



doi:10.5281/zenodo.18219905

Computational Medicine: Efficacy of Diabetes Detection Models Using Filter Technique

Omoghenemuko, Greg Imoniyovwe^{1*} & Edafeajiroke, Michael Favour²

¹Department of Computer Science,
College of Education, Warri, Delta State, Nigeria

²Department of Computer Science,
University of Port Harcourt, Rivers State, Nigeria

*Corresponding Author: Omoghenemuko Greg Imoniyovwe
E- mail: omoghenemukog@mail.com

ABSTRACT

The eighth-leading cause of high mortality rate in developing countries is diabetes mellitus commonly referred to as diabetes. Diabetes is the mother of all deadly diseases and it is caused by the malfunctioning of the insulin- a hormone that lowers the level of glucose in the blood. Healthy balanced routine, exercise, and early treatment, do increase the survival rate of diabetes patients. Different diabetes detection techniques including machine learning algorithmic technique, have been vital tools for early prediction. However, to ascertain the most promising detection technique from existing techniques remains a prime concern for researchers in the 21st century. The purpose of this study, was to investigate the performance of five (5) different diabetes detection filter techniques, which are, the Chi-Square, ANOVA, Correlation-based Feature Selection (CFS), Information Gain and F-Test. Feature selection was used to build a Random Forest (RF) classifier. The model was trained with Nine (9) attributes and 100001 responses. Results of experimentation revealed that F-Test-RF outperformed the other models with an accuracy of 96.83%, a competitive precision of 90.68%, specificity of 99.33% and sensitivity 70.08%. F-test-RF can be recommended for diabetic prediction since it outperformed the other models in terms of accuracy and sensitivity based on the used dataset.

Keywords: RF+diabetes prediction, Filter technique+diabetes prediction, Chi-Square, ANOVA+diabetes prediction, CFS+diabetes prediction, Information gain+diabetes prediction and F-test+diabetes prediction

1 INTRODUCTION

Diabetes is one of the most dangerous chronic diseases that is often referred to as mother of all diseases (Manivannan et al., 2023; Abdollahi & Nouri-Moghaddam, 2022). It is caused by deficiency or ineffective insulin production (Aggarwal, 2021). Diabetes disease can be categorized as Type 1 or Type 2 and gestational diabetes. Type I occurs when there is lack of insulin in the body (Alhussan et al., 2023). Type 2 can be traced to lack of effective use of insulin that was produced or inadequate amount of insulin that was released into the bloodstream and this affects adults over the age of 45 (Sivashankari et al., 2022). Gestational diabetes affects pregnant women and its diagnostic criteria vary by country, illustrating the lack of consensus around the most effective way to screen in routine clinical care (Altaher & Malebary, 2022).

Sugar in the blood increases when the glucose in the body remains undigested or is not metabolized properly (Saxena et al., 2022). And this after long time increases the sugar level in the blood (Vijiya, 2019). There is no known cure for diabetes to the best of the researchers' knowledge, but a diabetic patient can live a healthy life after following a balanced routine. However, if the proper treatment is not received by a diabetic at an appropriate time, it could lead to others complicated diseases such as acute myocardial infarction (heart attack), blindness, kidney failure, pneumonia, stroke, deterioration of vital organ and other chronic and deadly diseases that ultimately lead to excruciating deaths of these patients (Manivannan et al., 2023; Saxena, et al., 2022; Abdollahi & Nouri-Moghaddam

2022; Abe et al., 2021). Therefore, it is better to predict diabetes as early as possible so that all the parts of the body can function properly (Anggoro & Permatasari, 2023). Hence the study. In recent time attention have been drawn to the use of a robust machine learning algorithms for an efficient diabetes prediction process (Saxena, *et al.*, 2022). One major way this can be actualized is identifying diabetes and its risk factors and several classification algorithms such as Naïve Bayes (NB), random forest (RF), Ada boost (AB), multi-layer perceptron (MLP), decision tree (DT) have been proposed as a solution but sample size, feature selection, missing data differentiate are serious debacles to their performances.

Therefore, it is essential to compare and assess the performance of different algorithms using standardized metrics and datasets to identify the most effective approach for diabetes detection (Mohiddin *et al.*, 2022). As a result of the penumbra of uncertainty surrounding diabetes detection, pointing out the most effective approach for diabetes detection/prediction still remains an issue of heated debate by researchers. Therefore, the study employed several filter techniques, which are Chi-Square, ANOVA/ANCOVA models, correlation-based feature selection, information gain, F-Test, and RF-model using python programming language, to identify and remove irrelevant and redundant attributes from the dataset by performing feature selection on the diabetes dataset obtained from kaggle.com to achieve a formidable prediction. Optimal feature subset was used to build the random forest classifying and result obtained were compared for effective diabetic detection.

Purpose of the study

The purpose of this study was to investigate the performance of five (5) different filter techniques, which are the Chi-Square, ANOVA, Correlation-based Feature Selection (CFS), Information Gain and F-Test.

Scope of the study

The study covered the comparison of the efficacy of five diabetes detection models using filter technique.

Challenges of existing systems

There is no known cure for diabetes to the best of the researchers' knowledge, but a diabetic patient can live a healthy life after following a balanced routine. However, if measures to ameliorate effects of diabetes are not articulated and followed appropriately and timely, it could engender others complicated diseases like acute myocardial infarction (heart attack), blindness, kidney failure, pneumonia, stroke, deterioration of vital organ and other chronic and deadly diseases that ultimately lead to excruciating deaths of these patients (Manivannan et al., 2023; Saxena, et al., 2022; Abdollahi & Nouri-Moghaddam 2022; Abe et al., 2021). Therefore, it is better to predict diabetes as early as possible so that all the parts of the body can function properly (Anggoro & Permatasari, 2023). In recent times, attention has been drawn to the use of a robust machine learning algorithms for an efficient diabetes prediction process (Saxena, *et al.*, 2022). One major way this can be actualized is identifying diabetes and its risk factors, and several classification algorithms such as Naïve Bayes (NB), random forest (RF), Ada boost (AB), multi-layer perceptron (MLP), decision tree (DT) have been proposed as solution but sample size, feature selection, missing data differentiation, and lack of efficiency in previous models, are serious debacles to their performances.

2 LITERATURE REVIEW

Anggoro and Permatasari (2023) carried out a study on performance comparison of the kernels of the support vector machine (SVM) algorithm for diabetes mellitus classification, compared the accuracy, precision, recall, and F1-score values of the SVM algorithm with various kernels and data pre-processing. The pre-processing data used included data splitting, data normalization, and data oversampling. This research has the benefit of solving health problems based on the percentage of diabetes mellitus and can be used as material for accurate information. The results of this study revealed that the highest accuracy was obtained by 80% obtained from the polynomial kernel, the highest precision was obtained by 65% which was also obtained from the polynomial kernel, and the highest recall was obtained by 79% obtained from the RBF kernel and the highest F1-score was obtained by 70% obtained from RBF kernel. However, with pertinent feature employed to handle the pair wise classification model, this model can be improved.

