# ENSEMBLING OF DIABETES PRETICTION USING MACHINE LEARNING CLASSIFIERS

[1] S. Balu, [2]Shobitha.S, [3]Rathna Priya. K, [4]Rivitha.P, [5]Ravulapalli lakshmi
[1] professor, [2,3,4,5]UG Scholar
Department of Computer Science
Vivekanandha College of Engineering ForWomen, Namakkal,India
[1]balu@vcew.ac.in, [2]shobitha2801@gmail.com, [3]rathnapriyak@gmail.com
[4]rivithapalani@gmail.com, [5]susmitharavulapalli2000@gmail.com

## ABSTRACT

Diabetes mellitus is affecting an increasing number of families as its prevalence rises. Prior to diagnosis, most diabetics have limited knowledge of their health status or the risk factors they face. In this work, we suggested a novel model for predicting type 2 diabetes mellitus based on data mining approaches (T2DM). The primary issues we are attempting to address are increasing the accuracy of the prediction model and making the model adaptable to more than one dataset. The model is made up of MLP, NB and RF method and a set of pre-processing techniques. To compare our findings to those of other studies, we used the Pima Indians Diabetes Dataset and the Waikato Environment for Knowledge Analysis toolbox. The results suggest that the model has a 3.04% greater prediction accuracy than previous studies. Furthermore, our approach assures that the dataset quality is enough. We used our algorithm to two more diabetes datasets to further assess its performance. The results of both experiments reveal good performance. As a result, the model is demonstrated to be beneficial for realistic diabetic health management.

## 1. INTRODUCTION

### DIABETES CLASSIFIACITON:

Diabetes has been one of the most prevalent diseases in recent years, and its global incidence is continuously increasing. It is a catch-all phrase for a variety of metabolic disorders, the most common of which is chronic hyperglycaemia. Poor insulin secretion, impaired insulin action, or both are the causes. Diabetes' persistent hyperglycaemiais linked to long-term damage, malfunction, and failure of multiple organs, including the eyes, kidneys, nerves, heart, and blood vessels. The great majority of diabetes may be classified into two types: type 1 and type 2.

Type 1 diabetes is caused by a total lack of insulin secretion. Type 2 diabetes, on the other hand, is significantly more common, and the reason is a combination of insulin resistance and a poor compensatory insulin secretary response. Kind 2 diabetes is the most frequent type of diabetes. According to the sixth edition of the IDF (International Diabetes Federation) Diabetes Atlas, 382 million people are projected to have diabetes, with rapid rises recorded in nations all over the world, and Type 2 diabetes accounts for the vast majority of all diabetes.

### DATA MINING:

We live in a world where enormous volumes of data are collected on a daily basis. Manual data analysis is the conventional way of converting data into knowledge. This type of data processing is slow, costly, and subjective when data quantities expand fast. The old technique is becoming obsolete in many industries and cannot match the demands of data analysis. Data mining, also known as knowledge discovery in databases (KDD), may be able to fill this void by offering methods for extracting information from data. The practise of identifying intriguing patterns and knowledge from massive volumes of data is known as data mining. Databases, data warehouses, the Web, other information repositories, and data streamed into the system dynamically are examples of data sources. Data mining has been used in a number of disciplines during the last few decades, including marketing, finance (particularly investing), fraud detection, manufacturing, telecommunications, and many scientific domains, including medical data analysis. As the volume of medical data grows, there is an increasing need for fast data analysis to extract relevant, task-oriented information from massive amounts of data. This type of information might be useful in future medical decisions

RANDOM FOREST ALGORITHM: Random forest has approximately the same hyper parameters as decision trees and bagging classifiers. Random forest provides unpredictability to the model while the trees develop. When splitting a node, it looks for the best feature from a random group of characteristics rather than the most essential feature. Decision trees are a common way for performing a variety of machine learning tasks. Learning from trees ""It is invariant under scaling and numerous other transformations of feature values, is resilient to insertion of irrelevant features, and yields inspect able models," state Hastie et al. However, they are rarely correct ". Deep-grown trees, in particular, tend to acquire very irregular patterns: they overfat their training sets, resulting in low bias but very high variance. Random forests are a method of averaging numerous deep decision trees that have been trained on various regions of the same training set in order to reduce variation.

PCA

The quadratic programming (QP) problem that emerges during the training of support-vector machines is solved by the sequential minimum optimization (PCA) approach (SVM). The popular LIBSVM utility implements PCA, which is commonly, used for training support vector machines. The release of the PCA technique in has sparked a great deal of interest in the SVM field, as earlier approaches for SVM training were far more difficult and needed expensive third-party QP solvers. The "chunking algorithm" is what it's called. Starting with a random selection of the data, the algorithm solves the issue and iteratively adds cases that break the optimality constraints. One downside of this technique is that it requires solving QP-problems that scale with the number of SVs. PCA can be more than 1000 times quickerthan the chunking technique on real-world sparse data sets.

