

**The Superior University Lahore**

**Faculty of Computer Science & Information**

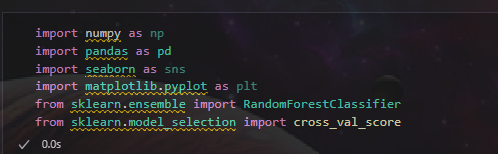
**Technology**

**Name: Alishba Haroon**

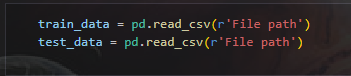
**Roll No: BSAI-116-4C**

**Date: 17 Feb 2025**

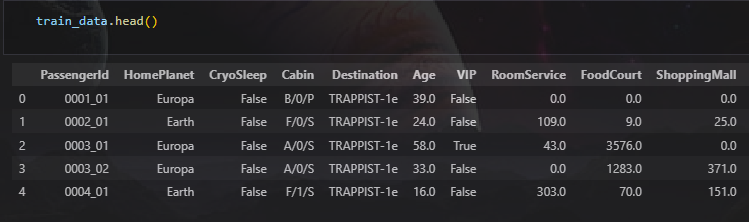
**Subject: PAI LAB**



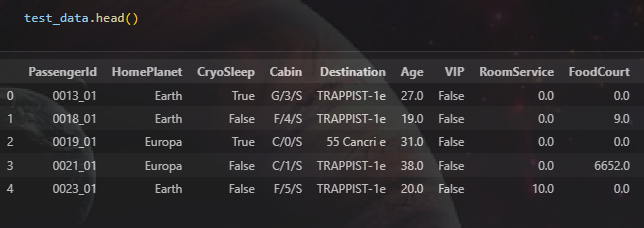
* **numpy (np):** Used for numerical operations and array handling.
* **pandas (pd):** Helps handle and analyze tabular data (DataFrames).
* **seaborn (sns):** A visualization library built on top of matplotlib, used for creating attractive statistical plots.
* **matplotlib.pyplot (plt):** A widely used library for creating static, animated, and interactive visualizations.
* **RandomForestClassifier:** A machine learning algorithm that operates by constructing multiple decision trees and averaging their predictions (ensemble learning).
* **cross\_val\_score:** A cross-validation function, helps evaluate model performance by splitting data into different training/testing sets.



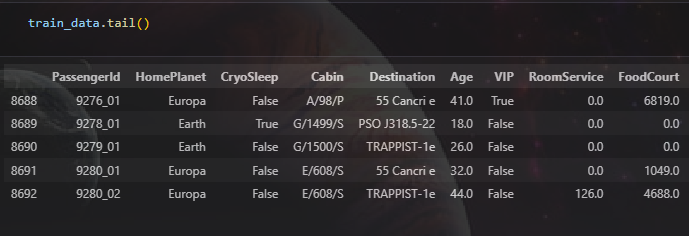
* pd.read\_csv(r'File path') loads **CSV data into a Pandas DataFrame**.



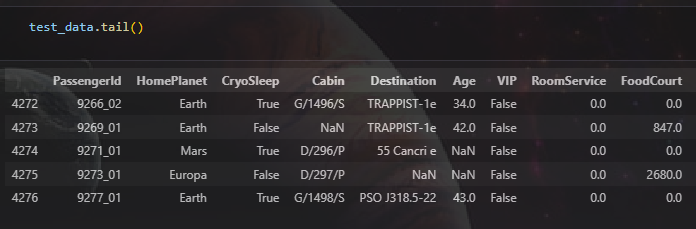
* train\_data.head() shows top first 5 rows



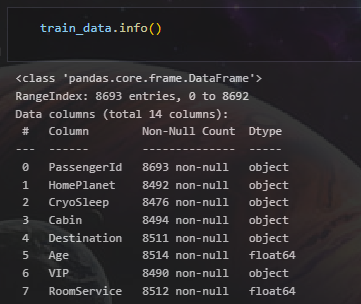
* test\_data.head() show top first 5 rows



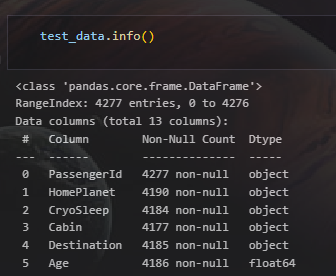
* train\_data.tail() show last 5 rows

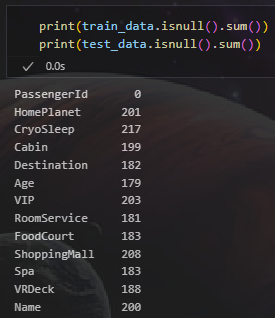


* test\_data.tail() show last 5 rows

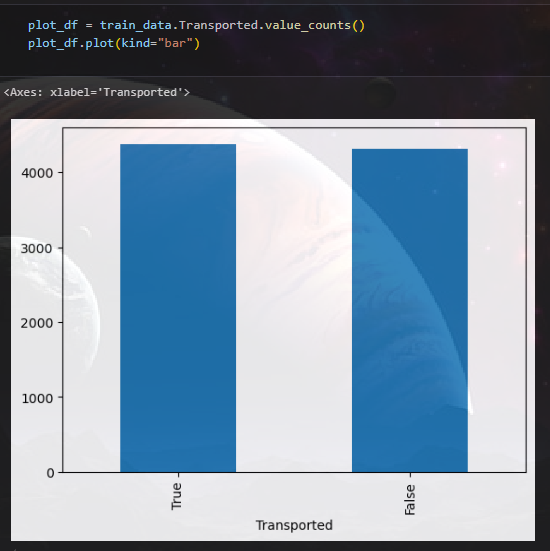


* train\_data.info() shows the dataset's structure, data types, and missing values.

