THE OPEN UNIVERSITY OF SRI LANKA

DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING

BACHELOR OF TECHNOLOGY / BACHELOR OF SOFTWARE ENGINEERING

EEX6377

Principles And Applications Of Data Mining

MINI PROJECT

In-Depth Study On Support Vector Machine (SVM) For Chronic Kidney Disease (CKD) Prediction

M.I.F. ISHKA

721430546 / S92070546

14.12.2024

CONTENT

1. INTRODUCTION

The steady deterioration of kidney function, which frequently leads to end-stage renal failure, is the hallmark of chronic kidney disease (CKD), a serious worldwide health concern. Although early detection is crucial for preventing serious consequences, conventional diagnostic techniques usually fall short in detecting CKD in its early stages because of weak or nonexistent symptoms.

By identifying intricate patterns in data, machine learning (ML) has the potential to revolutionize medical diagnoses. Support Vector Machines (SVM), one of the best machine learning approaches for early disease diagnosis, excel at accurately classifying high-dimensional data.

The Chronic Kidney Disease dataset from the UCI Machine Learning Repository is used in this study to create an SVM-based classification system for CKD prediction. Model performance will be improved through feature selection, and relative efficacy will be assessed through comparisons with classifiers like Decision Trees and Naive Bayes.

In order to facilitate early CKD detection, improve clinical decision-making, and lessen healthcare costs, the project intends to develop a reliable, effective diagnostic instrument.

2. PROBLEM BACKGROUND

Global health is being threatened by chronic kidney disease (CKD), a condition that is quiet but serious. Chronic kidney disease (CKD), which is characterized by a progressive decrease of kidney function, frequently goes undetected until it is advanced, increasing morbidity and mortality. Conventional diagnostic techniques mostly depend on intrusive testing and clinical knowledge, both of which are frequently unavailable in environments with low resources.

An opportunity to use machine learning (ML) for a quicker and more accurate diagnosis of chronic kidney disease (CKD) is presented by the growing availability of healthcare data. Through the examination of clinical characteristics like blood pressure, glucose levels, and creatinine, machine learning algorithms are able to spot minor trends that point to chronic kidney disease (CKD) early on.

However, problems like feature relevance and high-dimensional data call for dependable models like Support Vector Machines (SVM), which are excellent at managing complicated datasets and achieving high precision.

3. LITERATURE REVIEW

Recent years have seen a notable increase in interest in the use of machine learning (ML) in the prediction of chronic kidney disease (CKD). The efficiency of ML algorithms in detecting CKD patterns from clinical datasets has been shown in a number of studies. Among these, Naive Bayes, Decision Trees, and Support Vector Machines (SVM) are commonly cited for their robustness and classification accuracy.

Studies show that SVM works especially well with high-dimensional data and offers better accuracy when paired with feature selection methods. For example, research has demonstrated that when employing kernel functions like the Radial Basis Function (RBF), SVM performs better than other classifiers in CKD prediction. Decision trees are a useful option in therapeutic settings because they are also praised for their interpretability. Naive Bayes provides simplicity and computational economy, particularly when working with categorical data, but occasionally being less accurate.

A key factor in improving ML models' performance is feature selection. By determining the most pertinent qualities, methods like subset assessment and recursive feature elimination (RFE) have been demonstrated to increase classification accuracy. Characteristics such as blood pressure, blood sugar, hemoglobin, and creatinine are frequently found to be important predictors of chronic kidney disease.

The literature's comparative analyses show the benefits and drawbacks of different classifiers. Despite SVM's constant high accuracy, its computing demands may be a disadvantage. Naive Bayes, despite its simplicity, may have trouble with complex data linkages, while decision trees provide transparency but may overfit. These revelations highlight how crucial it is to use the right preprocessing methods and algorithms in order to maximize model performance.

By creating an SVM-based classification system for CKD prediction and using feature selection to handle high-dimensionality issues, this study expands on previous research. The results are intended to add to the expanding corpus of research on machine learning applications in healthcare, with an emphasis on enhancing patient outcomes and early diagnosis.

Beyond categorization, a variety of issues can be solved by Support Vector Machines (SVM). Other issues that SVM can resolve include the following:

Regression (SVR - Support Vector Regression): SVM can be applied to regression tasks, which aim to predict a continuous value instead of a class label (SVR, or Support Vector Regression). Finding the hyperplane that best fits the data while accounting for error tolerance is how SVR operates.

Outlier Detection (One-Class SVM): By training on a dataset that only contains normal data, SVM can be used to detect anomalies. Any data point outside of the border that defines "normal" data is regarded as an anomaly or outlier, which the model learns to recognize.

Multiclass Classification: Although SVM is a binary classifier by nature, it may be used to solve multiclass problems by combining several binary classifiers using techniques like "one-vs-one" or "one-vs-all," which classify data into more than two groups.

Sentiment analysis and text classification: SVM is frequently utilized in natural language processing applications such as document classification, sentiment analysis, and spam email detection. As is typical with text data, it works well in high-dimensional settings.

Image Classification: SVM is also used for image recognition tasks, such as recognizing faces, objects, or handwriting, in order to categorize images into several categories.

Bioinformatics: SVM has shown effective in bioinformatics applications involving the analysis of huge and complicated datasets, such as protein structure prediction and gene expression classification.

Time Series Forecasting: SVM can be used for time series forecasting difficulties, especially when the relationship between the time series data points is non-linear, even though it is less popular than other techniques like ARIMA or LSTMs.

4. METHODOLOGY

1. Data Collection:

The UCI Machine Learning Repository's Chronic Kidney Disease dataset is utilized. 400 records with 25 variables make up this dataset, which includes clinical and demographic data like blood pressure, hemoglobin, glucose, and creatinine levels.

Important Features

400 instances, each of which represents a distinct patient.

25 attributes (one target variable plus 24 characteristics).

Class is the target variable; it might be either ckd or notckd.

Numerical Attributes	Categorical Attributes
age	specific gravity (sg)
blood pressure (bp)	albumin (al)
blood glucose random (bgr)	sugar (su)
blood urea (bu)	red blood cells (rbc)
serum creatinine (sc)	pus cell (pc)
sodium (sod)	pus csll clumps (pcc)
potassium (pot)	bacteria (ba)
hemoglobin (hemo)	hypertension (htn)
packed cell volume (pcv)	diabetes mellitus (dm)
white blood cell count (wbcc)	coranary artery disease (cad)
red blood cell count (rbcc)	appetite (appet)
	peda edema (pe)
	anemia (ane)
	class (classification)

The Chronic Kidney Disease dataset used in this work was obtained from publicly available dataset: https://archive.ics.uci.edu/ml/datasets/Chronic_Kidney_Disease.

