Mathematical Statistician and Engineering Applications
ISSN: 2094-0343
2326-9865

Accurate Diabetic Prediction and Rank Prioritized Weight Improvised Voting Classifiers with Adaboosted Random Forest Algorithm

J. Revathy[1] and Dr. D. Selvanayagi[2]

[1] Research scholar, Department of Computer Science, Vellalar College for women, Erode, Tamil Nadu, India- 638012

[1]revamsc12@gmail.com

[2] Assistant professor, Department of Computer Applications, Vellalar College for women, Erode,
Tamil Nadu, India- 638012

[2]selvasubhika@gmail.com

Article Info
Page Number: 1764-1771
Publication Issue:
Vol. 71 No. 4 (2022)

Article History
Article Received: 25 March 2022
Revised: 30 April 2022
Accepted: 15 June 2022
Publication: 19 August 2022

Abstract
Effective diabetes categorization is essential for identifying a person's severe health status. This would facilitate taking prompt action on the pertinent Diabetes-related concerns. This would facilitate the prompt resolution of the pertinent Diabetes-related concerns. This study examines performance indicators for several categorization techniques, including Support Vector Machine (SVM) , Random Forest (RF) and K Nearest Neighbor (KNN). The major objective of this study is to develop a model that can correctly forecast a patient's likelihood of developing diabetes. The availability of a vast amount of duplicated, disorganised medical data presents another difficult implementation challenge. This experiment must be used using the resources at hand to identify diabetes early. We have used three distinct data mining methods in this research project. The experiments used the Pima Indians Diabetes Database (PIDD), which is sourced from the UCI machine learning repository, to test the performance of the Modified Multi Class K+KNN (MMVK+KNN), Radial basis Kernelized SVM classifier with PCA, and finally Rank prioritised weight improvised voting classifiers with adaboosted Random Forest Algorithm, RPWIVC RFA, which achieved 72 percent higher accuracy.

Keywords:- Diabetes prediction, Decision Tree, KNN, Modified Multi Class K+KNN (MMCK+KNN), Principle Component Analysis, PCA Support Vector Machine, SVM Adaboosted Random Forest, Rank prioritized weight improvised voting classifiers with adaboosted Random Forest Algorithm, RPWIVC_RFA

1 Introduction

Diabetes is a prolonged disease that disturb and changes the normal blood glucose level in the human body. Pancreatic cells alpha and beta controls the level of blood glucose which are regulated by two major hormones insulin and glucagon. Former decreases the level of blood glucose whereas later increases the blood glucose level concentration. The abnormal changes in the level of this hormone secretion pave the way for diabetes. Regrettably, approximately 1.5 million people are losing their lives due to this chronic disease every year [4].

Predictions of this deadly dangerous disease at early stages are very crucial to save the life of human beings. For accurate prediction of this diabetes and classifying the type of this disease,

collecting health information and generating data sorting and understanding reports from the large collection of data information is very tedious. Data mining techniques are much helpful in extracting the required information from enormous amount of large databases. Hence automatic prediction and classification algorithms will be much suitable for easy and earlier prediction of this chronic disease. To perform data generation for research purpose, there are surplus amount of various sensors and machines are been in use like Magnetic Resonance Imagery (MRI), digital microscopy, Computed Tomography (CT), mass spectrometry and the list goes on.

First section briefs us about the occurrence of diabetes and its importance of predicting the disease at early stages, introduction and advantages of data mining in data science in predicting diabetes. Second section deals with the significant methodologies and implementations that are done earlier[2]. Third section discusses the algorithmic implementation of Modified Multi Class K+ KNN Algorithm (MMCK+KNN), Radial Basis Kernelized functional SVM classifier with PCA (RBKFSVM_PCA) and Rank Prioritized Weight Improvised Voting Classifiers with Adaboosted Random Forest Algorithm (RPWIVCARFA). Fourth section talk over on the results and discussion of our analysis, findings on accuracy and other parameters and conclusion and future scope were done on fifth section.

2 Methodology

For Prediction of Diabetes we have chosen three most preferable algorithms like KNN, PCA and Random Forest Algorithms, using these we have proposed 3 new algorithms i.e., Modified Multi Class K+ KNN Algorithm (MMCK+KNN), Radial Basis Kernelized functional SVM classifier with PCA (RBKFSVM_PCA) and Rank Prioritized Weight Improvised Voting Classifiers with Adaboosted Random Forest Algorithm (RPWIVCARFA).