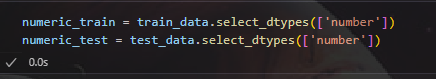




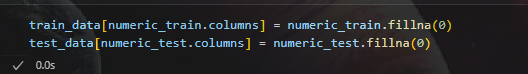
* isnull().sum() counts the number of missing (NaN) values in each column.



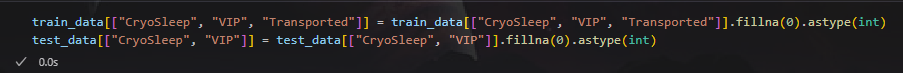
* value\_counts() counts occurrences of each category (e.g., True/False if binary).
* plot(kind="bar") creates a **bar chart** to show the frequency of each category.



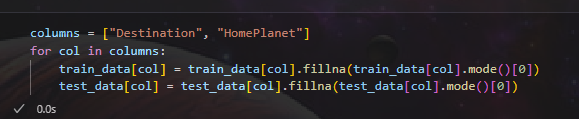
* select\_dtypes(['number']) filters out non-numeric data (like text or categorical values).



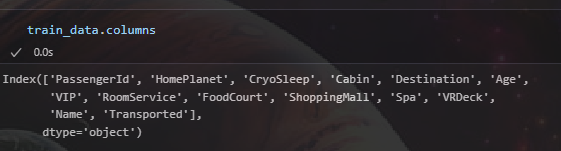
* fillna(0) replaces all NaN values with 0.



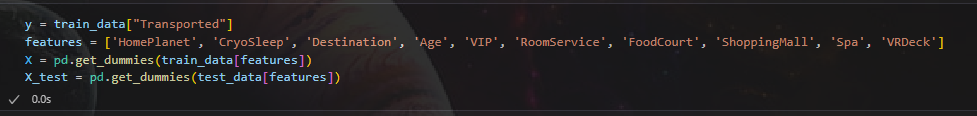
* astype(int): Converts boolean-like categorical values (True/False or NaN) into 0 and 1.



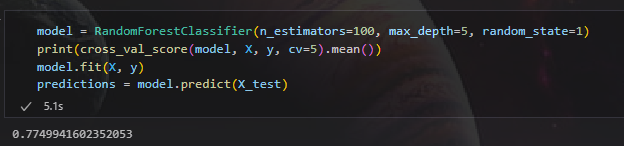
* mode()[0] gets the most common value in each column.
* fillna() replaces missing values with that mode.



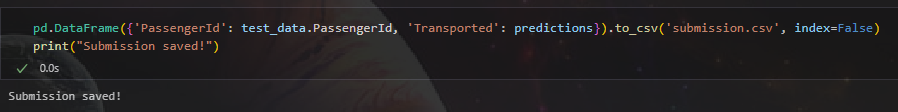
* Show all column name in train dataset.



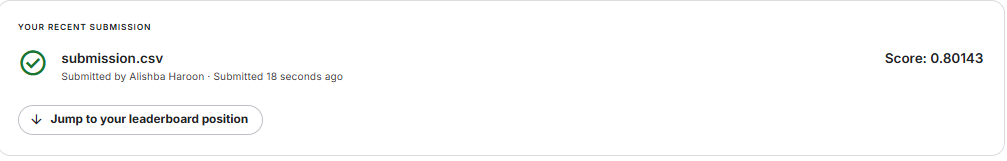
* This selects features, sets the target variable, and applies **one-hot encoding** to categorical data.

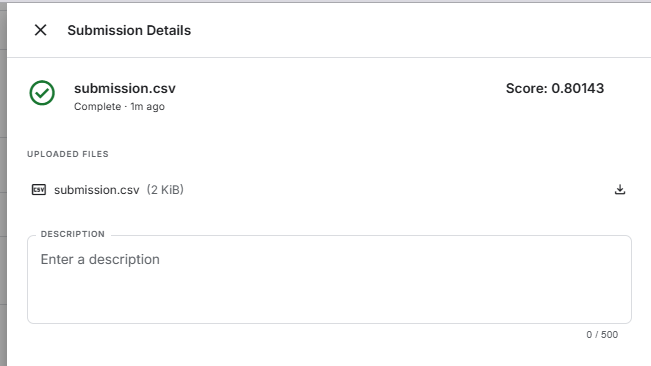


* **RandomForestClassifier():** Creates a Random Forest model with 100 trees and max depth 5.
* **cross\_val\_score().mean()**: Performs 5-fold cross-validation and prints the average score.
* **model.fit(X, y)**: Trains the model on the dataset.
* **model.predict(X\_test)**: Makes predictions on the test data.



* **Creates a DataFrame** with PassengerId and predicted Transported values.
* **Saves it as submission.csv** using to\_csv(), without the index.
* **Prints confirmation** that the file is saved.

****

****