{3}\cdot\log_2\left(\frac{1}{3}\right)+\frac{1}{3}\cdot\log_2\left(\frac{1}{3}\right)\right)\right) + \frac{2}{7}\left(-\left(\underbrace{\frac{1}{2}\cdot\log_2\left(\frac{1}{2}\right)+\frac{1}{2}\cdot\log_2\left(\frac{1}{2}\right)}_{=-1}\right)\right)+\frac{2}{7}\left(-\left(\underbrace{1\cdot\log_2(1)}_{=0}\right)\right)\right] =$

$= -\left(\frac{3}{7}\cdot\log_2\left(\frac{3}{7}\right)+\frac{4}{7}\cdot\log_2\left(\frac{2}{7}\right)\right) - \left[\frac{3}{7}\left(-\log_2\left(\frac{1}{3}\right)\right)+\frac{2}{7}+0\right] \simeq \boxed{0{,}592}$

$\rightarrow IG(y_3) = H(y_{out}) - H(y_{out}|y_3) = H(y_{out}) - \left[\frac{1}{7}\left(-\left(\underbrace{1\cdot\log_2(1)}_{=0}\right)\right) + \frac{2}{7}\left(-\left(\underbrace{\frac{1}{2}\cdot\log_2\left(\frac{1}{2}\right)+\frac{1}{2}\cdot\log_2\left(\frac{1}{2}\right)}_{=-1}\right)\right)+\frac{4}{7}\left(-\left(\underbrace{\frac{1}{2}\cdot\log_2\left(\frac{1}{2}\right)+\frac{1}{2}\cdot\log_2\left(\frac{1}{2}\right)}_{=-1}\right)\right)\right] =$

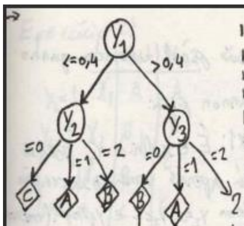
$= H(y_{out}) - \left(\frac{2}{7}+\frac{4}{7}\right) \simeq \boxed{0{,}6995}$

$\rightarrow IG(y_4) = H(y_{out}) - H(y_{out}|y_4) = H(y_{out}) - \left[\frac{2}{7}\left(-\left(\underbrace{\frac{1}{2}\cdot\log_2\left(\frac{1}{2}\right) + \frac{1}{2}\cdot\log_2\left(\frac{1}{2}\right)}_{=-1}\right)\right) + \frac{3}{7}\left(-\left(\frac{1}{3}\cdot\log_2\left(\frac{1}{3}\right)+\frac{2}{3}\cdot\log_2\left(\frac{2}{3}\right)\right)\right)+\frac{2}{7}\left(-\left(\underbrace{\frac{1}{2}\cdot\log_2\left(\frac{1}{2}\right)+\frac{1}{2}\cdot\log_2\left(\frac{1}{2}\right)}_{=-1}\right)\right)\right]$

$= H(y_{out}) - \left[\frac{2}{7}+\frac{3}{7}\left(-\left(\frac{1}{3}\cdot\log_2\left(\frac{1}{3}\right)+\frac{2}{3}\cdot\log_2\left(\frac{2}{3}\right)\right)\right)+\frac{2}{7}\right] \simeq \boxed{0{,}592}$

· Como $IG(y_3) > IG(y_2) = IG(y_4)$, então o próximo nó será y_3.

→



↳ É escolhido B, porque há apenas uma observação com $y_3 = 0$ e $y_1 > 0{,}4$ (menor que o mínimo de 4 observações) em x_{10}. Sendo $y_{out} = B$.

↳ É escolhido A porque há duas observações com $y_3 = 1$ e $y_1 > 0{,}4$ (menor que o mínimo de 4 observações) em x_7 e x_9, sendo A e B os respetivos y_{out}. Logo, obdecendo à regra de ordem alfabética escolhe-se A.

