ISSN: 2096-3246
Volume 53, Issue 02, June 2021

https://doi.org/10.5281/zenodo.12726456

Machine learning based Diabetics classification and prediction

SIRIKONDA ANANTHNAG, MANGALI ANIL KUMAR, VIJAYA BHASKAR MADGULA

Assistant Professor [1,2,3]

ananthnagsvit9f@gmail.com, anilkumarcse02@gmail.com, vijaya.bhaskar2010@gmail.com

Department of CSE, Sri Venkateswara Institute of Technology,

N.H 44, Hampapuram, Rapthadu, Anantapuramu, Andhra Pradesh 515722

Keywords:

Machine learning techniques, Diabetes, prediction, Dataset

ABSTRACT

An excess of glucose in the bloodstream leads to the disease known as diabetes. Ignoring diabetes for an extended period of time may lead to serious complications, including but not limited to: renal disease, high blood pressure, vision loss, and heart difficulties. When caught early enough, diabetes is manageable. In order to accomplish this goal, our organisation will use a variety of machine learning approaches to improve the accuracy of early diabetes predictions in human or patient settings. By constructing models from patient data sets, machine learning approaches provide a stronger foundation for prediction. In this study, we will use the dataset to forecast the occurrence of diabetes by using clustering and classification algorithms from machine learning. These include Random Forest (RF), Gradient Boost (GB), K-Nearest Neighbour (KNN), Decision Tree (DT), and Logistic Regression (LR). When comparing several models, accuracy may vary significantly. Based on the results of the experiment, it seems that the version may accurately forecast the onset of diabetes. Compared to other strategic vision computers, Random Forest outperforms them in terms of accuracy.

https://doi.org/10.5281/zenodo.12726456

INTRODUCTION

Worldwide, diabetes is a major health concern. Obesity, excessive blood sugar, and other similar conditions may lead to diabetes mellitus. As a result of its effect on insulin, it improves blood sugar levels and causes crabs to have an aberrant metabolism. When insulin production drops below normal levels, diabetes sets in. Based on data from the World Health Organisation Worldwide Health Organisation (WHO), some 422 million people suffer from diabetes, disproportionately in countries with poor per capita income or inadequate healthcare services. A total of 490 billion will be reached by all twelve months of 2030. Nevertheless, diabetes is a major health concern in many nations, including India, Canada, China, and many more. With over 100 million people calling India home today, the real India has 40 million people living with diabetes. Worldwide, diabetes ranks as the top killer. A person's way of life may be impacted by early diabetes prognosis and management. This panel aims to do this by examining the many characteristics linked to diabetes in order to make predictions about the disease. To achieve this goal, we use a combination of machine learning algorithms, classification observations, and the Pima Indian Diabetes dataset to make diabetes predictions [1]. The goal of machine learning is to deliberately train computers or other automated systems. Machine learning algorithms generate various group and category models from the obtained data set, providing an environmentally friendly foundation for information acquisition. With these information in hand, diabetes may be better predicted. Predictions may be successfully made using a variety of machine learning approaches, but selecting a suitable methodology can be challenging. Hence, in order to forecast this stimulus, we use aggregation algorithms in conjunction with data set categorization [2].

A large body of comparable studies has already predicted the development of diabetes. While prior studies did utilise information from the PIMA Indians Database, which is given by the National Institutes of Diabetes and Digestive and Kidney Diseases, this experiment is unique in that it is based on data collected via the Kaggle Competition.

gathering information both online and off. However, in contrast to other studies, we used dynamic recordings collected from the individuals' trajectories to build our machine learning models. Also, type 1 and type 2 diabetes have been used as shams in conventional research on a frequent basis. Having said that, we saw that both type 1 and type 2 diabetes were treated as if they were one and the same [3]. Just so you know, walking may be dangerous for those with diabetes and peripheral neuropathy. Findings from research examining this component may vary greatly, however, since diabetics are quite diverse in terms of the frequency and severity of diabetes complications. Blood analysis or questionnaires and surveys were the mainstays of earlier research that attempted to predict diabetes. The writers' non-invasive volume-based classification of health and diabetes-specific workouts is a pastime of theirs. So, instead of taking blood samples or having patients fill out surveys and questionnaires, the present study's logic shifted to propose a method for determining whether an elderly person has diabetes using data obtained almost exclusively on foot.

