A MINI-PROJECT REPORT ON

“Titanic Survival Prediction”

SUBMITTED TO THE SAVITRIBAI PHULE PUNE UNIVERSITY, PUNE,IN THE PARIAL FULFILLMENT OF THE REQUIREMENTS FOR SUBJECT

OF

LEARNING PRACTICAL – III

BY

Name: Tanishka Deepak Kadam **Roll No.: BECB126**

Name: Vaishnavi Vivekanand More **Roll No.: BECB140**

Name: Preeti Prakash Pingale **Roll No.: BECB145**

UNDER THE GUIDANCE

OF

PROF. NILESH KAMBLE

Nutan Maharashtra Vidya Prasarak Mandal's

NUTAN MAHARASHTRA INSTITUTE OF ENGINEERING & TECHNOLOGY, PUNE

Samarth Vidya Sankul, Vishnupuri,

Talegaon Dabhade, Maharashtra 410507

CERTIFICATE

This is to certify that the Mini-Project entitled

"Titanic Survival Prediction"

Submitted By

Name: Vaishnavi More

Roll No: BECB140

is a bonafide work carried out by him under the supervision of Prof. Nilesh Kamble and it is approved for the partial fulfillment of the requirement of subject Learning Practical- III

The Mini-Project work has not been earlier submitted to any other institute or university for the award of degree or diploma.

Prof. Nilesh Kamble	**Prof. Rohini Hanchate**	**Dr. Pramod Patil**
(Subject Co-ordinator)	(Head of Department)	(Principal)

Place : Pune

Date :

INDEX

Table of Contents

1. 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

2. Objectives:

- To build a machine learning model that predicts whether a passenger survived the Titanic shipwreck.
- To analyze passenger data such as name, age, gender, and socio-economic class to identify key survival factors.
- To perform data preprocessing (handling missing values, encoding categorical data, normalization, etc.) for better model accuracy.
- To apply supervised learning algorithms like Logistic Regression, Decision Tree, or Random Forest for prediction.
- To evaluate the model's performance using metrics such as accuracy, precision, recall, and F1-score.
- To gain insights into how different features influenced the survival chances of passengers.
- To demonstrate the practical application of machine learning techniques in solving real-world classification problems..

3. Methedology:

3.1 Data Collection

The dataset used for this project is obtained from the Kaggle Titanic Competition**.**

It contains information about passengers such as Name, Age, Gender, Passenger Class (Pclass), Ticket Fare, Cabin, Embarked port, and Survival status (0 = Not Survived, 1 = Survived).

The dataset is divided into two parts:

train.csv – used for training the model.

test.csv – used for evaluating the model's prediction performance.

3.2 Data Preprocessing

Data preprocessing is a crucial step to handle inconsistencies and prepare data for training.

Handling Missing Values: Missing data in columns like Age, Cabin, and Embarked were filled using suitable strategies such as mean/median/mode imputation.

Encoding Categorical Features: Categorical variables (like Sex and Embarked) were converted into numerical format using label encoding or one-hot encoding.

Feature Selection: Unnecessary or non-numerical columns (like Name, Ticket, Cabin) were removed.

Feature Scaling: Normalization or standardization was applied to ensure that all features contribute equally to the model.

3.3 Exploratory Data Analysis (EDA)

Conducted statistical and visual analysis to identify correlations between passenger attributes and survival chances.

Insights found include:

Females had a higher survival rate than males.

Passengers in 1st class had better survival rates compared to those in lower classes.

Younger passengers were more likely to survive.

Visualization tools such as matplotlib and seaborn were used for analysis.

3.4 Model Building

Multiple supervised machine learning algorithms were applied for classification, such as:

Logistic Regression
Decision Tree Classifier
Random Forest Classifier
Support Vector Machine (SVM)

The training dataset was split into training and validation sets to evaluate model performance.

Hyperparameter tuning was performed to improve accuracy.

3.5 Model Evaluation

Model performance was measured using metrics such as:

Accuracy
Precision
Recall
F1-Score

Confusion Matrix and ROC Curve were also used to visualize model results.
The model with the highest accuracy and balanced performance across metrics was selected as the final prediction model.

3.6 Prediction

The trained model was tested on the test dataset to predict the survival status of passengers. Results were submitted to Kaggle for leaderboard evaluation.

