A Project Report On

**Machine Learning (Mini Project II)**

**CLASS: BE-2**

GUIDED BY:

**Dr. Prajakta Khadkikar**



**DEPARTMENT OF COMPUTER ENGINEERING**

PUNE INSTITUTE OF COMPUTER TECHNOLOGY DHANKAWADI, PUNE-43

SAVITRIBAI PHULE PUNE UNIVERSITY 2020-21

Rohit James – 41266

Sufiya Sayyed - 41278

Aditya Wanjale - 41281

**Title:**

Logistic Regression

**Problem Statement:**

Build a machine learning model that predicts the type of people who survived the Titanic shipwreck using passenger data (i.e. name, age, gender, socio-economic class, etc.).

Dataset Link: https://www.kaggle.com/competitions/titanic/data

**Objectives:**

To predict the type of people who survived the Titanic shipwreck using the titanic dataset.

**Development Environment:**

Software – Python, Jupyter Notebook

Hardware – 64-bit Windows-10 OS

**Libraries Used:**

Pandas – used to import the training and testing datasets from the .csv files

Numpy – used for mathematical operations and computations

matplotlib.pyplot – used to plot various plots

seaborn - used for data visualization

sklearn – used for machine learning task

**Theory:**

1. Titanic dataset

The titanic.csv file contains data for 891 of the real Titanic passengers. Each row represents one person. The columns describe different attributes about the person including whether they passengerId, survived (S), their name, class, age (A), their passenger-class (C), their sex (G) and the fare they paid (X).

1. Logistic Regression –

This type of statistical model (also known as logit model) is often used for classification and predictive analytics. Logistic regression estimates the probability of an event occurring, such as voted or didn’t vote, based on a given dataset of independent variables. Since the outcome is a probability, the dependent variable is bounded between 0 and 1. In logistic regression, a logit transformation is applied on the odds—that is, the probability of success divided by the probability of failure.

1. Boxplots –

A boxplot is a standardized way of displaying the distribution of data based on a five number summary (“minimum”, first quartile [Q1], median, third quartile [Q3] and “maximum”). It tells about the outliers and what their values are.

1. Barplots –

A barplot is basically used to aggregate the categorical data according to some methods and by default it’s the mean. It can also be understood as a visualization of the group by action. To use this plot we choose a categorical column for the x-axis and a numerical column for the y-axis, and we see that it creates a plot taking a mean per categorical column

1. Correlation Matrix –

A correlation matrix is a table containing correlation coefficients between variables. Each cell in the table represents the correlation between two variables. The value lies between -1 and 1. A correlation matrix is used to summarize data, as a diagnostic for advanced analyses and as an input into a more advanced analysis. The two key components of the correlation are:

