Abstract

A series of experiments performed to find the best model that predicts if a child with pharyngitis should be recommended to a rapid antigen test by a physician.

Team members

JiaTong Lyu  
Meher Satya Swaroop Malina

676 cases of pharyngitis in children

CLASSIFICATION MODEL

**PROPOSAL:**

Our goal is to predict the result of the rapid antigen detection test for Group A streptococcus (GAS) infection using signs and symptoms in children with pharyngitis. The model built will help the physicians decide whether they can rely on signs and symptoms to select pediatric patients with pharyngitis who should undergo rapid antigen detection testing (RADT) for GAS.

**DATASET DESCRIPTION:**

The dataset contains signs and symptoms of 676 cases of children with pharyngitis. It has 20 attributes, and one of them is the result of the rapid antigen detection test for GAS, which is the target variable. Features include symptoms, signs, and patient info like cough, tender, temperature, age, diarrhea, headache, etc.

Below is the **detailed description** of each attribute and its type

|  |  |  |
| --- | --- | --- |
| Attribute | Type | Description |
| age\_y | Numeric | Age of the patient |
| **radt** | Category | Result of the RADT, 0 – negative, 1 – positive |
| pain | Category | Indicates pain, 0 –no pain, 1 – pain |
| swollenadp | Numeric | Indicates swollen level |
| tender | Category | Indicates tenderness, 0 –no tenderness, 1 – tenderness |
| tonsillarswelling | Category | Indicates tonsillar swelling, 0 – no swell, 1 – swollen |
| exudate | Category | Indicates if exudative, 0 – not exudative, 1 – exudative |
| temperature | Numeric | Body temperature of the patient |
| sudden | Category | Indicates the nature of time taken to develop the disease, 0 – not sudden, 1 – sudden |
| cough | Category | Indicates cough, 0 –no cough, 1 – cough |
| rhinorrhea | Category | Indicates rhinorrhea, 0 –no rhinorrhea, 1 – rhinorrhea |
| conjunctivitis | Category | Indicates conjunctivitis, 0 –no conjunctivitis, 1 – conjunctivitis |
| headache | Category | Indicates headache, 0 –no headache, 1 – headache |
| erythema | Category | Indicates erythema, 0 –no erythema, 1 – erythema |
| petechiae | Category | Indicates petechiae, 0 –no petechiae, 1 – petechiae |
| abdopain | Category | Indicates abdominal pain, 0 –no abdominal pain, 1 – abdominal pain |
| diarrhea | Category | Indicates diarrhea, 0 –no diarrhea, 1 – diarrhea |
| nauseavomit | Category | Indicates nausea vomit, 0 –no nausea vomit, 1 – nausea vomit |
| scarlet | Category | Indicates if the color is scarlet, 0 –not scarlet, 1 – scarlet |

**Target Variable**: radt (categorical)

**Dataset**: Pharyngitis in children dataset

(<https://www.kaggle.com/datasets/yoshifumimiya/pharyngitis>)

**TOOLS USED:**

**WEKA:**

WEKA is an open-source software provides tools for data preprocessing, implementation of several Machine Learning algorithms, and visualization tools so that you can develop machine learning techniques and apply them to real-world data mining problems. [1]

**PYTHON:**

Python is a high-level, general-purpose programming language. Its design philosophy emphasizes code readability with the use of significant indentation.

Python is dynamically typed, and garbage collected. It supports multiple programming paradigms, including structured (particularly procedural), object-oriented and functional programming. It is often described as a "batteries included" language due to its comprehensive standard library. [2]

**Packages used in this project:**

**Pandas:** pandas is a Python package providing fast, flexible, and expressive data structures designed to make working with “relational” or “labeled” data both easy and intuitive. It aims to be the fundamental high-level building block for doing practical, real-world data analysis in Python. [3]

**NumPy:** NumPy is the fundamental package for scientific computing in Python. It is a Python library that provides a multidimensional array object, various derived objects (such as masked arrays and matrices), and an assortment of routines for fast operations on arrays, including mathematical, logical, shape manipulation, sorting, selecting, I/O, discrete Fourier transforms, basic linear algebra, basic statistical operations, random simulation and much more. [4]

**Math:** This module provides access to the mathematical functions defined by the C standard. [5]

**Sklearn:** A set of python modules for machine learning and data mining. [6]

**Statistics:** This module provides functions for calculating mathematical statistics of numeric (Real-valued) data. [7]

**Seaborn:** Seaborn is a Python data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics. [8]

**Matplotlib:** Matplotlib is a comprehensive library for creating static, animated, and interactive visualizations in Python. [9]

**CLASSIFICATION ALOGRITHMS USED:**

1. **Random forest:** The random forest is a classification algorithm consisting of many decisions trees. It uses bagging and feature randomness when building each individual tree to try to create an uncorrelated forest of trees whose prediction by committee is more accurate than that of any individual tree. [10]
2. **Adaboost with random forest as the base classifier:** AdaBoost is a boosting ensemble model and works especially well with the decision tree. It learns from the mistakes by increasing the weight of misclassified data points. Base classifier used here was random forest. [11]
3. **Adaboost with j48 as the base classifier:** AdaBoost is a boosting ensemble model and works especially well with the decision tree. It learns from the mistakes by increasing the weight of misclassified data points. Base classifier used here was J48 decision tree. [11]
4. **KNN (k Nearest Neighbors):** The k-nearest neighbors algorithm, also known as KNN or k-NN, is a non-parametric, supervised learning classifier, which uses proximity to make classifications or predictions about the grouping of an individual data point. [12]
5. **Logistic regression:** The logistic model (or logit model) is a statistical model that models the probability of an event taking place by having the log-odds for the event be a linear combination of one or more independent variables. In regression analysis, logistic regression (or logit regression) is estimating the parameters of a logistic model (the coefficients in the linear combination). [13]
6. **J48 Decision tree:** The J48 algorithm is a classification algorithm which produces decision trees based on information theory.
7. **Naïve Bayes:** Naive Bayes classifier is a supervised learning algorithm based on applying Bayes’ theorem with the “naive” assumption of conditional independence between every pair of features given the value of the class variable. [14]
8. **Support Vector Machines (SVM):** SVM is a supervised learning method that looks at data and sorts it into categories. An SVM outputs a map of the sorted data with the margins between the two as far apart as possible. [15]

**ATTRIBUTE SELECTION METHODS USED:**

1. **Chi-Square Test for feature selection:** A chi-square test on correlation is performed between each attribute and the target variable. The top n attributes with the lowest p-values are selected in this method. **n=11 in this project.**
2. **Mutual info feature selection:** Information gain of each attribute is calculated with respect to the target variable. The top n attributes with the highest information gain are selected in this method.
3. **ANOVA test for feature selection:** Like chi-square test, an F-test on variance is performed between each attribute and the target variable. The top n attributes with the lowest p-values are selected in this method.
4. **Recursive feature elimination:** This method uses a model to select the best set of features. Each possible combination of attributes of various sizes is trained on the model, and the feature set with which a model trained has the highest accuracy is selected. We can specify a lower limit on the size of features up to which recursive elimination should occur to decrease computation.
5. **Feature selection with random forests:** This method uses random forests to select the best set of features. The features with the highest average information gain from different trees in the random forest are selected in this method.

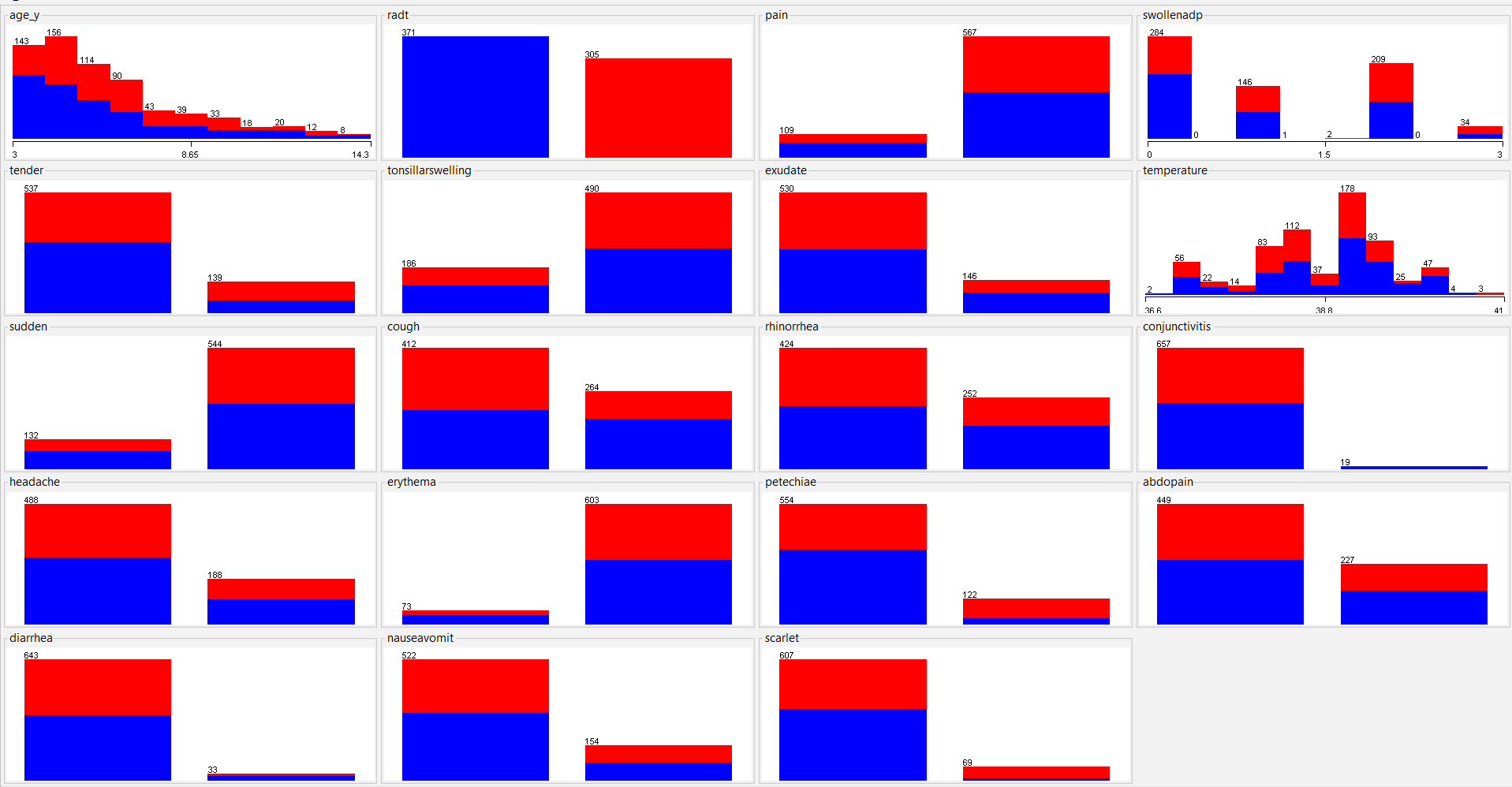
**FEATURES SELECTED:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Attribute** | **Chi-square** | **Mutual info** | **ANOVA test** | **Recursive elimination** | **Random forests** |
| age\_y | Y |  | Y | Y | Y |
| **radt** | Target | Target | Target | Target | Target |
| pain |  | Y | Y | Y |  |
| swollenadp | Y |  | Y | Y | Y |
| tender | Y | Y | Y |  |  |
| tonsillarswelling |  | Y |  |  |  |
| exudate | Y |  | Y |  |  |
| temperature | Y | Y | Y | Y | Y |
| sudden |  | Y |  |  |  |
| cough | Y |  | Y | Y |  |
| rhinorrhea | Y | Y | Y |  |  |
| conjunctivitis | Eliminated due to low variance | Eliminated due to low variance | Eliminated due to low variance | Eliminated due to low variance | Eliminated due to low variance |
| headache |  | Y |  | Y |  |
| erythema | Y |  | Y |  |  |
| petechiae | Y | Y | Y | Y | Y |
| abdopain |  | Y |  |  |  |
| diarrhea | Eliminated due to low variance | Eliminated due to low variance | Eliminated due to low variance | Eliminated due to low variance | Eliminated due to low variance |
| nauseavomit | Y | Y |  |  |  |
| scarlet | Y | Y | Y | Y | Y |

**Note:** Number attribute has been eliminated from the dataset as it is a constant and used only to identify each case.

**EXPLORATORY DATA ANALYSIS:**

1. Each attribute has been visualized to know the class distribution. 0 – blue, 1 – red.
2. There are 371 cases with a negative RADT result and 305 cases with a positive RADT result in the entire dataset.

