Machine Learning Model Comparison: Evaluating Top Performers with WEKA

Abstract

The project aims to develop machine learning models to discover the best performer among several algorithms. Within the paper, the project would include the pre-processing stage of cleaning the data via WEKA and the step to separate the dataset into training and test sets to evaluate the models. With the experimenter result from WEKA, the project found the top performers were Random Forest, Decision Tree, and IBK. The paper will further discuss the parameters implemented towards the models along with the comparison. The comparison would be divided into two sections. The first section would experiment with the models on their default settings (no parameter change implemented) to explore the best models. The second section will explore the models with parameter changes applied to the model.

1.0 Introduction

With rapid urbanization occurring in Malaysia in the past decade. The country as a whole would be experiencing improvement from the past. Job prospects within the last decade have experienced significant transformation (Firus Khan et al, 2022). Implying the butterfly effect, small changes within certain conditions would lead to significant changes in the outcome. Suggesting that the lifestyle from 20 years ago would be drastically varied. Focusing on the ordinary Malaysian lifestyle, the citizen has adopted a sedentary environment where the rise of white-collar jobs within the market due to technological advancement (Firus Khan et al, 2022). The lifestyle shift correlated with morbidity-inducing behaviors such as alcohol consumption, unhealthy diet, and low physical activity as mentioned by the researcher. These bad behaviors do indicate a strong correlation with cardiovascular diseases (CVD) within the country. Furthermore, factors such as age, smoking, waist circumference, low amount of physical activity, and total fat intake were several variables to analyze. The result reflects a strong association with coronary heart disease (CHD) Wan Musa et al (2022). As per Firus Khan et al (2022), only 422.7 million cases of CVD were identified along with 17.9 million deaths in 2015 with the projection of an increase of 23.6 million death by 2030. For this reason, identifying heart diseases would be crucial to allow citizens of Malaysia to receive treatment earlier. It is within the project scope to build and strengthen a machine learning model and implement the model in the medical industry. In doing so, data are inserted into the model, allowing medical practitioners to view the predicted result along with their own judgment. It alerts medical practitioners to conduct due diligence quickly, creating higher productivity within the workforce.

1.1 Problem Statement

This project has concluded two main problem statements revolving around the heart prediction models.

- The reliability of the machine learning model prediction on a small dataset

The statement refers to the previous paper "*A comprehensive approach to predict heart diseases using data mining"* conducted by Kaur (2017). The research paper conducted focused solely on developing machine learning models for heart disease prediction while utilizing a small dataset. Therefore, the project aims to increase the size of the dataset to allow the machine learning models to develop a more robust model.

- Improve the accuracy of heart disease prediction

The statement refers to the previous studies conducted by Kaur (2017). The highest accuracy recorded within the research was 81.85%. Thus, the aim of the project would be to improve the accuracy of the prediction by selecting the best models for the prediction.

1.2 Aims

- To discover and determine the most reliable machine learning model to conduct heart disease prediction.
- To record the influences of implementing changes of parameters within the machine to achieve better results.

1.3 Objective

- To develop the robustness of the algorithm by implementing 10-fold cross-validation.

Kaur (2017) utilizes the training set within the classifier in WEKA to determine the overall accuracy of the models. This study aims to utilize the 10-fold cross-validation to allow the models to better develop while reducing overfitting the model.

2.0 Overview of Dataset Selected

The dataset utilized in this research is the Heart Disease dataset (https://ieee-dataport.org/open-access/heart-disease-dataset-comprehensive) from the Institute of Electrical and Electronics Engineers (IEEE). It is a combination of 5 popular heart disease datasets that comprises 1190 samples with 11 attributes. The details of the attributes are explained in Table 1:

Feature Name	Description
Age	Age of patients in years
Sex	Gender of patients
Chest Pain Type	Type of chest pain which 1: Typical, 2: Typical Angina, 3: Non-Anginal Pain, 4: Asymptomatic
Resting BP	Blood pressure level at resting mode (mm/HG)
Cholesterol	Serum cholesterol (mg/dl)
Fasting Blood Sugar	Blood sugar levels on fasting > 120 mg/dl which 1: true, 0: false
Max Heart Rate	Maximum heart rate
Resting ECG	Result of electrocardiogram while resting, 0: Normal, 1: Abnormal in ST, 2: Left ventricular hypertrophy
Exercise Angina	Chest pain induced by exercise which 0: NO, 1: Yes
Oldpeak	Comparison of ST-depression during exercise and resting
ST Slope	ST segment measured in terms of slope during peak exercise, 0: Normal 1: Upsloping 2: Flat 3: Downsloping
Heart Attack (target)	0: Normal, 1:Diagnose of Heart Disease

Table 1: Description of attributes and the target class

3.0 Exploratory Data Analysis (EDA)

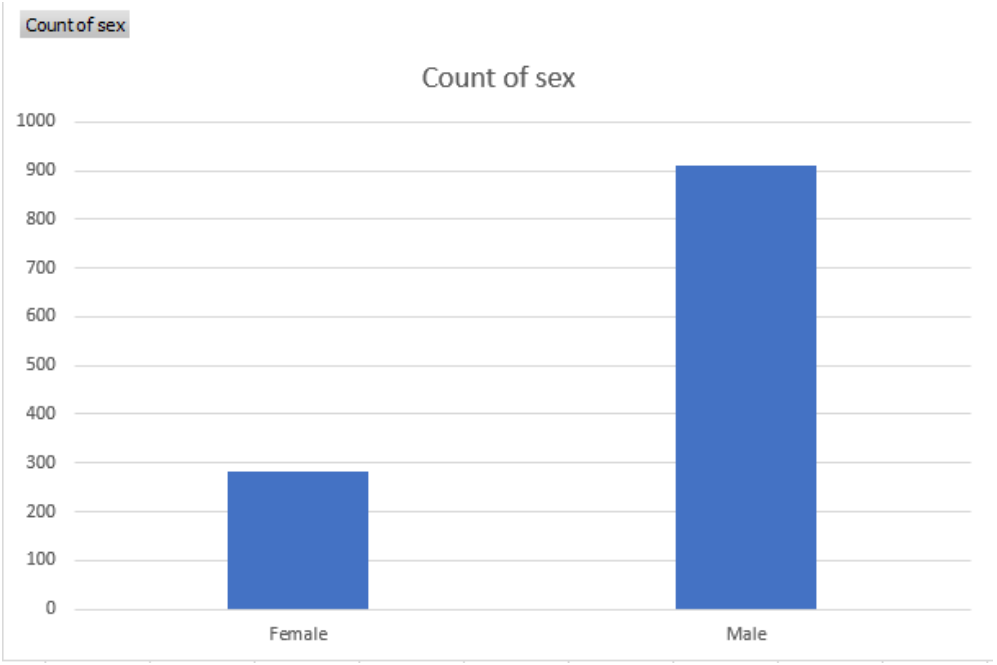


Figure 1 (Distribution of male & female subjects)

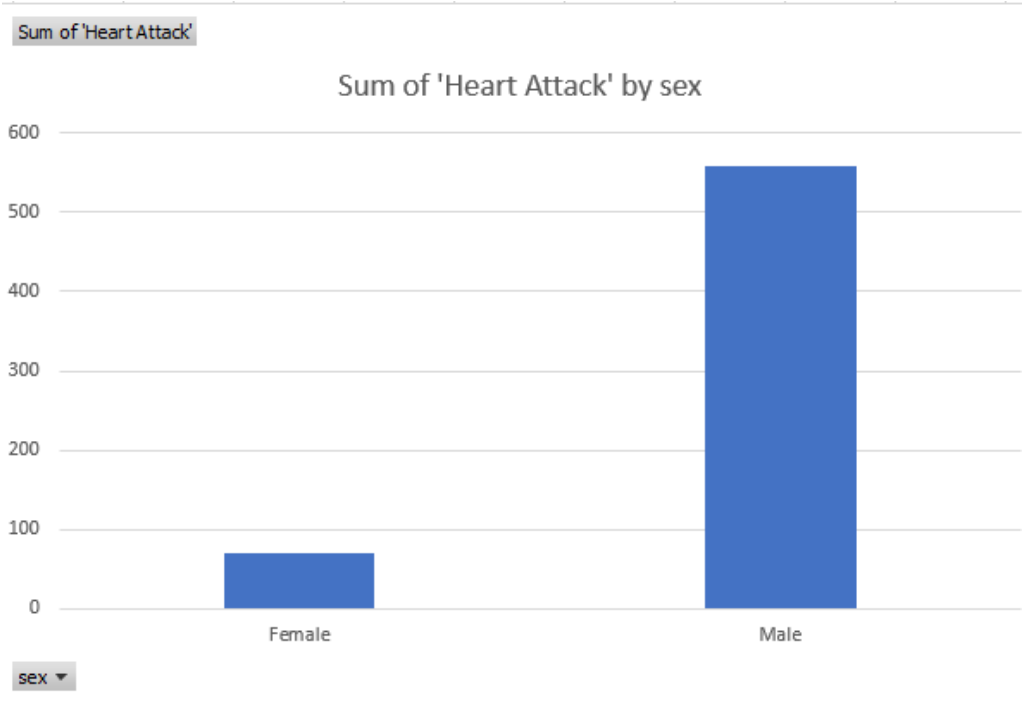


Figure 2 (Distribution of Female and Male by heart attack)

As seen in Figures 1 and 2, the dataset indicates that the majority (76.39%) were male while 23.61% were female. Followed by males had a higher probability of being diagnosed with heart attack by a significant amount when in comparison to females.

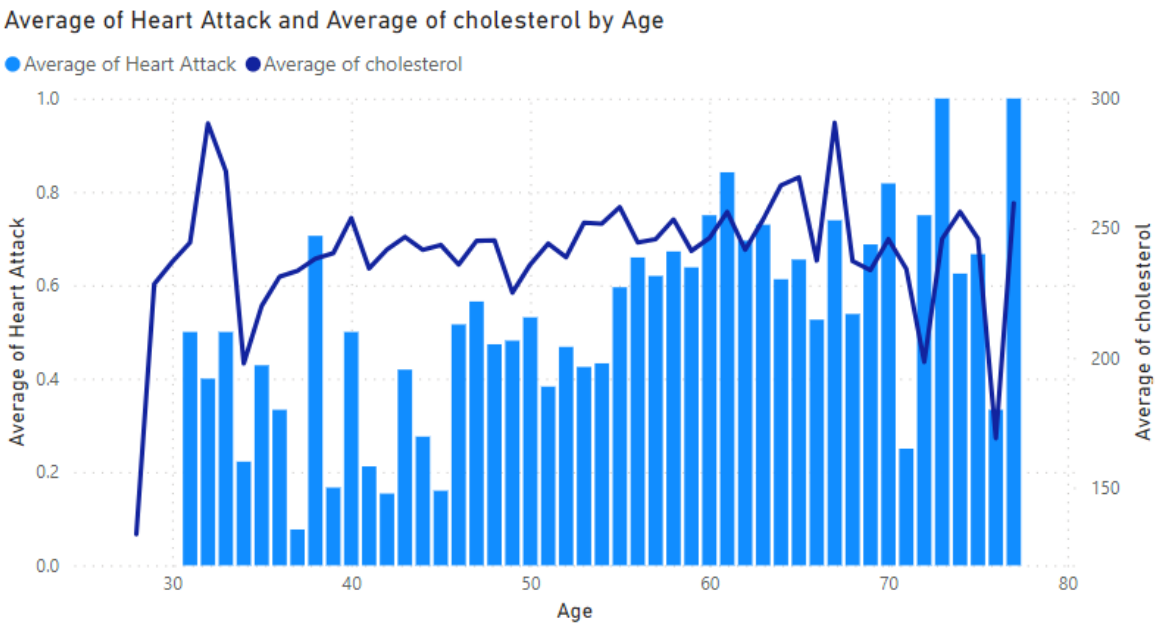


Figure 3 (Distribution of heart attack based on age & cholesterol)

Upon closer inspection of Figure 3, the chart clearly displays a strong correlation between age and heart attack. Despite the average cholesterol in the early thirties being substantially high, the average heart attack increases accordingly. Reaching the age of 40 and 50, the number of heart attack cases ascend gradually despite the level of cholesterol remaining relatively consistent. The older generation displayed the highest cases of heart attack during the sixties. Surprisingly, heart attack cases remain relatively high despite the line downtrend in the average cholesterol.

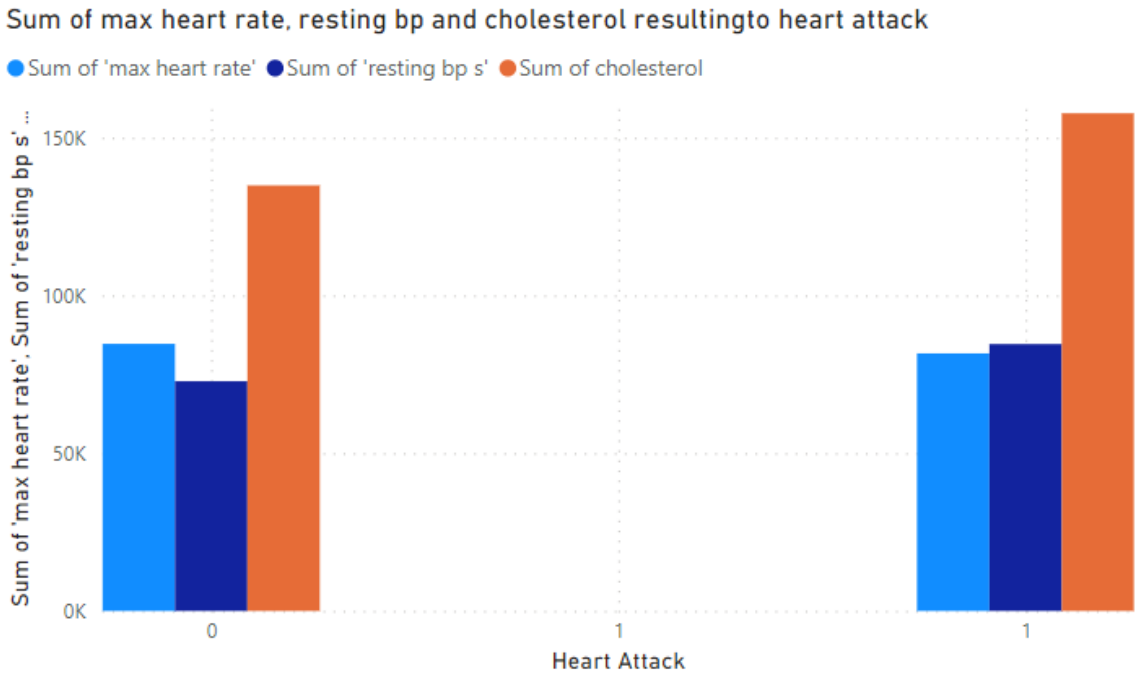


Figure 4 (Correlation between max heart rate, resting bp & cholesterol with heart attack).

Furthermore, Figure 4 shows the sum of the maximum heart rate with the average resting bp and max heart rate along with the heart attack class. The max heart rate & resting bp do prove to have a significant relation towards heart attack. Patients without heart attack display a higher max heart rate and lower resting bp, suggesting a healthy heart whereas the opposite applies to a weaker heart.

3.1 Model Selection Based on EDA

- Random Forest

Random forest would be an excellent model for non-linear relationships. As seen in Figure 3, the cholesterol relationship with the heart attack cases rise has a strong correlation. However, upon reaching the older demographic. The relationship was negatively correlated. Therefore, the ensemble learning method allows machine learning to combine multiple decision trees with a voting system to predict the result. As a result, it allows novice users to reconfigure the parameter due to its robust system (Goldstein, Polley & Polley, 2011)

- K-Nearest Neighbors (KNN)

K-Nearest Neighbours (KNN) would prove to be an effective tool for prediction for clustering like behavior. Within Figure 3, the age group could be divided into six clusters. In this context, having an unknown data point allows the algorithm to compare similar data points and implement the Euclidean formula to measure the closeness to create a prediction (Rahim and Ahmar, 2022).

- Decision Tree

Due to the non-linear relationship, the decision tree would be suitable for classification prediction. The core would be random forest is an ensemble learning model where the decision tree is an individual in comparison. Machine learning would utilize the Gini impurity and information gain to reorganize the feature by importance to make predictions.

4.0 Methodology

4.1 Data Preprocessing: Data Problem.

- Data Skewness

The imbalance of class presents a problem, it influences the accuracy matrix of the result. Wu & Fang (2020) dataset was unbalanced, and it was tested for prediction. The predictions were measured through accuracy, sensitivity, specificity & AUC. The prediction displayed high accuracy, but the other measurements were low or close to zero. Thus, resulting in an overall weak model. The dataset does display an imbalance issue.

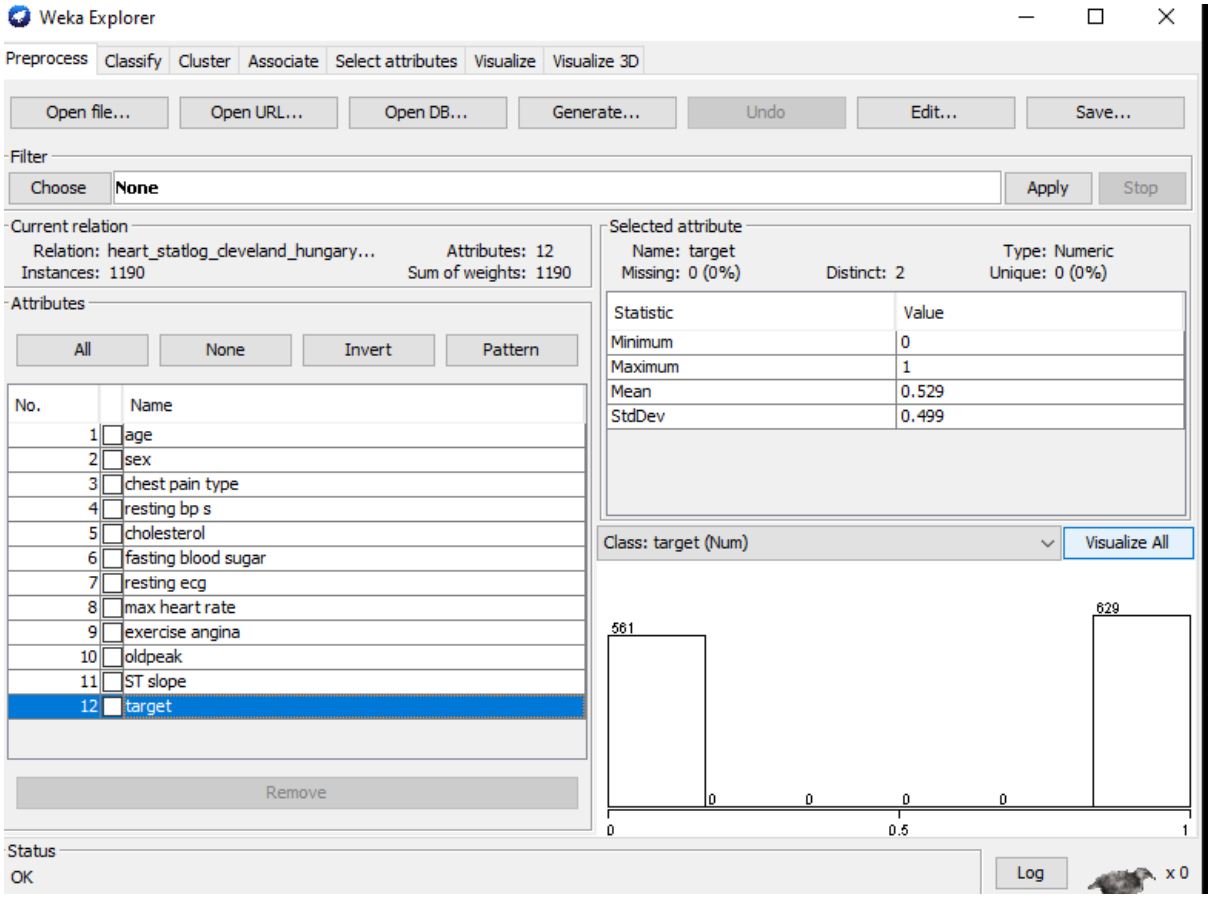
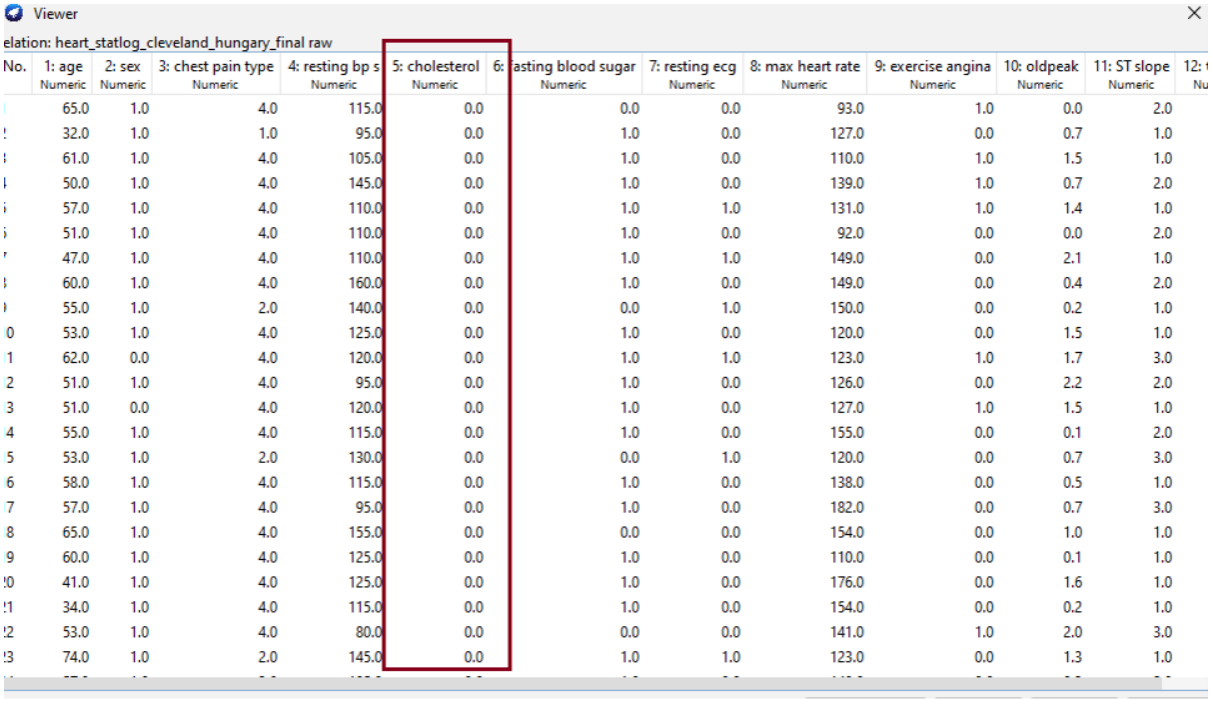


Figure 5 (Imbalance of class)

- Missing Value

Within the dataset, there were missing values. Although WEKA displayed zero, the missing value was disguised as 0. Therefore, WEKA reads the data as numeric, allowing it to bypass the WEKA missing value detection.



