RESEARCH ARTICLE OPEN ACCESS

Comparative Analysis Of Diabetes Prediction Using Logistic Regression and KNN

Prof.Preeti.B[1],Aishwarya S Bingeri[2] , Akshata Sajjan[3] ,Muskan Khazi[4] , Nisha M Patil[5]

[1]Department of Electronics and Communication Engineering,SDMCET,Dharwad,India,[2] Department of Electronics and Communication Engineering,SDMCET,Dharwad,India,[3]Department of Electronics and Communication Engineering,SDMCET,Dharwad,India,[4]Department of Electronics and Communication Engineering,SDMCET,Dharwad,India,[5]Department of Electronics and Communication Engineering, SDMCET,Dharwad,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 heart disease, 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 a 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 paper, we 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. we are using the KNN and Logistic Regression algorithms to predict the level of diabetes with future risk and compare the results obtained.

Key Words: Diabetes Mellitus, Machine learning , KNN, Logistic Regression, Confusion Matrix, Accuracy Score.

I. INTRODUCTION

Diabetes is deadly diseases in the world. Diabetes caused because of obesity or high blood glucose level, and so forth. It affects the hormone insulin, resulting in abnormal metabolism of crabs and improves level of sugar in the blood. Diabetes occurs when body does not make enough insulin. According to (WHO) World Health Organization about 422 million people suffering from diabetes particularly from low or idle income countries. And this could be increased to 490 billion up to the year of 2030. However prevalence of diabetes is found among various Countries like Canada, China, and India etc. Population of India is now more than 100 million so the actual number of diabetics in India is 40 million. Diabetes is major cause of death in the world. Early prediction of disease like diabetes can be controlled and save the human life. To accomplish this, this work explores prediction of diabetes by taking various attributes related to diabetes disease. For this purpose we use the dataset collected from different hospitals, we apply various Machine Learning classification and ensemble techniques to predict diabetes. Machine Learning is a method that is used to train computers or machines explicitly. Machine Learning Techniques such as Logistic Regression and KNN provide appropriate result to collect Knowledge by building classification models from collected dataset. Such collected data can be useful to predict diabetes..

II. LITERATURE SURVEY

A massive data is generated from scientific institutes each 12 months. Many people have proposed one-of-a-kind structures for the prediction of diabetics. Orbietal is one among them who proposed a machine for the prediction of diabetics.

Many data sets are to be had for different diseases. Mining of those datasets offers beneficial facts. The principle goal of this device is to predict diabetes primarily based on the candidate struggling at unique age, with higher accuracy the use of Logistic Regression. Usually choice timbers are supervised studying strategies used for each category and regression. KNN is also used for the classification and regression. Genetic programming (GP) is utilized by Pradhan for testing and education of the database for the prediction of the diabetes. However the output effects received within the Genetic programming approach has less accuracy whilst as compared to other techniques and additionally designed a prediction model for the diabetes disease with two models specifically ANN(synthetic Neural Networks) and the second one is FBS(Fasting Blood Sugar. The algorithms at the danger of diabetes mellitus become proposed by Nongyao. He proposed 4 famend machine gaining knowledge of classifications techniques specifically decision Tree, artificial Neural Networks, Logistic Regression and Naive Bayes. Bagging and Boosting strategies are used for growing the robustness of the designed version.

Sonu Kumari and Archana Singh proposed an intelligent and effective methodology for the automated detection of Diabetes Mellitus using Neural Network. The paper approached the aim of diagnoses by using ANNs and demonstrated the need for pre-processing and replacing missing values in the dataset being considered. Through the Modified training set, a better accuracy was achieved with lesser time required for training the set. Sajida by using CPCSSN(Canadian primary care sentinel surveillance Network) dataset and three machine learning methods to predict the diabetes Disses (DD) in early stage to safe human life at from early death . Sadri used Naive Bayes, RBF Network and J48 data mining algorithms for diagnosing type II diabetes. They used WEKA tool. Finally they found Naive Bayes, having the accuracy rate of 76.96% than other algorithms. In this paper, Prediction of diabetes is done using ensemble voting classifiers for pima Indian diabetes dataset, in comparison with different classification algorithms, the highest accuracy of 80% and 81% is achieved for data set by using 10-fold cross validation and by spitting data into 20% testing and 80% training. J. Pradeep Kandhasamy, S. Balamurali [58] this research study compare the performance of algorithms those are used to predict diabetes using data mining techniques. Also authors classifiers J48 Decision Tree, K-Nearest Neighbors, and Random Forest, Support Vector Machines to classify patients with diabetes mellitus. Authors compared four prediction models for predicting diabetes mellitus using 8 important attributes under

two different situations. One is before pre-processing the dataset. After pre-processing, the dataset given more accurate result when compared to the previous studies. In our case, both KNN (k=1) and logistic Regression performance much better classifiers and they provide nearly 85% accuracy. From this we can come to know that after removing the noisy data from our dataset it will provide good result for our problem.