****

1. The number of rows with missing values for each attribute is presented below.



**Correlation between the attributes:**



**DATA PREPROCESSING:**

**Tool used:** Python

**Handling Missing Data**:

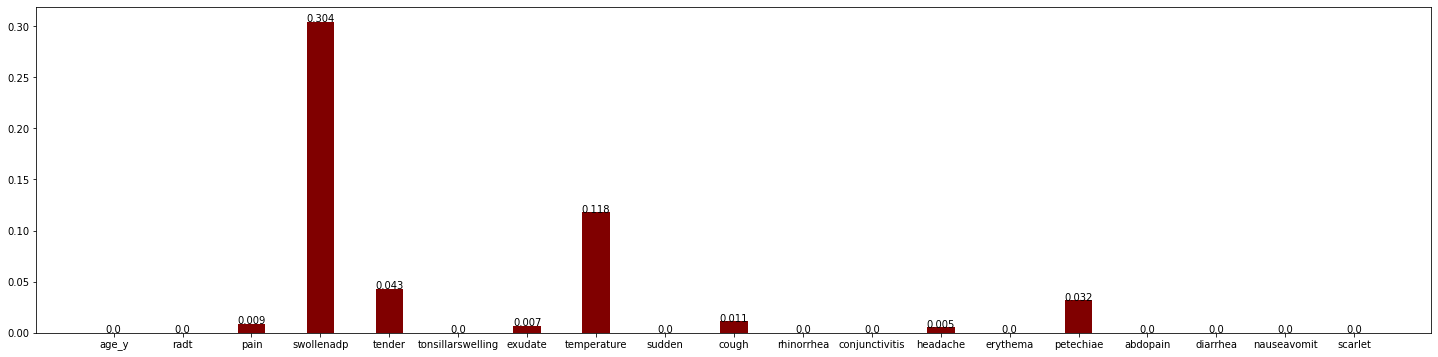
There are 164 rows with missing data in the dataset. Removing the rows with missing data does not give the best results when the dataset is small. Hence, we decided to preform imputation on the dataset and fill in the missing value to make the best use of the data we have.

We have used 3 different imputation methods and selected the best by comparing the datasets based on Euclidean distance for numeric attributes and Jaccard similarity for categorical attributes.

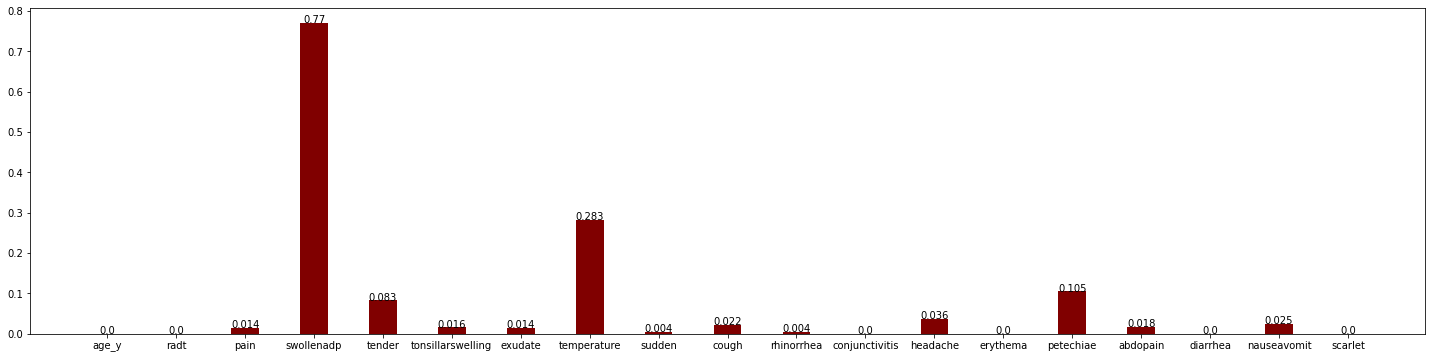
1. **k-NN imputation (k=1):** It is an iterative process where each column with missing data is considered as the target column and its value is filled with the value of the target column of the first nearest tuple. Nearness is calculated by the Euclidean distance.
2. **k-NN imputation (k=5):** It is an iterative process where each column with missing data is considered as the target column and its value is filled with the average value of the target column taken from the 5 nearest tuples. Weights have been applied to each nearest tuple based on the distance.
3. **Logistic regression imputation:** It is an iterative process where each column with missing data is considered as the target column and its value is filled with the predicted value from logistic regression model trained for each attribute. The data has been normalized before imputation.

Below is the comparison of the 3 different datasets derived from the original dataset by the above 3 imputation methods.

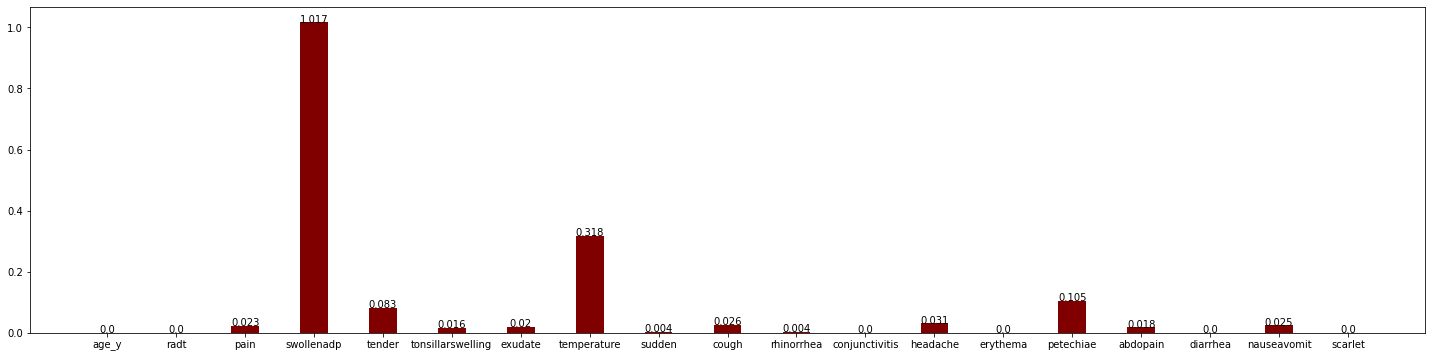
**k-NN imputation (k=1) vs k-NN imputation (k=5):**



**Logistic regression imputation vs k-NN imputation (k=5):**



**Logistic regression imputation vs k-NN imputation (k=1):**



Above graphs the **dissimilarity of each column between two different datasets**.

From the above graphs, we can tell that datasets produced from **k-NN imputation (k=1)** and **k-NN imputation (k=5)** are closer to each other than they are to the dataset produced from **Logistic regression imputation.**

Hence, on an average, if we must select the best dataset after imputation, it would be the dataset produced from **k-NN imputation (k=5)** as it falls in between the 3 datasets created in terms of dissimilarity and also with k=5, it would be gathering data from more of similar tuples than k=1 and need to be normalized like we needed for logistic regression imputation in this case (variance is preserved).

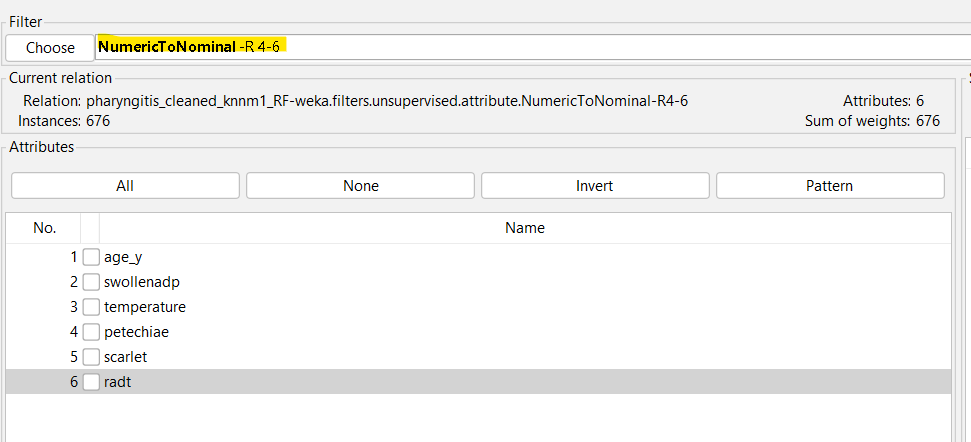
**Note:** **Variance** **do not affect the models we have chosen** for this project as most of them are decision trees and the end goal is just classification.

**Converting csv to arff:**

The datasets produced after the imputation techniques used are stored in csv format. The issue here is a csv file do not preserve the datatype of a categorical variable. It stores the categorical data as numeric data.

Hence, we used WEKA to convert numeric to nominal attributes and then to convert it to a arff file.

Below is the screenshot of the filter used to convert numeric attributes to nominal attributes.



**DATA MINING:**

**Tool used:** WEKA

Initially, 8 models were trained using the 8 predetermined algorithms (refer page 5) with the entire dataset using 10-fold cross validation.

Now that we have the final 5 datasets generated with imputation and 5 attribute selection, we have trained the 40 (8 x 5) models using the 8 algorithms after splitting 66% of each dataset as training sets and the rest 34% as test sets in **WEKA**.

