ISSN: 0974-5823 Vol. 7 (Special Issue, Jan.-Mar. 2022)

International Journal of Mechanical Engineering

Detection of Diabetics at an Early Stage through the Machine Learning Algorithms

Chandrashekhar Kumbhar[1] , Dr. Abid Hussain[2]

[1] Research Scholar, School of Engineering and Technology, Career Point University, Kota, Rajasthan, India

[2] Associate Professor, School of Computer Applications, Career Point University, Kota, Rajasthan, India

Abstract **-**These days in India, diabetics make up a sizable proportion of the population, and the disease's prevalence is steadily expanding. Diabetes is becoming an increasingly serious problem in India, with an estimated 8.7 percent of the population between the ages of 20 and 70 suffering from the condition. The growing incidence of diabetes and other noncommunicable illnesses may be attributed to a confluence of causes, including rapid urbanization, sedentary lifestyles, bad diets, the use of tobacco products, and an increase in life expectancy. Diabetes is responsible for the deaths of one million individuals each and every year. Although we are unable to stop people from passing away, we can lessen the number of deaths that occur by identifying diabetics at an earlier stage. The disease known as diabetes mellitus is characterized by sugar levels in the body that remain excessively elevated throughout time. As a consequence of this, [5] it causes harm to a significant number of the systems in the body, including the neurons and blood vessels. This disease has a good chance of being detected in its early stages, which is a factor that may contribute to the prevention of deaths in people. In order to get insights that may be used to decision-making, it is necessary to first gather the data to be analyzed and then evaluate the data. Massive data sets and measurements may be seen with the use of charts, graphs, and other types of visualizations.

In this study, we will explore several machine learning techniques for diabetes prediction. If you look at the architecture shown below, you'll find that the data is gathered from all over the world, but the vast majority of the entries [8] come from the local area. In this work, we will investigate the feasibility of using machine learning methods to forecast the number of diabetes. due to the fact that we obtained the information via the use of a survey form and one of the pathology labs in Pune. In the R programming language, we have used the SVM and Random Forest algorithms. The R programming language is a statistical programming language that is extensively used in the data science arena for the implementation of machine learning algorithms.

Index Terms - Machine learning, diabetics, Data visualisation, R Programming, detection

I. Introduction

Diabetes mellitus, sometimes known as diabetes or just diabetes, is a set of conditions in which the body is unable to maintain a balance between the amount of sugar that is present in the blood. Diabetes may also be referred to as diabetes. Insulin is only one of a number of hormones that cooperate with one another to maintain a healthy level of blood sugar in a person. The pancreas is a very small organ that may be found in close proximity to the liver and the stomach. It is the organ that is in charge of producing insulin. The pancreas is responsible for the release of extra vital enzymes that assist in [3]the digestion of food, and it is one of the organs that is involved in this process. Insulin makes it easier for glucose to travel from the blood into the cells of the liver, muscle, and fat, where it may be used as fuel. Glucose is one of the main sources of energy in the body. Diabetes is a life-threatening condition that may be caused by having high blood pressure. High blood pressure can induce diabetes. If the condition is not treated at the appropriate time, it might lead to a number of consequences. The discipline of machine learning has now entered the scene as an important component of this journey. The goal of the study is to ascertain the possibility that specific diseases would manifest themselves throughout the formative years of a person's life at various points in time.

There has been a significant increase in the number of people living with diabetes in India since the beginning of the previous decade. Diabetes affects something in the vicinity of 77 million [2] people in India at this point,

as determined by the results of a research that was carried out by the International Diabetes Federation (IDF). The estimates published by the IDF place the crude prevalence rate in India's urban regions somewhere in the vicinity of 9 percent. Diabetes affects a far higher percentage of the population in India than it does in any other country on the face of the planet.[9] India has the second highest number of children with type 1 diabetes in the world, behind only the United States of America.

The continuous hyperglycemia that is caused by diabetes is associated to the long-term damage, breakdown, and failure of various systems. These organs include the eyes, kidneys, nerves, heart, and veins.

