International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 06 Issue: 07 | July - 2022 Impact Factor: 7.185 ISSN: 2582-3930

END TO END DIABETES PREDICTION APPLICATION USING MACHINE LEARNING

M. Kanagalakshmi and N.Gowri M.C.A., B.Ed.,
Final year PG student and assistant professor,
Department of Information Technology
G. Venkataswamy Naidu college, kovilpatti, Tamil Nadu, India

ABSTRACT

Diabetes Mellitus is among critical diseases and lots of people are suffering from this disease. Age, obesity, lack of exercise, hereditary diabetes, living style, bad diet, high blood pressure, etc. can cause Diabetes Mellitus. People having diabetes have high risk of diseases like diabetes symptom, kidney disease, stroke, eye problem, nerve damage, etc. Current practice in hospital is to collect required information for diabetes diagnosis through various tests and appropriate treatment is provided based on diagnosis. Big Data Analytics plays an significant role in healthcare industries. Healthcare industries have large volume databases. Using big data analytics one can study huge datasets and find hidden information, hidden patterns to discover knowledge from the data and predict outcomes accordingly. In existing method, the classification and prediction accuracy is not so high. In this project, I have proposed a diabetes prediction model for better classification of diabetes which includes few external factors responsible for diabetes along with regular factors like Glucose, BMI, Age, Insulin, etc. Classification accuracy is boosted with new dataset compared to existing dataset. Further with imposed a pipeline model for diabetes prediction intended towards improving the accuracy of classification.

I. INTRODUCTION

Diabetes is a common chronic disease which can pose great threat to human health. Diabetes can be identified when blood glucose is higher than normal level, which is caused by high secretion of insulin or biological effects. Diabetes can cause various damage to our body and can function tissues, kidneys, eyes and blood vessels. Diabetes can be divided into two categories, type 1 diabetes and type 2 diabetes. Patients with type 1 diabetes are normally younger with an age less then 30 years old. The clinical symptoms are increase thirst and frequent urination this type of diabetes cannot be cleared by medications as it requires therapy. Type 2 diabetes occurs more commonly on middle-aged and old people, which can show hypertension, obesity and other diseases. With our living standards diabetes has increased commonly in people's daily life. So how to analyze diabetes is worth studying. As we get the diagnosis earlier we can control it. Machine learning can make preliminary judgment on diabetes mellitus according to physical examination data, and by reference with doctors. Recently, many algorithms are used to predict diabetes, including machine learning methods like Random Forest, (KNN) K-Nearest Neighbor, Decision Tree and so on. With this machine learning techniques we are able to predict diabetes by constructing predicting models which are obtained by medical datasets. By extracting such knowledge we are able to predict

diabetic patient. We use the best technique to predict based on our attributes of the given datasets in order to get the perfect accuracy to predict diabetes mellitus

II. LITERATURE REVIEW

Defusal Faruque and Asaduzzaman, Iqbal H.Sarker has discussed that diabetes is one of the most common disorder of the human body it is caused due the metabolic disorder .Hence that they used various and important ML algorithms that are Support Vector machine, NB,KNN and DT to predict the diabetes[1]. Sidong Wei,Xuejiao Zhao and Chunyan Miao presented that diabetes is commonly called as disorder in which glucose level in body is high. In this paper they use popular methods such as SVM and deep neural network for identify the disease and data processing. [2]. END TO END DIABETES PREDICTION APPLICATION USING MACHINE LEARNING P a g e 3 | 59 Lakshmi K.S and G.Santhosh Kumar according to them Hospital databases serve as wealthy information source for the fruitful medication diagnosis. IN this they used NLP tools along with combined with data mining algorithms for the extraction of rules [3]. Jian-xunChen , Shih-LiSu and Che-Ha Chang discussed about Ontology that generate a primary care planning to the medical professional's for the accustoming. The result of the research paper shows the model can be provided personalize diabetes mellitus care planning efficiently [4]. MM Alotaib, RSH.Istepanian, and A.Sungoor they are present a clever based mobile polygenic disease control system & tutoring model for the patients with diabetes. In this, system is able to store the clinical information about the diabetes system, such an often blood sugar level and BP measured and hypo glycaemia event [5].

III. METHODOLOGY

1.PROPOSED METHODOLOGY

Goal of the paper is to investigate for model to predict dia- betes with better accuracy. We experimented with different classification and ensemble algorithms to predict diabetes. In the following, we briefly discuss the phase.

