****

**Heart Disease Prediction Using Machine Learning**

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

Student Name: Emamul Hassan  
Student ID: 1731250042  
Email: emamul.hassan@northsouth.edu  
  
Date prepared: 16/09/2021

**ABSTRACT:**

The sinking of the RMS Titanic caused the death of thousands of passengers and crew is one of the deadliest maritime disasters in history. One of the reasons that the shipwreck led to such loss of life was that there were not enough lifeboats for the passengers and crew. The interesting observation which comes out from the sinking is that some people were more likely to survive than others, like women, children were the one who got the priority to rescue. The objective is to first explore hidden or previously unknown information by applying exploratory data analytics on available dataset and then apply different machine learning models to complete the analysis of what sorts of people were likely to survive. After this the results of applying machine learning models are compared and analyzed on the basis of accuracy.

**INTRODUCTION:**

The most infamous disaster which occurred over a century ago on April 15, 1912, that is well known as sinking of “The Titanic”. The collision with the iceberg ripped off many parts of the Titanic. Many classes of people of all ages and gender where present on that fateful night, but the bad luck was that there were only few life boats to rescue. The dead included a large number of men whose place was given to the many women and children on board. The men travelling in second class were dead on the vine.

**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.

An instance-based prediction of heart-disease presence with the Cleveland database produced C4: 74.8% accuracy in 1989. (David, 1988)

**METHODOLOGY:**

In the project we implemented five classification algorithms which were asked in the question. 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.

**EVALUATION METRICS:**

**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**

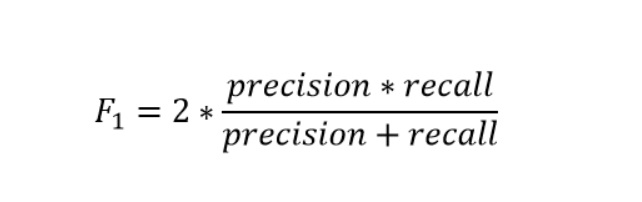
Let’s start with precision, which answers the following question: what proportion of predicted Positives is truly positive?

**Recall**

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

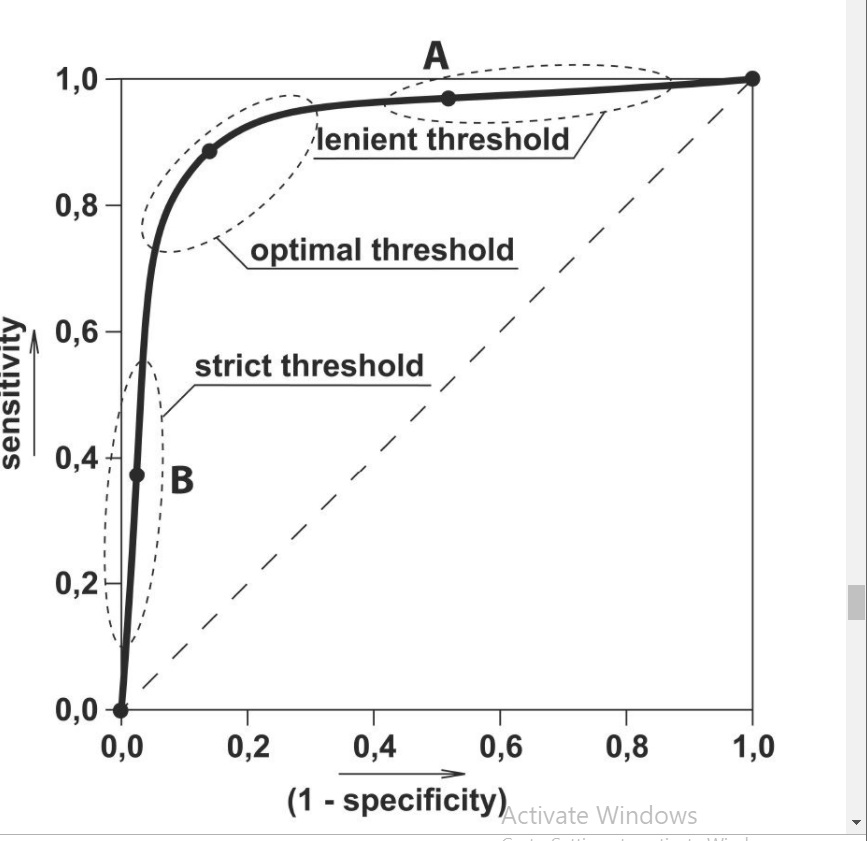
**F1 score:**

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 | 81% | 0.80 | 0.66 | 0.72 |
| Naïve Bayes | 81% | 0.77 | 0.71 | 0.74 |
| SVM | 79% | 0.78 | 0.64 | 0.70 |
| Decision Tree | 75% | 0.68 | 0.66 | 0.67 |
| Random Forest | 80% | 0.79 | 0.66 | 0.72 |

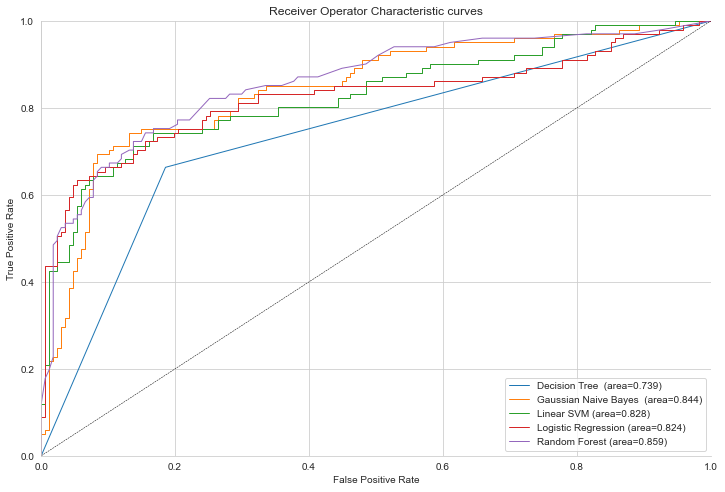


Fig: ROC and AUC

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 given dataset and I will be using Machine Learning to predict whether any person will survive the Titanic incident. The observation was that Logistic Regression and Naïve Bayes, both has the highest accuracy of 81% In future, we can collect latest data from online resource and train our models for better results.