Around the world, numerous chronic illnesses are widespread, both in developing as well as industrialised countries. diabetes is a metabolic disease that affects blood sugar levels by increasing or decreasing the quantity of insulin produced. [2]Human bodily components such as the eyes, kidneys, heart and nerves are all affected by diabetes.

II. Literature Survey

The goal of this work is to develop an intelligent system that will be known as the DeeDee system and will be able to direct a diabetic through their typical activities. This system would be implemented on a mobile phone, which is a portable electronic device that one may carry on themselves at all times of the day. The development of an original method for estimating glycemia levels is the fundamental contribution made by this system. They have said that the currently available techniques for glycaemia prediction are dependent on continuous monitoring of the patient. This implies that the patient is put into a controlled environment (such as a hospital). On the other hand, his system wants to anticipate glycaemia and, as a consequence, to support the patient in his or her day-to-day life. On the other hand, on the other hand, his system aims to predict glycaemia. In his illustration, the amount of information collected from the patient is a lot lower than what it would be in a situation that included continuous monitoring of the patient. His method relies on the prior experiences of diabetics that were comparable to the one that is being experienced now in order to arrive at an accurate estimation of the value of glycaemia. In order for the system as a whole to be able to carry out the activities for which it was designed, one of the fundamental principles that must be met is the generation and management of diabetics' profiles. [15] Each profile is produced in such a manner that it explains and combines similar scenarios that have happened throughout the diabetic's life, i.e., how the diabetic's body has responded in certain settings. These scenarios have been collected throughout the course of the diabetic's lifetime. This is handled in an automated manner. In order to effectively identify the occurrences that are similar to one another, they need the context components that were collected at that time point. The patient's cell phone, which the patient carries with them at all times, is continuously gathering information on these traits. This article offers a number of recommendations for enhancing the standard of living of diabetics, one of which is the creation of a unified system that is compatible with a variety of different kinds of *medical technology*.

The primary objective of this investigation [16] is to develop a web application that is predicated on the increased accuracy of prediction provided by a highly effective machine learning algorithm. They employed a benchmark dataset called Pima Indian, which is capable of forecasting the start of diabetes based on diagnostics approach. This dataset was used by these researchers. An artificial neural network (ANN) exhibits a considerable increase in accuracy with a prediction rate of 82.35 percent, which is what leads us to design an interactive web application for diabetes prediction.

To categorize a person's data into two categories, "Yes" and "No," this [17]paper uses 10 criteria, including age, family history of alcoholism, smoking, etc. to do just that. SVM, KNN, ANN and Naive Bayes are four machine learning algorithms used in this study. The cumulative result is derived from the mode of each of the four outputs.

The objective of this proposed technique [18] is to zero in on the characteristics that have a role in the diagnosis of diabetes mellitus in its earliest stages utilizing predictive analysis. According to the findings, the decision tree algorithm and the Random forest model are the ones that perform the best when used to the analysis of diabetes data. The respective specificities of these two models are 98.20 and 98.00 percent. According to the result of a naive Bayesian analysis, the level of accuracy achieved is 82.30 percent. This study additionally generalizes the selection of appropriate characteristics from the dataset in order to increase the accuracy of the classification.

At an early stage, experiments are carried out using the Pima Indians Diabetes Database (PIDD), which is obtained from a computer at the University of California, Irvine. The effectiveness of each of the three algorithms is assessed using a variety of metrics, including precision, accuracy, F-measure, and recall, amongst others. The number of occurrences that are properly and wrongly categorised is used to assess accuracy. The findings that were collected indicate that Naive Bayes surpasses other algorithms, with the maximum accuracy of 76.30 percent being achieved by it. Receiver Operating Characteristic (ROC) curves are used in an appropriate and methodical manner to carry out the verification process for these outcomes.