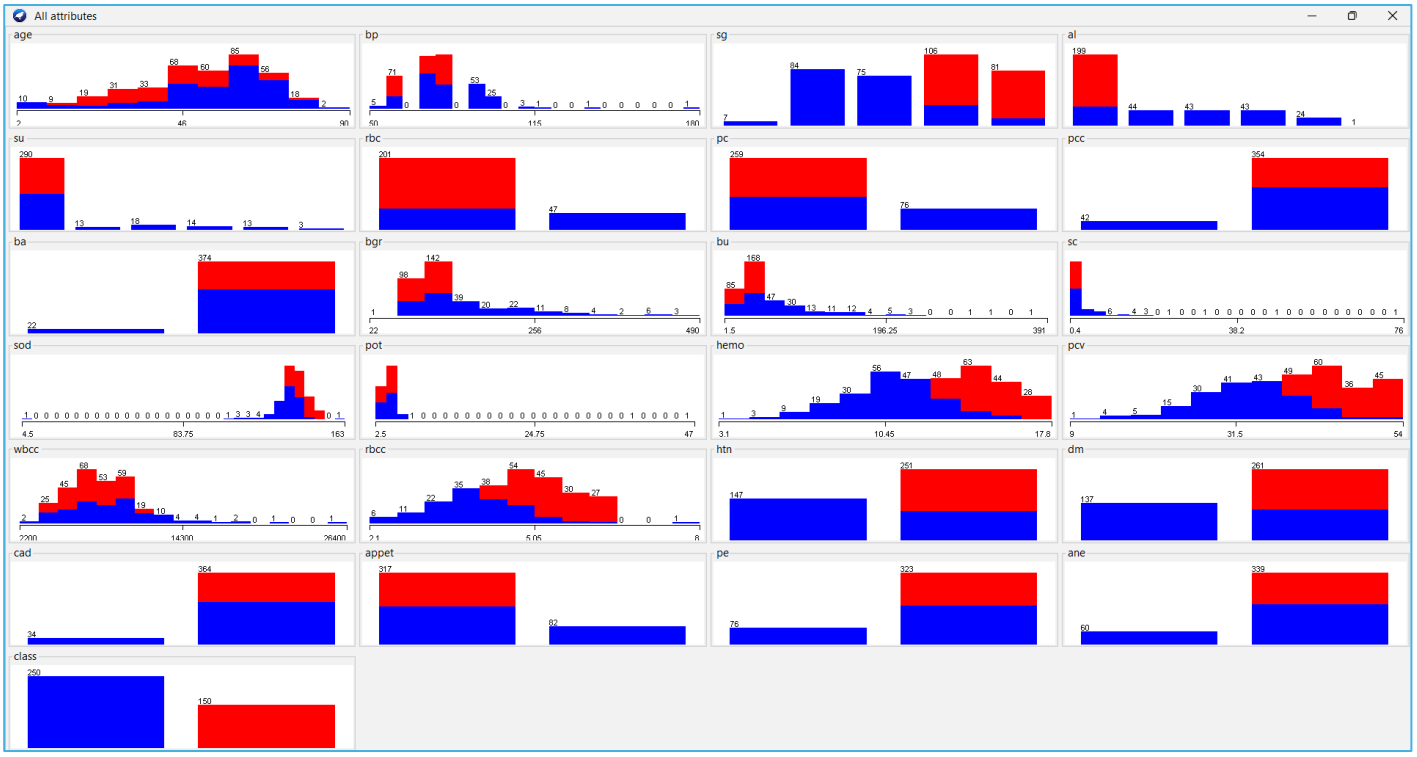

Image 1: Visual representation of the attributes

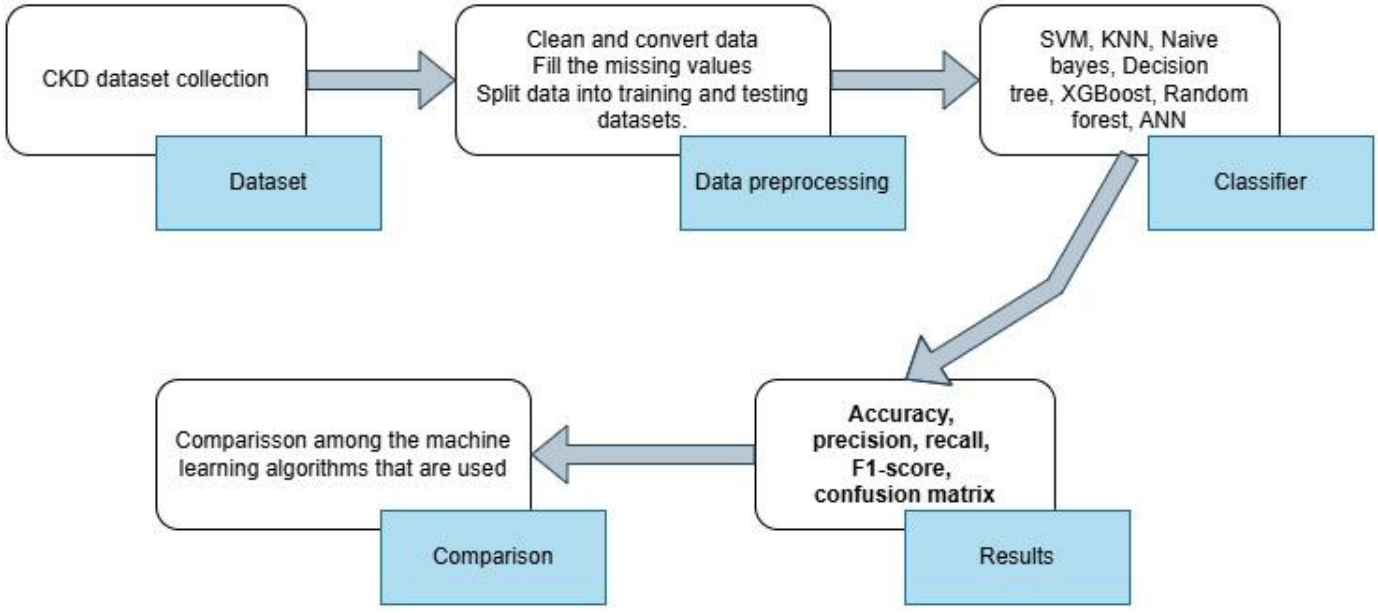

Image 2: Structure of the CKD diagnostic process

2. Data Preprocessing:

Use mode imputation for categorical features and median imputation for numerical attributes to deal with missing values. To guarantee uniformity in model training, normalize numerical features to a predetermined range. To reduce dimensionality and increase model performance, use Recursive Feature Elimination (RFE) to find and keep the most pertinent features.

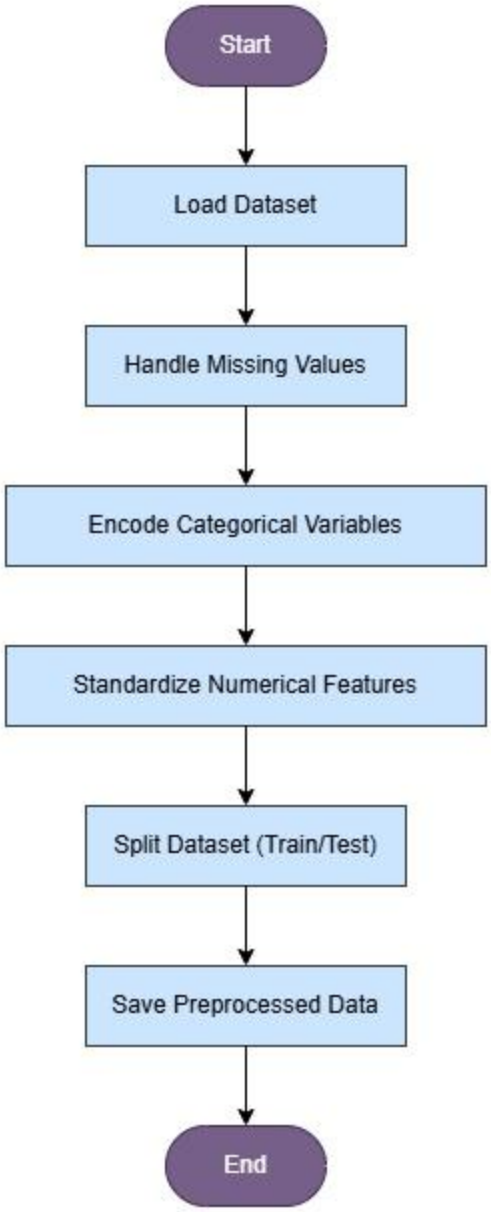

Image 3: Preprocessing Flowchart

1) Loading the dataset

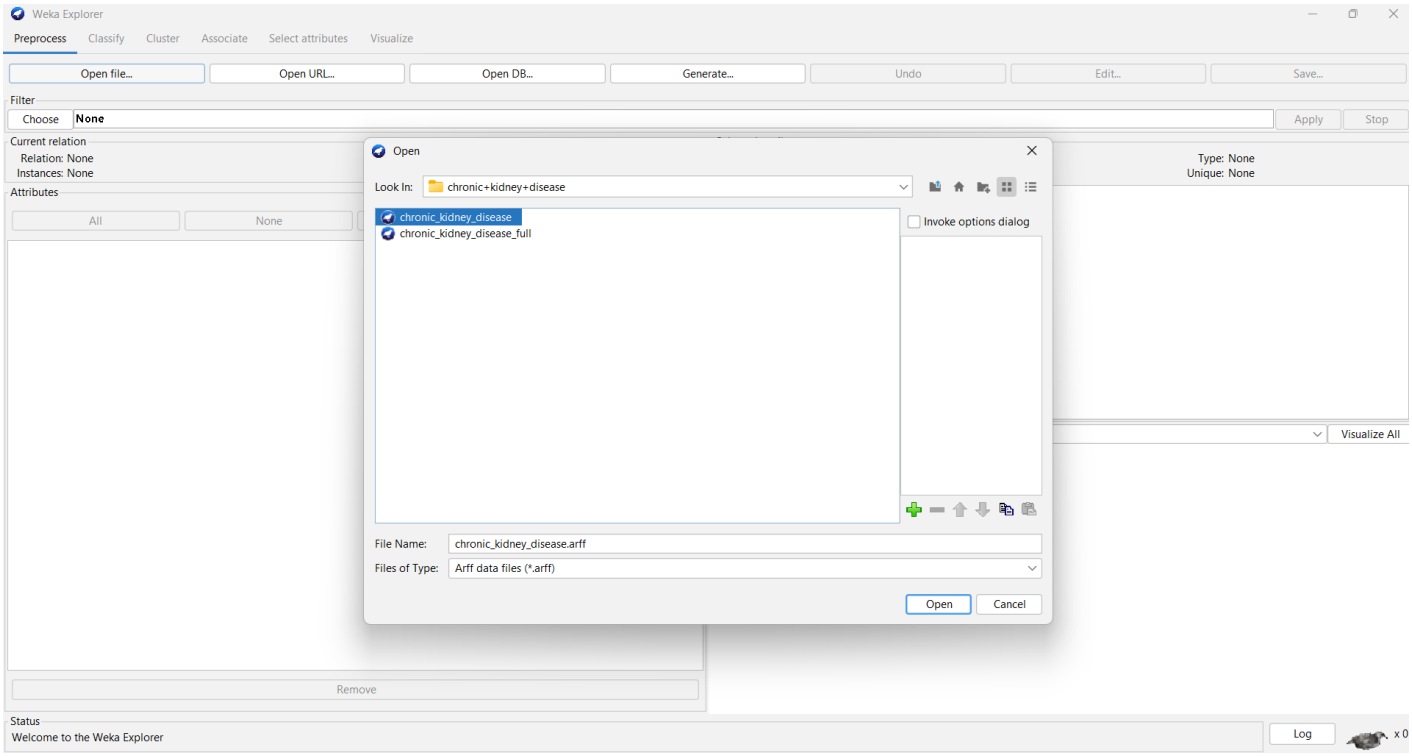

Image 4: Loading the dataset to the Weka tool

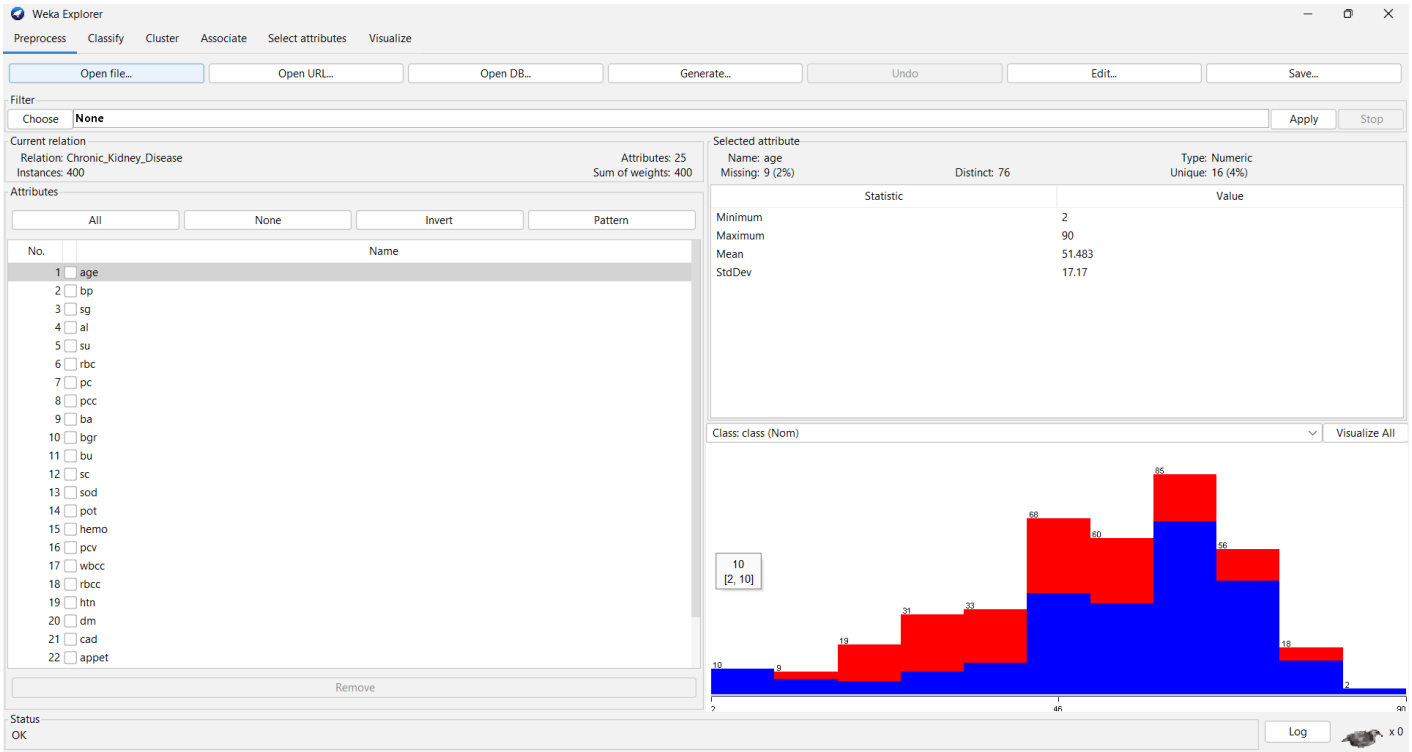

Image 5: View the dataset which includes 25 attributes

2) Handling missing values

Image 5 itself show that there are 9 missing values which is 2%. So, we have to handle it first to make sure there are no missing values.