### 1.5 DATA WAREHOUSING

Data warehousing is a set of strategies, techniques, and tools that allow knowledge workers—senior managers, directors, managers, and analysts— conduct data analytics that aid in decision-making and the improvement of information resources. Data warehousing is a phenomena that arose as a result of the massive quantity of electronic data that has been saved in recent years, as well as the pressing need to use that data to achieve goals that go beyond the normal duties associated with everyday processing. A major organisation, in a typical situation, has several branches, and top management must measure and analyse how each branch contributes to the overall business performance. The corporate database contains complete information on the work performed by branches. To fulfil the demands of the managers, tailor-made queries may be issued to get the necessary data. To make this approach work, database administrators must first develop the appropriate query (usually an aggregate SQL query) after thoroughly researching database catalogues. The query is then processed. Because of the large amount of data, the query complexity, and the concurrent impacts of other normal workload queries on data, this might take many hours. Finally, a spreadsheet report is prepared and distributed to high management.

## 2. **LITERATURE REVIEW**

### CURRENT ISSUES AND GUIDELINES IN PREDICTIVE DATA MINING IN CLINICAL MEDICINE

In this paper, Ricardo B, Blaz Z, and colleagues offered Predictive data mining is quickly becoming an indispensable tool for medical researchers and clinicians. Understanding the fundamental concerns behind these methodologies, as well as following agreed-upon and defined protocols, is required for their implementation and distribution of results. Data mining is the act of choosing, examining, and modelling huge volumes of data in order to identify unknown patterns or correlations that offer the data analyst with a clear and meaningful conclusion. The phrase data mining, coined in the mid-1990s, has today become a synonym for 'Knowledge Discovery in Databases,' emphasising the data analysis process rather than the usage of specific analytical methodologies. Data mining issues are frequently tackled by combining techniques from computer science, such as multidimensional databases, machine learning, soft computing, anddata visualization, and statistics, such as hypothesis testing, clustering, classification, and regression algorithms.

## DIABETES MANAGEMENT THROUGH A MOBILE HEALTH CONSULTATION APPLICATION

Diabetes, according to Mechelle Gittens, Reco King, et al, is today one of the most serious dangers to human life, with an increase in the number of identified cases globally. Because type 2 diabetes accounts for the bulk of cases identified, this abrupt growth has been connected to changes in human lifestyle. Mobile health (m-health) technologies are being applied across the health industry to help people live better lives. The culture we picked for our study has a predominantly African-descent population and is in crisis, since it has one of the highest rates of diabetes and amputation in the world. A data acquisition module (DAM), a mobile phone, and a health data server are all part of the proposed system. The DAM collects data from the patient using a variety of sensors and feeds it to the mobile device through Bluetooth. When the readings reach the mobile phone, they are sent through an IP network (similar to the Internet) to a distant health data centre. Health care providers, who can then respond properly, may then view the readings. The technology monitors patients throughout the clock, which, according to the authors, might eliminate the necessity for face-to- face discussions between physicians and patients. This enables patients to obtain the care they require from the comfort of their own homes. Unlike much of the previous research, this technique can be deployed without the necessity for people to own a smartphone.

## A REVIEW OF DECISION SUPPORT SYSTEMS FOR PREDICTING DIABETES MELLITUS

In this study, VeenaVijayan V et al claimed that diabetes is induced by an increase in blood sugar levels. This can lead to a number of consequences, including renal failure, stroke, cancer, heart disease, and blindness. Early identification and diagnosis aid in identifying and avoiding these problems. A variety of computerised information systems were created to predict and diagnose diabetes using various classifiers. Choosing appropriate algorithms for categorization definitely improves the system's accuracy and efficiency. The primary goal of this work is to compare the benefits of various pre-processing strategies for diabetic decision support systems based on Support Vector Machine (SVM), Naive Bayes classifier, and Decision Tree. Computational approaches statistical methods, clustering, classification, pattern recognition, and transformation are all part of data mining. Medical data mining entails extracting hidden patterns from massive amounts of heterogeneous data, hence opening up a vast supply of medical data analysis.

