SCTR's Pune Institute of Computer Technology
Dhankawadi, Pune

A MINI PROJECT REPORT ON

Titanic Survival Prediction using Machine Learning

Under the guidance of
Prof. Sumit Shevtekar

DEPARTMENT OF COMPUTER ENGINEERING
ACADEMIC YEAR 2022-23

Title:
Titanic Survival Prediction using Machine Learning

Team:

1. Gouri Gupta - 41227
2. Sakshi Harode- 41228
3. Pratik Kadale – 41235

Abstract

This project is based on the Titanic dataset given on Kaggle. The sinking of the Titanic is one of the most infamous shipwrecks in history. On April 15, 1912, the widely considered “unsinkable” Titanic sank after colliding with an iceberg. Unfortunately, there weren’t enough lifeboats for everyone on board, resulting in the **death.** In this project, we see how we can use machine-learning techniques to predict survivors of the Titanic. With a dataset of 891 individuals containing features like sex, age, and class, we attempt to predict the survivors of a small test group of 418. We are using Logistic Regression Model for the same.

Introduction

Machine learning means the application of any computer-enabled algorithm that can be applied against a data set to find a pattern in the data. This encompasses basically all types of data science algorithms, supervised, unsupervised, segmentation, classification, or regression". few important areas where machine learning can be applied are Handwriting Recognition, Language Translation, Speech Recognition, Image Classification, Autonomous Driving. Some features of machine learning algorithms can be observations that are used to form predictions for image classification, the pixels are the features, For voice recognition, the pitch and volume of the sound samples are the features and for autonomous cars, data from the cameras, range sensors, and GPS.
Using data provided by www.kaggle.com, our goal is to apply machine-learning techniques to successfully predict which passengers survived the sinking of the Titanic. Features like ticket price, age, sex, and class will be used to make the predictions. Using Logistic Regression methods, we try to predict the survival of passengers using different combinations of features. The challenge boils down to a classification problem given a set of features.

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

Motivation

To predict what type of people survived the Titanic Shipwreck using passanger data and build its prediction model is the main motive to study this mini project.

Objective

Goal: Build a predictive model that answers the question: “what sorts of people were more likely to survive?” using passenger data like age, gender, class, etc.

Theory

Data Set :

The data we used for our project was provided on the Kaggle website. We were given 891 passenger samples for our training set and their associated labels of whether or not the passenger survived. For each passenger, we were given his/her passenger class, name, sex, age, number of siblings/spouses aboard, number of parents/children aboard, ticket number, fare, cabin embarked, and port of embarkation.
For the test data, we had 418 samples in the same format. The dataset is not complete, meaning that for several samples, one or many of fields were not available and marked empty (especially in the latter fields – age, fare, cabin, and port). However, all sample points contained at least information about gender and passenger class.
To normalize the data, we replace missing values with the mean of the remaining data set or other values.

Understanding the Titanic Dataset

So first we will understand our titanic dataset. This is a dataset of Titanic ship passengers & here

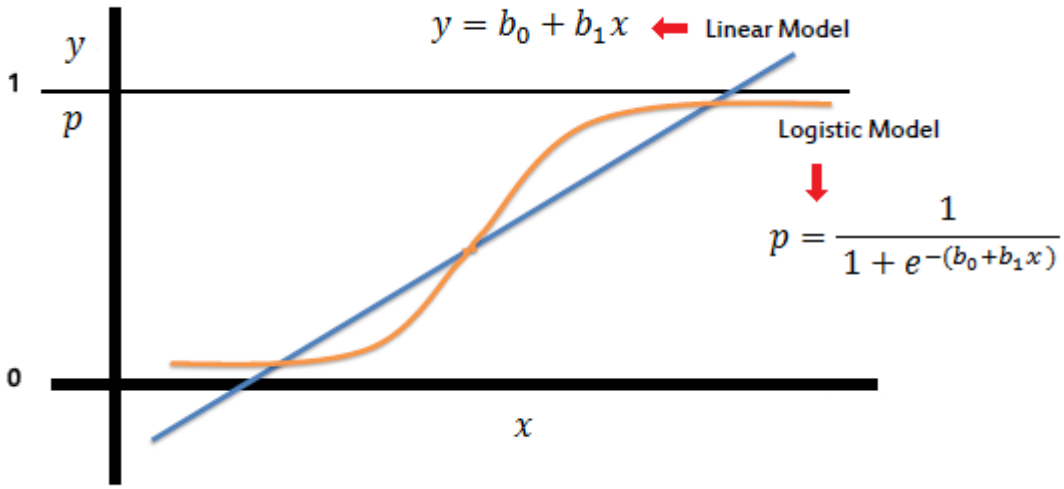
- Each row represents the data of 1 passenger.
- Columns represent the features. We have 10 features/ variables in this dataset.