Modified Multi class K+ KNN Algorithm (MMCK+KNN)

The K Nearest Neighbor (KNN) algorithm is performed as part of the data mining technique to implement the extraction of relevant information by mining data transformation from large dataset. In other words, we can term it as prediction of new classes based on similarity measures from the original dataset. Here we have used distance as the measure of factor for predicting the classes[5]. In our modified multi class K+KNN we have performed based on the distance measure and the classes are categorized based on the K+ value. In regular KNN algorithm we choose a numerical value for K as the closest neighbor to the referenced data point which needs to be categorized. In the modified Multi class K+KNN algorithm, the K value is fixed with the help of a programmable distance variable that identify the nearest neighbor of the specific data point. The degree of closeness of the data point is calculated based on the overlapping attributes here in this case, say for example level of glycogen, blood sugar, patient age and other parametric points that corresponds to the classification of diabetes. The KNN algorithms choose a number k as the closest neighbor to the data point to be categorized. If k is set to 5, it will search for the 5 closest Neighbors to that data point.

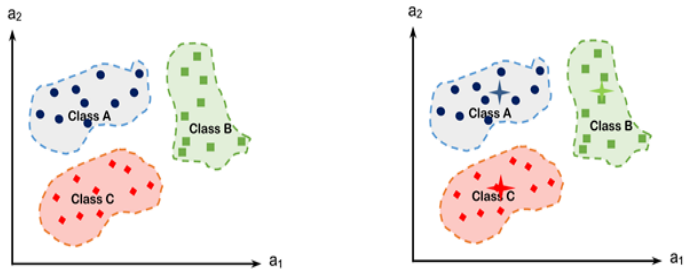

Fig. 4. a) Multi class training data image b) Class center point determination

If suppose k=4 in this case, KNN locates the four closest neighbors. In our modified multi class KNN algorithm, Figure 4 a shows the training data images that are classified as Class A, B and C. The proposed algorithm will initially identify the center point for each class and the Euclidean distance between the center point of each class to each data point of the class which is shown in figure 4b.

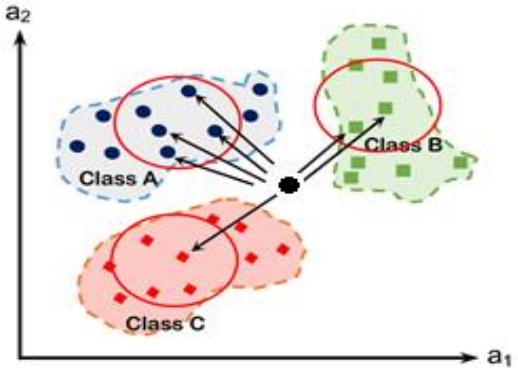

Fig. 5. Classification of target data point

Then average distance for each class will be determined to finally classify the target data point. This is how the K value will be fixed and the highest possible class will be considered further as to estimate the weighted average of its neighbor which is shown in flowchart below in figure 6. Using that k we have to classify the multi values. For these we can easily do classification, So, K has more numbers of nearest values. Using this way we can easily find the classification and can perform prediction

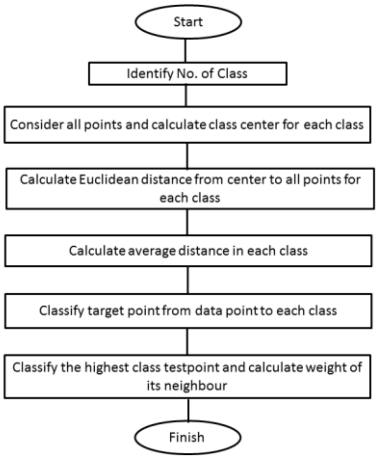

Fig. 6. Flowchart diagram for Modified Multiclass K+KNN

To identify K+ value, first we need to know the K value in KNN, which is the nearest neighbor of KNN, the below algorithm explains on finding the K+ value.

```
•  Import the diabetes dataset
•  Get the training and test model dataset
•  Make a fit the model to KNN :
        •  knn = KNeighborsClassifier(n_neighbors=i)
           ore=cross_val_score(knn,df_feat,data2['cluster'],cv=20)
           accuracy_rate.append(score.mean())

•  for(int i = 0; i<kplus.length;i++)
       for(int i = 0; i<kn.length;i++)
           kn.[i] = KNeighborsClassifier(n_neighbors=10)

•  print(confusion_matrix(y_test,pred))
•  print(classification_report(y_test,pred))
•  End
```

Fig. 7. Algorithm for Modified Multiclass K+KNN

Radial Basis Kernelized functional SVM classifier with PCA (RBKFSVM_PCA)

