ISSN No:-2456-2165

Early Diabetic Risk Prediction using Machine Learning Classification Techniques

Adetunji Olusogo Julius[1], Ayeni Olusola Ayokunle[2], Fasanya Olawale Ibrahim[3]
[1,3] Department of Computer Engineering, Ajayi Polytechnic Ikere- Ekiti, Ekiti State, Nigeria.
[2] Department of Computer Science, Federal Polytechnic Ado Ekiti, Ekiti State, Nigeria.

Abstract:- Diabetes is a metabolic disorder that results from deficiency of the insulin secretion to control high sugar contents in the body system. At early stage, diabetes can be managed and controlled. Prolong diabetes leads to complication disorders such as diabetes retinopathy, angina, heart attack, stroke, atherosclerosis and even death. Therefore, assessment of diabetic risk prediction is necessary at early stage by using machine learning classification techniques based on observed sample features. The dataset used for this paper was obtained from Irvine (UCI) repository of machine learning databases and was analyzed on WEKA application platform. The dataset contains 520 samples with 17 distinct attributes. Machine learning algorithms used as classifier are K-Nearest Neighbors algorithm (KNN), Support Vector Machine (SVM), Functional Tree (FT). The evaluated results were based on parameters such as accuracy, specificity and precision. KNN has the highest accuracy, specificity and precision of 98.08%, 99% and 99.36% respectively.

***Keywords**:- Machine Learning Algorithm, WEKA, Diabetes, KNN, SVM, FT.*

I. INTRODUCTION

Diabetes poses threat to human lives and attributed by hyperglycemia (Zou *et al.*, 2008). Diabetes is one of the chronic disorders resulted from deficiency of insulin secretion to control high sugar contents in the body system and it affects both sexes worldwide (Mahboob *et al.*, 2018). Based on World Health Organization (WHO) 2021 report, four hundred and twenty two million (422 000 000) people are diabetic patients worldwide with one million and six hundred thousand deaths each year, most of which are from underdeveloped or developing countries. Diabetes is not only affected by the major factor of "high sugar concentration" in the body system but with some other factors such as obesity, partial paresis, age, delayed healing, muscle stiffness, hereditary factors and deficiency in insulin secretion (Deepti and Dilip, 2018). However, early risk prediction of diabetes is a remedy from its complications (Vijayan and Anjavili, 2015). Early risk prediction of diabetes would safe medical practitioners from wasting their time and energy from clinical diagnosis of diabetic patients (Muhammad *et al*., 2018).

The dataset is obtained from the research paper (Faniqul Islam *et al.*, 2020). The dataset contains 520 samples with 17 distinct attributes.

II. RELATED WORK

Sakshi Gujral *et al.* (2017) adopted the application of classification techniques namely Support Vector Machine and Decision Trees for the prediction of diabetes mellitus. The dataset employed for this paper was obtained from PIMA Indian Diabetes Data-set. PIMA India is concerned with women's health. Support vector machine has the higher accuracy of 82%.

Deepti Sisodia *et al.*(2018) employed support vector machine, decision tree and naïve bayes algorithms. The research was performed on the Pima Indians Diabetes Database (PIDD) which is sourced from UCI machine learning repository. The results obtained showed that naïve bayes outperforms other algorithms with the accuracy of 76.30% comparatively other algorithm.

Sneha and Gangil (2019) applied decision tree, Naïve Bayesian and random forest algorithms for the early prediction of diabetes mellitus using optimal features selection. The proposed method focused on selecting the attributes for the early detection of diabetes using predictive analysis. Decision tree and random forest algorithms have the highest specificity of 98.20% and 98.00%, respectively while Naïve Bayesian has the best accuracy of 82.30%.

Hassan, Malaserene and Leema (2020) conducted prediction of diabetes mellitus using classification techniques like Decision Tree, KNearest Neighbors, and Support Vector Machines. It was observed that support vector machine (SVM) outperforms decision tree and KNN with highest accuracy of 90.23%.