4. Implementation:

Titanic dataset is one of the most popular datasets used for understanding machine learning basics. It contains information of all the passengers aboard the RMS Titanic, which unfortunately was shipwrecked. This dataset can be used to predict whether a given passenger survived or not.
The csv file can be downloaded from Kaggle

CODE:

1. PassengerId: Unique Id of a passenger
2. Survived: If the passenger survived(0-No, 1-Yes)
3. Pclass: Passenger Class (1 = 1st, 2 = 2nd, 3 = 3rd)
4. Name: Name of the passenger
5. Sex: Male/Female
6. Age: Passenger age in years
7. SibSp: No of siblings/spouses aboard
8. Parch: No of parents/children aboard
9. Ticket: Ticket Number
10. Fare: Passenger Fare
11. Cabin: Cabin number
12. Embarked: Port of Embarkation (C = Cherbourg; Q = Queenstown; S = Southampton)

Code: Loading data using Pandas

Python3

```
#importing pandas library
import pandas as pd

#loading data
titanic = pd.read_csv('...\input\train.csv')
```

Seaborn:

It is a Python library used to statistically visualize data. Seaborn, built over Matplotlib, provides a better interface and ease of usage. It can be installed using the following command,

pip3 install seaborn

Code: Printing data head

Python3

```
# View first five rows of the dataset
titanic.head()
```

Output :

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S

Code: Checking the NULL values

Python3

```
titanic.isnull().sum()
```

Output:

```
PassengerId      0
Survived         0
Pclass           0
Name             0
Sex              0
Age            177
SibSp            0
Parch            0
Ticket           0
Fare             0
Cabin          687
Embarked         2
dtype: int64
```

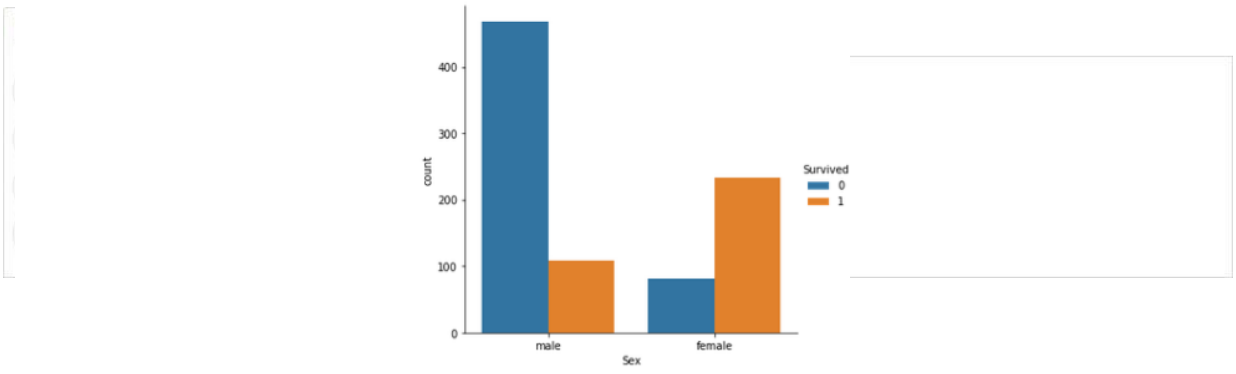
The columns having null values are: Age, Cabin, Embarked. They need to be filled up with appropriate values later on.

Features: The titanic dataset has roughly the following types of features:

- **Categorical/Nominal**: Variables that can be divided into multiple categories but having no order or priority.
  Eg. Embarked (C = Cherbourg; Q = Queenstown; S = Southampton)
- **Binary**: A subtype of categorical features, where the variable has only two categories.
  Eg: Sex (Male/Female)
- **Ordinal**: They are similar to categorical features but they have an order(i.e can be sorted).
  Eg. Pclass (1, 2, 3)
- **Continuous**: They can take up any value between the minimum and maximum values in a column.
  Eg. Age, Fare
- **Count**: They represent the count of a variable.
  Eg. SibSp, Parch
- **Useless**: They don't contribute to the final outcome of an ML model.
  Here, *PassengerId*, *Name*, *Cabin* and *Ticket* might fall into this category.

Code: Graphical Analysis

Output :



Just by observing the graph, it can be approximated that the survival rate of men is around *20%* and that of women is around *75%*. Therefore, whether a passenger is a male or a female plays an important role in determining if one is going to survive.