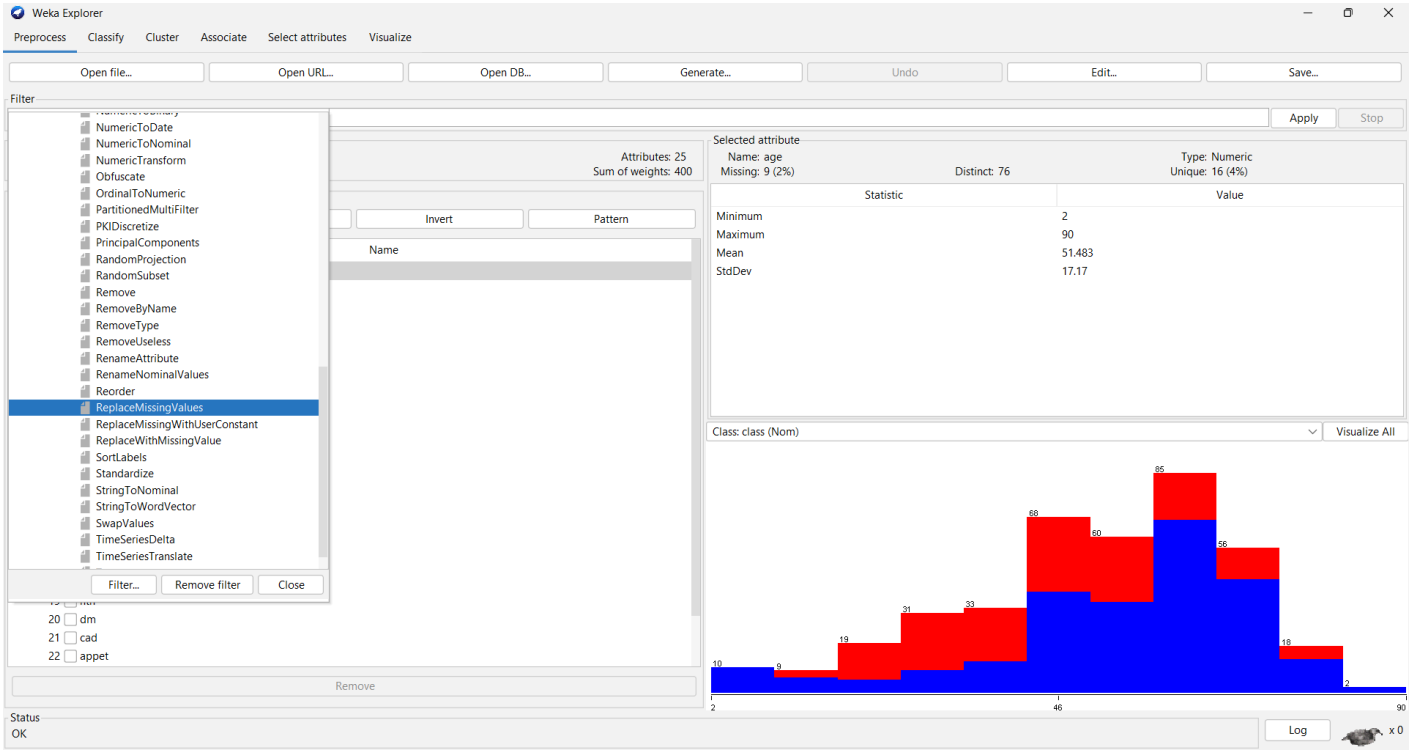

Image 6: weka→filters→Unsupervised→attribute→ReplaceMisssingValues

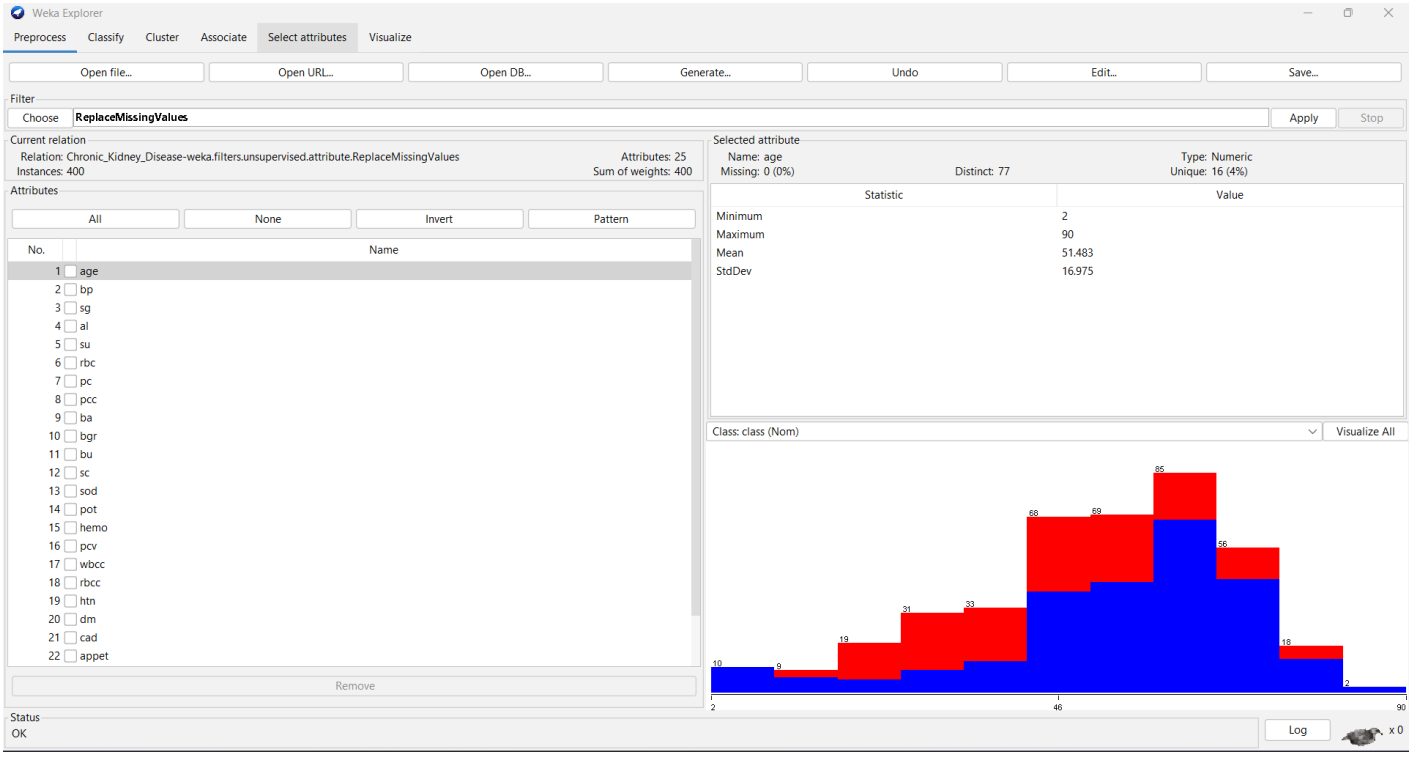

Image 7: After clicking apply we can see missing value is 0

3. Encoding categorical variables

In order to convert categorical variables into a numerical format that machine learning algorithms can comprehend and interpret, we encode the data. Non-numerical (categorical) data cannot be directly used with the majority of machine learning models, particularly those that rely on mathematical calculations like SVM, Logistic Regression, and Neural Networks.

Benefits of Discretization

- Makes Numerical Data Simpler

Transforms continuous data into useful classifications, such as high, medium, and low.

- Enhances Interpretability

Ranges are easier to understand than absolute numbers.

- Decreases Sensitivity

Aids in lessening sensitivity to minute variations in numerical data.

