****

**Titanic Coding Assignment Report**

**Course Number: CSE 445  
Section: 04  
Semester: Summer 2021  
Faculty Name: Intisar Tahmid Naheen (ITN)**

Student Name: Md Abir Hossain  
Student ID: 1731597042  
Email: abir.hossain04@northsouth.edu

**ABSTRACT:**

The sinking of the RMS Titanic is one of the most infamous shipwrecks in history. On April 15, 1912, during her maiden voyage, the Titanic sank after colliding with an iceberg, killing 1502 out of 2224 passengers and crew. This sensational tragedy shocked the international community and led to better safety regulations for ships.

**INTRODUCTION:**

The goal of the project was to predict the survival of passengers based off a set of data. We used Kaggle competition "Titanic: Machine Learning from Disaster" (see https://www.kaggle.com/c/titanic/data) to retrieve necessary data and evaluate accuracy of our predictions. The historical data has been split into two groups, a 'Training set' and a 'test set'. For the training set, we are provided with the outcome (Whether or not a passenger survived). We used this set to build our model to generate predictions for the test set. For each passenger in the test set, we had to predict whether or not they survived the sinking. Our score was the percentage of correctly predictions.

**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. One Of the relevant works is:

Titanic: Machine Learning from Disaster

M.A.D.-Python team: Dylan Kenny, Matthew Kiggans, Aleksandr Smirnov

Louisiana State University, MATH 4020, Professor Peter Wolenski

**METHODOLOGY:**

In our project we wanted to implement several classification algorithm discussed throughout the course of the semester. The classification algorithms are as follows:

**Decision Tree:** One of the most commonly used algorithms in medical application is the decision tree. It makes the use of graphs to represent a decision tree. Research has shown that amongst three diverse data mining techniques, decision tree is the best with 99.6% approximation. Within decision trees, further research on various medical datasets showed that CART is the best in terms of accuracy and time complexity

**Support Vector Machines:** Support Vector Machine is a supervised learning algorithms and it reduces the overflowing of trained data. Its goal is to find the optimized decision boundaries to help predict heart disease at the earlier stage.

**Random Forest Classifier:** Random forests Classifier is a supervised learning algorithm. It does both classification and regression. It is also very flexible and easily usable algorithm. A forest is contains multiple trees. Robustness of the forest depends on the number of trees in it. Random forests obtains the best solution by creating decision trees on randomly selected data samples, gets prediction from each tree and then voting. It also helps to find feature importance.

**Gaussian Naïve Bayes:** Naive Bayes are a group of supervised machine learning classification algorithms based on the Bayes theorem. It is a simple classification technique, but has high functionality. They find use when the dimensionality of the inputs is high. Complex classification problems can also be implemented by using Naive Bayes Classifier.

**Logistic Regression:** Logistic regression is another technique borrowed by machine learning from the field of statistics. It is the go-to method for binary classification problems (problems with two class values).

**Experiments:**

The data has been split into two groups:

training set (train.csv)

test set (test.csv)

The training set should be used to build your machine learning models. For the training set, we provide the outcome (also known as the “ground truth”) for each passenger. Your model will be based on “features” like passengers’ gender and class. You can also use feature engineering to create new features.

The test set should be used to see how well your model performs on unseen data. For the test set, we do not provide the ground truth for each passenger. It is your job to predict these outcomes. For each passenger in the test set, use the model you trained to predict whether or not they survived the sinking of the Titanic.

We also include gender\_submission.csv, a set of predictions that assume all and only female passengers survive, as an example of what a submission file should look like.

1. Survived: Passenger Survived or not (Categorical) it’s the outcome

2. Sex: Gender of patient (Male, Female)

3. Pclass: Passenger class (1, 2 or 3)

4. Age: Passenger Age

5. SibSip: Sibling or spouse number

6. Parch: Number of parents / children aboard the Titanic

7. Ticket: Ticket number

8. Fare: Passenger fare

9. Cabin: Gender of patient

10. Embarked: Port of Embarkation (C = Cherbourg, Q = Queenstown,

S = Southampton)

The evaluation metrics that we will use to test the performance of various models are:

**Accuracy**

Accuracy is the quintessential classification metric. It is pretty easy to understand. And easily suited for binary as well as a multiclass classification problem.

**Precision**

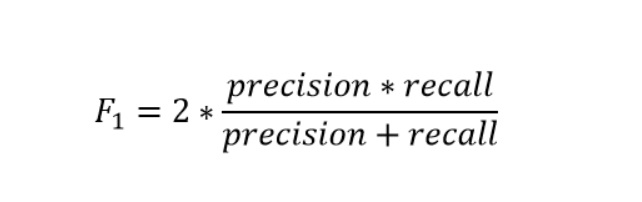
Let’s start with precision, which answers the following question: what proportion of predicted Positives is truly positive?

**Recall**

Another very useful measure is recall, which answers a different question: what proportion of actual Positives is correctly classified?