· Sendo $y_1 > 0{,}4$ e $y_3 = 2$:

$\rightarrow$ ~~$IG(y_{out})$~~

$\rightarrow IG(y_2) = H(y_{out}) - H(y_{out}|y_2) =$

$= -\left(\underbrace{\frac{1}{2}\cdot\log_2\left(\frac{1}{2}\right)+\frac{1}{2}\cdot\log_2\left(\frac{1}{2}\right)}_{=-1}\right) - \left[\frac{1}{2}\left(-\left(\underbrace{1\cdot\log_2(1)}_{=0}\right)\right)+\frac{1}{4}\left(-\left(\underbrace{1\cdot\log_2(1)}_{=0}\right)\right) + \frac{1}{4}\left(-\left(\underbrace{1\cdot\log_2(1)}_{=0}\right)\right)\right] = \boxed{1}$

$\rightarrow IG(y_4) = H(y_{out}) - H(y_{out}|y_4) =$

$= H(y_{out}) - \left[\frac{1}{2}\left(-\left(\underbrace{\frac{1}{2}\cdot\log_2(1)+\frac{1}{2}\cdot\log_2(1)}_{=-1}\right)\right) + \frac{1}{4}\left(-\left(\underbrace{1\cdot\log_2(1)}_{=0}\right)\right)+\frac{1}{4}\left(-\left(\underbrace{1\cdot\log_2(1)}_{=0}\right)\right)\right] =$

$= H(y_{out}) - \frac{1}{2} = \boxed{\frac{1}{2}}$

· Como $IG(y_2) > IG(y_4)$, para $y_1 > 0{,}4$ e $y_3 = 2$, então o próximo nó a ser escolhido será y_2.

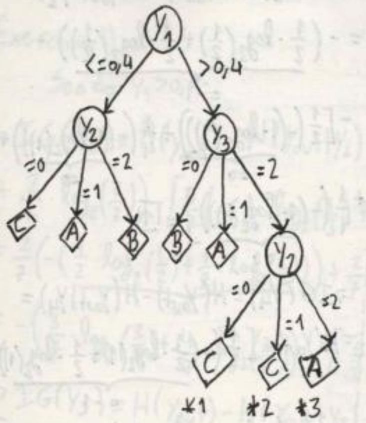
→

Aprendizagem 2023/24

Homework I - Decision Trees and Evaluation

→

· Conclusão: A árvore de decisão final usando ganho de informação com entropia de Shannon fica:



*1: É escolhido C, porque há apenas uma observação com $y_2=0$, $y_3=2$ e $y_1>0{,}4$ (menor que o mínimo de 4 observações) em X_{12}. Sendo $Y_{out}=C$.

*2: É escolhido C, porque há apenas uma observação com $y_2=1$, $y_3=2$ e $y_1>0{,}4$ (menor que o mínimo de 4 observações) em X_{11}. Sendo $Y_{out}=C$.

*3: É escolhido A porque há apenas duas observações com $y_2=2$, $y_3=2$ e $y_1>0{,}4$ (menor que o mínimo de 4 observações) em $X=6$ e $X=8$. Sendo que ambos têm o mesmo $Y_{out}=A$.

2) [2.5v] Draw the training confusion matrix for the learnt decision tree.

Exercício 2)

		Y_{out}	$\hat{Y}_{out}$
$y_2=1$	X_1	A	A
$y_2=2$	X_2	B	B
$y_2=0$	X_3	B	C
	X_4	C	C
	X_5	C	C
$y_2=2 \wedge y_3=2$	X_6	A	A
$y_3=1$	X_7	A	A
$y_2=2 \wedge y_3=2$	X_8	A	A
$y_3=1$	X_9	B	A
$y_3=0$	X_{10}	B	B
$y_2=1 \wedge y_3=2$	X_{11}	C	C
$y_2=0 \wedge y_3=2$	X_{12}	C	C

· Conclusão: A matriz confusão para a árvore de decisão do exercício anterior fica da seguinte forma:

		real (Y_{out}) A	B	C
previsto ($\hat{Y}_{out}$)	A	4	1	0
	B	0	2	0
	C	0	1	4

Aprendizagem 2023/24

Homework I - Decision Trees and Evaluation

3) [1.5v] Identify which class has the lowest training F1 score.