In this particular study[20], the researchers attempted to forecast cases of diabetes mellitus by using decision trees, random forests, and neural networks. The data collection contains information about patients' physical examinations at a hospital in Luzhou, China. It has fourteen characteristics in all. In this particular investigation, the models were evaluated using five separate rounds of cross validation. We picked certain ways that have the superior performance to conduct independent test trials using in order to verify the methods' potential for universal application. For the training set, we chose at random the data of 68994 healthy persons and 68994 diabetes patients. As a result of the imbalance in the data, we arbitrarily extracted 5 sets of data. The conclusion may be found by taking the average of these five separate tests. In this investigation, we decreased the dimensionality by using the techniques of principal component analysis (PCA) and minimal redundancy maximum relevance (mRMR). When all of the criteria were considered, the findings indicated that random forest prediction could achieve the maximum level of accuracy (ACC = 0.8084) possible.

The purpose of this work is to develop a diabetes diagnosis software that is simple to use, accurate, and effective at a low cost. This software will have a graphical user interface and will be able to predict diabetes.[21] Non-governmental organizations will be able to use this software to diagnose people who come from economically disadvantaged sections.This document makes reference to a project that has the goal of categorizing a person's data into two groups, 'Yes' and 'No,' depending on 10 different characteristics. Some of these factors include age, family history, being a smoker or an alcoholic, etc.In this work the author aggregated result is generated by taking the mode of the four outputs from four different machine learning methods. These algorithms include the Naive Bayes algorithm, the SVM algorithm, the KNN algorithm, and the ANN algorithm.

In this article[22], author presents a Diabetes Prediction Decision Support System (DSS) that is built on Machine Learning (ML) approaches. The traditional methods of machine learning were contrasted with the deep learning techniques. For the traditional approach to machine learning, we looked at two of the most popular classifiers: the Support Vector Machine, sometimes known as SVM, and the Random Forest (RF). On the other hand, for Deep Learning (DL), they made use of a completely Convolutional Neural Network (CNN) in order to predict and identify diabetic patients. The suggested method was tested using the publicly accessible Pima Indians Diabetes database, which had a total of 768 samples, each of which had 8 characteristics. 500 of the samples were classified as not being from diabetes individuals, whereas the remaining 268 were. The results that were achieved using DL, SVM, and RF in terms of their overall accuracy were as follows: 76.81 percent, 65.38 percent, and 83.67 percent accordingly. The findings of the experiments indicate that RF is superior than deep learning and SVM approaches in terms of its ability to accurately forecast diabetes.

Utilizing relevant characteristics, developing a prediction algorithm by utilizing machine learning, and determining the best classifier to utilize in order to get the most accurate results possible when compared to clinical outcomes are the goals of this study as per the author [23]. The objective of the suggested strategy is to zero in on the characteristics that are useful in the early diagnosis of diabetes mellitus by making use of predictive analysis. According to the result of author, the decision tree algorithm and the Random forest model are the ones that perform the best when used to the analysis of diabetes data. The respective specificities of these two models are 98.20 and 98.00 percent. According to the result of a naive Bayesian analysis, the level of accuracy achieved is 82.30 percent. In addition, this study generalizes the selection of the most useful characteristics from the dataset in order to increase the classification accuracy.

According to author[24], the purpose of this study is to evaluate the various classifiers in order to determine which ones are best able to determine the likelihood of illness in patients with the highest level of specificity and precision. In other words, the objective is to find out which ones are most accurate. Experimentation using classification algorithms such as K Nearest Neighbor (KNN), Decision Tree (DT), Naive Bayes (NB), Support Vector Machine (SVM), Logistic Regression (LR), and Random Forest has been carried out on the Pima Indians Diabetes dataset, which can be accessed online at the UCI Repository. The results of this work can be found in the UCI Repository (RF). The dataset was subjected to these categorization methods, which made use of nine

distinct characteristics. The performance of a classifier is evaluated with the help of metrics such as precision, recall, and accuracy, and the results are estimated with the help of both correct and incorrect instances. The results showed that the Logistic Regression (LR) method produces a greater degree of accuracy than other algorithms, with a score of 77.6 percent, in contrast to the performances of those other algorithms. Support-Vector Machine (SVM), Logistic Regression, Classification, K-Nearest-Neighbor (KNN), Decision Tree (DT), Naive Bayes (NB), and Support-Vector Machine (KNN) (LR).