No.	1: age Numeric	2: sex Numeric	3: chest pain type Numeric	4: resting bp s Numeric	5: cholesterol Numeric	6: asting blood sugar Numeric	7: resting ecg Numeric	8: max heart rate Numeric	9: exercise angina Numeric	10: oldpeak Numeric	11: ST slope Numeric	12: Nu
	65.0	1.0	4.0	115.0	0.0	0.0	0.0	93.0	1.0	0.0	2.0	
	32.0	1.0	1.0	95.0	0.0	1.0	0.0	127.0	0.0	0.7	1.0	
	61.0	1.0	4.0	105.0	0.0	1.0	0.0	110.0	1.0	1.5	1.0	
	50.0	1.0	4.0	145.0	0.0	1.0	0.0	139.0	1.0	0.7	2.0	
	57.0	1.0	4.0	110.0	0.0	1.0	1.0	131.0	1.0	1.4	1.0	
	51.0	1.0	4.0	110.0	0.0	1.0	0.0	92.0	0.0	0.0	2.0	
	47.0	1.0	4.0	110.0	0.0	1.0	1.0	149.0	0.0	2.1	1.0	
	60.0	1.0	4.0	160.0	0.0	1.0	0.0	149.0	0.0	0.4	2.0	
	55.0	1.0	2.0	140.0	0.0	0.0	1.0	150.0	0.0	0.2	1.0	
0	53.0	1.0	4.0	125.0	0.0	1.0	0.0	120.0	0.0	1.5	1.0	
1	62.0	0.0	4.0	120.0	0.0	1.0	1.0	123.0	1.0	1.7	3.0	
2	51.0	1.0	4.0	95.0	0.0	1.0	0.0	126.0	0.0	2.2	2.0	
3	51.0	0.0	4.0	120.0	0.0	1.0	0.0	127.0	1.0	1.5	1.0	
4	55.0	1.0	4.0	115.0	0.0	1.0	0.0	155.0	0.0	0.1	2.0	
5	53.0	1.0	2.0	130.0	0.0	0.0	1.0	120.0	0.0	0.7	3.0	
6	58.0	1.0	4.0	115.0	0.0	1.0	0.0	138.0	0.0	0.5	1.0	
7	57.0	1.0	4.0	95.0	0.0	1.0	0.0	182.0	0.0	0.7	3.0	
8	65.0	1.0	4.0	155.0	0.0	0.0	0.0	154.0	0.0	1.0	1.0	
9	60.0	1.0	4.0	125.0	0.0	1.0	0.0	110.0	0.0	0.1	1.0	
0	41.0	1.0	4.0	125.0	0.0	1.0	0.0	176.0	0.0	1.6	1.0	
1	34.0	1.0	4.0	115.0	0.0	1.0	0.0	154.0	0.0	0.2	1.0	
2	53.0	1.0	4.0	80.0	0.0	0.0	0.0	141.0	1.0	2.0	3.0	
3	74.0	1.0	2.0	145.0	0.0	1.0	1.0	123.0	0.0	1.3	1.0	

Figure 6 (missing value disguised as 0)

4.2 Data Solutions

- Data Skewness

With the emphasis on improving accuracy, the SMOTE technique was implemented to balance the data.

- Missing Value

The projects would be deleting the zero within the column and leaving it blank. Thus, allowing WEKA to read the missing value. As it identifies the missing value, the filter 'ReplaceMissingValue' filter would be applied to fill in the missing value with the mode and mean from the training data.

4.3 Filtering the training set

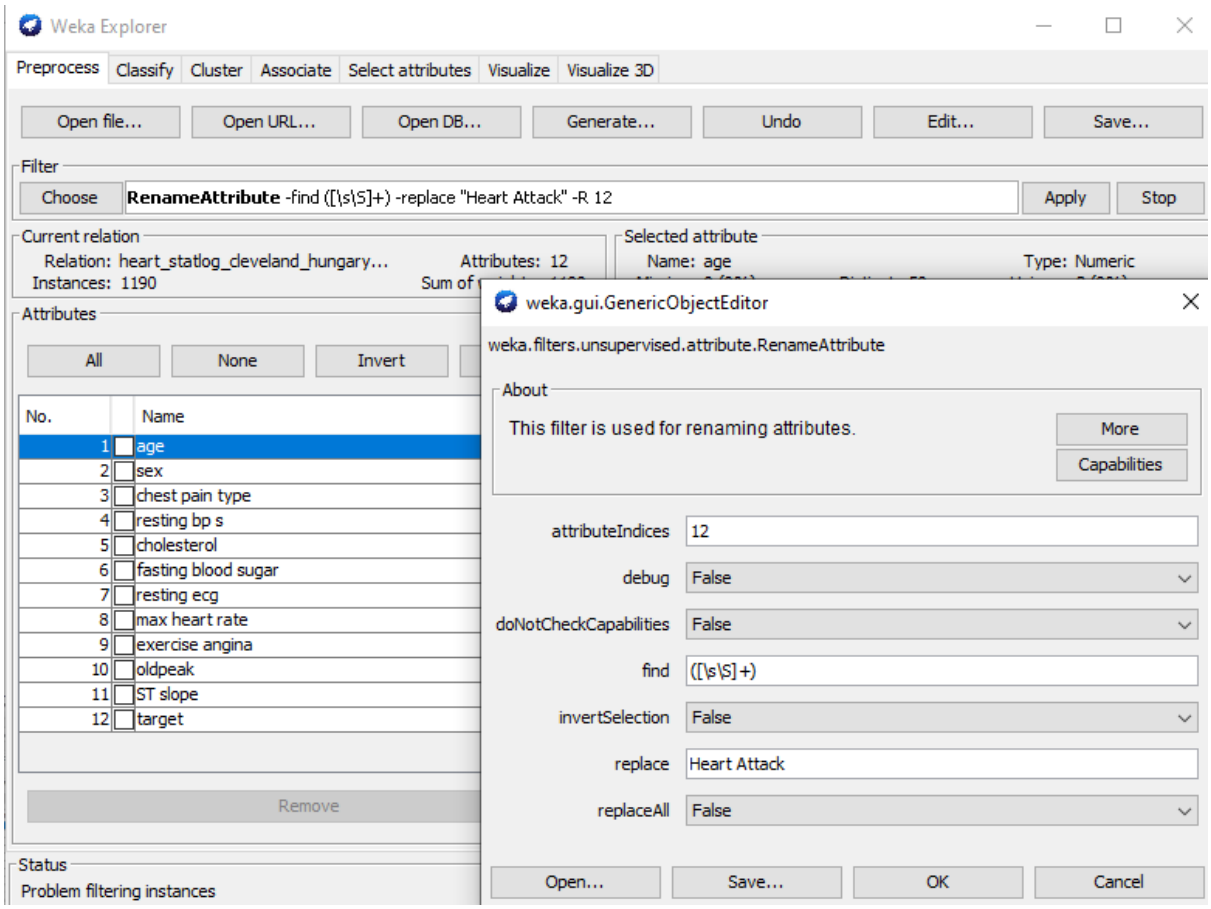


Figure 7 (Step 1: Renaming the target attribute to Heart Attack)

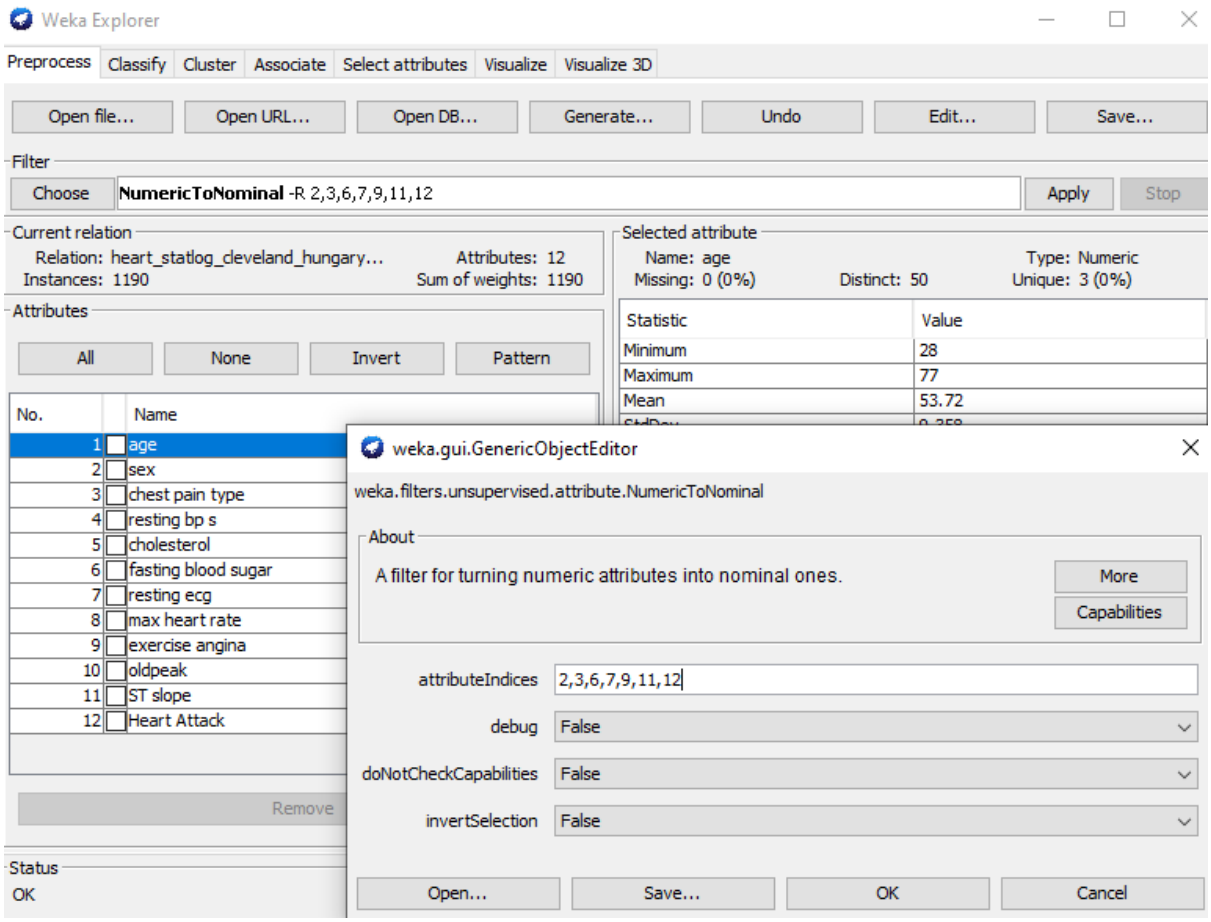


Figure 8 (Step 2: Converting numeric to nominal type due to the data selected are categorical types)

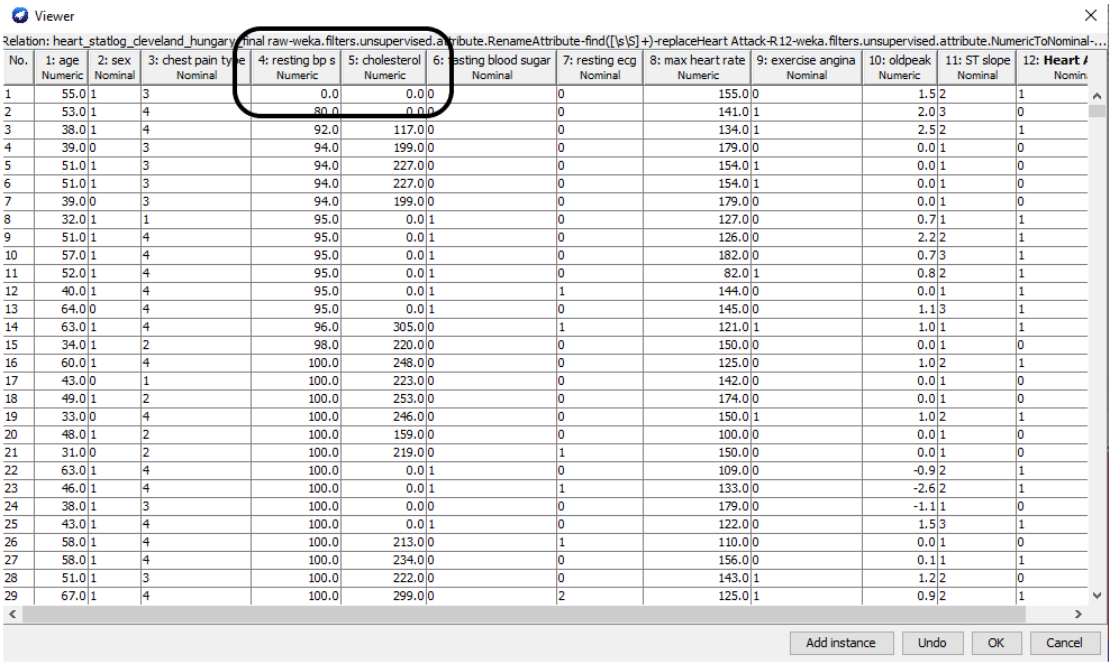


Figure 9 (Step 3.1: Replacing all 0 values with missing values as the system is unable to detect 0 as a missing value)

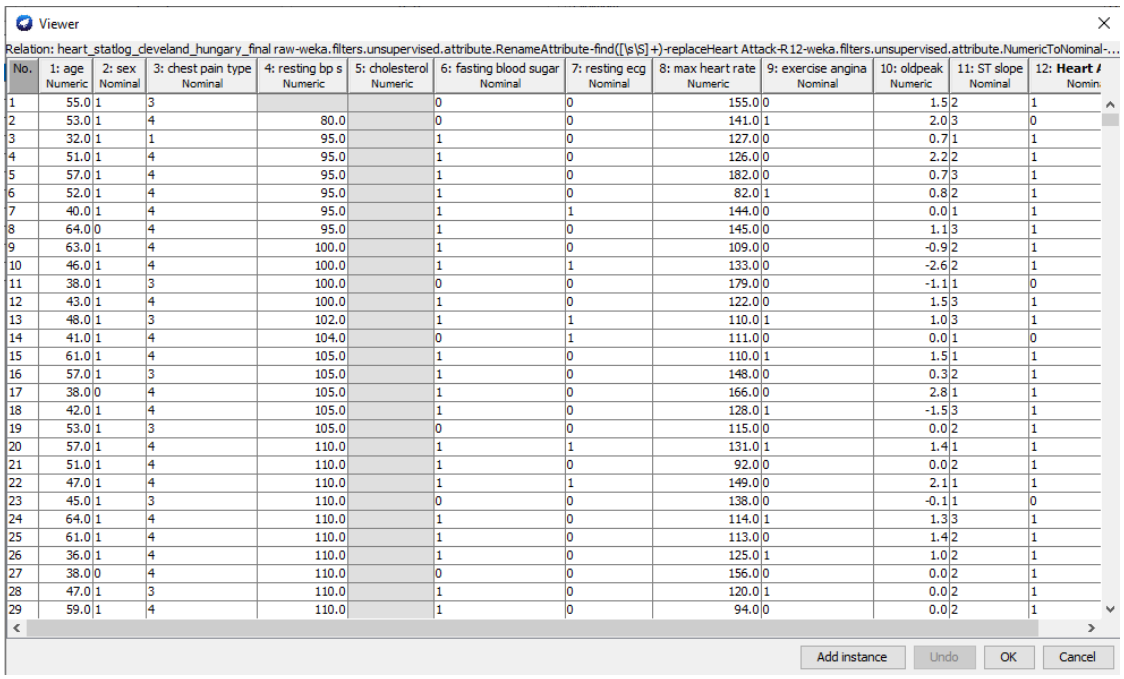


Figure 10 (Step 3.2 Result of removing all 0 values)

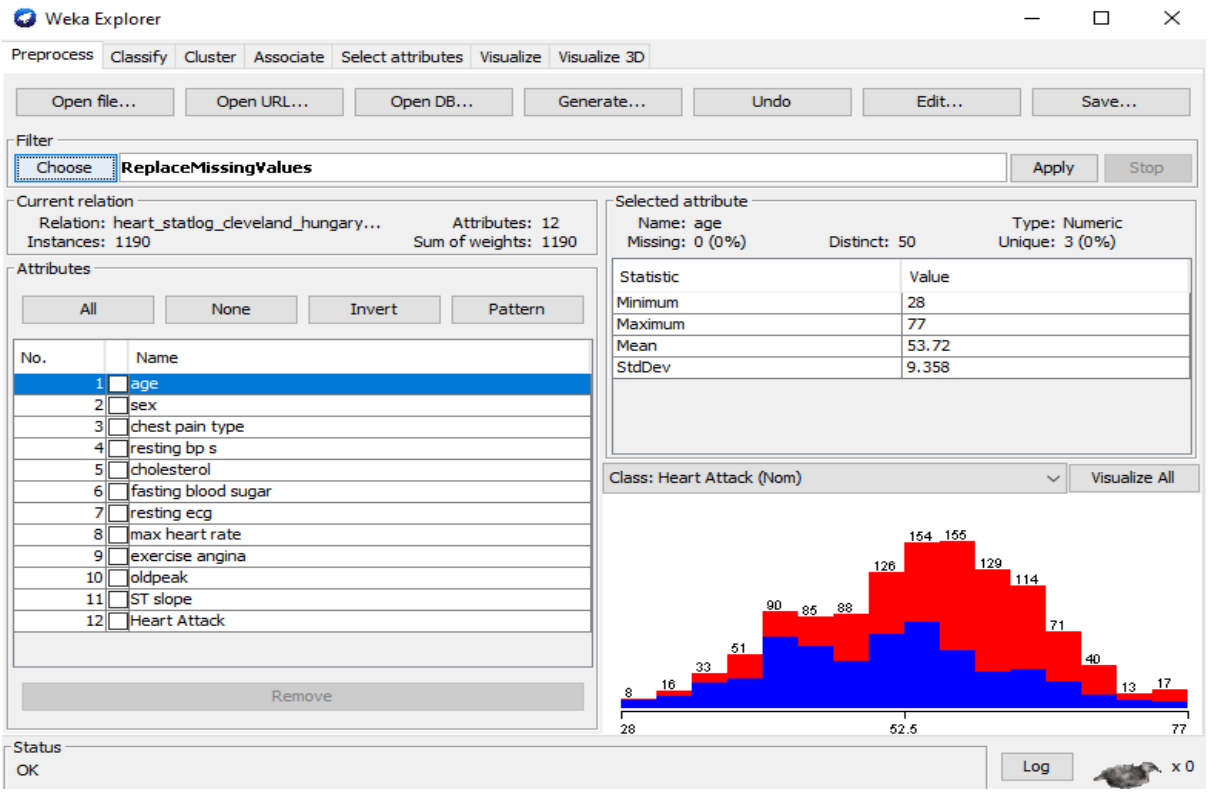


Figure 11 (Step 4: Applying 'ReplaceMissingValue' filter with mean and mode.)