Exercício 3)

$$F1\text{-}Score = 2 \cdot \frac{1}{\frac{1}{P}+\frac{1}{R}}, \quad P = \frac{TP}{TP+FP} \quad e \quad R = \frac{TP}{TP+FN}$$

. Class A

TP	FP	FP
FN	TN	TN
FN	TN	TN

$TP=4$ $TN=7$
$FP=1$ $FN=0$
$P=\frac{4}{5}$ $R=1$

. Class B

TN	FN	TN
FP	TP	FP
TN	FN	TN

$TP=2$ $TN=8$
$FP=0$ $FN=2$
$P=1$ $R=\frac{1}{2}$

. Class C

TN	TN	FN
TN	TN	FN
FP	FP	TP

$TP=4$ $TN=7$
$FP=1$ $FN=0$
$P=\frac{4}{5}$ $R=1$

$$\cdot F_1\text{-}Score\,(Class\,A) = 2 \cdot \frac{1}{\frac{1}{\frac{4}{5}}+\frac{1}{1}} \simeq 0{,}89$$

$$\cdot F_1\text{-}Score\,(Class\,B) = 2 \cdot \frac{1}{\frac{1}{1}+\frac{1}{\frac{1}{2}}} \simeq \boxed{0{,}67}$$

$$\cdot F1\text{-}Score\,(Class\,C) = 2 \cdot \frac{1}{\frac{1}{\frac{4}{5}}+\frac{1}{1}} \simeq 0{,}89$$

R: Class B has the lowest training F1-Score.

4) [1v] Considering y_2 to be ordinal, asses if y_1 and y_2 are correlated using the Spearman coefficient.

Exercício 4)

	y_1	y_2	rank(y_1)	rank(y_2)
x_1	0,24	1	3	8
x_2	0,06	2	2	11
x_3	0,04	0	1	3,5
x_4	0,36	0	5	3,5
x_5	0,32	0	4	3,5
x_6	0,68	2	10	11
x_7	0,9	0	12	3,5
x_8	0,76	2	11	11
x_9	0,46	1	7	8
x_{10}	0,62	0	9	3,5
x_{11}	0,44	1	6	8
x_{12}	0,52	0	8	3,5

$$\sum rank(y_1)\cdot rank(y_2) = 24+22+3{,}5+(5\cdot 3{,}5)+(4\cdot 3{,}5)+110+(12\cdot 3{,}5)+121+(7\cdot 8)+(9\cdot 3{,}5)+(6\cdot 8)+(8\cdot 3{,}5) = 517{,}5$$

$$\sum rank(y_1) = 3+2+1+5+4+10+12+11+7+9+6+8 = 78$$

$$\sum rank(y_2) = 8+11+3{,}5+3{,}5+3{,}5+11+3{,}5+11+8+3{,}5+8+3{,}5 = 78$$

$$\sum rank(y_1)^2 = 9+4+1+25+16+100+144+121+49+81+36+64 = 650$$

$$\sum rank(y_2)^2 = 64+121+12{,}25+12{,}25+12{,}25+121+12{,}25+121+64+12{,}25+64+12{,}25 = 628{,}5$$

$$\left(\sum rank(y_1)\right)^2 = \left(\sum rank(y_2)\right)^2 = 6084$$

$$Spearman(y_1,y_2) = PCC(rank(y_1), rank(y_2)) = \frac{cov(rank(y_1), rank(y_2))}{\sigma(rank(y_1))\cdot\sigma(rank(y_2))} =$$

$$= \frac{\sum (rank(y_1)\cdot rank(y_2)) - \frac{\sum rank(y_1)\cdot \sum rank(y_2)}{n}}{\sqrt{\left(\sum rank(y_1)^2 - \frac{(\sum rank(y_1))^2}{n}\right)\cdot\left(\sum rank(y_2)^2 - \frac{(\sum rank(y_2))^2}{n}\right)}}, \quad n=12 =$$