## A SYSTEM BASED ON MACHINE LEARNING FOR PREDICTING DIABETES RISK USING MOBILE DEVICES

Diabetes mellitus (DM) is approaching potentially pandemic proportions in India, according to Ms. K Sowjanya et al Diabetes and its possible consequences cause immense sickness and devastation, resulting in a significant health-care cost on both households and society. The troubling element is that diabetes is now being connected to a number of problems and is appearing at a younger age in the country. Diabetes Mellitus (DM), often known as diabetes, is a condition that occurs when the pancreas stops producing insulin or when body cells no longer react to insulin. Insulin, a sort of hormone that functions as a key that permits glucose from the blood to enter our cells, fuels our body cells. If the insulin-producing beta cells in the pancreas are suppressed, the glucose in the blood is not appropriately controlled, and the glucose level in the blood rises quickly, causing a diabetic. There are essentially four forms of diabetes at that period. Prediabetes is defined by a glucose level that is greater than usual but not yet high enough to be classified as diabetes.

DIABETES RISK ASSESSMENT MODEL DESIGN AND IMPLEMENTATION BASED ON MOBILE THINGS

In this study, Gang Shi, Shanshan Liu, et al. proposed the main risk factors for diabetes and established the diabetes risk assessment model, which was placed on a mobile terminal with a backstage where the data of personal situation collected by questionnaire could be analysed, achieving lifestyle interventions and exercise habit proposals aimed at the selected high-risk diabetes. As a result, this approach has the benefits of being simple to use, broad, and efficient. According to a survey conducted by the World Public Health Organization. It is a significant risk factor for mortality and a slew of nonfatal problems that will place a significant cost on patients, their families, and the health-care system. Several recent intervention trials have unmistakably demonstrated that in high-risk patients, lifestyle changes can effectively prevent type 2 diabetes. In the early 1990s, industrialised nations such as Finland, Europe, and the United States began to research the assessment model of diabetes risks, implementing efforts in terms of screening among diabetes populations, resulting in a considerable reduction in the number of diabetics in foreign countries. Diabetes, on the other hand, is becoming more common in China. According to a poll, the number of diabetics and pre-diabetics is the highest in the globe, and the trend of attacking diabetics is becoming younger

## 3. EXISTING SYSTEM

Diabetes is one of the most well-known non-transmittable illnesses in the world, according to Exiting System. It is ranked as the sixth biggest cause of death. Diabetes will affect 642 million individuals globally by 2040, according to projections. Diabetes detection in patients at an early stage has long been a priority for medical researchers and experts. With the availability of huge technical progress in computer science, joint research have demonstrated that by applying computer skills and algorithms (such as data mining), efficient, cost-effective, and speedy procedures for diabetes diagnosis may be generated.Many academics have used data mining to construct various prediction models to predict and diagnose diabetes. Results achieved after the experiment proves that, Adaboost machine learning ensemble technique outperforms well comparatively bagging as well as a J48 decision tree.In designed system for diabetes prediction, whose main aim is the prediction of diabetes a candidate is suffering at a particular age.

## 4. PROPOSED SYSTEM

The major goal of this proposed effort is to categorise data as diabetes or non-diabetic and increase classification accuracy. For many classification problems, the greater the number of samples picked, the worse the classification accuracy. This survey examined several categorization strategies for diabetes and non-diabetic data. As a result, it is discovered that approaches like as MLP, NB AND RF are most suited for constructing the Diabetes prediction system. With the use of new computational approaches, machine learning has the potential to change diabetes risk prediction. A significant number of epidemiological and genetic diabetes risk datasets are available. Diabetes must be detected early in order to be treated. This paper proposed a machine learning method for predicting diabetes levels. The method may also assist researchers in developing an accurate and useful tool that will reach physicians' tables to assist them in making better decisions about illness state.

In many cases, the performance of algorithm is high in the context of speed but the accuracy of data classification is high. The performance of the model can be described by Confusion Matrix as False Negative (FN), False Positive (FP), True Negatives (TN), True Positives (TP). These models that were used to fit the training set were compared and estimated their performance in terms of precision, recall , accuracy and ROC- AUC-Score.The results obtained in this study have achieved high accuracy rate for predicting diabetes when compared with other existing methods. Improved Accuracy. Better Performance with Minimal Area Under Curve Values Appeared The correctness of the model in predicting the instances is measured in terms of accuracy

PREPROCESSING

It is used in the Data Selection process to detect issues such as missing numbers, incorrect content, and data inconsistency. It computes the data to achieve the desired findings in the data analysis stage by evaluating the datasets with a particular tool like WEKA. Depicts the pre-processing approach comprises transformation. Bringing back the missing values.

NORMALIZING THE DATA

Identifying the incorrect values because real-world data is filthy, fragmentary, and noisy, we must use data preparation techniques. This approach entails identifying mistakes and missing data values from a large dataset. Pre-processing makes it simple to return missing values and fix erroneous data.