The purpose of this research is to devise a technique that, [25] if put into practice, would make it possible for a patient's level of risk for developing diabetes to be evaluated with a greater degree of accuracy. Strategies for classification are used in the process of creating models. Some examples of these techniques are decision trees, artificial neural networks (ANN), naïve bayes networks (NBN), and support vector machine (SVM) algorithms. Precisions of 85 percent are provided by the models for the Decision Tree method, 77 percent are provided by the Naive Bayes algorithm, and 77.3 percent are provided by the Support Vector Machine algorithm. The findings provide conclusive evidence that the protocols ensure a high level of data precision.

This inquiry was conducted by the author with the intention of providing a novel conceptual framework for the classification of ideas. Using the angiosome concept of the foot, we categorized the plantar thermographic patterns that were found into twenty separate categories. Thermographic images obtained from 32 healthy volunteers and 129 diabetes patients without ulcers who were recruited from the Diabetes Foot Outpatient Clinic at the University of Tokyo Hospital were classified according to the framework categories stated earlier. The thermographic patterns of more than 65 percent of feet in the normal group were assigned to the two typical categories, which included the 'butterfly pattern' among the 20 categories; on the other hand, the thermographic patterns of 225 feet (87.2 percent) in the diabetic groups were variously assigned to 18 out of the 20 categories. The butterfly pattern was one of the typical thermographic patterns. This is the first study to present accurate plantar thermographic patterns, and the results suggest that diabetes patients experience more significant alterations than the normal subjects. The measurements taken from the plantar temperature have shown these conclusions. Thermography is one of the screening techniques that may be used to examine the patient's blood circulation during normal foot care as well as during surgical operations. [26] This may be one of the screening methods that is employed.

III. Dataset Information

For instructional purposes, we have gathered data from a Google form and some data from a pathology lab. The most essential and novel parameter that we have taken into consideration is the HBA1C test information, as well as family history, a few symptoms, and a few general inquiries from users.

The purpose for this is to get rid of an outdated PIMA dataset, which was the motivation for it. Since it is an ancient method and has been used by a great number of researchers for the purpose of identifying diabetics, despite the fact that it lacks some essential characteristics that were meant to be included but, alas, were not, the method is considered to be unreliable. In order to circumvent this, we have compiled our own own dataset.

We have total 19 attributes and almost 750 entries of diabetics and non diabetics individuals.

IV. Dataset Visualisation in R

3D scatter Plot

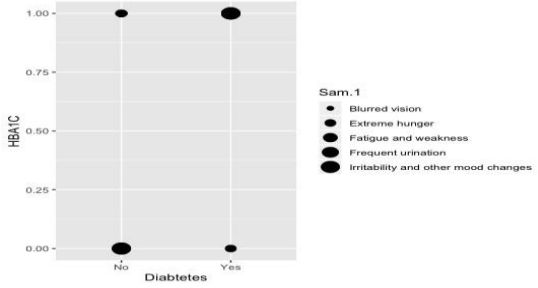


Figure 1 : 3D scatter plot

A scatter plot is similar to a scatter plot, however a three-dimensional scatter plot has three variables instead of two. With the proviso that x, y, and z or f (x, y) are all real numbers. A scatter plots shows the relationship between two variables, it represents how two variables are correlated with each other. Here we have used 3D scatter plot which using x axis as Diabetes , y axis as Sam 2 and z axis as HBA1C.

Bubble chart

A bubble chart is a kind of the scatter plot that may be used to investigate the connections between three different numerical factors [16](also known as a bubble plot). In a bubble chart, each dot stands for a single data point, and the horizontal position, vertical position, and dot size of each dot, in that order, display the corresponding values of the variables that correlate to that point