A study by Alhussan et al. (2023) investigated the classification of diabetes using feature selection and hybrid Al-Biruni earth radius and dipper throated optimization. The study proposed a new feature selection algorithm based on a dynamic Al-Biruni earth radius and dipper-throated optimization algorithm (DBERDITO). The selected features were classified using a random forest classifier with its parameters optimized using the proposed DBERDITO. Results: The proposed methodology is evaluated and compared with recent optimization methods and machine learning models to prove its efficiency and superiority. The overall accuracy of diabetes classification achieved by the study was 98.6%. On the other hand, statistical tests were conducted to assess the statistical significance and the statistical difference of the proposed approach based on the analysis of variance (ANOVA) and Wilcoxon signed-

rank tests. Conclusions: The results of these tests confirmed the superiority of the proposed approach compared to the other classification and optimization methods.

Almutairi and Abbod (2023) carried out a study on machine learning methods for diabetes prevalence classification in Saudi Arabia. The study investigated the ability of different classification methods to classify diabetes prevalence rates and the predicted trends in the disease according to associated behavioural risk factors (smoking, obesity and inactivity) in Saudi Arabia. Classification models for diabetes prevalence were developed using different machine learning algorithms, including linear discriminant (LD), support vector machine (SVM), K-nearest neighbour (KNN), and neural network pattern recognition (NNPR). Four kernel functions of SVM and two types of KNN algorithms were used, namely linear SVM, Gaussian SVM, quadratic SVM, cubic SVM, fine KNN, and weighted KNN. The performance evaluation in terms of the accuracy of each developed model were determined, and the developed classifiers were compared using the Classification Learner App in MATLAB, according to prediction speed and training time. The experimental results on the predictive performance analysis of the classification models showed that weighted KNN performed well in the prediction of diabetes prevalence rate, with the highest average accuracy of 94.5% and less training time than the other classification methods, for both men and women datasets. Several studies have obtained a better performance hence the study.

Mohiddin et al. (2022), study was based on An Approach for Early Prediction of Diabetes using Firefly Optimization Algorithm. The study evaluates the performance of the firefly algorithm using four wide metrics for evaluation: accuracy, precision, recall, and F-score. Our experiments were conducted on a real-world dataset consisting of 768 individuals, of which 268 had diabetes. The training and testing sets were randomly divided into two groups with an 80:20 ratio. The study performed the firefly algorithm for feature selection. It was used to optimize the parameters using the firefly algorithm. Then the optimized parameters were then used to train the firefly algorithm on the entire training set. The experimental results demonstrate that the firefly algorithm achieves competitive performance compared to other machine learning algorithms in terms of precision, accuracy, F-score, and recall, the firefly method outperforms other algorithms.

Jader and Sadegh (2022) study designed an artificial neural network detection model for recognizing diabetes mellitus. The dataset was obtained from the Kurdistan region, which collected information from pregnant women with and without diabetes. The criterion of this study was to minimize the error function in neural network training using a neural network model design. After designing the ANN model, the error rate of the neural network decreased during training, based on its designing process, and the accuracy of the prediction increased to 91%. The results showed that the ANN algorithm improves prediction accuracy based on its network design.

Ghabousiana et al. (2022) carried out a study on Hybrid of particle swarm optimization algorithm and fuzzy system for diabetes diagnosis. The study utilized fuzzy systems to binary the particle swarm algorithm. The achieved model was applied to the diabetes dataset and then evaluated using a neural network classifier. The results indicated an increase in classification accuracy to 95.47% compared to other existing methods.

A study by Le (2021) was based on a Novel Wrapper Based Feature Selection for Early Diabetes Prediction Enhanced with a Metaheuristic. In this study, a machine learning model was used to predict the early onset of diabetes patients. It is a novel wrapper-based feature selection utilizing Grey Wolf Optimization (GWO) and an Adaptive Particle Swarm Optimization (APSO) to optimize the Multilayer Perceptron (MLP) to reduce the number of required input attributes. Moreover, study also compared the results achieved using this method and several conventional machine learning algorithms approaches such as Support Vector Machine (SVM), Decision Tree (DT), K-Nearest Neighbour (KNN), Naïve Bayesian Classifier (NBC), Random Forest Classifier (RFC), Logistic Regression (LR). Computational results of our proposed method show not only that much fewer features are needed, but also higher prediction accuracy can be achieved (96% for GWO - MLP and 97% for APGWO - MLP). Although this work has the potential to be applicable to clinical practice and become a supporting tool for doctors/physicians but an enhance prediction can be obtained if pertinent features are employed hence the study.

A study by Aslam (2021) attempted to develop an early prediction model for diabetes readmission and identified the significant factors that lead to readmission of diabetes patients. The early prediction will reduce the risk of hospital readmission. Several machine learning classifiers, such as Logistic Regression (LR), Decision Tree (DT), and Random Forest (RF), were applied. The Firefly bio-inspired technique was used for feature selection and model optimization. Synthetic Minority Oversampling Technique (SMOTE) was applied to alleviate the data imbalance problem. The performance of the classifiers was compared using different feature sets. Experiments showed that RF outperformed the other models using reduced features selected by the Firefly algorithm. The study achieved the highest accuracy, precision, recall, and Area Under Curve (AUC) of 0.99, 0.99, 0.94, and 0.98, respectively. The results show the significance of the proposed model in diabetes readmission prediction. As a result, it is suggested that other system models and multiple data sets be investigated in order to obtain better results and identify significant features for early readmission prediction in diabetic patients. Hence the study.

Sivaranjani et al. (2021) employed Machine learning algorithms Support Vector Machine (SVM) & Random Forest (RF) to identify the potential chances of getting affected by Diabetes Related Diseases. After pre-processing the data, features which influences the prediction were selected by implementing step forward and backward feature selection. The Principal Component Analysis (PCA) dimensionality reduction method is analysed after the selection of specific features and the accuracy of the prediction is 83% implementing Random Forest (RF) which is significant in comparison with Support Vector Machine (SVM) with accuracy of 81.4%. However, with comparison of different approached an enhanced model can be obtained hence this study.

Aggarwal (2021) analysed various features selection techniques used in predicting diabetes. They included: The Recursive Feature Elimination (RFE), Principal Component Analysis (PCA), Chi-Square Test, and Correlation-based Feature Selection (CFS), to improving diabetes prediction using random forest. We used the publicly-available data set from the NIH for the analysis. It included 768 specimens and eight features (Pima Indian Diabetes dataset). The data collected for the study included 8 features and 768 samples. After processing the data, the study evaluated the performance of the four feature selection methods and also compared the results with the importance ranking of the selected features to identify the underlying biological factors. The results of the study revealed that the four different techniques used improved the performance of the random forest model when compared to the use of all features. For instance, the RFE and PCA techniques had the highest accuracy, while the Chi-Square Test and the CFS techniques had the highest specificity and sensitivity. The selected features also varied in appearance, with the former appearing in all four techniques. The findings of the study indicate that the use of feature selection can help improve the accuracy of the prediction of diabetes. The four techniques studied had weaknesses and strengths, which suggests that researchers should pay attention to the characteristics of the dataset when choosing a technique. The analysis's conclusion suggests that future studies should expand the scope of the techniques and test them on larger and more diverse sets of data. Hence the study.

Vijiya (2019) developed a system which can perform early prediction of diabetes for a patient with a higher accuracy by using Random Forest algorithm in machine learning technique. The Random Forest algorithms are often used for each classification and regression tasks and it is also a type of ensemble learning method. The accuracy level is greater when compared to other algorithms. The proposed model gives the best results for diabetic prediction and the result showed that the prediction system is capable of predicting the diabetes disease effectively, and most importantly, instantly. However, the efficiency of this model can be improved when feature selection technique are employed to select optimal feature subset hence the study.