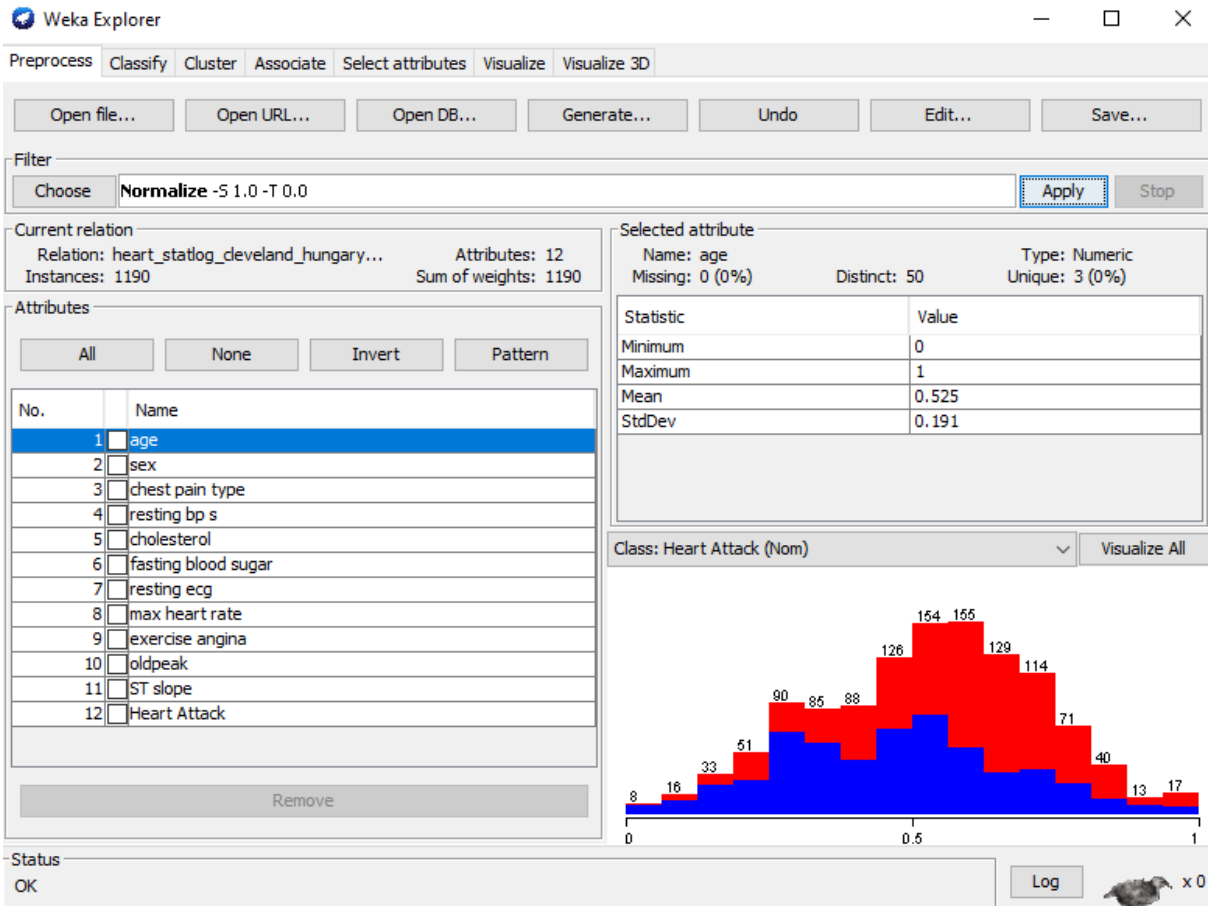


Figure 12 (Step 5: Apply the 'Normalize' filter)

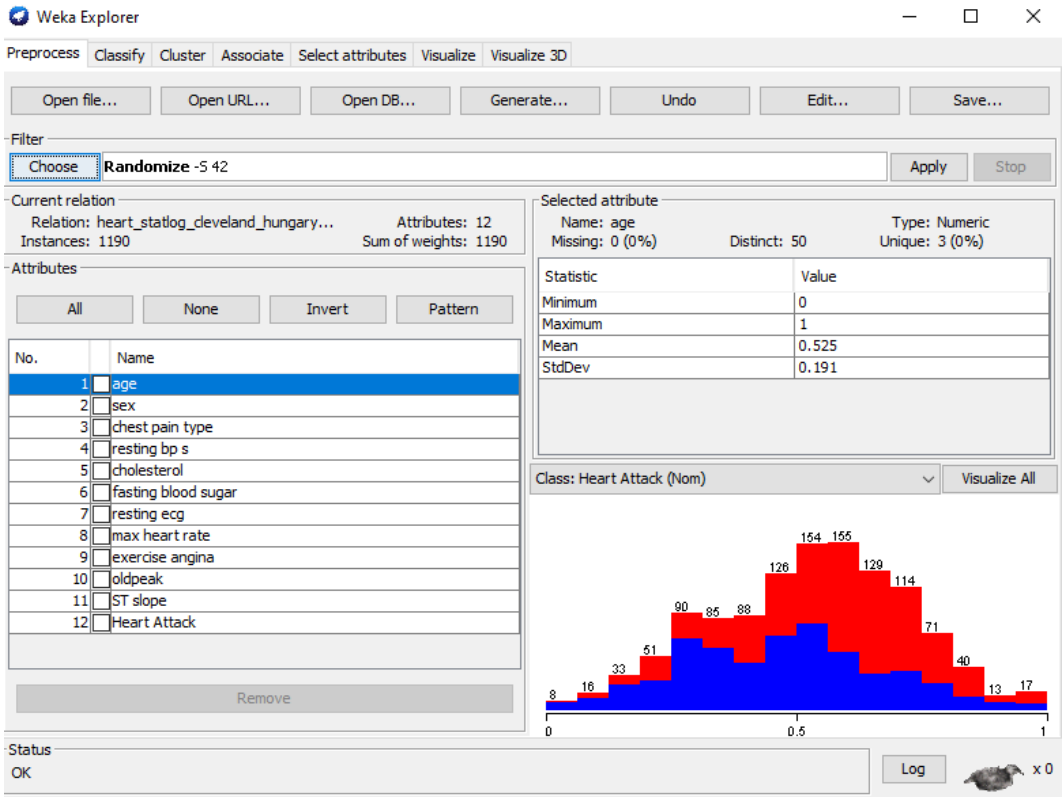


Figure 13 (Step 6: Apply the 'Randomize' filter)

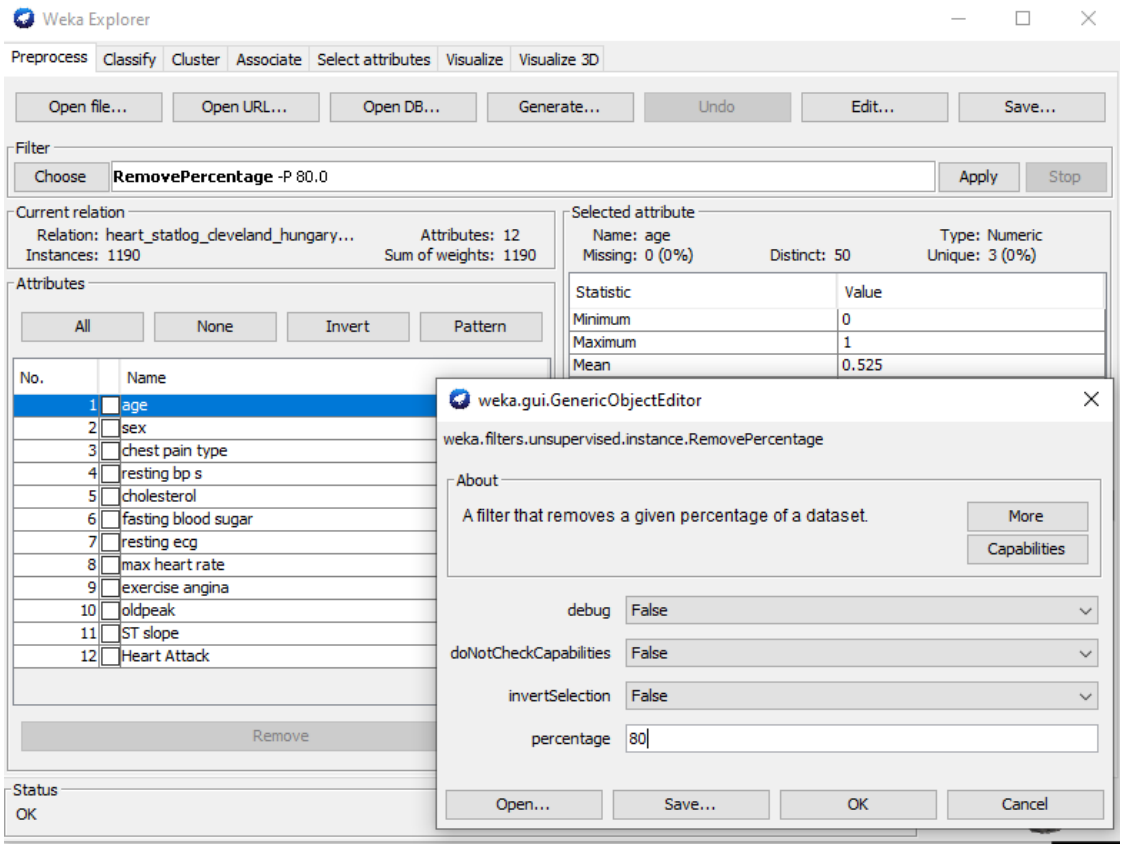


Figure 14 (Step 7: Remove 80% of the dataset to separate the test set)

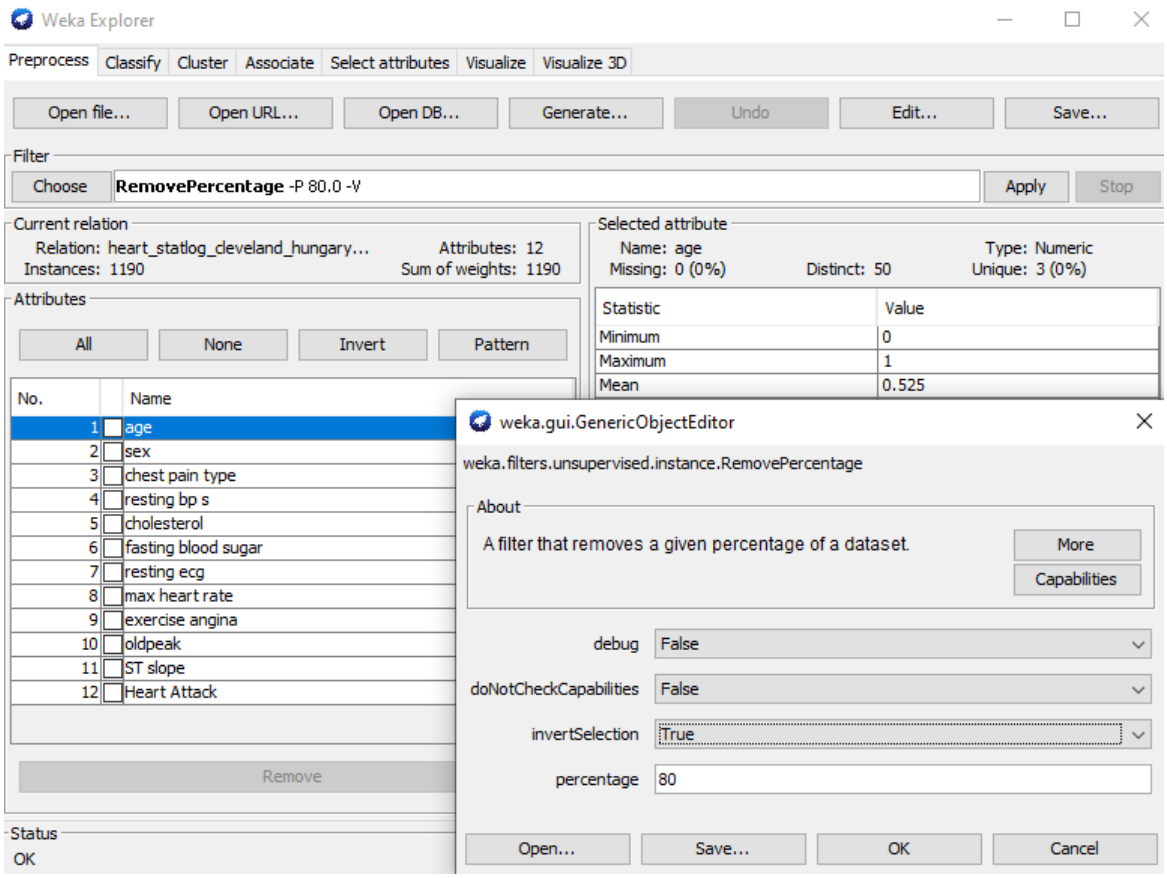
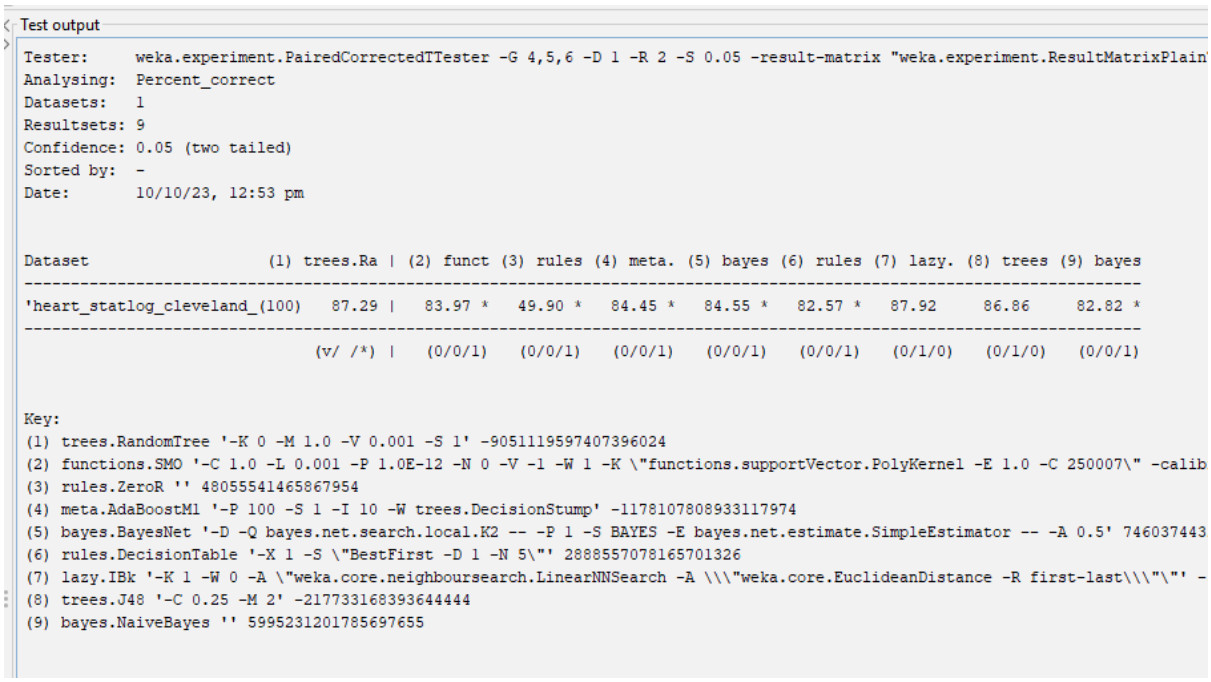


Figure 15 (Step 8: Undo the action and change the invert the 'True' within the remove percentage filter to save the training set.)

4.4 Experimenter



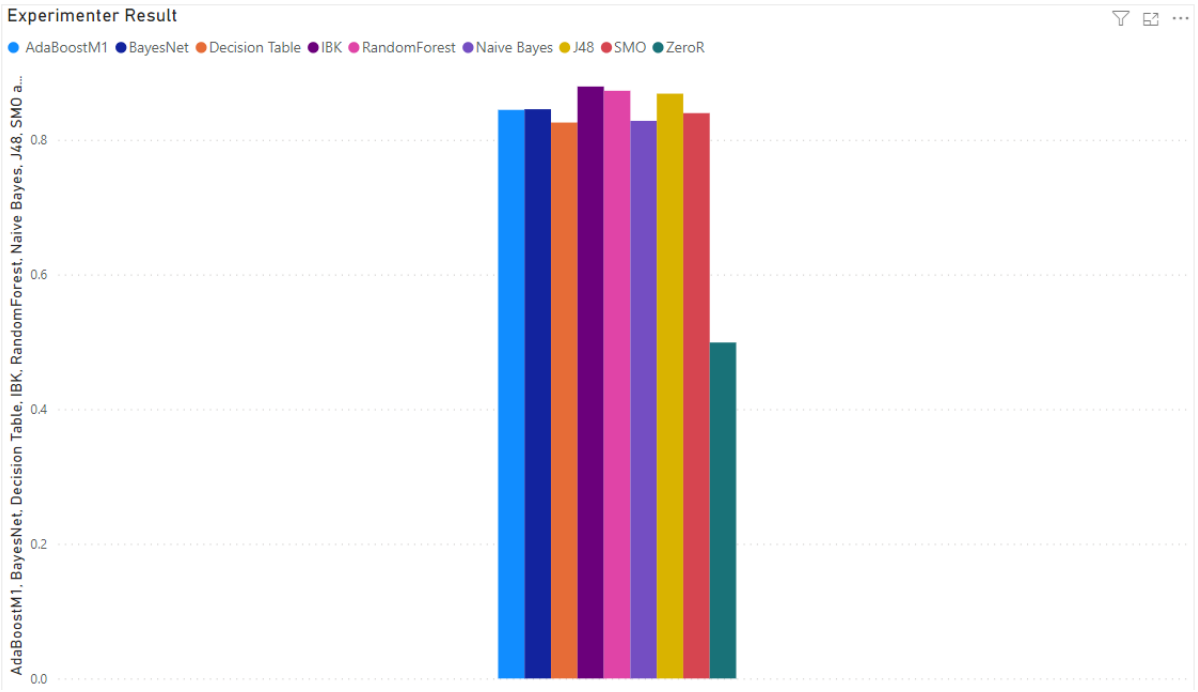
Test output

```
Tester:     weka.experiment.PairedCorrectedTTester -G 4,5,6 -D 1 -R 2 -S 0.05 -result-matrix "weka.experiment.ResultMatrixPlain
Analysing:  Percent_correct
Datasets:   1
Resultsets: 9
Confidence: 0.05 (two tailed)
Sorted by:  -
Date:       10/10/23, 12:53 pm


Dataset                   (1) trees.Ra | (2) funct (3) rules (4) meta. (5) bayes (6) rules (7) lazy. (8) trees (9) bayes
-----------------------------------------------------------------------------------------------------------------------
'heart_statlog_cleveland_(100)   87.29 |   83.97 *   49.90 *   84.45 *   84.55 *   82.57 *   87.92     86.86     82.82 *
-----------------------------------------------------------------------------------------------------------------------
                              (v/ /*) |   (0/0/1)   (0/0/1)   (0/0/1)   (0/0/1)   (0/0/1)   (0/1/0)   (0/1/0)   (0/0/1)


Key:
(1) trees.RandomTree '-K 0 -M 1.0 -V 0.001 -S 1' -9051119597407396024
(2) functions.SMO '-C 1.0 -L 0.001 -P 1.0E-12 -N 0 -V -1 -W 1 -K \"functions.supportVector.PolyKernel -E 1.0 -C 250007\" -calib
(3) rules.ZeroR '' 48055541465867954
(4) meta.AdaBoostM1 '-P 100 -S 1 -I 10 -W trees.DecisionStump' -1178107808933117974
(5) bayes.BayesNet '-D -Q bayes.net.search.local.K2 -- -P 1 -S BAYES -E bayes.net.estimate.SimpleEstimator -- -A 0.5' 746037443
(6) rules.DecisionTable '-X 1 -S \"BestFirst -D 1 -N 5\"' 2888557078165701326
(7) lazy.IBk '-K 1 -W 0 -A \"weka.core.neighboursearch.LinearNNSearch -A \\\"weka.core.EuclideanDistance -R first-last\\\"\"' -
(8) trees.J48 '-C 0.25 -M 2' -217733168393644444
(9) bayes.NaiveBayes '' 5995231201785697655
```

Figure 16 (Experimenter on training set to determine the top algorithms)



Figureure 17 (Clustered column chart for the algorithms)

IBK	Random Forest	J48	BayesNet	AdaBoost M1	SMO	Naive Bayes	Decision Table	ZeroR
87.92%	87.29%	86.86%	84.55%	84.45%	83.97%	82.82%	82.57%	49.90%

Table 2 (The result of the experimenter on the training set)

Based on the experimenter in WEKA, it has shown that the top three performing models were IBK, Random Forest and J48. Based on these observations, IBK outperformed Random Forest and J48 by 0.63% and 1.06% respectively. The result section will further elaborate on the three model performances in detail.