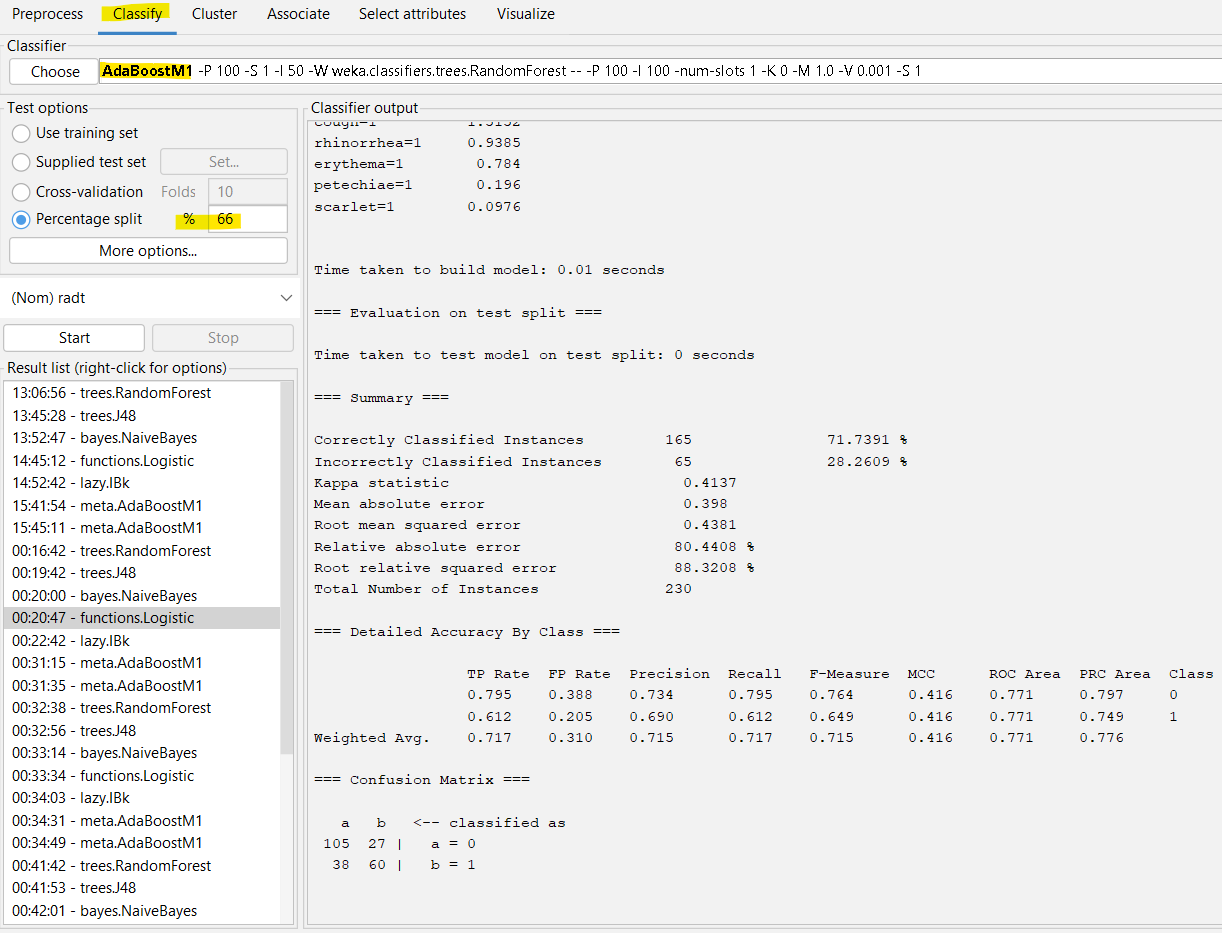
**Hyperparameter tuning:**

**Hyperparameters** such as

1. k in k-NN
2. random seed for Xval/%split in WEKA
3. No of iterations
4. Batch size

have been selected based on the **random search** method and the best set for each algorithm has been used for the model training and evaluation.

Below is a sample screenshot to show the usage of WEKA for data mining.



**RESULT AND EVALUTION METRICS:**

Below are the accuracies and the evaluation metrics of the 5 models generated.

1. **Random Forest:**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | TP Rate | FP Rate | Precision | Recall | F-Measure | MCC | ROC Area | PRC Area | Class | Acc |
|  | 0.985 | 0.841 | 0.912 | 0.985 | 0.947 | 0.258 | 0.787 | 0.966 | 0 |  |
|  | 0.159 | 0.015 | 0.55 | 0.159 | 0.247 | 0.258 | 0.787 | 0.332 | 1 |  |
| Weighted Avg. | 0.901 | 0.756 | 0.875 | 0.901 | 0.876 | 0.258 | 0.787 | 0.901 |  | 90.0888 |

**Confusion matrix:**

|  |  |  |
| --- | --- | --- |
| Classified as Actual | a=0 | b=1 |
| a = 0 | 598 | 9 |
| b=1 | 58 | 11 |

1. **J48:**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | TP Rate | FP Rate | Precision | Recall | F-Measure | MCC | ROC Area | PRC Area | Class | Acc |
|  | 0.984 | 0.884 | 0.907 | 0.984 | 0.944 | 0.187 | 0.704 | 0.94 | 0 |  |
|  | 0.116 | 0.016 | 0.444 | 0.116 | 0.184 | 0.187 | 0.704 | 0.242 | 1 |  |
| Weighted Avg. | 0.895 | 0.796 | 0.86 | 0.895 | 0.866 | 0.187 | 0.704 | 0.869 |  | 89.497 |

**Confusion matrix:**

|  |  |  |
| --- | --- | --- |
| Classified as Actual | a=0 | b=1 |
| a = 0 | 597 | 10 |
| b=1 | 61 | 8 |

1. **Naïve Bayes:**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | TP Rate | FP Rate | Precision | Recall | F-Measure | MCC | ROC Area | PRC Area | Class | Acc |
|  | 0.97 | 0.797 | 0.915 | 0.97 | 0.942 | 0.247 | 0.792 | 0.964 | 0 |  |
|  | 0.203 | 0.03 | 0.438 | 0.203 | 0.277 | 0.247 | 0.792 | 0.306 | 1 |  |
| Weighted Avg. | 0.892 | 0.719 | 0.866 | 0.892 | 0.874 | 0.247 | 0.792 | 0.897 |  | 89.2012 |

**Confusion matrix:**

|  |  |  |
| --- | --- | --- |
| Classified as Actual | a=0 | b=1 |
| a = 0 | 589 | 18 |
| b=1 | 55 | 14 |

1. **Logistic Regression:**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | TP Rate | FP Rate | Precision | Recall | F-Measure | MCC | ROC Area | PRC Area | Class | Acc |
|  | 0.984 | 0.899 | 0.906 | 0.984 | 0.943 | 0.164 | 0.795 | 0.967 | 0 |  |
|  | 0.101 | 0.016 | 0.412 | 0.101 | 0.163 | 0.164 | 0.795 | 0.317 | 1 |  |
| Weighted Avg. | 0.893 | 0.809 | 0.855 | 0.893 | 0.863 | 0.164 | 0.795 | 0.901 |  | 89.3491 |

**Confusion matrix:**

|  |  |  |
| --- | --- | --- |
| Classified as Actual | a=0 | b=1 |
| a = 0 | 597 | 10 |
| b=1 | 62 | 7 |

1. **KNN:**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | TP Rate | FP Rate | Precision | Recall | F-Measure | MCC | ROC Area | PRC Area | Class | Acc |
|  | 0.998 | 1 | 0.898 | 0.998 | 0.945 | -0.013 | 0.739 | 0.953 | 0 |  |
|  | 0 | 0.002 | 0 | 0 | 0 | -0.013 | 0.739 | 0.215 | 1 |  |
| Weighted Avg. | 0.896 | 0.898 | 0.806 | 0.896 | 0.849 | -0.013 | 0.739 | 0.878 |  | 89.645 |

**Confusion matrix:**

|  |  |  |
| --- | --- | --- |
| Classified as Actual | a=0 | b=1 |
| a = 0 | 606 | 1 |
| b=1 | 69 | 0 |

1. **AdaBoost with J48:**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | TP Rate | FP Rate | Precision | Recall | F-Measure | MCC | ROC Area | PRC Area | Class | Acc |
|  | 0.947 | 0.783 | 0.914 | 0.947 | 0.93 | 0.196 | 0.714 | 0.947 | 0 |  |
|  | 0.217 | 0.053 | 0.319 | 0.217 | 0.259 | 0.196 | 0.714 | 0.278 | 1 |  |
| Weighted Avg. | 0.873 | 0.708 | 0.853 | 0.873 | 0.862 | 0.196 | 0.714 | 0.879 |  | 87.278 |

**Confusion matrix:**

|  |  |  |
| --- | --- | --- |
| Classified as Actual | a=0 | b=1 |
| a = 0 | 575 | 32 |
| b=1 | 54 | 15 |

1. **AdaBoost with Random Forest:**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | TP Rate | FP Rate | Precision | Recall | F-Measure | MCC | ROC Area | PRC Area | Class | Acc |
|  | 0.988 | 0.87 | 0.909 | 0.988 | 0.947 | 0.237 | 0.785 | 0.966 | 0 |  |
|  | 0.13 | 0.012 | 0.563 | 0.13 | 0.212 | 0.237 | 0.785 | 0.332 | 1 |  |
| Weighted Avg. | 0.901 | 0.782 | 0.874 | 0.901 | 0.872 | 0.237 | 0.785 | 0.901 |  | 90.0888 |

**Confusion matrix:**

|  |  |  |
| --- | --- | --- |
| Classified as Actual | a=0 | b=1 |
| a = 0 | 600 | 7 |
| b=1 | 60 | 9 |

1. **SVM:**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | TP Rate | FP Rate | Precision | Recall | F-Measure | MCC | ROC Area | PRC Area | Class | Acc |
|  | 1 | 1 | 0.898 | 1 | 0.946 | ? | 0.5 | 0.898 | 0 |  |
|  | 0 | 0 | ? | 0 | ? | ? | 0.5 | 0.102 | 1 |  |
| Weighted Avg. | 0.898 | 0.898 | ? | 0.898 | ? | ? | 0.5 | 0.817 |  | 89.7929 |

**Confusion matrix:**

|  |  |  |
| --- | --- | --- |
| Classified as Actual | a=0 | b=1 |
| a = 0 | 607 | 0 |
| b=1 | 69 | 0 |

Below are the accuracies and the evaluation metrics of the 40 models generated.

