# Diabetes Prediction using Machine learning and Data Mining Methods

**Anant Ram, Honey Vishwakarma**

GLA University, Mathura
anant.ram@gla.ac.in, honeyvishwakarma8787@gmail.com

**Abstract.** Diabetes mellitus, commonly known as diabetes, is a metabolic disease. It is an extremely regular disease to the humankind from young to oldster. A persistent disease appears when blood glucose level is too high. Hence, to reduce the increasing rate of diabetes, diagnosing diabetes is very important. Data Analytics is a methodical procedure of examining and recognizing the concealed pattern from huge measure of information to reach conclusions. In medical science, this methodical procedure is implemented by using different machine learning algorithms to analyze the medical data like K-Nearest Neighbors, Support Vector Classifier, Logistic Regression, Gaussian Naive Bayes, and Random Forest. The objective of this research is to utilize significant features rather than using all the features. Therefore, we performed the data cleaning along with the potential feature selection and then used the Logistic Regression. Proposed approach outperform with some existing approaches that are using the machine learning algorithms.

**Keywords:** Diabetes, Machine Learning, Support Vector Classifier, Logistic Regression.

## 1. Introduction

HUMAN body needs stamina for actuation. The carbohydrate is separated into glucose, which is the significant energy source for the cells of the human body. Insulin is expected to move the glucose into body cells. The blood glucose is provided with insulin and glucagon hormones created by the pancreas. Insulin hormones delivered by the beta cells of the islets of Langerhans and glucagon hormones are created by the alpha cells of the islets of Langerhans in the pancreas [1]. At the point when the blood glucose builds, beta cells are invigorated and insulin is given to the blood. Insulin empowers blood glucose to get into the cells and this glucose is utilized for vitality. So blood glucose is kept in a restricted range. Diabetes is an interminable illness with the possibility to cause an overall medicinal services emergency [2]. In any case, early forecast of diabetes is a very testing errand for restorative professionals because of the complex relationship on different variables. Diabetes influences human organs, for example, heart, nerves, foot, kidney, eye and so forth.

Data Mining is a procedure to extract helpful information from huge dataset, as there are extremely large and vast data is available related to diabetes in hospitals. It is a multidimensional field of Information technology, which includes computational procedure, Artificial Intelligence, Classification techniques, Statistical technique, Clustering and recognizing patterns. In recent time,
Data mining techniques have been broadly utilized in forecasting the data and this is done by using different data mining algorithms to predict the disease with higher precision so as to spare the human life and decrease the treatment cost. The outcome of data mining techniques also depends on potential features associated with multidimensional data. Therefore, in this paper we proposed framework that extract the potential features from the datasets to predict the diabetes. The proposed framework deals with a bigger dataset to increase the accuracy. We utilize five different machine learning algorithms, namely K-Nearest Neighbors, Support Vector Classifier, Logistic Regression, Gaussian Naive Bayes, and Random Forest lead our analysis. We also utilize feature selection method to improve the accuracy of the result.

Paper is organized as follows, Section 2 presents literature review, section 3 presents experimental evaluation of already existing approaches, section 4 presents proposed model, section 5 presents the experimental evaluation and finally section 6 conclude the paper.

2. Related Work

Lot many researchers are focusing the on diagnosis of Diabetes. In [3] authors have introduced a technique on prediction of Diabetes. In this Research, for Diabetes Prediction, Decision tree, random forest and neural network were used and for feature Engineering, Principle Component analysis (PCA) and Minimum Redundancy Maximum Relevance (mRMR) techniques were used. The results showed that prediction with random forest gives the highest accuracy (ACC = 0.8084) when all the attributes such as pregnancy, plasma glucose, diastolic blood pressure, triceps skin fold thickness, 2-h serum insulin, body mass index, diabetes pedigree function and age were used.

In [4] authors have proposed a technique based on predictive analytics. Predictive analytics in healthcare can change the way how medical researchers and practitioners gain insights from medical data and take decisions. In this paper, they used six popular machine learning algorithms namely SVM, KNN, LR, DT, RF and NB for predictive analytics. From the experimental results obtained, it can be seen that SVM and KNN gives highest accuracy for predicting diabetes. Both these algorithms provide 77% accuracy which is highest as compared to other four algorithms used in this paper.