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

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

2.Data Preprocessing- Data preprocessing is most im- portant process. Mostly healthcare related data contains missing vale and other impurities that can cause effective- ness of data. To improve quality and effectiveness obtained after mining process, Data preprocessing is done. To use Machine Learning Techniques on the dataset effectively ths process is essential for accurate result and successful prediction. For Pima Indian diabetes dataset we need to perform pre processing in two steps.

1.Missing Values removal- Remove all the instances that have zero (0) as worth. Having zero as worth is not possi- ble. Therefore this instance is eliminated. Through elimi- nating irrelevant features/instances we make feature subset and this process is called features subset selection, which reduces diamentonality of data and help to work faster.

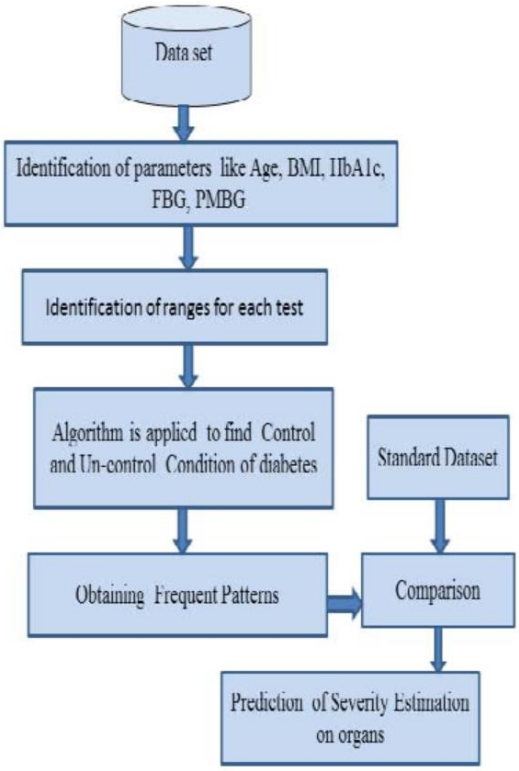
2.Splitting of data- After cleaning the data, data is nor- malized in training and testing the model. When data is spitted then we train algorithm on the training data set and keep test data set aside. This training process will produce the training model based on logic and algorithms and val- ues of the feature in training data. Basically aim of normal- ization is to bring all the attributes under same scale.

3.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 accura- cy of them, and also been able to figure out the responsi- ble/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 algo- rithm. Svm is most popular classification technique. Svm creates a hyperplane that separate two classes. It can create a hyperplane or set of hyperplane in high dimensional 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 sup- ported 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 bet- ter.
- To find the better hyper plane you have to calcu- late 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 Neighbors – KNN is also a supervised ma- chine learning algorithm. KNN helps to solve both the classification and regression problems. KNN is lazy predic-tion 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 train- ing data set its nearest neighbors. Here K= Number of nearby neighbors, its always a positive integer. Neighbors value is chosen from set of class. Closeness is

mainly de-fined in terms of Euclidean distance. The Euclidean dis- tance between two points P and Q i.e. P (p1,p2, . Pn) and Q (q1, q2,..qn) is defined by the following equation:-

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.
- Find out the same output values.
- If the values are same, then the patient is diabetic, other- wise not.

3.Logistic Regression- Logistic regression is also a su- pervised learning classification algorithm. It is used to es- timate the probability of a binary response based on one or more predictors. They can be continuous or discrete. Lo- gistic regression used when we want to classify or distin- guish some data items into categories.

It classify the data in binary form means only in 0 and 1 which refer case to classify patient that is positive or nega- tive 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 neg- ative class.

Sigmoid function P = 1/1+e – (a+bx) Here P = probability, a and b = parameter of Model.

Ensembling- Ensembling is a machine learning technique Ensemble means using multiple learning algorithms to- gether for some task. It provides better prediction than any other individual model thats why it is used. The main cause of error is noise bias and variance, ensemble methods help to reduce or minimize these errors. There are two popular ensemble methods such as Bagging, Boosting, ada-boosting, Gradient boosting, voting, averaging etc. Here In these work we have used Bagging (Random forest) and Gradient boosting ensemble methods for predicting diabetes.

5.Random Forest It is type of ensemble learning meth- od 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. Ran- dom Forest is developed by Leo Bremen. It is popular en- semble Learning Method. Random Forest Improve Perfor- mance 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 indi- vidual 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 until l number of nodes has been reached.
- Built forest by repeating steps a to d for a num- ber of times to create n number of trees.

The random forest finds the best split using the Gin- Index Cost Function which is given by:

The first step is to need the take a glance at choices and use the foundations of each indiscriminately created decision tree to predict the result and stores the anticipated outcome at intervals the target place. Secondly, calculate the votes for each predicted target and ultimately, admit the high voted predicted target as a result of the ultimate prediction from the random forest formula. Some of the options of Random Forest does correct predictions result for a spread of applications are offered.