III. METHODOLOGY

This section is comprised of the following steps: the data description, pre-processing technique and the classification algorithm. The proposed model is designed and implemented by comparing the benefit of applying KNN and Logistic regression. A new methodology is then proposed by using data pre-processing technique to transform the initial set of features, thereby solving the problem of correlation, which makes it difficult for the classification algorithm to find relationships among the data. The data pre-processing technique helps to filter out irrelevant features and also increases model performance. After performing data pre-processing, the result is then passed for supervised classification using Logistic regression and KNN algorithms.

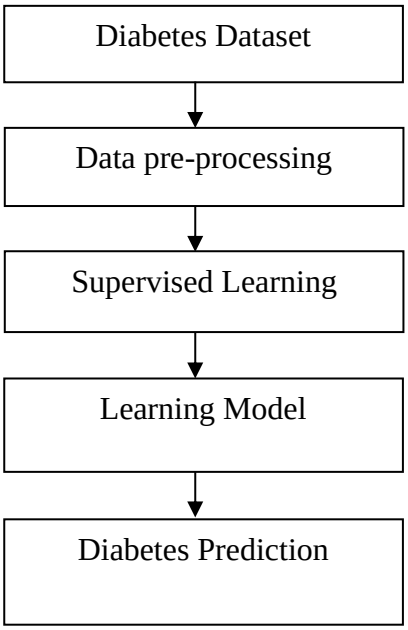

Figure 1: Flowchart

Figure 1 shows flowchart for diabetes prediction model. This model has 5 different modules. These modules include

1. Dataset collection
2. Data Pre-processing
3. Supervised Learning Techniques
4. Build model
5. Diabetes prediction

1. Dataset Collection

The dataset contains 8 attributes which are:

- Number of pregnancy
- Plasma glucose concentration after an oral glucose tolerance test
- Diastolic blood pressure(mm HG)
- Skin thickness(mm)
- Insulin(mu U/ml)
- Body mass index
- Diabetes pedigree function and age.

2. Data Pre-processing

- Data pre-processing is a process of preparing the raw data and making it suitable for a machine learning model.
- It is the first and crucial step while creating a machine learning model.
- When creating a machine learning project, it is not always a case that we come across the clean and formatted data.
- Thus while doing any operation with data, it is mandatory to clean it and put in a formatted way.

3. Supervised Learning Technique:

A. 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 classify 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.

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

two points in Euclidean n-space

p, q =

Euclidean vectors, starting from

qi, pi = the origin of the space (initial point)

n = n-space

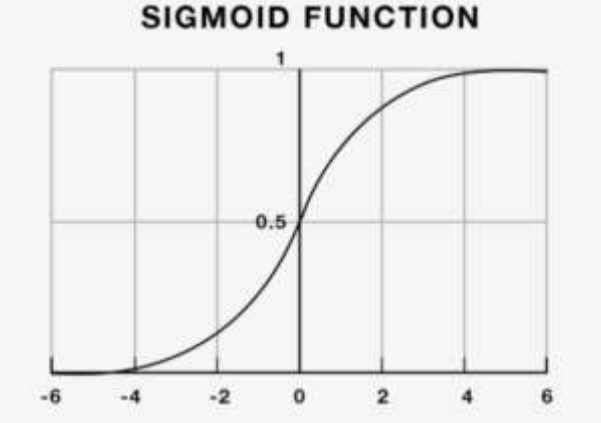

Figure 2:Sigmoid Function

B. KNN

K-Nearest Neighbor – KNN is also a supervised ma- chine 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 its nearest neighbors.

Here K= Number of nearby neighbors, it's always a positive integer. Neighbors 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:-

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

$$d(\mathbf{p},\mathbf{q}) = \sqrt{\sum_{i=1}^{n}(q_i - p_i)^2}$$

- 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.