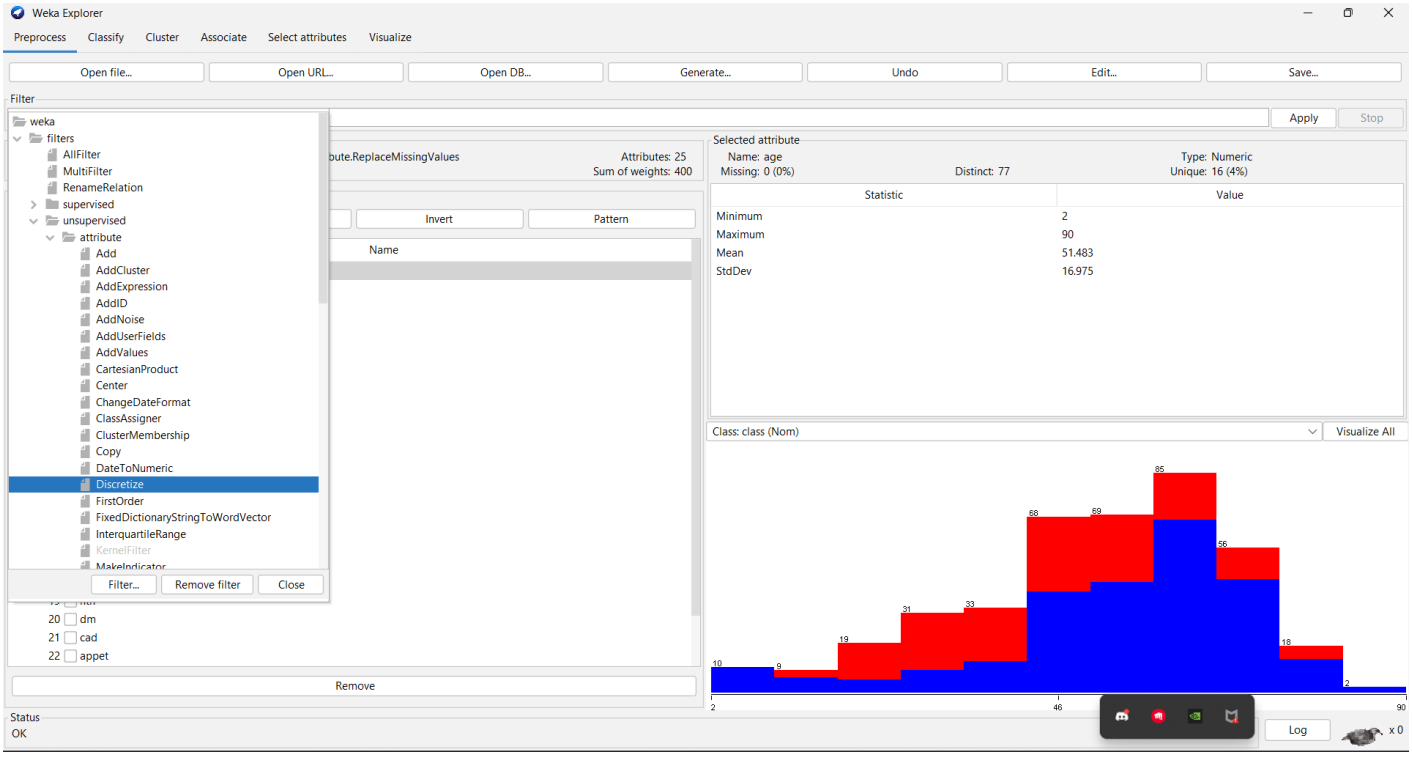

Image 8: weka→filters→Unsupervised→attribute→Descritize

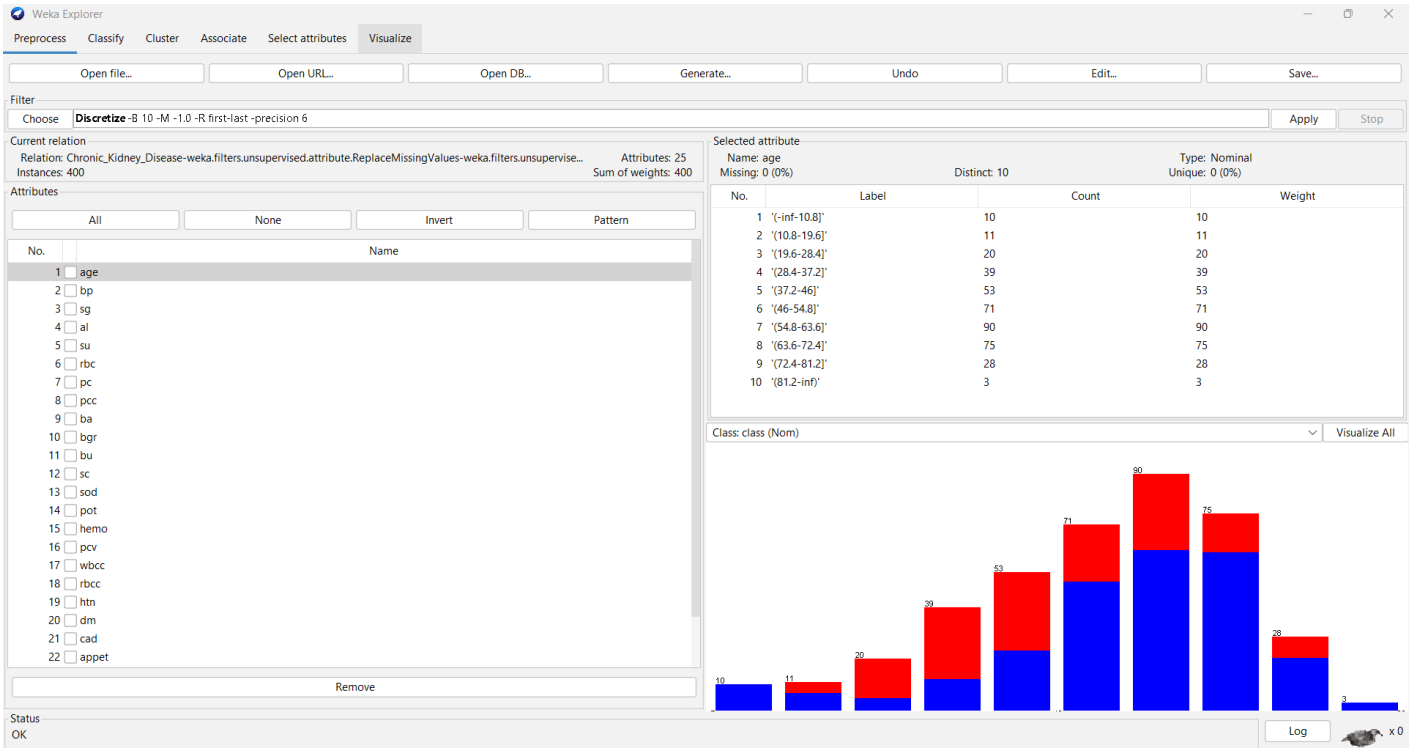

Image 9: After clicking apply we can see the output

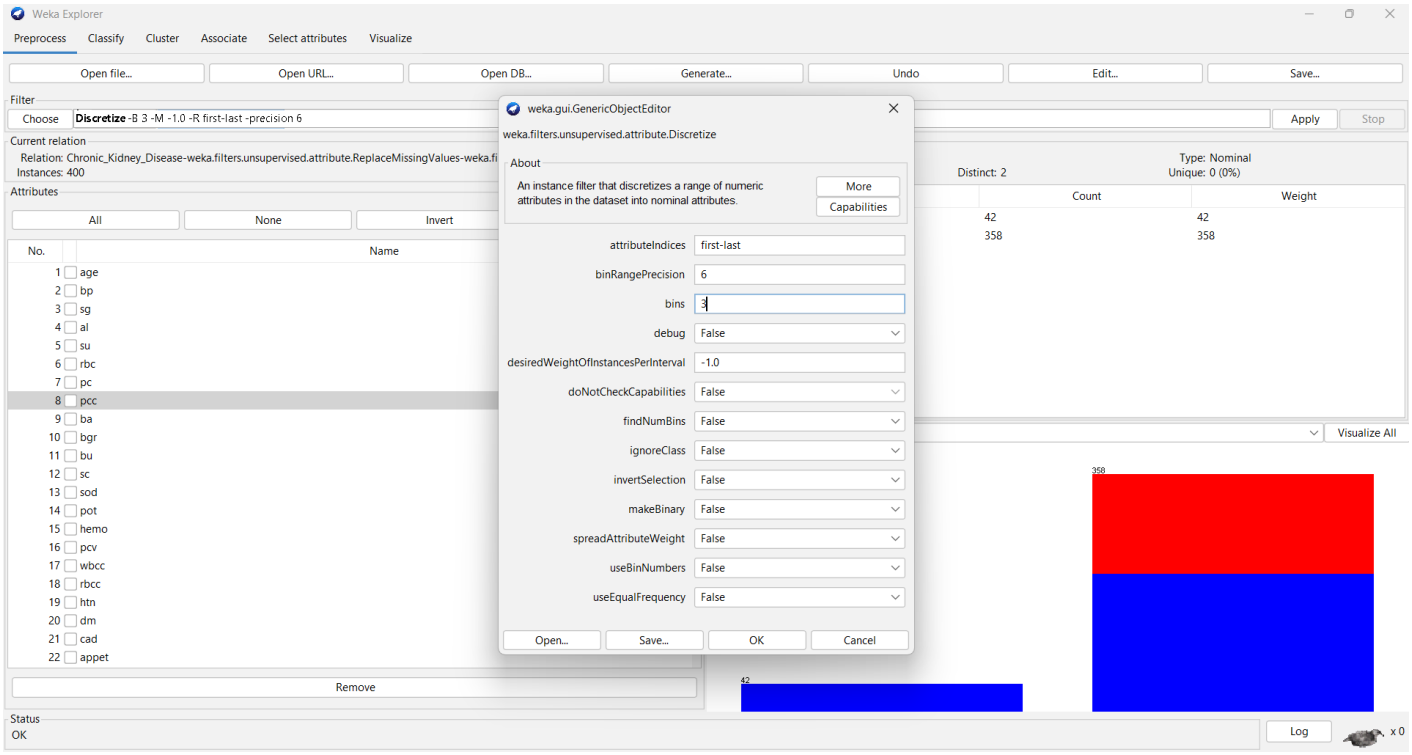

Image 10: Set bin to 3 (Depending on the configuration, each numerical attribute will be separated into three equal-width or equal-frequency bands)

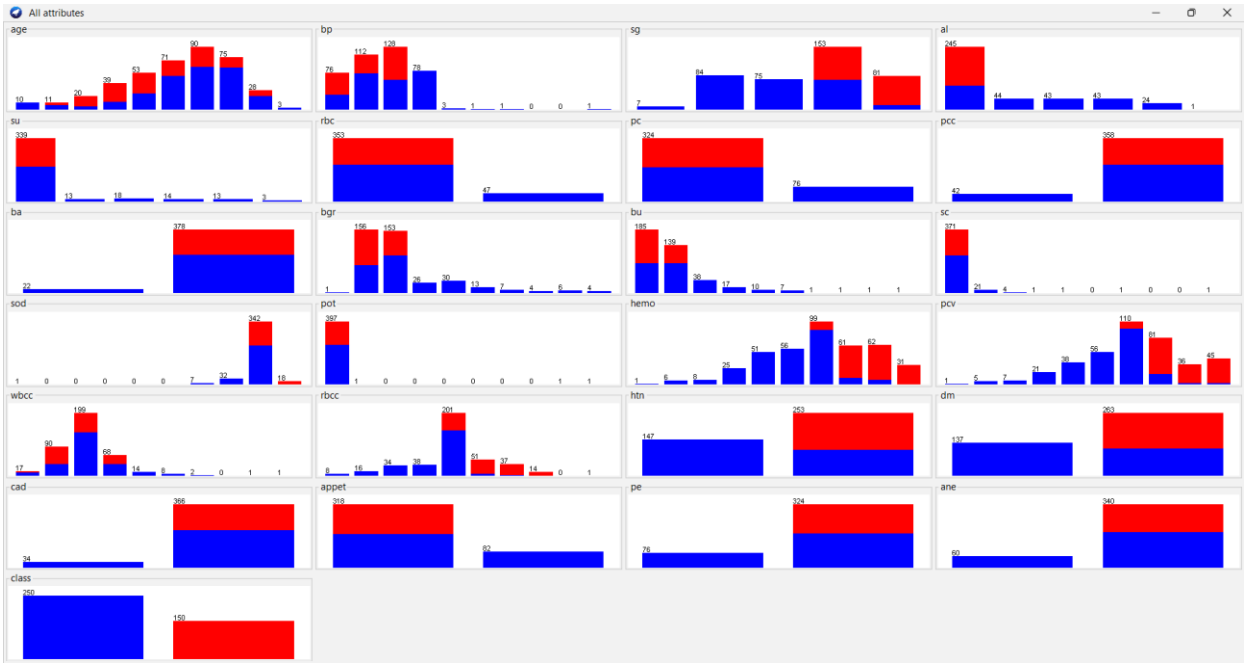

Image 11: Click Visualize All tab to see how the bins are distributed across the dataset

Run the Apriori test

Apriori can be used to better comprehend the dataset and investigate attribute correlations prior to classification. It can assist in locating important characteristic combinations that may have an impact on classification. When used alone, apriori can be helpful in identifying trends in transactional or categorical datasets.