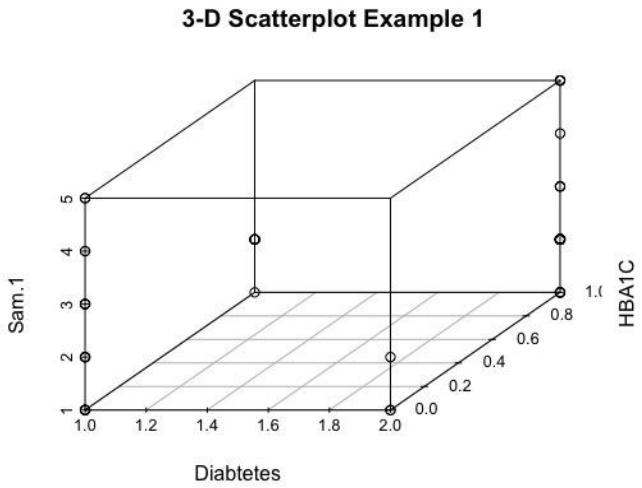


Figure 2 : Bubble chart

V. Random Forest

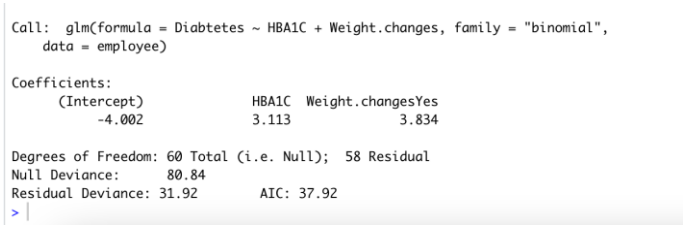
A supervised machine learning algorithm known as a random forest is formed from other machine learning algorithms in the form of decision trees. This method is used in a variety of fields, including as banking and e-commerce, to make predictions on behavior and results.

In the realm of machine learning, a technique known as a random forest may be used to address difficulties with regression and classification. It achieves this by using a technique known as ensemble learning, which is a procedure that pulls together [18] a number of different classifiers in order to solve complex situations. This allows it to more accurately determine which categories a given piece of data belongs to.

A huge number of decision trees are used in the process that is known as a random forest. The "forest" that the random forest algorithm generates may be trained using either the bagging or bootstrap aggregating approach. Both of these techniques are examples of possible applications. Bagging is a meta-algorithm that may be used to increase the accuracy of machine learning algorithms. [22] This meta-algorithm is an ensemble of various machine learning algorithms.

After taking into consideration the forecasts provided by the decision trees, the outcome is established by the use of the approach known as random forest. It does this by computing the average, often known as the mean,

of the output from each of the distinct trees. [24] It's possible that if there were more trees in the forest, the outcome would be more accurate.



```
Call:  glm(formula = Diabtetes ~ HBA1C + Weight.changes, family = "binomial",
    data = employee)

Coefficients:
      (Intercept)             HBA1C  Weight.changesYes
           -4.002             3.113             3.834

Degrees of Freedom: 60 Total (i.e. Null);  58 Residual
Null Deviance:      80.84
Residual Deviance: 31.92        AIC: 37.92
>
```

Figure 3: Evaluate variable importance I

By using a random forest, an algorithm that is based on a decision tree may have its inadequacies addressed and resolved. It enhances accuracy while at the same time reducing the number of datasets that are overfit. It does this without requiring a significant lot of configuration in the packages, and it generates predictions.

Here we have [24] used all the attributes for detection of diabetics. We do have option to select the appropriate attributes before applying the actual algorithm. R

```
$ Diabtetes     : Factor w/ 2 levels "No","Yes": 2 1 1 2 2 2 2 2 1 2 ...
$ HBA1C         : int  1 0 0 1 1 1 0 1 1 1 ...
$ Gender        : Factor w/ 2 levels "Female","Male": 1 2 2 2 1 1 1 2 2 2 ...
$ Family.member : int  1 1 0 1 1 1 1 1 0 1 ...
$ Type          : int  1 0 0 2 2 2 1 1 0 1 ...
$ Sam.1         : Factor w/ 5 levels "Blurred vision",..: 4 4 4 2 3 2 2 2 2 5 ...
$ Sam.2         : Factor w/ 6 levels " Blurred vision",..: 2 2 6 6 4 1 3 6 6 3 ...
$ Insulin       : int  1 0 0 1 1 1 1 1 0 1 ...
$ Issues        : Factor w/ 12 levels "Afraid","Afraid, Confused, Sad",..: 3 6 1 5 11
4 12 8 9 9 ...
$ Health.problem: Factor w/ 6 levels "Heart Disease",..: 2 5 5 1 6 2 5 5 2 4 ...
$ Weight.changes: Factor w/ 2 levels "No","Yes": 2 1 1 2 1 2 2 2 2 2 ...
```

