# PREDICTING DIABETES MELLITUS USING FEATURE SELECTION AND CLASSIFICATION TECHNIQUES IN MACHINE LEARNING ALGORITHMS

*Ambily Merlin Kuruvilla* [1] *Dr.N.V.Balaji* [2]

**ABSTRACT**

Diabetes is a disease that is now spreading like an epidemic around the globe. Diabetics is a chronic disease that occurs when the blood sugar or glucose in the body is not controlled or broken down. It may be caused either when the body does not react to the insulin produced naturally in the body or when the produced insulin is insufficient. The latest WHO statistics points diabetics as a life-threatening disease condition with an estimated 1.6 million deaths worldwide. The word diabetics mellitus is of Greek origin that means 'to pass through honey or sweet'.

Constant high blood sugar in blood stream termed hyperglycemia is a serious condition that can adversely affect the health of an individual. A patient may experience loss of energy with fatigue and brokenness. Uncontrolled levels threaten body organs which include kidneys, heart, eyes and nervous system. Taking into account the widespread nature of the disease, finding a cure using latest computer advancements has been a topic of study for many researchers and scientists worldwide. This research focuses on creating a forecast or a prediction algorithm that can sort out an optimal classifier. The optimal classifier must be able to deliver near close results to real world clinical outcomes when it is juxtaposed to a validity of its accuracy. Sorting out attributes that trouble early detection of the disease is the objective of the study.

The dataset used for the prediction is the PIMA Indian dataset. Naïve Bayesian, J48, Random tree, random forest and SMO are the algorithms used for this research. The conclusions and findings of this work extend to feature selection mechanism for improving classification accuracy. The outcomes of Naïve Bayesian and SMO algorithms prove themselves to be the best for the purpose. PIMA Indian dataset is used for the prediction.

**Keywords:** WHO, World health organization, Naïve Bayesian, J48, Random forest and SMO, Multilayer Perception

## I INTRODUCTION

As per a recent analysis of World Health Association, around 442 million individuals are diagnosed with diabetics every year. Diabetes mellitus is a chronic disease that indicating a high sugar level in the blood stream caused by the inefficient functioning of the pancreatic beta cells. A person suffering from diabetics is prone to various health risks namely pancreas glitch, heart diseases, blood pressure, kidney failure and risks to other sensitive organs of the body.

Like any other disease, early prediction is the key to controlling and balancing the effects of diabetes. The utilization of machine learning and its application methods deliver efficient results to excerpt useful information by excogitation of prediction models from medical diagnostic datasets that are collected from a diverse group of diabetic patients. Selected information from these datasets can be useful to predict and analyze diabetic sufferers. The tools of machine learning have the ability to predict diabetes mellitus. However, the constraints lie in the ability to select the best technique in machine learning to predict based on such attributes. Therefore, in this work four different classification algorithms are used for the analysis and prediction of diabetes.

[1]Research Scholar ,Department of CS,CA & IT
Karpagam Academy of Higher Education, Coimbator

[2]Research Supervisor, Department of CS,CA & IT
Karpagam Academy of Higher Education, Coimbatore

## II RELATED STUDIES

Six different [4] classification tools along with PIMA Indian diabetes dataset is used for the prediction. WEKA tool is used for the analysis and it is found that MLP is showing better performance.

## III IMPLEMENTATION METHODS

### *A. DATASET DESCRIPTION*

The dataset used to study is gathered from UCI repository (PIMA Indian Dataset). It contains attributes such as age, sex, body mass index, etc. It includes test results of both diabetic and non-diabetic patients. To form the dataset, HbA1c, FBG and PMBG test results from patients are used. According to the latest test reports of diabetic patients, the identification of attributes can be done and various parameters such as Age, Body Mass Index, HbA1, etc. are included.

