*DOI: https://doi.org/10.17762/turcomat.v11i1.13586*

# Data Mining Techniques for Early Prediction of Diabetes Mellitus

**Sushant Chamoli**
Department of Computer Science & Information Technology, Graphic Era Hill University, Dehradun Uttarakhand India 248002

**Abstract**
The prevalence of diabetes mellitus (DM) as a disease is considered to be a leading cause of mortality rate in the world and tends to effect eye sight, kidney and heart of a human body. Since its occurrence triggers diversity in health issues; detecting it at the right phase is mandatory. For this reason, multiple research scholars have contributed their work in this field of study by adapting to various data mining techniques and thereby reducing the overall workload of medical practitioners. Detection of diabetes mellitus using such techniques have also resulted in its early diagnosis and thereby enhanced the overall treatment of detection at the right stage. The phenomenon of diabetes mellitus can however be categorized as Type 1 and Type 2 diabetes; wherein Type 2 diabetes is responsible to cause heart diseases. Therefore, the primary aim of the research study is to detect the occurrence of DM by utilizing techniques of data mining. For this reason, the authors have implemented the back propagation technique to classify whether an individual is diabetic positive or not. In addition to the back propagation technique, the authors have also implemented Naïve Bayes, Random Forest and J48 with input of neural networks having 8 parameters. The execution of these algorithms is performed using 6 hidden layers of neuron on the PIMA dataset and is further applied on R studio. Throughout the implementation, it has been observed that the back propagation technique generated highest accuracy in comparison to the working implementation of Naïve Bayes, Random Forest and J48.

**Keywords**: Back propagation, diabetes mellitus, data mining, J48, PIMA

## Introduction

The occurrence of diabetes mellitus (DM) is identified by high levels of sugar and glucose in blood and is therefore declared to be as a chronic disease that conducts to serious impairment of kidney and heart diseases in a human body. As per survey reports conducted by diabetes federation [1] an approximate number of 400 million people suffer from diabetes and are yet unaware about it. The number is however expected to double by 2035. Hence, its early detection in the right stage has become a mandatory task for not only medical practitioners, but also research scholars. The occurrence of diabetes can however be characterized into four types: the first category being referred to as Type 1 diabetes that majorly occurs in young age group of people. This category of diabetes is also labelled to as juvenile diabetes wherein the occurrence of the disease tends to destruct the immune system of a human body by releasing cells that deteriorates the level of insulin in the blood [2]. This triggers to less production of insulin in the human body. The second category of DM is referred to as Type 2 diabetes and can therefore occur at any age amongst individuals. This category of diabetes is also labelled to as insulin independent diabetes; wherein the human body becomes completely resistant to insulin and stops its production in the human blood [3]. The third category of diabetes is referred to as gestational diabetes that majorly tends to occur and affect pregnant women [4]. According to statistics, 18 percent of pregnant women suffer from gestational diabetes and is majorly induced due to fluctuations in sugar levels. The final category of diabetes is referred to as pregestational diabetes and occurs during pregnancy wherein the body becomes resistant to production of insulin the blood.

Therefore, such instances of DM occurrence tends to have a major effect on the human body, including to damage of retina, loss of vision, liver problems and cardiovascular diseases. Hence in such a scenario; detection of diabetes mellitus (DM) becomes a significant challenge. For this purpose, the implementation of data mining techniques is majorly used wherein; the data is initially extracted from a repository and patterns of prediction are formed from historic events that have occurred in the past. Such historic events are stored in the database and thereby contribute in various domains of banking and healthcare [5]. The adoption of such techniques occurs on a larger set of data that helps to predict the occurrence of the disease at the right stage with optimized results. However, there are multiple algorithms and techniques that can be used to detect the occurrence of the same. Therefore, the primary aim of the research study is to detect the occurrence of diabetes mellitus using such existing data mining techniques and thereby predicting the disease at the right stage with generation of optimised results. For this purpose, following are the contributions of the study:

- To generate a novelty in the proposed system
- To conduct a thorough literature survey and analyse the limitations of the existing system
- To implement data mining techniques and generate optimised results

The organization of the research thus proposed is sequenced by an introduction of diabetes mellitus followed by literature survey of various authors performed in the same field of domain. The research then mentions the proposed methodology along with the algorithms thus used to predict the same. A detailed architecture of the system model is briefed followed by conduction of experimental results. The paper finally comes to an end by summarizing the conclusion in accordance with the references.