4.5 Top Algorithms

1. Decision Tree

The Decision Tree allows easy interpretability for researchers to comprehend the justification of the classification by implementing a basic structure (root node, internal node & transient terminal node). Thus, it creates a flowchart in a hierarchical structure allowing the researchers to analyze the correlation between attributes (Mabu, Obayashi & Kuremoto, 2016).

With robust ML models available to the market, it requires higher computational power and longer duration to complete the task. (Rodríguez, Alesanco, Mehavilla & García, 2022). Within the business context, higher quantities of commodities such as time and capital would be required. Thus, the Decision Tree leans towards a simple ML model thus allowing lower computation power. Leveraging large organizations to opt for a simple ML model with the mentality of allocating resources in an effective & efficient manner.

2. Random Forest

It implements an ensemble learning algorithm that is derived from a decision tree. It allows each node to randomly select a subset of attributes to determine the best split between parent & child nodes (Gini impurity). Thus, it creates a numeric weightage on certain attributes during the prediction of heart attack (Cutler et al, 2007; Ansarullah, Saif, Kumar & Kirmani., 2022). On top of that, random forest machine learning involves proximities (information gain) to group information (Cutler et al, 2007). A few applications would be:

- Impute missing data
- Facilitate graphical representations
- Traditional Multivariate

Random forest allows the flexibility of multi-level classification. Hence, suggesting that random forest would. Although, the project does not contain three or more classifications, it is a beneficial feature to be aware of for future projects (Kerr et al., 2016).

In comparison to the Decision Tree (J48), as the ML executes an ensemble learning technique, it minimizes the likelihood of overfitting. With cross-validation, it builds the random forest model by weighing the weightage of the attribute therefore minimizing selection bias (Tong et al, 2004)

3. K-nearest neighbors

By utilizing similarity-based grouping, it measured the similarity by using the Euclidean distance between data points. The ML model allows the user/researcher to adjust the neighbor density by fine-tuning the K value parameter. Bhattacharya, Ghosh & Chowdhury (2015) researcher further mentioned that KNN would be less sensitive towards outliers by increasing the K value. Since the ML calculates the similarity distance to group data points, outliers have less influence on the classification. The size of the dataset influences the ML model. KNN strives within a smaller dataset (Nababan, Sitompul & Tulus, 2018). KNN is a nonparametric model thus no assumption is made about the underlying data distribution.

4. BayesNet

Also known as Bayesian Networks, uses graphics to represent the probability relationship between a set of variables. These variables are in the form of nodes while the relationship between them is known as edge (Spencer, Thabtah, Abdelhamid & Thompson, 2020). Knowing that the edges indicate a probabilistic dependency between nodes, the probability relationships can be understood based on the conditional probability given its parent node (Korb & Nicholson, 2003). The ML model capable of performing advanced probabilistic reasoning is particularly useful in answering “what-if” questions, allowing making decisions in uncertain situations (Qian, 2014).

5. AdaBoostM1

Utilizing the boosting techniques, AdaBoostM1 is an ensemble learning algorithm. This ML starts by distributing equal weights to all data points. In each iteration, this ML continues to assign higher weights to those that are wrongly classified, allowing them to be prioritized. The process will end when the desired accuracy is achieved (Shahri, Lai, Mohamad, Rahman, & Rambli, 2021).

As an ensemble learning model, AdaBoost is less sensitive to the overfitting problem, but it is likely to be influenced by outliers and noisy data which is a common problem faced by boosting techniques (Cao, Kwong, & Wang, 2012).

4.6 Cross-validation Method

With the machine learning model, 10-fold cross-validation was selected. It allows the training set to conduct multiple iterations on a subset of data based on the parameter set and provides a robust estimation for the model outcome, allowing the available data to be utilized to the maximum capability. (Simon, Subramanian., Li, & Menezes, 2011). It also allows multiple subsets of data for the ML model to train effectively, reducing bias selection and overfitting (Lemm., Blankertz., Dickhaus & Müller, 2011). In comparison, holdout methods excel in requiring less computation power. However, within the project dataset and objective, the size would be considered relatively small compared to a larger dataset, and achieving higher accuracy would be the project's main priority. Thus, 10-fold cross-validation was selected and applied to all ML models. Within the

4.7 Selection of Top 3 Models for Evaluation with cross validation

As seen within Table 2, the top algorithms performing were Decision Tree, Random Forest, and IBK. As mentioned in the selected method of method to evaluate the models. All three models were tested via a 10-fold cross-validation method to ensure robust evaluation before hyperparameter tuning would be applied. The three algorithms will include the training and test set table to indicate the variance between both sets. Variance less than 5% between the training and test sets would be accepted and will be considered an adequate level of robustness on the model.

1. Decision tree

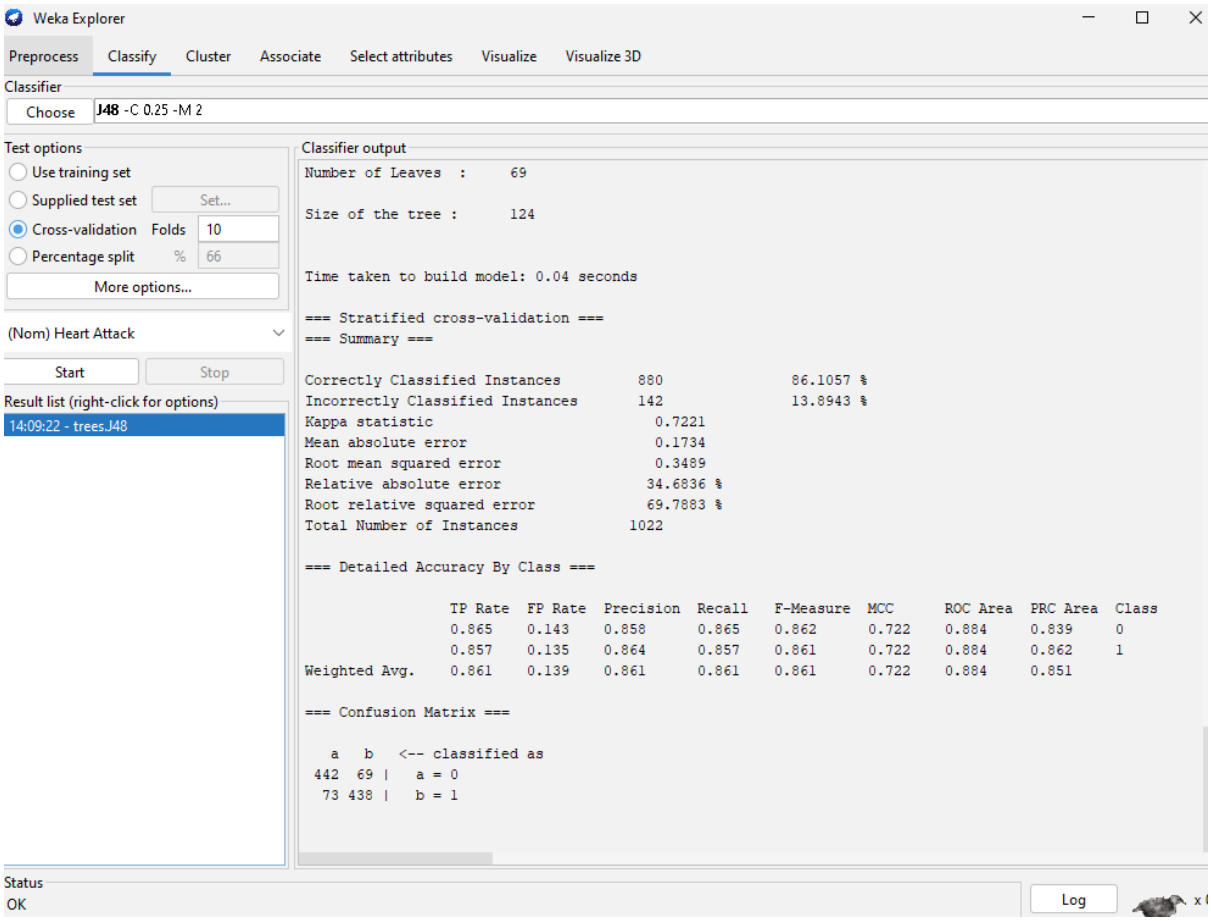


Figure 18

It displays the decision tree algorithm on the training set without any adjustment to the parameter. The purpose would be to display the authentic result of comparing the test and training set.

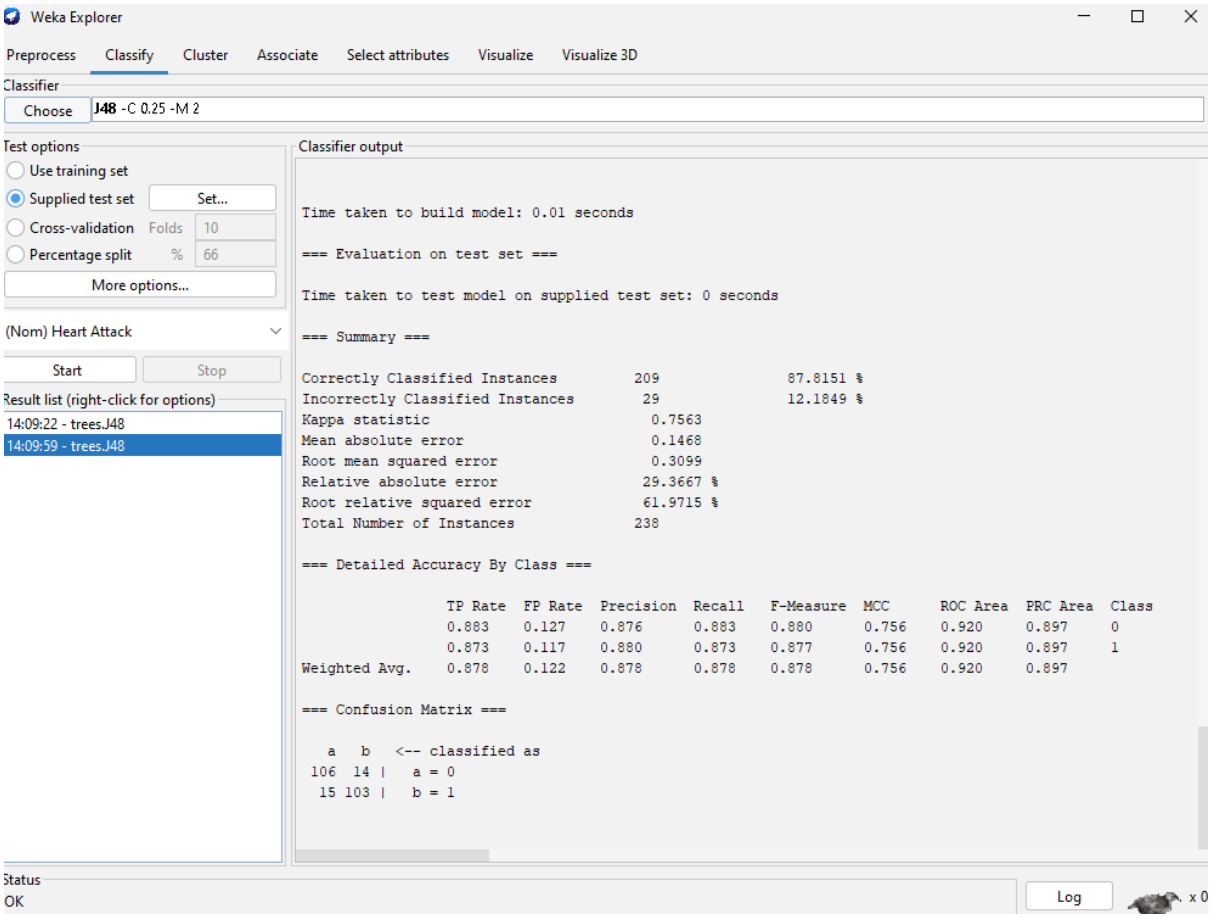


Figure 19

It displays the decision tree algorithm on the test set without any adjustment to the parameter. The purpose would be to display the authentic result of comparing the test and training set.

Metrics	**Training Set**	**Test Set**
Accuracy	86.10%	87.82%
Mean Absolute Error (MAE)	0.1734	0.1468
Root Mean Square Error (RMSE)	0.3489	0.3099

Table 3 (Summary for Decision Tree in training set in comparison to test set)

Within Table 3, the decision tree algorithm illustrates that the test set had a higher result (87.82%) compared to the training set along with lower MAE and RMSE.

2. Random Forest

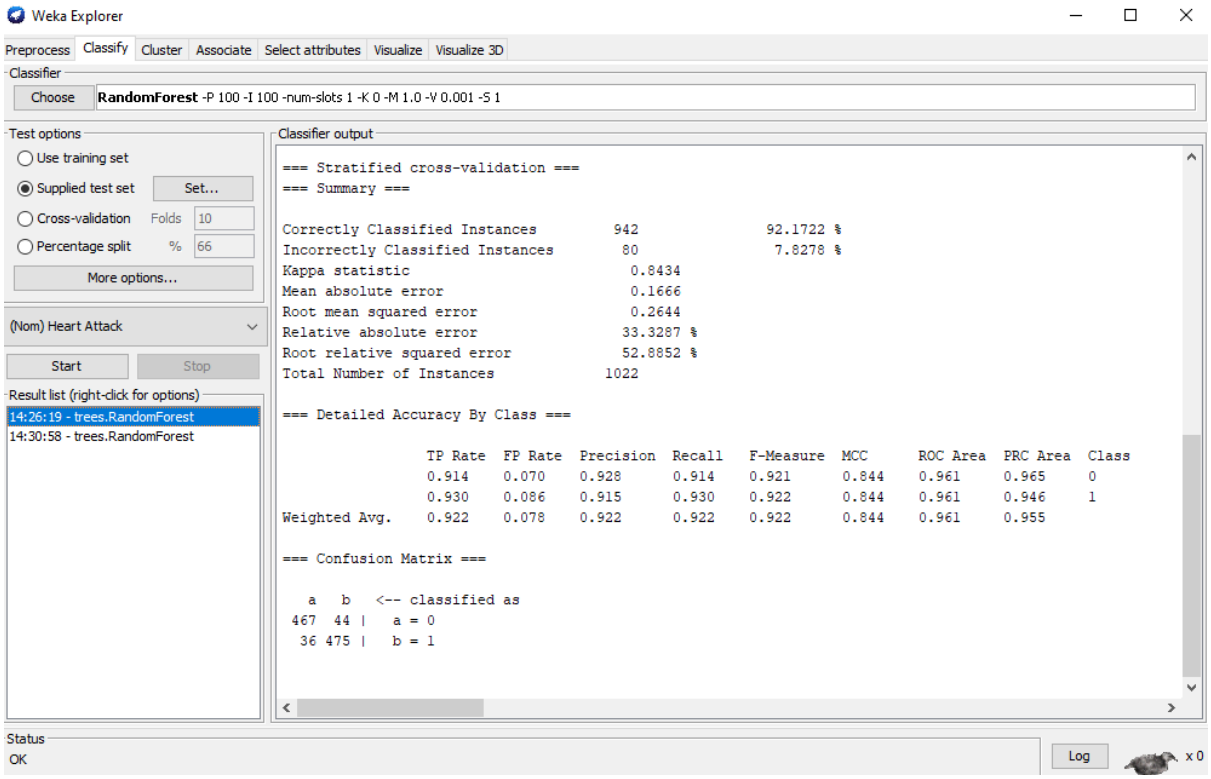


Figure 20

It displays the Random Forest algorithm on the training set without any adjustment to the parameter. The purpose would be to display the authentic result of comparing the test and training set.

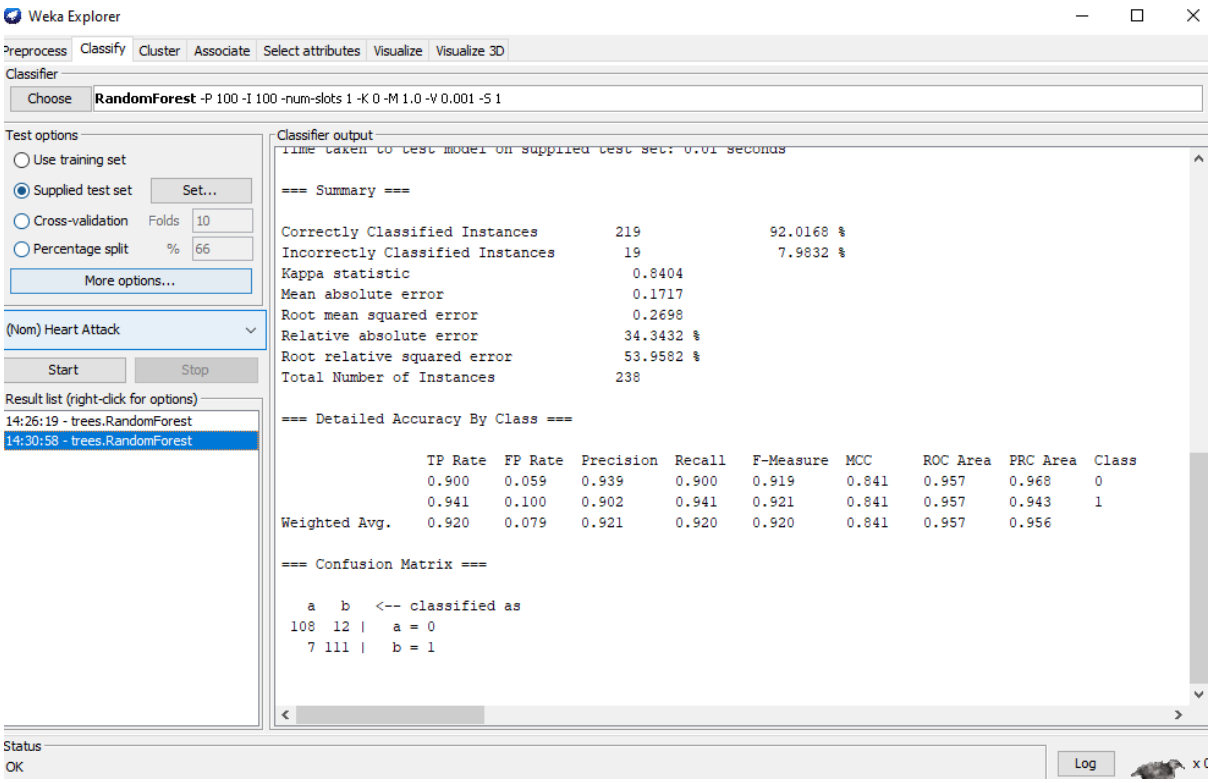


Figure 21

It displays the Random Forest algorithm on the test set without any adjustment to the parameter. The purpose would be to display the authentic result of comparing the test and training set.

Metrics	**Training Set**	**Test Set**
Accuracy	92.17%	92.02%
Mean Absolute Error (MAE)	0.1666	0.1717
Root Mean Square Error (RMSE)	0.2644	0.2698

Table 4 (Summary for Random Forest in training set in comparison to test set)

Random Forest displays minimal variance between the training and test set. The test set performed lower than the group expected as the accuracy dropped by 0.15% shown in Table 4. The factors could be reflected in an increase of MAE and RMSE by 0.0051 and 0.0054 respectively. Suggesting the increase of residue decreases the model accuracy on the performance.

