International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 08 | Aug 2020 www.irjet.net p-ISSN: 2395-0072

# Diabetes Prediction Using Machine Learning Classification Algorithms

**Jitranjan Sahoo[1], Manoranjan Dash[2], Abhilash Pati [3]**

*[1]Assistant Professor, Interscience Institute of Management & Technology, Bhubaneswar, Odisha, India.*
*jitranjansahoo@gmail.com*
*[2]Associate Professor, Siksha O Anusandhan(Deemed to be University),Bhubaneswar,Odisha,India.*
*[3]Research Scholar, Department of CSE, Siksha O Anusandhan(Deemed to be University),Bhubaneswar,Odisha,India.*

---

**Abstract** Diabetes or Diabetes Mellitus (DM) a major metabolic disorder, can be caused due to age, obesity, lack of exercise, hereditary diabetes, living style, bad diet, high blood pressure, etc., entire body system can be influenced harmfully which will lead to creating diseases in heat, kidney, eye and other organs in the body. Hence early diagnosis and treatment is required in order to prevent the diseases. Recent Machine Learning (ML) techniques are used in accurate predictions and in improving the performance. The paper focuses on ML classification techniques in PIDD (Pima Indian Diabetes Dataset) sourced from UCI ML repository to forecast the likelihood of diabetes in patients with utmost correctness using Python. Six ML techniques were used in the experiment to detect diabetes at an early stage and the performance of these algorithms is validated using measures i.e. Error Rate, Accuracy, Precision, Recall and F-Measure. Logistic regression was found outperform all the ML algorithm showing the maximum accuracy of 79.17% in comparison to other algorithm.

***Key Words***: Diabetes Mellitus (DM), Machine Learning, Logistic Regression, Naive Bayes, K- Nearest Neighbors, Decision Trees, Random Forest, SVM

## 1. INTRODUCTION

Diabetes or Diabetes Mellitus (DM) may be a set of metabolic problems known by high blood sugar levels over a protracted period of our time. Diabetes (DM) is outlined as a bunch of metabolic disorders in the main caused by abnormal insulin secretion and/or action [1]. Symptoms of high aldohexose incorporate excessive voiding, continually feeling thirsty and enlarged hunger [2]. If not treated on time, diabetes will cause serious health problems in a person like diabetic acidosis, hyperosmolar hyperglycemic state, or maybe result in death. This could result in time period complications as well as vas upset, brain stroke, failure, ulcers within the foot, and eye complications etc [3]. Diabetes is caused once the duct gland within the body is unable to come up with insulin in enough amounts or once the cells and tissues within the body fail to utilize the insulin created. Diabetes exists in 3 forms [4]: Diabetes Mellitus Type-1 is characterized by duct gland generating insulin but what's needed by the body, a condition conjointly referred to as "insulin-subordinate diabetes mellitus" (IDDM). Folks littered with type-1 DM need external insulin indefinite quantity to form up for the less insulin created by the duct gland. Diabetes Mellitus Type-2 is marked by the body resisting insulin because the body cells react otherwise to insulin than they traditional would. this could ultimately result in no insulin within the body. This can be otherwise referred to as "non-insulin subordinate diabetes mellitus" (NIDDM) or "adult beginning diabetes". This sort of diabetes is often found in folks with high BMI or people who lead associate degree inactive manner. Gestational diabetes is that the third principle structure that's ascertained throughout physiological state. Generally, for a traditional person, aldohexose levels vary from seventy to ninety-nine milligrams per deciliter. An individual is taken into account diabetic providing the fast aldohexose level is found to be over 126 mg/dL. Within the practice, an individual having an aldohexose concentration of a hundred to one hundred twenty-five mg/dL is taken into account as pre-diabetic [5]. Such an individual is susceptible to the event of sort two diabetes. Over the years, it's been found that folks with the subsequent health characteristics face a larger risk against diabetes:

• A Body Mass Index worth larger than twenty five

• Members of the family littered with diabetes

•People with cholesterol concentration within the body but forty mg/dL prolonged high blood pressure having physiological condition diabetes

• People World Health Organization have suffered from polycystic ovary disorder within the past
• People happiness to ethnic teams like African American, or Native American, or Spanish American, or Asian-pacific aged over forty five years

• Having associate degree inactive manner

When a doctor diagnoses that a person has prediabetes, they recommend the individual higher their manner. Adopting a fitness regime and an honest diet arrange will facilitate forestall diabetes [6]. This analysis aims to work out the danger of development of diabetes in a person. So in this study, we used logistical Regression, Naive Bayes, K- Nearest Neighbors, Decision Trees, Random Forest and SVM machine learning classification algorithms are used and evaluated on the PIDD dataset to seek out the prediction of diabetes during a patient. Experimental performance of all the 6 algorithms is compared on numerous measures and achieved sensible accuracy [7].