I. LITERATURE SURVEY

the group headed by K.Vijiya Kumar and colleagues [4] Utilising the random forest rule specified in the tool learning method, the suggested algorithm for diabetes prediction enhances an existing tool's ability to accurately forecast a patient's risk of developing diabetes at an earlier stage. Thrillingly, the suggested model can accurately and rapidly forecast diabetes, and the end result proved that the predictor is capable of making such predictions.
Nonso Nnamoko et al. [5] The use of a supervised preparation procedure has been made in order to forecast the start of diabetes. Each group is given one of five popular classifiers, and their output is combined using a meta-classifier. The findings are contrasted with other studies that have used the

same dataset in their literature reviews. The suggested method allows for more precise prediction of when diabetes will start.
This work is by Tejas N. Joshi and colleagues. [6] Using three distinct methods for identifying monitoring devices—support vector machines (SVMs), logistic regression (RR), and artificial neural networks (ANNs)—aspirational machine learning approaches to diabetes prediction propose. A robust technique for the early diagnosis of diabetes is proposed in this mission.

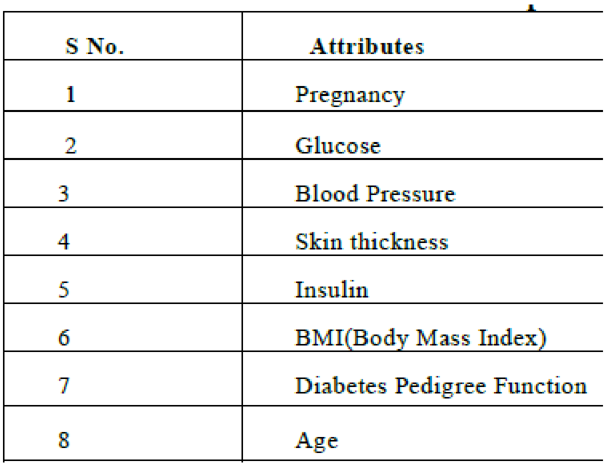
"Deeraj Shetty et al." [7] With the use of data mining, we can put together a smart diabetes prediction system that can analyse diabetes records in a database. With this gadget, you may analyse a diabetes database using various diabetes variables and use techniques like Bayesian and KNN (K-Nearest Neighbour) to forecast the onset of diabetes.
*Sarwar, Muhammad Azeem, et al. 8 Using healthcare study tool algorithms for diabetes prediction: a suggested perspective They have studied the tool using six different algorithms, and they compare and contrast the algorithms' accuracy and performance. Which diabetes prediction method is more relevant is shown by comparing the device-only research techniques employed in this study. Researchers now consider diabetes prediction a hobby, and they train algorithms to determine whether a patient has diabetes or not by applying the appropriate classifier to the dataset. The categorization approach isn't necessarily state-of-the-art, according to earlier studies. In order to address the issues highlighted by prior research, a system is required for diabetes prediction in computers.

II. PROPOSED METHODOLOGY

The work aims to explore a model for predicting diabetes with higher accuracy. We experimented with unique classification and clustering algorithms to predict diabetes. Below we will talk briefly about the section.

A. Dataset Description- the data is gathered from UCI repository which is named as Pima Indian Diabetes Dataset. The dataset has many attributes of 768 patients.

Table.1 Dataset Description



S No.	Attributes
1	Pregnancy
2	Glucose
3	Blood Pressure
4	Skin thickness
5	Insulin
6	BMI(Body Mass Index)
7	Diabetes Pedigree Function
8	Age

The 9th attribute is class variable of each data points. This class variable shows the outcome 0 and 1 for diabetics which indicates positive or negative for diabetics.

ISSN: 2096-3246
Volume 53, Issue 02, June 2021

https://doi.org/10.5281/zenodo.12726456

Distribution of Diabetic patient- We made a model to predict diabetes however the dataset was slightly imbalanced having around 500 classes labeled as 0 means negative means no diabetes and 268labeled as 1 means positive means diabetic.