**Dataset with chi2 feature selection:**

1. **Random Forest:**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | TP Rate | FP Rate | Precision | Recall | F-Measure | MCC | ROC Area | PRC Area | Class | Acc |
|  | 0.712 | 0.316 | 0.752 | 0.712 | 0.732 | 0.393 | 0.727 | 0.763 | 0 |  |
|  | 0.684 | 0.288 | 0.638 | 0.684 | 0.66 | 0.393 | 0.727 | 0.649 | 1 |  |
| Weighted Avg. | 0.7 | 0.304 | 0.703 | 0.7 | 0.701 | 0.393 | 0.727 | 0.714 |  | 70 |

**Confusion matrix:**

|  |  |  |
| --- | --- | --- |
| Classified as Actual | a=0 | b=1 |
| a = 0 | 94 | 38 |
| b=1 | 31 | 67 |

1. **J48:**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | TP Rate | FP Rate | Precision | Recall | F-Measure | MCC | ROC Area | PRC Area | Class | Acc |
|  | 0.644 | 0.276 | 0.759 | 0.644 | 0.697 | 0.365 | 0.699 | 0.715 | 0 |  |
|  | 0.724 | 0.356 | 0.602 | 0.724 | 0.657 | 0.365 | 0.699 | 0.592 | 1 |  |
| Weighted Avg. | 0.678 | 0.31 | 0.692 | 0.678 | 0.68 | 0.365 | 0.699 | 0.663 |  | 67.8261 |

**Confusion matrix:**

|  |  |  |
| --- | --- | --- |
| Classified as Actual | a=0 | b=1 |
| a = 0 | 85 | 47 |
| b=1 | 27 | 71 |

1. **Naïve Bayes:**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | TP Rate | FP Rate | Precision | Recall | F-Measure | MCC | ROC Area | PRC Area | Class | Acc |
|  | 0.78 | 0.398 | 0.725 | 0.78 | 0.752 | 0.389 | 0.734 | 0.746 | 0 |  |
|  | 0.602 | 0.22 | 0.67 | 0.602 | 0.634 | 0.389 | 0.734 | 0.716 | 1 |  |
| Weighted Avg. | 0.704 | 0.322 | 0.702 | 0.704 | 0.702 | 0.389 | 0.734 | 0.733 |  | 70.4348 |

**Confusion matrix:**

|  |  |  |
| --- | --- | --- |
| Classified as Actual | a=0 | b=1 |
| a = 0 | 103 | 29 |
| b=1 | 39 | 59 |

1. **Logistic Regression:**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | TP Rate | FP Rate | Precision | Recall | F-Measure | MCC | ROC Area | PRC Area | Class | Acc |
|  | 0.788 | 0.378 | 0.738 | 0.788 | 0.762 | 0.417 | 0.774 | 0.8 | 0 |  |
|  | 0.622 | 0.212 | 0.685 | 0.622 | 0.652 | 0.417 | 0.774 | 0.75 | 1 |  |
| Weighted Avg. | 0.717 | 0.307 | 0.715 | 0.717 | 0.715 | 0.417 | 0.774 | 0.779 |  | 71.7391 |

**Confusion matrix:**

|  |  |  |
| --- | --- | --- |
| Classified as Actual | a=0 | b=1 |
| a = 0 | 104 | 28 |
| b=1 | 37 | 61 |

1. **KNN:**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | TP Rate | FP Rate | Precision | Recall | F-Measure | MCC | ROC Area | PRC Area | Class | Acc |
|  | 0.808 | 0.45 | 0.7 | 0.808 | 0.75 | 0.372 | 0.681 | 0.684 | 0 |  |
|  | 0.55 | 0.192 | 0.688 | 0.55 | 0.611 | 0.372 | 0.681 | 0.614 | 1 |  |
| Weighted Avg. | 0.696 | 0.338 | 0.695 | 0.696 | 0.69 | 0.372 | 0.681 | 0.654 |  | 69.5652 |

**Confusion matrix:**

|  |  |  |
| --- | --- | --- |
| Classified as Actual | a=0 | b=1 |
| a = 0 | 105 | 25 |
| b=1 | 45 | 55 |

1. **AdaBoost with J48:**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | TP Rate | FP Rate | Precision | Recall | F-Measure | MCC | ROC Area | PRC Area | Class | Acc |
|  | 0.674 | 0.378 | 0.706 | 0.674 | 0.69 | 0.295 | 0.688 | 0.707 | 0 |  |
|  | 0.622 | 0.326 | 0.587 | 0.622 | 0.604 | 0.295 | 0.688 | 0.629 | 1 |  |
| Weighted Avg. | 0.652 | 0.355 | 0.655 | 0.652 | 0.653 | 0.295 | 0.688 | 0.674 |  | 65.2174 |

**Confusion matrix:**

|  |  |  |
| --- | --- | --- |
| Classified as Actual | a=0 | b=1 |
| a = 0 | 89 | 43 |
| b=1 | 37 | 61 |

1. **AdaBoost with Random Forest:**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | TP Rate | FP Rate | Precision | Recall | F-Measure | MCC | ROC Area | PRC Area | Class | Acc |
|  | 0.689 | 0.286 | 0.765 | 0.689 | 0.725 | 0.399 | 0.73 | 0.757 | 0 |  |
|  | 0.714 | 0.311 | 0.631 | 0.714 | 0.67 | 0.399 | 0.73 | 0.649 | 1 |  |
| Weighted Avg. | 0.7 | 0.296 | 0.708 | 0.7 | 0.702 | 0.399 | 0.73 | 0.711 |  | 70 |

**Confusion matrix:**

|  |  |  |
| --- | --- | --- |
| Classified as Actual | a=0 | b=1 |
| a = 0 | 91 | 41 |
| b=1 | 28 | 70 |

1. **SVM:**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | TP Rate | FP Rate | Precision | Recall | F-Measure | MCC | ROC Area | PRC Area | Class | Acc |
|  | 0.788 | 0.367 | 0.743 | 0.788 | 0.765 | 0.426 | 0.71 | 0.707 | 0 |  |
|  | 0.633 | 0.212 | 0.689 | 0.633 | 0.66 | 0.426 | 0.71 | 0.592 | 1 |  |
| Weighted Avg. | 0.722 | 0.301 | 0.72 | 0.722 | 0.72 | 0.426 | 0.71 | 0.658 |  | 72.1739 |

**Confusion matrix:**

|  |  |  |
| --- | --- | --- |
| Classified as Actual | a=0 | b=1 |
| a = 0 | 104 | 28 |
| b=1 | 36 | 62 |

**Dataset with f-test feature selection:**

1. **Random Forest:**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | TP Rate | FP Rate | Precision | Recall | F-Measure | MCC | ROC Area | PRC Area | Class | Acc |
|  | 0.697 | 0.398 | 0.702 | 0.697 | 0.700 | 0.299 | 0.710 | 0.746 | 0 |  |
|  | 0.602 | 0.303 | 0.596 | 0.602 | 0.599 | 0.299 | 0.710 | 0.634 | 1 |  |
| Weighted Avg. | 0.657 | 0.358 | 0.657 | 0.657 | 0.657 | 0.299 | 0.710 | 0.698 |  | 65.6522 |

**Confusion matrix:**

|  |  |  |
| --- | --- | --- |
| Classified as Actual | a=0 | b=1 |
| a = 0 | 92 | 40 |
| b=1 | 39 | 59 |

1. **J48:**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | TP Rate | FP Rate | Precision | Recall | F-Measure | MCC | ROC Area | PRC Area | Class | Acc |
|  | 0.667 | 0.337 | 0.727 | 0.667 | 0.696 | 0.327 | 0.694 | 0.69 | 0 |  |
|  | 0.663 | 0.333 | 0.596 | 0.663 | 0.628 | 0.327 | 0.694 | 0.624 | 1 |  |
| Weighted Avg. | 0.665 | 0.335 | 0.671 | 0.665 | 0.667 | 0.327 | 0.694 | 0.662 |  | 66.5217 |

**Confusion matrix:**

|  |  |  |
| --- | --- | --- |
| Classified as Actual | a=0 | b=1 |
| a = 0 | 88 | 44 |
| b=1 | 33 | 65 |

1. **Naïve Bayes:**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | TP Rate | FP Rate | Precision | Recall | F-Measure | MCC | ROC Area | PRC Area | Class | Acc |
|  | 0.788 | 0.398 | 0.727 | 0.788 | 0.756 | 0.398 | 0.734 | 0.748 | 0 |  |
|  | 0.602 | 0.212 | 0.678 | 0.602 | 0.638 | 0.398 | 0.734 | 0.719 | 1 |  |
| Weighted Avg. | 0.709 | 0.319 | 0.706 | 0.709 | 0.706 | 0.398 | 0.734 | 0.736 |  | 70.8696 |

**Confusion matrix:**

|  |  |  |
| --- | --- | --- |
| Classified as Actual | a=0 | b=1 |
| a = 0 | 104 | 28 |
| b=1 | 39 | 59 |

1. **Logistic Regression:**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | TP Rate | FP Rate | Precision | Recall | F-Measure | MCC | ROC Area | PRC Area | Class | Acc |
|  | 0.795 | 0.388 | 0.734 | 0.795 | 0.764 | 0.416 | 0.771 | 0.797 | 0 |  |
|  | 0.612 | 0.205 | 0.690 | 0.612 | 0.649 | 0.416 | 0.771 | 0.749 | 1 |  |
| Weighted Avg. | 0.717 | 0.310 | 0.715 | 0.717 | 0.715 | 0.416 | 0.771 | 0.776 |  | 71.7391 |

**Confusion matrix:**

|  |  |  |
| --- | --- | --- |
| Classified as Actual | a=0 | b=1 |
| a = 0 | 105 | 27 |
| b=1 | 38 | 60 |

1. **KNN:**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | TP Rate | FP Rate | Precision | Recall | F-Measure | MCC | ROC Area | PRC Area | Class | Acc |
|  | 0.833 | 0.327 | 0.775 | 0.833 | 0.803 | 0.516 | 0.766 | 0.777 | 0 |  |
|  | 0.673 | 0.167 | 0.750 | 0.673 | 0.710 | 0.516 | 0.766 | 0.684 | 1 |  |
| Weighted Avg. | 0.765 | 0.258 | 0.764 | 0.765 | 0.763 | 0.516 | 0.766 | 0.737 |  | 76.5217 |

**Confusion matrix:**

|  |  |  |
| --- | --- | --- |
| Classified as Actual | a=0 | b=1 |
| a = 0 | 110 | 22 |
| b=1 | 32 | 66 |

1. **AdaBoost with J48:**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | TP Rate | FP Rate | Precision | Recall | F-Measure | MCC | ROC Area | PRC Area | Class | Acc |
|  | 0.598 | 0.347 | 0.699 | 0.598 | 0.645 | 0.249 | 0.697 | 0.746 | 0 |  |
|  | 0.653 | 0.402 | 0.547 | 0.653 | 0.595 | 0.249 | 0.697 | 0.648 | 1 |  |
| Weighted Avg. | 0.622 | 0.370 | 0.634 | 0.622 | 0.624 | 0.249 | 0.697 | 0.704 |  | 62.1739 |

**Confusion matrix:**

|  |  |  |
| --- | --- | --- |
| Classified as Actual | a=0 | b=1 |
| a = 0 | 598 | 9 |
| b=1 | 58 | 11 |

1. **AdaBoost with Random Forest:**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | TP Rate | FP Rate | Precision | Recall | F-Measure | MCC | ROC Area | PRC Area | Class | Acc |
|  | 0.652 | 0.347 | 0.717 | 0.652 | 0.683 | 0.302 | 0.702 | 0.735 | 0 |  |
|  | 0.653 | 0.348 | 0.582 | 0.653 | 0.615 | 0.302 | 0.702 | 0.625 | 1 |  |
| Weighted Avg. | 0.652 | 0.348 | 0.659 | 0.652 | 0.654 | 0.302 | 0.702 | 0.688 |  | 65.2174 |