### *B. DATA PREPROCESSING AND FEATURE SELECTION*

Feature selection is the method where the features that contribute most to your prediction variable or output you are interested in are automatically or manually selected. The dataset bearing non-essential features can result in the model losing accuracy and making it depend on immaterial features. In this, ChiSquareAttributeEval is used for feature selection. From the dataset, 8 attributes are selected. These attributes were used for the prediction. During the Preprocessing stage missing and incorrect values are replaced with the mean and the median.

### C. APPLYING MACHINE LEARNING TECHNIQUES

#### *1)* NAÏVE BAYES

In NAVIE BAYES between the predictors, a probabilistic classifier from the Bayes Theorem is implemented with independent assumption between the predictors. Naïve Bayesian approach uses Bayes Theorem as the input in the dataset, conducts analysis and predicts the category label. It calculates a class probability in the input data which is useful for predicting the unrevealed data sample class.

#### *2)* RANDOM FOREST

Random Forest is a supervised learning method that is used for both classification and Regression. The scheme behind the random forest is that it acts as a bagging technique used to create random sample features. The random forest functions as a bagging strategy to establish random sample characteristics. The distinctive feature between the decision tree and the random forest is that the method to scan the root node and split the function node would run randomly in Random Forests.

#### *3) J48*

J48 is an algorithm that is a supervised learning method. J48 helps in classification by allowing a decision tree. The decision trees rendered by this algorithm can be used for classification. Decision tree is a method that continuously divides the given dataset into two or more sample data. The aim of this method is to predict the class value of the target variable.

#### *4)* MULTILAYER PERCPTION

A multilayer perception (MLP) falls within an artificial neural network feed forward class. The supervised learning method used by MLP for practice is back propagation (BP). BP is a supervised learning technique that MLP utilizes for training. MLP can differentiate data that are not linearly separable. The multiple layers and non-linear activation set MLP apart from a linear perception.

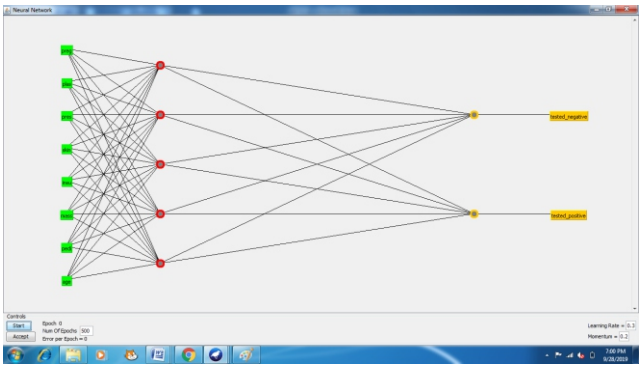

*Fig 1 : Result Of Multilayer Perception algorithm with PIMA Indian dataset.*

## IV EXPERIMENTAL RESULTS

For doing the comparison, sensitivity and specificity metrics are observed for various machine learning algorithms. Percentage of correctly classified tuples by a classifier will determine the accuracy of the classifier. Here Kappa statistics is also used for the performance measurement. Kappa statistics is used as a metric and used to compare the observed accuracy with the expected accuracy. So, here Kappa statistics is used not only to evaluate a single classifier but to compare various evaluators. Accuracy is measured by using the following formulae:

$$\text{Accuracy} = \frac{(\text{True Positive} + \text{True Negative})}{(\text{True Positive} + \text{True Negative} + \text{False Positives} + \text{False Negatives})}$$

***Table: 1 Results of the Classification Performance Analysis.***

| Algorithm | CCI | ICCI | RMASE | Test Options |
|---|---|---|---|---|
| SMO | 594 | 174 | 0.476 | 10FoldsCross Validation |
| SMO | 123 | 31 | 0.4487 | Percentage Split |
| Navie Bayes | 586 | 182 | 0.4168 | 10 Folds Cross Validation |
| Navie Bayes | 119 | 35 | 0.3927 | Percentage Split |
| Multilayer Perception | 579 | 189 | 0.4215 | 10 Folds Cross Validation |
| Multilayer Perception | 114 | 40 | 0.4071 | Percentage Split |
| J48 | 567 | 201 | 0.4463 | 10 Folds Cross Validation |
| J48 | 117 | 37 | 0.43 | Percentage Split |
| Random Forest | 568 | 200 | 69.3575 | 10 Folds Cross Validation |
| Random Forest | 117 | 37 | 68.1159 | Percentage Split |