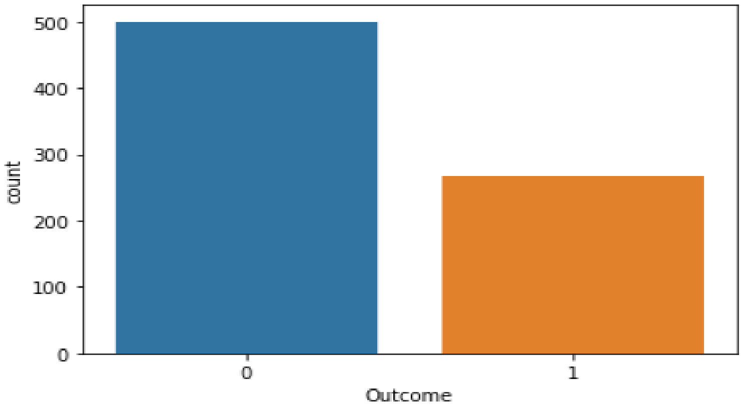


Fig. 1 Ratio of Diabetic and Non-Diabetic Patient

Data Preprocessing- The most crucial step is data pre-processing. The majority of data sets pertaining to healthcare have errors or missing values that can compromise their usefulness. Data pre-processing is carried out to enhance the efficacy and quality of the results acquired from the mining process. This step is critical for getting accurate results and making good predictions when using Machine Learning Techniques on the dataset. Two phases of pre-processing are required for the Pima Indian diabetes dataset.

Missing Values removal- Remove all the instances that have zero (0) as worth. Having zero as worth is not possible. Therefore, this instance is eliminated. Through eliminating irrelevant features/instances we make feature subset and this process is called features subset

selection, which reduces dimensionality of data and help to work faster.

1). Splitting of data- After cleaning the data, data is normalized in training and testing the model. When data is spitted then we train algorithm on the training data set and keep test data set aside. This training process will produce the training model based on logic and algorithms and values of the feature in training data. Basically, aim of normalization is to bring all the attributes under same scale.

C. Apply Machine Learning- When data has been ready, we apply Machine Learning Technique. We use different classification and ensemble techniques, to predict diabetes. The methods applied on Pima Indians diabetes dataset. Main objective to apply Machine Learning Techniques to analyze the performance of these methods and find accuracy of them, and also been able to figure out the responsible/ important feature which play a major role in prediction.

The Techniques are follows-

1) Support Vector Machine- Support Vector Machine also known as svm is a supervised machine learning algorithm. Svm is most popular classification technique. Svm creates a hyperplane that separate two classes. It can create a hyperplane or set of hyperplanes in highdimensional space. This hyper plane can be used for classification or regression also. Svm differentiates instances in specific classes and can also classify the entities which are not supported by data. Separation is done by through hyperplane performs the separation to the closest training point of any class.

Algorithm-

- Select the hyper plane which divides the class better.
- To find the better hyper plane you have to calculate the distance between the planes and the data which is called Margin.
- If the distance between the classes is low then the chance of miss conception is high and vice versa. So, we need to
- Select the class which has the high margin. Margin = distance to positive point + Distance to negative point.

2) K-Nearest Neighbor - KNN is also a supervised machine learning algorithm. KNN helps to solve both the classification and regression problems. KNN is lazy prediction technique. KNN assumes that similar things are near to each other. Many times, data points which are similar are very near to each other. KNN helps to group new work based on similarity measure. KNN algorithm record all the records and classify them according to their similarity measure. For finding the distance between the points uses tree like structure. To make a prediction for a new data point, the algorithm finds the closest data points in the training data set — it's nearest neighbors. Here **K**= Number of nearby neighbors, it's always a positive integer. Neighbor's value is chosen from set of class. Closeness is mainly defined in terms of Euclidean distance. The Euclidean distance between two points P and Q i.e. P (p1,p2, …. Pn) and Q (q1, q2,..qn) is defined by the following equation:-

$$d(P,Q) = \sum_{i=1}^{n} (P_i - Q_i)^2$$

Algorithm-

- Take a sample dataset of columns and rows named as Pima Indian Diabetes data set.

- Take a test dataset of attributes and rows.