3. K-Nearest Neighbors (KNN)

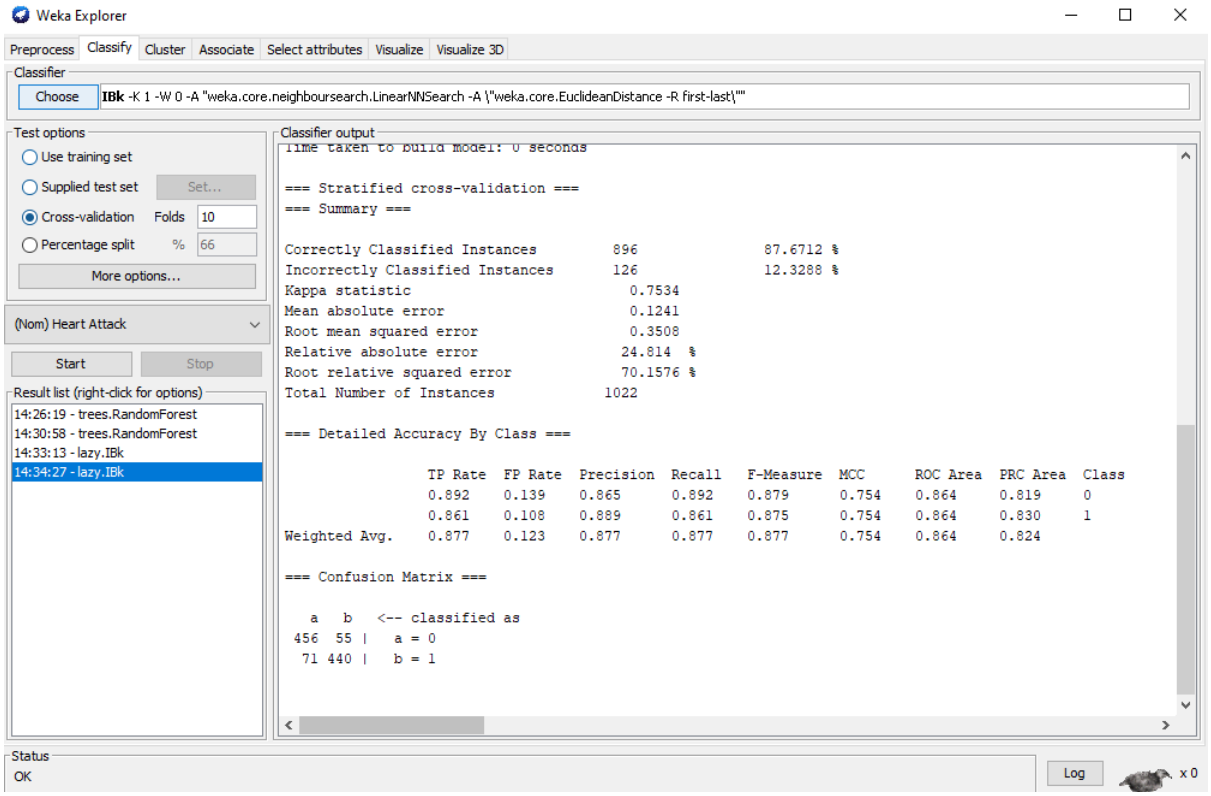


Figure 22

It displays the KNN algorithm on the training set without any adjustment to the parameter. The purpose would be to display the authentic result of comparing the test and training set.

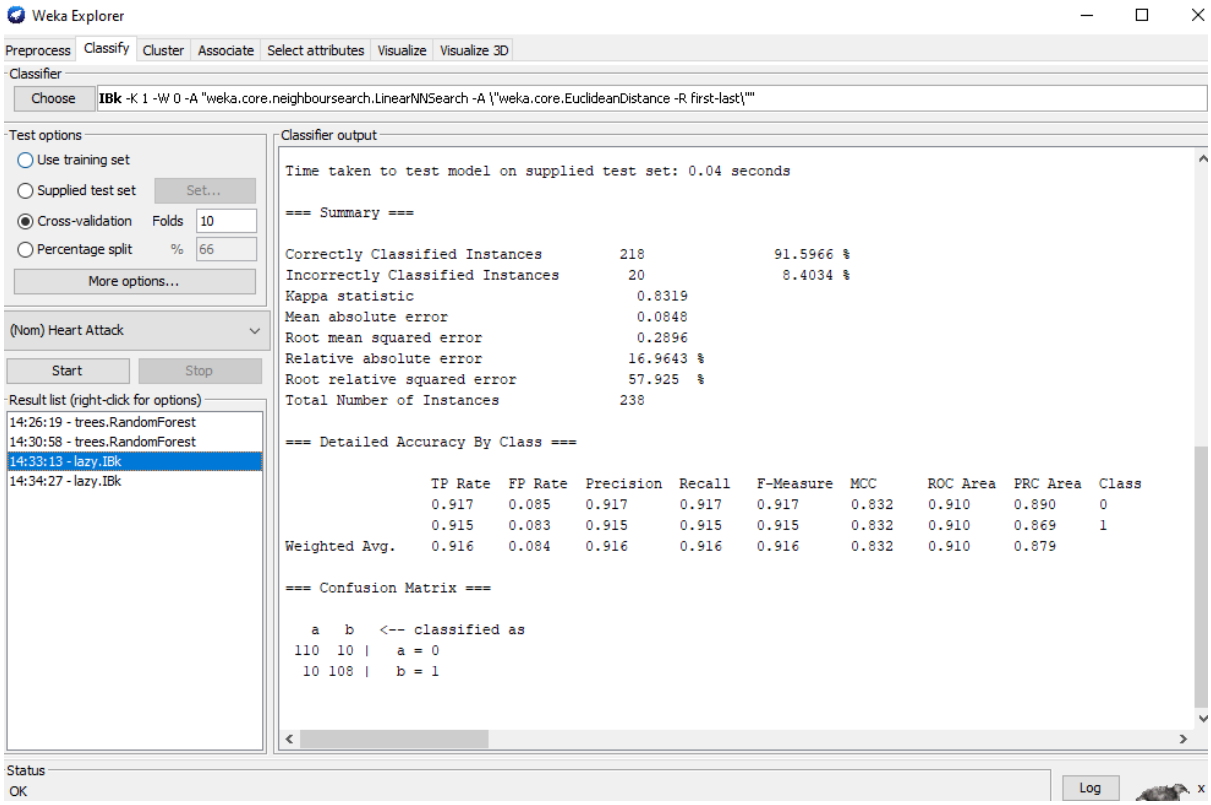


Figure 23

It displays the KNN algorithm on the test set without any adjustment to the parameter. The purpose would be to display the authentic result of comparing the test and training set.

Metrics	**Training Set**	**Test Set**
Accuracy	87.67%	91.60%
Mean Absolute Error (MAE)	0.1241	0.0848
Root Mean Square Error (RMSE)	0.3508	0.2896

Table 5 (Summary for KNN in training set in comparison to test set)

Table 5 shows a larger variance in terms of accuracy of a 3.93% increase in the test set. It is reflected accordingly by the reduction of MAE and RMSE by 0.0393 and 0.0612 respectively. Although the variance exhibited on the higher boundary, it is still below 5% thus it is within an acceptable range.

4.8 Best Pre-selected Machine Model

Metrics	Decision Tree	Random Forest	IBK
Accuracy	87.82%	92.02%	91.60%
Mean Absolute Error (MAE)	0.1468	0.1717	0.0848
Root Mean Square Error (RMSE)	0.3099	0.2698	0.2896

Table 6 (Summary for three algorithms based on the test set result)

As seen in Table 6, Random Forest would be the best-performing model among the three algorithms, with the highest accuracy of 92% Moreover, the RMSE was the lowest for Random Forest at 0.2698, while Decision Tree performed the worse at 0.3099 (a difference of 0.0401). Despite that. Random Forest has the highest MAE of 0.1717 whereas IBK had the lowest (0.0848) which is a difference of 0.0869. Within Table 6, the algorithm was compared before any modification to the parameter was implemented. Thus, the following section will include the modification of parameters on the machine learning model based on journal articles to potentially improve the overall performance.

5.0 Result & Discussion

To allow the reader to understand the result and discussion in the following section, the glossary of each metric will be included. In doing so, it allows the reader to fully understand the importance of each metric selected within the discussion section.

1. Accuracy
    a. $Accuracy\ formula\ = \frac{TP + TN}{TP + TN + FP + FN}$
    b. As seen in the formula for accuracy, it would be a suitable measurement of accuracy when the classes are balanced within the dataset.

2. Precision
    a. $Precision\ = \frac{TP}{TP + FP}$
    b. Precision focuses on informing users of the importance of accuracy in relevance to the positive predictions.
    c. Precision proves to be critical within the medical industry. Due to the fact that it would lead medical practitioners with their next action impacting the patient's life.

3. Recall
    a. $Recall\ = \frac{TP}{TP + FN}$
    b. Recall focus on the metric on capturing all the true positive predictions within the dataset.

4. F1-score
    a. $F1\ score\ formula\ = \frac{TP}{TP + 0.5(FP+FN)}$
    b. Within the metric, the F1-score would be an excellent choice for measuring the model performance on asymmetric class. The researchers initially opted for balancing the data first. However, it was noticed that the test result was underperforming. Thus, the researchers had to revert the metric towards the F1 score (Shin, Park, Lee, Yang & Park, 2021).

5. The Matthews Correlation Coefficient (MCC)
    a. $MCC\ formula\ = \frac{TP\ X\ TN - FP\ X\ FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$
    b. MCC provides a better metric to measure the binary classification outcome. The score system ranges from -1 to 1. With an alternate scoring system,

obtaining an excellent result within the four categories in the confusion matrix. It takes all categories into consideration to inform users of the overall performance of the model (Chicco and Jurman, 2020).

c. Although it is mainly used for the imbalance class, MCC exhibits the application of the balance class as an alternated metric (Chicco and Jurman, 2020).

6. Receiver Operating Characteristic (ROC) Area
    a. ROC analysis allows the various models to quantitatively demonstrate their discriminatory ability. With continuous measurement within the clinical sector, it is often converted to a dichotomous test. Allowing the ROC to illustrate the tradeoff between sensitivity & specificity (Zou, O'Malley & Mauri, 2007).
    b. Moreover, Zou et al (2007) elaborate that the smoothness of the line illustrates a correlation with the accuracy of the prediction (Zou, O'Malley & Mauri, 2007).

7. Precision Recall Curve (PRC)
    a. On top of ROC, PRC would be similar however, it lacks handling of imbalanced datasets (Saito and Rehmsmeier, 2015).
    b. PRC's discriminatory ability would be to quantitatively recall precision and recall.

5.1 Decision Tree

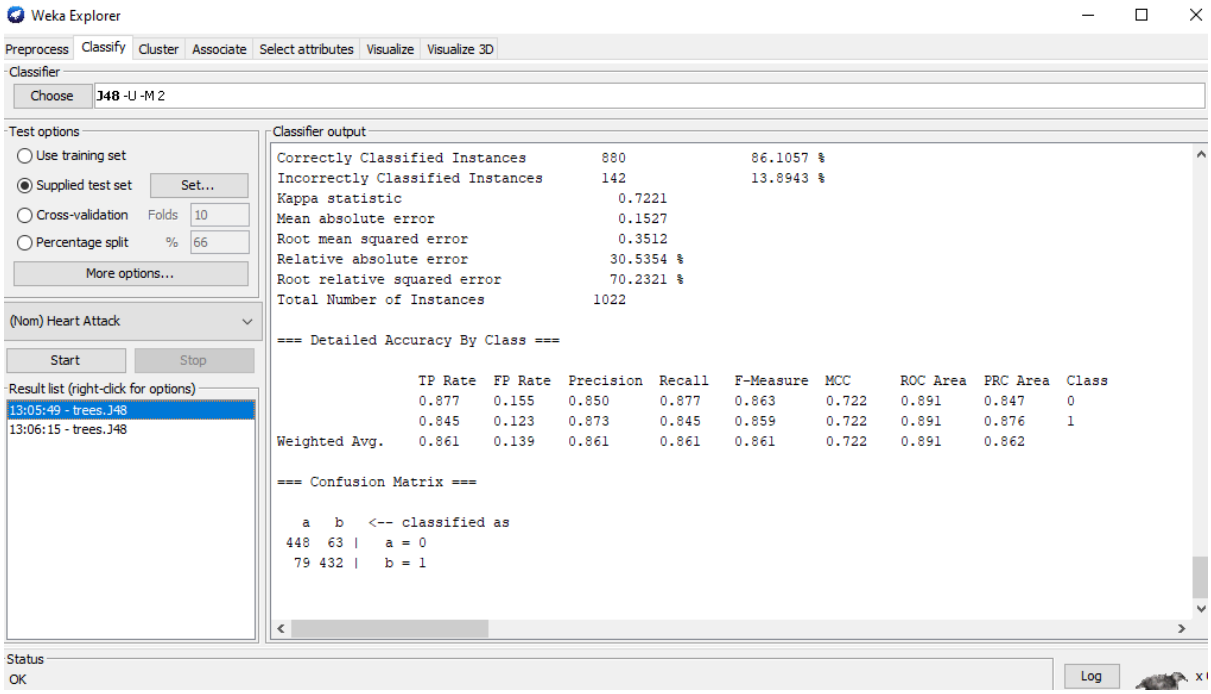


Figure 24 (The parameter was adjusted on the algorithm, it was performed on the training set and achieved an accuracy of 86%.)

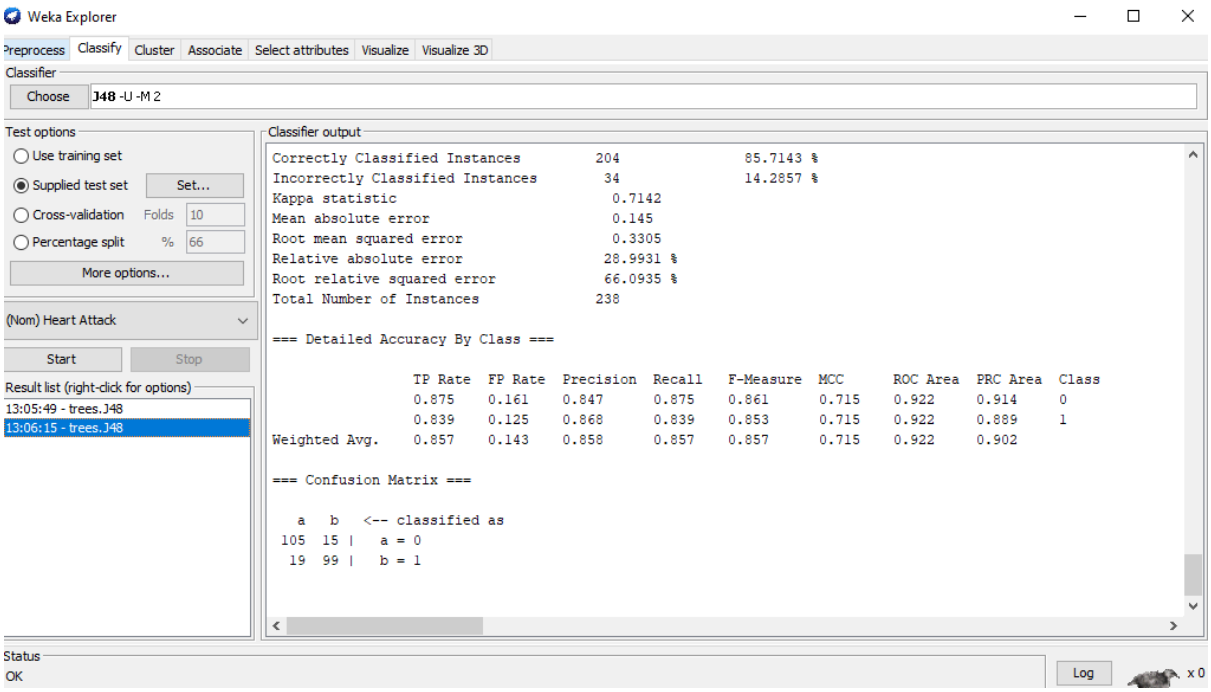


Figure 25 (The parameter was adjusted on the algorithm, it was performed on the test set, and achieved an accuracy of 85%)

To achieve the best result, the project included several suggestions for tuning the parameter. Several research studies will be included to illustrate the various methods of tuning parameters within the industry. Although some parameters may not be applicable to WEKA, it expands the reader's knowledge of the different approaches implemented.

Patil, Wadhai, and Gokhale (2010) research revolves around different techniques of pruning with the decision tree. The researchers found that pruning the decision tree significantly improved the accuracy. Overcomplexing due to allowing the tree to grow deeply contains flaws. Overanalyzing would cause the ML model to capture noise/outlier data. Resulting in the ML model to study the training data pattern to closely and negatively affect the prediction. The researchers discussed several techniques for pruning:

- Reduced Error Pruning
- Pessimistic Error Pruning
- Cost-Complexity Pruning
- Minimum Error Pruning
- Critical Value Pruning
- Optimal Pruning
- Cost-Sensitive Decision Tree Pruning

Although the techniques were introduced, WEKA was unable to utilize them due to its limited capability thus the project chose the generic option ‘unprune’ parameter to prune the decision tree.

Hastie, Tibshirani, and Friedman (2009), the researchers discussed the confidence factor and minimum number of instances per leaf were key parameters to avoid over or under-fitting.

Mantovani, Horvath, Cerri, Vanschoren & Carvalho (2016) suggested the recommended parameter range:

- Prune Confidence: {0.001 ~ 5)
- numFold {2 - 10}

The recommended parameter tuning was:

- Enable the Unprune feature within WEKA.
- Confidence Factor: 0.50.
- Other hyperparameters remained the same.

	No adjusted parameter on the machine learning model (Default setting)	**Recommended tuning by Journal Article**
Accuracy	87.81%	85.71%
Mean Absolute Error (MAE)	0.1468	0.145
Root Mean Square Error (RMSE)	0.3099	0.3305

Table 7(Summary of Decision Tree)

As seen on Table 7, Decision Tree displayed the highest accuracy when no parameter was adjusted to the model. It achieved 87% while the model with the implementation of tuning parameters only achieved 85%. Decision Tree achieved the highest for MAE and RMSE as well. The difference between both models for MAE and RMSE were 0.0018 and 0.0206, RMSE shows the biggest variances among MAE and RMSE.

	No adjusted parameter on the machine learning model (Default setting)	**Recommended tuning by Journal Article**
TP Rate	0.878	0.875
FP Rate	0.122	0.143
Precision	0.878	0.858
Recall	0.878	0.857
F- Measurement	0.878	0.857
MCC	0.756	0.715
ROC Area	0.920	0.922
PRC Area	0.897	0.902

Table 8 (Accuracy by class - Weighted Avg - Decision Tree)

Table 8 displayed that Decision Tree with no parameter tuning outperformed model with parameter tuning. The default setting model outperformed metrics such as TP Rate, ROC Area, and PRC Area by a small margin. MCC has the largest difference when comparing both models. A significant decrease of 0.041 points was noticed after implementing parameter adjustment. While FP Rate, Precision, Recall, and F-measurement drop an average of 0.02 points within the model with the adjustment parameter implemented onto the model.