These radial basis Kernelized functional SVM classifier will be much powerful kernel, as they are much suitable even for complex dataset. The support vector points are the vector points closest to the hyper plane, because only these two points contribute to the algorithm's result, whereas the other points do not. Removing a data point that isn't a support vector has no influence on the model. The hyper plane's size is determined by the number of features. If there are only two input characteristics, the hyper plane is merely a line. The hyper plane becomes a two-dimensional plane when the number of input features reaches three. It becomes impossible to imagine, when the number of features exceeds three. In our dataset, we have filtered out the top features through PCA. Only two components are used since they aid in the visualization of the boundary in a display. It is even harder to imagine greater than 2 components and for decision function contours. set up the SVM classifier with a radial basis function kernel and the best fit values from grid-search cross-validation for the 'C' and 'gamma' parameters. Create contours with x, y (eventually the chosen principal components), and Z by defining a function (the decision function for SVM). That x and y here variable 1 and 2 can fit the SVM Model after the training and test method is converted.

Rank Prioritized Weight Improvised Voting Classifiers with Adaboosted Random Forest Algorithm (RPWIVCARFA).

This prediction method is a hybrid method of combining Voting classifiers. Adaboost and Random forest algorithm to efficiently improve the accuracy of the prediction and diagnosis system. The role of Adaboost algorithm is to transform a weak algorithm to a strong linear algorithm based on its repetitive iterations[3].

The Random forest algorithm executes the subsets in parallel which is the key difference between AdaBoost and random forest. This RFA executes special overfitting reduction with the help of a combination of weak learners. Both the boosting and nagging concepts combined together will help in efficiently predicting the diabetic disease and as the weight of the logic is improvised by including all features in the class as shown in figure 9. The training set will be sent through the classification models as the combination of new data and classification models will help in prediction of diagnostic. Then the predicted features will be assigned rank based on their weight and the voting classifiers help in selecting the majority feature predicted data and thus the accuracy and efficiency of the algorithm will also be much improved.

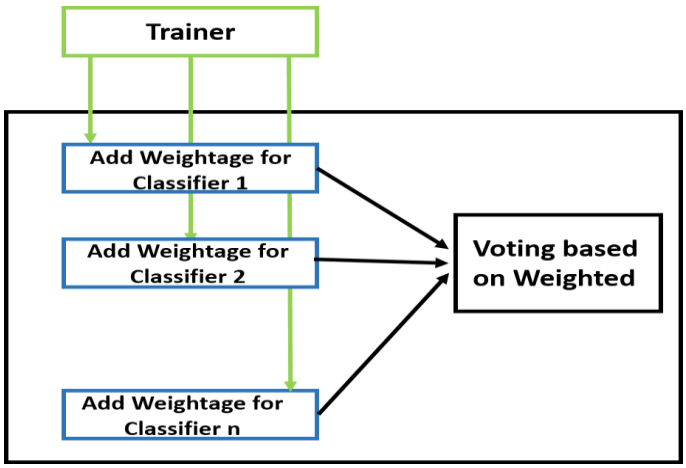

Fig. 9. Proposed RPWIVCARFA Implementation

3 Implementation Analysis

The algorithms discussed in section 3 are implemented and tested with The Pima Indians Diabetes Database (PIDD) which are obtain from the UCI machine learning repository. These three methods implement in python. The main objective of considering this dataset is to predict accurately whether the patient has diabetes or not. This dataset has some unique features where all the patients where female with minimum 21 years of age. The entire dataset includes various classes like pregnancies, Glucose, blood pressure, skin thickness, Insulin, Body mass index (BMI) age and the list goes on. The diagnosis happens based on the certain features on the dataset. The dataset graph used for the experimentation is shown in figure 11.