3 METHOD

Design of the study

The design of the study is object-orientation and experimental design. The experimental framework of the study is expressly stated in sequential steps in view of achieving the stated objectives. The experiments were designed to compare different filter algorithms like the Chi-Square, ANOVA, Correlation-Based Feature Selection, Information Gain and F-Test Approaches with Random Forest (RF) as the classifier on the diabetes dataset that was obtained from kaggle repository. Best feature subset obtained was used to build a Random Forest classifier for each Filter techniques. The system was designed to get the optimal feature subset from the dataset so as to avoid outliers, under-sampling and oversampling in the dataset designed to simulate a data mining sequential process. RF Classifier has been good classification and regression algorithm on the diabetes dataset but to get a more promising results five (5) filter algorithms were explored. The different filter techniques were introduced to determine the best features pertinent enough for obtaining an efficient model for diabetes prediction processes.

Figures 3.1 and 3.2 show a flowchart and a flow diagram that describe the steps involved in the collection of diabetes dataset from kaggle dataset repository. Filter algorithm was employed for feature selection by optimizing the high dimensional dataset for better classifier performance. This way, main factors and the life style responsible for diabetes Mellitus were determined, the best features were then used to build the developed model. The dataset was partitioned into two-fold: The training and testing set at a percentage ratio of 75% to 25% respectively. Results obtained from each developed model were compared to ascertain the best model for diabetic prediction on the diabetes dataset. The results were evaluated based on ML statistical metrics like the classification accuracy, specificity, sensitivity and precision. These processes were implemented on a python Jupiter Notebook in a Google coolab high-performance language for Machine Learning Platform.

Figure 3.1
Flowchart of feature selection using filter technique for evaluation of diabetes detection models' performance using random forest classification algorithm

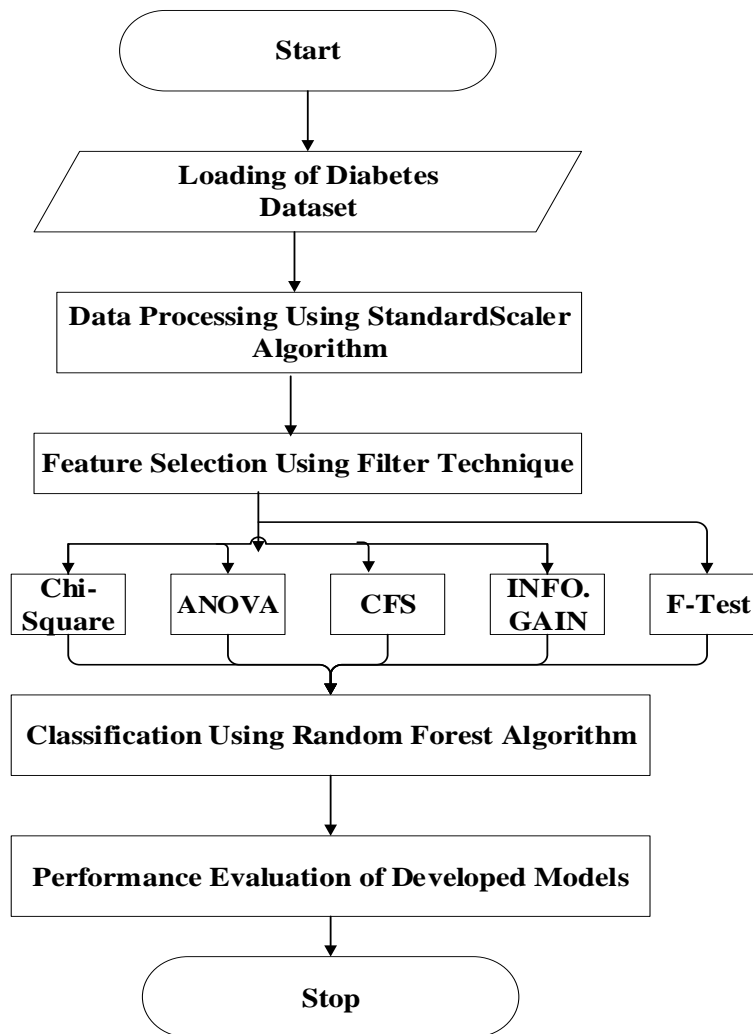
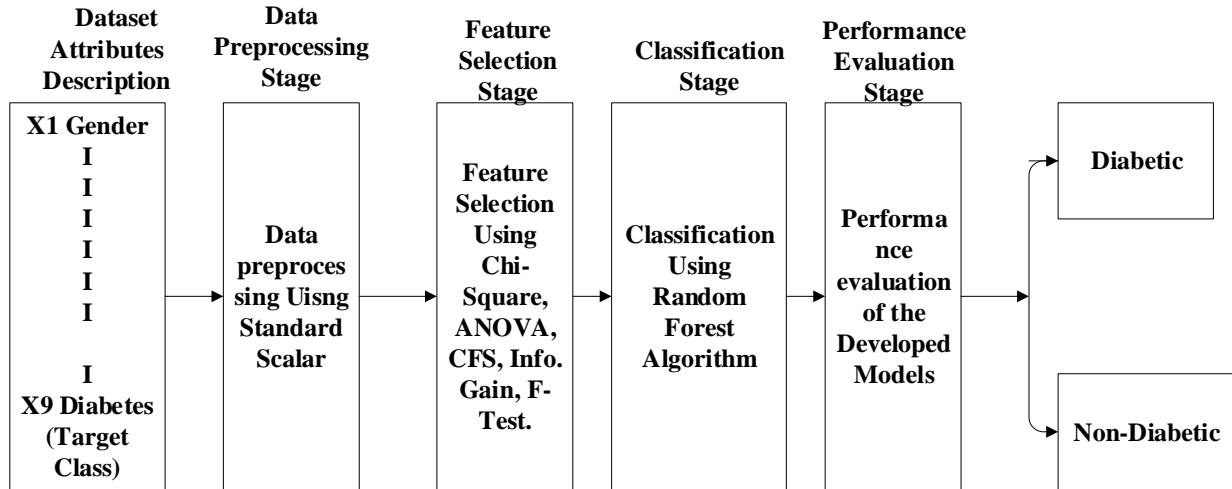


Figure 3.2 Framework for the developed system



Dataset Acquisition

A diabetes dataset was used to formulate a knowledge base for the system so as to build an efficient system. The dataset was obtained from the National Institute of Diabetes and Digestive and Kidney Diseases. The Dataset consists of Nine (9) different attributes of which one (1) was taken as the target variables or the class label and then Ten Thousand and One responses (10001). This dataset was passed unto the filter algorithm for feature selection processes as shown in Figures 3.1 and 3.2. This feature selection process, reduced computational overhead to the efficiency of the developed system. Attribute description of the various dataset are shown in Table 3.1.

