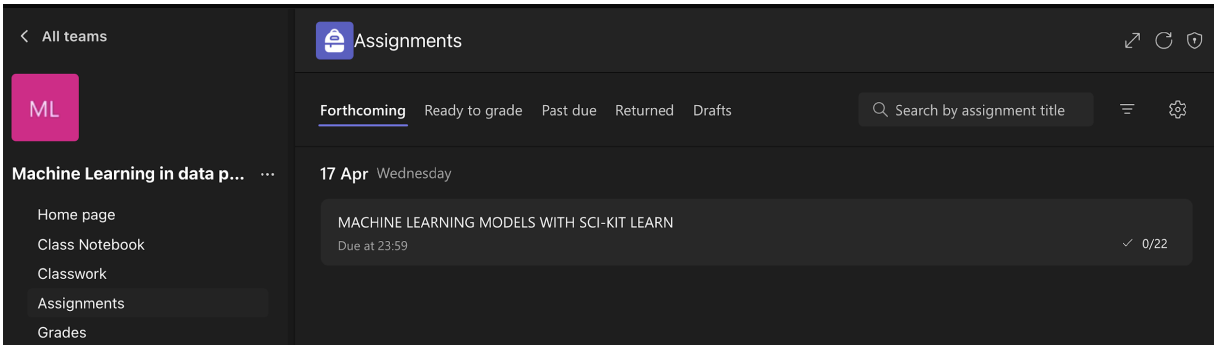
Machine Learning in Data	ERASMUS
Decision Tree	

Decision Tree
The objectives of these assignments are:
- practice loading data from files,
- working with the decision tree

The data we will be working on concerns passengers of the Titanic ship. The models created will predict the probability of survival.

Data files and solution scripts should be submitted **WITHIN 7 DAYS FROM THIS MEETING** today in **IPYNB** format via the **MS Teams platform**.



EXERCISE 1

Prepare data.

Load a sample dataset. For this, use the Titanic dataset.

You can get prepared data from the following RAW link:
https://raw.githubusercontent.com/datasciencedojo/datasets/master/titanic.csv

Tasks:
1. Load the dataset into a pandas DataFrame.
2. Display the first 5 rows and basic information about the data (info(), describe()).
3. Select the following columns for further analysis:
   - Survived (target variable)
   - Pclass, Sex, Age, SibSp, Parch, Fare, Embarked
4. Handle missing values:
   - Fill missing Age values with the median age.
   - Fill missing Embarked values with the most frequent value (mode).

```
PassengerId,Survived,Pclass,Name,Sex,Age,SibSp,Parch,Ticket,Fare,Cabin,Embarked
1,0,3,"Braund, Mr. Owen Harris",male,22,1,0,A/5 21171,7.25,,S
2,1,1,"Cumings, Mrs. John Bradley (Florence Briggs Thayer)",female,38,1,0,PC 17599,71.2833,C85,C
3,1,3,"Heikkinen, Miss. Laina",female,26,0,0,STON/O2. 3101282,7.925,,S
4,1,1,"Futrelle, Mrs. Jacques Heath (Lily May Peel)",female,35,1,0,113803,53.1,C123,S
5,0,3,"Allen, Mr. William Henry",male,35,0,0,373450,8.05,,S
6,0,3,"Moran, Mr. James",male,,0,0,330877,8.4583,,Q
7,0,1,"McCarthy, Mr. Timothy J",male,54,0,0,17463,51.8625,E46,S
8,0,3,"Palsson, Master. Gosta Leonard",male,2,3,1,349909,21.075,,S
9,1,3,"Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)",female,27,0,2,347742,11.1333,,S
10,1,2,"Nasser, Mrs. Nicholas (Adele Achem)",female,14,1,0,237736,30.0708,,C
11,1,3,"Sandstrom, Miss. Marguerite Rut",female,4,1,1,PP 9549,16.7,G6,S
12,1,1,"Bonnell, Miss. Elizabeth",female,58,0,0,113783,26.55,C103,S
13,0,3,"Saundercock, Mr. William Henry",male,20,0,0,A/5. 2151,8.05,,S
14,0,3,"Andersson, Mr. Anders Johan",male,39,1,5,347082,31.275,,S
15,0,3,"Vestrom, Miss. Hulda Amanda Adolfina",female,14,0,0,350406,7.8542,,S
16,1,2,"Hewlett, Mrs. (Mary D Kingcome) ",female,55,0,0,248706,16,,S
17,0,3,"Rice, Master. Eugene",male,2,4,1,382652,29.125,,Q
18,1,2,"Williams, Mr. Charles Eugene",male,,0,0,244373,13,,S
19,0,3,"Vander Planke, Mrs. Julius (Emelia Maria Vandemoortele)",female,31,1,0,345763,18,,S
20,1,3,"Masselmani, Mrs. Fatima",female,,0,0,2649,7.225,,C
```