*CCI: Correctly Classified instances,*
*ICCI: Incorrectly Classfied Instances,*
*RMASE: Root Mean Absolute Square Error*

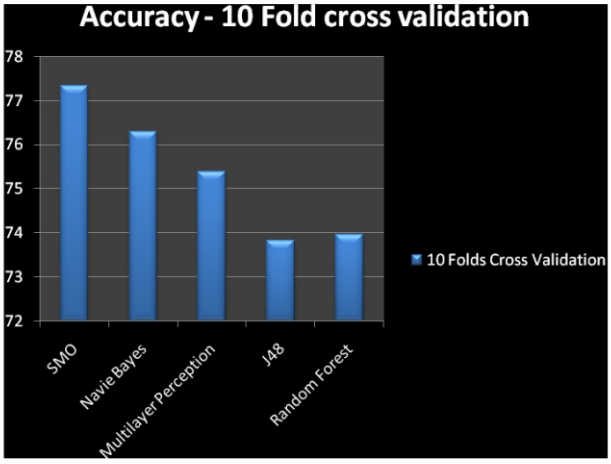


***Fig.2: Result of 10-Fold cross validation***

***Table2: Results of the Classification Performance Analysis.***

| Algorithm | Kappa Statistics | Accuracy | Test Options |
|---|---|---|---|
| SMO | 0.4682 | 77.3438 | 10 Folds Cross Validation |
| SMO | 0.5007 | 79.8701 | Percentage Split |
| Navie Bayes | 0.4664 | 76.3021 | 10 Folds Cross Validation |
| Navie Bayes | 0.4675 | 77.2727 | Percentage Split |
| Multilayer Perception | 0.4484 | 75.3906 | 10 Folds Cross Validation |
| Multilayer Perception | 0.3741 | 74.026 | Percentage Split |
| J48 | 0.4164 | 73.8281 | 10 Folds Cross Validation |
| J48 | 0.4493 | 75.974 | Percentage Split |
| Random Forest | 0.4052 | 73.9583 | 10 Folds Cross Validation |
| Random Forest | 0.4371 | 75.974 | Percentage Split |

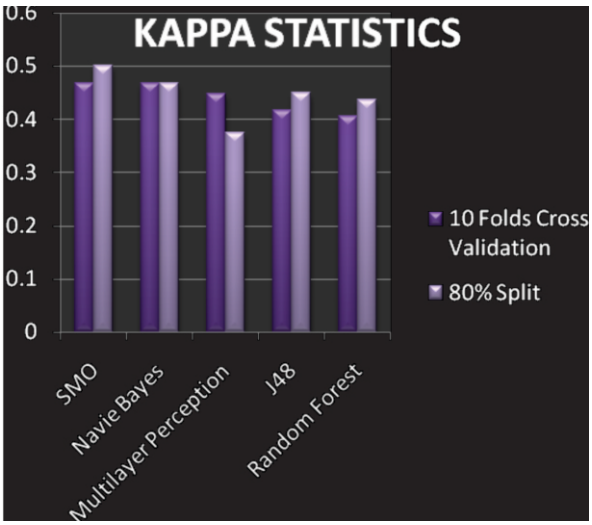


***Fig.3: Evaluation with Kappa Statistics for analyzing the performance of various classifiers in diabetic detection.***