5.2 Random Forest

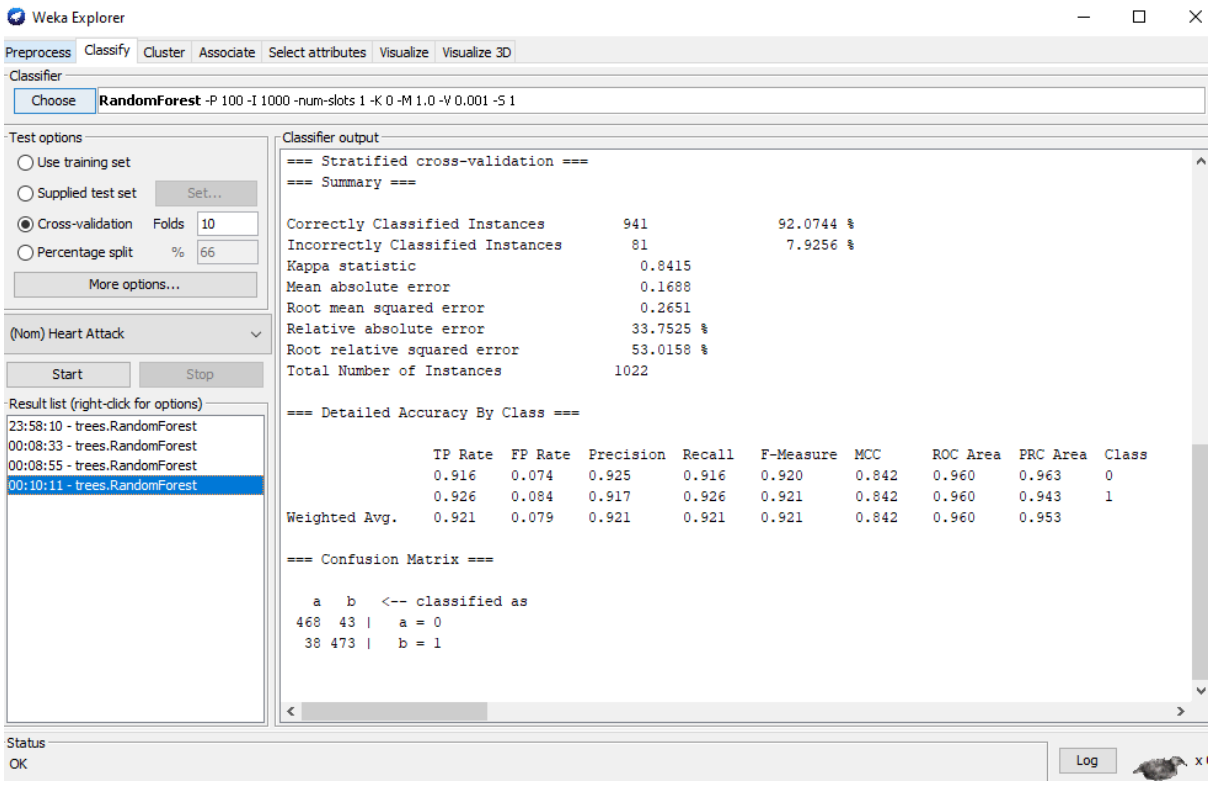


Figure 26 (The parameter was adjusted on the algorithm, it was performed on the training set and achieved an accuracy of 92%)

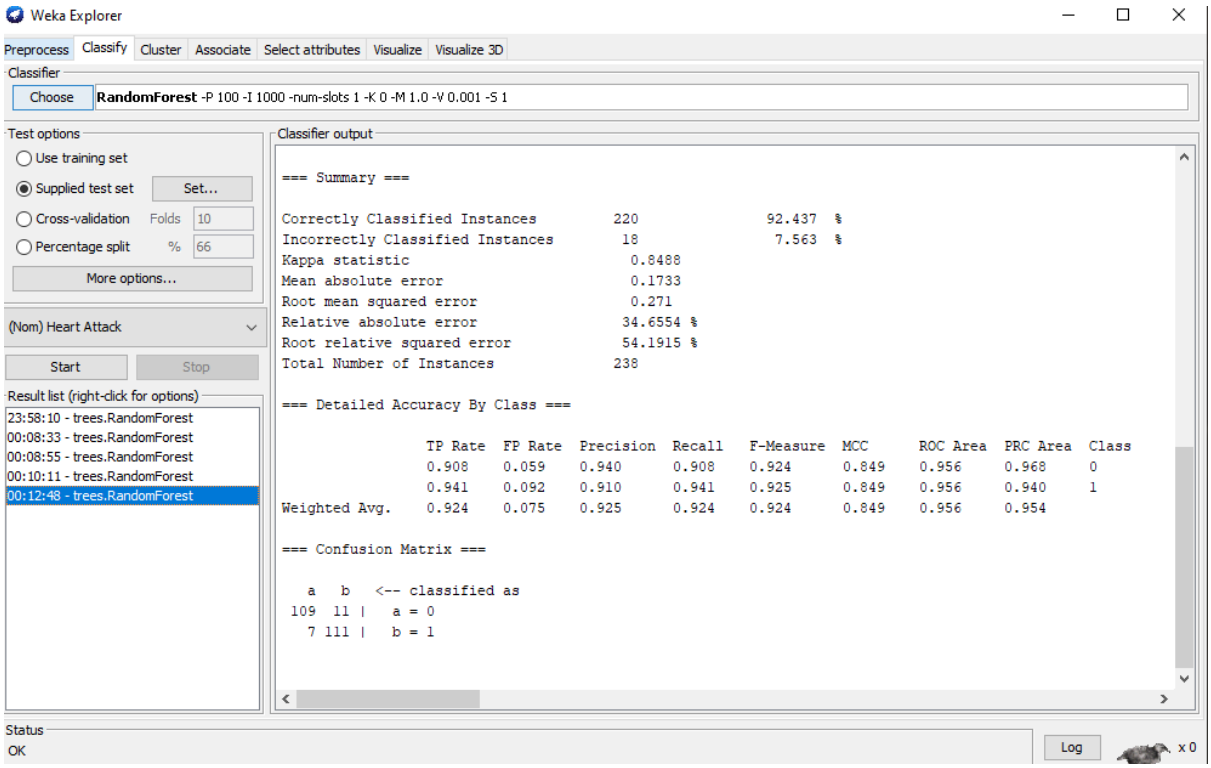


Figure 27 (The parameter was adjusted on the algorithms, it was performed on the test set and achieved an accuracy of 92%)

To achieve the best result, the tuning parameters suggested by the researchers are shown below. Although some parameters may not be applicable to WEKA, it expands the reader's knowledge of the different approaches implemented.

Probst, Wright, and Boulesteix (2019) research suggested fine-tuning several parameters which are:

- Number of forests
- Mtry (randomness)
- Sample size
- Replacement
- Node Size
- Splitting Rule

However, WEKA does not consist of all the parameters within. Thus, the only parameter present would be the number of trees.

- The suggested value for the number of trees was 1000.
- Other hyperparameters remained the same.

Metrics	**No adjusted parameter on the machine learning model (Default setting)**	**Recommended tuning by Journal Article**
Accuracy	92.02%	92.44%
Mean Absolute Error (MAE)	0.1717	0.1733
Root Mean Square Error (RMSE)	0.2698	0.271

Table 9(Summary of Random Forest)

Table 9 exhibits the result of both models. In comparison, both models' results were fairly consistent. Both models achieved an average accuracy of 92%. However, the model with the parameter changes had the highest accuracy of 92.44%. The tuning recommended by the journal article displays an increase of 0.42% in terms of accuracy. The differences between MAE and RMSE were relatively small. Although the recommended tune parameter had a better performance overall.

Metrics	No adjusted parameter on the machine learning model (Default setting)	Recommended tuning by Journal Article
TP Rate	0.920	0.924
FP Rate	0.079	0.075
Precision	0.921	0.925
Recall	0.920	0.924
F- Measurement	0.920	0.924
MCC	0.841	0.849
ROC Area	0.957	0.956
PRC Area	0.956	0.954

Table 10 (Accuracy by class- Weighted Avg - Random Forest)

Table 10 indicates that the model with the parameter tuning had the better result in comparison to the default setting of the model. Although the average differences for TP Rate, FP Rate, Precision, Recall & F-measurement were an increase of 0.004 points on the models with the parameter tuning implemented. Moreover, ROC Area and PRC Area performed better on the default setting of 0.001 & 0.002 points respectively. On the models with tuning parameters adjusted, MCC has shown the largest increase of 0.008 points when compared to the default setting model.

5.3 KNN

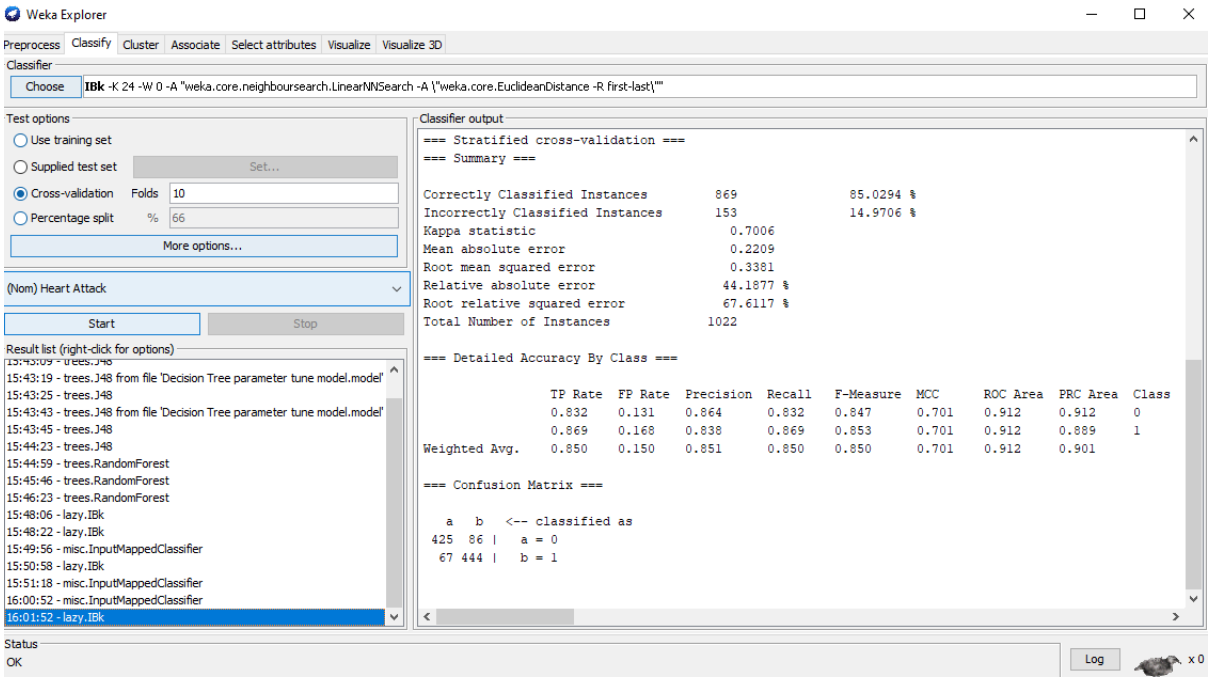


Figure 28 (The parameter was adjusted on the algorithms, it was performed on the training set and achieved an accuracy of 85%)

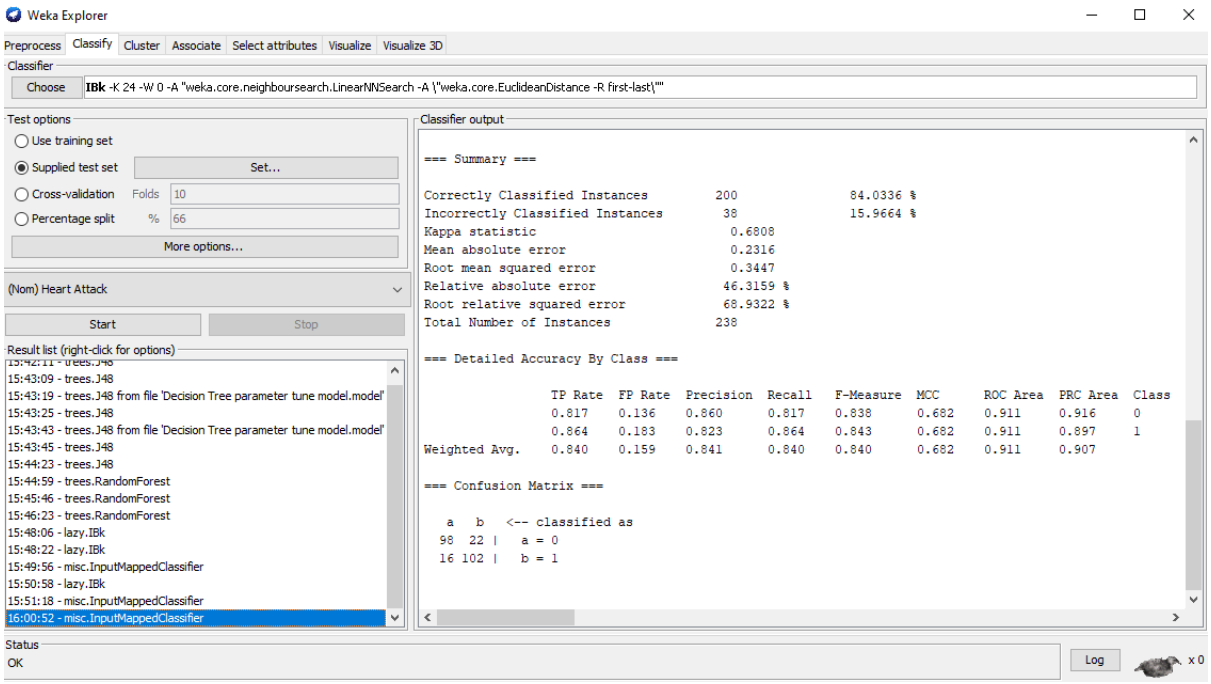


Figure 29 (The parameter was adjusted on the algorithms, it was performed on the test set and achieved an accuracy of 84%)

Due to the nature of KNN, the researchers propose to implement a feature selection to achieve better results. It explores the different types of feature selection within WEKA. Furthermore, another researcher suggested adjusting the number of neighbors (K) to improve the model.

KNN model relies on the similarity feature with several data points for prediction and heavily relies on the feature present within the dataset. Tahir, Bouridane, and Kurugöllü (2004) elaborate on the different types of attribute selection.

- Feature selection (Filter selection)
- Feature Weighting (Wrapped selection)

With both methods in comparison, the researcher recommends the wrapped method over the filter method due to influences on the result. However, the filter method triumphs in terms of time efficiency (Li et al, 2021). Battineni, Chintalapudi, Amenta, & Traini (2020) research emphasizes the importance of utilizing feature weighting as the importance of predicting Alzheimer's disease would be transformative towards an individual life.

Wrapped selection was implemented. The 11 attributes were reduced to 7 attributes which are:

- Sex
- Chest Pain type
- Resting Bps
- Cholesterol
- Exercise-induced Angina
- Oldpeak
- ST Slope

Udovychenko, Popov, & Chaikovsky (2015) explore the effectiveness of tuning the parameter of the number of neighbors (k) along with several distance metrics. The recommended parameter was:

- KNN range: 19 - 27
    - Among the KNN range: 24 displayed the highest.
    - With the inclusion of wrapped attributes.

Metrics	**No adjusted parameter & no attribute reduction was applied on the machine learning model (Default setting)**	**Recommended tuning by Journal Article**
Accuracy	91.60%	84.03%
Mean Absolute Error (MAE)	0.0848	0.2316
Root Mean Square Error (RMSE)	0.2896	0.3447

Table 11 (Summary of KNN)

In Table 11, KNN had the largest drop in accuracy after the implementation of parameters into the model. A significant decrease of 7.57% was observed from the default model. The highest accuracy was achieved with the default setting (without any parameter applied with the exclusion of attribute reduction) of the model at 91.60% while the other model achieved an accuracy of only 84.03%. A steep increase of 0.1468 and 0.0551 was observed within MAE and RMSE respectively.

Metrics	**No adjusted parameter on the machine learning model (Default setting)**	**Recommended tuning by Journal Article**
TP Rate	0.916	0.840
FP Rate	0.0084	0.159
Precision	0.916	0.841
Recall	0.916	0.840
F- Measurement	0.916	0.840
MCC	0.832	0.682
ROC Area	0.910	0.911
PRC Area	0.879	0.907

Table 12 (Accuracy by class- Weighted Avg - KNN)

Across Table 12 metric, the majority of the individual class accuracy dropped sharply after the parameter's changes were implemented. The largest drop in the individual accuracy class was the FP Rate, an increase from 0.0084 to 0.159, a change of 0.1506 points. The average decrease for TP rate, Precision, Recall & F-measurement was 0.075. Surprisingly, the two individual class accuracy that outperformed the default setting was ROC Area & PRC Area. The implementation of the parameter changes displays an increase of 0.001 & 0.008 points.

5.4 Best Machine Model

Metrics	**Decision Tree- Default Setting**	**Random Forest Recommended by Journal Article**	**KNN - Default Setting**
Accuracy	87.81%	92.44%	91.60%
Mean Absolute Error (MAE)	0.1468	0.1733	0.0848
Root Mean Square Error (RMSE)	0.3099	0.271	0.2896

Table 13 (Summary of top performing models)

In comparison with the three algorithms, Random Forest (with the implementation of parameter changes) displayed the highest accuracy of 92% along with lowest RMSE of 0.271. However, within the three models, it had the highest MAE value of 0.1733 while KNN had the lowest of 0.0848. Decision Tree had the lowest accuracy of 87% within the three models while having the highest RMSE value of 0.3099.

Metrics	Decision Tree- Tuning made by team	Random Forest Recommended by Journal Article	KNN - Default Setting
TP Rate	0.878	0.924	0.807
FP Rate	0.122	0.075	0.193
Precision	0.878	0.925	0.808
Recall	0.878	0.924	0.807
F- Measurement	0.878	0.924	0.807
MCC	0.756	0.849	0.614
ROC Area	0.920	0.956	0.888
PRC Area	0.897	0.954	0.887

Table 14 (Accuracy by class- Weighted Avg - Top performing models)

Table 14 shows that the Random Forest (with the implementation of parameter changes) significantly outperformed among the three algorithms in all individual class accuracy. Random Forest outperformed Decision Tree and IBK by an average of 0.046 points and 0.117 points within the following five categories: TP Rate, FP Rate, Precision, Recall & F-measurement. Moreover, the highest TP rate was Random Forest of 0.924 while the lowest was 0.807. Once again, Random Forest had the best score for an FP Rate of 0.075 while the other two models displayed significantly extreme values exceeding the 0.100 threshold. The respective values for Decision Tree and KNN were 0.122 & 0.193.

In terms of precision, only Random Forest exceeded the 0.900 threshold and achieved 0.925. While the two other models were hovering around the 0.800 thresholds of 0.878 and 0.808 for Decision Tree and KNN respectively. Recall and F-Measurement show almost similar results exhibiting random forest exceeding the 0.900 mark while the other two models average on the 0.800 mark. The biggest difference would be Random Forest in comparison with IBK's MCC value, the difference was 0.235 points. While the comparison of Random Forest and Decision tree MCC value was only 0.093 points.

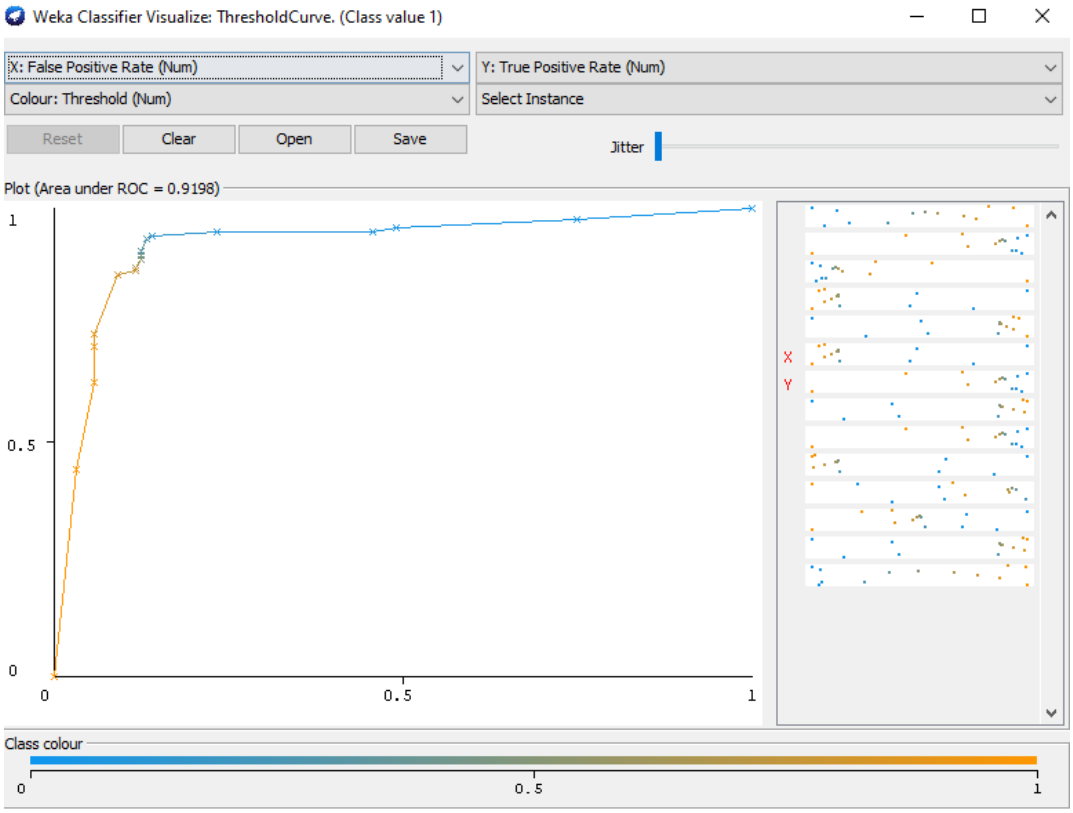


Figure 30 (ROC-AUC Curve for the Decision Tree)