**Confusion matrix:**

|  |  |  |
| --- | --- | --- |
| Classified as Actual | a=0 | b=1 |
| a = 0 | 86 | 46 |
| b=1 | 34 | 64 |

1. **SVM:**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | TP Rate | FP Rate | Precision | Recall | F-Measure | MCC | ROC Area | PRC Area | Class | Acc |
|  | 0.795 | 0.408 | 0.724 | 0.795 | 0.758 | 0.397 | 0.694 | 0.693 | 0 |  |
|  | 0.592 | 0.205 | 0.682 | 0.592 | 0.634 | 0.397 | 0.694 | 0.578 | 1 |  |
| Weighted Avg. | 0.709 | 0.321 | 0.706 | 0.709 | 0.705 | 0.397 | 0.694 | 0.644 |  | 70.8696 |

**Confusion matrix:**

|  |  |  |
| --- | --- | --- |
| Classified as Actual | a=0 | b=1 |
| a = 0 | 105 | 27 |
| b=1 | 40 | 58 |

**Dataset with mutual info feature selection:**

1. **Random Forest:**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | TP Rate | FP Rate | Precision | Recall | F-Measure | MCC | ROC Area | PRC Area | Class | Acc |
|  | 0.652 | 0.357 | 0.711 | 0.652 | 0.680 | 0.292 | 0.711 | 0.7 | 0 |  |
|  | 0.643 | 0.348 | 0.578 | 0.643 | 0.609 | 0.292 | 0.711 | 0.701 | 1 |  |
| Weighted Avg. | 0.648 | 0.353 | 0.654 | 0.648 | 0.650 | 0.292 | 0.711 | 0.701 |  | 64.7826 |

**Confusion matrix:**

|  |  |  |
| --- | --- | --- |
| Classified as Actual | a=0 | b=1 |
| a = 0 | 86 | 46 |
| b=1 | 35 | 63 |

1. **J48:**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | TP Rate | FP Rate | Precision | Recall | F-Measure | MCC | ROC Area | PRC Area | Class | Acc |
|  | 0.803 | 0.408 | 0.726 | 0.803 | 0.763 | 0.406 | 0.696 | 0.702 | 0 |  |
|  | 0.592 | 0.197 | 0.690 | 0.592 | 0.637 | 0.406 | 0.696 | 0.617 | 1 |  |
| Weighted Avg. | 0.713 | 0.318 | 0.711 | 0.713 | 0.709 | 0.406 | 0.696 | 0.666 |  | 71.3043 |

**Confusion matrix:**

|  |  |  |
| --- | --- | --- |
| Classified as Actual | a=0 | b=1 |
| a = 0 | 106 | 26 |
| b=1 | 40 | 58 |

1. **Naïve Bayes:**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | TP Rate | FP Rate | Precision | Recall | F-Measure | MCC | ROC Area | PRC Area | Class | Acc |
|  | 0.833 | 0.459 | 0.710 | 0.833 | 0.767 | 0.395 | 0.751 | 0.76 | 0 |  |
|  | 0.541 | 0.167 | 0.707 | 0.541 | 0.613 | 0.395 | 0.751 | 0.728 | 1 |  |
| Weighted Avg. | 0.709 | 0.335 | 0.708 | 0.709 | 0.701 | 0.395 | 0.751 | 0.747 |  | 70.8696 |

**Confusion matrix:**

|  |  |  |
| --- | --- | --- |
| Classified as Actual | a=0 | b=1 |
| a = 0 | 110 | 22 |
| b=1 | 45 | 53 |

1. **Logistic Regression:**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | TP Rate | FP Rate | Precision | Recall | F-Measure | MCC | ROC Area | PRC Area | Class | Acc |
|  | 0.795 | 0.388 | 0.734 | 0.795 | 0.764 | 0.416 | 0.769 | 0.814 | 0 |  |
|  | 0.612 | 0.205 | 0.690 | 0.612 | 0.649 | 0.416 | 0.769 | 0.741 | 1 |  |
| Weighted Avg. | 0.717 | 0.310 | 0.715 | 0.717 | 0.715 | 0.416 | 0.769 | 0.783 |  | 71.7391 |

**Confusion matrix:**

|  |  |  |
| --- | --- | --- |
| Classified as Actual | a=0 | b=1 |
| a = 0 | 105 | 27 |
| b=1 | 38 | 60 |

1. **KNN:**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | TP Rate | FP Rate | Precision | Recall | F-Measure | MCC | ROC Area | PRC Area | Class | Acc |
|  | 0.841 | 0.490 | 0.698 | 0.841 | 0.763 | 0.376 | 0.746 | 0.745 | 0 |  |
|  | 0.51 | 0.159 | 0.704 | 0.510 | 0.592 | 0.376 | 0.746 | 0.693 | 1 |  |
| Weighted Avg. | 0.7 | 0.349 | 0.701 | 0.700 | 0.690 | 0.376 | 0.746 | 0.723 |  | 70 |

**Confusion matrix:**

|  |  |  |
| --- | --- | --- |
| Classified as Actual | a=0 | b=1 |
| a = 0 | 111 | 21 |
| b=1 | 48 | 50 |

1. **AdaBoost with J48:**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | TP Rate | FP Rate | Precision | Recall | F-Measure | MCC | ROC Area | PRC Area | Class | Acc |
|  | 0.644 | 0.398 | 0.685 | 0.644 | 0.664 | 0.244 | 0.690 | 0.704 | 0 |  |
|  | 0.602 | 0.356 | 0.557 | 0.602 | 0.578 | 0.244 | 0.690 | 0.63 | 1 |  |
| Weighted Avg. | 0.626 | 0.380 | 0.631 | 0.626 | 0.628 | 0.244 | 0.690 | 0.673 |  | 62.6087 |

**Confusion matrix:**

|  |  |  |
| --- | --- | --- |
| Classified as Actual | a=0 | b=1 |
| a = 0 | 85 | 47 |
| b=1 | 39 | 59 |

1. **AdaBoost with Random Forest:**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | TP Rate | FP Rate | Precision | Recall | F-Measure | MCC | ROC Area | PRC Area | Class | Acc |
|  | 0.652 | 0.367 | 0.705 | 0.652 | 0.677 | 0.282 | 0.684 | 0.713 | 0 |  |
|  | 0.633 | 0.348 | 0.574 | 0.633 | 0.602 | 0.282 | 0.684 | 0.623 | 1 |  |
| Weighted Avg. | 0.643 | 0.359 | 0.649 | 0.643 | 0.645 | 0.282 | 0.684 | 0.675 |  | 64.3478 |

**Confusion matrix:**

|  |  |  |
| --- | --- | --- |
| Classified as Actual | a=0 | b=1 |
| a = 0 | 86 | 46 |
| b=1 | 36 | 62 |

1. **SVM:**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | TP Rate | FP Rate | Precision | Recall | F-Measure | MCC | ROC Area | PRC Area | Class | Acc |
|  | 0.886 | 0.602 | 0.665 | 0.886 | 0.76 | 0.332 | 0.642 | 0.654 | 0 |  |
|  | 0.398 | 0.114 | 0.722 | 0.398 | 0.513 | 0.332 | 0.642 | 0.544 | 1 |  |
| Weighted Avg. | 0.398 | 0.114 | 0.722 | 0.398 | 0.513 | 0.332 | 0.642 | 0.544 |  | 67.8261 |

**Confusion matrix:**

|  |  |  |
| --- | --- | --- |
| Classified as Actual | a=0 | b=1 |
| a = 0 | 117 | 15 |
| b=1 | 59 | 39 |

**Dataset with recursive feature elimination:**

1. **Random Forest:**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | TP Rate | FP Rate | Precision | Recall | F-Measure | MCC | ROC Area | PRC Area | Class | Acc |
|  | 0.606 | 0.327 | 0.714 | 0.606 | 0.656 | 0.277 | 0.673 | 0.726 | 0 |  |
|  | 0.673 | 0.394 | 0.559 | 0.673 | 0.611 | 0.277 | 0.673 | 0.58 | 1 |  |
| Weighted Avg. | 0.635 | 0.355 | 0.648 | 0.635 | 0.637 | 0.277 | 0.673 | 0.664 |  | 63.4783 |

**Confusion matrix:**

|  |  |  |
| --- | --- | --- |
| Classified as Actual | a=0 | b=1 |
| a = 0 | 80 | 52 |
| b=1 | 32 | 66 |

1. **J48:**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | TP Rate | FP Rate | Precision | Recall | F-Measure | MCC | ROC Area | PRC Area | Class | Acc |
|  | 0.591 | 0.296 | 0.729 | 0.591 | 0.653 | 0.292 | 0.672 | 0.726 | 0 |  |
|  | 0.704 | 0.409 | 0.561 | 0.704 | 0.624 | 0.292 | 0.672 | 0.548 | 1 |  |
| Weighted Avg. | 0.639 | 0.344 | 0.657 | 0.639 | 0.641 | 0.292 | 0.672 | 0.65 |  | 63.913 |

**Confusion matrix:**

|  |  |  |
| --- | --- | --- |
| Classified as Actual | a=0 | b=1 |
| a = 0 | 78 | 54 |
| b=1 | 29 | 69 |

1. **Naïve Bayes:**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | TP Rate | FP Rate | Precision | Recall | F-Measure | MCC | ROC Area | PRC Area | Class | Acc |
|  | 0.818 | 0.551 | 0.667 | 0.818 | 0.735 | 0.29 | 0.72 | 0.732 | 0 |  |
|  | 0.449 | 0.182 | 0.647 | 0.449 | 0.53 | 0.29 | 0.72 | 0.703 | 1 |  |
| Weighted Avg. | 0.661 | 0.394 | 0.658 | 0.661 | 0.648 | 0.29 | 0.72 | 0.719 |  | 66.087 |

**Confusion matrix:**

|  |  |  |
| --- | --- | --- |
| Classified as Actual | a=0 | b=1 |
| a = 0 | 108 | 24 |
| b=1 | 54 | 44 |

1. **Logistic Regression:**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | TP Rate | FP Rate | Precision | Recall | F-Measure | MCC | ROC Area | PRC Area | Class | Acc |
|  | 0.826 | 0.52 | 0.681 | 0.826 | 0.747 | 0.328 | 0.758 | 0.789 | 0 |  |
|  | 0.48 | 0.174 | 0.671 | 0.48 | 0.56 | 0.328 | 0.758 | 0.726 | 1 |  |
| Weighted Avg. | 0.678 | 0.373 | 0.677 | 0.678 | 0.667 | 0.328 | 0.758 | 0.762 |  | 67.8261 |

**Confusion matrix:**

|  |  |  |
| --- | --- | --- |
| Classified as Actual | a=0 | b=1 |
| a = 0 | 109 | 23 |
| b=1 | 51 | 47 |