III. METHODOLOGY

The proposed system includes Data collection, Data Pre-processing, System Architecture and System Evaluation.

3.1 Data Collection

The data set used in this paper was obtained from Irvine (UCI) repository of machine learning databases. The dataset contains **520** samples with **17** distinct attributes.

3.2 Data Pre – Processing

The dataset is in csv format and is being converted to arff format which is form suitable for WEKA application.

3.3 System Architecture

The figure 3. 1 depicts the architecture of the proposed system in which the early stage diabetic risk prediction dataset is classified with K-Nearest Neighbors algorithm (KNN), Random Forest (RF), Support Vector Machine (SVM) and Functional Tree (FT) classifiers.

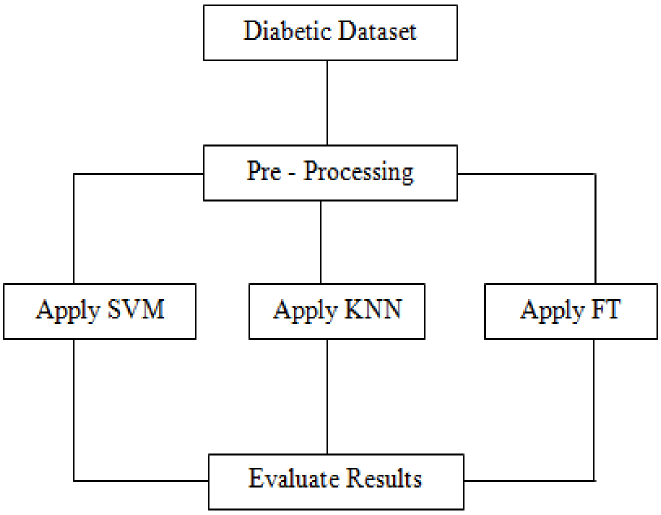

Figure 3.1 System Architecture

3.4 System Pseudo code

The pseudo code of the system is enumerated in the arrangement below:
Step 1: load Dataset of diabetes into weka application
Step 2: Pre-processing of the dataset
Step 3: Classification using support vector machine (SVM) and record the result
Step 4: Classification using K-Nearest Neighbors algorithm (KNN) and record the result
Step 5: Classification using Functional Tree (FT) algorithm and record the result.
Step 6: Evaluation of the results.

Step 1 and step 2 involved loading of the dataset into the weka application and conversion of the dataset from csv format to arff format respectively. Steps 3, 4 and 5 are the classifications using support vector machine (SVM), K-Nearest Neighbors and Functional Tree (FT) algorithms respectively. **Step 6** is the evaluation of the results based on metrics namely accuracy, specificity and precision.

3.5 Metrics used for classification

i. **Accuracy:** The percentage of the correctly classified instances i.e. accuracy, is obtained by subtracting the percentage of incorrectly classified instances from 100. Accuracy is obtained by

$$\text{Accuracy} = \frac{\text{Tp} + \text{TN}}{\text{TP+TN+FN+FP}} \times \frac{100}{1}\% \,.$$

TP = True Positive
TN = True Negative
FN = False Negative
FP = False Positive

ii. **Specificity:** Specificity is defined as the proportion of actual negatives, which got predicted as the negative (or true negative).

$$\text{Specificity} = \frac{\text{True Negative}}{\text{True Negative } + \text{False Positive}}$$

iii. **Precision:** Precision is calculated as the number of true positives divided by the total number of true positives and false positives.

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}}$$

3.6 Experimental Set Up

The experimental set up involved.

i. **WEKA Software:** Waikato Environment for Knowledge Analysis (WEKA) is a collection of machine learning algorithms for tasks in mining of data. It implements algorithms for data pre-processing, classification, regression, clustering and association rules, visualization tools are also included.