Table 3.1
Attribute description and definition

Feature Index	Feature Name
1	(string) Gender
2	(int) Age
3	(bool) hypertension
4	(bool) heart_disease
5	(string) smoking_history
6	(Float) bmi
7	(Float) HbA1c_level
8	(int) blood_glucose_level
9	(bool) diabetes: target variable or Class Label

Performance Evaluation Parameters

To determine the effectiveness of the developed model, statistical yardstick was used to measure the effectiveness of the model in terms of the positive predictive rate, negative predictive rate as well as the classification accuracy (Zhu, et al., 2017). Example are the True positive (TP) rate, True Negative (TN), False Positive (FP) rate, False Negative (FN) rate, from this classification computational timing, accuracy, sensitivity (recall) and specificity were derived. If TP Prediction is +ve and patient is diabetic, this is a desirable result, TN Prediction is -ve and patient is non-diabetic, it is also a desirable result with FP Prediction is +ve and patient is non-diabetic, this is a false alarm, and can be regarded as bad and lastly with FN Prediction is -ve and patient is diabetic, this is worst of all and must be avoided (Zhu, et al., 2017). This study detects diabetics by giving its output as either diabetic (+ve) or non-diabetic (-ve). To achieve that if it starts with True then the prediction was correct whether diabetic or not, so TP is a diabetic patient correctly predicted and a TN is a healthy patient correctly predicted. Oppositely, if it starts with False then the prediction was incorrect, so FP is a healthy patient incorrectly predicted as diabetic (+) and a FN is a diabetic patient incorrectly predicted as healthy. Positive or negative indicates the output of our program (Baratloo et al., 2015). correct or incorrect output is judged by true or false. Based on all these the performance Parameters are as described.

i. Classification Accuracy (CA)

The consistency of a test is its ability to accurately distinguish between the patient and stable cases. In order to estimate the accuracy of a test, in all assessed cases, we can determine the proportion of true positive and true negative (Baratloo et al., 2015). See equation (3.1) mathematical illustration of CA:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.1}$$

ii. Sensitivity (Recall)

The number of True Positives divided by the number of True Positives plus the number of False Negatives is called Recall. Put another way, the number of positive predictions in the test data is separated by the number of positive class values. It is also called Sensitivity or the True Positive Rate (Baratloo et al., 2015). Mathematically, this can be described as shown equation (3.2).

$$\text{Sensitivity} = \frac{TP}{TP + FN} \tag{3.2}$$

iii. Specificity

A test done to determine the ability of healthy cases correctly is termed specificity. To estimate it, proportion of true negative in healthy cases are calculated. Mathematically, this can be stated as shown in equation (3.3).

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{3.3}$$

iv. Precision

The sum of True Positives divided by the number of True Positives plus False Positives is precision. In other words, it is the number of positive predictions separated by the total number of predicted positive class values. Good predictive value (PPV) is often referred to. Mathematically, this can be stated as shown equation (3.4).

$$\text{Precision} = \frac{TP}{TP + FP} \tag{3.4}$$

4 RESULTS AND DISCUSSIONS

Overview to analysis of the results of the developed system

This section presents the results analysis of the data science approach- which involves object-orientation and experimentation used in achieving the stated objectives. The experiment was designed to analyse the performance of different filter algorithm on random forest classifier for effective diabetic prediction process. The results of each stage were implemented to obtain the best model which was in turn compared with the different models. Accuracy

of the different models were compared, the results were evaluated based on machine learning statistical metrics like the classification accuracy, true positive rate, false negative rate, error rate, specificity, sensitivity and training, and testing time. The experimental setup was successfully carried out and developed with the Python programming language in Google coolab with interwoven connected component in the platform so as to create a friendly user interface. The developed systems used various components' environments in the platform for successful presentation of the result from the various data mining stages, namely, data pre-processing, feature selection, classification and performance evaluation.

The Python Programming Command Window

The python command window helps to display result of the data mining task in a console screen for readability and easy expression of output. See Figure 4.1.

Figure 4.1

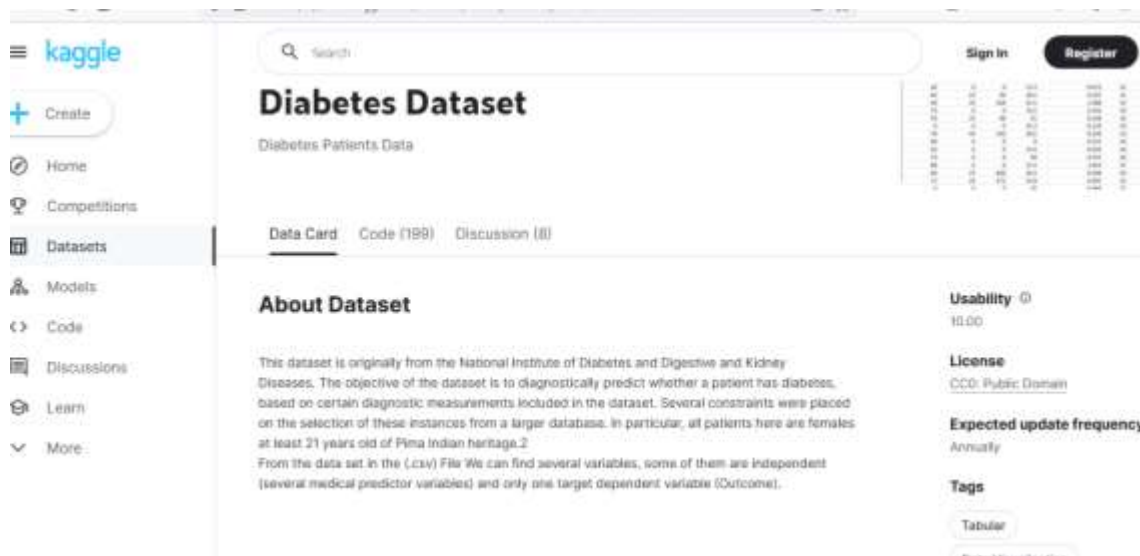
The Python Programming Command Window on Google Coolabs



Figure 4.2 shows an outlook of the diabetes dataset acquisition for the developed system, the dataset was acquired through the Kaggle dataset reposition. at run time all modules vital to actualizing the system aim were automated into the working process of developed platform. The interactive graphical user interface automates the loading of data via the load button which loaded the dataset into the platform.

Figure 4.2

Diabetes dataset acquisition interface for the developed system



When the dataset was pre-processed and best features selected using Chi-Square, ANOVA, Correlation-Based Feature Selection, Information Gain and F-Test resulted to successful selection of optimal feature for building the random forest classifier models. Developed models were compare to ascertain the most effective models for diabetes prediction process. Figure 4.3 and Table 4.1 show that only Five (5) attributes were selected after optimization. However, to detect diabetes, the order of importance of this risk factors were described in descending order (order of importance) as obtained from the feature selection algorithm.