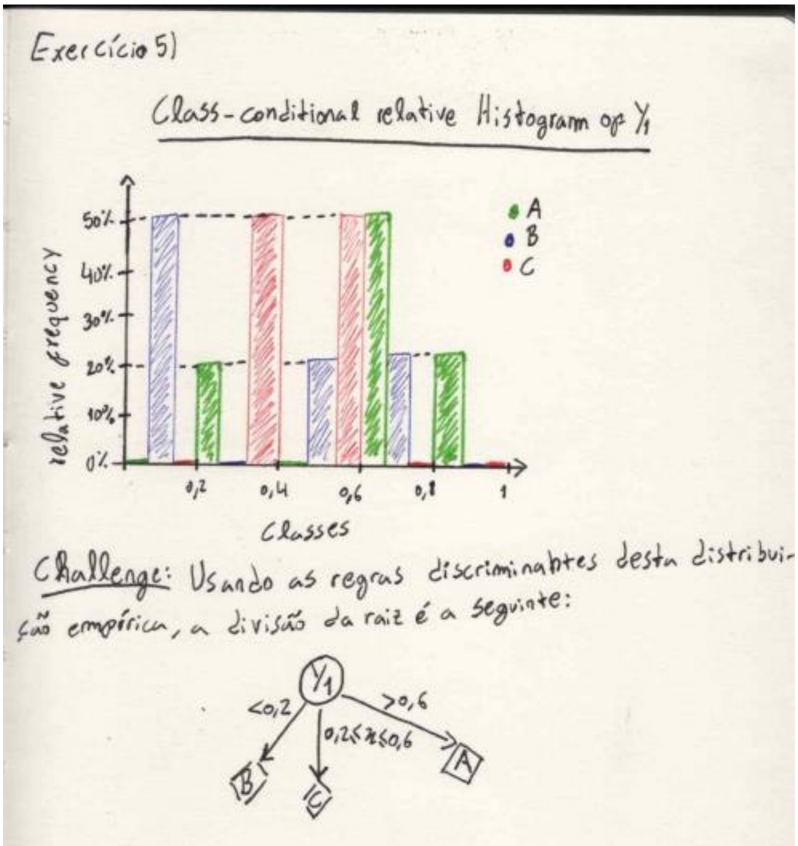
$$= \frac{517 - \frac{78\cdot 78}{12}}{\sqrt{\left(650-\frac{6084}{12}\right)\left(628-\frac{6084}{12}\right)}} \Leftrightarrow \boxed{Spearman(y_1,y_2) \simeq 0{,}076}$$

TÉCNICO LISBOA

Aprendizagem 2023/24

Homework I - Decision Trees and Evaluation

5) [1v] Draw the class-conditional relative histogram of y_1 using 5 equally spaced bins in [0,1]. Challenge: find the root split using the discriminant rules from these empirical distributions.



Aprendizagem 2023/24

Homework I - Decision Trees and Evaluation

II. Programming [9v]

To answer the following questions, consider using the sklearn API documentation and the notebooks in the coursewebpageasguidance.ShowinyourPDFreportboththecodeandthecorrespondingresults.

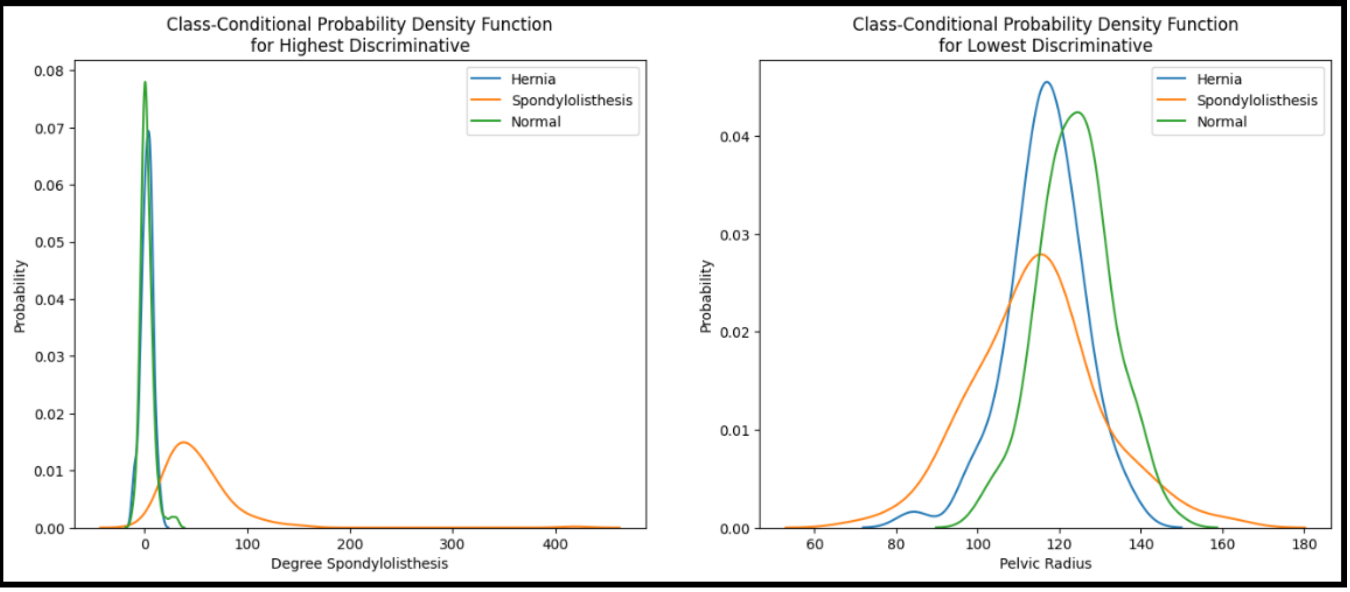
Consider the column_diagnosis.arff data available at the homework tab, comprising 6 biomechanical features to classify 310 orthopaedic patients into 3 classes (normal, disk hernia, spondilolysthesis).