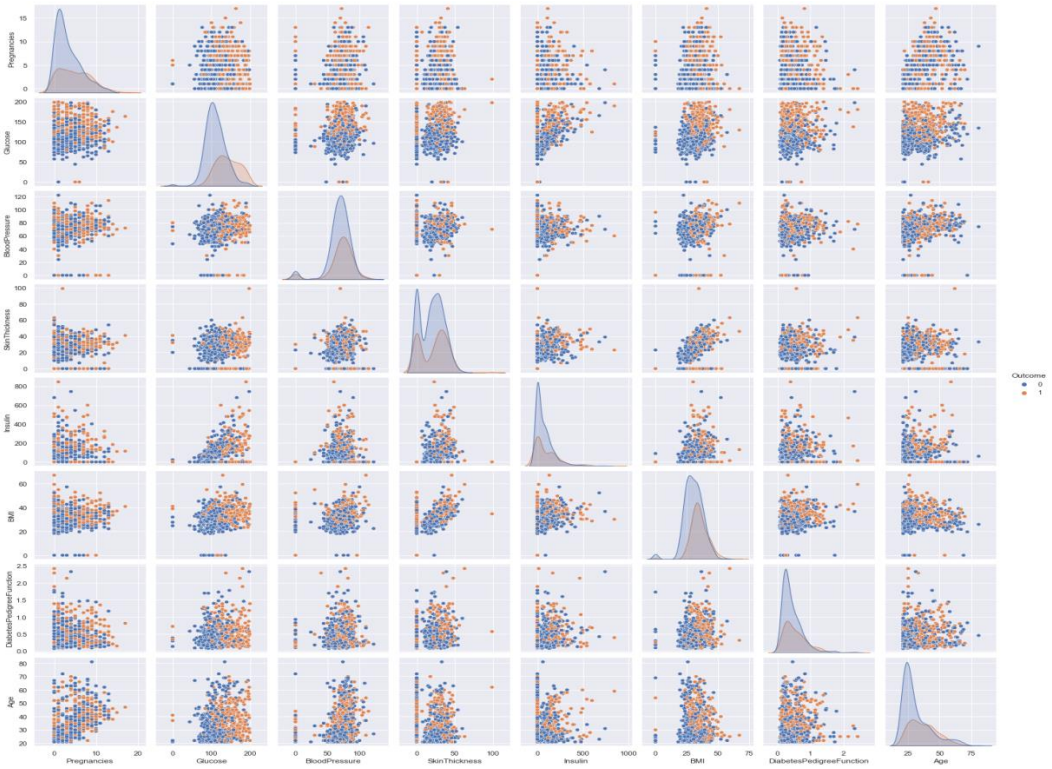

Fig. 11. Dataset graph used for experimental evaluation

The diagnosis of diabetes is detected by the level of glucose concentration along with the density concentration to determine the existence of diabetes is present or not. The experimental results are run with the help of RFA algorithm and the corresponding analysis of various classes are mined from the original database separately as shown in figure 12, in order to predict the database regularly and effectively

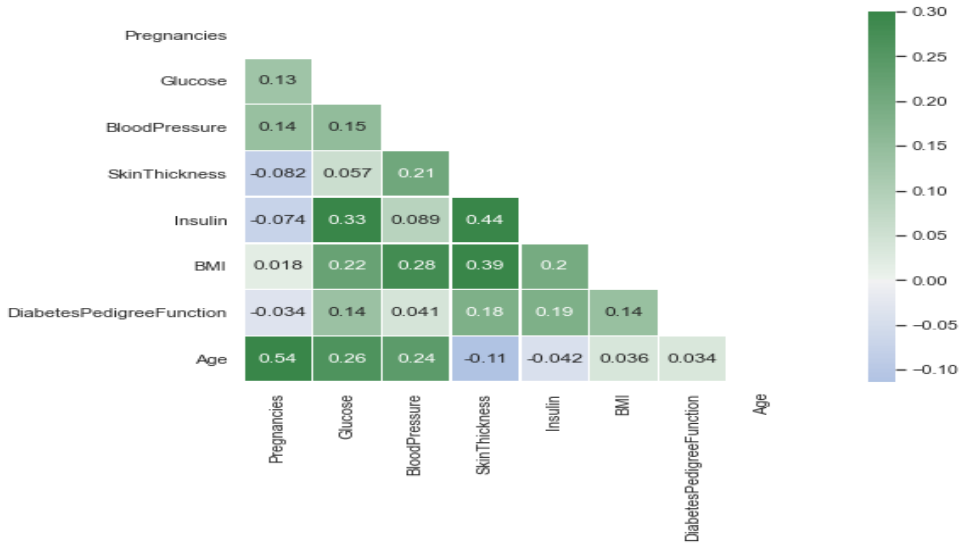

Fig. 12. Diabetic Data classification using proposed algorithm

The original dataset has enormous amount of data which are classified into various classes like pregnancies, Glucose, Blood pressure, Skin thickness, Insulin, BMI, Diabetes pedigree function and age. All these classes will be implemented with the help of all the three algorithms and the proposed rank prioritized weight improved voting classifiers with RFA algorithm shows better results in terms of accuracy, macro average and weighted average. The results obtained from the experimental results are tabulated in table 1 below and from the results it has been observed that prediction of diabetes can be done within stipulated time with greater performance metrics.