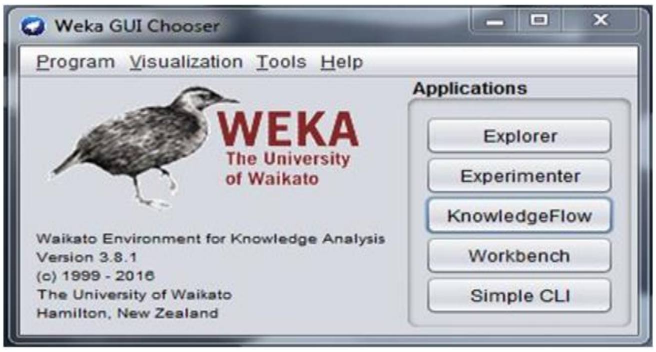

Figure 3.2: WEKA tool Applications interface (GUI Chooser)

ii. **Dataset:** The dataset employed in this paper was obtained from Irvine (UCI) repository of machine learning databases and the paper that generated the dataset is (Faniqul Islam *et al.*, 2020).

iii. **K-Fold Cross Validation:** The experimental set up using weka application in this paper is made up of 10-fold cross validation. The training dataset is randomly partitioned into 10 groups, the first 9 groups are used for training the classifier and the other group was used as the dataset to test on.

IV. RESULTS

4.1 Experimental Results

Figure 4.1 demonstrates how to load the dataset into WEKA application.