In [5], authors introduced a technique for Diabetes Prediction. The goal of this examination is to structure a model which can guess the probability of diabetes in patients with greatest accuracy. In this they were used three classification techniques namely Decision Tree, Naïve Bayes, and Support Vector Machine to early forecast diabetes at initial stage. Different estimates like Accuracy, Precision, F-Measure and Recall were used to calculate the performance of all three models. Result showed that Naïve Bayes Model gives highest accuracy of 77.30% in comparison of other two models.

In [6], authors proposed an effective predictive model with high affectability and selectivity to more readily identify Canadian patient on the basis of laboratory result and analytical data of patient. In this they were used 13309 Canadian patient records and built a predictive model by using Logistic Regression and Gradient Boosting Machine (GBM) techniques. The unfair ability of these models is calculated by using area under the ROC curve (AROC). They also compare these models with other machine learning techniques such as Random Forest and Decision tree. The GBM with accuracy 84.7% and Logistic Regression model with accuracy 84.0% gives better result as compare to Random Forest and Decision Tree Models.

In [7], paper introduced a technique to assess the classifiers that can predict the probability of disease in patients with the greatest accuracy. This work has been done using different machine learning algorithm namely KNN, Naive Bayes, Support Vector Machine, Decision Tree, Logistic Regression, Random Forest on Pima Indians Diabetes dataset. Performance of these models is calculated in terms on recall, precision, accuracy and ROC-AUC-Score. In comparison of other
algorithms Logistic Regression performed better with the accuracy of 77.6 %, ROC-AUC score of 73.6 and f1-score of 75.

In [8], authors has used data mining techniques to analyze the performance of various algorithms i.e. Random Forest, Support Vector Machines and J48 Decision Tree, K-Nearest Neighbors for diabetes prediction. They analyze algorithm performance with the dataset (from UCI machine learning data repository) in both instances i.e. before and after pre-processing and perform the comparison with respect to Sensitivity, Specificity and Accuracy.

In [9], authors has used a machine learning algorithm i.e. Support Vector Machine as the classifier for diagnosing diabetes disease and for testing Pima Indian diabetes dataset has been used. In the proposed technique, SVM with Radial basic function is used for classification. The investigational outcome proves that the support vector machine can be effectively used for diagnosis of diabetes.

In [10] authors have examined Hybrid Prediction Model which utilize K-mean clustering algorithm for prediction of Type-2 diabetes. In this study, they subsequently applied clustering algorithm on the data to extract a pattern from large dataset and at last C4.5 algorithm was used to construct the ultimate classifier model. They have used sensitivity and specificity for analyzing the performance of suggested method and give the accuracy of 92.38%.

In [11], authors performed feature selection as pre-processing to remove redundant and irrelevant data to improve the accuracy and this was done by using F-score method and K-mean clustering algorithm. They evaluate the performance of SVM classifier with respect to various factors i.e. Accuracy (98%), Sensitivity (97.77%) and Specificity (97.79%).

In [12], authors introduced a decision support system that uses AdaBoost Algorithm with Decision Stump as base classifier for classification. Along with this Support Vector Machine, Naive Bayes and Decision Tree were also applied for verification of accuracy. AdaBoost algorithm gives best accuracy i.e. 80.72% in comparison of Support Vector Machine, Naive Bayes and Decision Tree.

3. Existing Algorithms

We performed the experimental evaluation with existing methods [13] by using the same data set directly.

K-Nearest Neighbors: is a Supervised learning Algorithm figure 1 that can be used to implement both regression and classification problems. According to this, identical things exist in close proximity.
Logistic Regression: Logistic Regression Model: A Statistical model that is used to explain data and to describe the relationship between one dependent variable and at least one ordinal, nominal or ratio-level independent variable. This approach performs well when the test is based on two values.

\[ n(t) = \frac{1}{1 + e^{-t}} \]  \hspace{1cm} (1)

Support Vector Classifier: It is a supervised machine learning algorithm that examines data used for regression and classification analysis. In this, every data item is plotted as a point in n-dimension (where n indicates the number of features) along with each feature value, shown in figure 2. After that, to separate the two classes very well we find hyper-plane by performing classification.
Gaussian Naïve Bayes: It is classification model based on Bayes Theorem. It is not a single algorithm but the collection of different algorithms where all the algorithm shares same objective. It requires that each feature of dataset must be independent and not related with one another. The main advantage of this model is that it provides feasible solution on large dataset.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$  \hspace{1cm} (2)

Random Forest: A classification algorithm containing more no. of decision trees. It is like a bootstrapping algorithm. On data sample it creates an individual decision trees and then gets the forecast from every decision tree and then finally select the best feasible solution by means of voting. It gives more accurate result in comparison of individual decision tree, represented in figure 3.