Figure 4.3
Features selected from the filter algorithms

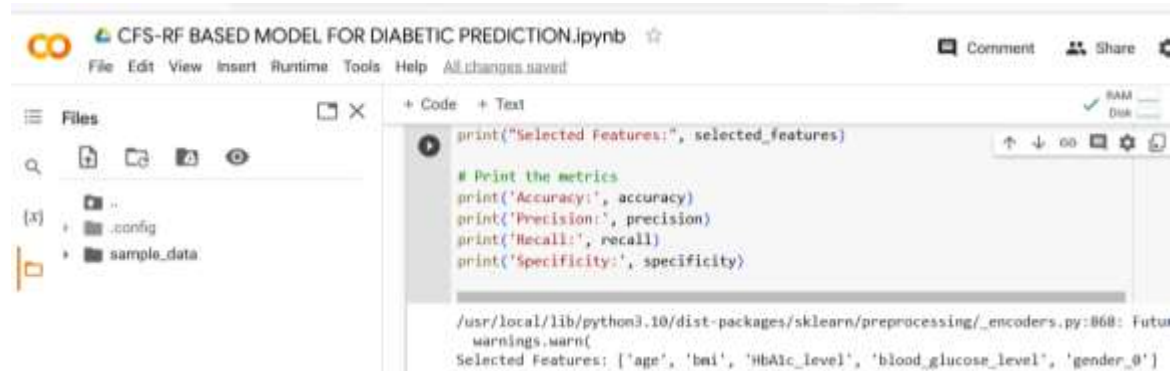


Table 4.1 Features selected from the filter algorithms

No of features	Diabetes: Target Variable
1	age
2	bmi
3	HbA1c_Level
4	blood_glucose_level
5	gender

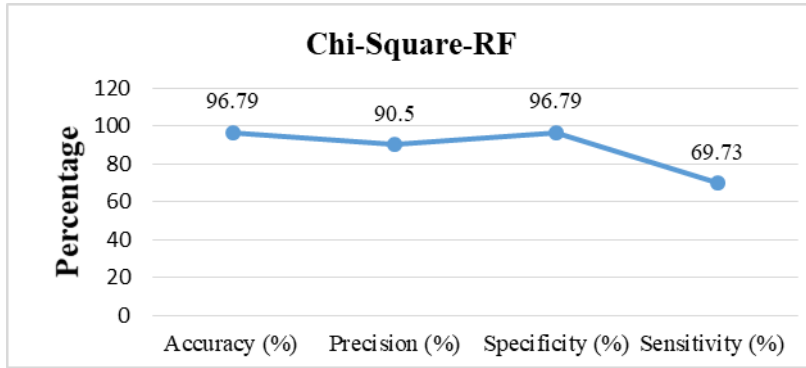
Training and Classification of the Developed Models

The model was trained and classified by splitting the data into training and testing set of 75% and 25% respectively, and then each model was evaluated. This section gives the exact values for the different confusion matrixes. The different result obtained for Chi-Square, ANOVA, Correlation-Based Feature Selection, Information Gain and F-Test were noted for comparative analysis for recommendation on the diabetic dataset. The table 4.2 shows the result for the evaluation parameters of the Chi-Square-RF for diabetes prediction. Result obtained in percentage for Accuracy, Precision, Specificity and Sensitivity are 96.79%, 90.50%, 96.79% and 69.73% respectively. While Fig. 4.4 describe the Performance Analysis of Chi-Square-RF Model for Diabetic Prediction.

Table 4.2
Performance analysis of Chi-square-RF model for diabetic prediction

Technique	Accuracy (%)	Precision (%)	Specificity (%)	Sensitivity (%)
Chi-Square-RF	96.79	90.50	96.79	69.73

Figure 4.4 Graphical illustration of the performance analysis of Chi-Square-RF model for diabetic prediction



The Table 4.3 shows the result for the evaluation parameters of the ANOVA-RF for diabetes prediction. While Fig. 4.5 provides the graphical illustration of the Performance Analysis of ANOVA-RF Model for Diabetic Prediction. Result obtained in percentage for Accuracy, Precision, Specificity and Sensitivity are 96.79%, 90.50%, 96.79% and 69.73% respectively.

Table 4.3 Performance analysis of ANOVA-RF model for diabetic prediction

Technique	Accuracy (%)	Precision (%)	Specificity (%)	Sensitivity (%)
ANOVA-RF	96.79	90.50	96.79	69.73

Figure 4.5 Graphical illustration of the Performance Analysis of ANOVA-RF Model for Diabetic Prediction

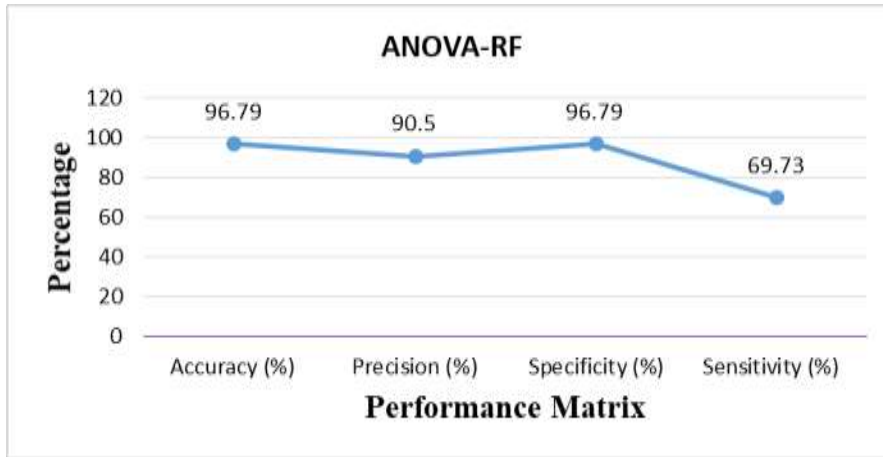


Table 4.4 indicates the results for the evaluation parameters of the Information Gain-RF for diabetes prediction. Fig. 4.6 shows the graphical illustration of the Performance Analysis of Information Gain-RF Model for Diabetic Prediction. Result obtained in percentage for Accuracy, Precision, Specificity and Sensitivity are 96.80%, 91.13%, 99.37% and 69.20% respectively.

Table 4.4 Performance analysis of information gain-RF model for diabetic prediction

Technique	Accuracy (%)	Precision (%)	Specificity (%)	Sensitivity (%)
INFO. GAIN-RF	96.80	91.13	99.37	69.20

Figure 4.6 Graphical illustration of the performance analysis of information gain-RF model for diabetic prediction

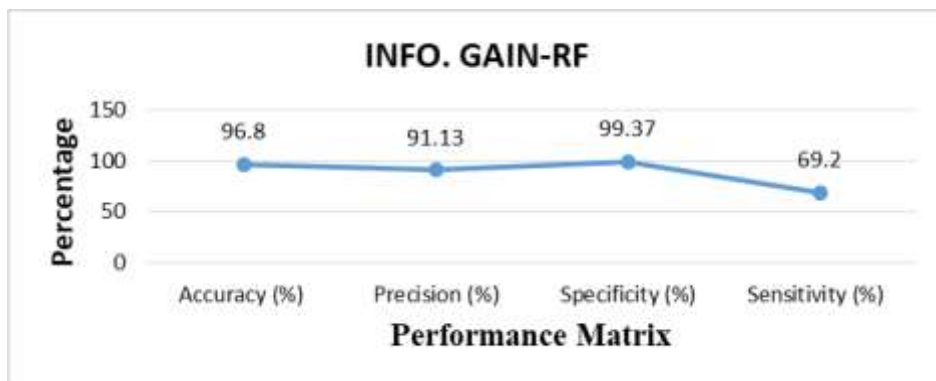
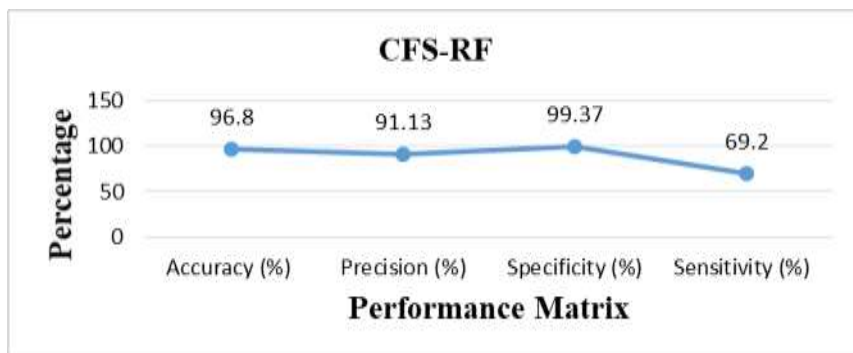


Table 4.5 shows the result for the evaluation parameters of the Correlation-Based Feature Selection-RF for diabetes prediction. Fig. 4.7 graphically illustrate the performance analysis of CFS-RF Model for Diabetic Prediction. Result obtained in percentage for Accuracy, Precision, Specificity and Sensitivity are 96.80%, 91.13%, 99.37% and 69.20% respectively.