## Literature Survey

In a research work proposed by authors Sajida Perveena et.al in [5]; they performed data mining classification techniques to predict the occurrence of diabetes mellitus in the early stages. For this purpose, they obtained the PIMA dataset from Kaggle repository and worked on 516 patients. The analysis thereby included the examination of glucose levels, blood pressure and BMI of the registered patient. The entire classification of the patient as diabetes positive or not was based on the implementation of three data mining algorithms namely; AdaBoost, bagging ensemble and J48. The surveillance data was initially obtained and went through the pre-processing stage. This stage however comprised of labelling the data using binary notions of 0 and 1; wherein 0 represented diabetes negative and 1represented diabetes positive patient. After the stage of pre-processing the data; the dataset from the repository was further visualized using count plots. This visualization helped to classify DM positive and negative patients. The data was then further sent for the training phase; with 80 percent of the overall data and used for testing purpose on the above mentioned algorithms. 20 percent of the overall dataset was implemented on data mining algorithms. Throughout the execution of the dataset, it was observed that the implementation of J48 resulted in the generation of highest levels of accuracy with a precision factor of 81 percent. A similar work was further extended by authors Megha Borse et.al in [6] wherein they combined the implementation of data mining techniques along with neural networks. The usage of neural networks helped to enhance the overall system through its hidden layers. The hidden layers thereby consisted of 8 parameters of neurons which were further trained using 10 cross-fold validation techniques. The overall implementation was also performed using back propagation technique. This was primarily done to trigger the neuron in the forward direction with its initial weight matching the biases thus dedicated to each neuron. The overall implementation was deployed using MATLAB as the software tool. The authors also created a GUI as a user friendly interaction; wherein the patient could enter his personal details such as glucose levels, blood pressure, age BMI etc. the GUI would then classify the patient as DM positive or negative based on his health factors. The data mining techniques however used by the authors included the execution of SVM, KNN and J48. On experimental analysis; it was witnessed that the implementation of SVM generated optimised levels of accuracy and gave a precision factor of 85.23 percent.

Authors Kumari Deepika et.al in [7] conducted their experiments on the PIMA dataset which was obtained from the Kaggle repository. The dataset comprised of 786 diabetic and non-diabetic patients. The overall dataset was spread over two csv files that consisted of train and test data. 70 percent of the dataset was used for training purpose and 30 percent of the dataset was used for testing purpose. The author further trained the dataset on four data mining algorithms namely; SVM, KNN, Logistic Regression and a stacking model. The stacking model was thus built using Meta classifiers and based classifiers. The Meta classifiers used by the author were the execution of AdaBoost algorithm and logistic regression; whereas the execution of base classifier was fulfilled through the execution of J48. Once the data was trained and tested; the dataset further underwent the process of validation. This stage included validating the train dataset using 10 fold cross validation which was run for 20 epochs. Adam was used as the optimiser with ReLu as the activation function. Through the executional run; it was observed that the validation error was eventually reduced with an increase in validation accuracy. The overall precision factor thus obtained was witnessed to be 89.23 percent and was achieved through the implementation of stacking algorithm.