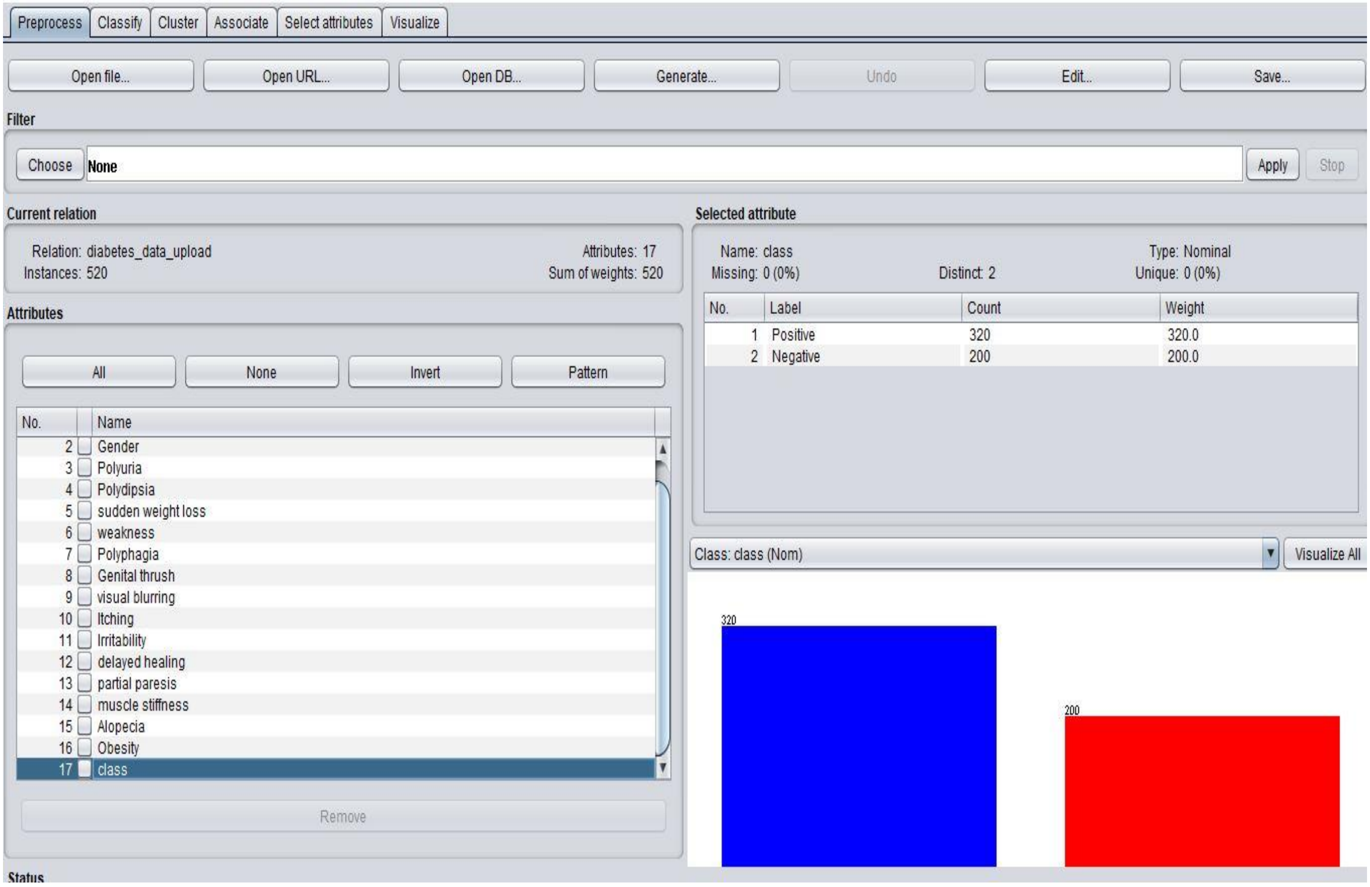

Figure 4.1: Loading dataset into WEKA Application

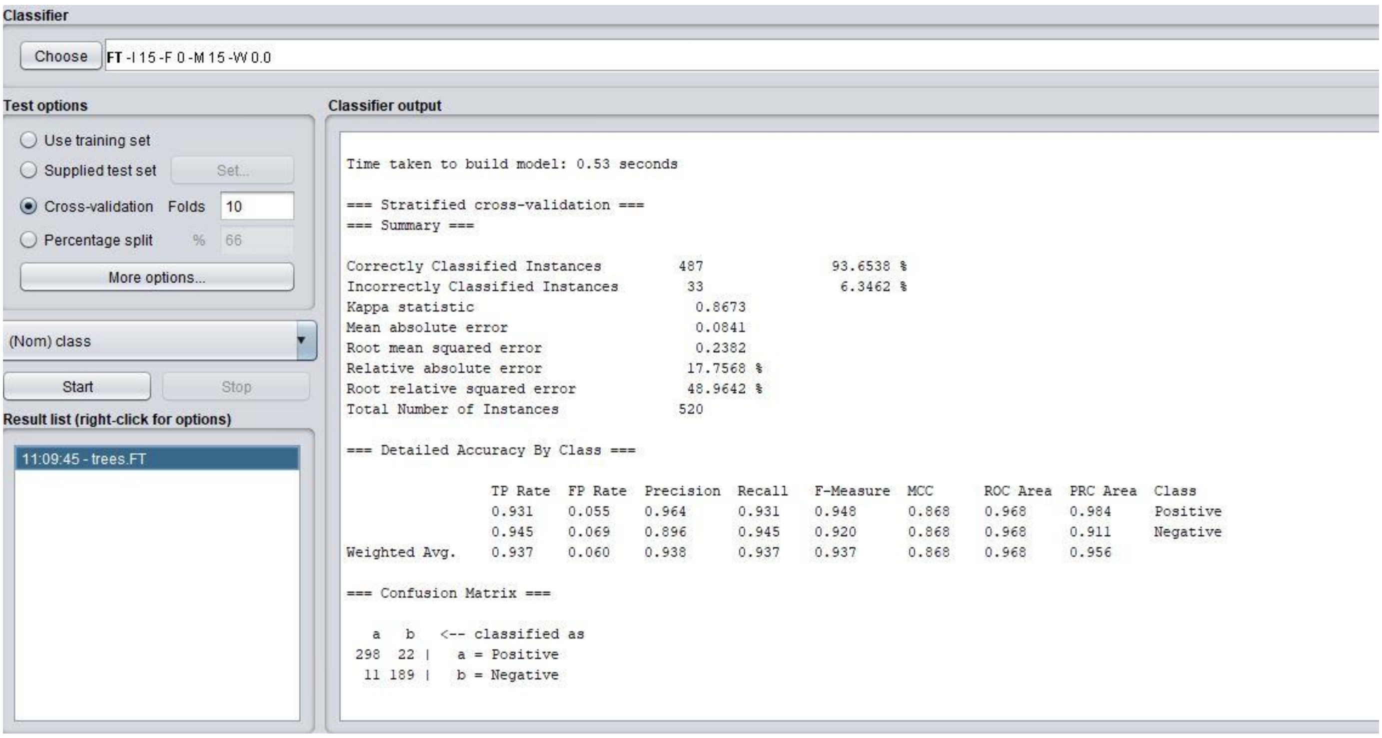

Figure 4.2: Application of Functional Tree (FT) Algorithm

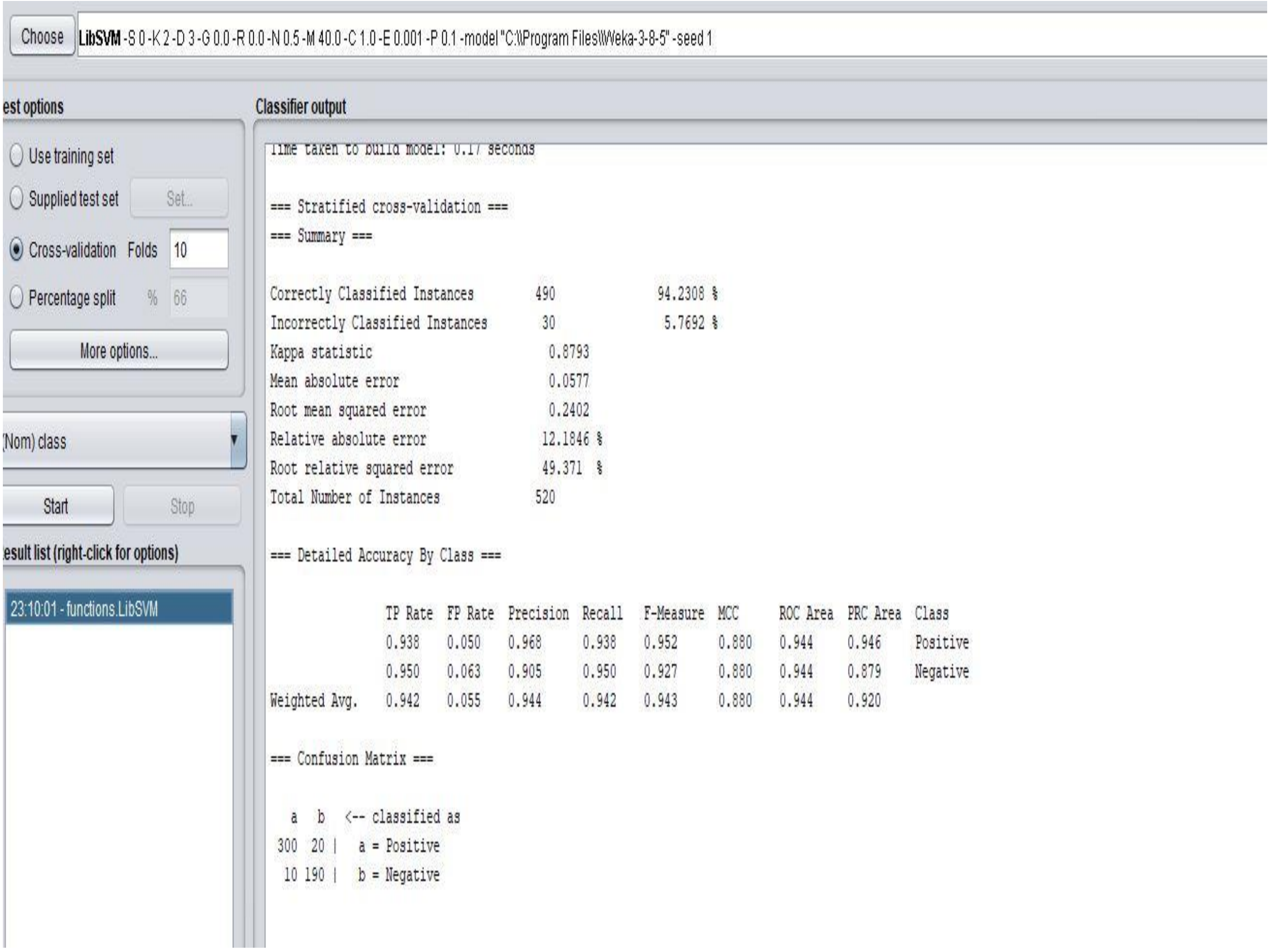

Figure 4.3: Application of Support Vector Machine (SVM) Algorithm

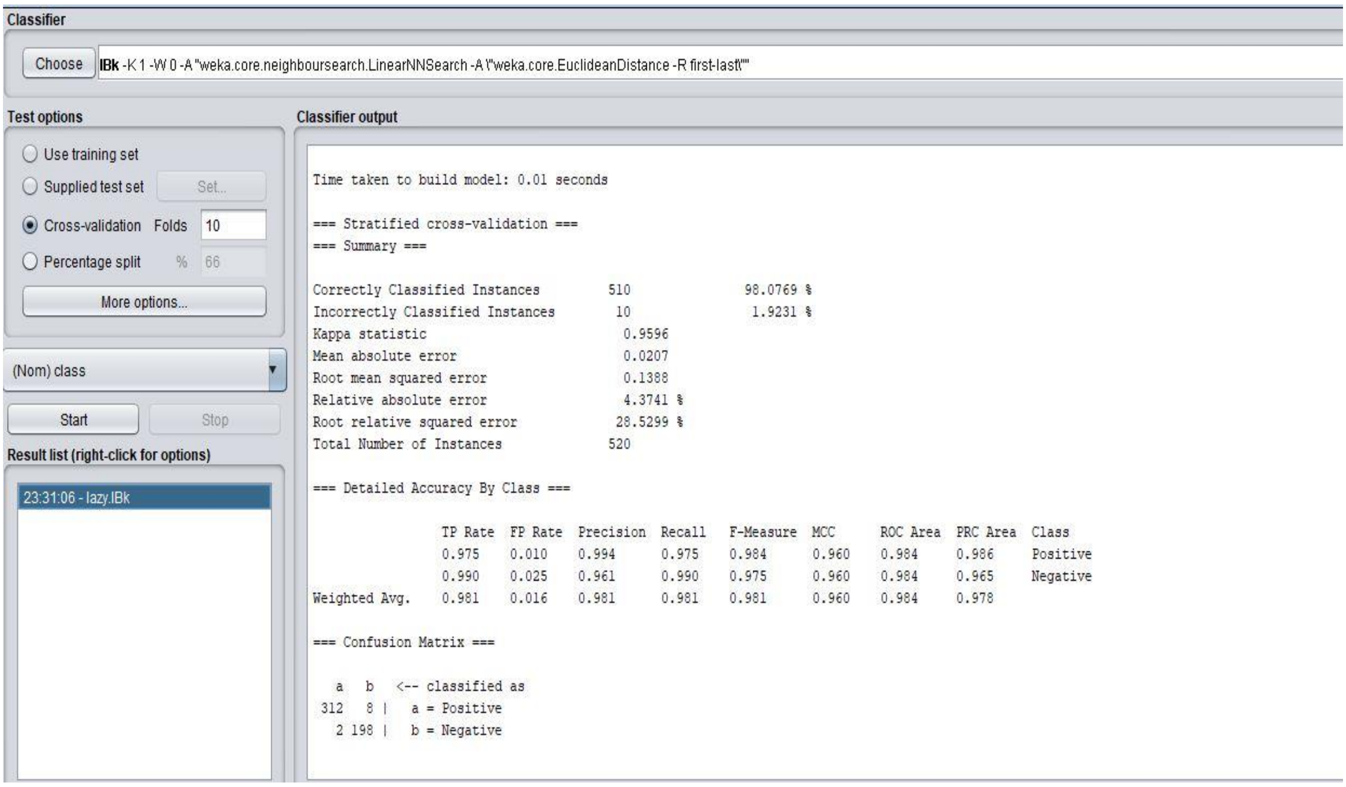

Figure 4.4: Application of K-Nearest Neighbors Algorithm

ISSN No:-2456-2165

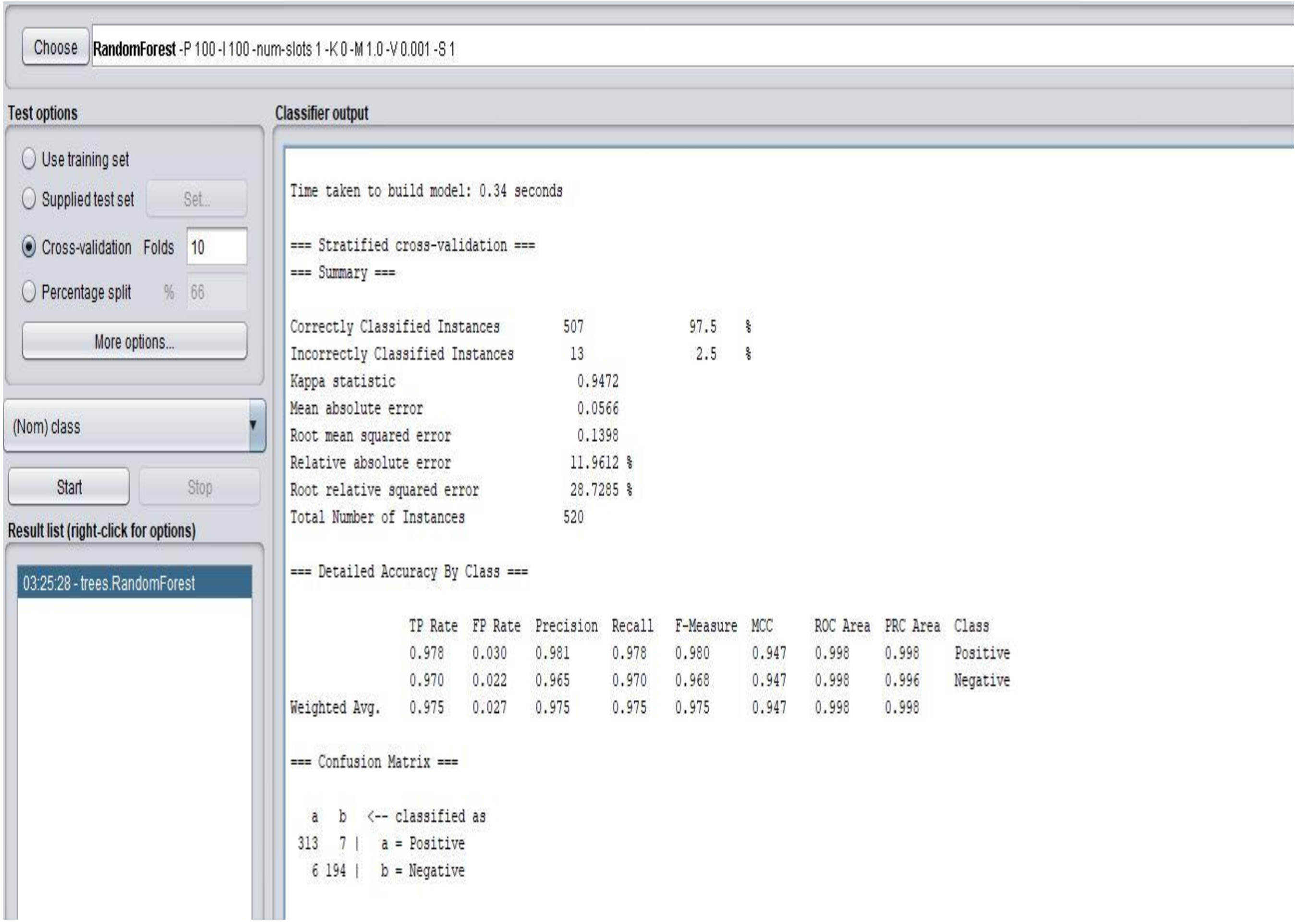

Figure 4.5: Application of Random Forest Algorithm

Table 4.1: Summary of confusion matrix and Results of evaluation metrics for the classification using machine learning algorithms

