**PYTHON DATA MANIPULATION**

This is a CT7201 class assessment that focuses on data manipulation, exploration, analysis, machine learning models and prediction on the Titanic dataset from Kaggle using python as the primary programming language completed in a Jupyter notebook.

The result is using machine learning to create a model that predicts which group of passengers survived the infamous Titanic shipwreck.

**Files Included Within This Project**

* titanic.csv
* assignment\_CT7201.ipynb
* assignment\_CT7201.pdf

**Libraries Used Within This Project**

**Pandas**

This is a python package providing fast, flexible, and expressive data structures built to make working with relational data intuitive [1]

Uses within the Project

This was used to

* Load, merge, join and reshape the Titanic csv file data set
* Handle missing data (represented as NaN) in floating point as well as non-floating point data.
* Aggregate and transform data by using the group by functionality to perform split-apply-combine operations on the data set.

**NumPy**

This is a python library that provides a multidimensional array objects, matrices and an assortment of routines such as mathematical, logical, shape manipulation for fast operations on arrays[2].

Uses within the Project

* Random simulation/generation on the data set.

**Matplotlib**

A comprehensive library for creating static, animated, and interactive visualizations in Python[3]

Uses within the Project

* Visualize the data exploratory analysis on age, passenger class and ports embarked within the Titanic dataset.

**Seaborn**

A library built on top of Matplotlib and integrates closely with Pandas data structure which helps to explore and understand the data set through visualization[4].

Its plotting functions operates on data frames and arrays containing whole datasets and internally perform the necessary semantic mapping to produce informative plots.

Uses within the Project

* Visualize and plot the timeline of events on the titanic voyage and data exploratory analysis on age, passenger class and ports embarked within the Titanic dataset.

**Scikit-learn**

This is a machine learning library for the python programming language for predictive data analysis built on NumPy and Matplotlib[5]

Uses within the Project

* Leveraged on its classifications, regression, and clustering algorithms to predict which passengers survived the titanic shipwreck

**SETUP TASK**

* Creation of the environment
* Imported the dataset using the Pandas library
* Imported the libraries used for exploration and machine learning such as Pandas, NumPy and scikit-learn

**DATASET**

The data set used for this project it’s the titanic onboarding dataset which includes passenger information like name, age, gender, socio-economic class, port of onboarding and it can be obtained [here](https://www.kaggle.com/c/titanic/data) or pasting this url - ( <https://www.kaggle.com/c/titanic/data> ) on a web browser.

**REFERENCES**

[1] Pandas-docs, "Pandas Documentation." [Online]. Available: <https://pandas.pydata.org/pandas-docs/stable/getting_started/index.html>.

[2] Numpy, "Numpy documentation reference." [Online]. Available: <https://numpy.org/doc/stable/reference/index.html#reference>.

[3] Matplotlib, "Matplotlib Refernce." [Online]. Available: <https://matplotlib.org/stable/api/index>.

[4] Seaborn, "Seaborn Reference." [Online]. Available: <https://seaborn.pydata.org/index.html>.

[5] S. Learn, "Scikit-learn Library." [Online]. Available: <https://scikit-learn.org/stable/index.html>.