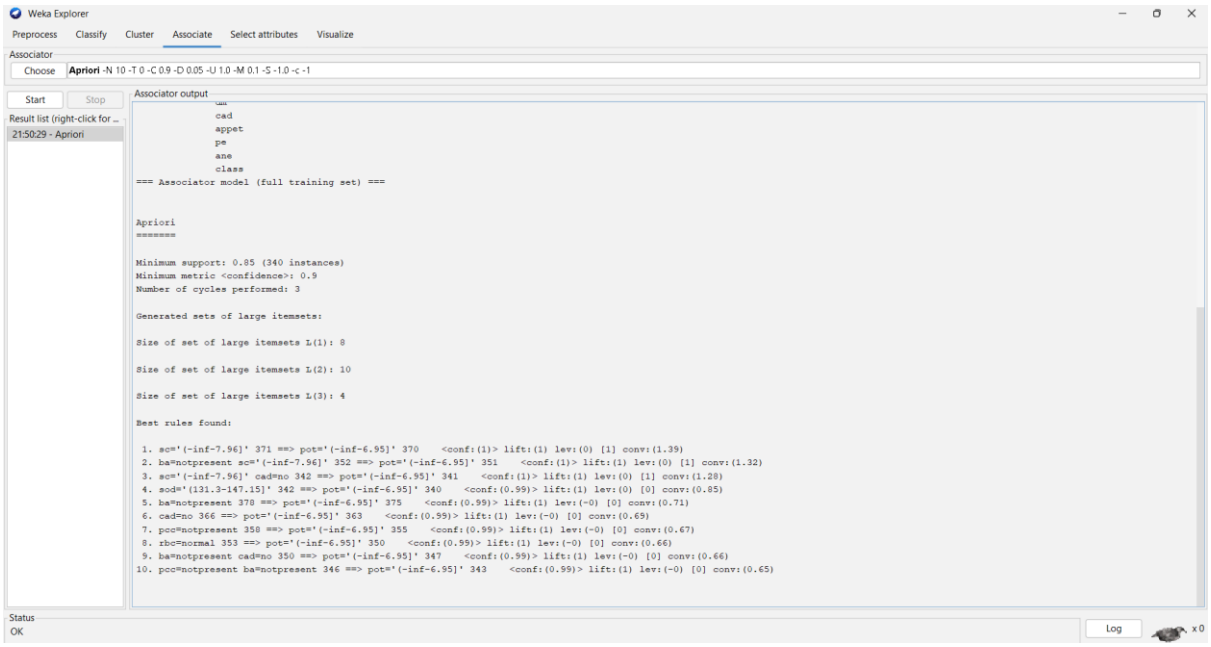

Image 12: Go to associate tab and select apriori and run

5. DESIGN

The following elements are included in the classification system's design:

Pipeline for Data:

CKD dataset as input.

Preprocessing includes feature selection, normalization, and imputation of missing values.

Framework for Evaluation:

For model validation, use cross-validation.

comparison to alternative classifiers.

Model Structure:

RBF-based SVM classifier.

including feature selection to improve efficiency.

Visualization :

performance measures shown graphically, such as confusion matrices and ROC curves.

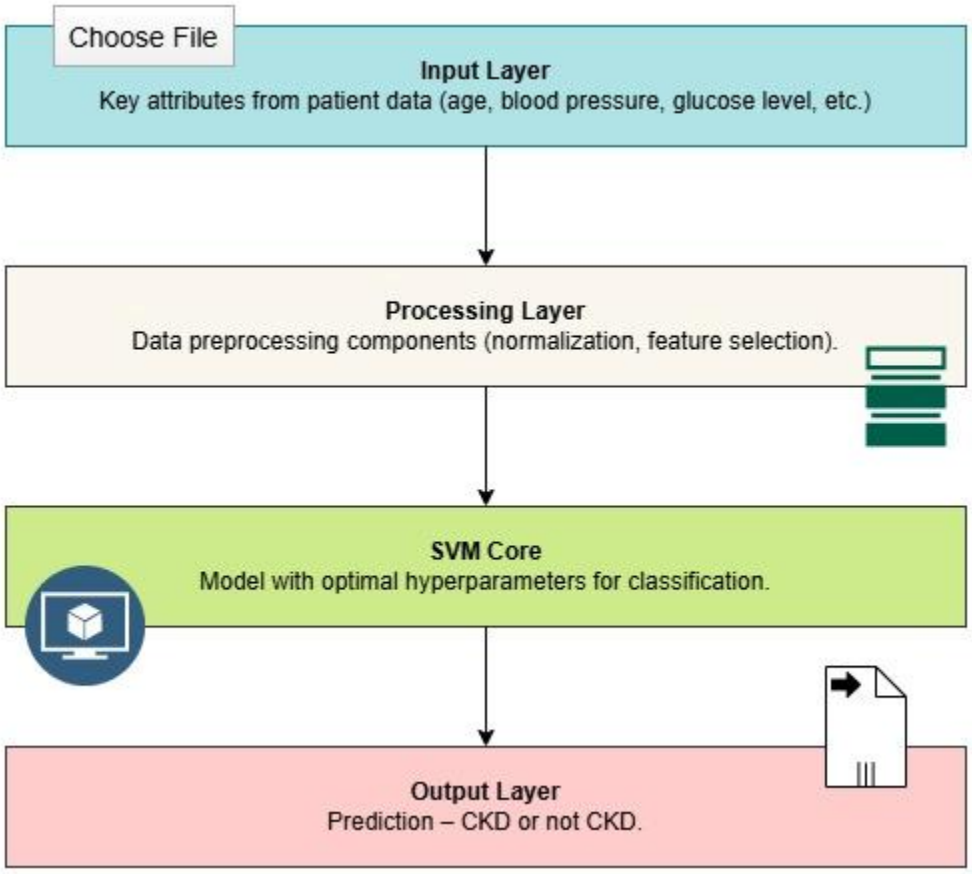

Image 12: SVM Classifier Architecture

6. DEVELOPMENT

The actions consist of: CKD dataset loading into WEKA. applying filters for feature selection and normalization as part of the preprocessing step. Using optimal parameters, train the SVM classifier. creating analysis-ready visualizations and applying validation procedures to assess the model's performance.

3) Model Creation

To deal with non-linear correlations in the data, train an SVM classifier with the Radial Basis Function (RBF) kernel. Optimize parameters like the kernel coefficient (gamma) and regularization term (C) by performing hyperparameter tuning using grid search.

4) Model Validation

Use 10-fold cross-validation to assess how well the model performs when applied to unknown data. To evaluate relative performance, compare the SVM classifier's output with that of other algorithms, such as Decision Trees and Naive Bayes.