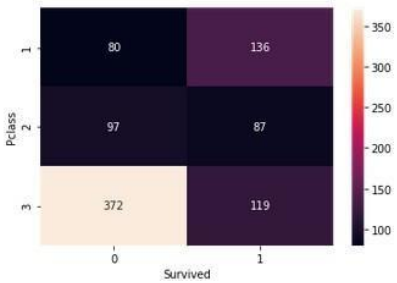
Code : Pclass (Ordinal Feature) vs Survived

Python3

```
# Group the dataset by Pclass and Survived and then unstack them
group = titanic.groupby(['Pclass', 'Survived'])
pclass_survived = group.size().unstack()

# Heatmap - Color encoded 2D representation of data.
sns.heatmap(pclass_survived, annot = True, fmt ="d")
```

Output:



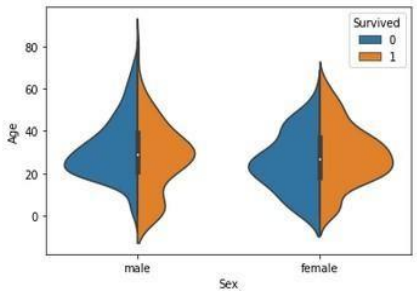
It helps in determining if higher-class passengers had more survival rate than the lower class ones or vice versa. *Class 1* passengers have a higher survival chance compared to *classes 2 and 3*. It implies that *Pclass* contributes a lot to a passenger's survival rate.

Code : Age (Continuous Feature) vs Survived

Python3

```
# Violinplot Displays distribution of data
# across all levels of a category.
sns.violinplot(x ="Sex", y ="Age", hue ="Survived",
data = titanic, split = True)
```

Output :

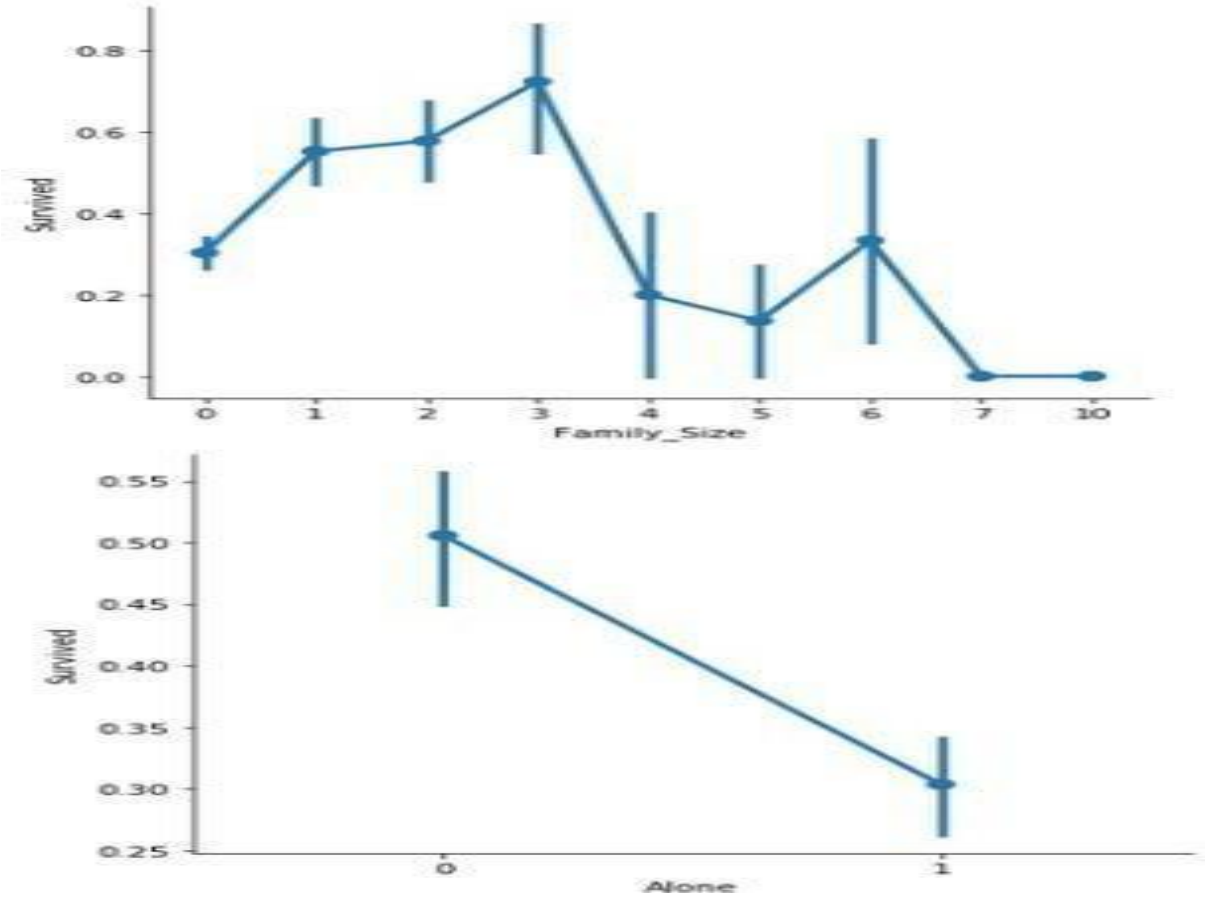


This graph gives a summary of the age range of men, women and children who were saved. The survival rate is –

- Good for children.
- High for women in the age range 20-50.
- Less for men as the age increases.

