Titanic Assignment Report

Abstract:

The titanic dataset is a popular public dataset to practice classification in machine learning. Titanic was a sunken ship with a few survivors. The survivors were taken to a lifeboat based on age, sex, passenger class, etc. We will analyze the data and try to predict the survival probability of a person.

Introduction:

We have three .csv files as our dataset. One is train.csv, on which we will train our machine learning models. Another one is test.csv, on which we will test some passenger's data and predict their survival. Finally, there is a submission.csv file where we will store our prediction. The train.csv has ten features: a passenger's name, age, gender, port of embarkment, sibling/spouse, parent/child, passenger class, ticket fair, etc.

Relevant Work:

The titanic dataset is one of the most practiced datasets in Kaggle. There were a lot of machine learning enthusiasts working on this dataset in competition. They have experimented with many Data preprocessing techniques with the missing values, have tried to implement and fine-tune many different machine learning models. A huge repository of codes can be found in Kaggle using this dataset.

Methodology:

We will first do EDA (Explanatory Data Analysis) to understand the dataset, find important features and leave out the unnecessary data. Then we will train various classification models, such as Logistic Regression, Gaussian Naive Bayes, Decision Tree, Support Vector Machine, and Random Forest Classifier. Then we will test the model using test data and submit the predictions in the submission file.

Experiments:

While doing the EDA, we have found out features that involve the survival of the passengers in the titanic. We also found many missing values. We have handled the missing values either by replacing it with an average value or removing the feature altogether because that was not important. We were lucky that the survival column of the training data was balanced, so we didn't have to balance the dataset by either upsampling or downsampling. After preparing the data for learning with data preprocessing techniques such as Data Normalization and One Hot

Encoding, we feed data into various machine learning algorithms. We have calculated different evaluation metrics such as accuracy, precision, recall, f1_score, etc. We also have drawn ROC curves of the models to evaluate their efficiency with testing data.

Results and Discussion:

The evaluation metrics we got from various machine learning models are as below:

Model	Accuracy	Precision	Recall	F1_score
Logistic Regression	57%	0.56	0.93	0.52
Naïve Bayes	45%	0.44	0	0.62
SVM	63%	0.56	0.85	0.67
Decision Tree	75%	0.97	0.99	0.60
Random Forest	78%	0.88	0.94	0.70

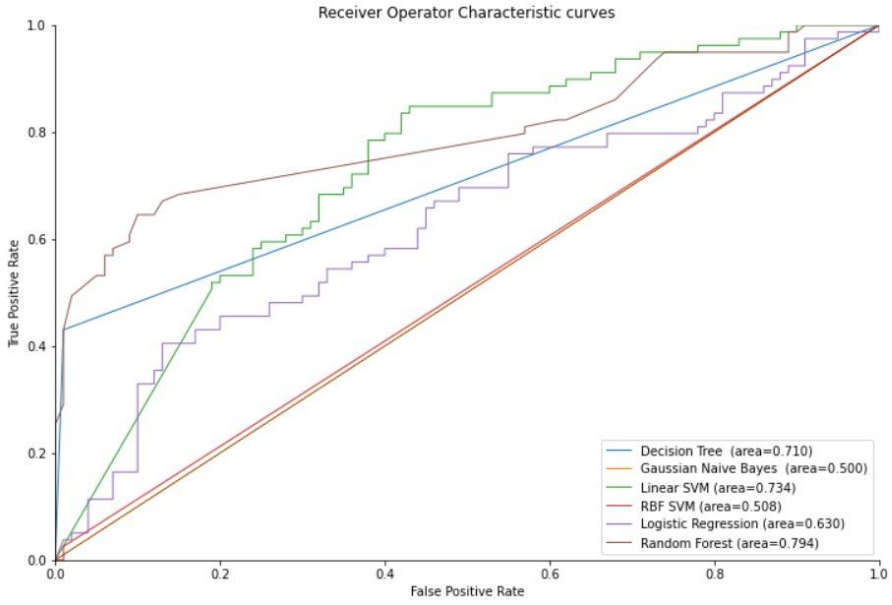

Fig: ROC curve and AUC

The reason for relatively lower accuracy in some models is that not all machine learning models are suitable in every dataset. There were many missing values in the Age column of the training data, which we replaced by average age. This decision greatly affected the accuracy of the models. We have also seen the Radial Basis Function of Support Vector Machine performing very poorly in this case, possibly because of the exponential factor in the algorithm of RBF.

Searching for a solution of lower accuracy in our models, we have found several solutions. One of the approaches was to remove the rows of the missing values, thus shrinking the training data and then using cross-validation multiple folds on the training data to improve accuracy.

Conclusion:

Our assignment was to test our ability to understand a given dataset and use classification models to predict a feature. The titanic dataset helps us to experiment and learn about fundamental machine learning. The accuracy could be more improved with better data preprocessing.