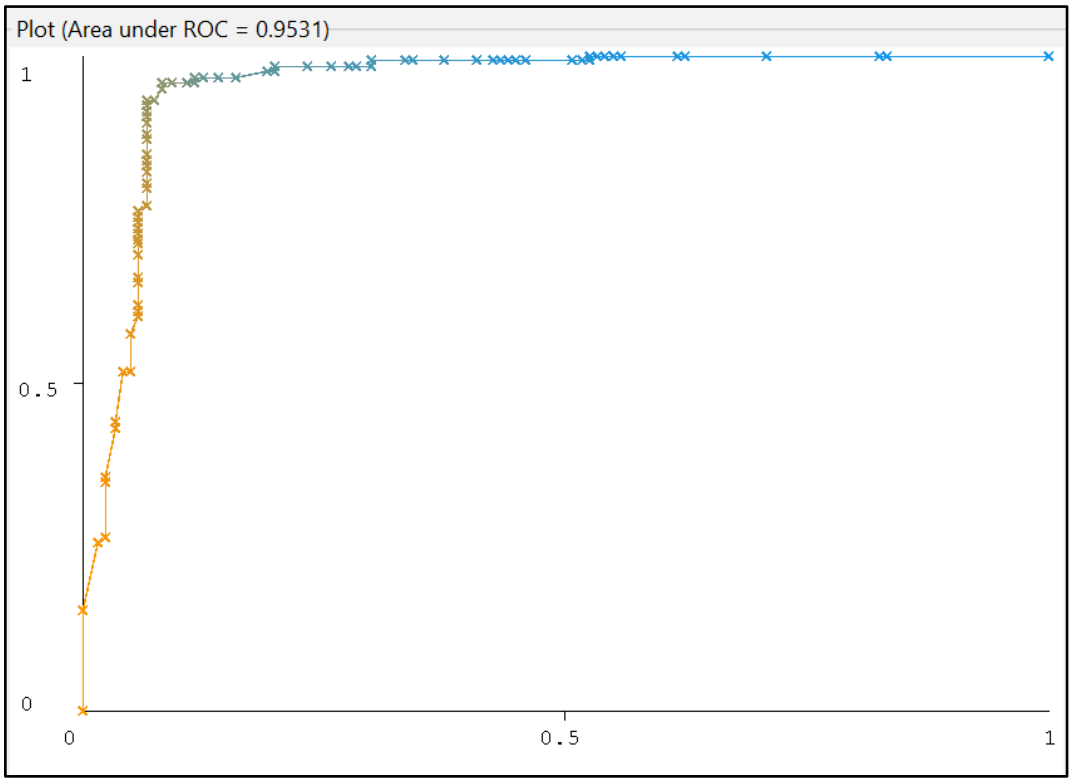


Figure 31 (ROC-AUC Curve for Random Forest)

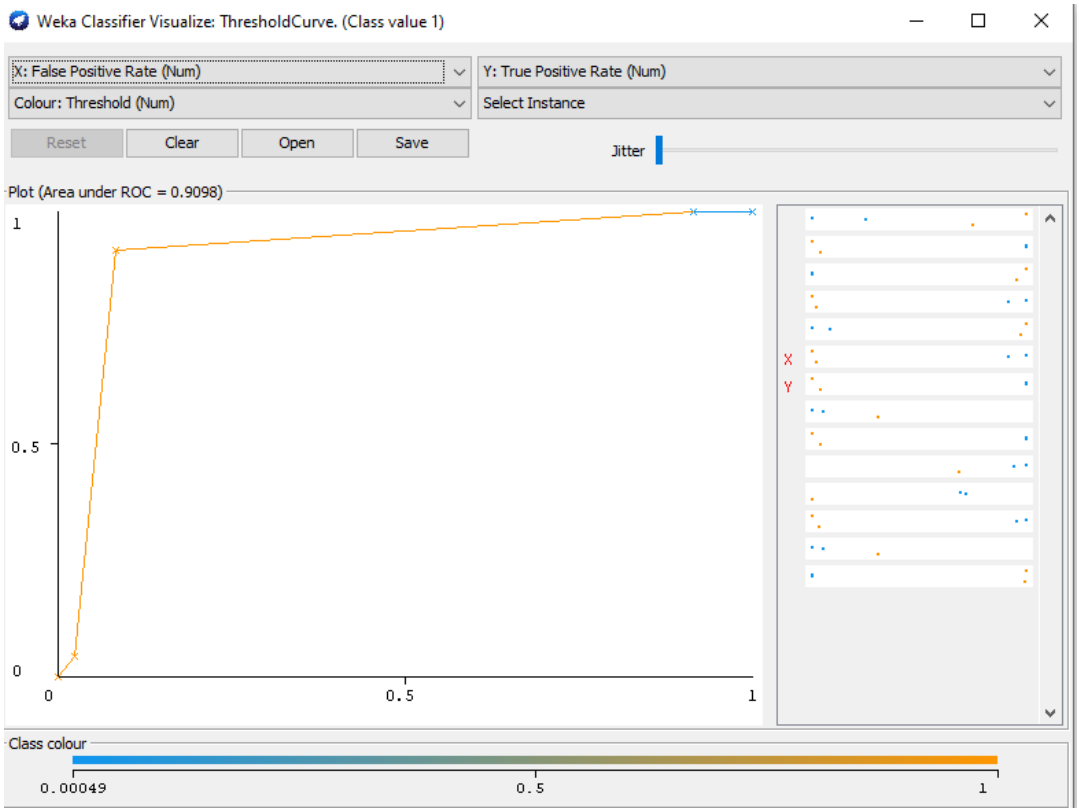


Figure 32 (ROC-AUC Curve for KNN)

To further evaluate the ROC-AUC Curve, Zou et al (2007) state '*When the true ROC curve is a smooth function, the precision of statistical inferences based on the empirical ROC curve is reduced relative to a model-based estimator'*. Thus, upon the evaluation of the ROC-AUC Curve, the inclusion of determining the curve smoothness could provide researchers with a deeper understanding of the diagram. Upon reviewing Figures 30, 31, and 32, the differences between the three were easily distinguishable. The Decision Tree as seen in Figure 30, it illustrates the curve was jagged among the three models. The curve was nonlinear when compared. Therefore, according to Zou et al (2007), it displayed that the model signifies that the precision of statistical inferences would be greatly reduced. Interestingly, KNN (Figure 32) exhibits two straight lines connected within the graph. This means that precision within the model is robust. However, it had the lowest threshold point among the three. Lastly, Random Forest (Figure 31) had the highest threshold point among the three while maintaining a smooth curve within the graph. Therefore, it means that the algorithm was able to create predictions with the highest True Positive while minimizing False Positive instances. Moreover, the curve smoothness minimizes the reduction of precision.

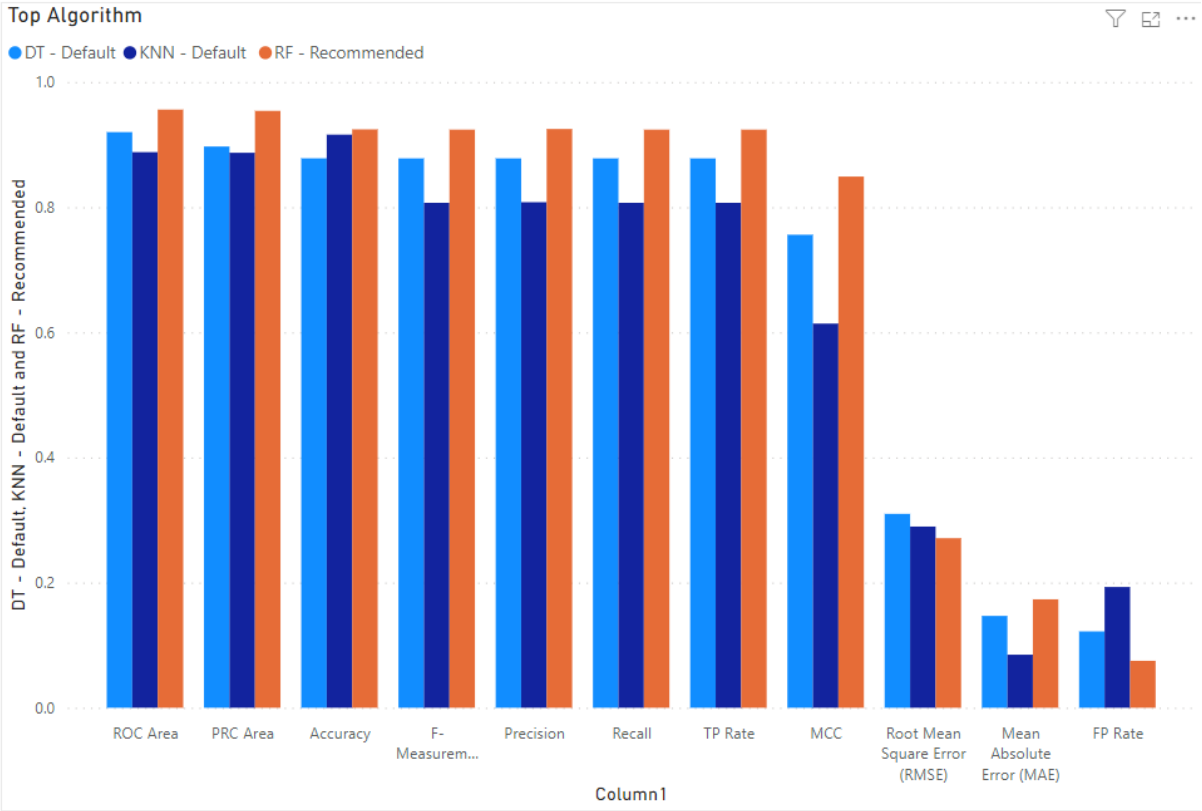


Figure 34 (Clustered column chart of the top three algorithms)

As seen in Figure 30, the clustered column chart allows users to visualize each component uncomplicatedly. Within the figure, it is easily distinguishable that the Random Forest model excels in most of the measurement metrics. As illustrated on the right side, a lower rating for RMSE, MAE, and FP Rate would be an ideal scenario. Although the Decision Tree and Random Forest are generally on par, the Random Forest outperforms the Decision Tree due to implementing an ensemble learning model nature. The ensemble learning model allows Random Forest to overcome the influences of noise on the dataset. Due to the sensitivity of the Decision Tree, small changes could cause the model to overfit (Bertsimas and Dunn, 2017). The benefits of Random Forest create several base trees and by randomly selecting a subset of attributes, the tree would be built. Decision Tree splits the parent and child nodes based on the Gini Impurity and information gain (Cutler et al, 2007). Whereas, with several base trees built within the model, it allows attribute weightage which improves the heart prediction (Ansarullah et al, 2022). Among the three models, KNN performs the poorest. According to Garcia, Debreuve & Barlaud (2008) research, it suggested that tuning the K value and finding the optimal value proved to be time-consuming. Although the project followed the suggested K value proposed by a journal article, the result underperformed when compared to the model without any parameter changes. Thus, in the search for the optimal value, a grid search methodology would be recommended to improve the performance (Bergstra and

Bengio, 2012). Furthermore, the attribute selection was conducted to improve the model. From a medical perspective, attribute reduction may not prove to be ideal as every attribute holds a certain amount of weightage.

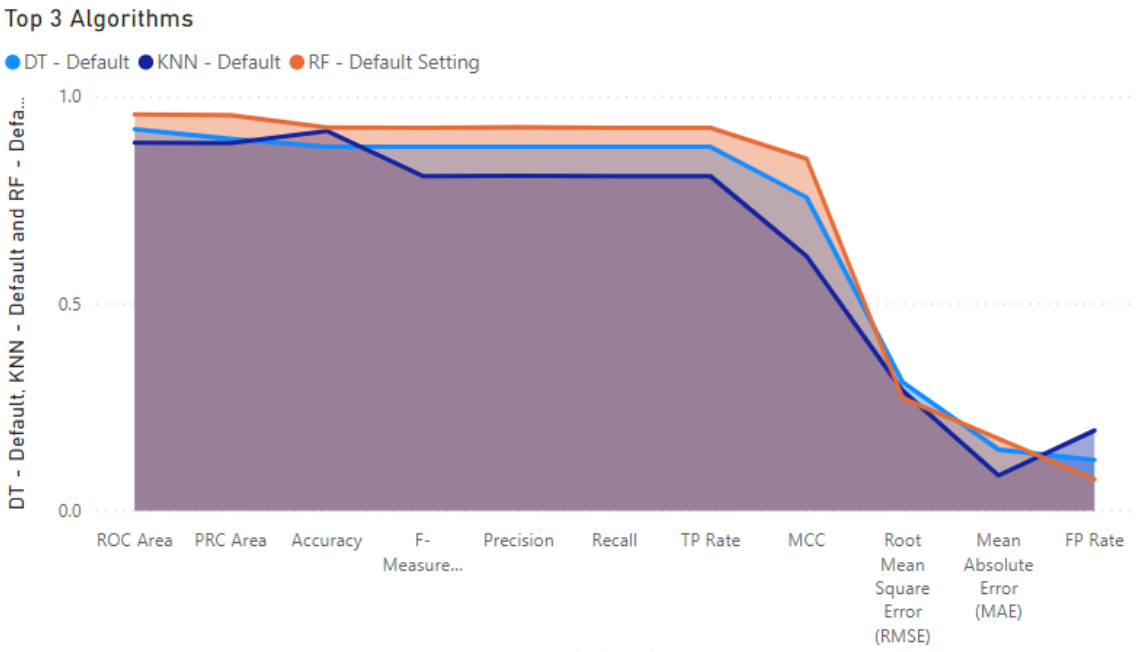


Figure 35 (Area chart of the top three algorithms)

With Figure 35, it allows the reader to visualize the algorithm's performance in an area chart. Allowing the visualization of how each algorithm category is compared.

6.0 Limitation

Upon obtaining the dataset, the data was not complete. It is not an ideal scenario as imputing data (ReplaceMissingValue Filter) was necessary. Furthermore, only 1190 instances were available within the dataset. Ideally, the project would prefer a larger dataset to develop the machine learning models. In comparison to the medical industry, the actual statistic of patients having heart diseases would be significantly higher than 1190 instances. Moreover, a limitation of attribution within the dataset. With tons of data stored within the hospital data warehouse, the project assumed that these 11 attributes had the largest weightage. With the vast number of patients recorded within a database, it is crucial for the models to learn with larger datasets along with multiple dimensions to develop robust algorithms. As the aim of the project would be to develop a machine-learning model for Malaysia, certain features were not included within. Thus, the machine learning model could be biased toward the demographic for Cleveland, Hungarian, Switzerland, and Long Beach VA. With different regions, the lifestyle of the average individual would vary drastically hence the prediction of the heart disease detection model requires further tuning to cater to the Malaysian demographic.

7.0 Conclusion

Upon reviewing the model, the project discovered that Random Forest was the best model. With the highest performance among the various machine learning models. Moreover, the training and test set indicates that the algorithms proved to be robust as the accuracy was 93.37% and 93.28% respectively. In contrast, KNN had the largest discrepancy of 10.93% between the training and test sets. Suggesting the model to be overfitted despite various journal article recommendations. Due to the shortcomings of KNN, searching the optimal K value requires utilizing a grid search approach. However, the project uses the recommended K value range suggested. The recommended range could be optimal for that research however it may not apply to the project case. Lastly, the Decision Tree was the second best-performing model, achieving an accuracy of 87.81%. Due to the sensitivity of the model towards the dataset, the model requires pruning to stop the models from overanalyzing the pattern & trend from the noise & outlier present within the data. Due to WEKA limitation, the project was unable to try various methods of pruning suggested by the journal article. In conclusion, despite the limitation, the model has achieved satisfactory accuracy.

References

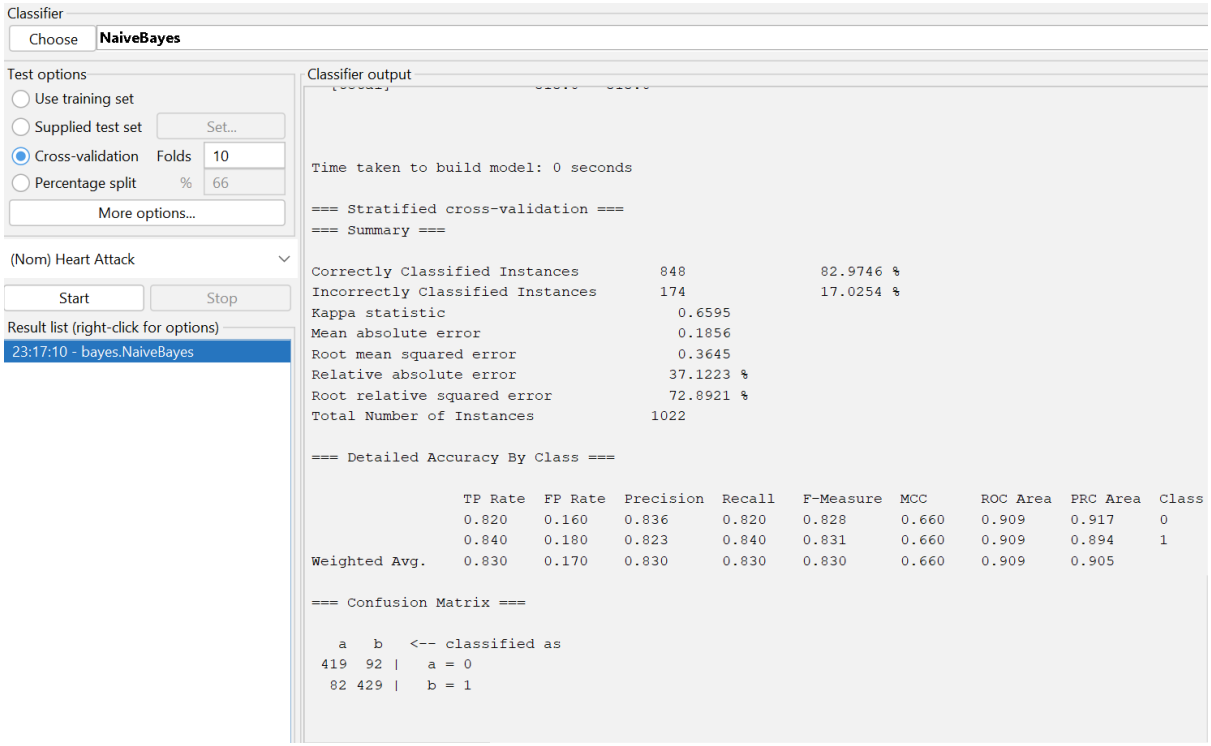
1. Manu Siddhartha, November 5, 2020, "Heart Disease Dataset (Comprehensive)", IEEE Dataport, doi: https://dx.doi.org/10.21227/dz4t-cm36.
2. Ansarullah, S., Saif, S., Kumar, P., & Kirmani, M. (2022). Significance of visible non-invasive risk attributes for the initial prediction of heart disease using different machine learning techniques. Computational Intelligence and Neuroscience, 2022, 1-12. https://doi.org/10.1155/2022/9580896
3. Battineni, G., Chintalapudi, N., Amenta, F., & Traini, E. (2020). A comprehensive machine-learning model applied to magnetic resonance imaging (mri) to predict alzheimer's disease (ad) in older subjects. Journal of Clinical Medicine, 9(7), 2146. https://doi.org/10.3390/jcm9072146
4. Bhattacharya, G., Ghosh, K., & Chowdhury, A. (2015). A probabilistic framework for dynamic k estimation in knn classifiers with certainty factor. 2015 Eighth International Conference on Advances in Pattern Recognition (ICAPR). https://doi.org/10.1109/icapr.2015.7050683
5. Cao, J., Kwong, S., & Wang, R. (2012). A noise-detection based adaboost algorithm for mislabeled data. Pattern Recognition, 45(12), 4451-4465. https://doi.org/10.1016/j.patcog.2012.05.002
6. Cutler, D. R., Edwards, T. C., Beard, K. H., Cutler, A., Hess, K. T., Gibson, J., & Lawler, J. J. (2007). Random forests for classification in ecology. Ecology, 88(11), 2783–2792. https://doi.org/10.1890/07-0539.1
7. Firus Khan, A. Y., Ramli, A. S., Abdul Razak, S., Mohd Kasim, N. A., Chua, Y.-A., Ul-Saufie, A. Z., Jalaludin, M. A., & Nawawi, H. (2022). The Malaysian Health and Wellbeing Assessment (myhebat) study protocol: *An initiation of a national registry for Extended Cardiovascular Risk Evaluation in the community. International Journal of Environmental Research and Public Health*, 19(18), 11789. https://doi.org/10.3390/ijerph191811789
8. Goldstein, B. A., Polley, E. C., & Briggs, F. (2011). Random forests for genetic association studies. Statistical Applications in Genetics and Molecular Biology, 10(1). https://doi.org/10.2202/1544-6115.1691
9. Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: data mining, inference, and prediction. Springer Science & Business Media.
10. Karabulut, E., Özel, S., & Ibrikci, T. (2012). A comparative study on the effect of feature selection on classification accuracy. Procedia Technology, 1, 323-327. https://doi.org/10.1016/j.protcy.2012.02.068