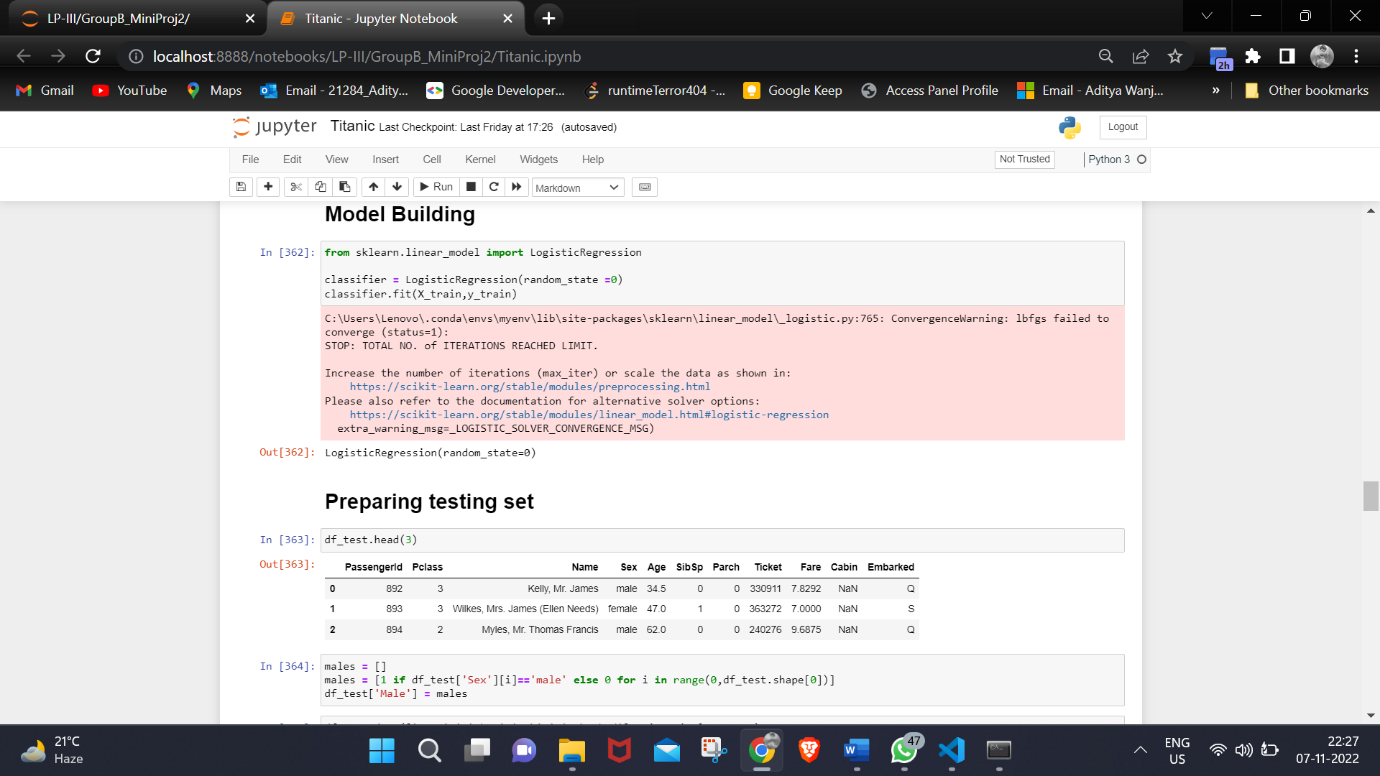
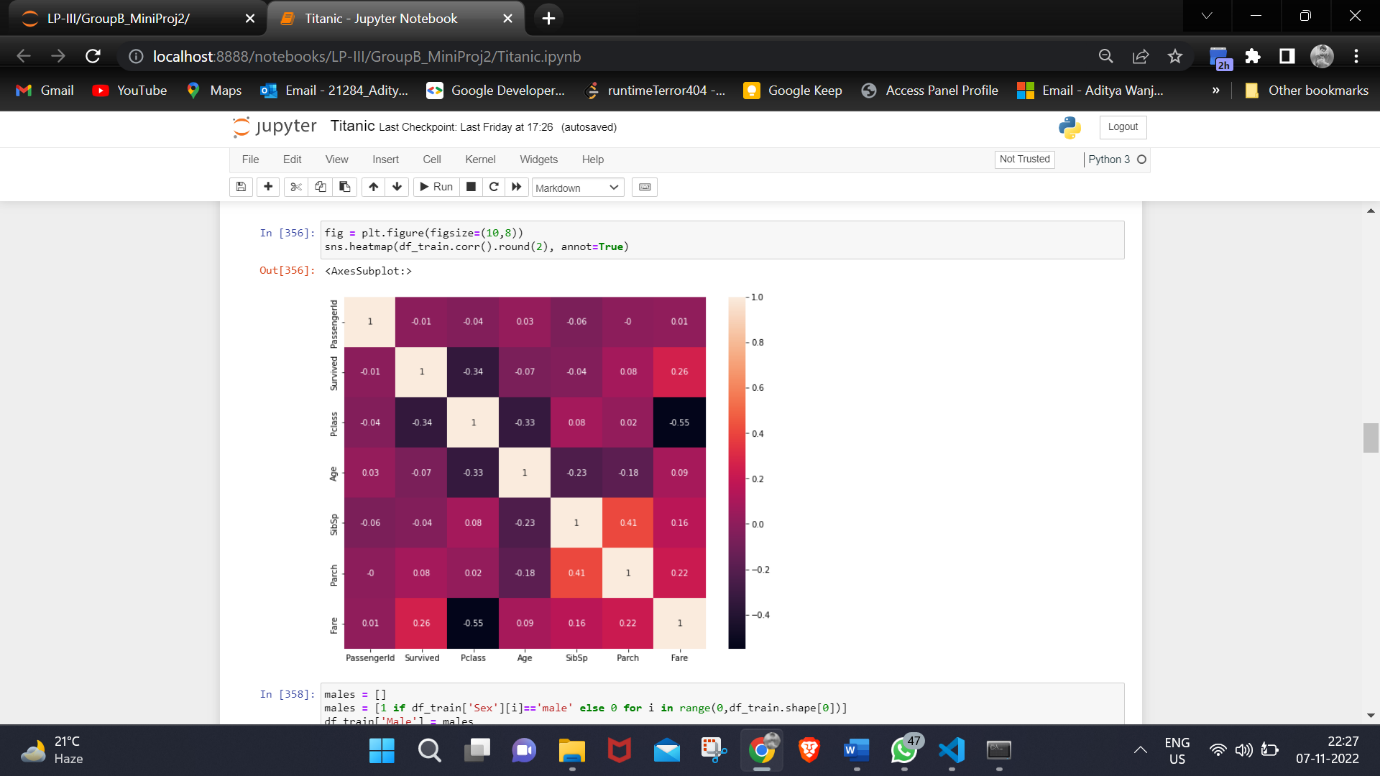
