Project Title

Predicting Blood Donations

Driven Data competition

Team members

Dhawal Parmar (ddp160330)

Komal Mukadam (kjm160030)

Nikita Kothari (nrk160530)

Sonali Mishra (sxm161931)

Number of free late days used: 2

Note: You are allowed a **total** of 4 free late days for the **entire semester**. You can use at most 2 for each assignment. After that, there will be a penalty of 10% for each late day.

1. Introduction and Problem Description

1.1. Introduction

The Blood Transfusion Service Center drives to different universities and collects blood as part of a blood drive. We want to predict the chances of a person donating blood when the vehicle comes to campus next time. To ensure that more patients get the blood transfusions as and when they require, we need to build good data-driven systems so that we can track and predict donations and supply needs that will help to improve the entire supply chain.

1.2. Problem Description

The Goal of this project is to predict whether he/she donated blood in March 2007 on given UCI dataset. This competition is for learning and exploring different machine learning models to predict best results based on performance evaluation metrics. We suggest using more than one evaluation parameter to avoid overfitting of the model. In this dataset, we are predicting the probability of blood donations, so participants can use supervised learning methods to achieve desired results.

2. Dataset Description

2.1. Description

The dataset contains Multivariate Characteristics. It contains 748 donor instances randomly selected from the donor database and each instance has 5 attributes. The data represents business level orientation and was recorded in 2008 with associated tasks as classification and it contains no missing values.

2.2. Source

The donor database of Blood Transfusion Service Center in Hsin-Chu City in Taiwan was used as source for this study. This study is used to demonstrate the RFMTC marketing model which is a RFM model proposed by I-Cheng et al in 2009.

2.3. Attribute Information

This dataset contains 5 attributes and the details are as follows:

- **Recency- Months since last donation**: This is the number of months since the donor's most recent blood donation.
- **Frequency - Number of donations**: This is the total number of blood donations that the donor has made.
- **Monetary – Total volume donated**: This is the total amount of blood that the donor has donated in cubic centimeters.
- **Time - Months since first donation**: This is the number of months since the donor's first blood donation.
- **Whether he/she donated blood in March 2007**: It is represented in terms of a binary variable. 1 denotes donating blood; 0 stands for not donating blood.

The features are defined as below:

Variables	**Definition**	**Key**
Recency	Months since last donation	Continuous
Frequency	Total number of donation	Continuous
Monetary	Total blood donated in c.c.	Continuous
Time	Months since first donation	Continuous
Whether he/she donated blood in March 2007	Donating blood or not	Binary 1 stand for donating blood; 0 stands for not donating blood

Figure 3.1 - Data Description table

	Recency (months)	Frequency (times)	Monetary (c.c. blood)	Time (months)	whether he/she donated blood in March 2007
count	748.000000	748.000000	748.000000	748.000000	748.000000
mean	9.506684	5.514706	1378.676471	34.282086	0.237968
std	8.095396	5.839307	1459.826781	24.376714	0.426124
min	0.000000	1.000000	250.000000	2.000000	0.000000
25%	2.750000	2.000000	500.000000	16.000000	0.000000
50%	7.000000	4.000000	1000.000000	28.000000	0.000000
75%	14.000000	7.000000	1750.000000	50.000000	0.000000
max	74.000000	50.000000	12500.000000	98.000000	1.000000

Figure 3.2 - Metadata of the features

2.4. Data Analysis

2.4.1. Density Plot

Density plots represent the distribution of each attribute which helps to determine the visualization of each graph. It is represented in terms of a line which is an abstracted histogram. Density plot of recently(month), frequency(times), Monetary (c. C. blood) are right-skewed whereas the density plot of whether he/she donated blood in March 2007 is bimodal because it represents binary value for the corresponding predicate attribute. Density plot for time is multimodal which represents data distributed over time.