1) [1.5v] Apply f_classif from sklearn to assess the discriminative power of the input variables. Identify the input variable with the highest and lowest discriminative power. Plot the class-conditional probability density functions of these two input variables.

```
import pandas as pd
from scipy.io.arff import loadarff
from sklearn.feature_selection import f_classif
import matplotlib.pyplot as plt
import seaborn as sns

# Read ARFF file
data = loadarff('column_diagnosis.arff')
df = pd.DataFrame(data[0])
df['class'] = df['class'].str.decode('utf-8')

X = df.drop('class', axis=1)
y = df['class']

fimportance = f_classif(X, y)

# Create a DataFrame to store variable names, F-values, and p-values
values_df = pd.DataFrame({'Attribute': X.columns.values, 'F-Value': fimportance[0], 'P-Value': fimportance[1]})

# Identify the input variable with the highest and lowest discriminative power
highest_d = values_df['Attribute'][values_df['F-Value'].idxmax()]
lowest_d = values_df['Attribute'][values_df['F-Value'].idxmin()]
print(f'Highest discriminative power: {highest_d}')
print(f'Lowest discriminative power: {lowest_d}')


classes = df['class'].unique()
plt.figure(figsize=(16, 6))

plt.subplot(1, 2, 1)
for cls in classes:
    sns.kdeplot(X[y == cls][highest_d], label= cls)
plt.title('Class-Conditional Probability Density Function\nfor Highest Discriminative')
plt.xlabel(highest_d.title().replace("_", " "))
plt.ylabel('Probability')
plt.legend()

plt.subplot(1, 2, 2)
for cls in classes:
    sns.kdeplot(X[y == cls][lowest_d], label= cls)
plt.title('Class-Conditional Probability Density Function\nfor Lowest Discriminative')
plt.xlabel(lowest_d.title().replace("_", " "))
plt.ylabel('Probability')
plt.legend()
plt.show()
```

Homework I - Decision Trees and Evaluation

```
Highest discriminative power: degree_spondylolisthesis
Lowest discriminative power: pelvic_radius
```