## 2. LITERATURE REVIEW

The analysis of connected work provides results on numerous tending datasets, wherever analysis and predictions were disbursed mistreatment numerous ways and techniques. Numerous prediction models are developed and enforced by numerous researchers' mistreatment variants of knowledge mining techniques, machine learning algorithms or additionally combination of those techniques. Dr Saravana Kumar N M, Eswari, Sampath P and Lavanya S (2015) enforced a system mistreatment Hadoop and Map scale back technique for analysis of Diabetic information. This method predicts form of diabetes and additionally risks related to it. The system is Hadoop primarily based and is economical for any tending organization.[8] AiswaryaIyer (2015) used classification technique to check hidden patterns in diabetes dataset. Naïve Thomas Bayes and Decision Trees were employed in this model. Comparison was created for performance of each algorithms and effectiveness of each algorithms was shown as a result.[9] K. Rajesh and V. Sangeetha (2012) used classification technique. They used C4.5 Decision Tree algorithmic program to search out hidden patterns from the dataset for classifying with efficiency.[10] Humar Kahramanli and NovruzAllahverdi (2008) used Artificial neural network (ANN) together with formal logic to predict diabetes.[11] B.M. Patil, R.C. Joshi and Hindu deity Toshniwal (2010) projected Hybrid Prediction Model which incorporates straightforward K-means clump algorithmic program, followed by application of classification algorithmic program to the result obtained from clump algorithmic program. so as to create classifiers C4.5 Decision Tree algorithmic program is employed.[12] Mani Butwall and Shraddha Kumar (2015) projected a model mistreatment Random Forest Classifier to forecast diabetes behaviour.[13] Nawaz Mohamudally1 and Dost Muhammad (2011) used C4.5 Decision Tree algorithmic program, Neural Network, K-means clump algorithmic program and image to predict diabetes. Orabi et al [16] diabetes prediction system designed used the ML decision tree on predicting diabetes at a specific age and the results were satisfactory in predicting the diabetes occurrence at a particular age , with higher accuracy. Pradhan et al [19] coaching and testing of the info for prediction of diabetes was done using Genetic Programming (GP) which was sourced from UCI database.GP was provided optimum accuracy in comparison to other different enforced techniques was helpful for diabetes prediction at low price. Rashid et al. [22] ANN was employed to predict diabetes –chronic sickness and they designed system consisting of two modules and AN was employed with the initial module and the Fasting Blood sugar (FBS) was employed within the second module. Decision tree was employed in finding the symptoms of diabetes on patient's health [23]. Nongyap et al [24] used associate algorithm which classifies the risk of DM and also used the decision tree algorithm

## 3. CLASSFICATION OF MACHINE LEARNING ALGORITHMS

Machine learning (ML) which is a subset of AI that can learn to perform a task with extracted data or models. ML is a set of algorithms that have the capability to learn to perform tasks such as prediction and classification effectively using data. "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E."

Logistic Regression: Logistic regression is a sort of supervised learning which estimates the connection between a binary dependent variable and at least one independent variable by evaluating probabilities with the help of sigmoid function. In contrary to its name, logistic regression is not used for regression problems rather is a type of machine learning classification problem where the dependent variable is dichotomous (0/1, - 1/1, true/false) and experimental variable can binominal, ordinal, interval or ratio-level.

Naïve Bayes: Naive Bayes classifiers work well in many real-world situations like document classification and spam filtering. Naive Bayes classification method is a probabilistic machine learning algorithm based on Bayes theorem described in probability. Even with its simplicity it outperforms other classifiers; hence, it is one of the best classifiers.

K-Nearest Neighbors: Classification is computed from an easy majority vote of the k nearest neighbors of every point.K-Nearest Neighbor (KNN) method can be used to solve problems pertaining to both regression as well as classification, though it is generally being used to solve classification problems in business. Its major advantage is simplicity of translation and low computation time.

Decision Tree: A Decision tree produces a sequence of rules which will be wont to classify the information. A Decision Tree is works on the principle of decision making. It can be described in form of tree and provides high accuracy and stability.