In a similar work proposed by authors VeenaVijayan et.al in [8]; they predicted the occurrence of diabetes mellitus amongst patients using the BMI index level, glucose levels and blood pressure levels in a human body. The prediction mechanism also involved categorizing the patient with Type 1 or Type 2 diabetes. This categorization helped to assist the algorithms wherein; the final classification of the patient as DM positive or DM negative could be done. The diagnosis of diabetes mellitus was further enhanced through the implementation of four data mining based algorithms which included the execution of AdaBoost, gradient boost, logistic regression and a hybrid algorithm. The implementation of the hybrid algorithm was similar to the one occurring in the research work proposed by authors in [7]. However, the Meta classifier used in this research comprised of J48 and SVM whereas the base classifier used comprised of KNN. The overall dataset was split into training, testing and validation phase; with the split ration of 70:20:10 respectively. The dataset used was obtained from Kaggle repository and thereby consisted of csv files as train and test. The dataset further included historic data of 750 patients which were further categorised as DM positive or negative patients. Throughout the implementation, it was observed that the executional run of hybrid algorithm generated highest levels of accuracy and produced a precision of 91.56 percent.

Authors Jianxin et.al in [9] conducted their analysis based on patient's historic data. This data included his personal details such as BMI index, age, weight, glucose levels and cardiovascular movements. The details also included his body attributes such as that of waist and hips. PIMA dataset was used for the detection and prediction of diabetes mellitus in a patient. However, the research work of the author included a detailed study being provided on Type 2 wherein the production of insulin is completely stopped in a human body and the body becomes resistant to the same. Since the occurrence of Type 2 diabetes can take place in people from any age group; patients who were aged above 60 were also diagnosed. It was observed that majority of the patients were still unaware of the presence of the disease in their body that eventually led to loss of vision and triggered kidney diseases. Hence the detection was highly made a compulsion. To execute the proposed research; the authors implemented three data mining algorithms along with a combination of neural network that comprised of multi-layer perceptron (MLP). The usage of an MLP led to the generation of hidden layers based on evaluation of 7 parameters. The layers however comprised of neurons that eventually enhanced the overall working implementation of the proposed model. The data mining algorithms included the execution of SVM, KNN and AdaBoost. On implementation, it was observed that the MLP generated optimised and better results and gave a precision factor of 91.23 percent.

## Proposed Methodology

The primary aim of the proposed work is to detect the occurrence of diabetes mellitus in a patient and further classify him as DM positive or negative. For this purpose, we have used data mining techniques to detect the same. The usage of data mining technique enables to gain insights from the past and take into consideration the patients' health records such as his age, BMI index, glucose levels, blood pressure, eye vision, cardiovascular status etc. Once the algorithms thus used gets access to this patient data; the respective algorithms are further trained and tested using cross fold validation technique. The entire process of data mining thus helps to retrieve past data and further predict the occurrence of the disease in the future. Therefore, this process helps to analyse and identify the disease in the early stages of its occurrence. However for the execution of the proposed work, the authors acquire the repository from PIMA dataset and further perform its analysis on the same. The obtained dataset then undergoes the stage of pre-processing wherein; the noisy and redundant data is discarded. This process is done to filter necessary attributes of patient data and further use t for training purpose. The pre-processing phase also involves a data cleaning technique wherein irrelevant data is filtered so as to create a subset of relevant data. After this stage; a process of normalization is followed wherein; mathematical equations to calculations of min-max values are obtained. The system model then undergoes the process of cross validation wherein the model is run for 20 epochs using Adam as the optimizer and ReLu as the activation function. A 5 cross validation process is performed at this stage. The process of back propagation is thus followed after cross validation technique. Back propagation is primarily done to classify the patient as DM positive or negative. This classification is done on the PIMA dataset thus obtained. The system is then trained and tested on respective data mining techniques thus used. For the purpose of implementation of the proposed research study; the authors have used Random Forest, J48 and Naïve Bayes as data mining algorithms. Once the entire system model is trained and tested on the above mentioned PIMA dataset; the system then undergoes the process of performance evaluation. This step is mandatory so as to declare the optimised model amongst the three algorithms thus executed. The declaration is however done on the basis of accuracy thus obtained. The entire executional process and the workflow of the same is depicted in figure 1 below:

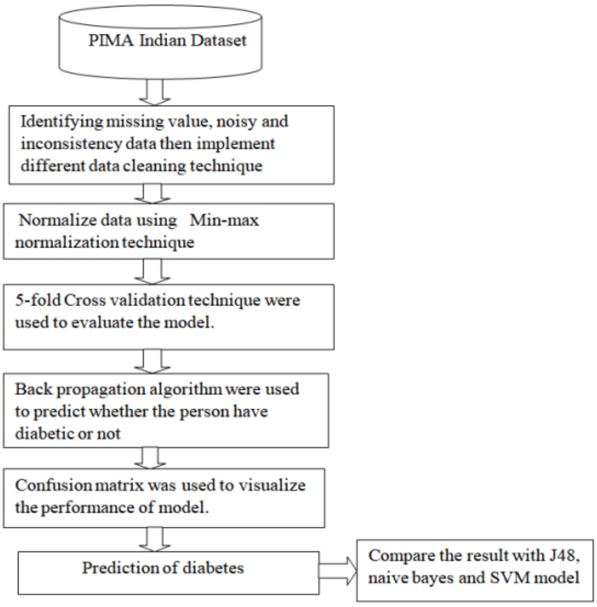

PIMA Indian Dataset
Identifying missing value, noisy and inconsistency data then implement different data cleaning technique
Normalize data using Min-max normalization technique
5-fold Cross validation technique were used to evaluate the model.
Back propagation algorithm were used to predict whether the person have diabetic or not
Confusion matrix was used to visualize the performance of model.
Prediction of diabetes
Compare the result with J48, naive bayes and SVM model

**Figure 1: Architecture of Proposed Methodology**

*A. Description of the dataset used*

The entire study of the research has been executed on the PIMA dataset thus obtained from Kaggle repository. The dataset however, consists of 786 DM instances of patients with 8 attributes belonging to each patient. The attributes are however labelled using two classes as either numeric or normal. The attributes thus used is depicted in the table below:

**Table 1: attributes of DM patients used in the dataset**

| Attribute | Type of Data | ***Description*** |
|---|---|---|
| Npreg | Numeric | Number of pregnancy |
| Glu | Numeric | Glucose concentration in the body |
| Bp | Numeric | Pressure Diastolic blood pressure |
| Skin | Numeric | Skin fold thickness(mm) insulin |
| Serum | Numeric | 2-hour serum insulin |
| Bmi | Numeric | Body mass index |
| Ped | Numeric | Pedigree diabetic function |
| Age | Numeric | Age of patient |
| Type | Nominal | Class variable of diabetes |

*B. Data Pre-Processing*

Data pre-processing is one of the most significant steps that is carried out throughout the implementation of data mining technique. It is in this step; that the inconsistency in the data is thus removed. Data pre-processing is primarily done so as to balance the data obtained from the dataset. The data balance is performed to generate optimised results during the training and testing phase of the system model. On the other hand, the pre-processing stage also includes identifying missing and NULL values from the dataset and thereby discarding it so as to maintain overall consistency.

### *C. Data Normalization*

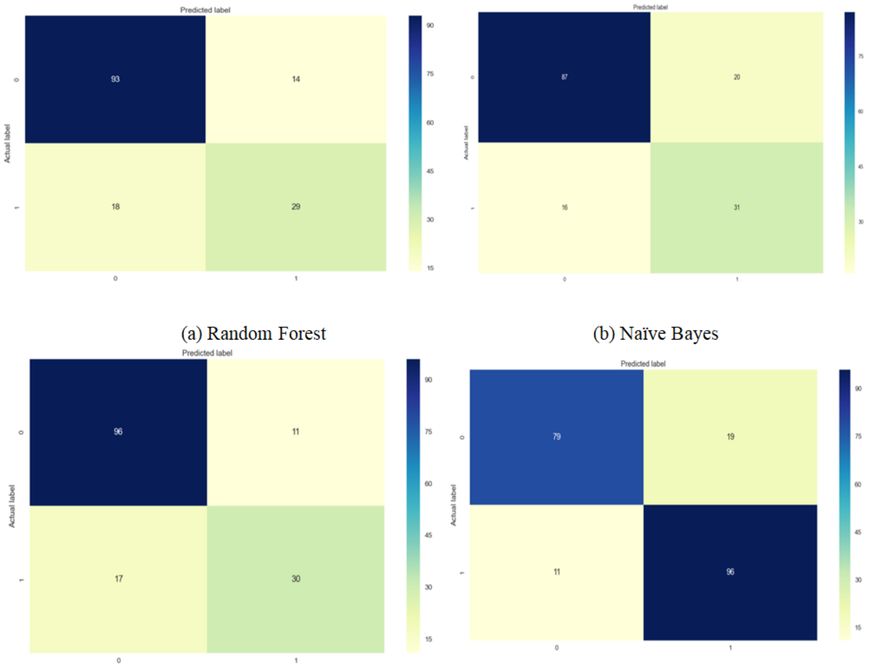
Data normalization is primarily done to convert the obtained data from the dataset into a format that is readable by data mining algorithms. This conversion and transformation of data is however carried out using aggregation techniques such as mathematical calculation of min-max values. For the implementation of the proposed thesis; the authors have used the normalization technique of converting the data into suitable format using min-max values performed through aggregation.