Figure 4 :Check types of variables

```
> importance(rf)
                        No        Yes MeanDecreaseAccuracy MeanDecreaseGini
HBA1C            6.6066466  6.4368060             8.547939       2.90112998
Gender           1.0010015  1.4039987             1.641614       0.02682895
Family.member    2.5404272  2.8938017             3.531109       0.34588562
Type            20.5038502 19.4753309            21.375378      10.51960944
Sam.1            0.8969422  1.6512088             1.972663       0.12077773
Sam.2            0.9759219  2.6255563             2.484865       0.27968376
Insulin         18.1856595 16.8910932            19.250683       8.62818558
Issues           6.2252355  2.5196550             6.058732       1.28565000
Health.problem   3.2824829  0.9438615             3.302986       0.59208677
Weight.changes   6.4637327  6.2726059             8.026460       3.32526989
```

Figure 5 : Evaluate variable importance visualisation

By using the Random forest algorithm we have secured the

90% of accuracy.

VI. Decision Tree

Because of its malleability and flexibility, the predictive modeling technique known as "Decision Tree Analysis" may be used to a wide range of settings and problems according to the aforementioned qualities. In most cases, decision trees are constructed by using an[19] algorithmic approach. This method identifies methods to partition a data set in accordance with certain criteria. A decision-tree partitioning

algorithm is the name given to this specific approach. It is one of the most common methods of supervised learning, and it is also one of the methods that is the most applicable in real-world situations. Decision trees are a kind of non-parametric supervised learning that may be used to classification as well as regression issues. They can also be used to solve problems with missing data. This project aims

to develop a model that is capable of predicting the value of a target variable through the discovery and application of straightforward decision rules that are derived from the characteristics of the data. This will be accomplished by the discovery and application of simple decision rules.

```
> model <- glm(formula = Diabtetes ~ HBA1C + Weight.changes, data = employee ,family =
 "binomial")
> model

Call:  glm(formula = Diabtetes ~ HBA1C + Weight.changes, family = "binomial",
    data = employee)

Coefficients:
      (Intercept)              HBA1C  Weight.changesYes
           -4.002              3.113              3.834

Degrees of Freedom: 60 Total (i.e. Null);  58 Residual
Null Deviance:      80.84
Residual Deviance: 31.92        AIC: 37.92
>
```

Figure 6 : Decision tree

The majority of the [15]time, the guidelines for making judgments are presented in the form of if-then-else statements. Going further down the tree causes the rules to grow more convoluted, but it also makes the model more accurate.

By using the decision tree we have secured the 85% of accuracy which are lower than random forest. We have not used the entire data for the prediction, only limited amount of data have been used.

VII. Logistic Regression :

The method of categorisation known as logistic regression is one of several possible approaches.[23] It produces a prediction about one of two probable outcomes by taking into account a variety of different elements that operate independently of one another.

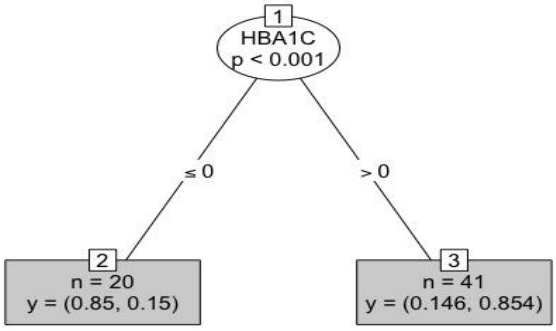


Figure 7 : Training model of logistic regression

```
> summary(model)

Call:
glm(formula = Diabtetes ~ HBA1C + Weight.changes, family = "binomial",
    data = employee)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4481  -0.1904   0.3202   0.3202   1.5703

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)        -4.002      1.172  -3.413 0.000642 ***
HBA1C               3.113      1.004   3.100 0.001938 **
Weight.changesYes   3.834      1.042   3.680 0.000233 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 80.837  on 60  degrees of freedom
Residual deviance: 31.924  on 58  degrees of freedom
AIC: 37.924

Number of Fisher Scoring iterations: 6
```

Figure 8 : Summary model of logistic regression

In such scenario, what does this indicate about the situation? A result is said to be binary if there are only two possible outcomes: either the event takes place (1) or it does not take place (0). Another definition of a binary result is a result in which there are only two possible outcomes (0). The term "independent variable" refers to any component or variable other than the one being studied that has the potential to influence the outcome (or dependent variable).