Table 4.5 Performance Analysis of CFS-RF Model for Diabetic Prediction

Technique	Accuracy (%)	Precision (%)	Specificity (%)	Sensitivity (%)
CFS-RF	96.80	91.13	99.37	69.20

Table 4.7 Graphical illustration of the performance analysis of CFS-RF model for diabetic prediction

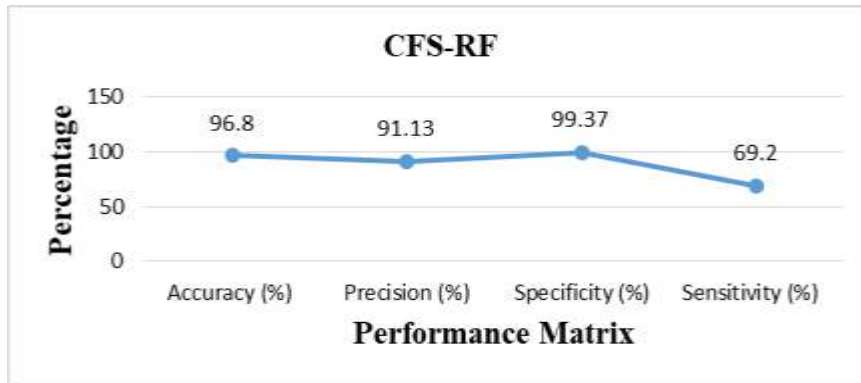


The Table 4.6 shows the result for the evaluation parameters of the F-Test-RF for diabetes prediction. Fig. 4.8. Graphically illustrate the Performance Analysis of F-Test-RF Model for Diabetic Prediction. Result obtained in percentage for Accuracy, Precision, Specificity and Sensitivity are, 96.83%, 90.68%, 99.33% and 70.08% respectively.

Table 4.6 Performance analysis of F-Test-RF model for diabetic prediction

Technique	Accuracy (%)	Precision (%)	Specificity (%)	Sensitivity (%)
F-Test-RF	96.83	90.68	99.33	70.08

Figure 4.8 Graphical illustration of the performance analysis of F-Test-RF model for diabetic prediction



Performance Evaluation Analysis of Different Filter Techniques

In order to ascertain the best system to recommend for diabetic prediction process, various models were developed with a Filter Algorithms such as Chi-Square, ANOVA, Correlation-Based Feature Selection, Information Gain and F-Test. System performance was compared using the different performance matrices. This session therefore gives Result obtained in percentage for Accuracy, Precision, Specificity and Sensitivity for the developed models. Table 4.7 showed a description of result obtained for Performance Evaluation Analysis of the filter Techniques. Results showed that F-Test-RF outperformed the other model with an accuracy level of 96.83% followed by the INFO. GAIN-RF and CFS-RF with an accuracy of 96.80%, Chi-Square-RF and ANOVA-RF have a competitive performance of 96.79%. in terms of precision INFO. GAIN-RF and CFS-RF performed better with a value of 91.13%, followed by the F-Test-RF model with 90.68%. Chi-Square-RF and ANOVA-RF had a precision of 90.50% in each model. In terms of Specificity the performance of INFO. GAIN-RF and CFS-RF outperformed the other models with a value of 99.37%. Chi-Square-RF and ANOVA-R had 96.79% in each case. While F-Test-RF had 99.33%. In terms of Sensitivity F-Test-RF out-performed the other models with a value of 70.08%. This is followed by the Chi-Square-RF and ANOVA-RF models with a sensitivity of 69.73%. result showed that the performance of all the model were highly impressive and were closely related in their performance, F-test-Rf can be recommended for diabetic detection process since it outperformed the other models in terms of accuracy and sensitivity. Results in all the models shows confidence in its knowledge discovery from the dataset used.

Table 4.7 Performance evaluation analysis of the different filter techniques

Technique	Accuracy (%)	Precision (%)	Specificity (%)	Sensitivity (%)
Chi-Square-RF	96.79	90.50	96.79	69.73
ANOVA-RF	96.79	90.50	96.79	69.73
INFO. GAIN-RF	96.80	91.13	99.37	69.20
CFS-RF	96.80	91.13	99.37	69.20
F-Test-RF	96.83	90.68	99.33	70.08

Graphical Analysis

The graphical analysis shows a comparative result of the classification accuracy, precision specificity, sensitivity, the various developed model achieved with the five different filter techniques. this section gives a graphical representation of comparative analysis of the five different filter techniques employed in this study which were Chi-Square, ANOVA, Correlation-Based Feature Selection, Information Gain and F-Test.

Figure 4.4. shows a comparative analysis of the different filter techniques based on classification accuracy. Result showed that the classification accuracy obtained were 96.79%, 96.79%, 96.8%, 96.8% and 96.83% respectively. F-Test-RF obtained a higher performance in terms of the accuracy hence F-Test-RF model can be recommended for diabetes prediction process.

Figure 4.4 Comparative analysis of the different filter techniques based on classification accuracy

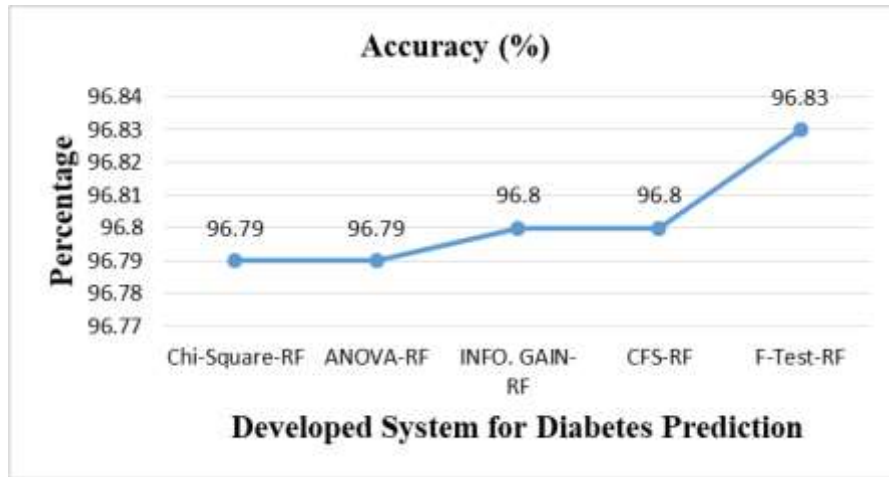


Figure 4.5 gives a graphical representation of comparative analysis of the five different filter technique employed in this study which were Chi-Square, ANOVA, Correlation-Based Feature Selection, Information Gain and F-Test. Result showed that the precision obtained were 90.50%, 90.50%, 91.13%, 91.13% and 90.68% respectively. The Info-Gain-RF and the CFS-RF obtained the same performance and prove to outperformance other models although Result obtained from F-test-RF Model was competitive with other models.