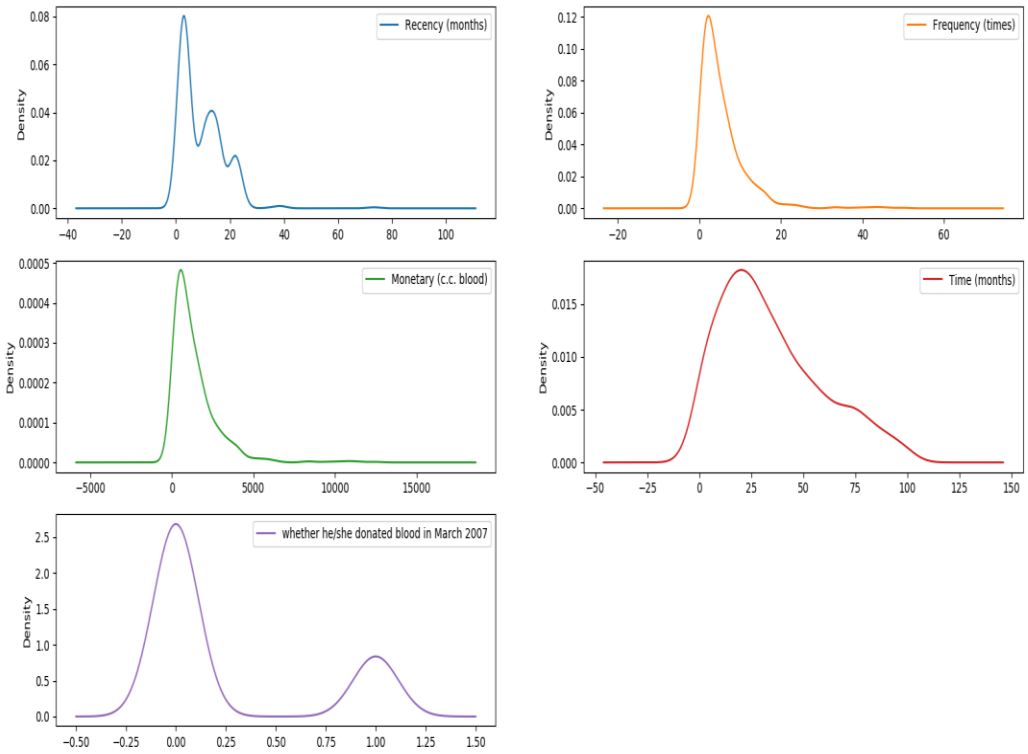

Figure 3.3 - Density plot for all given attributes

2.4.2. Correlation and Scatter Plots

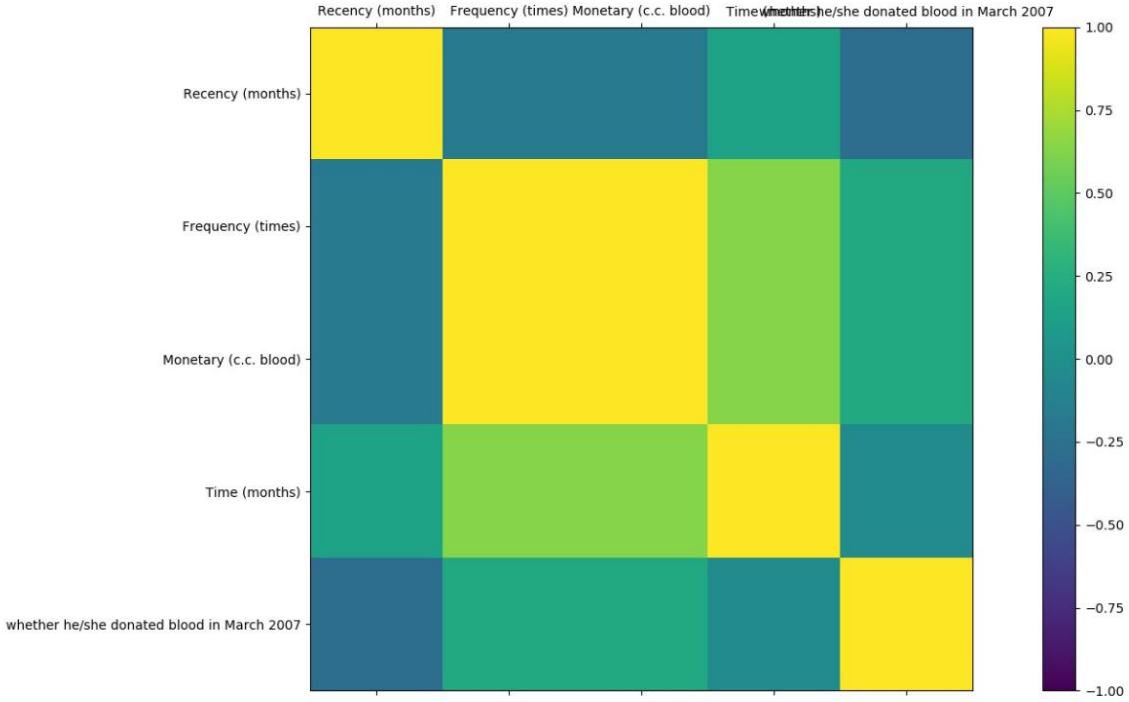

Figure 3.4 - Correlation plot for all given attributes

The relation between two features is plotted along two-dimensional axes. Correlation between two features can be seen along the axis. If the value along y-axis increases with x-axis, then the two features are correlated.