1. **KNN:**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | TP Rate | FP Rate | Precision | Recall | F-Measure | MCC | ROC Area | PRC Area | Class | Acc |
|  | 0.765 | 0.429 | 0.706 | 0.765 | 0.735 | 0.343 | 0.699 | 0.693 | 0 |  |
|  | 0.571 | 0.235 | 0.644 | 0.571 | 0.605 | 0.343 | 0.699 | 0.634 | 1 |  |
| Weighted Avg. | 0.683 | 0.346 | 0.68 | 0.683 | 0.68 | 0.343 | 0.699 | 0.668 |  | 68.2609 |

**Confusion matrix:**

|  |  |  |
| --- | --- | --- |
| Classified as Actual | a=0 | b=1 |
| a = 0 | 101 | 31 |
| b=1 | 42 | 56 |

1. **AdaBoost with J48:**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | TP Rate | FP Rate | Precision | Recall | F-Measure | MCC | ROC Area | PRC Area | Class | Acc |
|  | 0.561 | 0.316 | 0.705 | 0.561 | 0.624 | 0.243 | 0.649 | 0.684 | 0 |  |
|  | 0.684 | 0.439 | 0.536 | 0.684 | 0.601 | 0.243 | 0.649 | 0.608 | 1 |  |
| Weighted Avg. | 0.613 | 0.369 | 0.633 | 0.613 | 0.614 | 0.243 | 0.649 | 0.651 |  | 61.3043 |

**Confusion matrix:**

|  |  |  |
| --- | --- | --- |
| Classified as Actual | a=0 | b=1 |
| a = 0 | 74 | 58 |
| b=1 | 31 | 67 |

1. **AdaBoost with Random Forest:**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | TP Rate | FP Rate | Precision | Recall | F-Measure | MCC | ROC Area | PRC Area | Class | Acc |
|  | 0.636 | 0.357 | 0.706 | 0.636 | 0.669 | 0.276 | 0.679 | 0.714 | 0 |  |
|  | 0.643 | 0.364 | 0.568 | 0.643 | 0.603 | 0.276 | 0.679 | 0.599 | 1 |  |
| Weighted Avg. | 0.639 | 0.36 | 0.647 | 0.639 | 0.641 | 0.276 | 0.679 | 0.665 |  | 63.913 |

**Confusion matrix:**

|  |  |  |
| --- | --- | --- |
| Classified as Actual | a=0 | b=1 |
| a = 0 | 84 | 48 |
| b=1 | 35 | 63 |

1. **SVM:**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | TP Rate | FP Rate | Precision | Recall | F-Measure | MCC | ROC Area | PRC Area | Class | Acc |
|  | 0.788 | 0.439 | 0.707 | 0.788 | 0.746 | 0.359 | 0.675 | 0.679 | 0 |  |
|  | 0.561 | 0.212 | 0.663 | 0.561 | 0.608 | 0.359 | 0.675 | 0.559 | 1 |  |
| Weighted Avg. | 0.691 | 0.342 | 0.688 | 0.691 | 0.687 | 0.359 | 0.675 | 0.628 |  | 69.1304 |

**Confusion matrix:**

|  |  |  |
| --- | --- | --- |
| Classified as Actual | a=0 | b=1 |
| a = 0 | 104 | 28 |
| b=1 | 43 | 55 |

**Dataset with Random Forest:**

1. **Random Forest:**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | TP Rate | FP Rate | Precision | Recall | F-Measure | MCC | ROC Area | PRC Area | Class | Acc |
|  | 0.606 | 0.378 | 0.684 | 0.606 | 0.643 | 0.226 | 0.679 | 0.742 | 0 |  |
|  | 0.622 | 0.394 | 0.54 | 0.622 | 0.578 | 0.226 | 0.679 | 0.571 | 1 |  |
| Weighted Avg. | 0.613 | 0.385 | 0.622 | 0.613 | 0.615 | 0.226 | 0.679 | 0.669 |  | 61.3043 |

**Confusion matrix:**

|  |  |  |
| --- | --- | --- |
| Classified as Actual | a=0 | b=1 |
| a = 0 | 80 | 52 |
| b=1 | 37 | 61 |

1. **J48:**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | TP Rate | FP Rate | Precision | Recall | F-Measure | MCC | ROC Area | PRC Area | Class | Acc |
|  | 0.795 | 0.49 | 0.686 | 0.795 | 0.737 | 0.32 | 0.699 | 0.702 | 0 |  |
|  | 0.51 | 0.205 | 0.649 | 0.51 | 0.571 | 0.32 | 0.699 | 0.632 | 1 |  |
| Weighted Avg. | 0.674 | 0.368 | 0.671 | 0.674 | 0.666 | 0.32 | 0.699 | 0.672 |  | 67.3913 |

**Confusion matrix:**

|  |  |  |
| --- | --- | --- |
| Classified as Actual | a=0 | b=1 |
| a = 0 | 105 | 27 |
| b=1 | 48 | 50 |

1. **Naïve Bayes:**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | TP Rate | FP Rate | Precision | Recall | F-Measure | MCC | ROC Area | PRC Area | Class | Acc |
|  | 0.879 | 0.561 | 0.678 | 0.879 | 0.766 | 0.36 | 0.677 | 0.691 | 0 |  |
|  | 0.439 | 0.121 | 0.729 | 0.439 | 0.548 | 0.36 | 0.677 | 0.684 | 1 |  |
| Weighted Avg. | 0.691 | 0.374 | 0.7 | 0.691 | 0.673 | 0.36 | 0.677 | 0.688 |  | 69.1304 |

**Confusion matrix:**

|  |  |  |
| --- | --- | --- |
| Classified as Actual | a=0 | b=1 |
| a = 0 | 116 | 16 |
| b=1 | 55 | 43 |

1. **Logistic Regression:**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | TP Rate | FP Rate | Precision | Recall | F-Measure | MCC | ROC Area | PRC Area | Class | Acc |
|  | 0.879 | 0.582 | 0.671 | 0.879 | 0.761 | 0.34 | 0.755 | 0.788 | 0 |  |
|  | 0.418 | 0.121 | 0.719 | 0.418 | 0.529 | 0.34 | 0.755 | 0.725 | 1 |  |
| Weighted Avg. | 0.683 | 0.385 | 0.691 | 0.683 | 0.662 | 0.34 | 0.755 | 0.761 |  | 68.2609 |

**Confusion matrix:**

|  |  |  |
| --- | --- | --- |
| Classified as Actual | a=0 | b=1 |
| a = 0 | 116 | 16 |
| b=1 | 57 | 41 |

1. **KNN:**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | TP Rate | FP Rate | Precision | Recall | F-Measure | MCC | ROC Area | PRC Area | Class | Acc |
|  | 0.636 | 0.337 | 0.718 | 0.636 | 0.675 | 0.296 | 0.701 | 0.711 | 0 |  |
|  | 0.663 | 0.364 | 0.575 | 0.663 | 0.616 | 0.296 | 0.701 | 0.634 | 1 |  |
| Weighted Avg. | 0.648 | 0.348 | 0.657 | 0.648 | 0.65 | 0.296 | 0.701 | 0.678 |  | 64.7826 |

**Confusion matrix:**

|  |  |  |
| --- | --- | --- |
| Classified as Actual | a=0 | b=1 |
| a = 0 | 84 | 48 |
| b=1 | 33 | 65 |

1. **AdaBoost with J48:**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | TP Rate | FP Rate | Precision | Recall | F-Measure | MCC | ROC Area | PRC Area | Class | Acc |
|  | 0.818 | 0.52 | 0.679 | 0.818 | 0.742 | 0.319 | 0.696 | 0.707 | 0 |  |
|  | 0.48 | 0.182 | 0.662 | 0.48 | 0.556 | 0.319 | 0.696 | 0.628 | 1 |  |
| Weighted Avg. | 0.674 | 0.376 | 0.672 | 0.674 | 0.663 | 0.319 | 0.696 | 0.673 |  | 67.3913 |

**Confusion matrix:**

|  |  |  |
| --- | --- | --- |
| Classified as Actual | a=0 | b=1 |
| a = 0 | 108 | 24 |
| b=1 | 51 | 47 |

1. **AdaBoost with Random Forest:**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | TP Rate | FP Rate | Precision | Recall | F-Measure | MCC | ROC Area | PRC Area | Class | Acc |
|  | 0.711 | 0.368 | 0.733 | 0.711 | 0.722 | 0.341 | 0.708 | 0.779 | 0 |  |
|  | 0.632 | 0.289 | 0.606 | 0.632 | 0.619 | 0.341 | 0.708 | 0.585 | 1 |  |
| Weighted Avg. | 0.678 | 0.336 | 0.68 | 0.678 | 0.679 | 0.341 | 0.708 | 0.699 |  | 67.8261 |

**Confusion matrix:**

|  |  |  |
| --- | --- | --- |
| Classified as Actual | a=0 | b=1 |
| a = 0 | 96 | 39 |
| b=1 | 35 | 60 |

1. **SVM:**

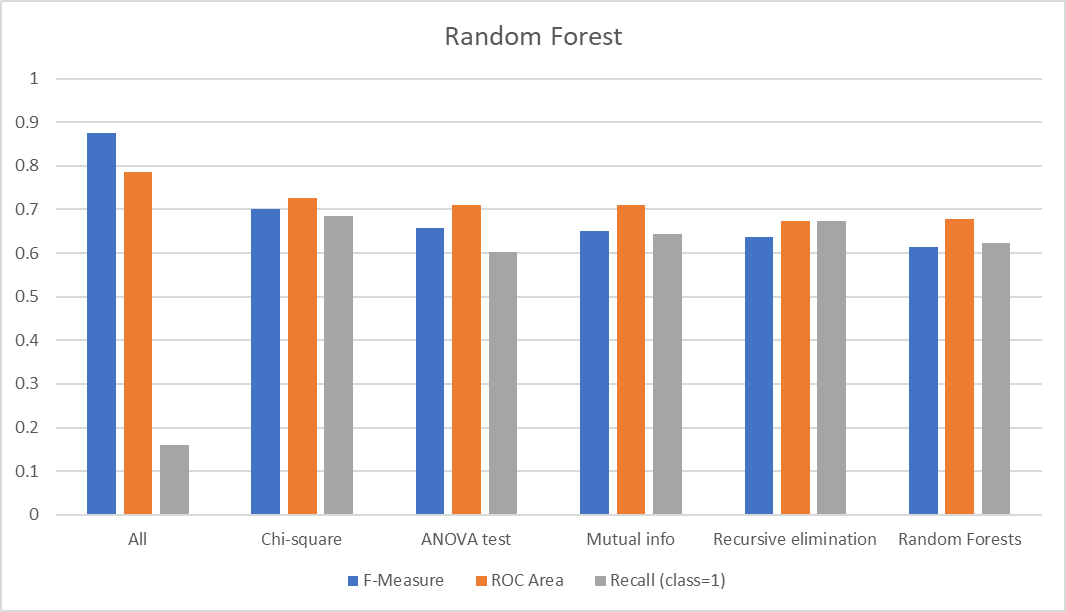
|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | TP Rate | FP Rate | Precision | Recall | F-Measure | MCC | ROC Area | PRC Area | Class | Acc |
|  | 0.803 | 0.469 | 0.697 | 0.803 | 0.746 | 0.349 | 0.667 | 0.673 | 0 |  |
|  | 0.531 | 0.197 | 0.667 | 0.531 | 0.591 | 0.349 | 0.667 | 0.554 | 1 |  |
| Weighted Avg. | 0.687 | 0.353 | 0.684 | 0.687 | 0.68 | 0.349 | 0.667 | 0.622 |  | 68.6957 |