Figure 4.5 Comparative analysis of the different filter techniques based on precision

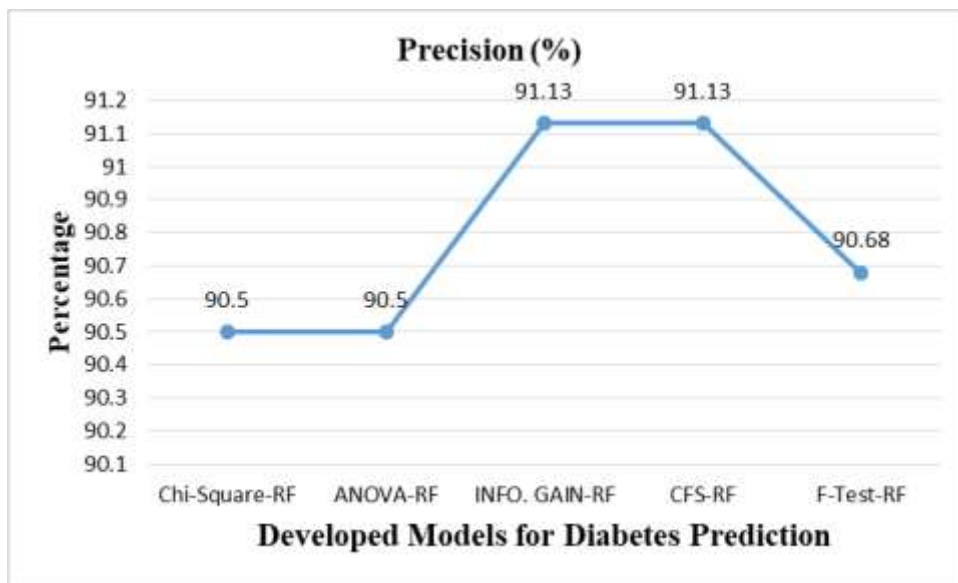


Figure 4.6 gives a graphical representation of comparative analysis of the five different techniques employed in this study which were Chi-Square, ANOVA, Correlation-Based Feature Selection, Information Gain and F-Test. Result showed that the specificity obtained were 96.79%, 96.79%, 99.37%, 99.37% and 99.33% respectively. Result obtained from F-test-RF Model was competitive with other models.

Figure 4.6
Comparative analysis of the different filter techniques based on specificity

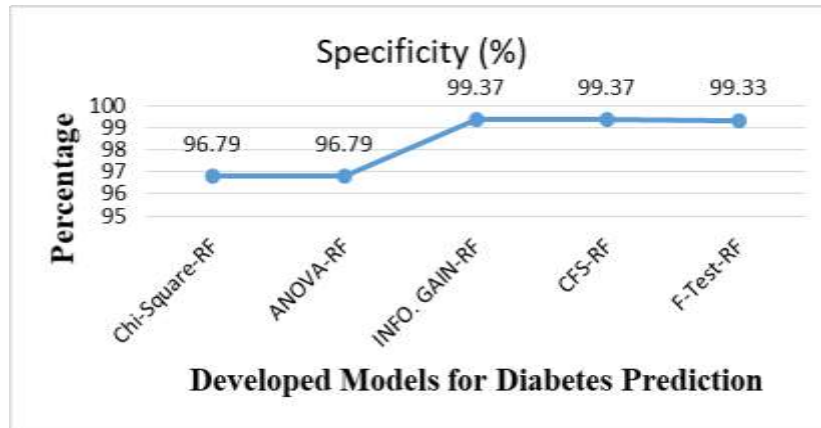
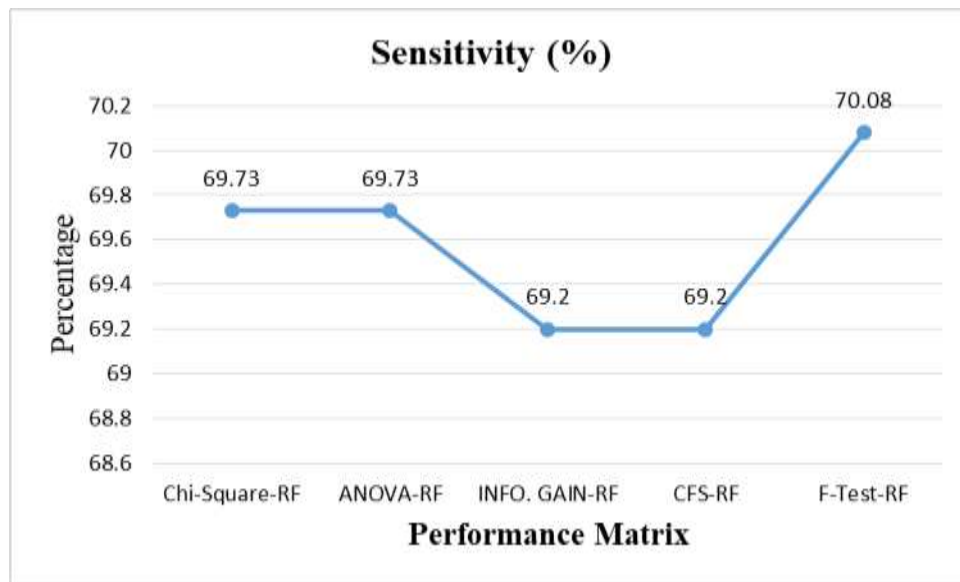


Figure 4.7 gives a graphical representation of comparative analysis of the five different techniques employed in this study which were Chi-Square, ANOVA, Correlation-Based Feature Selection, Information Gain and F-Test. Result showed that the sensitivity obtained were 69.73%, 69.73%, 66.20%, 69.20% and 70.08% respectively. Hence in terms of sensitivity, F-Test-RF outperformed other models hence based on sensitivity F-Test-RF can be recommended for diabetes prediction process.

Figure 4.7
Comparative analysis of the different filter techniques based on sensitivity



CONCLUSION

In this study, different filter techniques were employed for diabetic prediction process. The developed models were used to achieve a high success rate in diabetic prediction process, the models followed preprocessing, feature selection and classification technique of a data mining process. Standard scaler technique was employed for preprocessing stage. Chi-Square, ANOVA, Correlation-Based Feature Selection, Information Gain and F-Test were used as filter selection techniques. Optimal feature subset (risk factors) necessary for diabetes detection were partitioned into training and testing set of 75% and 25% respectively. Optimal feature subset was used to build a random forest model for classification purpose. Results showed that F-Test-RF outperformed the other model with

an accuracy level of 96.83% followed by the INFO. GAIN-RF and CFS-RF with an accuracy of 96.80%, Chi-Square-RF and ANOVA-RF had a competitive performance of 96.79% in terms of precision INFO. GAIN-RF and CFS-RF performed better with a value of 91.13%, followed by the F-Test-RF model with 90.68%, Chi-Square-RF and ANOVA-RF had a precision of 90.50% in each model. In terms of Specificity the performance of INFO. GAIN-RF and CFS-RF outperformed the other models with a value of 99.37%, Chi-Square-RF and ANOVA-R had 96.79% in each case. While F-Test-RF had 99.33%. In terms of Sensitivity F-Test-RF outperformed the other models with a value of 70.08%. This is followed by the Chi-Square-RF and ANOVA-RF models with a sensitivity of 69.73%. result showed that the performance of all the model were highly impressive and were closely related in their performance, F-test-Rf can be recommended for diabetic detection process since it outperformed the other model in terms of accuracy and sensitivity. Results in all the models shows confidence in its knowledge discovery from the dataset used. The developed model proved to be efficient for diabetic detection process. However other technique can be employed to enhanced the model as well as providing a real time system for consultation purpose.