From the graph, we can see that Frequency and Monetary have a strong linear correlation with each other. So, one of them can be eliminated.

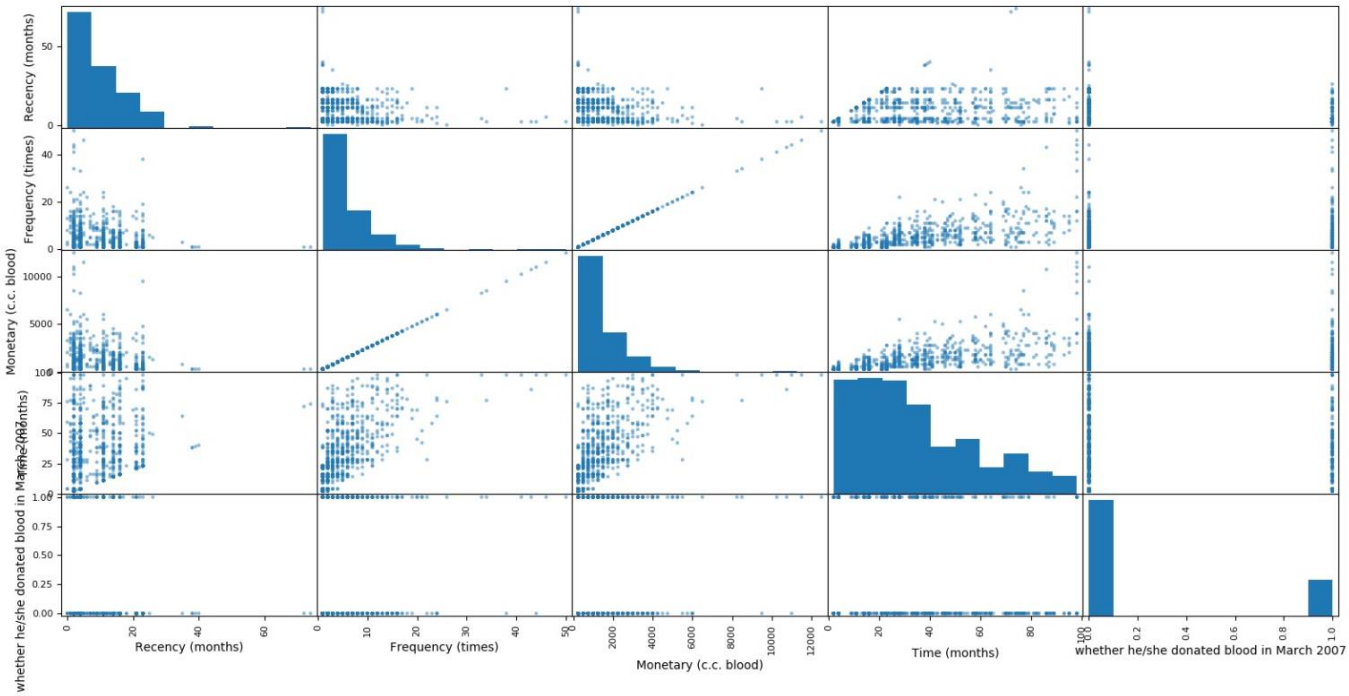

Figure 3.5 - Scatter plot for all given attributes

3. Pre-processing Techniques

3.1. Standardization of Data

As different features have data varying in different ranges, standardization of data is required to scale all the data to the same scale. The distribution of standardized data is Gaussian having mean as zero and variance as one. We have used scikit learn's 'StandardScaler' method for standardization of the data.

3.2. Feature Selection based on correlation

As observed from the scatter plot the Frequency and Monetary features have a strong correlation. Thus, selecting either of them will not affect the performance of the model. Elimination of one of the feature will also reduce the redundancy in the data. We eliminated Monetary (c.c. blood) feature from the dataset.

4. Proposed solutions and methods

4.1. Solutions

According to the problem statement our aim is to create probabilistic models to obtain the probability of an individual donating blood. Since we are provided with the class labels, we have used supervised learning model for classification.