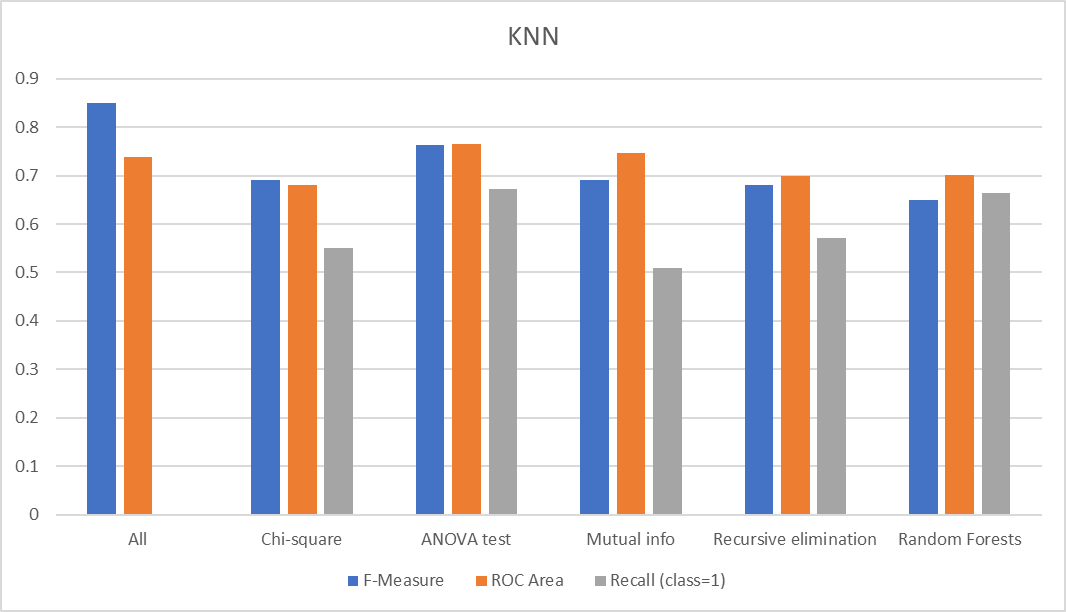
**Confusion matrix:**

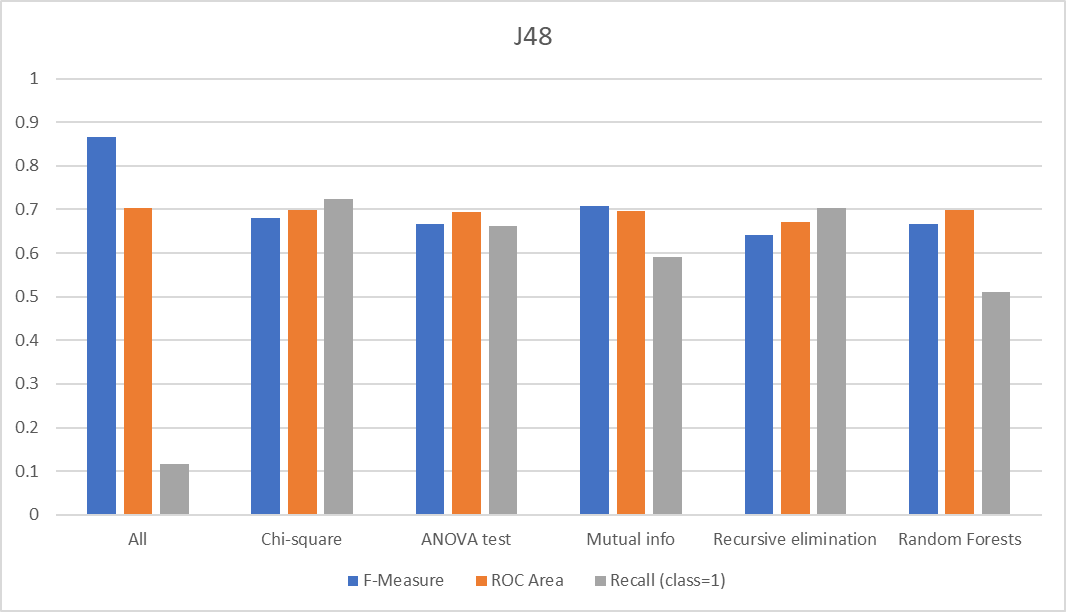
|  |  |  |
| --- | --- | --- |
| Classified as Actual | a=0 | b=1 |
| a = 0 | 106 | 26 |
| b=1 | 46 | 52 |

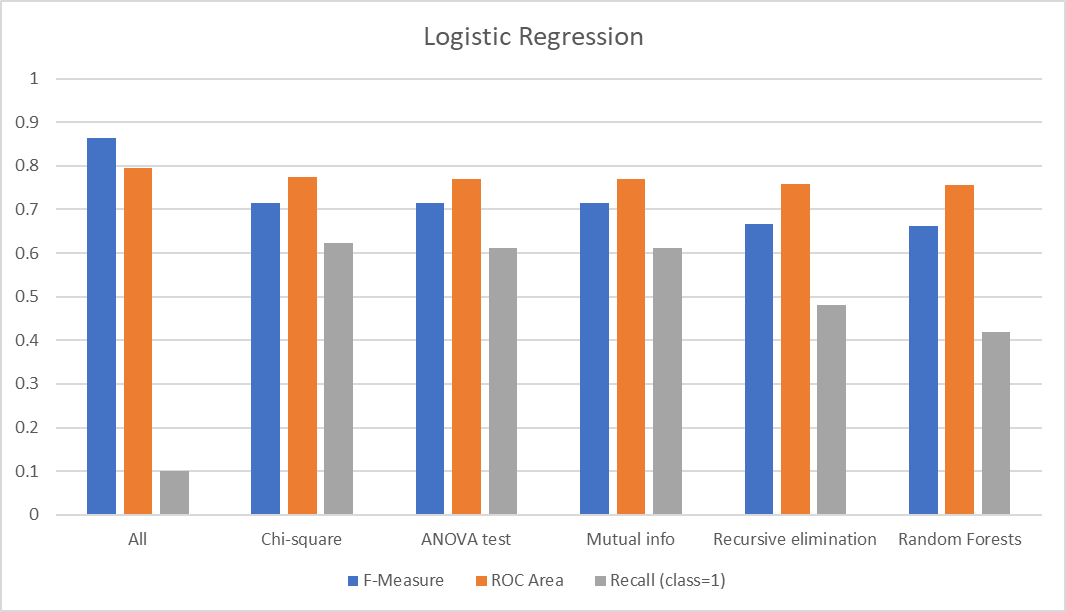
**RESULT VISUALIZATION**

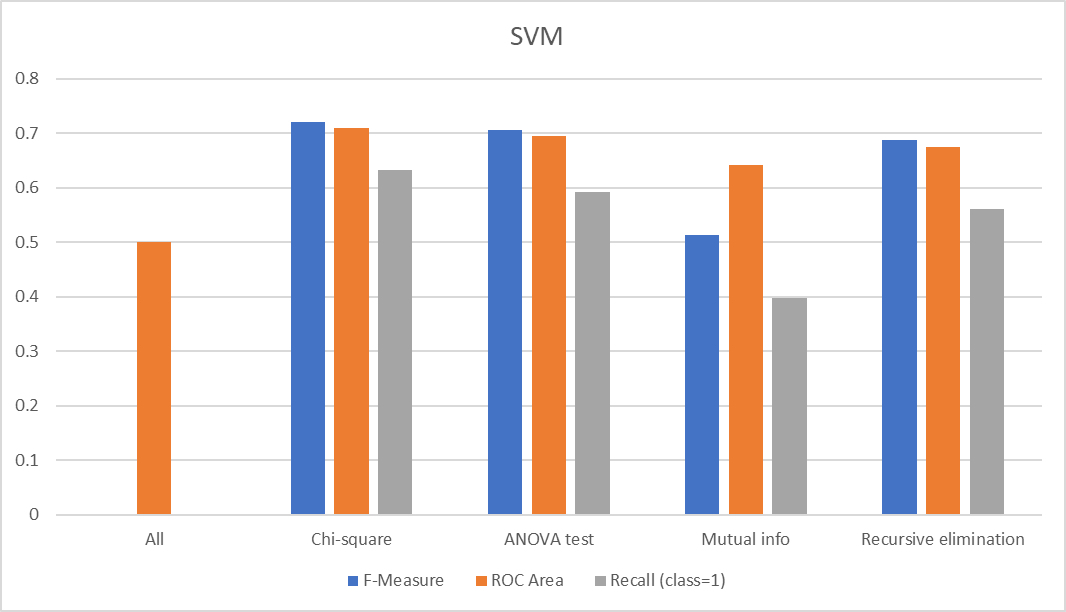
More than accuracy, **recall (class=1), F-measure and ROC** **area** play an important role in determining the best model for this problem.