- Find the Euclidean distance by the help of formula-

Then, Decide a random value of K. is the no. of nearest neighbors

- Then with the help of these minimum distance and Euclidean distance find out the nth column of each.

3) Find out the same output values

4) **Decision Tree-** Decision tree is a basic classification method. It is supervised learning method. Decision tree used when response variable is categorical. Decision tree has tree like structure-based model which describes classification process based on input feature. Input variables are any types like graph, text, discrete, continuous etc. Steps for Decision Tree **Algorithm-**

- Construct tree with nodes as input feature.

- Select feature to predict the output from input feature whose information gain is highest.

- The highest information gain is calculated for each attribute in each node of tree.

- Repeat step 2 to form a subtree using the feature which is not used in above node.

5) **Logistic Regression**- Logistic regression is also a supervised learning classification algorithm. It is used to estimate the probability of a binary response based on one or more predictors. They can be continuous or discrete. Logistic regression used when we want to classify or distinguish some data items into categories. It classifies the data in binary form means only in 0 and 1 which refer case to classify patient that is positive or negative for diabetes. Main aim of logistic regression is to best fit which is responsible for describing the relationship between target and predictor variable. Logistic regression is a based on Linear regression model. Logistic regression model uses sigmoid function to predict probability of positive and negative class.

6) Random Forest – It is type of ensemble learning method and also used for classification and regression tasks. The accuracy it gives is grater then compared to other models. This method can easily handle large datasets. Random Forest is developed by Leo Bremen. It is popular ensemble Learning Method. Random Forest Improve Performance of Decision Tree by reducing variance. It operates by constructing a multitude of decision trees at training time and outputs the class that is the mode of the classes or classification or mean prediction (regression) of the individual trees.

Algorithm-

- The first step is to select the ―R‖ features from the total features ―m‖ where R<<M.
- Among the ―R‖ features, the node using the best split point.
- Split the node into sub nodes using the best split.
- Repeat a to c steps untill‖ l‖ number of nodes has been reached.
- Built forest by repeating steps a to d for

―a‖ number of times to create ―n‖ number of trees

7) Gradient Boosting - Gradient Boosting is most powerful ensemble technique used for prediction and it is a classification technique. It combines week learner together to make strong learner models for prediction. It uses Decision Tree model. it classifies complex data sets and it is very effective and popular method. In gradient boosting model performance improve over iterations.

Algorithm-

- Consider a sample of target values as P
- Estimate the error in target values.
- Update and adjust the weights to reduce error M.

ISSN: 2096-3246
Volume 53, Issue 02, June 2021

https://doi.org/10.5281/zenodo.12726456

SYSTEM ARCHITECTURE

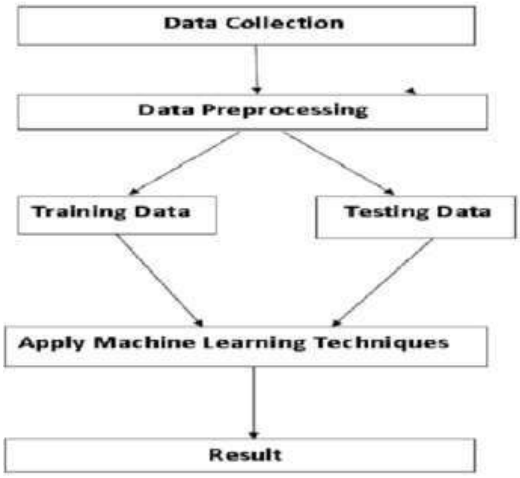


Fig.2 System Architecture

EXPERIMENTAL RESULTS

Various approaches were used in this project. The suggested solution is based on python and employs several ensemble and classification techniques. These strategies are often used in Machine Learning to provide the highest level of accuracy while working with data. Based on the results of this study, the random forest classifier is the most effective. Our top-tier performance and predictions are the result of our utilisation of state-of-the-art Machine Learning algorithms. These Machine Learning algorithms produced the outcome shown in the figure.