The following classifiers were used at initial stage to train our model:

1. Artificial Neural Network
2. SVM with 'rbf' kernel
3. k-Nearest Neighbors
4. Bagging
5. Random Forests
6. Adaboost

4.2. Methods

We have used different model evaluation techniques to predict the performance of the models using various metrics for performance evaluation and methods for model comparisons. Since the dataset is small, we have used k-fold cross validation technique to test on various datasets and thereby increasing the performance.

4.2.1. Performance Evaluation Metrics

We have used accuracy, precision, recall and AUC as the different metrics for evaluating the classifiers. After analyzing the metric values for all the classifiers, we selected Artificial Neural Network, Adaboost and K-Nearest Neighbors classifier as the best suited classifier for the given dataset. We have used enough performance metrics to avoid overfitting on a metric.

4.2.2. Performance Evaluation Methods

After analyzing the data, we observed that the class label data is not equally distributed. We found that the class label 0, indicating that the person did not donate blood in 2007 appeared more than class label 1 which indicates that the person donated blood. Out of 748 instances, 548 instances have class label 0 and remaining 170 instances have class label 1. Due to the imbalance in class label distribution, we used Stratified k-Fold technique for creating different datasets. This technique ensures that in each fold, the test dataset is made by preserving the percentage of samples for each class. We have used 10-fold cross validation.

5. Experimental Results and Analysis

5.1. Results

As mentioned above we have used various supervised learning techniques to train the model. The results obtained during the experimental phase are as follows:

Models	Accuracy (%)	Precision (%)	Recall (%)	AUC	Log Loss
Neural Network	79.95	74.66	65.21	0.65	6.924
K-NN	79.01	74.68	60.72	0.61	7.240
Ada Boosting	78.21	71.25	62.33	0.62	7.525
Bagging	73.52	61.88	59.0	0.59	9.144
SVM	78.48	78.32	58.0	0.57	7.433
Random Forest	73.92	62.89	60.0	0.59	9.006

Figure 6.1 - Comparisons between different supervised learning models

Based on our initial experiments and results, we found Adaboost Classifier, Artificial neural network, and K-Nearest Neighbors as the best classifier on our dataset and hence we tried to tune parameters based on the accuracy, Area under curve(AuC) and Recall achieving better results.

5.1.1. Adaboost Classifier

The aim of the Adaboost classifier is to create a strong classifying model by combining many weak models. Adaboosting is a re-weighting scheme for data instance where misclassified points are given higher weightage and the correctly classified points are assigned lower weights.

To train our data for the adaboosting model, we have set the number of estimators to 20 and learning rate to 1 which gave us the best results of accuracy around 78%.