### *D. Methodologies Used*

- Back Propagation
  The implementation of a back propagation technique involves the functioning of a conventional neural network. However, back propagation is considered to be a feed-forward based neural network that is responsible to adjust the weights and thereby assign respective bias to the neurons thus involved. The working of a neural network occurs through neurons wherein all the layers are stacked between input vectors. Every input is being passed through a neuron; a function is applied to it and later fed to the next stages. The entire procedure majorly comprises of the input layer, hidden layer and the output layer. For the purpose of implementation of the proposed thesis; 4 hidden layers with each layer having 8 neurons is implemented. The hidden layer is further responsible to inhibit large epoch values using Adam as the optimiser [10].
- Naïve Bayes
  In order to handle classification issues, a Naive Bayes algorithm is typically used. This algorithm is known as a naive algorithm since each feature's principal occurrence in it occurs independently of all other feature occurrences. But, the algorithm's overall forecast is based on probability and the correlation between the estimated probabilities and the likelihood that the event will occur. The main variables that make up this algorithm's implementation are entirely determined by the likelihood that the symptoms will emerge [10].
- Random Forest
  The application of random forests is one of the most successful ensemble methods used for classification in data mining. This method is well recognised for its tendency for prediction and estimation based on probability. A group of numerous trees is typically referred to as a "random forest," in which each decision tree within the group contributes to the generation of a vote. The majority of the votes cast are accountable for indicating a choice about the object's class. As a result, random forests are frequently referred to as a hierarchical group of trees. The dataset used in the experimental investigation for our research is large enough to contain several attributes for a single instance [10].
- J48
  The implementation of J48 is a data mining technique which is considered to be an extension to the ID3 data mining technique. The working implementation of J48 involves feature detection of missing values and thereby building a classification model with value ranges that falls under the continuous category of PIMA dataset.

## Results

This section of the research illustrates the results thus obtained on conduction of the experiments.



(a) Random Forest

(b) Naïve Bayes

(c) J48

(d) Back Propagation

**Figure 2: Confusion Matrix**

The confusion matrix created in this manner for diabetes prediction is shown in the figure above. The values from the Random Forest approach, which produces 93 true positive cases and 14 false positive cases for predicted patients with diabetes, are shown in the aforementioned figure (a). On the other hand, for patients without diabetes, the RF technique produces 18 false negatives and 29 real negatives cases. Similar to this, confusion matrices are used to create all of the values for different algorithms. Nonetheless, it is crucial to highlight that back propagation implementation, which yields an overall testing accuracy of 81.81 percent as seen in table 2, provides the maximum accuracy.