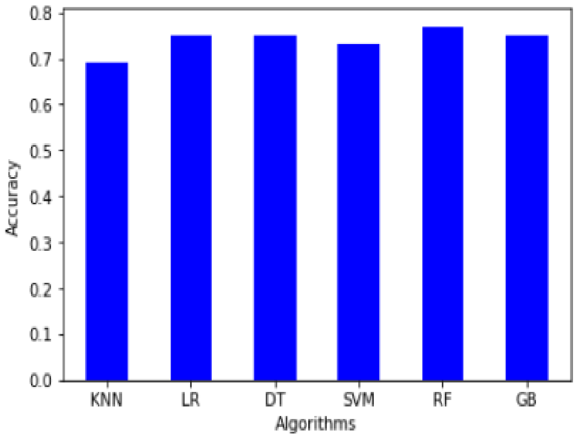


Fig.3 Accuracy Result of Machine learning methods

Here feature played important role in prediction is presented for random forest algorithm. The sum of the importance of each feature playing major role for diabetes have been plotted, where X-axis represents the importance of each feature and Y-Axis the names of the features.

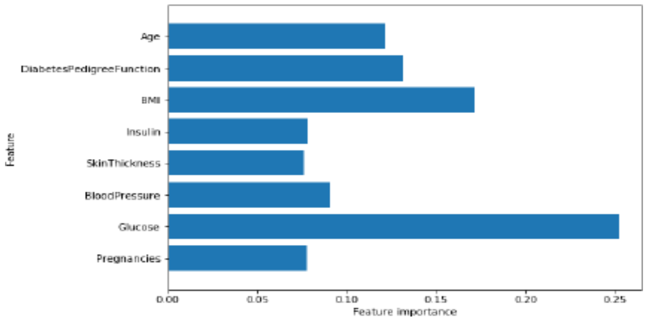


Fig.4 Feature Importance Plot for Random Forest

CONCLUSION

The project's overarching goal—the development and implementation of a system for diabetes prediction using machine learning methods—and the evaluation of that system's performance—were both accomplished. Classifiers such as Support Vector Machines (SVMs), K-NNs, Random Forests, Decision Trees, Logistic Regression, and Gradient Boosting are used in the proposed approach's ensemble learning and classification procedures. Additionally, a classification accuracy of 77% was achieved. In order to find a cure for diabetes and save lives, the experimental findings may help health care providers make early predictions and decisions.

REFERENCES

"Diabetes fact sheet in Korea 2018," written by KS Park and published by the Korean Diabetes Association in 2018.

Studying and forecasting diabetes patients' outcomes with the use of judgement tree, pp. 829–833, 2013 (The Institute of Electronics and Information Engineers).

A data-driven strategy to predicting diabetes and cardiovascular disease using machine learning was published in BMC Med. Inform. Decis. Mak., volume 19, issue 211, November 2019, by A Dinh, S Miertschin, A Young, and SD Mohanty.

[4] In the 24th International Conference on Automation and Computing (ICAC) held in Newcastle upon Tyne, United Kingdom, in September 2018, pp. 1-6, MA Sarawr, N. Kamal, W. Hamid, and MA Shah discuss the use of machine learning algorithms for diabetes prediction in healthcare.

5. "Predicting Diabetes Onset: an Ensemble Supervised Learning Approach" by Nonso Nnamoko, Abir Hussain, and David England presented at the 2018 IEEE Congress on Evolutionary Computation (CEC).

[6] "Diabetes Disease Prediction Using Data Mining" (Deeraj Shetty, Kishor Rit, Sohail Shaikh, and Nikita Patil, 2017). International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS).

In their article "Intelligible support vector machines for diagnosis of diabetes mellitus," Nahla B., Andrew et al. addressed this topic. "Information Technology in Biomedicine," (July 2010), 14th IEEE Transactions,

In their 2015 article "Classification of Diabetes Mellitus Using Machine Learning Techniques," A.K. Dewangan and P. Agrawal of the International Journal of Engineering and Applied Sciences published their findings.

Using machine learning classification approaches for the prediction of type 2 diabetes, Prerna and Garg (2020) published in Procedia Computer Science, volume 167, pages 706-716.

The authors of the following work: "Development of customised senior fitness service technology based on sports evaluation" (JY Lee, JH Jun, KD Joong, KK Kim, CH Yeom, HK Min, and YK Kim, 2010). Corporate-Academic Partnership, Dong-A University