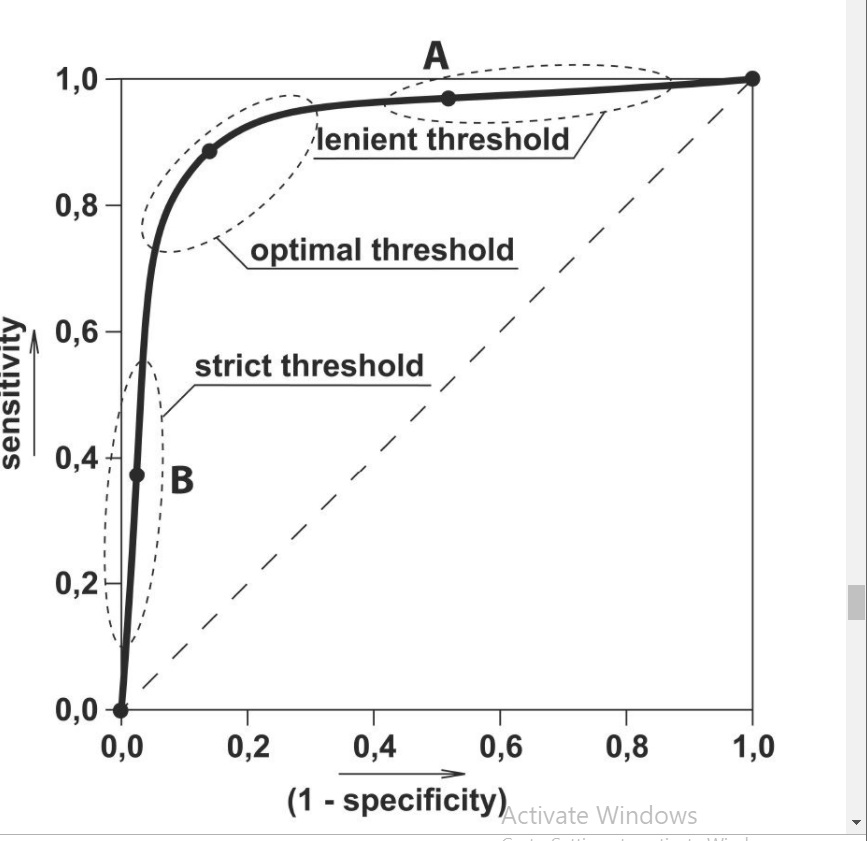
**F1 score:**

The F1 score is a number between 0 and 1 and is the harmonic mean of precision and recall.



**ROC and AUC:**

AUC ROC indicates how well the probabilities from the positive classes are separated from the negative classes



**Results and Discussion:**

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

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Accuracy | Precision | Recall | F1\_score |
| Logistic Regression | 76% | 0.78 | 0.63 | 0.69 |
| Naïve Bayes | 76% | 0.75 | 0.66 | 0.70 |
| SVM | 77% | 0.78 | 0.64 | 0.71 |
| Decision Tree | 74% | 0.73 | 0.63 | 0.67 |
| Random Forest | 82% | 0.86 | 0.71 | 0.77 |

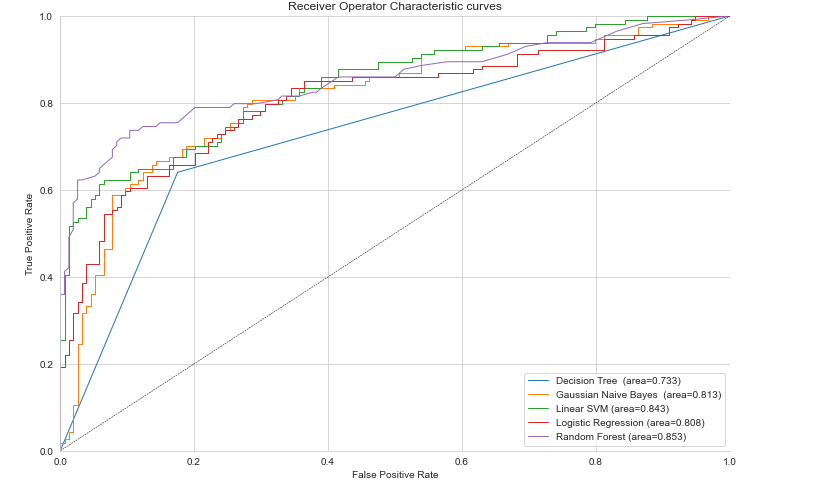


Fig: ROC and AUC

We have seen different performance metrics from different machine learning algorithm. It clearly shows that no one classification can outright perform better than the other algorithms. Each has some advantage and disadvantage depending on the data preparation, data cleaning, etc. A way to improve the performance of the models could be ensemble method, where we can join multiple classification algorithms together.

**Conclusion:**

The five algorithms were implemented on the dataset from Kaggle and I will be using Machine Learning to predict whether any person will suffer or is suffering from heart disease. The observation was that Random Forest Classifier has the highest accuracy of 85.3% however Naïve Bayes has a good accuracy of 81.3% and if the dataset is larger, the computational costs of the Random Forest Classifier will increase. In future, we can collect latest data from local hospitals or international organizations and train our models better.