MLP, NB and RF

MLP (Multi-Layer Perceptron) is a type of artificial neural network that uses multiple layers of interconnected nodes to classify input data. It is often used for complex Classification problems and can handle non-linear relationships between variables. NB (Naive Bayes) is a probabilistic algorithm that is based on Bayes' theorem. It works by assuming that the features in the data are independent of each other, which is often not true, but still performs well on many real- world datasets. It is a simple and fast algorithm, and it is often used for text classification and spam filtering. RF (Random Forest) is an ensemble learning algorithm that creates multiple decision trees and combines their outputs to make a final classification decision. Each decision tree is built on a random subset of the data, and a random subset of the features. This randomness helps to reduce over fitting and improve the accuracy of the model. RF is often. Used for complex classification problems and is known for its high accuracy and robustness.

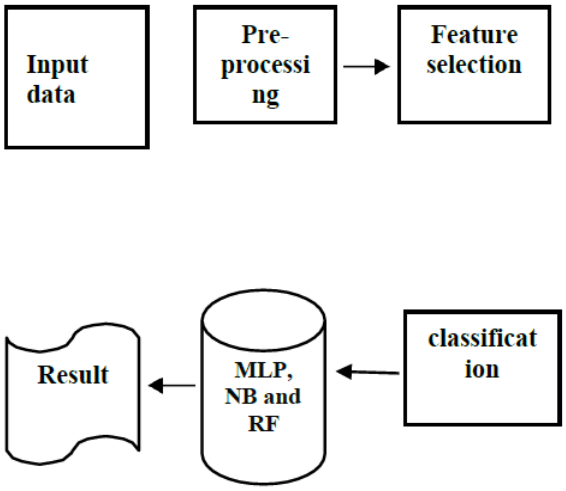

Figure 4.1 process flowchart

RESULT PREDICTION

We discovered the following three values in the PIMA Indian Diabetes data set: 1. Kappa Statistic: This is a benchmark that combines observed and expected accuracy (random chance). 2. MAE-Mean Absolute Error: the average of the absolute error between the observed and predicted values. 3.
RMSE-Root Mean Squared Error: This metric calculates the difference between sample and population values. It might be simply approximated using a prototype or predictor, as well as observed data.

## 5. EXPERIMENTAL SETUP

A significant outcome revealed from the application of MLP, NB and RF is that the procedure assisted in decreasing the disadvantage of having duplicate features that are useless for clustering. Because reducing the number of variables in the original data set aided in dealing with noisy and outlier data, MLP, NB and RF enhanced our k-means result. The main advantage of MLP, NB and RF is that once we have identified these Principal Components from the data and can compress the data, i.e., reduce the number of dimensions without losing much information, it becomes an essential process in order to determine the number of clusters and provide a statistical framework to model the cluster structure.Any predictive and diagnostic model's efficiency and accuracy are critical and should be guaranteed before it is launched for implementation. We investigated and assessed the output of our model using several evaluation measures, and the results are displayed below.

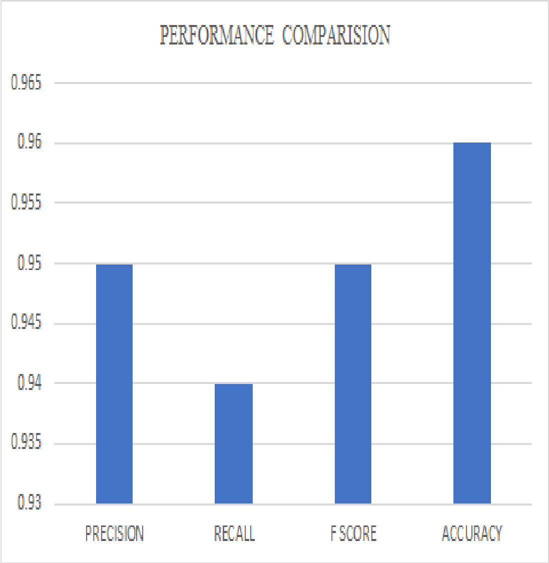

Table 5.1 performance comparision

To begin, we used the fold cross validation approach to measure the performance of our model, which allows us to estimate how well our model would perform when given new and previously unlearned data. Because we used 10-fold cross validation, our dataset was partitioned into ten subgroups. On each trial, one subset serves as the test set, while the other nine serve as the training set. The average error across all ten trials was then computed to determine our model's overall performance. This strategy addresses two issues: first, it decreases the problem of bias because practically all of the data is utilised for fitting, and second,it considerably reduces the problem variance

| MLP, NB and RF PREDICTION AND CLASSIFICATION | EXECUTION TIME (IN S) |
|---|---|
| PRECISION | 0.95 |
| RECALL | 0.94 |
| F SCORE | 0.95 |
| ACCURACY | 0.96 |