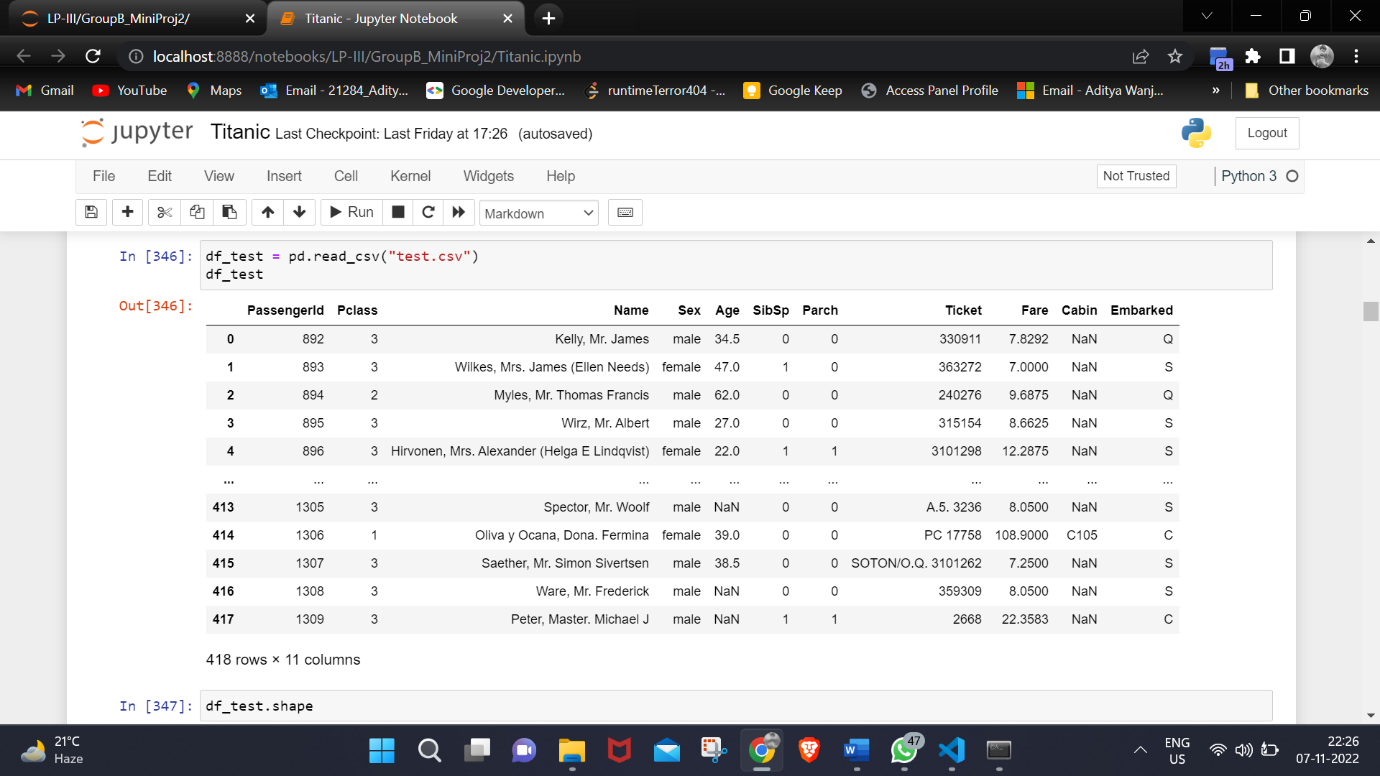
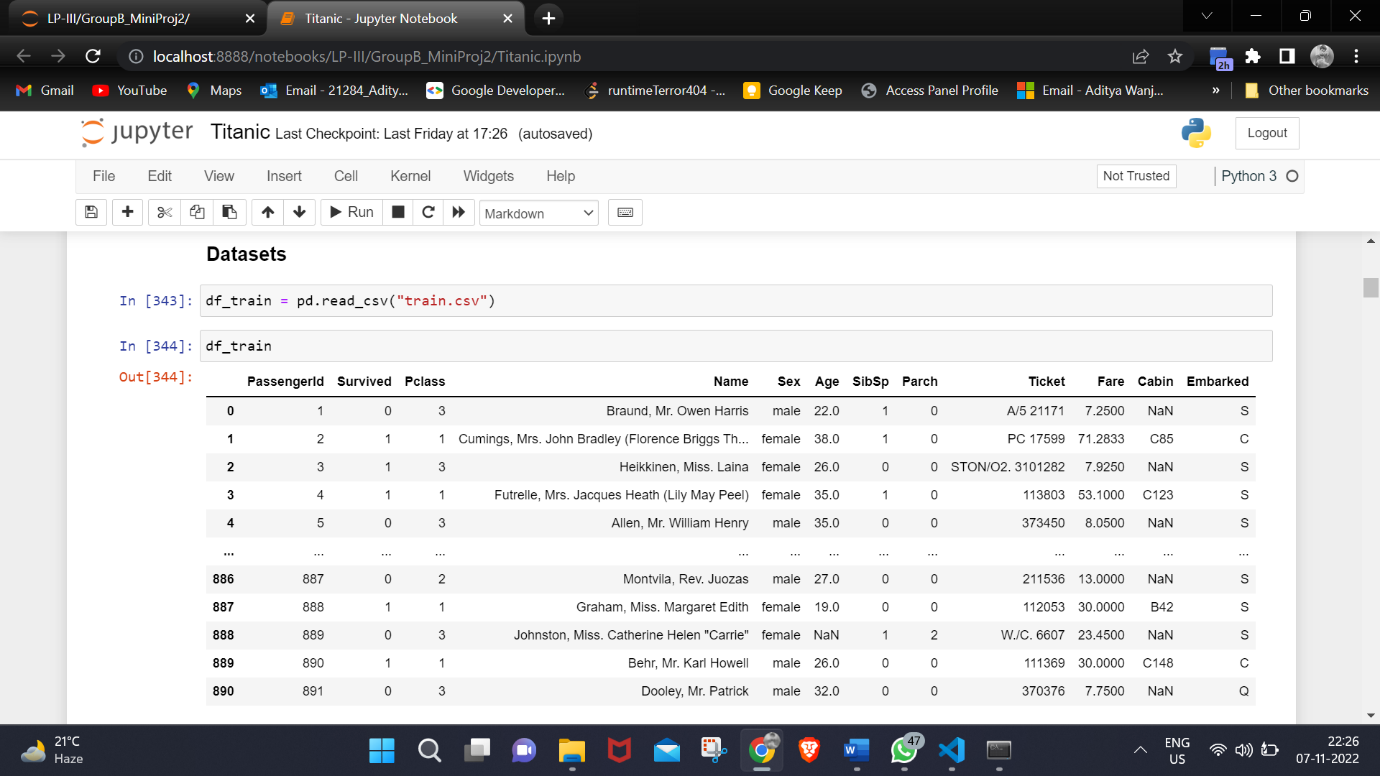
Magnitude: larger the magnitude, stronger the correlation.