4.MODEL BUILDING

This is most important phase which includes model build- ing for prediction of diabetes. In this we have implemented various machine learning algorithms which are discussed above for diabetes prediction.

Procedure of Proposed Methodology-

Step1: Import required libraries, Import diabetes dataset.

Step2: Pre-process data to remove missing data.

Step3: Perform percentage split of 80% to divide dataset as Training set and 20% to Test set.

Step4: Select the machine learning algorithm i.e. K-Nearest Neighbor, Support Vector Machine, Decision Tree, Logistic regression, Random Forest and Gradient boosting algorithm.

Step5: Build the classifier model for the mentioned ma- chine learning algorithm based on training set.

Step6: Test the Classifier model for the mentioned ma- chine learning algorithm based on test set.

Step7: Perform Comparison Evaluation of the experi- mental performance results obtained for each classifier.

Step8: After analyzing based on various measures con- clude the best performing algorithm.

5.EXPERIMENTAL RESULTS

In this work different steps were taken. The proposed ap- proach uses different classification and ensemble methods and implemented using python. These methods are stand- ard Machine Learning methods used to obtain the best ac- curacy from data. In this work we see that random forest classifier achieves better compared to others. Overall we have used best Machine Learning techniques for prediction and to achieve high performance accuracy. Figure shows the result of these Machine Learning methods.

Here feature played important role in prediction is present- ed 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.

IV. CONCLUSION

One of the significant impediments with the progression of technology and medicine is the early detection of a disease, which is in this case, diabetes. However, in this study, systematic efforts were made into designing a model which is accurate enough in determining the onset of the disease. With the experiments conducted on the Pima Indians Diabetes Database, we have readily predicted this disease. Moreover, the results achieved proved the adequacy of the system, with an accuracy of 76% using the K-Nearest Neighbours classifiers. With this being said, it is

hopeful that we can implement this model into a system to predict other deadly diseases as well. There can be room for further improvement for the automation of the analysis of diabetes or any other disease in the future. In future, we will try to create a diabetes dataset in collaboration with a hospital or a medical institute and will try to achieve better results. We will be incorporating more Machine Learning and Deep learning models for achieving better results as well.

V.REFERENCES

[1] "Automate The Boring Stuff With Python, 2Nd Edition: Practical Programming For Total Beginner", SL Swegart, 2020.

[2]"Learning Python 5ed: Powerful Object-Oriented Programming" O'Relly – 2014.

[3]"Elements of Programming Interviews in Python: The Insiders' Guide", Adan Aziz, Tsung-Hsen Lee – 2021.

[4] Debadri Dutta, Debpriyo Paul, Parthajeet Ghosh, "Analyzing Feature Importances for Diabetes Prediction using Machine Learning". IEEE, pp 942-928, 2018.

[5]K.VijiyaKumar, B.Lavanya, I.Nirmala, S.Sofia Caroline, "Random Forest Algorithm for the Prediction of Diabetes ".Proceeding of International Conference on Systems Compu- tation Automation and Networking, 2019.

[6]Md. Faisal Faruque, Asaduzzaman, Iqbal H. Sarker, "Perfor- mance Analysis of Machine Learning Techniques to Predict Diabetes Mellitus". International Conference on Electrical, Computer and Communication Engineering (ECCE), 7-9 Feb- ruary, 2019.

[7]Tejas N. Joshi, Prof. Pramila M. Chawan, "Diabetes Prediction Using Machine Learning Techniques".Int. Journal of Engineer- ing Research and Application, Vol. 8, Issue 1, (Part -II) Janu- ary 2018, pp.-09-13

[8]Nonso Nnamoko, Abir Hussain, David England, "Predicting Diabetes Onset: an Ensemble Supervised Learning Approach ". IEEE Congress on Evolutionary Computation (CEC), 2018.

[9]Deeraj Shetty, Kishor Rit, Sohail Shaikh, Nikita Patil, "Diabe- tes Disease Prediction Using Data Mining ".International Con- ference on Innovations in Information, Embedded and Com- munication Systems (ICIIECS), 2017.

[10]Nahla B., Andrew et al,"Intelligible support vector machines for diagnosis of diabetes mellitus. Information Technology in Biomedicine", IEEE Transactions. 14, (July. 2010), 1114-20.

[11]A.K., Dewangan, and P., Agrawal, Classification of Diabetes Mellitus Using Machine Learning Techniques, International Journal of Engineering and Applied Sciences, vol. 2, 2015.