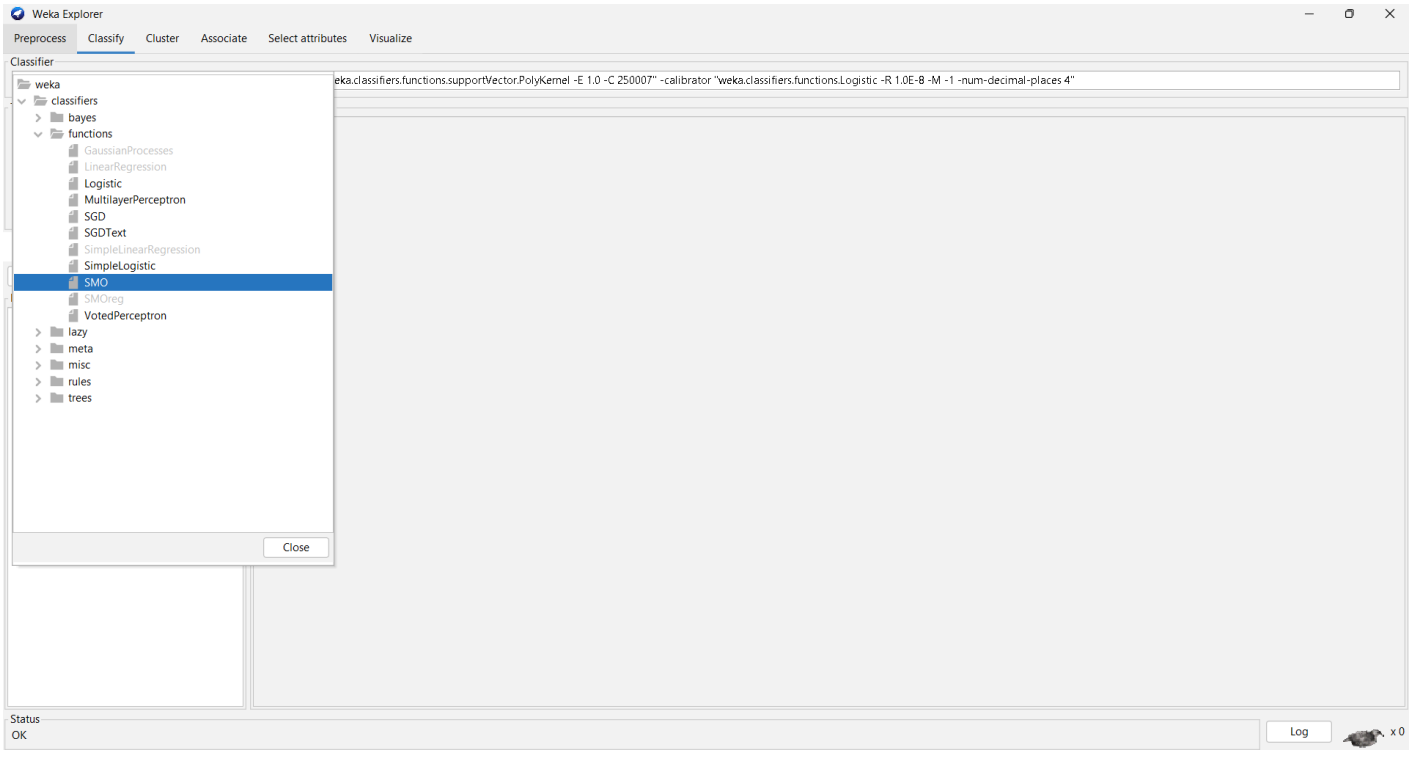

Image 13: Go to classify tab and select SMO to run the SVM classifier

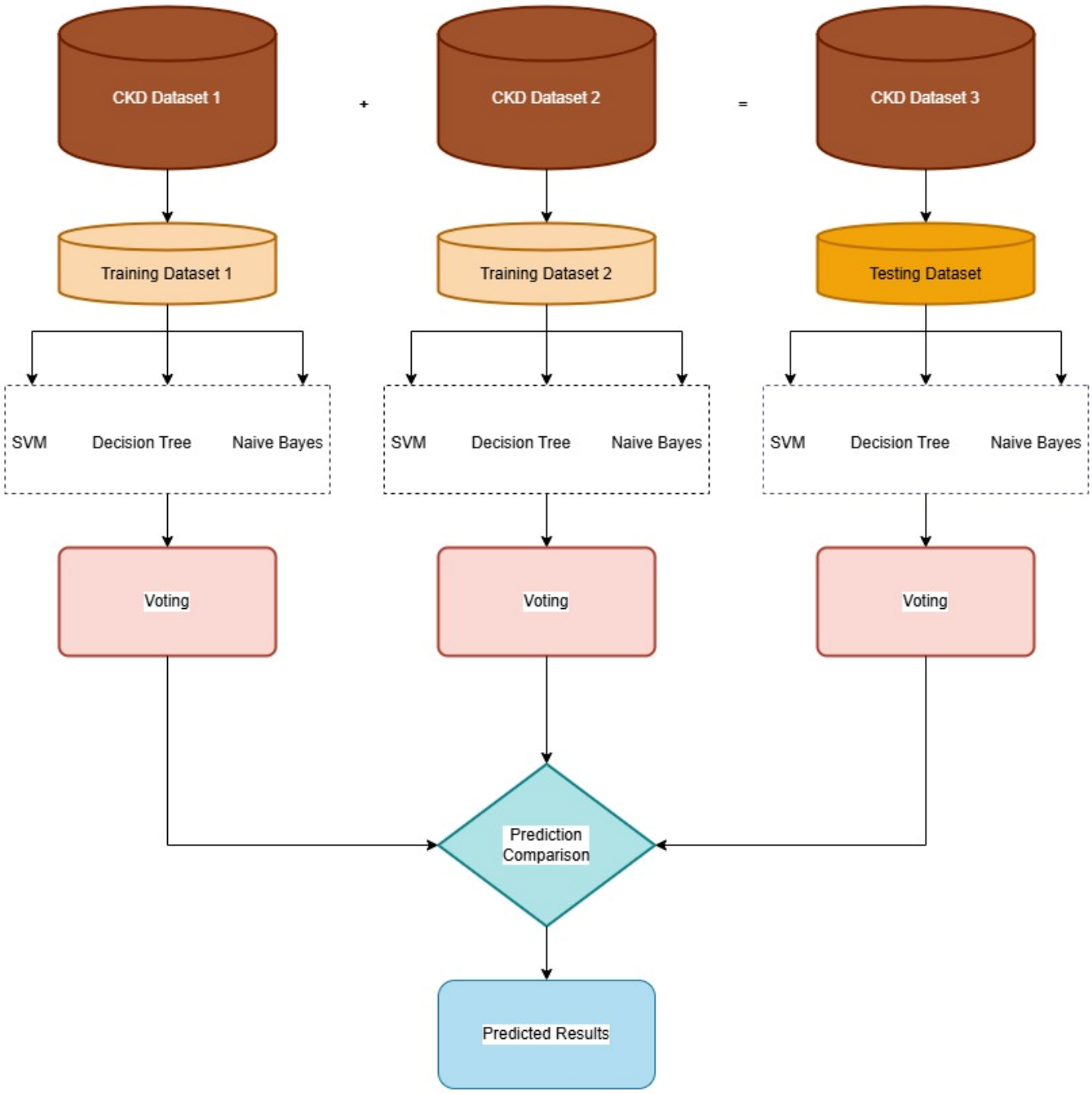

Image 14 : Flowchart of the development process

Theory of the classifications that have been used

Support Vector Machine (SVM)

The Support Vector Machine (SVM) is a supervised learning technique that is employed for problems involving regression and classification. It operates by locating the hyperplane in a high-dimensional space that best divides data points of several kinds. The hyperplane is selected to optimize the distance between each class's nearest data points, also known as support vectors.

Decision Tree

A decision tree is a structure that resembles a flowchart, with each leaf node denoting a class label or value and each inside node representing a choice based on a feature. Until it reaches a stopping requirement, it iteratively divides the data according to the feature that offers the best split (using metrics like entropy or Gini impurity).

Naïve Bayes

Based on the Bayes theorem, Naive Bayes is a probabilistic classifier that makes the assumption that features

Based on Bayes' theorem, the Naive Bayes classifier is a probabilistic algorithm that assumes that characteristics are conditionally independent given the class label. Given the feature values, it determines the likelihood of each class and designates the class with the highest probability. It works especially well for jobs involving text classification.

Random Forest

Using random selections of the data and features, Random Forest is an ensemble learning technique that creates several decision trees. Every tree produces a prediction on its own, and the total of all the trees' predictions is used to make the final prediction (for example, by majority vote for classification). Compared to a single decision tree, it increases accuracy and lessens overfitting.

7. TEST RESULT

5) Performance metrices

Use metrics like accuracy, precision, recall, F1 score, and ROC-AUC to assess the model. To determine the classifier's advantages and disadvantages in predicting CKD-positive and CKD-negative instances, examine the confusion matrix.