K Fold Cross Validation: Cross Validation is a valuable strategy to estimates the performance of Machine learning models. It helps in understanding how the models would perform on independent data set. It is sometime known as Rotation Estimation.
Working: This method divides the data set into k equal folds and then utilize one fold as the testing set and the union of other folds as training set, as shown in figure 4. This procedure continues k times and at each time it uses different fold as the testing. The testing accuracy is the average testing accuracy of the procedure.

Table 1: Average Testing Accuracy of Models

| S.No. | Model Name | Score   |
|-------|------------|---------|
| 1     | KNN        | 0.759787|
| 2     | SVC        | 0.808581|
| 3     | LG         | 0.847021|
| 4     | RF         | 0.772260|
| 5     | GNB        | 0.817797|

As per the accuracy score, we can observe that the Logistic Regression performs better than the other algorithm.

4. Proposed Model

As we are aware that machine learning algorithms are applied on data to achieve the accurate result and it requires first to preprocess data. Therefore we proposed the model, presented in the following figure 5. This framework presents the sequence and flow of the experimental process followed in this work.
4.1 Dataset Preparation: Pima Indians Diabetes Database is the widely used data set for experimental evaluation. This data set is available at UCI Machine Learning Repository. This dataset contains 768 observations with nine attributes.

4.2 Data Exploration: Data exploration is an approach similar to initial data analysis with the data and get some understanding about the significant features. In this data analyst utilize visual exploration tools to know what is actually present in a dataset and the qualities of the data.
4.3 Data Cleaning: Data Cleaning is the next phase of the workflow. It is viewed to be one of the vital steps of the workflow. In this, we remove the irrelevant or duplicate data, missing values and unexpected outliers. After analyzing the dataset, we observed that there are some fields that contain zero values of Blood Pressure, BMI and Glucose. The following table shows the no. of fields that contain zero value. We removed these observations from the data set and after that there are 724 observations with nine attributes.

| Attributes       | No. of counts contains zero values |
|------------------|-----------------------------------|
| Blood Pressure   | 35                                |
| BMI              | 11                                |
| Glucose          | 05                                |

4.4 Algorithm Selection: This phase is the most emerging phase of the workflow. In this phase we select the model which gives the highest accuracy result with the dataset. We applied the logistic regression by considering all the features that are available in the dataset. We also performed the experimental evaluation using logistic regression with selected five features and observed that feature selection play significant role in accuracy.

4.5 Feature Engineering: In this phase, we used the recursive feature elimination approach to find the potential features to improve the accuracy. Recursive Feature Elimination approach removes the features in a recursive manner and then builds a model by utilizing the remaining attributes.

![Logistic Regression CV score vs No of Features](image)

Figure 7 Feature Accuracy of Logistic Regression

We applied the same on above considered dataset. The figure 7 represents that accuracy with potential features returned by the recursive feature elimination approach along with Logistic regression. The observation over the above plot shows that the five features ['Pregnancies', 'Glucose', 'BMI', 'Diabetes Pedigree Function', 'Age'] to the model gives better accuracy rather then considering all features.
5. Results and Discussion

We simulated the proposed model by considering the above mentioned dataset. We also implemented a number of already existing approach and compared the result with our proposed model. Since we used the recursive feature elimination approach to find the potential features to improve the accuracy, while already existing approaches considers the complete feature space for the calculation of accuracy. We also performed the experimental evaluation by considering all features and selected features and observed that all the feature consideration produced 0.8470210711728514 accuracy while the optimized feature space produced the 0.8505877119643279 accuracy. The experimental evaluation of existing approaches and proposed approach is given Table 3. The proposed approach outperform with respect to the existing approaches, in both the situations i.e. by considering complete feature space or the potential feature space that improve the accuracy level.

