PROJECT REPORT
TITANIC SURVIVAL PREDICTION MODEL

Author: Apoorva

The RMS Titanic was a British passenger liner that sank in the North Atlantic Ocean in the early morning hours of 15 April 1912, after it collided with an iceberg during its maiden voyage from Southampton to New York City. There were an estimated 2,224 passengers and crew aboard the ship, and more than 1,500 died, making it one of the deadliest commercial peacetime maritime disasters in modern history. The RMS Titanic was the largest ship afloat at the time it entered service and was the second of three Olympic-class ocean liners operated by the White Star Line. The Titanic was built by the Harland and Wolff shipyard in Belfast. Thomas Andrews, her architect, died in the disaster.This sensational tragedy shocked the international community and led to better safety regulations for ships.

Project aim :

Design various Machine Learning models based on a given training data set (the famous Titanic dataset) and test them on the test dataset to predict the survival of passengers, and compare the various models for accuracy.

IDE used: Jupyter Notebook

Dataset link: https://www.kaggle.com/c/titanic/data

Libraries used:

```
import pandas as pd
import numpy as np

import matplotlib.pyplot as plt
import seaborn as sns

from sklearn import metrics
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.preprocessing import StandardScaler
from sklearn.metrics import classification_report
from sklearn.svm import SVC,LinearSVC
from sklearn.ensemble import RandomForestClassifier
from sklearn.neighbors import KNeighborsClassifier
from sklearn.naive_bayes import GaussianNB
from sklearn.linear_model import Perceptron
from sklearn.linear_model import SGDClassifier
from sklearn.tree import DecisionTreeClassifier
from sklearn.model_selection import GridSearchCV , cross_val_score

from sklearn.metrics import plot_confusion_matrix
from sklearn.metrics import precision_recall_curve
```

Data set Information:

The data has been split into two groups:

- training set (train.csv)
- test set(test.csv)

Variable Definition Key:

- Survival
  - 0= No
  - 1= Yes
- pclass (Class of Travel)
  - 1=1st
  - 2=2nd
  - 3=3rd
- Sex (Gender)
- Age
- sibsp : siblings / spouses aboard the Titanic
- parch :parents / children aboard the Titanic
- tickets
- fare
- cabin
- Embarked : Port of Embarkation.
  - C = Cherbourg,
  - Q = Queenstown,
  - S = Southampton
- pclass: A proxy for socio-economic status (SES)
- PassengerId

Process Report :

1. Data collection:

Data set used: https://www.kaggle.com/c/titanic/data

```
test_df = pd.read_csv("test.csv")
train_df = pd.read_csv(“train.csv")
```

The data obtained is then observed using different functions *(df.shape ,df.info(), df.describe())*

2. Data cleaning (Data Preprocessing):

Missing Values:

- The missing values in the Age column are filled in with appropriate mean value.
- The missing values in Cabin are filled in as ‘Info unavailable’
- Since only 3 values are missing in Embarked , the most common value (’S’) is filled in.

Since, certain values don’t influence the data or the missing values can’t be filled in , the columns are draped. Overhere, the PassengerId, ticket columns are dropped

3. Data Exploratory analysis:(Data visualisation+ML Models)

The following graphs are plotted for the mentioned features to study correlation :

- Distplot : Sex vs Survived
- FacetGrid : Embarked vs Class vs Sex
- Heatmap : Correlations among different features of the dataset
- Catplot: SibSp and Parch (Relatives vs Survived)
- Kdeplot :
    1. Pclass vs Survived
    2. Age vs Survived

The character and string attributes are converted accordingly to int values and categorised .

Model Phase :

The following Machine Learning Models are built and the different precisions are observed:

- Support Vector Machines
- KNN
- Logistic Regression
- Random Forest
- Naive Bayes
- Perceptron
- Stochastic Gradient Decent
- Decision Tree

4. Model Evaluation :

Since it is observed that the Random Forest Model has the best accuracy (86.84), it is implemented. Accordingly , hyper parameter tuning is carried out .

Note: Random Forest is chosen over Decision tree as Decision tree combines some decisions, whereas a Random Forest combines several decision trees.

Conclusion: The Random Forest model has the highest accuracy.