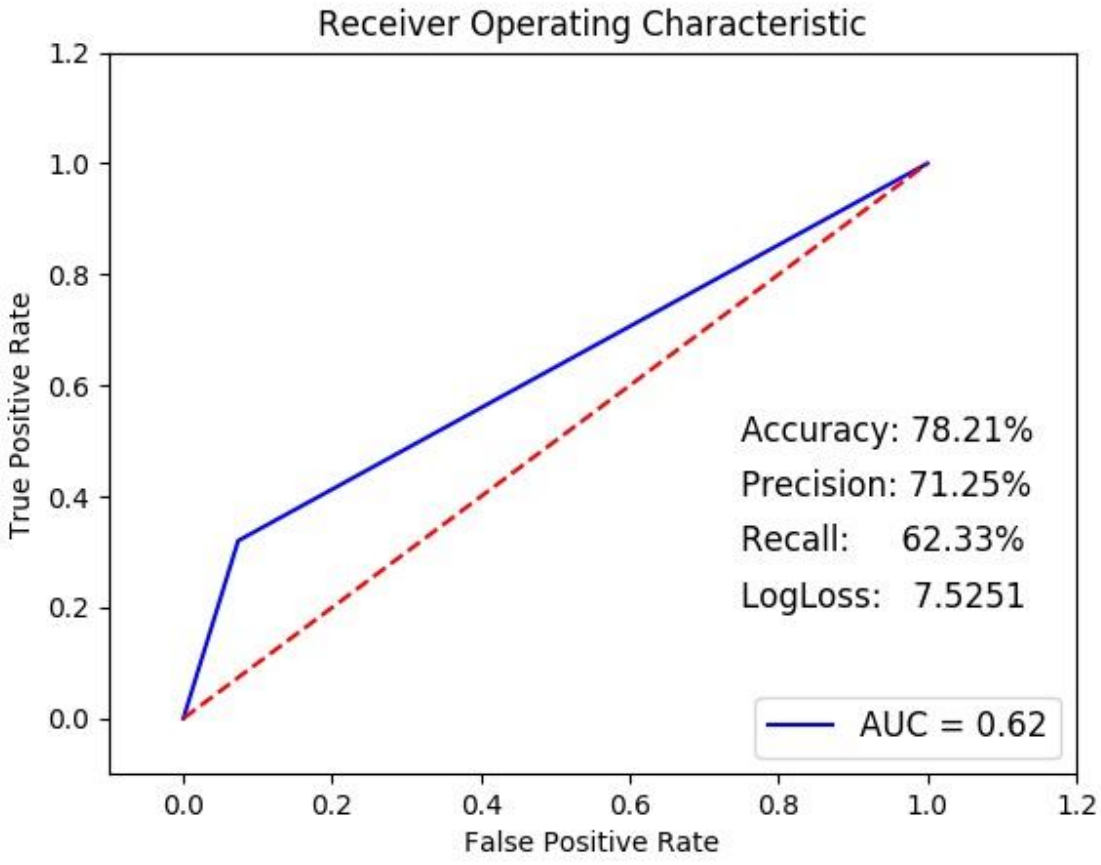

Figure 6.2 - ROC curve for Adaboost classifier

5.1.2. Artificial Neural Network

Artificial Neural Network are multilayer perceptron's that are universal estimators used to approximate any function. ANN has a large number of weighted connections between the neuron like processing elements that are highly parallel and have distributed control. We have trained our data using 200 hidden layers and achieved an accuracy of approximately 80% which was the best result achieved.

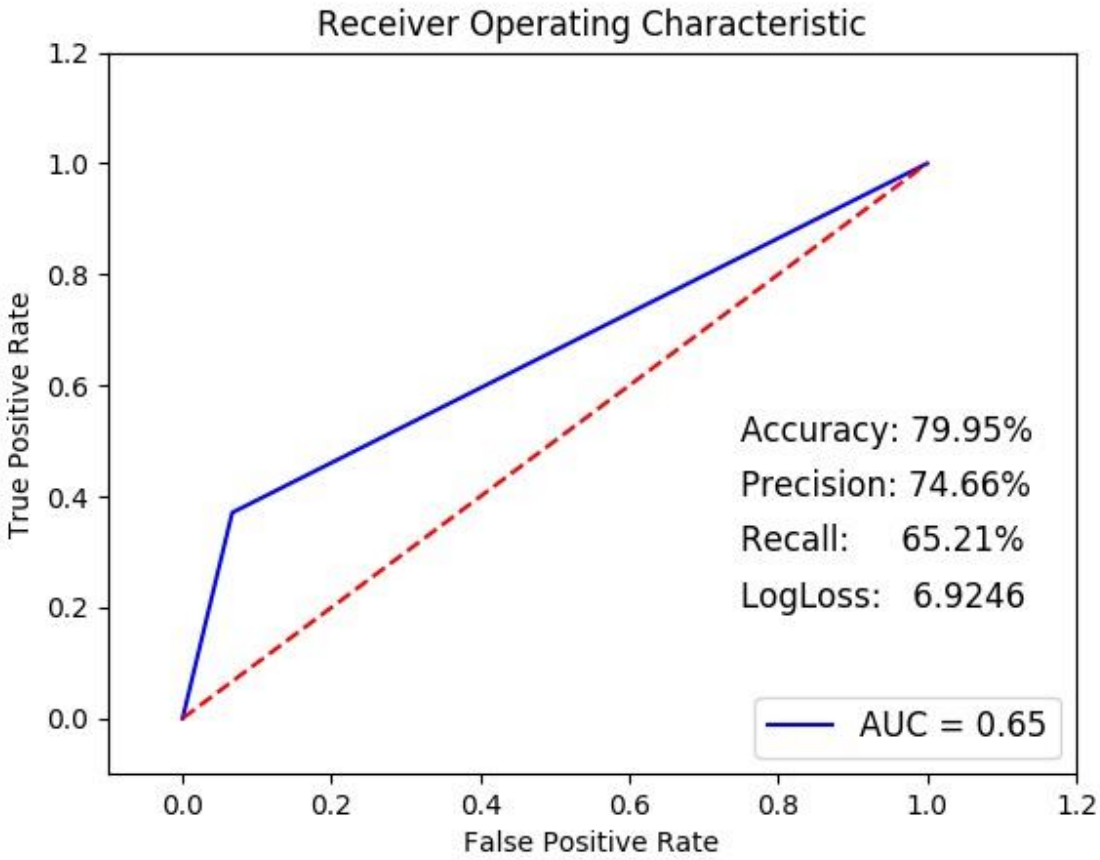

Figure 6.3 - ROC curve for Artificial Neural Network

5.1.3. K Nearest Neighbors

K nearest neighbor is a type of instance based learning. It stores all the training results and classifies new cases. Predictions are made for a new instance by searching through the entire training set for the K most similar neighboring instances and summarizing the output variable for those instances. We have used 10 nearest neighbors to predict our test data and have achieved an accuracy of 79%