1. **Survival:** This variable shows whether the person survived or not. This is our target variable & we have to predict its value. It's a binary variable. *0 means not survived and 1 means survived.*
2. **pclass:** The ticket class of passengers. 1st (upper class), 2nd (middle), or 3rd (lower).
3. **Sex:** Gender of passenger
4. **Age:** Age (in years) of a passenger
5. **sibsp:** The no. of siblings/spouses of a particular passenger who were there on the ship.
6. **parch:** The no. of parents/children of a particular passenger who were there on the ship.
7. **ticket:** Ticket Number
8. **fare:** Passenger fare (like 1st class ticket fare must be greater than 2nd pr 3rd class ticket right)
9. **cabin:** Cabin Number
10. **embarked:** Port of Embarkation; From where that passenger took the ship. (C = Cherbourg, Q = Queenstown, S = Southampton)

Logistic Regression:

A simple yet crisp description of Logistic Description would be, "it is a supervised learning classification algorithm used to predict the probability of a target variable. The nature of target or dependent variable is dichotomous, which means there would be only two possible classes." as stated in the tutorial points article.

The graph of logistic regression is as shown below:



What is Training Dataset?

The *training data is the biggest (in -size) subset of the original dataset, which is used to train or fit the machine learning model*. Firstly, the training data is fed to the ML algorithms, which lets them learn how to make predictions for the given task.

What is Test Dataset?

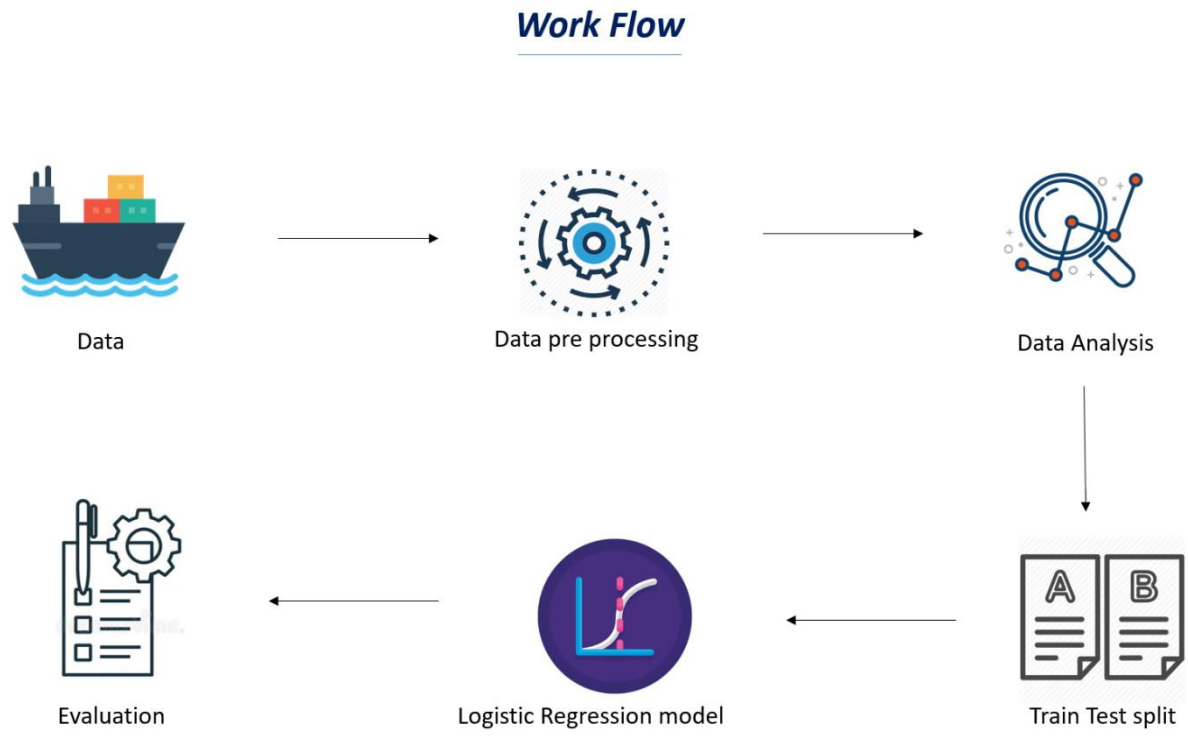
Once we train the model with the training dataset, it's time to test the model with the test dataset. This dataset evaluates the performance of the model and ensures that the model can generalize well with the new or unseen dataset. *The test dataset is another subset of original data, which is independent of the training dataset*. However, it has some similar types of features and class probability distribution and uses it as a benchmark for model evaluation once the model training is completed. Test data is a well-organized dataset that contains data for each type of scenario for a given problem that the model would be facing when used in the real world. Usually, the test dataset is approximately 20-25% of the total original data for an ML project.