Random Forest: Random forest classifier may be a meta-estimator that matches variety of decision trees on various sub-samples of datasets and uses average to enhance the predictive accuracy of the model and controls over-fitting. The sub-sample size is usually an equivalent because the original input sample size but the samples are drawn with replacement. The Random forest classifier creates multiple decision trees from randomly selected subset of training dataset.

Support Vector Machine: SVM is a supervised classifier in machine learning algorithms that can be used both for regression and classification. It is majorly applied in solving classification problems. The goal of SVM is to classify data points by an appropriate hyperplane in a multidimensional space. A hyperplane is decision boundary to classify data points. The hyperplane classifies the data points with maximum margin between the classes and the hyperplane.

## 4. MATERIALS AND METHODS

The Pima Indian Diabetes Dataset (PIDD) has been employed in this study, provided by the UCI Machine Learning Repository. The dataset has been originally collected from the National Institute of diabetes and organic process and urinary organ Diseases. The dataset consists of some medical distinct variables, like gestation record, BMI, internal secretion level, age, aldohexose concentration, heartbeat force per unit area, skeletal muscle skin fold thickness, diabetes pedigree function etc. This Dataset has 768 patient's data wherever all the patients are feminine and a minimum of twenty-one years previous. The quantity of true cases is 268 (34.90%) and therefore the varieties of false cases are 500 (65.10%), severally, within the dataset. Mahmud et al. [28] used six classification techniques, Artificial Neural Network (ANN), Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF), Logistic Regression (LR) and Naïve Bayes (NB). The main aim of this study is that the prediction of the patient affected by diabetes using the Python Programming codes by using the medical information PIDD. Table-1 shows a quick description of the dataset.

Table 1. Dataset Description

| Dataset Used | No. of Attributes | No. of Instances |
|---|---|---|
| PIDD | 9 | 768 |

The dataset conjointly includes numeric-valued 9 attributes wherever price of 1 category '0' treated as tested negative for diabetes and price of another category '1' is treated as tested positive for diabetes. Dataset description is outlined by Table-1 and therefore the Table-2 represents Attributes descriptions.

Logistic Regression, Naive Bayes, K- Nearest Neighbors, Decision Trees, Random Forest and SVM algorithms are employed in this analysis work.

Experiments are performed using internal cross-validation 10-folds. Accuracy, F-Measure, Recall, Precision and Error Rate are used for the classification of this work. Table-3 defines accuracy measures below:

The different attributes considered for the study are [ "Pg-No of times Pregnant", "Pl-Plasma Glucose Concentration","Pr-Diastolic Blood pressure(mm g)","Sk-Skin fold Thickness(mm)","In-2-Hr serum insulin(mu U/ml)","Ma-BMI(weight-in-kg/height-in-m)2 ","Pe-Diabetes pedigree function","Ag-Age in Years","Cl-Class 'o' or '1'"][30]

## 5. MEASUREMENT

The study used estimations like Accuracy(A) –"specifies the accuracy of the algorithm in predicting instances" , Precision (P)- "Specifies the Classifiers correctness/Accuracy", Recall-( R )-Measures the classifiers completeness or sensitivity , F-Measure(FI) –Weighted average of precision and recall., Error Rate (E)-Errors of the algorithm in predicting instances.

Different measures Formula is given as below

$$A=(TP+TN)/\text{Total no of Samples}$$
$$P=TP/(TP+FP)$$
$$R=TP/(TP+FN)$$
$$FI=2*(P*R)/P+R$$
$$E=(FP+FN)/\text{total no Samples}$$

## 6. RESULTS AND DISCUSSION

The various accuracy estimations are mentioned above and Corresponding classifiers performance over Accuracy, Precision, F-measure, Recall and Error Rate values are listed in Table-4. Where, TP defines True Positive, Tennessee defines True Negative, FP defines False positive, FN defines False Negative. Table-2 represents different performance values of all classification algorithms calculated on varied measures. From Table-2 it's analyzed that provision Regression showing the most accuracy. So, the provision Regression machine learning classifier will predict the possibilities of diabetes with a lot of accuracy as compared to alternative classifiers. So, provision Regression algorithmic rule is taken into account because the best supervised machine learning technique of this experiment as a result of it offers higher accuracy in various to alternative classification algorithms with Associate in measuring accuracy of 79.17 %.