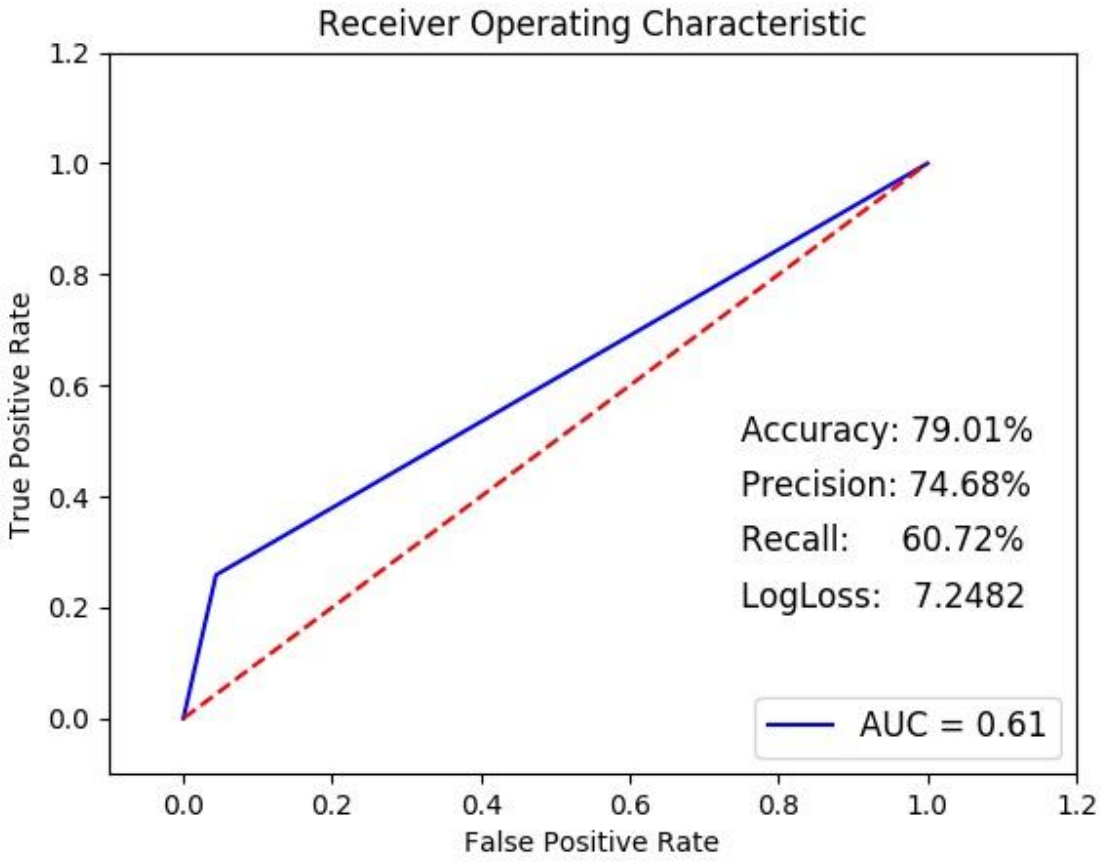

Figure 6.4 - ROC curve for K Nearest Neighbor

6. Summary

We have compared various data classification models and found that Ada boosting, Artificial neural network and k nearest neighbors generated best output model in terms of precision, accuracy, and area under the curve for the given data. Among the given three classifiers, the best classifier which suited the dataset was Artificial Neural Network.

7. References

7.1. Data Driven competition link

https://www.drivendata.org/competitions/2/warm-up-predict-blood-donations/page/5/

7.2. Database

https://archive.ics.uci.edu/ml/datasets/Blood+Transfusion+Service+Center

7.3. Scikit learn

http://scikit-learn.org/

7.4. Scikit learn - Graphical data analysis

https://machinelearningmastery.com/visualize-machine-learning-data-python-pandas/