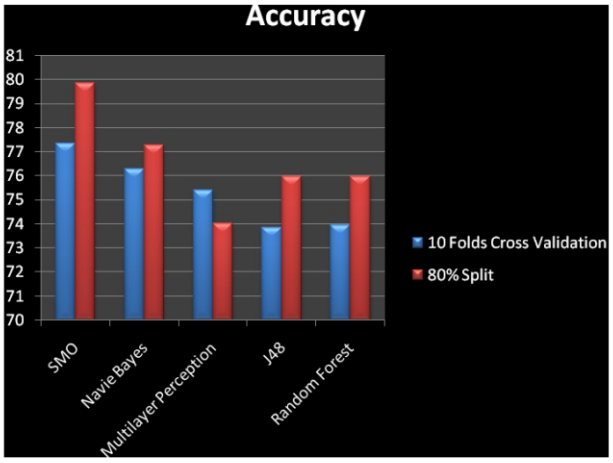


*Fig.4: Predicting the accuracy of various classifying algorithms using PIMA Indian dataset*

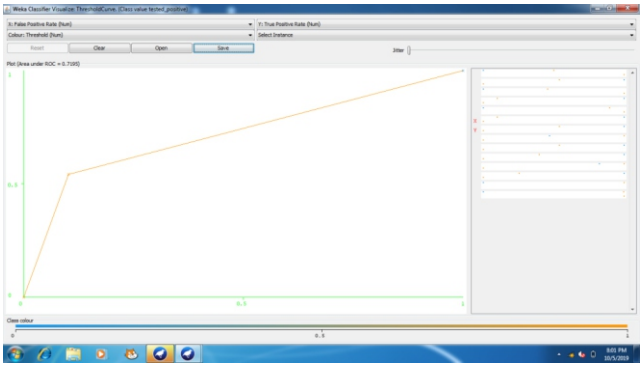

*Fig.5: Threshold curve of SMO algorithm*

*Table3: Detailed accuracy report of SMO Algorithm*

| | |
|---|---|
| TP Rate | 0.799 |
| FP Rate | 0.333 |
| Precision | 0.793 |
| Recall | 0.799 |
| F Measure | 0.789 |
| ROC Area | 0.733 |

Fig.2 and Fig.4 show the results of various machine learning algorithms such as SMO, Navie Bayes, Multilayer Perception, J48 and Random Forest. From these graphs we can find that results of SMO algorithm are better than those of other classifiers to predict diabetes mellitus. To calculate the accuracy True Positive Rate, False Positive Rate, F Measure, Recall, Precision and ROC curve measures are also used. From the above parameters it is observed that the accuracy of SMO algorithm is better than that of other algorithms

## CONCLUSION

In diabetes treatment detection of disease in the early stage is the key for treatment. In this work various machine learning approaches are used for predicting diabetes disease. SMO, Navie Bayes, Multilayer Perception, J48 and Random Forest algorithms were used for the prediction. Here diabetes diagnosis problem is investigated in terms of the accuracy of various classification algorithms.

In this modern day of technology and convenience, people don't bother to take care of their most precious wealth which is one's physical and mental health. This ignorance has resulted in the rise of chronic disease like diabetes. In this study various classification techniques were used for the analysis.

## REFERENCES

1. Kalaiselvi C and G. M. Nasira, "A New Approach for Diagnosis of Diabetes and Prediction of Cancer Using ANFIS", 2014 World Congress on Computing and Communication Technologies
2. Global report on diabetes by World Health Organisation. 2016, ISBN 978 92 4 156525 7.
3. VeenaVijayan V and Anjali C Prediction, "Diagnosis of diabetes mellitus—a machine learning approach", Recent Adv. 2015, https://doi.org/10.1109/raics.2015.7488400.
4. Kavakiotis I, Tsave O, Salifoglou A, Maglaveras N, Vlahavas I, Chouvarda I. Machine learning and data mining methods in diabetes research. Comput Struct Biotechnol J. 2017;15:104–16.
5. Kalaiselvi C and G. M. Nasira "Classification and Prediction of Heart Disease from Diabetes Patients using Hybrid Particle Swarm Optimization and Library Support Vector Machine Algorithm" Internatinal Journal of Computing Algorithm , 2015.
6. J. W. Smith, J. E. Everhart, W. C. Dickson, W. C. Knowler, and R. S. Johannes, "Using the ADAP Learning Algorithm to Forecast the Onset of Diabetes Mellitus," Proc. Annu Symp. Comput.