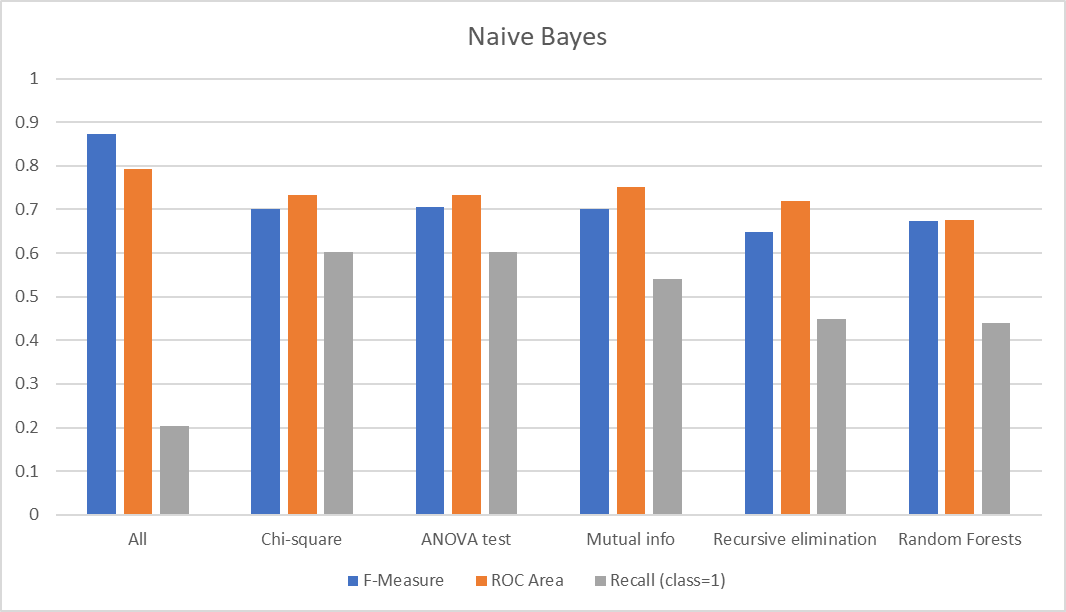
****

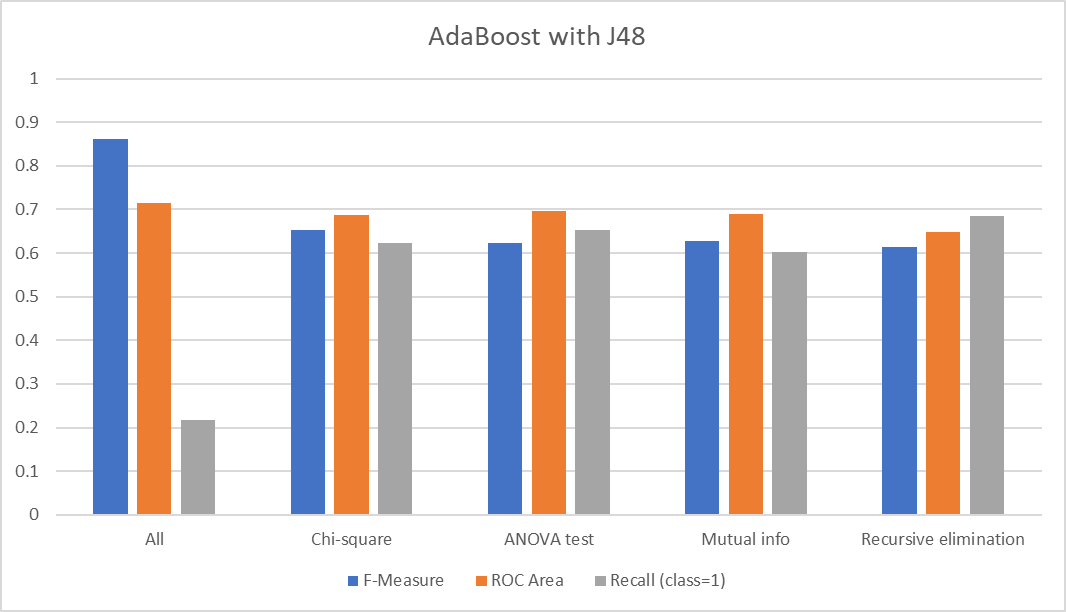
****

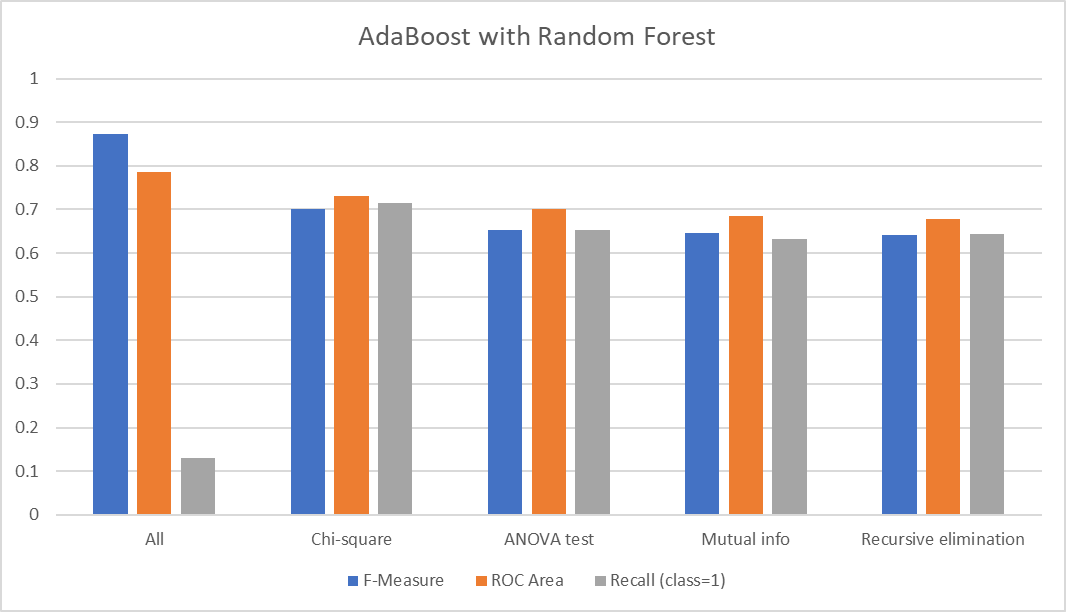
****

****

****

****

****

****

**Summary of the evaluation metrics:**



**FINAL MODEL SELECTION:**

Along with the important evaluation metrics like F-measure, ROC, and accuracy, this project requires a model to have high recall because a physician can’t afford to not recommend a RADT in a case that would actually be positive in RADT for GAS.

Hence, considering all the evaluation metrics, the best model is **KNN (k=13) when trained with ANOVA f-test features selected dataset on a 66% - 33% train-test split**, **with an accuracy of 76.52%.**

Also, it makes sense that KNN is the best model as most of the data is binary and the dataset is small. Getting an appropriate prediction depends on the k closest neighbors, and the binary data of the correlated attributes gives the best output in KNN as there’s a either 0% or 100% chance of getting it right.

**Comparing the best model trained with selected features and its base model:**

**KNN trained with all features:**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | TP Rate | FP Rate | Precision | Recall | F-Measure | MCC | ROC Area | PRC Area | Class | Acc |
|  | 0.998 | 1 | 0.898 | 0.998 | 0.945 | -0.013 | 0.739 | 0.953 | 0 |  |
|  | 0 | 0.002 | 0 | 0 | 0 | -0.013 | 0.739 | 0.215 | 1 |  |
| Weighted Avg. | 0.896 | 0.898 | 0.806 | 0.896 | 0.849 | -0.013 | 0.739 | 0.878 |  | 89.645 |

**Confusion matrix:**

|  |  |  |
| --- | --- | --- |
| Classified as Actual | a=0 | b=1 |
| a = 0 | 606 | 1 |
| b=1 | 69 | 0 |

**KNN trained with only features selected from ANOVA f-test:**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | TP Rate | FP Rate | Precision | Recall | F-Measure | MCC | ROC Area | PRC Area | Class | Acc |
|  | 0.833 | 0.327 | 0.775 | 0.833 | 0.803 | 0.516 | 0.766 | 0.777 | 0 |  |
|  | 0.673 | 0.167 | 0.750 | 0.673 | 0.710 | 0.516 | 0.766 | 0.684 | 1 |  |
| Weighted Avg. | 0.765 | 0.258 | 0.764 | 0.765 | 0.763 | 0.516 | 0.766 | 0.737 |  | 76.5217 |

**Confusion matrix:**

|  |  |  |
| --- | --- | --- |
| Classified as Actual | a=0 | b=1 |
| a = 0 | 110 | 22 |
| b=1 | 32 | 66 |

Even though the accuracy of the base model trained with all features is much higher than the model trained with features selected from ANOVA f-test, the recall for class=1 is 0 for the base model, which indicates it is not well trained (underfitted) for class = 1. This goes against the main goal of this project.

On the other hand, the KNN model trained with features selected from ANOVA f-test performs well, with the highest recall and accuracy among the models trained with features selected from various feature selection methods. This indicates the model learnt the most from the given amount of data.

**DISCUSSION AND CONCLUSION:**

Below are the points we observed and learned while doing this project.

1. A model with a high accuracy does not mean it is a good model, as it could be learning only one class, and test set could have a high distribution of that class. Cross validation creates class imbalance in a few cases. One must perform **stratified** **cross validation** when dealing with small amount of data though the initial dataset is not highly imbalanced.
2. Decision trees are susceptible to overfitting.
3. Independent features have a huge impact on processing and as well the prediction when it comes to modeling with small amounts of data. Feature selection is important.
4. Accuracies vary a lot even for a single misclassification when the dataset is small.
5. For this dataset, good models have accuracies below 80% as the dataset is small and these models have a huge role to play in medicine for this problem given there’s more data.
6. There is no best feature selection method. Just like there is no best set of input variables or best machine learning algorithm.

In conclusion, after a series of experiments conducted using various data imputation methods, scaling methods, attribute selection methods, 10-fold cross validation and 66-34 percentage train-test split, 8 different algorithms, and evaluation metrics on the 676 cases of pharyngitis in children dataset, we found out that physicians can rely on a **KNN(k=13) model** trained with features selected from **ANOVA f-test** to help them decide if to recommend a child with pharyngitis to rapid antigen detection test for Group A streptococcus (GAS) given that there’s a **76.52%** chancethey would get the result of the test right.

**ROLES OF TEAM MEMBERS:**

Malina Meher Satya Swaroop:

1. Handled data preprocessing, exploratory data analysis and hyperparameter tuning.
2. Handled 2 out 5 attribute selection methods.
3. Trained 4 out of 8 algorithms.
4. Visualized 4 out of 8 models based on the evaluation metrics.
5. Handled 70% of the report generation.

JiaTong Lyu:

1. Handled 3 out 5 attribute selection methods.
2. Trained 4 out of 8 algorithms.
3. Visualized 4 out of 8 models based on the evaluation metrics.
4. Handled 30 % of the report generation.

We both collectively decided the best model, discussion and conclusion based on the discussions we had.

**STEPS TO REPRODUCE THE DATASETS AND MODELS:**

1. Use the python script to generate the datasets.
2. Load them in WEKA and use filter NumericToNominal -R [nominal column numbers]
3. Go to classify tab and select the classifier
4. Select percentage split/cross validation. Select 5 in “random seed for XVal/ % Split” in more options.
5. Change the number of iterations to 50 for AdaBoost classifiers while running for all except dataset generated with chi-square feature selection.
6. Change the “random seed for XVal/ % Split” in more options to 3 for **only** AdaBoost with random forest when using the dataset generated with Random Forest feature selection.

**REFERENCES:**

1. <https://www.tutorialspoint.com/weka/what_is_weka.htm>
2. <https://en.wikipedia.org/wiki/Python_(programming_language)>
3. <https://pandas.pydata.org/docs/getting_started/overview.html>
4. <https://numpy.org/doc/stable/user/whatisnumpy.html>
5. <https://docs.python.org/3/library/math.html>
6. <https://pypi.org/project/sklearn/>
7. <https://docs.python.org/3/library/statistics.html>
8. <https://seaborn.pydata.org/>
9. <https://matplotlib.org/>
10. <https://towardsdatascience.com/understanding-random-forest-58381e0602d2>
11. <https://towardsdatascience.com/basic-ensemble-learning-random-forest-adaboost-gradient-boosting-step-by-step-explained-95d49d1e2725>
12. <https://www.ibm.com/topics/knn>
13. <https://en.wikipedia.org/wiki/Logistic_regression>
14. <https://scikit-learn.org/stable/modules/naive_bayes.html>
15. <https://www.techopedia.com/definition/30364/support-vector-machine-svm>