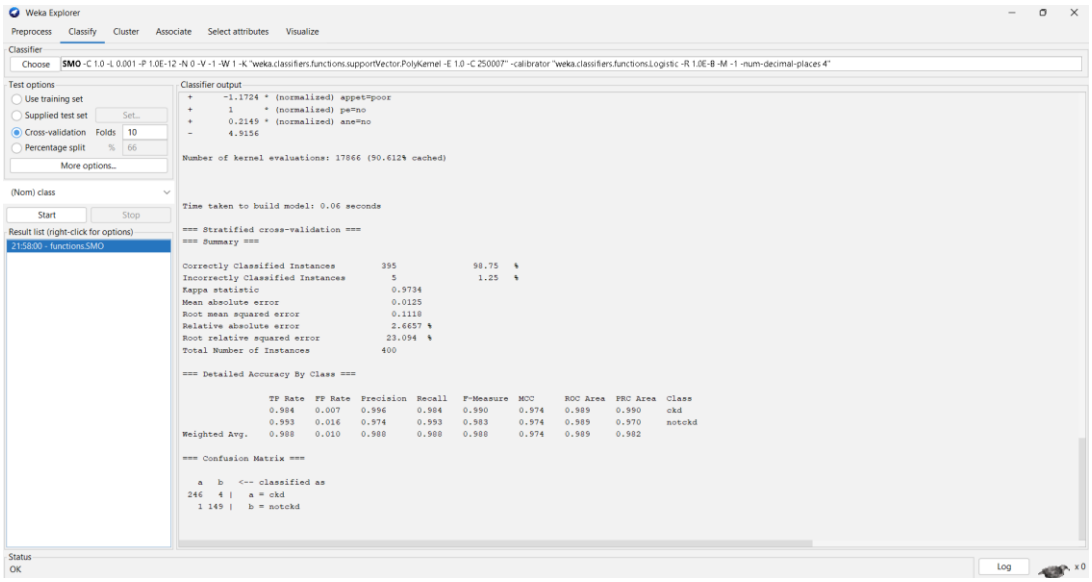

Image 15: Results of SVM classifier

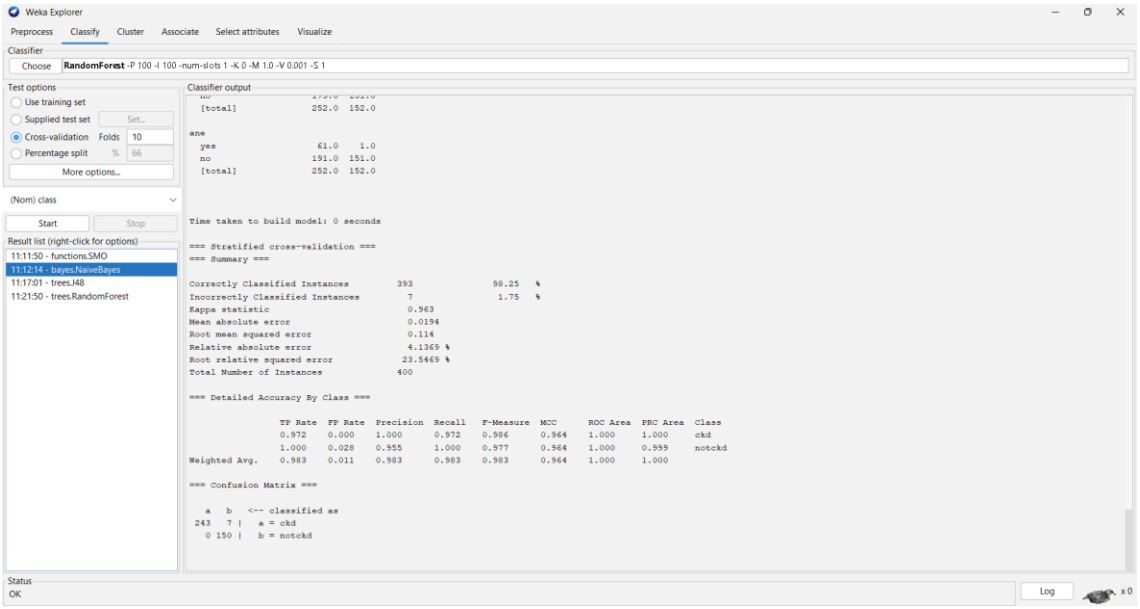

Image 16: Results of Naïve Bayes

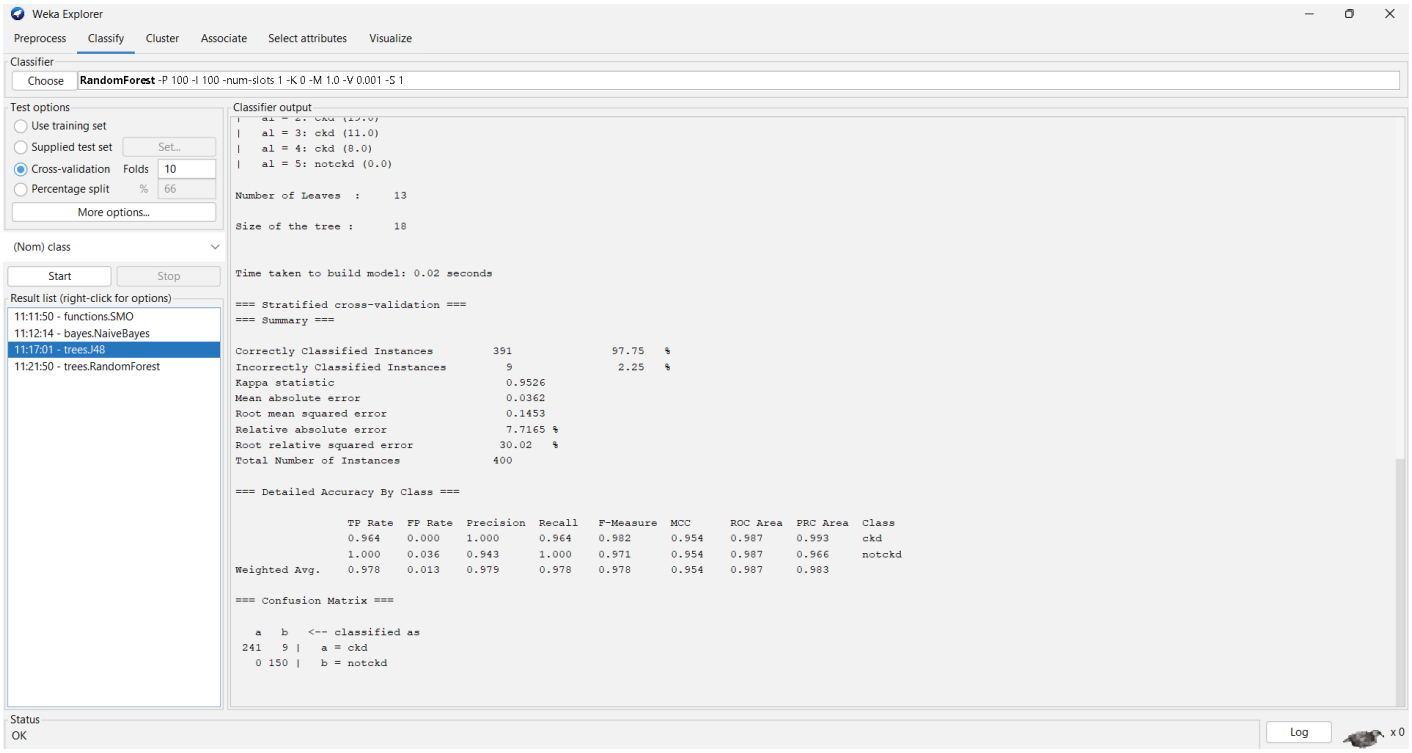

Image 17: Results of Decision Tree

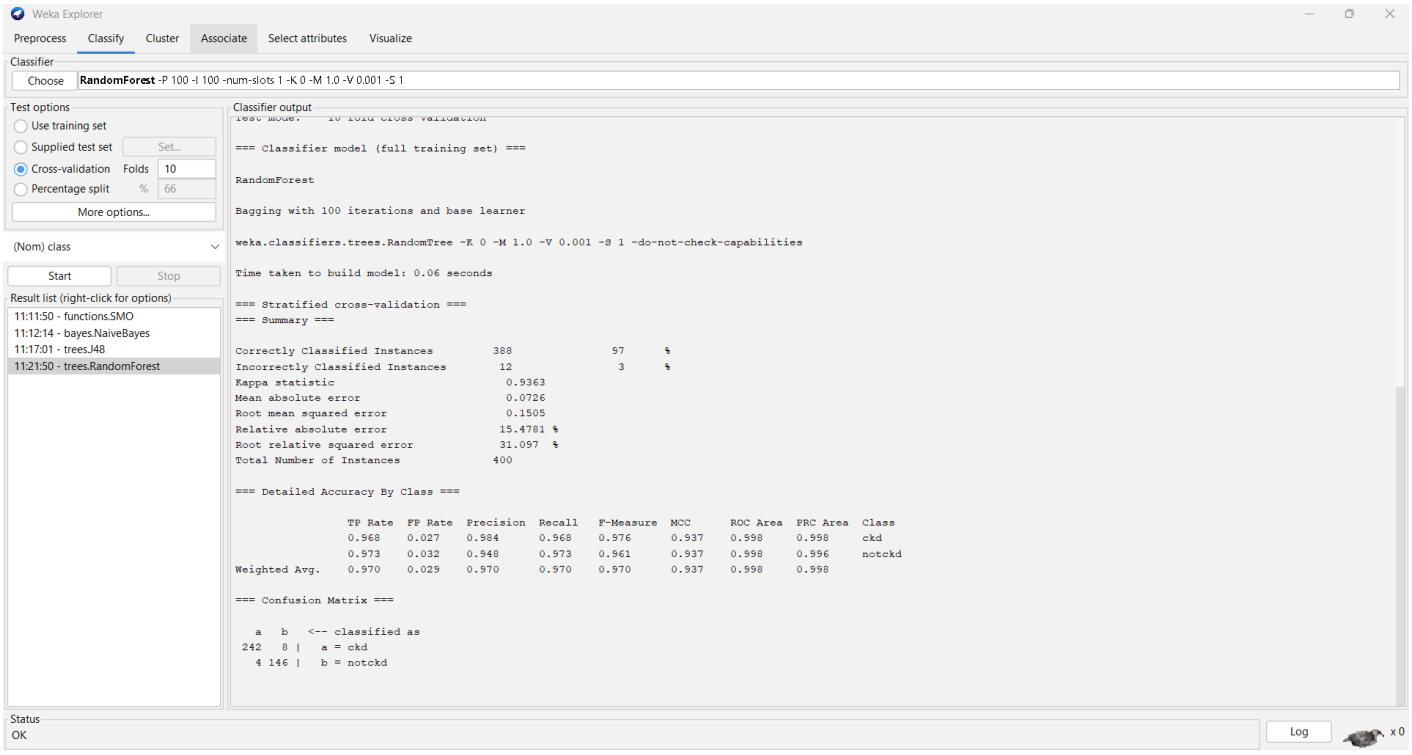

Image 18: Results of Random Forest Tree

Now let's summarize the results in a table for further understanding. It will help us in the analysis of the results under different classification techniques.

Classifier: SMO (Support Vector Machine)

Model Performance Metrics

Metric	Value
Correctly Classified Instances	98.75 %
Incorrectly Classified Instances	1.25 %
Kappa Statistic	0.9734
Mean Absolute Error (MAE)	0.0125
Root Mean Squared Error (RMSE)	0.1118
Relative Absolute Error (RAE)	2.6657 %
Root Relative Squared Error (RRSE)	23.094 %

Detailed Class Performance

Class	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
Ckd	0.984	0.007	0.996	0.984	0.990	0.989
notckd	0.993	0.016	0.974	0.993	0.983	0.989

Confusion Matrix

	Predicted: ckd	Predicted: notckd
Actual: ckd	**246**	**4**
Actual: notckd	**1**	**149**

Follow the same steps, use the required classification technique, and get the relevant test results for each classification separately is provided below.

Classifier: Decision Tree

Model Performance Metrics

Metric	Value
Correctly Classified Instances	97.75 %
Incorrectly Classified Instances	2.25 %
Kappa Statistic	0.9526
Mean Absolute Error (MAE)	0.0362
Root Mean Squared Error (RMSE)	0.1453
Relative Absolute Error (RAE)	7.7165 %
Root Relative Squared Error (RRSE)	30.02 %

Detailed Class Performance

Class	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
Ckd	0.964	0.000	1.000	0.964	0.982	0.987
notckd	1.000	0.036	0.943	1.000	0.971	0.987

Confusion Matrix

	Predicted: ckd	Predicted: notckd
Actual: ckd	**241**	**9**
Actual: notckd	**0**	**150**

Classifier: Naïve Bayes

Model Performance Metrics

Metric	Value
Correctly Classified Instances	98.25 %
Incorrectly Classified Instances	1.75 %
Kappa Statistic	0.963
Mean Absolute Error (MAE)	0.0194
Root Mean Squared Error (RMSE)	0.114
Relative Absolute Error (RAE)	4.1369 %
Root Relative Squared Error (RRSE)	23.5469 %

Detailed Class Performance

Class	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
Ckd	0.972	0.000	1.000	0.972	0.986	1.000
notckd	0.983	0.028	0.955	1.000	0.977	1.000

Confusion Matrix

	Predicted: ckd	Predicted: notckd
Actual: ckd	**243**	**7**
Actual: notckd	**0**	**150**

Classifier: Random Forest

Model Performance Metrics

Metric	Value
Correctly Classified Instances	97 %
Incorrectly Classified Instances	3 %
Kappa Statistic	0.9363
Mean Absolute Error (MAE)	0.0726
Root Mean Squared Error (RMSE)	0.1505
Relative Absolute Error (RAE)	15.4781 %
Root Relative Squared Error (RRSE)	31.097 %

Detailed Class Performance

Class	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
Ckd	0.968	0.027	0.984	0.968	0.976	0.998
notckd	0.970	0.032	0.948	0.973	0.961	0.998

Confusion Matrix

	Predicted: ckd	Predicted: notckd
Actual: ckd	**242**	**8**
Actual: notckd	**4**	**146**

8. DISCUSSION

Accuracy, precision, recall, F-measure, and ROC area were among the metrics used to assess the classifiers' effectiveness in identifying Chronic Kidney Disease (CKD). A thorough analysis of the findings is provided below:

Support Vector Machine (SVM)

Performance:

Among all classifiers, SVM had the greatest accuracy (98.75%), demonstrating its resilience in differentiating between cases of CKD and non-CKD.

Its capacity to reduce prediction mistakes was demonstrated by the lowest Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE), which were 0.0125 and 0.1115, respectively.

Class Performance:

CKD outperformed non-CKD by a small margin, although both classes had good precision, recall, and F-measure values.

With a ROC area of 0.989 for both classes, the discrimination ability was outstanding.

Confusion Matrix:

SVM's dependability for this dataset is demonstrated by the just 5 misclassifications (4 CKD and 1 non-CKD).

Conclusion:

With its superior precision and error reduction, SVM is the most dependable and accurate classifier for CKD detection in this investigation.

Decision Tree

Performance:

The Decision Tree's accuracy was 97.75%, which was little less than SVM's.

In comparison to SVM, the MAE and RMSE were greater at 0.0362 and 0.1453, respectively, suggesting a higher error rate.

Moderate prediction errors were indicated by the Relative Absolute Error (RAE), which was 7.116%.

Class Performance:

Although the F-measure (0.982) was lower than SVM, the precision for CKD was 1.000, indicating a trade-off in performance measurements.

The precision of the non-CKD class was lower, at 0.943.

Confusion Matrix:

Nine cases with CKD were incorrectly classified as non-CKD, which may have an effect on judgement in urgent medical situations.

Conclusion:

The Decision Tree performs well, although it is less dependable than SVM due to its higher error rates and misclassification of CKD cases.

Naïve Bayes

Performance:

With an accuracy of 98.25%, Naïve Bayes was comparable to SVM but marginally less accurate.

The low RMSE (0.1174) and MAE (0.0194) showed good error minimization.

Although greater than SVM, the RAE (4.113%) was still respectable.

Class Performance:

Both the CKD and non-CKD classes exhibited good ROC areas (1.000 for both classes) and high precision (1.000 and 0.955, respectively).

For CKD, the F-measure was 0.986, which was marginally lower than SVM.

Confusion Matrix:

Seven CKD cases were incorrectly identified in the confusion matrix; this was marginally higher than SVM but better than the decision tree.

Conclusion:

Naïve Bayes is a formidable competitor with competitive precision and accuracy; yet it is less optimal than SVM due to its higher rate of misclassification.

Random Forest

Performance:

Of all the classifiers, Random Forest had the lowest accuracy, at 97%.

Acceptable prediction errors were indicated by the moderate MAE (0.01726) and RMSE (0.1505).

There was potential for improvement in error reduction, as seen by the much higher RAE (13.4781%).

Class Performance:

CKD had a precision of 0.984, which was marginally less than that of SVM and Naïve Bayes.

Although the ROC area (0.998) remained high, indicating significant discrimination ability, the F-measure for both classes was below SVM and Naïve Bayes.

Confusion Matrix:

12 incorrect classifications (8 CKD and 4 non-CKD), the most of any classifier.

Conclusion:

Random Forest is a strong ensemble approach; however, it is not as appropriate for this dataset as SVM because of its greater misclassification rate and comparatively lower accuracy.