Accuracy

To find the accuracy of model in confusion matrix the formula is :

$$accuracy = \frac{true\ positives + true\ negatives}{true\ positives + true\ negatives + false\ positives + false\ negatives}$$

Work Flow



RESULT :

Project 1: . Titanic Survival Prediction using Machine Learning.ipynb

File Edit View Insert Runtime Tools Help Changes will not be saved

Share

+ Code + Text Copy to Drive

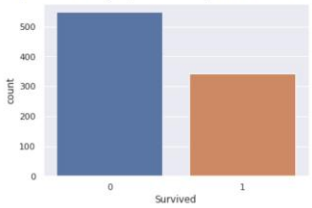
Reconnect Editing

Data Visualization

```
[ ] sns.set()
```

```
# making a count plot for "Survived" column
sns.countplot('Survived', data=titanic_data)
```

```
/usr/local/lib/python3.7/dist-packages/seaborn/_decorators.py:43: FutureWarning: Pass the following variable as a keyword arg: x. From version 0.12, the only valid positional argument
  FutureWarning
<matplotlib.axes._subplots.AxesSubplot at 0x7fd6c77f16d0>
```



Project 1: . Titanic Survival Prediction using Machine Learning.ipynb

File Edit View Insert Runtime Tools Help Changes will not be saved

Share

+ Code + Text Copy to Drive

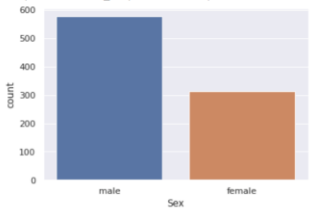
Reconnect Editing

```
[ ] titanic_data['Sex'].value_counts()
```

```
male      577
female    314
Name: Sex, dtype: int64
```

```
# making a count plot for "Sex" column
sns.countplot('Sex', data=titanic_data)
```

```
/usr/local/lib/python3.7/dist-packages/seaborn/_decorators.py:43: FutureWarning: Pass the following variable as a keyword arg: x. From version 0.12, the only valid positional argument
  FutureWarning
<matplotlib.axes._subplots.AxesSubplot at 0x7fd6cbeb1d90>
```



Project 1: Titanic Survival Prediction using Machine Learning.ipynb

File Edit View Insert Runtime Tools Help Changes will not be saved

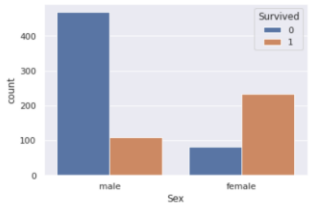
Share

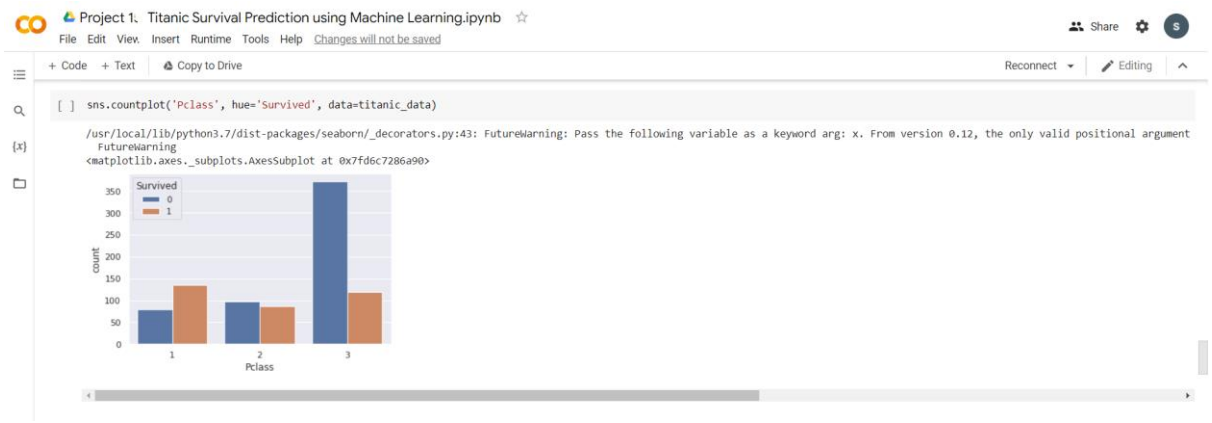
+ Code + Text Copy to Drive

Reconnect Editing

```
[ ] # number of survivors Gender wise
sns.countplot('Sex', hue='Survived', data=titanic_data)
```

```
/usr/local/lib/python3.7/dist-packages/seaborn/_decorators.py:43: FutureWarning: Pass the following variable as a keyword arg: x. From version 0.12, the only valid positional argument
  FutureWarning
<matplotlib.axes._subplots.AxesSubplot at 0x7fd6c77d0dd0>
```





Conclusion

The analysis revealed interesting patterns across individual-level features. Factors such as socioeconomic status, social norms and family composition appeared to have an impact on likelihood of survival. These conclusions, however, were derived from findings in the given data set.
It has been observed that female survival rates are very high (approx 74%) while male survival rates are very low. To make predictions in classification problem, the techniques of logistic regression is primarily used.
It would be interesting to play more with dataset and introducing more attributes which might lead to better results. Various other machine learning techniques like Naive Bayes, K-NN classification can be used to solve the problem.