Sign: if positive, there is a regular correlation. If negative, there is an inverse correlation.

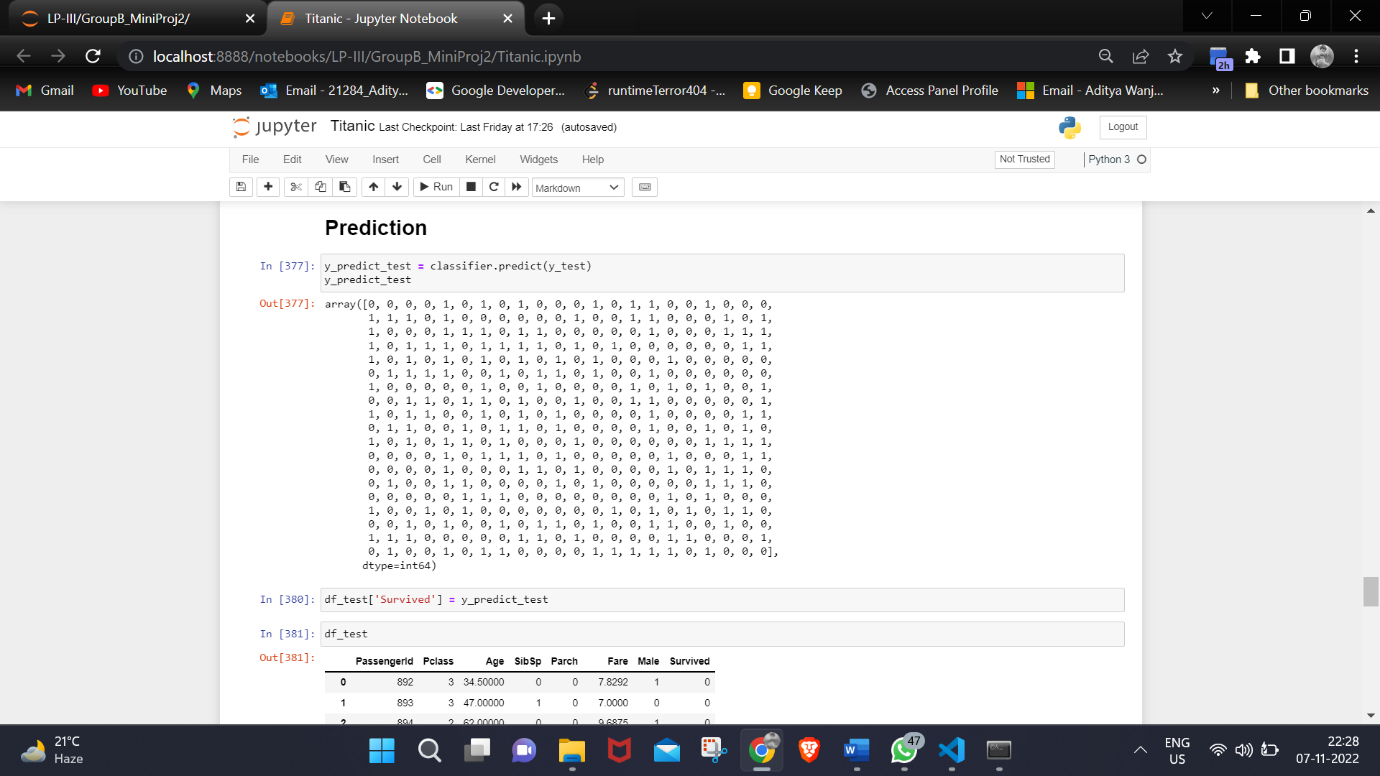
**Steps of implementation:**

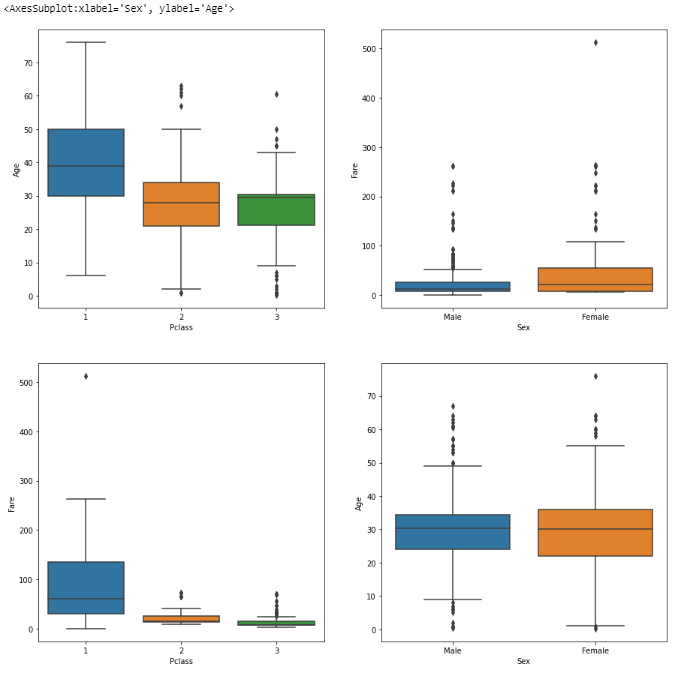
1. Import necessary libraries
2. Import training and testing dataset
3. Pre-process and clean the training and testing dataset
4. Remove outliers and null values
5. Visualize training dataset
6. Find correlation matrix
7. Build logistic regression model
8. Predict the type of people of who survived from the testing data
9. Visualize the predicted results

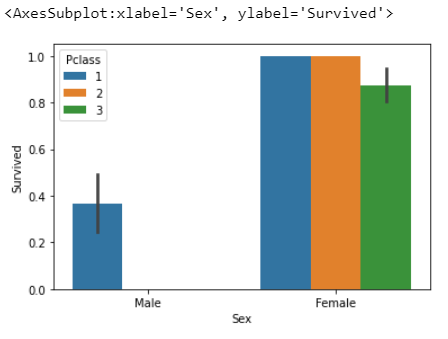
**Sample Screenshots:**

****

**Output:**

****

****

****

**Conclusion:**

Hence, we built a machine learning model that predicts the type of people that survived on the Titanic shipwreck.