| Reference         | Dataset used       | Proposed Model / Method | Accuracy     |
|-------------------|--------------------|-------------------------|--------------|
| Quan Zou          | PIMA Indian dataset| Random Forest           | 80.84 %      |
| Muhammad Azeem Sarwar | PIMA Indian dataset | 1. Support Vector Machine 2. K-Nearest Neighbors | 1. 77 %  2. 77 % |
| Deepti Sisodiya   | PIMA Indian dataset| Naïve Bayes             | 77.30 %      |
| Hang Lai ()       | PIMA Indian dataset| 1. Gradient Boosting Machine 2. Logistic Regression | 1. 84.7 %  2. 84.0 % |
| Proposed Framework(Model) | PIMA Indian dataset | 1. Logistic Regression with all nine features 2. Logistic Regression with five features | 1. 84.7 %  2. 85.05 % |

Table 3: Comparisons of our framework with other researchers

6. Conclusion

In this paper, a systematic experimental study is followed by using various algorithms to predict the diabetes. Based on the above experiment, Logistic regression with all features gives better result (ACC = 84.70 %) for prediction of diabetes in comparison of other algorithms. we performed feature engineering to select the suitable features, has a better performance with respect to other models. It means Logistic regression with some features gives the best accuracy of 85 %. In this paper, we cannot predict a person has which type of diabetes, so in future our aim is to predict the type of diabetes and analyze the feature of each indicator, which may improve the accuracy of diabetes prediction.
References

[1] David L. Krull, Richard A. Peterson, “Preclinical Applications of Quantitative Imaging Cytometry to Support Drug Discovery”, Methods in Cell Biology, vol. 102, pp. 291-308, 2011.

[2] John Wiley & Sons, “Practical Diabetes Care, 4th edition 20-Feb-2018.

[3] Zou Quan, Qu Kaiyang, Luo Yamei, Yin Dehui, Ju Ying, Tang Hua, “Predicting Diabetes Mellitus With Machine Learning Techniques”, rontiers in Genetics. vol 9, article 515.

[4] M. A. Sarwar, N. Kamal, W. Hamid, M. A. Shah, "Prediction of Diabetes Using Machine Learning Algorithms in Healthcare", 2018 24th International Conference on Automation and Computing (ICAC), no. September, pp. 1-6, 2018.

[5] Deepti Sisodiya and Dilip Singh Sisodiya, “Prediction of Diabetes using Classification Algorithms”, International Conference on Computational Intelligence and Data Science (ICCIDS 2018), 2018.

[6] Hang Lai, Huaxiong Huang, Karim Keshavjee, Aziz Guergachi &amp; Xin, “Predictive models for diabetes mellitus using machine learning techniques”, BMC Endocr Disord 19,October 2019.

[7] Aishwarya Jakka, Vakula Rani J, “Performance Evaluation of Machine Learning Models for Diabetes Prediction”, International Journal of Innovative Technology and Exploring Engineering (IJITEE), vol 8, no. 11, pp. 1976-1980, 2019.

[8] J. P. Kandhasamy and S. Balamurali, “Performance Analysis of Classifier Models to Predict Diabetes Mellitus,” Procedia Comput. Sci., vol. 47, pp. 45–51, 2015.

[9] V. Anuja Kumari, R.Chitra, "Classification of Diabetes Disease Using Support Vector Machine", International Journal of Engineering Research and Applications, vol. 3, no 2, pp.1797-1801, 2013.

[10] B. M. Patil, R. C. Joshi, and D. Toshniwal, “Hybrid prediction model for Type-2 diabetic patients,” Expert Syst. Appl., vol. 37, no. 12, pp. 8102–8108, 2010.

[11] K. K. Gandhi and P. N. B. Prajapati, “Diabetes prediction using feature selection and classification,” Int. J. Adv. Eng. Res. Dev., vol. 1, no. 05, pp. 1–7, 2014.

[12] V.V.Vijayan and C. Anjali, "Prediction and diagnosis of diabetes mellitus A machine learning approach", IEEE Recent Advances in Intelligent Computational Systems (RAICS) , pp. 122–127, 2015.

[13] Rashedur M. Rahman, Farhana Afroz, “Comparison of Various Classification Techniques Using Different Data Mining Tools for Diabetes Diagnosis”, Journal of Software Engineering and Applications, Vol 6,pp. 85-97,2013.