If one is working with binary data[22], then the statistical technique known as logistic regression is the one that should be used as the method of analysis. If the output or the dependent variable is dichotomous or categorical in nature, then you are dealing with binary data. To put it another way,

if it falls into one of two categories (such as "yes" or "no", "pass" or "fail," and so on), then you are not dealing with numeric data but rather with binary data.

By using the logistic regression we have secured the 87% of accuracy. Here we have used the limited number of parameters.

Conclusion

While working with all of the above algorithms we learned that random forest algorithm is giving us the good accuracy than the other algorithms. But while working with R programming or python programming we have limited scope to improved the algorithm because of lack of libraries, but still secured the good number of accuracy. Random forest gave us 90% accuracy also decision tree gave 85% and logistic regression gave 87% accuracy.

Future work

We found some difficulties while working with the algorithms in r programming i.e for preprocessing or splitting the data for training and testing purpose and lastly we do have limited number of algorithms to work with. So we can use Microsoft azure machine learning services or Microsoft learning studio to evaluate our this accuracy.

References

1. Allam, f & Nossair, Zaki & Gomma, Hesham & Ibrahim, Ibrahim & Abd-el Salam, Mona. (2011). Prediction of subcutaneous glucose concentration for type-1 diabetic patients using a feed forward neural network. Proceedings - ICCES'2011: 2011 International Conference on Computer Engineering and Systems. 129-133. 10.1109/ICCES.2011.6141026
2. Madhusmita Rout and Amandeep Kaur “Prediction of Diabetes Risk based on Machine Learning Techniques” International Conference on Intelligent Engineering and Management (ICIEM) 2020
3. Gaurav Tripathi and Rakesh Kumar “Early Prediction of Diabetes Mellitus Using Machine Learning” 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO) Amity University, Noida, India. June 4-5, 2020
4. Veena Vijayan V. And Anjali C “Prediction and Diagnosis of Diabetes Mellitus -A Machine Learning Approach” IEEE Recent Advances in Intelligent Computational Systems (RAICS) | 10-12 December 2015 | Trivandrum,2015
5. Samrat Kumar Dey , Ashraf Hossain , Md. Mahbubur Rahman “Implementation of a Web Application to Predict Diabetes Disease: An Approach Using Machine Learning Algorithm” 21st International Conference of Computer and Information Technology (ICCIT), 21-23 December, 2018
6. Vinod Maan, Jayati Vijaywargiya and Manya Srivastava "Diabetes Prognostication – An Aptness of Machine Learning” International Conference on Emerging Trends in Communication, Control and Computing (ICONC3) Mody University of Science and Technology, Lakshmangarh, Feb 21-22, 2020
7. Amani Yahyaoui, Akhtar Jamil , Jawad Rasheed and Mirsat Yesiltepe “A Decision Support System for Diabetes Prediction Using Machine Learning and Deep Learning Techniques” 978-1-7281-3992-0/19/$31.00 ©2019 IEEE
8. N. Sneha and Tarun Gangil “Analysis of diabetes mellitus for early prediction using optimal features selection” *(2019) 6:13* https://doi.org/10.1186/s40537-019-0175-6
9. Aishwarya Jakka, Vakula Rani J “Performance Evaluation of Machine Learning Models for Diabetes Prediction” International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-8 Issue-11, September 2019
10. Priyanka Sonar and Prof. K. JayaMalini “DIABETES PREDICTION USING DIFFERENT MACHINE LEARNING APPROACHES” Proceedings of the Third International Conference on Computing Methodologies and Communication (ICCMC 2019)