Since *Age* column is important, the missing values need to be filled, either by using the *Name* column(ascertaining age based on salutation – Mr, Mrs etc.) or by using a regressor.
After this step, another column – *Age_Range* (based on age column) can be created and the data can be analyzed again.

Code : Factor plot for Family_Size (Count Feature) and Family Size.



Family_Size denotes the number of people in a passenger's family. It is calculated by summing the **SibSp** and **Parch** columns of a respective passenger. Also, another column **Alone** is added to check the chances of survival of a lone passenger against the one with a family.
Important observations –

- If a passenger is alone, the survival rate is less.
- If the family size is greater than 5, chances of survival decrease considerably.

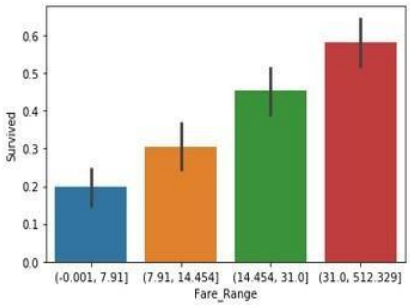
Code : Bar Plot for Fare (Continuous Feature)

Python3

```
# Divide Fare into 4 bins
titanic['Fare_Range'] = pd.qcut(titanic['Fare'], 4)

# Barplot - Shows approximate values based
# on the height of bars.
sns.barplot(x ='Fare_Range', y ='Survived',
data = titanic)
```

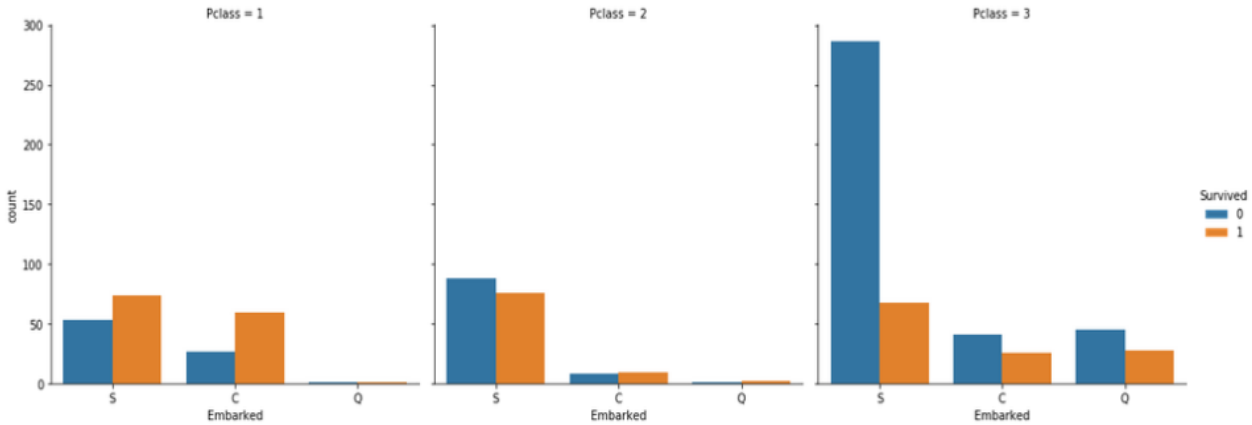
Output :



Fare denotes the fare paid by a passenger. As the values in this column are continuous, they need to be put in separate bins(as done for **Age** feature) to get a clear idea. It can be concluded that if a passenger paid a higher fare, the survival rate is more.

Python3

```
# Countplot
sns.catplot(x ='Embarked', hue ='Survived',
kind ='count', col ='Pclass', data = titanic)
```



Some notable observations are:

- Majority of the passengers boarded from *S*. So, the missing values can be filled with *S*.
- Majority of class 3 passengers boarded from *Q*.
- *S* looks lucky for class 1 and 2 passengers compared to class 3.

Conclusion :

- The columns that can be dropped are:
    - PassengerId, Name, Ticket, Cabin: They are strings, cannot be categorized and don't contribute much to the outcome.
    - Age, Fare: Instead, the respective range columns are retained.
- The titanic data can be analyzed using many more graph techniques and also more column correlations, than, as described in this article.
- Once the EDA is completed, the resultant dataset can be used for predictions.