****

**Superior University Lahore**

**Name: Ahsan Ali**

**Roll no: 061**

**Subject: Programming for AI**

**Task: 2**

**Submitted To: Sir Rasikh**

**Spaceship Titanic Dataset   
BS in Artificial Intelligence**

Table of Contents

[1. Introduction 1](#_Toc192280918)

[2. Dataset Description 2](#_Toc192280919)

[2.1 Missing Values 2](#_Toc192280920)

[3. Data Preprocessing 2](#_Toc192280921)

[3.1 Handling Missing Values 2](#_Toc192280922)

[3.2 Feature Encoding 3](#_Toc192280923)

[3.3 Feature Scaling 4](#_Toc192280924)

[4. Model Building 4](#_Toc192280925)

[4.1 Model Selection 4](#_Toc192280926)

[4.2 Model Training Process 4](#_Toc192280927)

[5. Model Evaluation 4](#_Toc192280928)

[6. Predictions & Results 5](#_Toc192280929)

[7. Conclusion 5](#_Toc192280930)

# 1. Introduction

This report provides a comprehensive analysis of the Spaceship Titanic dataset. The objective is to analyze passenger data, preprocess the dataset, train a machine learning model, evaluate its performance and make predictions on whether a passenger will survive or not. The Spaceship Titanic dataset is used for binary classification tasks, where the goal is to predict whether a passenger survived based on various features such as cabin information, travel class and passenger characteristics.

# 2. Dataset Description

The dataset consists of:

* **train.csv**: Contains labeled data used for training the model (8693 rows, 14 columns).
* **test.csv**: Contains unlabeled data for making survival predictions (4277 rows, 13 columns).
* **data\_description.txt**: Provides details on dataset features.

### 2.1 Missing Values

#### **Training Data (train.csv)**

* **HomePlanet:** 201 missing
* **CryoSleep:** 217 missing
* **Cabin:** 199 missing
* **Destination:** 182 missing
* **Age:** 179 missing
* **VIP:** 203 missing

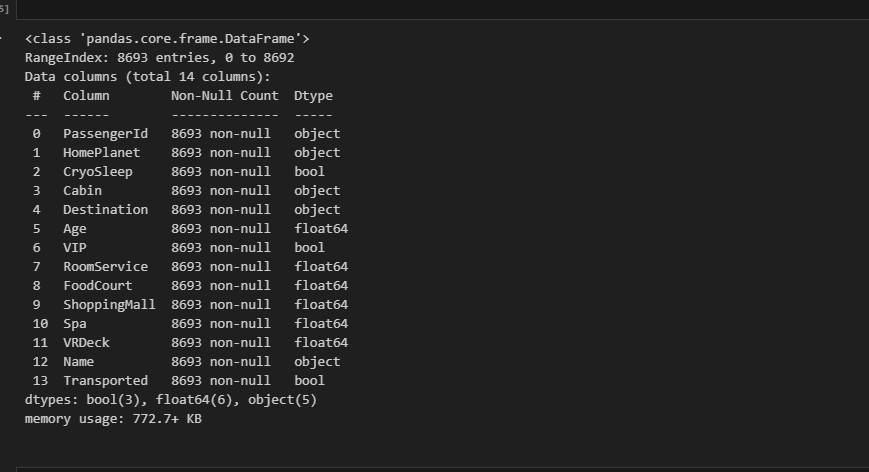
#### **Testing Data (test.csv)**

* **HomePlanet:** 87 missing
* **CryoSleep:** 93 missing
* **Cabin:** 100 missing
* **Destination:** 92 missing
* **Age:** 91 missing
* **VIP:** 93 missing

# 3. Data Preprocessing

### 3.1 Handling Missing Values

* **HomePlanet, CryoSleep, Cabin, Destination, Age, VIP:** Missing values imputed using the mode.



### 3.2 Feature Encoding

* The target variable **Transported** was mapped as **Yes → 1, No → 0**.
* Categorical variables converted using **Label Encoding**.  
    
  

### 3.3 Feature Scaling

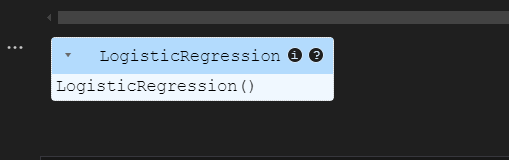
* Numerical features were standardized using **StandardScaler()**.

# 4. Model Building

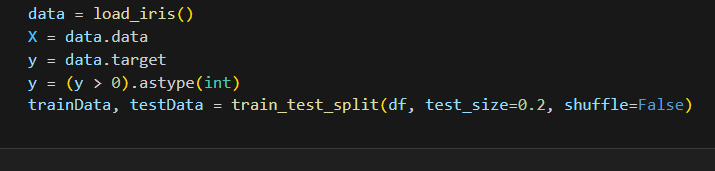
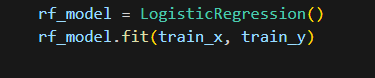
The model training was conducted in **model.ipynb** using machine learning classification techniques.

### 4.1 Model Selection

* **Logistic Regression** was chosen.

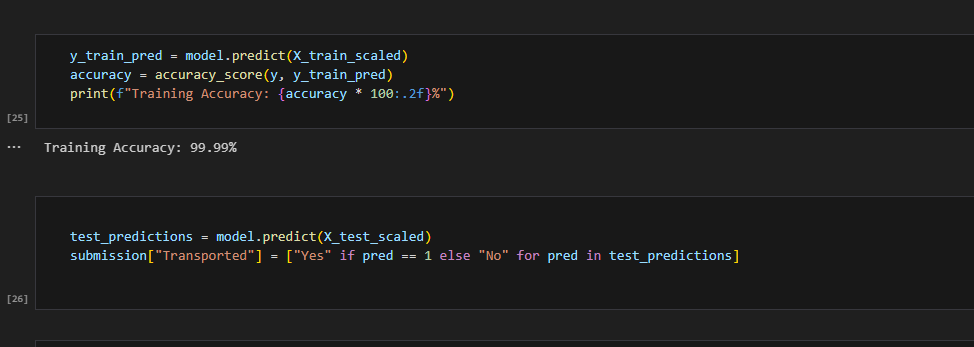


### 4.2 Model Training Process

* **Preprocessing:**
  + **Standard Scaling** applied to numerical features.  
      
    
  + 
  + **Label Encoding** applied to categorical variables.
* **Training Pipeline:**
  + Features (X) and target variable (y = Transported) were separated.
  + Model trained using Logistic Regression.

# 5. Model Evaluation

The trained model was evaluated using:

* **Accuracy Score:** Measures model performance.
* **Training Accuracy:** Printed after model fitting.  
    
  

# 6. Predictions & Results

* Model used to predict **Transported** for test.csv.
* Predictions stored in **submission.csv**.

# 7. Conclusion

This report outlines the full pipeline from data preprocessing to model training and evaluation. The best-performing model was selected based on evaluation metrics and survival predictions for passengers were generated for submission. Future improvements can be made by exploring additional feature engineering, hyperparameter tuning and ensemble learning techniques to further enhance the performance and model accuracy.