AMI23B – Business Intelligence
Lab 2

Task: Predictive Analysis, Supervised Learning – Titanic

Objectives:

- Understand the Titanic dataset characteristics and structure.
- Perform data exploration and analysis to identify patterns and insights.
- Pre-process and transform the dataset to prepare it for machine learning models.
- Develop a Decision Tree Classifier model to predict survival on the Titanic.
- Develop a Random Forest Classifier model to predict survival on the Titanic.
- Compare the performance of the Decision Tree and SVM models.
- Evaluate model performance on unseen test data.
- Visualize the results, key findings, and provide insights.

This task is about classifying a large set of data based on a set of pre-classified samples.

Your task: is to predict whether a passenger survived the Titanic shipwreck or not. You will use both a Decision Tree Classifier and a Random Forest Classifier to do this and compare the results. The general steps are data exploration and analysis, data pre-processing and transformation
(handling missing values, converting categorical features into numeric, converting discrete features into binary, etc.), and implementing your classifier.
The classic Titanic dataset provides information on the fate of passengers on the Titanic, summarised according to economic status (class), sex, age, and survival.

You will find two data files:
• Training set (titanic_train.csv) should be used to build your ML models.
• Test set (titanic_test.csv) should be used to see how well your model performs on unseen data.

Data Description and Notes:

- Pclass: A proxy for Socio-Economic Status (SES).
    - 1st = Upper
    - 2nd = Middle
    - 3rd = Lower
- Age: Age in years. It is fractional if less than 1. If the age is estimated, it is in the form of xx.5.
- SibSp: The number of siblings/spouses aboard the Titanic. The dataset defines family relations in this way:
- Sibling = brother, sister, stepbrother, stepsister
- Spouse = husband, wife (mistresses and fiancés were ignored)
- Parch: The number of parents/children aboard the Titanic. The dataset defines family relations in this way:
- Parent = mother, father
- Child = daughter, son, stepdaughter, stepson
- Some children travelled only with a nanny, therefore Parch = 0 for them.
- Embarked: The port of embarkation, C = Cherbourg, Q = Queenstown, S = Southampton.
- Ticket: The ticket numbers.

- Fare: The passenger fare.
- Cabin: The cabin number.

Submission:

Upload **Yourusername_Lab2.ipynb** file on Learn.

Main Python libraries to use:

- scikit-learn (a Python library that features various classification, regression, and clustering algorithms) https://scikit- learn.org/stable/
- pandas https://pandas.pydata.org/docs/
- NumPy https://numpy.org/
- Matplotlib https://matplotlib.org/
- seaborn: statistical data visualisation https://seaborn.pydata.org/

"You can have data without information, but you cannot have information without data."

~ Daniel Keys Moran