**Table 2: Accuracies of algorithms**

| Name of Algorithm | Precision | | Recall | | FI-Score | |
|---|---|---|---|---|---|---|
| Random Forest | Positive Cases | Negative Cases | Positive Cases | Negative Cases | Positive Cases | Negative Cases |
| | 0.84 | 0.67 | 0.87 | 0.62 | 0.85 | 0.64 |
| **Accuracy** | **0.79** | | | | | |
| Naïve Bayes | Positive Cases | Negative Cases | Positive Cases | Negative Cases | Positive Cases | Negative Cases |
| | 0.84 | 0.61 | 0.81 | 0.66 | 0.83 | 0.63 |
| **Accuracy** | **0.77** | | | | | |
| J48 | Positive Cases | Negative Cases | Positive Cases | Negative Cases | Positive Cases | Negative Cases |
| | 0.85 | 0.73 | 0.90 | 0.64 | 0.87 | 0.68 |
| **Accuracy** | **0.75** | | | | | |
| Back Propagation | Positive Cases | Negative Cases | Positive Cases | Negative Cases | Positive Cases | Negative Cases |
| | 0.85 | 0.73 | 0.90 | 0.64 | 0.87 | 0.68 |
| **Accuracy** | **0.81** | | | | | |

## Conclusions

The primary aim of the research study is to determine and predict whether the patient inhibits diabetes mellitus or not. For this purpose, the authors have used the conceptual theory of data mining algorithms and predicted the same using random forest, Naïve Bayes, J48 and techniques of back propagation. The dataset used for the same; is collected from the Kaggle repository and worked upon the PIMA dataset. The PIMA dataset however comprises of 768 patient instances with 8 attributes to be analysed. On experimentation analysis; it has been observed that the executional implementation of back propagation technique which involved the usage of hidden layers tends to generate an optimised result in comparison to other algorithms. However, the accuracy thus produced was witnessed to be 81.81 percent. The same work can be extended in the future with usage of larger datasets which can further be enhanced through data augmentation techniques.

## References

[1] S. Abhari, S. R. Niakan Kalhori, M. Ebrahimi, H. Hasannejadasl, and A. Garavand, "Artificial intelligence applications in type 2 diabetes mellitus care: focus on machine learning methods," Healthcare informatics research, vol. 25, no. 4, pp. 248–261, 2019.

[2] S. Larabi-Marie-Sainte, L. Aburahmah, R. Almohaini, T. Saba Current techniques for diabetes prediction: review and case study Appl. Sci., 9 (21) (2019), p. 4604

[3] S.I. Ayon, M.M. Islam Diabetes prediction: a deep learning approach Int. J. Inf. Eng. Electron. Bus., 12 (2) (2019), p. 21

[4] D. Sisodia, D.S. Sisodia Prediction of diabetes using classification algorithms Procedia Comput. Sci., 132 (2018), pp. 1578-1585

[5] S.Perveen , M.Shahbaz , A.Guergachi , and K.Keshavjee ,"Performance analysis of data mining classification techniques to predict diabetes,"Procedia Computer Science.pp.115-21,Dec 31 2016

[6] S.Joshi and M. Borse,"Detection and Prediction of Diabetes Mellitus Using Back-Propagation Neural Network,"International Conference on Micro-Electronics and Telecommunication Engineering, Ghaziabad India , 22-23 Sept ,2016, PP.110 – 113

[7] K.Deepika and S. Seema , "Predictive analytics to prevent and control chronic diseases," International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT),Banglore India, 21-23 July, 2016,pp. 381-386

[8] V.Vijayan and C.Anjali , "Prediction and Diagnosis of Diabetes Mellitus -A Machine Learning Approach II Diabetes,",IEEE Recent Advances in Intelligent Computational Systems (RAICS),Trivandrum India ,10-12 Dec,2015,Pp. 122 – 127

[9] W. Xu, J. Zhang, Q. Zhang, and X. Wei ,"Risk prediction of type II diabetes based on random forest model,"International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics , Chennai India, 27-28 Feb, 2017, Pp. 382 – 386

[10] Mrs. Nalini Jagtap1, Mrs. P. P. Shevatekar, and Mr. Nareshkumar Mustary "A comparative study of classification techniques in data mining algorithms," International Journal of Modern Trends in Engineering and Research, Volume 04, Issue 7, July– 2017