**Table 2. Performance Measures Based Various Classification Algorithms.**

| Classifications Algorithms | Accuracy (%) | Precision | Recall | F-Measure | Error Rate (%) |
|---|---|---|---|---|---|
| Logistic Regression | 79.17 | 0.879 | 0.813 | 0.845 | 20.81 |
| Naive Bayes | 74.48 | 0.814 | 0.795 | 0.805 | 25.52 |
| K- Nearest Neighbors | 74.48 | 0.831 | 0.786 | 0.808 | 25.52 |
| Decision Trees | 71.88 | 0.791 | 0.778 | 0.784 | 28.12 |
| Random Forest | 77.08 | 0.839 | 0.813 | 0.825 | 22.92 |
| SVM | 76.56 | 0.863 | 0.793 | 0.826 | 23.44 |

## 7. CONCLUSION:

There is a concern among physicians how to detect diabetes at its infancy stage. This study had made an effort in deigning the system in predicting the diabetes. The experimental work implemented six ML classification algorithms and the evaluation was done on various measures. The experiment was carried on the PIMA Indian Diabetes data set and the results confirmed the designed system had an accuracy of 79.17% using the supplying regression classification formula. The designed system using this ML algorithm can also be customized in predicting other alternative diseases. The research can be further enhanced in implementing other ML algorithm in improving the prediction of diabetes.

## REFERENCES

[1] American Diabetes Association. Diagnosis and classification of diabetes mellitus.Diabetes Care 2009;32(Suppl. 1):S62–7.

[2] http://diabetesindia.com/

[3] Anjana, R. M., Pradeepa, R., Deepa, M., Datta, M., Sudha, V., Unnikrishnan, R., Bhansali, A., Joshi, S. R., Joshi, P. P., Yajnik, C. S., Dhandhania, V. K. (2011) "Prevalence of diabetes and prediabetes (impaired fasting glucose and/or impaired glucose tolerance) in urban and rural India: Phase I results of the Indian Council of Medical Research–INdiaDIABetes (ICMR–INDIAB) study." Diabetologia 54 (12): 3022-3027.

[4] https://my.clevelandclinic.org/health/diseases/7104-diabetes-mellitus-an-overview

[5] https://www.diabetes.co.uk/diabetes_care/blood-sugar-level-ranges.html

[6] Kaveeshwar, S. A., Cornwall, J. (2014) "The current state of diabetes mellitus in India." The Australasian medical journal 7(1): 45.

[7] Iyer, A., S, J., Sumbaly, R., 2015. Diagnosis of Diabetes Using Classification Mining Techniques. International Journal of Data Mining & Knowledge Management Process 5, 1–14. doi:10.5121/ijdkp.2015.5101, arXiv:1502.03774.

[8] Dr Saravana kumar N M, Eswari T, Sampath P and Lavanya S," Predictive Methodology for Diabetic Data Analysis in Big Data", 2nd

[9] AiswaryaIyer, S. Jeyalatha and Ronak Sumbaly," Diagnosis of Diabetes Using Classification Mining Techniques", International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.5, No.1, January 2015.

[10] K. Rajesh and V. Sangeetha, "Application of Data Mining Methods and Techniques for Diabetes Diagnosis", International Journal of Engineering and Innovative Technology (IJEIT) Volume 2, Issue 3, September 2012.

[11] HumarKahramanli and NovruzAllahverdi,"Design of a Hybrid System for the Diabetes and Heart Disease", Expert Systems with Applications: An International Journal, Volume 35 Issue 1-2, July, 2008.

[12] B.M. Patil, R.C. Joshi and Durga Toshniwal,"Association Rule for Classification of Type-2 Diabetic Patients", ICMLC '10 Proceedings of the 2010 Second International Conference on Machine Learning and Computing, February 09 - 11, 2010.

[13] Mani Butwall and Shraddha Kumar," A Data Mining Approach for the Diagnosis of Diabetes Mellitus using Random Forest Classifier",

International Journal of Computer Applications, Volume 120 - Number 8,2015.

[14] Perveen, S., Shahbaz, M., Guergachi, A., Keshavjee, K., 2016. Performance Analysis of Data Mining Classification Techniques to Predict Diabetes. Procedia Computer Science 82, 115–121. doi:10.1016/j.procs.2016.04.016.

[15] Nai-Arun, N., Sittidech, P., 2014. Ensemble Learning Model for Diabetes Classification. Advanced Materials Research 931 - 932, 1427–1431. doi:10.4028/www.scientific.net/AMR.931-932.1427.