Or you can download .csv file, upload it on your Google Drive and read it as show below:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S

EXERCISE 2

Create a Decision Tree Model with Scikit-learn in Google Colab.

1. Import Necessary Libraries

- Import pandas for data manipulation.
- Import numpy for numerical operations.
- Import scikit-learn's DecisionTreeClassifier for creating the decision tree model.
- Import train_test_split for splitting the dataset into training and testing sets.

2. Encode categorical variables (Sex, Embarked) using one-hot encoding (e.g., pandas.get_dummies).

3. Split the dataset into features (X) and labels (y), where the label is the 'Survived' column.

Use **train_test_split** method to split the dataset into training and testing sets. Include 80% for training and 20% for testing.

4. Build a decision tree classifier (DecisionTreeClassifier):

Initialize a DecisionTreeClassifier object from scikit-learn.
Fit the model to the training data using the fit() method.

- Set a random_state for reproducibility.
- Optionally limit the max_depth of the tree.

5. Train the model on the training data.

6. Evaluate the model:
- Predict on the test set.
- Calculate accuracy, precision, recall and F1-score.

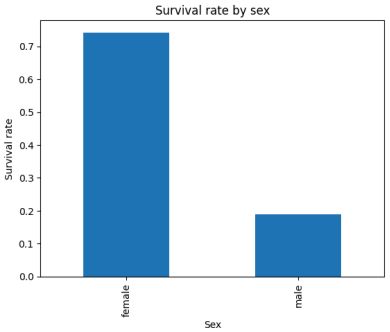
```
Accuracy: 0.7877094972067039

Classification report:
              precision    recall  f1-score   support

           0       0.77      0.94      0.84       110
           1       0.84      0.55      0.67        69

    accuracy                           0.79       179
   macro avg       0.81      0.74      0.76       179
weighted avg       0.80      0.79      0.78       179
```

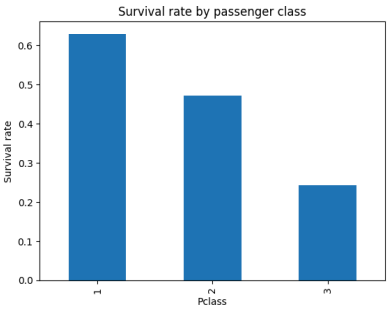
EXERCISE 3

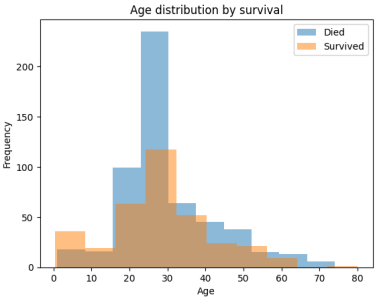
Data visualization.

Using matplotlib (and optionally seaborn), prepare the following charts based on the Titanic dataset:
1. Bar plot of survival rate by sex (percentage of survivors in each group: male, female).
2. Bar plot of survival rate by passenger class (Pclass).
3. Histogram (or density plot) of Age for survivors vs non-survivors.
4. (Optional) Bar plot of survival rate by port of embarkation (Embarked).

Add titles and axis labels to all plots, and briefly describe what you observe.