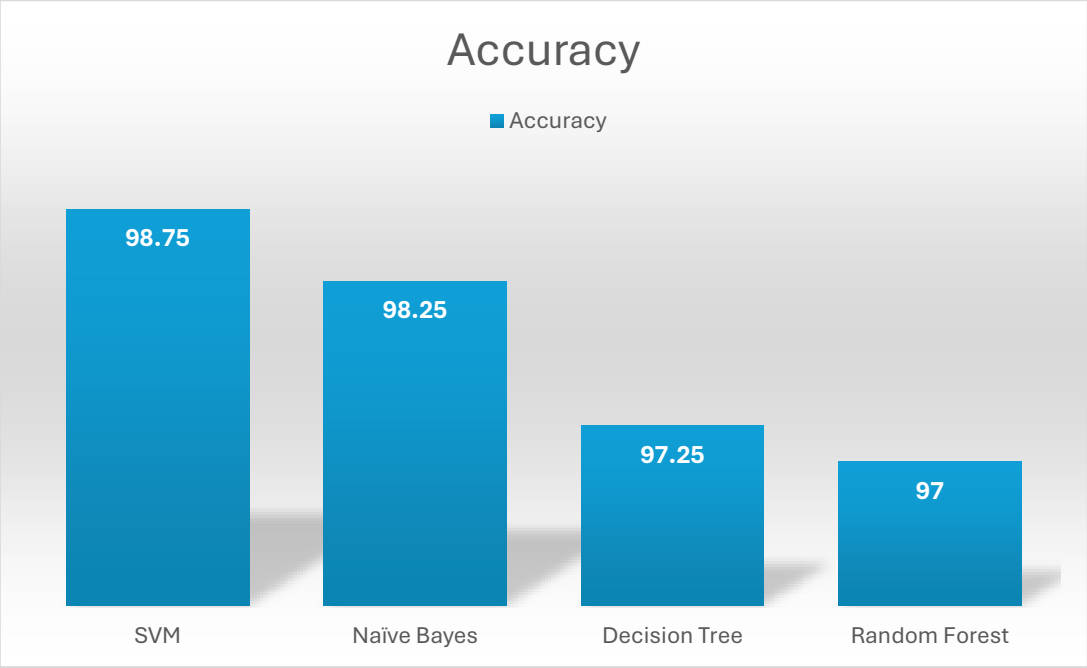

Chart 1: Accuracy Evaluation of all Classifiers

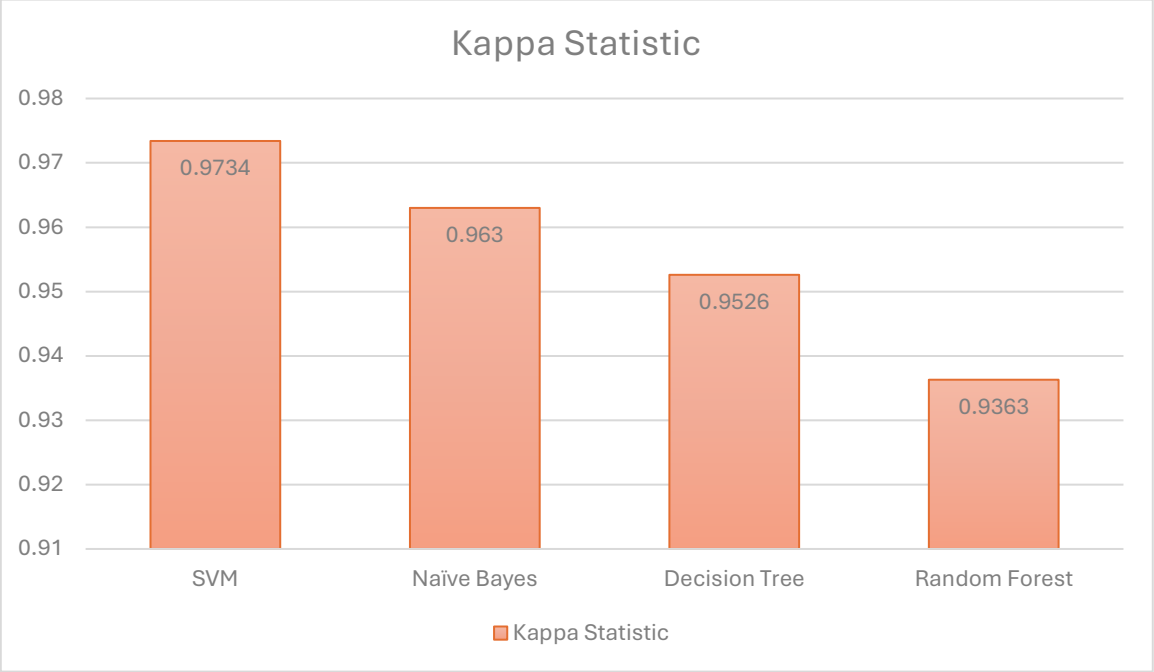

Chart 2: Kappa Value of all Classifiers

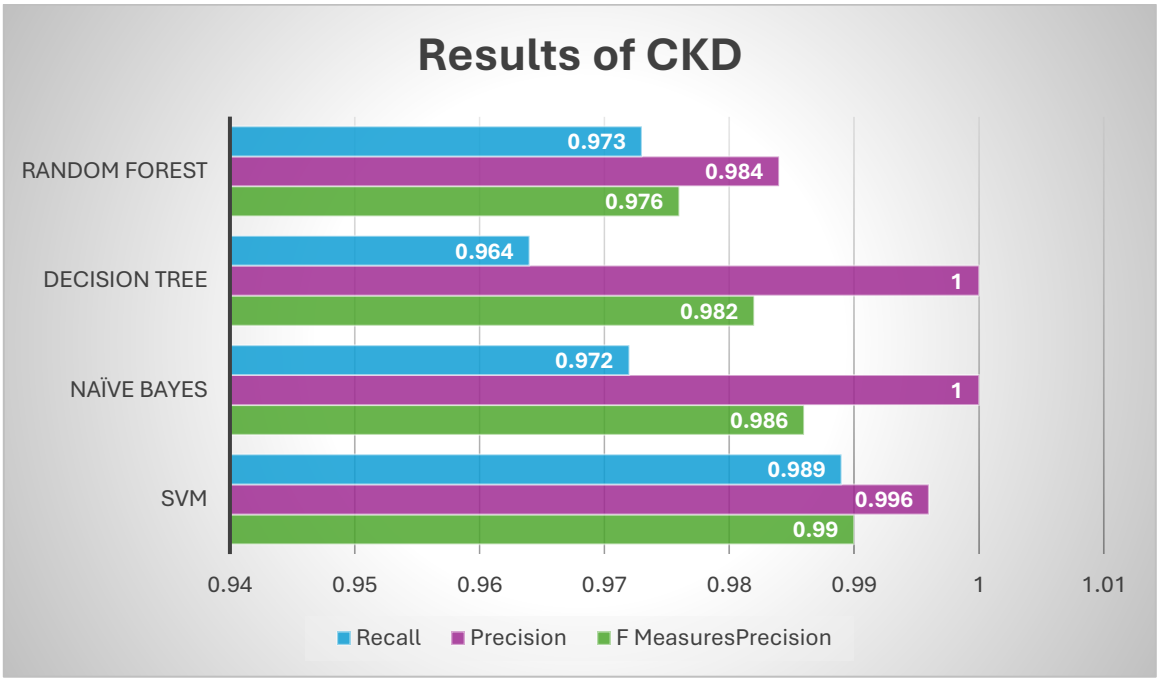

Chart 3: F-measure, precision, and Recall analysis representation of CKD

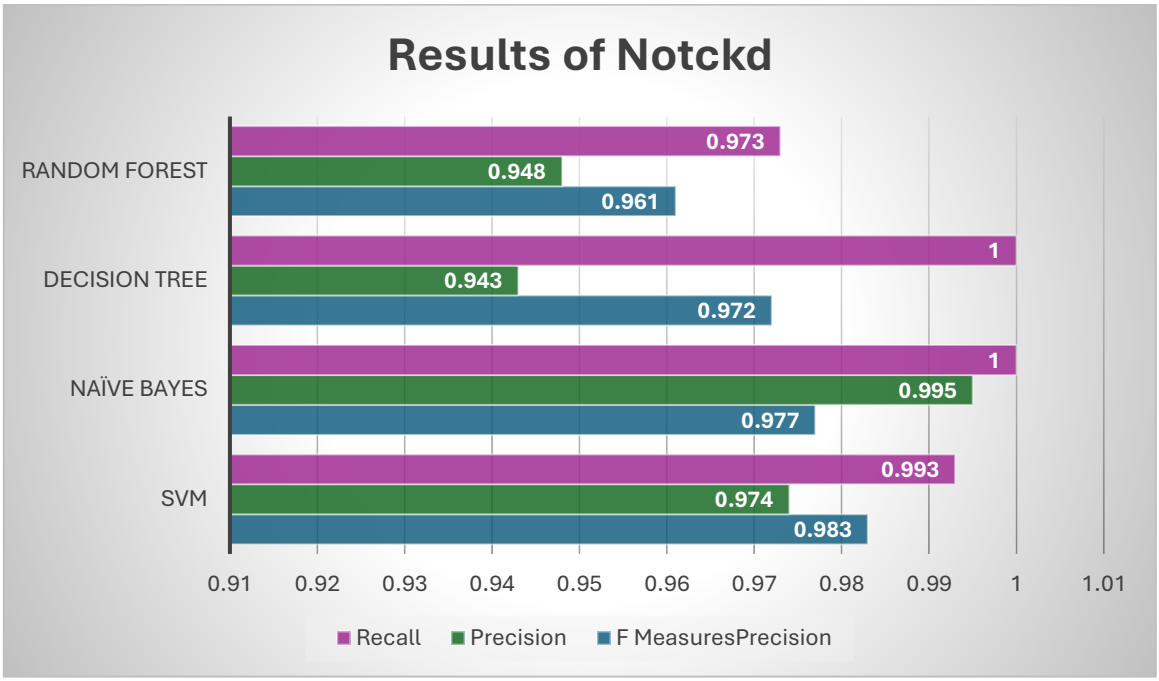

Chart 4: F-measure, precision, and Recall analysis representation of non-CKD

Overall Suggestion

With the highest accuracy, lowest error rates, and superior precision and recall for both classes, SVM is the best-performing classifier for CKD detection, according to the evaluation. A close second is Naïve Bayes, which offers competitive performance while having a little higher misclassification rate. Despite their effectiveness, Decision Tree and Random Forest are less dependable because of their greater error rates and tendency to misclassify cases of CKD.

SVM is the suggested option for real-world applications where precision is crucial, such medical diagnosis. On the other hand, Naïve Bayes might be a quicker option if processing power is scarce.

9. REFERENCES

Tutorialspoint. (n.d.). WEKA tutorial. Available at: https://www.tutorialspoint.com/weka/index.htm [Accessed 20 Dec. 2024].

Author(s). (Year). *Title of the article*. *Journal Name*, Volume(Issue), Page numbers. Available at: https://pmc.ncbi.nlm.nih.gov/articles/PMC9874070/ [Accessed 20 Dec. 2024].

Author(s). (Year). *Title of the article*. *Journal Name*, Volume(Issue), Page numbers. Available at: https://www.sciencedirect.com/science/article/pii/S2352914821001210 [Accessed 20 Dec. 2024].

Ali, H., Iftikhar, A., Farooq, M.U. and Mahmood, K., 2022. Machine learning techniques for chronic kidney disease risk prediction. *ResearchGate*. Available at: https://www.researchgate.net/publication/363541655_Machine_Learning_Techniques_for_Chronic_Kidney_Disease_Risk_Prediction [Accessed 20 December 2024].

Shailaja, K., Seetharamulu, B. and Jabbar, M.A., 2020. An empirical evaluation of machine learning techniques for chronic kidney disease prophecy. *ResearchGate*. Available at: https://www.researchgate.net/publication/339947166_An_Empirical_Evaluation_of_Machine_Learning_Techniques_for_Chronic_Kidney_Disease_Prophecy [Accessed 20 December 2024].

The article titled "Predicting the Risk of Chronic Kidney Disease (CKD) Using Machine Learning Algorithm" was published in *Applied Sciences* in January 2021. MDPI