[16] Orabi, K.M., Kamal, Y.M., Rabah, T.M., 2016. Early Predictive System for Diabetes Mellitus Disease, in: Industrial Conference on Data Mining, Springer. Springer. pp. 420–427.

[17] Priyam, A., Gupta, R., Rathee, A., Srivastava, S., 2013. Comparative Analysis of Decision Tree Classification Algorithms. International Journal of Current Engineering and Technology Vol.3, 334–337. doi:JUNE 2013, arXiv:ISSN 2277 - 4106.

[18] Esposito, F., Malerba, D., Semeraro, G., Kay, J., 1997. A comparative analysis of methods for pruning decision trees. IEEE Transactions on Pattern Analysis and Machine Intelligence 19, 476–491. doi:10.1109/34.589207.

[19] Bamnote, M.P., G.R., 2014. Design of Classifier for Detection of Diabetes Mellitus Using Genetic Programming. Advances in Intelligent Systems and Computing 1, 763–770. doi:10.1007/978-3-319-11933-5.

[20] Sharief, A.A., Sheta, A., 2014. Developing a Mathematical Model to Detect Diabetes Using Multigene Genetic Programming. International Journal of Advanced Research in Artificial Intelligence (IJARAI) 3, 54–59. doi:doi:10.14569/IJARAI.2014.031007.

[21] Pradhan, P.M.A., Bamnote, G.R., Tribhuvan, V., Jadhav, K., Chabukswar, V., Dhobale, V., 2012. A Genetic Programming Approach for Detection of Diabetes. International Journal Of Computational Engineering Research 2, 91–94.

[22] Tarik A. Rashid, S.M.A., Abdullah, R.M., Abstract, 2016. An Intelligent Approach for Diabetes Classification, Prediction and Description. Advances in Intelligent Systems and Computing 424, 323–335. doi:10.1007/978-3-319-28031-8.

[23] Han, J., Rodriguez, J.C., Beheshti, M., 2008. Discovering decision tree based diabetes prediction model, in: International Conference on Advanced Software Engineering and Its Applications, Springer. pp. 99–109.

[24] Nai-Arun, N., Moungmai, R., 2015. Comparison of Classifiers for the Risk of Diabetes Prediction. Procedia Computer Science 69, 132–142. doi:10.1016/j.procs.2015.10.014.

[25] Wilson RA, Keil FC. The MIT encyclopaedia of the cognitive sciences. MIT Press; 1999.

[26] Mitchell T. Machine learning. McGraw Hill0-07-042807-7; 1997 2.

[27] Russell, Stuart; Norvig, Peter (2003) [1995]. Artificial Intelligence: A Modern Ap-proach (2nd Ed.). Prentice Hall. ISBN 978-0137903955.

[28] Mahmud, S. H., Hossin, M. A., Ahmed, M. R., Noori, S. R. H., & Sarkar, M. N. I. (2018, August). Machine Learning Based Unified Framework for Diabetes Prediction. In Proceedings of the 2018 International Conference on Big Data Engineering and Technology (pp. 46-50).

[29] Kayaer, K., Tulay, 2003. Medical diagnosis on Pima Indian diabetes using general regression neural networks, in: Proceedings of the interna-tional conference on artificial neural networks and neural information processing (ICANN/ICONIP), pp. 181–184.

[30] Choubey, D.K., Paul, S., Kumar, S., Kumar, S., 2017. Classification of Pima Indian diabetes dataset using naive bayes with genetic algorithm as an attribute selection, in: Communication and Computing Systems: Proceedings of the International Conference on Communication and Computing System (ICCCS 2016), pp. 451–455.

**BIOGRAPHY**

**Mr.Jitranjan Sahoo** is presently working as the Assistant Professor at Interscience Institute of Management & Technology, Bhubaneswar. He has received the University Gold Medal for the best academic performance in MBA from "IBCS- Siksha 'O' Anusandhan University (Deemed to be)" on 16th November 2019. He is the Former Debt Manager at ICICI Bank-Durg and Former Intern at Vadodara Black Oil Terminal, IOCL. He has several research paper presentations at international & national Seminars and webinars as well, conducted by prestigious universities and management institutions. He has attended various management workshops and value added seminars and he has a passion for teaching, training, mentoring and article writing.
(EMAIL: jitranjansahoo@gmail.com)