ISSN:2231-0711
Thota Aditya Kumar| IJCSET(www.ijcset.net) | December 2016 | Vol 6, Issue 12, 419-422

Estimation of Diabetes Employing Classifiers and Machine Learning Methods

Thota Aditya Kumar

Student, Department of C.S.E, Geethanjali College of Engineering & Technology, Cheeryala(V) Keesara(M), Medchal Dist. Hyderabad, Telangana, INDIA

Abstract: Diabetes is considered to be as never-ending infection and deadliest disease which results in increase of glucose levels in the body. If it is not treated properly then it may cause many difficulties. The general approach of treating this disease is visiting of a patient to a specialist, but the result may not be accurate. But this problem will be solved by machine learning approaches. The objective of this study is to design a system that can predict the diabetes in the human body with maximum precision. Therefore, Machine learning classification algorithms namely SVM, KNN, Naïve Bayes and Logistic Regression are used in this experiment to detect diabetes at the starting stage. These experimentations are performed on Pima Indians Diabetes Database (PDID) which is drawn from UCI repository. The outcomes of the four algorithms are evaluated on diverse parameters such as Precision, Accuracy F-Measure, and Recall.

Key words: SVM, KNN, Naïve Bayes and Logistic Regression.

1. Introduction

Diabetes is most significant medical ailment throughout the world. Diabetes is grouped into four general types: type 1 and type 2 gestational diabetes and further explicit sorts. This normally create after numerous years (10-20). High blood glucose levels can cause genuine ailments influencing the heart and veins, eyes, kidneys, nerves and teeth. The key aim of this implementation is to examine the performance of numerous classification models like SVM, KNN, Naïve Bayes and Logistic Regression. That can give the probability of diseases in patients with the highest accuracy and exactness. General approach of several lab tests could obscure the crucial identification process and which fallout in delay in the judgment especially for a state where number of lab tests are mandatory. The current work is carried out on the basis of secondary data set drawn from the UCI and PIDD repository and machine learning mechanisms employed to identify the presence of ailment in patient at an early phase. The results emerging out of experimentation computes the accuracy, misclassification rate, precision, recall ad F1-score.

2.Proposed System

This system uses machine learning methods for predicting diabetes. Now a day's diabetes is known as one of the enduring diseases. It causes a growth in the concentration of sugar levels in the blood. Here in this system, we design a model which predicts the disease which gives maximum accuracy. The machine learning classification algorithms KNN, SVM and Naïve bayes and Logistic Regression are used for predicting the results [3][4].

Many factors such as age, insulin pregnancy, sugar levels are considered and in the data set which is obtained from the UCI Repository. The data obtained from the data set is used for predicting results. The proposed method mainly focusses on selecting of all attributes that ail I early detection of diabetes. The model selects the optimal features from data set in order to improve the accuracy. The attribute choice is most important step it decreases the intricacy of work. This phase always selects the optimum attributes from the data and it bypasses the data into machine learning program. The efficiency of all the algorithms are calculated based on some metrics like precision, accuracy, F-score of these classifiers are presented in the paper as shown in figure 2.

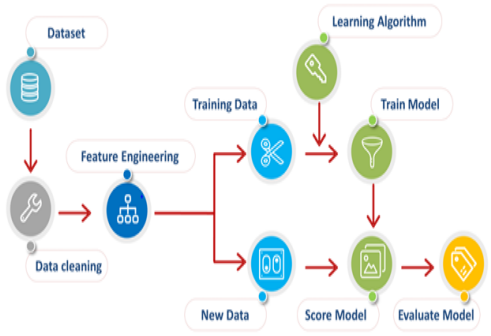

Fig. 2. Architecture to the system

3. Literature Survey

AishwaryaR, Gayathri P, JaisankarN., In 2013 had built a model that was developed using SVM andPCA (a preprocessing) method that helped in diagnosing of diabetes it showed accuracy around 95% [1].

In 2013, Aljumah, A.A., Ahamad, M.G., Siddiqui, carried out a research that focuses on employing a regression-based data mining technique to perform predictive analysis of diabetes treatment.[2]

In this research, Esposito, F., Malerba, D., Semeraro, G., Kay, they used a top-down strategy to solve the challenge of retrospectively pruning decision trees induced by data. In the fields of pattern recognition and machine learning, this subject has gotten a lot of attention, and many different methods have been offered in the literature [3].