RECOMMENDATION(s)

It is highly recommended that filter techniques be employed for feature selection, and that only the prominent features be obtained from vast features available for prognosis. Best feature obtained will subsequently be used to build an efficient classification model. Hence this model can be recommended in the diabetic center to aid medical practitioners for efficient medical delivery.

Further Research Direction(s)

Some promising areas future researchers can investigate include: A comparative analysis of filter and wrapper-based approach for diabetes detection; ensemble approach for diabetes detection using embedded approach; and optimized support vector machine (SVM) kernels comparison-based approach for diabetes prediction.

REFERENCES

- Abdollahi, J. & Nouri-Moghaddam, B. (2022). Hybrid stacked ensemble combined with genetic algorithms for diabetes prediction. *Iran Journal of Computer Science*, 5(3), 205–220. <https://doi.org/10.1007/s42044-022-00100>.
- Abdollahi, J., & Nouri-Moghaddam, B. (2022). Hybrid stacked ensemble combined with genetic algorithms for diabetes prediction. *Iran Journal of Computer Science*, 5(3), 205–220. <https://doi.org/10.1007/s42044-022-00100-1>
- Abe, O. S., Obe, O. O., Boyinbode, O. K. & Biodun, O. N. (2021). Classifier algorithms and ensemble models for diabetes mellitus prediction: A review. *International Journal of advanced Trends in Computer Science and Engineering*, 10(1). ISSN 2278-3091. <http://www.warse.org/IJATCSE/static/pdf/file/ijatcse641012021.pdf>
<https://doi.org/10.30534/ijatcse/2021/641012021>
- Aggarwal K. (2021). Comparison of feature selection techniques for improved diabetes prediction using random forest. *International Journal of Mechanical Engineering*, 6(3). ISSN: 0974-5823.
- Alhussan, A. A., Abdelhamid, A. A., Towfek, S.K., Ibrahim, A., Eid, M. M., Khafaga, D. S., Saraya, M. S. (2023). Classification of diabetes using feature selection and hybrid Al-birunearth radius and dipper throated optimization. *Diagnostics*, 13, 2-40, 2038. <https://doi.org/10.3390/diagnostics13122038>
- Almutairi, E. S. & Abbod, M. F. (2023). Machine learning methods for diabetes prevalence classification in Saudi Arabia. *Modelling* 2023, 4, 37–55. <https://doi.org/10.3390/>
- Altaher, A., & Malebary, S. J. (2022). A hybrid meta-classifier of fuzzy clustering and logistic regression for diabetes prediction. *Computers, Materials & Continua*, 71(3), 6089–6105. <https://doi.org/10.32604/cmc.2022.023848>
- Anggoro, D. A., & Permatasari, D. (2023). Performance comparison of the kernels of support vector machine algorithm for diabetes mellitus classification. *International Journal of Advanced Computer Science and Applications (IJACSA)*, 14(1), 580-585
- Aslam, N., Khan, I. U., Alkhalifah, S., AL-Sadiq, S. A., Bugararah, S. W., AL-Otobi, M. A., & AL-Odinie, Z. M. (2021). Predicting diabetic patient hospital readmission using optimized random forest and firefly. *Evolutionary Algorithm*, 11(5), 2088-5334.
- Baratloo A., Hosseini, M., Negida, A. & Ashal, G. E. (2015). Part 1: Simple definition and calculation of accuracy, sensitivity and specificity. This open-access article distributed under the terms of the Creative Commons Attribution Non-commercial, 3.0 License (CC BY-NC 3.0). Copyright © 2015 Shahid Beheshti University of Medical Sciences. All rights reserved. www.jemerg.com . *Emergency*, 3(2), 48-49.

- Ghabousiana, R., Farhang, Y., Majidzadeha, K., & Sangarha, A. B. (2022). Hybrid of particle swarm optimization algorithm and fuzzy system for diabetes diagnosis. *International Journal Nonlinear Anal. Appl.* In Press, 1–9 ISSN: 2008-6822 (electronic). <http://dx.doi.org/10.22075/ijnnaa.2022.29575.4196>
- Jader, R. & Sadegh, A. (2022). Fast and accurate artificial neural network model for diabetes recognition. *NeuroQuantology*, 20(10), 2187-2196. Doi: 10.14704/nq.2022.20.10.NQ55189eISSN.
- Manivannan, M. B. T., Thanuja K. M., Janani, H. G., & Poojith, R. (2023). Machine learning for diabetes prediction. *International Research Journal of Modernization in Engineering Technology and Science*, 5(3) 2582-5208, e-ISSN: Impact Factor- 7.868
- Manivannan, M., Balaji, T., Thanuja K. M. Janani, Himanth G. D. and Poojith, R. (2023). Machine learning for diabetes prediction. *International Research Journal of Modernization in Engineering Technology and Science*, 5(3), 2582-5208. e-ISSN: Impact Factor- 7.868
- Mohiddin, S. K., Kousar, H., Sharon, P., Krishna, V. S., & Anupriya, S. (2022). An approach for early prediction of diabetes using firefly optimization algorithm. *International Journal of Food And Nutritional Sciences (IJFANS)*, 11(12), 1718-1727
- Saxena, R., Sharma, S. K., Gupta, M., & Sampada, G. C. (2022). A novel approach for feature selection and classification of diabetes mellitus: Machine learning methods. *Hindawi Computational Intelligence and Neuroscience*, 2022, 1-11. Article ID3820360.
- Sivaranjani, S, Ananya, S., Aravinth, J., & Karthika, R. (2021). Diabetes prediction using machine learning algorithms with feature selection and dimensionality reduction. 7th International Conference on Advanced Computing and Communication Systems (ICACCS) 978-1-6654-0521-8/20/\$31.00 ©2021 IEEE | DOI: 10.1109/ICACCS51430.2021.9441935 P.141-148
- Vijiya, K. K. (2019). Random forest algorithm for the prediction of diabetes. *Proceeding of International Conference on Systems Computation Automation and Networking*. @IEEE 978-1-7281-1524-5 @IEEE
- Zhu, W., Zeng, N., & Wang, N. (2017). Sensitivity, specificity, accuracy, associated confidence interval and roc analysis with practical SAS® implementations.
- Le, T. M., Vo, T. M., Pham, T. N., & Dao, S. V. T. (2021). A novel wrapper-based feature selection for early diabetes prediction enhanced with a metaheuristic. *IEEE Engineering In Medicine And Biology Society Section*, 9, 7869 – 7883.