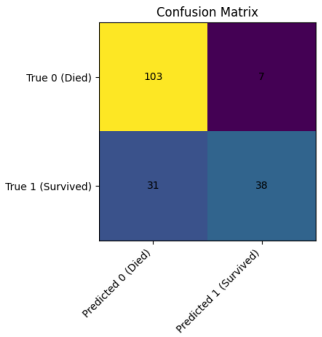


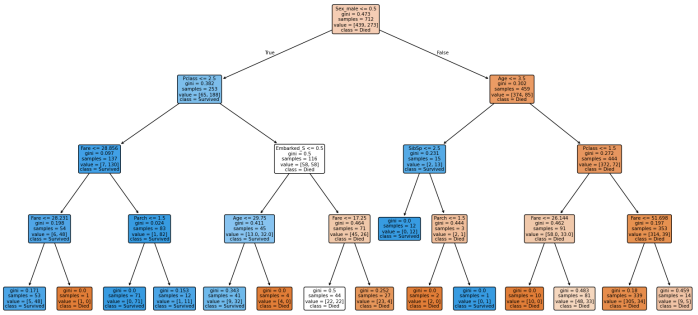


EXERCISE 4

Model evaluation and interpretation.

1. Prepare a confusion matrix for the decision tree model and display it.
2. Interpret the confusion matrix: how many passengers were correctly/incorrectly classified as survivors and non-survivors?
3. Visualize the trained decision tree (e.g., using sklearn.tree.plot_tree).
4. Based on the tree visualization, describe which features seem most important for predicting survival.


Confusion Matrix
True 0 (Died)
True 1 (Survived)
103
7
31
38
Predicted 0 (Died)
Predicted 1 (Survived)


Sex_male <= 0.5
gini = 0.472
samples = 712
value = [439, 273]
class = Died
True
False
Fcons <= 1.5
gini = 0.382
samples = 253
value = [65, 188]
class = Survived
Age <= 3.5
gini = 0.302
samples = 459
value = [374, 85]
class = Died
Age <= 28.69
gini = 0.097
samples = 107
value = [3, 104]
class = Survived
Embarked_S <= 0.5
gini = 0.5
samples = 116
value = [58, 58]
class = Died
Sibsp <= 2.5
gini = 0.211
samples = 15
value = [2, 13]
class = Survived
Pclass <= 1.5
gini = 0.272
samples = 444
value = [372, 72]
class = Died
Fare <= 23.233
gini = 0.199
samples = 54
value = [5, 49]
class = Survived
Sibsp <= 1.5
gini = 0.024
samples = 83
value = [1, 82]
class = Survived
Age <= 26.75
gini = 0.411
samples = 65
value = [13.0, 52.0]
class = Survived
Age <= 17.25
gini = 0.464
samples = 51
value = [45, 24]
class = Died
gini = 0.0
samples = 1
value = [0, 1]
class = Survived
Parch <= 1.5
gini = 0.444
samples = 3
value = [2, 1]
class = Died
Fare <= 26.146
gini = 0.462
samples = 91
value = [59.0, 33.0]
class = Died
Fare <= 51.606
gini = 0.107
samples = 353
value = [314, 39]
class = Died
gini = 0.171
samples = 41
value = [3, 38]
class = Survived
gini = 0.0
samples = 13
value = [2, 11]
class = Survived
gini = 0.0
samples = 71
value = [0, 71]
class = Survived
gini = 0.313
samples = 12
value = [1, 11]
class = Survived
gini = 0.363
samples = 43
value = [12, 31]
class = Survived
gini = 0.0
samples = 22
value = [1.0, 21.0]
class = Died
gini = 0.5
samples = 44
value = [22, 22]
class = Died
gini = 0.252
samples = 7
value = [23, 3]
class = Died
gini = 0.0
samples = 2
value = [0, 2]
class = Survived
gini = 0.0
samples = 1
value = [2, 0]
class = Died
gini = 0.0
samples = 1
value = [0, 1]
class = Survived
gini = 0.0
samples = 61
value = [59, 0]
class = Died
gini = 0.481
samples = 30
value = [18, 12]
class = Died
gini = 0.18
samples = 294
value = [28, 266]
class = Died
gini = 0.459
samples = 59
value = [36, 23]
class = Died