Table 1. Performance Metric analysis for diabetes prediction and classification

Algorithm	Accuracy	Macro Average	Weighted Average
KNN	73	72	72
MMCK+KNN	74	76	74
PCA	72	70	71
KFSVM_PCA	74	71	73
RFA	77.92	72	78
Proposed RPWIVCARFA	78.35	76	78

From the above table it is understood that the proposed Rank Prioritized Weight Improvised Voting Classifier based CA and Random Forest Algorithm (RPWIVCCARFA) has better improved performance in terms of accuracy and other parameters. The proposed work has achieved 71% of accuracy increment than KNN and 72% of accuracy increment than PCA and 70% accuracy increased than RFA which is clearly shown in figure 13 below.

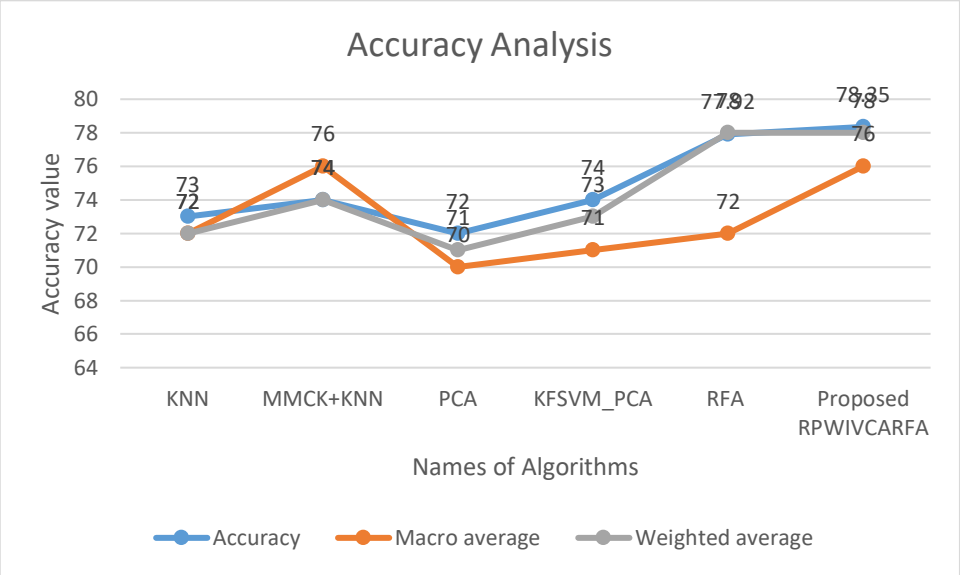

Fig. 13. Performance analysis chart for diabetes prediction and classification

Similarly the macro average also the proposed algorithm tops the table with 71% increment than KNN, 72% increment than PCA and 72% increment than existing RFA. Also the weighted average also tops the table for our proposed algorithm with similar 72% increment than the existing algorithms.

5 Conclusion

Early prediction of this disease though it doesn't cure the disease, the adverse effects can be avoided. Hence the early prediction and classification of this disease has become highly essential in the world where humans run behind technologies. For automatic diagnosis, prediction and classification from the large repository of databases, it is essential to extract only the relevant information for performing classification in less amount of time. For this we have implemented data mining techniques with modified versions on KNN, PCA and RFA and tested the performance using The Pima Indian Diabetics Database (PIDD). In this the proposed Rank Prioritized Weight Improved Voting Classifier based Principal component Analysis has achieved better performance of about 72% increment in terms of accuracy, weighted average and macro average. This algorithm can be much suitable for other datasets as well and the future enhancement of this work aims at further improvising the algorithm to improve accuracy percentage.

References

1. Hasan, M. K., Alam, M. A., Das, D., Hossain, E., & Hasan, M. (2020). Diabetes prediction using ensembling of different machine learning classifiers. IEEE Access, 8, 76516-76531.

2. S. Patikar, P. Saha, S. Neogy and C. Chowdhury, "An Approach towards prediction of Diabetes using Modified Fuzzy K Nearest Neighbor," 2020 IEEE International Conference on Computing, Power and Communication Technologies (GUCON), 2020, pp. 73-76, doi: 10.1109/GUCON48875.2020.9231066.
3. Rezaei Mohammadi, Zahra & Alizadeh, Hosein & Parvin, S. & Alinejad-Rokny, Hamid. (2012). 2.10. An extended MKNN modified K-nearest neighbor. ournal of Networking Technology. 2. 162-168.
4. https://www.who.int/health-topics/diabetes#tab=tab_1
5. J. M. Norris, R. K. Johnson and L. C. Stene, "Type 1 diabetes—Early life origins and changing epidemiology", Lancet Diabetes Endocrinol., vol. 8, no. 3, pp. 226-238, Mar. 2020.