4. Building Model

This is most important phase which includes model building for prediction of Diabetes. In this we have implemented Logistic Regression and KNN algorithms.

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, Logistic regression.
Step5: Build the classifier model for the mentioned machine learning algorithm based on training set.

Step6: Test the Classifier model for the mentioned machine learning algorithm based on test set.
Step7: Perform Comparison Evaluation of the experimental performance results obtained for each classifier.
Step8: After analyzing based on various measures conclude the best performing algorithm.

5.Diabetes Prediction

Figure 3: Data Pre-Processing

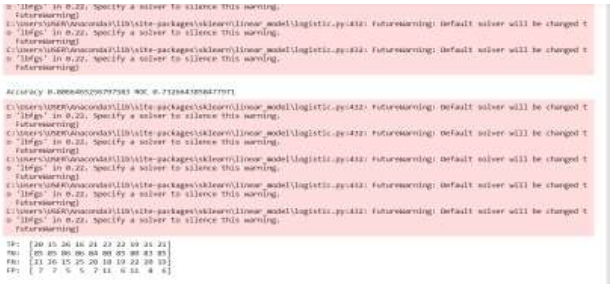

Figure 4: Logistic Regression

Figure 5: K-Nearest Neighbor

To select the best performance, we considered the AUC, sensitivity, specificity, plotted the ROC curve, and showed the average curve and range. The mean AUC for the KNN algorithm was 88% and Logistic Regression 80%.

IV. CONCLUSION

Before the comparison among two techniques, we thought machine learning technique would beat traditional statistical technique that should have been correct since the dataset to be analyzed was big and complex. K-Nearest Neighbor technique is still the best after data cleansing in our case. As a general rule of thumb, we recommend to apply Logistic Models at the beginning. Then a nice probabilistic interpretation is obtained. Honestly, choosing a model is always hard. If we would like to predict the response in a very high accurate, different classifiers are supposed to be applied. In the fact, data is the more important than model. This is one of the reasons why we achieve the best result performing K-Nearest Neighbors Model is that the raw dataset has been transformed enough.

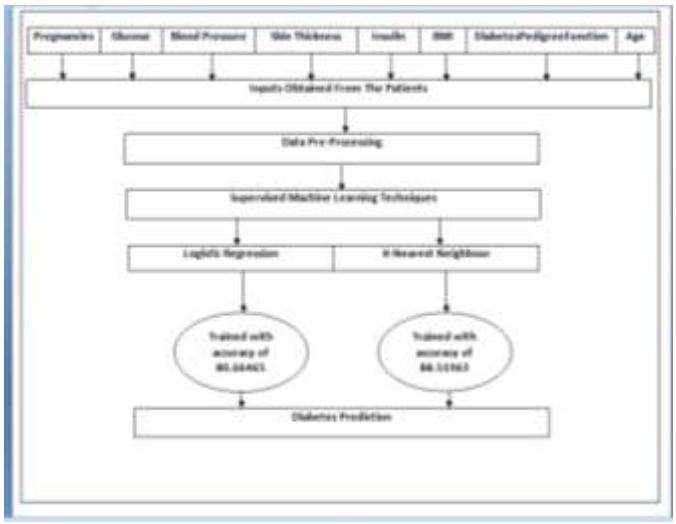

Figure 6: Architecture Design

REFERENCES

[1] Global Report on Diabetes 2016 by World Health Organisation. http://www.who.int/diabetes/publications/grd-2016/en/, ISBN 978 92 4 156525 7.

[2] Scott MG, Ivor JB, Gregory LB, Alan C, Robert HE, Barbara VH, William M, Sidney CS, James RS. Diabetes and cardiovascular disease a statement for healthcare professionals from the American Heart Association Circulation. 1999;100(10):1134–46.

[3] Komi, Zhai. 2017. Application of Data Mining Methods in Diabetes Prediction

[4] Dr Saravana kumar N M, Eswari T, Sampath P and Lavanya S," Predictive Methodology for Diabetic Data Analysis in Big Data", 2nd International Symposium on Big Data and Cloud Computing,2015. 22

[5] Aiswarya Iyer, 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.

[6] P. Suresh Kumar and S. Pranavi "Performance Analysis of Machine Learning Algorithms on Diabetes Dataset using Big Data Analytics", International Conference on Infocom Technologies and Unmanned Systems, 978-1-5386-0514-1, Dec. 18-20, 2017.

[7] 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.