Table 5.2 Execution Timings

## 6. RESULT AND DISCUSSION

The proposed model may be a useful tool for accurately predicting diabetes in patients for medical professionals. Medical professionals to identify patients, who are at a high risk of developing diabetes, allowing for early treatment and intervention, can use this model. The potential for incorporating additional machine learning techniques to further enhance the performance of this model can be evaluated in additional studies, as can its performance on larger and more diverse datasets. The objective of this work was to develop a reliable model for diabetes prediction.

## 7. REFERENCES

1. Riccardo B, Blaz Z. Predictive data mining in clinical medicine: current issues andguidelines. Int J Med Inf 2008;77:81– 97.

2. MechelleGittens, Reco King, Curtis Gittens and Adrian Als, Post-diagnosis Management of Diabetes through a Mobile Health Consultation Application, 2014 IEEE 16th International Conference on e-Health Networking, Applications and Services (Healthcom).

3. Marcano-Cede~no Alexis, Torres Joaquín, Andina Diego. A prediction model to diabetes using artificial metaplasticity. IWINAC 2011, Part II. LNCS 6687; 2011. p. 418–25.

4. VeenaVijayan V. and Anjali C., Decision support systems for predicting diabetes mellitus –a review. Proceedings of 2015 global conference on communication technologies (GCCT 2015).

5. Ms. K Sowjanya, MobDBTest: A machine learning based system for predicting diabetes risk using mobile devices. 2015 IEEE International Advance Computing Conference (IACC).

6. Gang Shi, Shanshan Liu and Ding Ye, Design and Implementation of Diabetes Risk Assessment Model Based On Mobile Things, 2015 7th International Conference on Information Technology in Medicine and Education.

7. Yanhui Sun, Liying Fang and Pu Wang, Improved k-means clustering based on Efros distance for longitudinal data, 2016 Chinese Control and Decision Conference (CCDC).

8. PhattharatSongthung and KunwadeeSripanidkulchai, Improving Type 2 Diabetes Mellitus Risk Prediction Using Classification, 2016 13th International

JointConference on Computer Science and Software Engineering (JCSSE).

9. Han Longfei, LuoSenlin. An intelligible risk stratification model based on pairwise and size constrained Kmeans. IEEE J Biomed Health Inf 2016;21(5):1288–96.

10. ArunaPavate and Nazneen Ansari, Risk Prediction of Disease Complications in Type 2 Diabetes Patients Using Soft Computing Techniques, 2015 Fifth International Conference on Advances in Computing and Communications.

11. NagannaChetty, An Improved Method for Disease Prediction using Fuzzy Approach, 2015 Second International Conference on Advances in Computing and Communication Engineering.

12. Michie D, Spiegelhalter DJ, Taylor CC. Machine learning, neural and statistical classification. Ellis Horwood; 1994.

13. Humar K, Novruz A. Design of a hybrid system for the diabetes and heart diseases. Expert SystAppl 2008;35:82–9.

14. RojalinaPriyadarshini, Nilamadhab Dash and Rachita Mishra, A Novel approach to Predict Diabetes Mellitus using Modified Extreme Learning Machine.

15. Li Huan, Zhang Qi, Lu Kejie. Integrating mobile sensing and social network for personalized health-care application. Health care information systems; 2016.

16. Yan Luo, Charles Ling, Jody Schuurman and Robert Petrella, GlucoGuide: An Intelligent Type-2 Diabetes Solution Using Data Mining and Mobile Computing, 2014 IEEE International Conference on Data Mining Workshop.

17. Schnall Rebecca, Rojas Marlene. A user- centered model for designing consumer mobile health (mHealth) applications (apps). J Biomed Inf 2016;60:243–51.

18. MdAbulBasar, Hassan NomaniAlvi, Gazi, A Review on Diabetes Patient Lifestyle Management Using Mobile Application, 18th International Conference on Computer and Information Technology (ICCIT), 21-23 December, 2015.

19. QasimMajeed, HayderHbail and AbdolahChalechale, A Comprehensive Mobile EHealthcare System, IKT2015 7th International Conference on Information and Knowledge Technology.

20. Muhammad H. Aboelfotoh, Patrick Martin and Hossam S. Hassanein, A mobilebased architecture for integrating personal health record data, 2014 IEEE 16th International Conference on e-Health Networking, Applications and Services (Healthcom).