Kumari, V.A., Chitra, R, in 2013, used datasets for diabetic disease from the University of California, Irvine's machine learning lab. SVM is used to train all of the patient data. For a given amount of data, selecting the appropriate parameter values for a specific kernel is crucial. With simple clinical data and no laboratory testing, the SVM technique can successfully detect a common disease[4].

4. Implementation

4.1. Collecting and preprocessing the patient's data set

Patients' data is collected. It must include attributes like sugar level, insulin, age, sex, pregnancy. The data is taken from UCI repository and divide it into training data and testing data set. 80% of the data is considered as training data and 20% of data is testing data. Data preprocessing is the step in which data is transformed into a state in which machine can easily parse. If the data contains too many formats, missing values, duplicate values than this cannot be used for predicting because it may not produce better results. So the data must be preprocessed [5].

4.2. Application of classification algorithms

Now after preprocessing the data set machine learning classification algorithms are applied. In this system SVM, K-nearest Neighbor and Naïve Bayes and Logistic Regression algorithms are used.

4.3. Performance evaluation based on measures

Next step is to evaluate the performance after applying the all algorithms. Corresponding classifiers performance is measured over accuracy values in %.

A. Algorithm Logistic Regression

Step1: import libraries and load the dataset
Step2: Visualize the data using histogram
Step3: Identify Outliers and remove them
Step4: Split the Training data into 80 20 ratio and Build the Model using Logistic Regression
Step5: Find Accuracy and ROC_AOC Curve

B. Algorithm Support Vector Machine

Step1: Load SVC library
Step2: Build Model using SVM
Step3: Find Accuracy and ROC_AOC Curve
Step4: Find the Crass value score

C. Algorithm KNN

Step1: Load the KNeighborsClassifier library
Step2: Build the Model using KNN
Step3: Find Accuracy and ROC_AOC Curve
Step4: Find the Crass value score

D. Algorithm Naïve Bayes

Step1: Load GaussianNB library
Step2: Find accuracy and ROC_AOC Crve
Step3: Find the Crass value score

Next Plot the Bar graph for the Logistic Regression, SVM, KNN and Naïve Bayes algorithm result. For that give the Label as Accuracy Score and Algorithms in X and Y axes[6].

Sample Data Set:

Collected the dataset with following attributes from the UCI repository and Machine learning classification algorithms are used on this to predict the presence of diabetes with maximum accuracy shown in table I.

Table I : Sample Data Set Used

Pregnancies	Glucose	Blood Pressure	Skin Thickness	Insulin	BMI	Diabetes pedigree function	Age	Outcome
2	138	62	35	0	3.3.6	0.127	47	1
0	84	62	31	125	38.2	0.233	23	0
0	145	82	0	0	44.2	0.63	31	1
0	135	0	42	250	42.3	0.365	24	1
1	139	62	41	480	40.7	0.536	21	0
4	173	78	32	265	46.5	1.159	58	0
8	99	72	17	0	25.6	0.294	28	0
2	194	80	0	0	26.1	0.551	67	0
2	83	65	28	66	36.8	0.629	24	0
4	89	90	30	0	33.5	0.292	42	0
4	99	68	38	0	32	0.145	33	0
3	125	70	18	122	28.9	1.144	45	1

5. Results and Discussion

1.Checking if any null values are present in the dataset after loading the dataset and then visuals of dataset is presented in histograms, Figure 5.1.

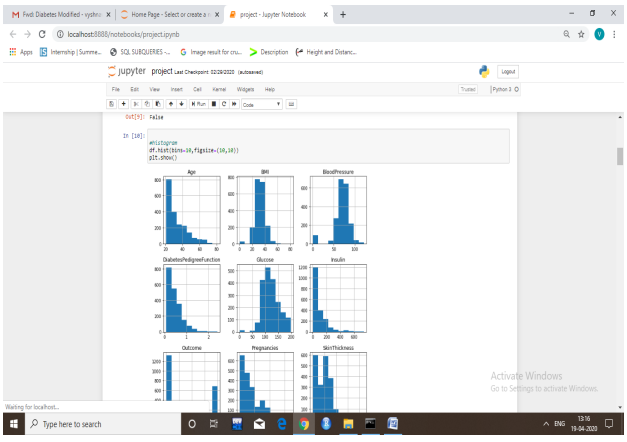

Fig. 5.1Visualizing the dataset using histograms

2. Next find out the correlation using heat map and count total outcome in each target as 0 and . That is 0 means no diabetes and 1 means with diabetes, Figyre 5.2.