Appl. Med. Care, pp. 261–265, 1988.

7. B. M. K Prasad, K. K Singh, N. Ruhil, K. Singh, and R. O'Kennedy, "Communication and Computing Systems", Proceedings of the International Conference on Communication and Computing Systems (ICCCS 2016), Gurgaon, India, 9-11 September, 2016. CRC Press, 2017
8. H. Wu, S. Yang, Z. Huang, J. He, and X. Wang, "Type 2 diabetes mellitus prediction model based on data mining," Informatics Med. Unlocked, vol. 10, pp. 100–107, Jan. 2018.
9. D. M. Renuka and J. M. Shyla, "Analysis of Various Data Mining Techniques to Predict Diabetes Mellitus", Int. J. Appl. Eng. Res. ISSN, vol. 11, no. 1, pp. 973–4562, 2016.
10. A. Jindal, A. Dua, N. Kumar, A. K. Das, A. V. Vasilakos, and J. J. P. C. Rodrigues, "Providing Healthcare-as-a-Service Using Fuzzy Rule-Based Big Data Analytics in Cloud Computing", IEEE J. Biomed. Heal. Informatics, pp. 1–1, 2018.
11. Rahul Karthik Sivagaminathan, Sreeram Ramakrishnan (2007), "A hybrid approach for feature subset selection using neural networks and ant colony optimization", Expert Systems with Applications. Vol. 33. pp. 49-60.
12. Liangjun Ke, Zuren Feng, Zhigang Ren, "An efficient ant colony optimization approach to attribute reduction in rough set theory".
13. A.Sheik Abdullah, R.Suganya, S.Rajaram, V.Rajendran, "A Novel Feature Selection Method for Predicting Heart Diseases with Data Mining Techniques", Asian Journal of Information Technology, 2016. DOI: 10.3923/ajit.2016.1314.1321.
14. Rana Forsati, Alireza Moayedikia and A.Keikha 2012, "A Novel Approach for Feature Selection based on the Bee Colony Optimization", International Journal of Computer Applications, Vol.43, pp. 1-40. Esposito F., Malerba D., Semeraro G. and Kay J.
15. "A comparative analysis of methods for pruning decision trees", IEEE Transactions on Pattern Analysis and Machine Intelligence, 19 (1997), pp. 476-491doi:10.1109/34.589207. Fatima M and Pasha M. "Survey of Machine Learning Algorithms for Disease Diagnostic", Journal of Intelligent Learning Systems and Applications, 09 (2017), pp. 1-16 doi:10.4236/jilsa.2017.91001.
16. Han, J., Rodriguez, J.C and Beheshti M, "Discovering decision tree based diabetes prediction model", in: International Conference on Advanced Software Engineering and Its Applications, Springer. pp. 99-109.
17. Aishwarya R., Gayathri P and Jaisankar N, "A Method for Classification Using Machine Learning Technique for Diabetes", International Journal of Engineering and Technology (IJET), 5 (2013), pp. 2903-2908.
18. Arora and R. Suman, "Comparative Analysis of Classification Algorithms on different datasets using WEKA", International Journal of Computer Applications, 54 (2012),pp. 21-25 i:10.5120/8626-2492.
19. "Prediction and Diagnosis of Diabetes Mellitus A Machine Learning Approach", 2015 IEEE Recent Advances in Intelligent Computational Systems (RAICS) (2015), pp. 122-127 doi:10.1109/ RAICS .2015.7488400.
20. "An Empirical Study of the Naive Bayes Classifier IJCAI 2001 workshop on empirical methods in artificial intelligence", IBM (2001), pp. 41-46
21. Priyam A, Gupta R, Rathee A and Srivastava S "Comparative Analysis of Decision Tree Classification Algorithms" International Journal of Current Engineering and Technology" Vol., 3 (2013), pp. 334-337 doi:June 2013, arXiv:ISSN 2277-4106.