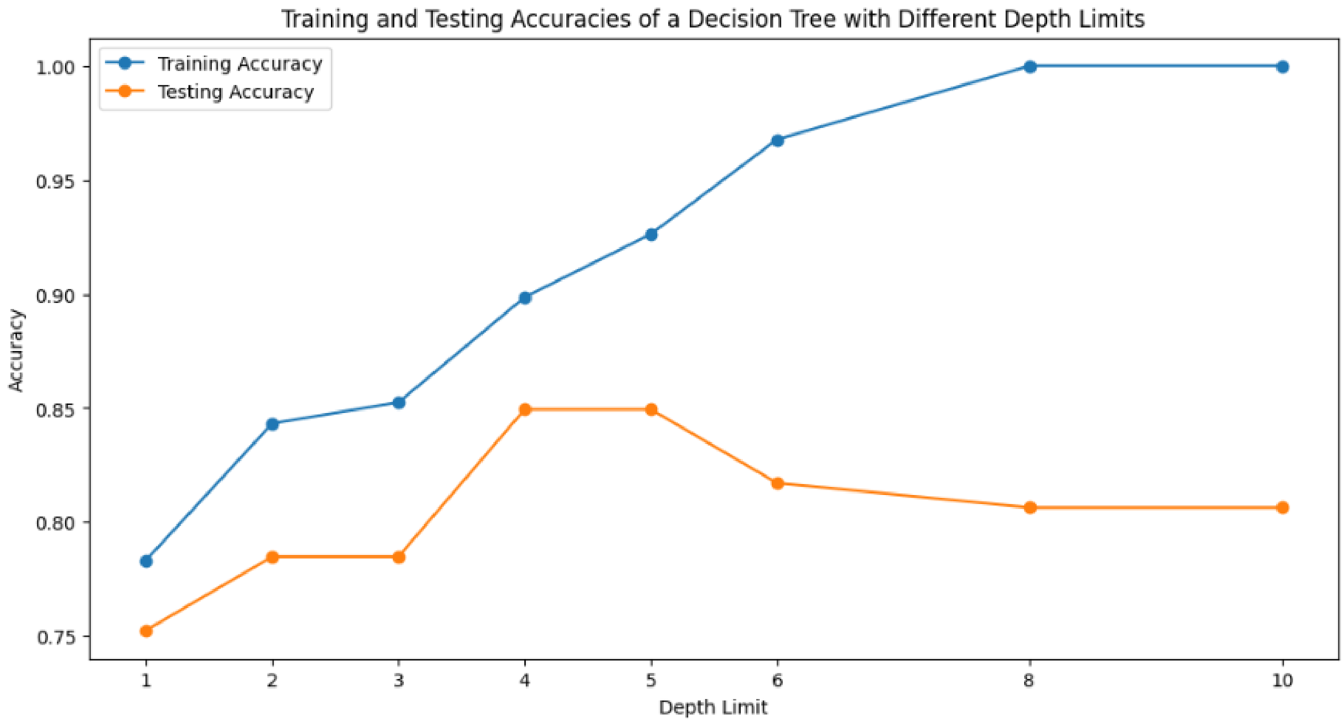
2) [4v] Using a stratified 70-30 training-testing split with a fixed seed (random_state=0), assess in a single plot both the training and testing accuracies of a decision tree with depth limits in {1,2,3,4,5,6,8,10} and the remaining parameters as default.

[optional] Note that split thresholding of numeric variables in decision trees is non-deterministic in sklearn, hence you may opt to average the results using 10 runs per parameterization.

```python
from sklearn import tree
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score

training_accuracies = []
testing_accuracies = []
depth_limits = [1, 2, 3, 4, 5, 6, 8, 10]
n_runs = 10

for depth_limit in depth_limits:
    train_sum = 0
    test_sum = 0

    for i in range(n_runs):
        X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, stratify=y, random_state=0)

        clf = tree.DecisionTreeClassifier(max_depth=depth_limit, random_state=0)
        clf.fit(X_train, y_train)

        train_sum += accuracy_score(y_train, clf.predict(X_train))
        test_sum += accuracy_score(y_test, clf.predict(X_test))

    # Calculate average accuracies for each depth limit
    avg_train_acc = train_sum / n_runs
    avg_test_acc = test_sum / n_runs

    training_accuracies.append(avg_train_acc)
    testing_accuracies.append(avg_test_acc)

plt.figure(figsize=(12, 6))
plt.plot(depth_limits, training_accuracies, label='Training Accuracy', marker='o')
plt.plot(depth_limits, testing_accuracies, label='Testing Accuracy', marker='o')
plt.title('Training and Testing Accuracies of a Decision Tree with Different Depth Limits')
plt.xlabel('Depth Limit')
plt.ylabel('Accuracy')
plt.xticks(depth_limits)
plt.legend()
plt.show()
```

Aprendizagem 2023/24

Homework I - Decision Trees and Evaluation



3) [1.5v] Comment on the results, including the generalization capacity across settings.

À medida que a profundidade máxima da árvore aumenta, a precisão do treinamento também aumenta. No entanto, a precisão do teste começa a diminiur após um certo ponto. Isso sugere que, com uma profundidade máxima muito alta, o modelo produz árvores que superajustam os dados de treinamento, ou seja, captura o ruído nos dados de treinamento, em vez de aprender relações gerais que podem ser aplicadas a novos dados. Assim conclui-se que o modelo tem baixa capacidade de generalização.