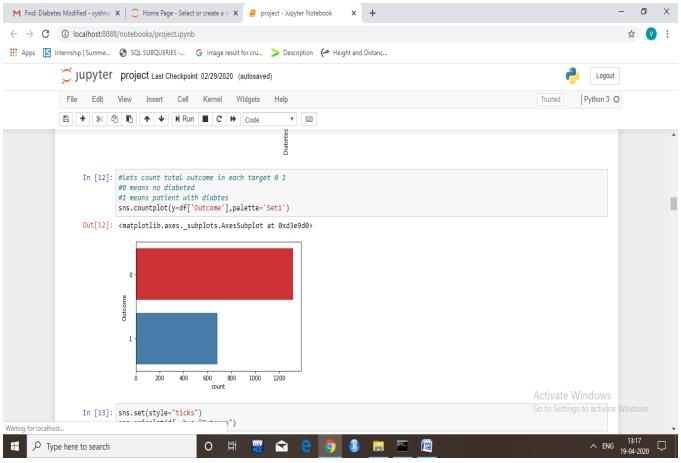

Fig. 5.2 Visualizing the output feature

3. Draw the accuracy curve for four algorithms. So, that we can find which algorithm is giving most accurate values for the diabetes. Here KNN is highest figure 5.3.

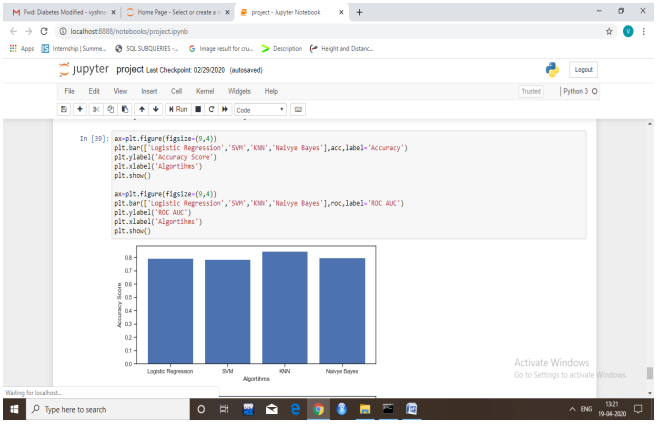

Fig. 5.3 Accuracy curve for different algorithms

6. Conclusion

Diabetes happens to be the dreadest real-world medical issue and its detection at an early stage is the key concerns of this study. Efforts are made to design a system that results in the prediction of diabetes disease. In this work , it uses KNN, SVM, NAÏVE BAYES and LOGISTIC REGRESSION classification algorithms. With the help of these algorithms the dataset is studies and evaluated on various metrics. Experiments are implemented on

PIDD which is take from UCI repository. It has been observed that KNN classifier yields the highest classification accuracy which is of 91% when compared with other classification algorithms. This system can aid in making more operational and reliable disease prediction process which would contribute towards developing the best healthcare structure by reducing cost, time which includes reduction in morality rate also .

References:

[1]. Aishwarya, R., Gayathri, P., Jaisankar, N., 2013. A Method for Classification Using Machine Learning Technique for Diabetes. International Journal of Engineering and Technology (IJET) 5, 2903–2908.

[2]. Aljumah, A.A., Ahamad, M.G., Siddiqui, M.K., 2013. Application of data mining: Diabetes health care in young and old patients. Journal of King Saud University - Computer and Information Sciences 25, 127–136. doi:10.1016/j.jksuci.2012.10.003.

[3]. 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.

[4]. Kumari, V.A., Chitra, R., 2013. Classification Of Diabetes Disease Using Support Vector Machine. International Journal of EngineeringResearch and Applications (IJERA) www.ijera.com 3, 1797–1801.

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

[6]. American Diabetes Association (2012). Diagnosis and classification of diabetes mellitus. Diabetes Care 35(Suppl. 1), S64–S71. doi: 10.2337/dc12-s064.