11. Madhava Prabhu S and Seema Verma “A Systematic Literature Review for Early Detection of Type II Diabetes” 5th International Conference on Advanced Computing & Communication Systems (ICACCS) 2019
12. T. Nagase, H. Sanada, K. Takehara, M. Oe, S. Iizaka, Y. Ohashi, et al., Variations of plantar thermographic patterns in normal controls and non- ulcer diabetic patients: novel classification using angiosome concept, J. Plastic, Reconstr. Aesthetic Surg. 64 (7) (2011) 860–866.
13. L. Vilcahuaman, R. Harba, R. Canals, M. Zaquera, C. Wilches, M. Arista, et al., "Detection of Diabetic Foot Hyperthermia by Infrared Imaging", IEEE EMB Conference , 2014, pp. 4831-4834.
14. Vanessa J Houghton, Virginia M Bower, David C Chant, "Is an increase in skin temperature predictive of neuropathic foot ulceration in people with diabetes? A systematic review and meta- analysis", JOURNAL OF FOOT AND ANKLE RESEARCH, vol. 6, no. 31, pp. 1-13, 2013.
15. VeenaVijayan V, Anjali C. Prediction and diagnosis of diabetes mellitus—a machine learning approach. Recent Adv. 2015. https://doi.org/10.1109/raics.2015.7488400.
16. Kavakiotis I, Tsave O, Salifoglou A, Maglaveras N, Vlahavas I, Chouvarda I. Machine learning and data mining methods in diabetes research. Comput Struct Biotechnol J. 2017;15:104–16.
17. Hina S, Shaikh A, Sattar SA. Analyzing diabetes datasets using data mining. J Basic Appl Sci. 2017;13:466–71.
18. Kevin P, Razvan B, Cindy M, Jay S, Frank S. A machine learning approach to predicting blood glucose levels for diabetes management. In: Modern artificial intelligence for health analytics. Papers from the AAAI-14. 2014.
19. Jegan Chitra. Classification of diabetes disease using support vector machine. Int J Eng Res Appl. 2013;3:1797–801.
20. Polat K, Güneş S, Arslan A. A cascade learning system for classification of diabetes disease: generalized discriminant analysis and least square support vector machine. Expert Syst Appl. 2008;34(1):482–7.
21. P. Goyal and S. Jain, "Prediction of Type-2 Diabetes using Classification and Ensemble Method Approach," 2022 International Mobile and Embedded Technology Conference (MECON), 2022, pp. 658-665, doi: 10.1109/MECON53876.2022.9752268.
22. M. M, G. B, D. K and N. S, "Diabetic Patient Prediction using Machine Learning Algorithm," 2021 Smart Technologies, Communication and Robotics (STCR), 2021, pp. 1-5, doi: 10.1109/STCR51658.2021.9588925.
23. B. S. MURTHY and J. SRILATHA, "Comparative Analysis on Diabetes Dataset Using Machine Learning Algorithms," 2021 6th International Conference on Communication and Electronics Systems (ICCES), 2021, pp. 1416-1422, doi: 10.1109/ICCES51350.2021.9488954.
24. Nurjahan, M. A. T. Rony, M. S. Satu and M. Whaiduzzaman, "Mining Significant Features of Diabetes through Employing Various Classification Methods," 2021 International Conference on Information and Communication Technology for Sustainable Development (ICICT4SD), 2021, pp. 240-244, doi: 10.1109/ICICT4SD50815.2021.9397006.
25. slam, M.M.F., Ferdousi, R., Rahman, S., Bushra, H.Y. (2020). Likelihood Prediction of Diabetes at Early Stage Using Data Mining Techniques. In: Gupta, M., Konar, D., Bhattacharyya, S., Biswas, S. (eds) Computer Vision and Machine Intelligence in Medical Image Analysis. Advances in Intelligent Systems and Computing, vol 992. Springer, Singapore.
26. Aditya Saxena, Megha Jain, Prashant Shrivastava. “Data Mining Techniques Based Diabetes Prediction” Indian Journal of Artificial Intelligence and Neural Networking (IJAINN) ISSN: 2582-7626 (Online), Volume-1 Issue-2, April 2021