

**The Superior University**

***Session 2023-2025***

***Department of Software Engineering***

***Faculty of Computer Science & Information Technology***

***The Superior University, Lahore***

***Course: Programming For Artificial Intelligence***

***Course Instructor:Sir Rasikh ali***

***Semester 4 BSAI***

***Fall 2025***

***Name:Sara hanif***

***Roll no:082***

***LAB TASK 1***

# Introduction

The Spaceship Titanic dataset is a Kaggle competition dataset that involves predicting whether passengers were transported to another dimension. This report explains the structure of the Jupyter Notebook, the terms used, and its applications.

# Overview of the Notebook

* Data loading and preprocessing
* Exploratory Data Analysis (EDA)
* Feature engineering
* Model training using machine learning techniques
* Predictions and evaluation

# Key Terms and Their Definitions

1. **NumPy (numpy):** A library for numerical computations in Python, particularly useful for handling arrays and mathematical operations.
2. **Pandas (pandas):** A library for data manipulation, allowing users to read, modify, and analyze tabular data.
3. **Operating System Module (os):** A Python module used to interact with the operating system, such as reading files from directories.
4. **CSV (pd.read\_csv):** A file format used to store tabular data, which can be loaded into a DataFrame using pandas.
5. **Exploratory Data Analysis (EDA):** A process used to understand the data through visualizations, statistical summaries, and data cleaning.
6. **Feature Engineering:** The process of creating new input features from existing data to improve model performance.
7. **Machine Learning (tensorflow, tensorflow\_decision\_forests):** Libraries used for training and implementing predictive models, including decision forests and neural networks.
8. **Classification Task:** A type of machine learning problem where the goal is to assign categories (e.g., predicting whether a passenger was transported or not).
9. **Model Training**: The process of teaching a machine learning model to recognize patterns in the dataset.
10. **Evaluation Metrics:** Methods used to measure the performance of the model, such as accuracy, precision, recall, and F1-score.

# *Code Explanation*

## 1. Importing Libraries

Essential Python libraries such as NumPy, Pandas, OS, TensorFlow, and TensorFlow Decision Forests are imported to handle data processing, machine learning, and system operations.

## 2. Loading the Dataset

The dataset is loaded from the Kaggle environment using pd.read\_csv(), allowing further analysis and processing.

## 3. Data Exploration and Preprocessing

* Checking missing values and handling them using techniques like imputation.
* Converting categorical features into numerical formats for machine learning models.
* Scaling numerical features for better model performance.

## 4. Feature Engineering

* Extracting meaningful features from existing data to enhance model predictions.
* Creating new columns based on domain knowledge.

## 5. Model Selection and Training

* Implementing machine learning models using TensorFlow and Decision Forests.
* Training models on the processed dataset.
* Tuning hyperparameters to improve model accuracy.

## 6. Model Evaluation

* Using classification metrics such as accuracy, precision, recall, and F1-score to assess the model’s performance.
* Visualizing results using plots and confusion matrices.

## 7. Prediction Process

* The trained model is used to make predictions on new or test data.
* The test dataset is preprocessed in the same way as the training data to ensure consistency.
* The model’s predict() function is applied to generate predictions.
* Predictions are typically stored in a new DataFrame and may be saved as a CSV file for submission.
* **Example:**
* predictions = model.predict(test\_data)
* submission = pd.DataFrame({'PassengerId': test\_data['PassengerId'], 'Transported':

predictions})

* submission.to\_csv('submission.csv', index=False)
* The output is used to determine the accuracy and effectiveness of the model. **Applications of the Notebook**
* **Predictive Analytics:** Understanding trends and making forecasts based on historical data.
* **Data Cleaning & Preprocessing:** Handling missing values, encoding categorical data, and preparing data for modeling.
* **Model Deployment:** Using trained models in real-world applications such as customer segmentation and recommendation systems.
* **Feature Engineering:** Extracting meaningful information from raw datasets to improve prediction accuracy.
* **Deep Learning Implementation:** Applying advanced machine learning techniques to improve model accuracy.

## Conclusion

The Spaceship Titanic notebook demonstrates a structured approach to solving a classification problem using machine learning. The notebook covers data processing, exploratory analysis, feature engineering, and predictive modeling using TensorFlow. Understanding these techniques is essential for data science projects involving structured datasets.