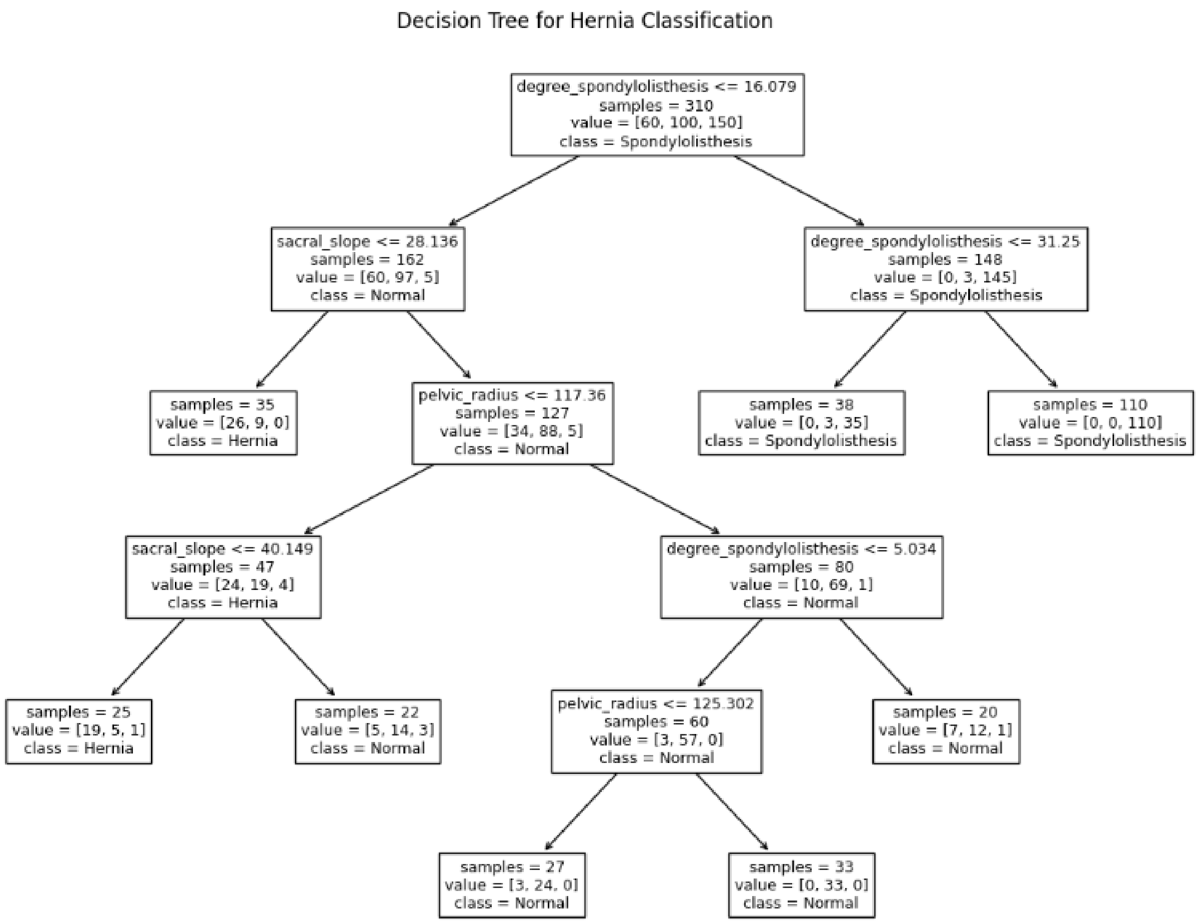
Aprendizagem 2023/24

Homework I - Decision Trees and Evaluation

4) [2v] To deploy the predictor, a healthcare team opted to learn a single decision tree (random_state=0) using *all* available data as training data, and further ensuring that each leaf has a minimum of 20 individuals in order to avoid overfitting risks.

    i. Plot the decision tree.

```
clf = tree.DecisionTreeClassifier(random_state=0, min_samples_leaf=20)
clf.fit(X, y)

plt.figure(figsize=(14, 10))
tree.plot_tree(clf, feature_names=X.columns.tolist(), class_names=clf.classes_.tolist(), impurity=False)
plt.title("Decision Tree for Hernia Classification")
plt.show()
```



ii. Characterize a hernia condition by identifying the hernia-conditional associations.

Pelas associações condicionais da árvore obtida, verifica-se a classe Hernia nos casos em que:

- degree spondylolisthesis assume valores inferiores a 16.079 e sacral slope assume valores inferiores a 28.136 ;
- degree spondylolisthesis assume valores inferiores a 16.079, sacral slope tem valores no intervalo]28.136 , 40.149] e pelvic radius assume valores inferiores a 117.36.

END