11. Kaur, A. (2017). A comprehensive approach to predict heart diseases using data mining. *International Journal of Innovations in Engineering and Technology,* 8(2). https://doi.org/10.21172/ijiet.82.003
12. Kerr, J., Patterson, R., Ellis, K., Godbole, S., Johnson, E., Lanckriet, G., & Staudenmayer, J. (2016). Objective assessment of physical activity. Medicine & Science in Sports & Exercise, 48(5), 951-957. https://doi.org/10.1249/mss.0000000000000841
13. Korb, K. B. and Nicholson, A. E. (2003). Bayesian artificial intelligence.. https://doi.org/10.1201/9780203491294
14. Li, Q., Liu, Y., Zhu, J., Chen, Z., Yang, S., Zhu, G., … & Chen, L. (2021). Upper-limb motion recognition based on hybrid feature selection: algorithm development and validation. Jmir Mhealth and Uhealth, 9(9), e24402. https://doi.org/10.2196/24402
15. Mabu, S., Obayashi, M., & Kuremoto, T. (2016). An evolutionary algorithm for making decision graphs for classification problems. *Journal of Robotics, Networking and Artificial Life,* 3(1), 45. https://doi.org/10.2991/jrnal.2016.3.1.11
16. Nababan, A. A., Sitompul, O. S., & Tulus, .. (2018). Attribute weighting based k-nearest neighbor using gain ratio. *Journal of Physics: Conference Series*, 1007, 012007. https://doi.org/10.1088/1742-6596/1007/1/012007
17. Patil, D., Wadhai, V., & Gokhale, J. (2010). Evaluation of decision tree pruning algorithms for complexity and classification accuracy. International Journal of Computer Applications, 11(2), 23-30. https://doi.org/10.5120/1554-2074
18. Probst, P., Wright, M., & Boulesteix, A. (2019). Hyperparameters and tuning strategies for random forest. Wiley Interdisciplinary Reviews Data Mining and Knowledge Discovery, 9(3). https://doi.org/10.1002/widm.1301
19. Qian, Z. (2014). Learning bayes nets for relational data with link uncertainty. Lecture Notes in Computer Science, 123-137. https://doi.org/10.1007/978-3-319-04534-4_9
20. Rahim, R. and Ahmar, A. S. (2022). Cross-validation and validation set methods for choosing k in knn algorithm for healthcare case study. *JINAV: Journal of Information and Visualization,* 3(1), 57-61. https://doi.org/10.35877/454ri.jinav1557
21. Rodríguez, M. A., Alesanco, Á., Mehavilla, L., & García, J. M. M. (2022). Evaluation of machine learning techniques for traffic flow-based intrusion detection. Sensors, 22(23), 9326. https://doi.org/10.3390/s22239326
22. Shahri, N. H. N. B. M., Lai, S. B. S., Mohamad, M., Rahman, H. A. B. A., & Rambli, A. (2021). Comparing the performance of adaboost, xgboost, and logistic regression for imbalanced data. Mathematics and Statistics, 9(3), 379-385. https://doi.org/10.13189/ms.2021.090320

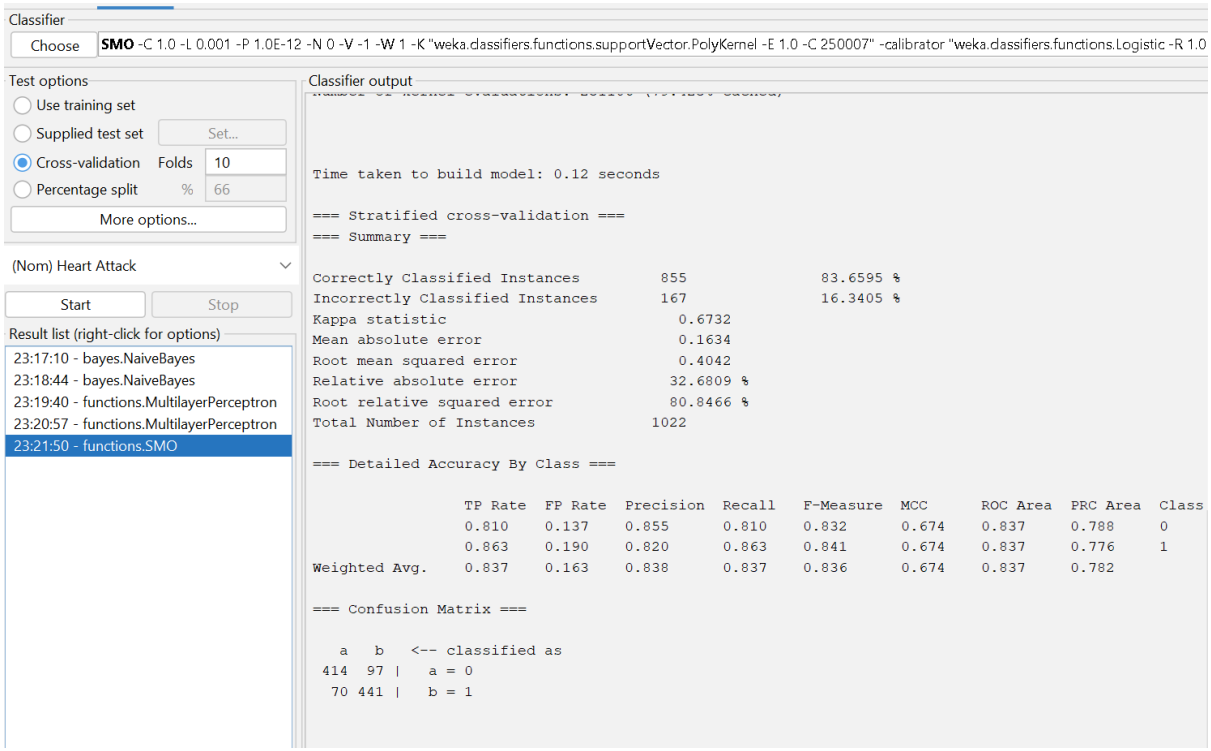
23. Spencer, R., Thabtah, F., Abdelhamid, N., & Thompson, M. P. (2020). Exploring feature selection and classification methods for predicting heart disease. Digital Health, 6, 205520762091477. https://doi.org/10.1177/2055207620914777
24. Tahir, M. A., Bouridane, A., & Kurugöllü, F. (2004). Simultaneous feature selection and weighting for nearest neighbor using tabu search. Lecture Notes in Computer Science, 390-395. https://doi.org/10.1007/978-3-540-28651-6_57
25. Tong, W., Xie, Q., Hong, H., Fang, H., Shi, L., Perkins, R., … & Petricoin, E. (2004). Using decision forest to classify prostate cancer samples on the basis of seldi-tof ms data: assessing chance correlation and prediction confidence. Environmental Health Perspectives, 112(16), 1622-1627. https://doi.org/10.1289/ehp.7109
26. Udovychenko, Y., Popov, A., & Chaikovsky, I. (2015). Binary classification of heart failures using k-nn with various distance metrics. International Journal of Electronics and Telecommunications, 61(4), 339-344. https://doi.org/10.2478/eletel-2015-0044
27. Wan Musa, W. Z., Ahmad, A., Abu Bakar, N. A., Arfah, N. W.-, Ramli, A. W., & Naing, N. N. (2022). Predictors of coronary heart disease (CHD) among Malaysian adults: Findings from Mydiet-CHD Study. *Malaysian Journal of Medicine and Health Sciences*, 18(6), 259–269. https://doi.org/10.47836/mjmhs.18.6.34
28. Wu, Y. and Fang, Y. (2020). Stroke prediction with machine learning methods among older chinese. *International Journal of Environmental Research and Public Health,* 17(6), 1828. https://doi.org/10.3390/ijerph17061828
29. Xing, W. and Bei, Y. (2020). Medical health big data classification based on knn classification algorithm. IEEE Access, 8, 28808-28819. https://doi.org/10.1109/access.2019.2955754
30. Simon, R., Subramanian, J., Li, M., & Menezes, S. (2011). Using cross-validation to evaluate predictive accuracy of survival risk classifiers based on high-dimensional data. Briefings in Bioinformatics, 12(3), 203-214. https://doi.org/10.1093/bib/bbr001
31. Lemm, S., Blankertz, B., Dickhaus, T., & Müller, K. R. (2011). Introduction to machine learning for brain imaging. NeuroImage, 56(2), 387-399. https://doi.org/10.1016/j.neuroimage.2010.11.004
32. Shin, S. J., Park, J., Lee, S. H., Yang, K., & Park, R. W. (2021). Predictability of mortality in patients with myocardial injury after noncardiac surgery based on perioperative factors via machine learning: retrospective study. JMIR Medical Informatics, 9(10), e32771. https://doi.org/10.2196/32771
33. Chicco, D. and Jurman, G. (2020). The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. BMC Genomics, 21(1). https://doi.org/10.1186/s12864-019-6413-7

34. Zou, K. H., O'Malley, A. J., & Mauri, L. (2007). Receiver-operating characteristic analysis for evaluating diagnostic tests and predictive models. Circulation, 115(5), 654-657. https://doi.org/10.1161/circulationaha.105.594929
35. Saito, T. and Rehmsmeier, M. (2015). The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. Plos One, 10(3), e0118432. https://doi.org/10.1371/journal.pone.0118432
36. Mantovani, R. G., Horvath, T., Cerri, R., Vanschoren, J., & Carvalho, A. C. P. L. F. (2016). Hyper-parameter tuning of a decision tree induction algorithm. 2016 5th Brazilian Conference on Intelligent Systems (BRACIS), 37–42. https://doi.org/10.1109/bracis.2016.018
37. Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. Journal of Machine Learning Research, 13(Feb), 281-305.
38. Bertsimas, D. and Dunn, J. (2017). Optimal classification trees. Machine Learning, 106(7), 1039-1082. https://doi.org/10.1007/s10994-017-5633-9

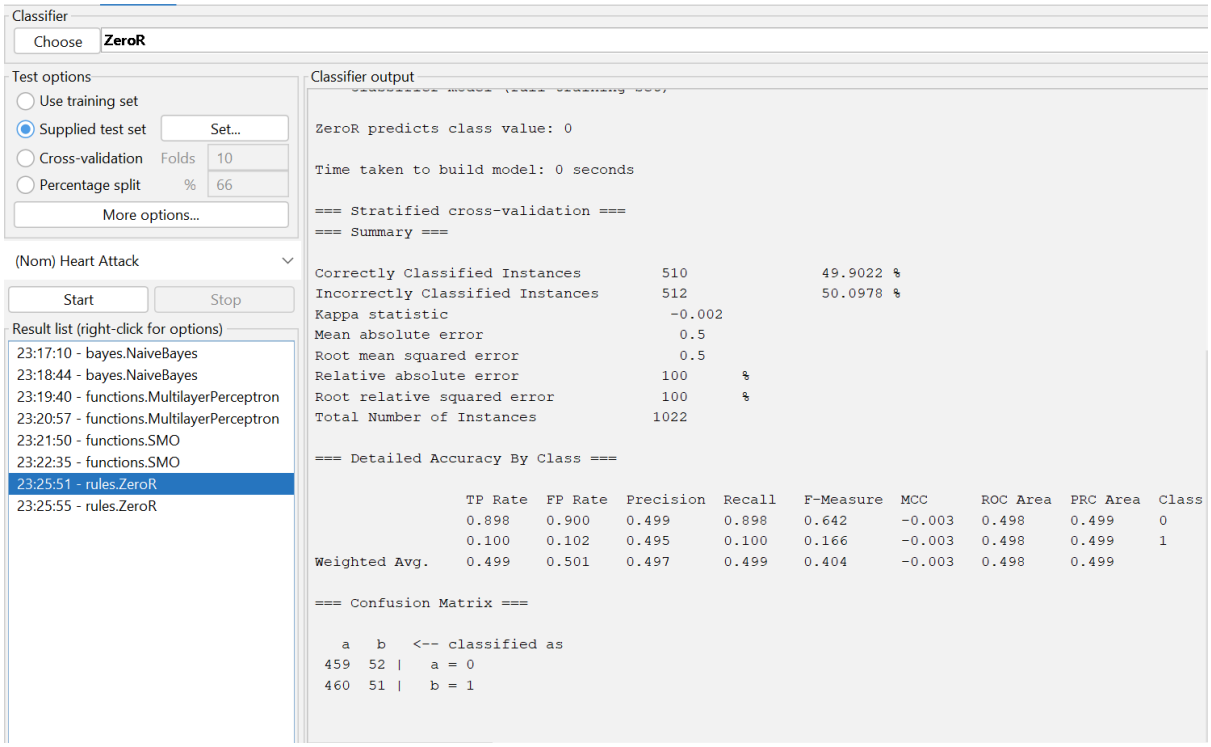
Appendix



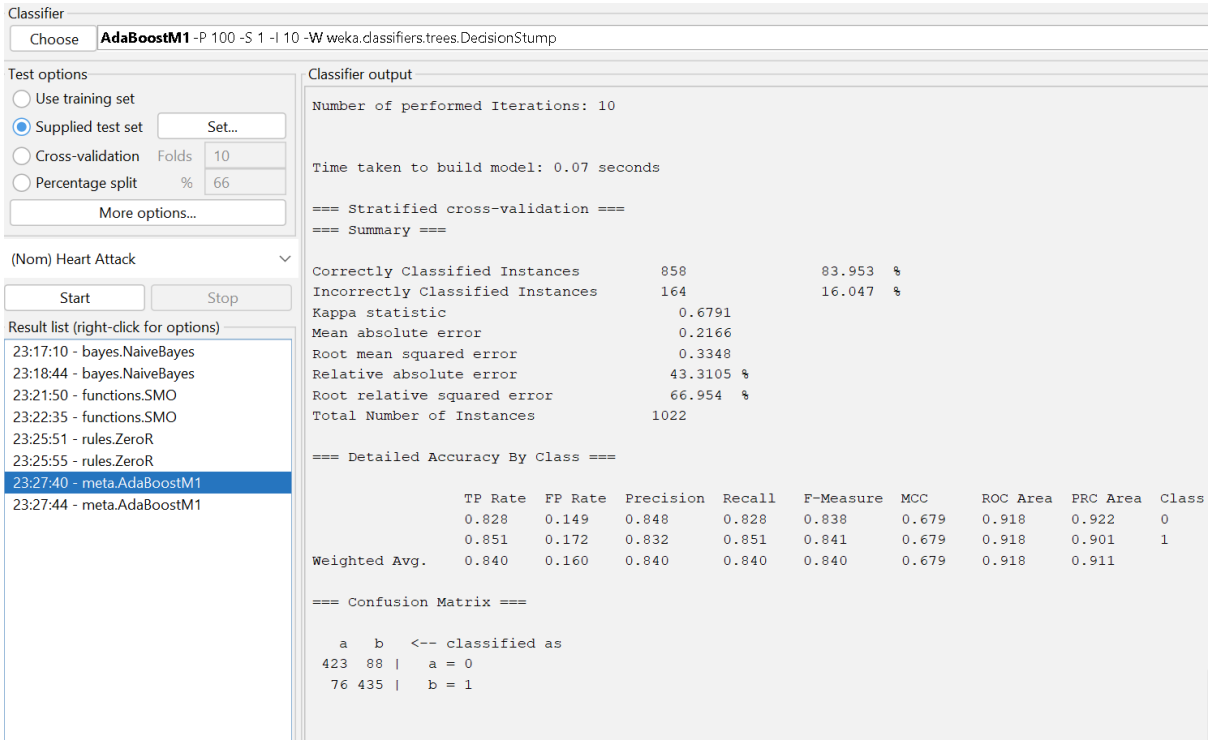
Figureure 36 (Naives Bayes model on the training set)



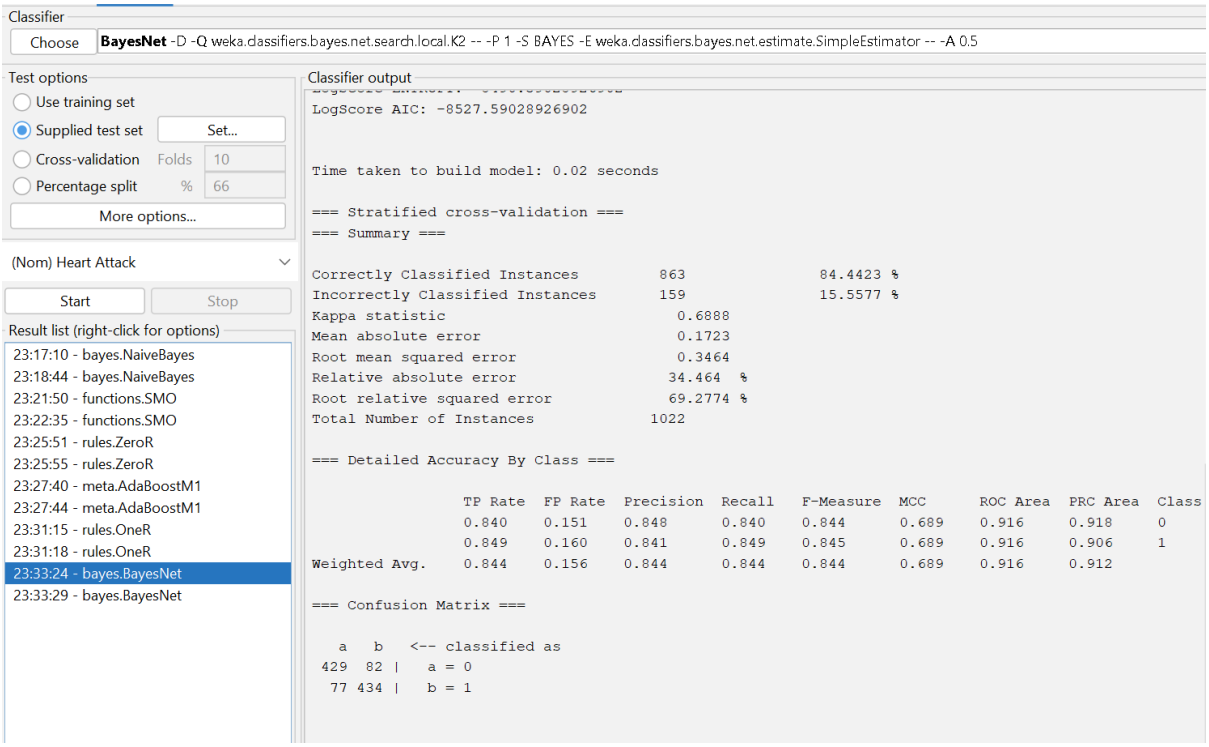
Figureure 37 (Support Vector Machine model on the training set)



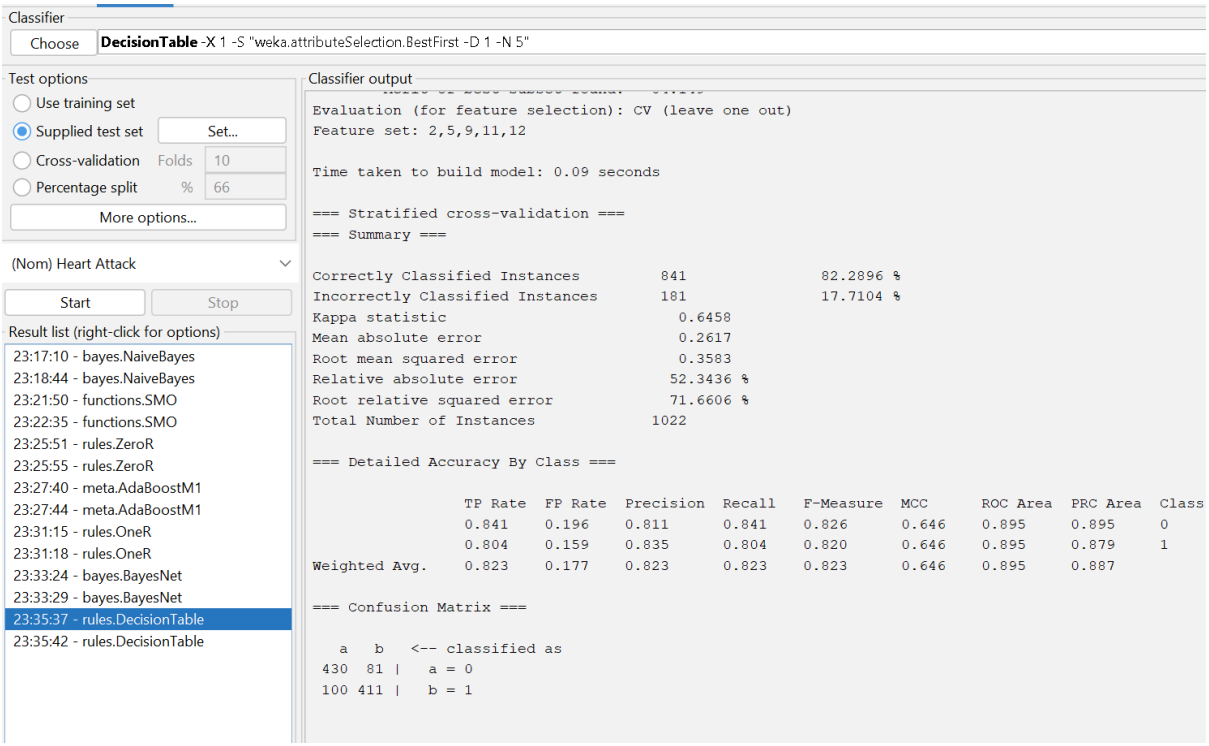
Figureure 38 (ZeroR model on the training set)



Figureure 39 (AdaBoostM1 model on the training set)



Figureure 40 (BayesNet model on the training set)



Figureure 41 (DecisionTable model on the training set)