Number	Algorithms	True Positive	False Negative	False Positive	True Negative	Accuracy (%)	Precision (%)	Specificity (%)
1	K-Nearest Neighbors	312	8	2	198	98.08	99.36	99.00
2	Support Vector Machine (SVM)	300	20	10	190	94.23	96.77	95.00
3	Functional Tree (FT)	298	22	11	189	93.65	96.44	94.50
4	Random Forest (RF)	313	7	6	194	97.50	98.12	97.00

V. CONCLUSION

The research work was conducted using support vector machine (SVM), K-Nearest Neighbors (KNN), Functional Tree (FT) and Random Forest (RF) Algorithms as classifiers in which K-Nearest Neighbors perform better in terms of accuracy, precision and specificity. The research work produced highest accuracy, specificity and precision through K-Nearest Neighbors algorithm.

FUTURE WORK

Further studies can be carried out using other classification techniques, adoption of feature extraction/ feature selection techniques or using other datasets.

REFERENCES

[1]. Deepti, S. and Dilip S. (2018). Prediction of Diabetes using Classification Algorithms. International Conference on Computational Intelligence and Data Science.

[2]. Faniqul Islam M.M et al. (2020) 'Likelihood prediction of diabetes at early stage using data mining techniques.' Computer Vision and Machine Intelligence in Medical Image Analysis. Springer, Singapore, 113-125.

[3]. Hassan A.S *et al.* (2020). Diabetes Mellitus Prediction using Classification Techniques. International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-9 Issue-5.

[4]. Mahboob T. A *et al.* (2018). A model for early prediction of diabetes. Informatics in Medicine Unlocked.

[5]. Muhammad A. *et al.*(2018). Prediction of Diabetes Using Machine Learning Algorithms in Healthcare. Proceedings of the 24th International Conference on Automation & Computing, Newcastle University.

[6]. Sakshi Gujral *et al.* (2017). Detecting and Predicting Diabetes Using Supervised Learning: An Approach towards Better Healthcare for Women.International Journal of Advanced Research in Computer Science Volume 8, No. 5.

[7]. Naveen K.G *et al.* (2020). Prediction of Diabetes Using Machine Learning Classification Algorithms. International Journal of Scientific and Technology Research.

[8]. Vijayan, V.V and Anjavili, C. (2015). Prediction and diagnosis of diabetes mellitus. A machine learning approach IEEE Recent advances in Intelligent Computational Systems.

[9]. Zou *et al.*, (2018). Predicting Diabetes Mellitus with Machine Learning Techniques. Frontiers in genetics.