International Journal of Research in Engineering, Science and Management
Volume-3, Issue-2, February-2020
www.ijresm.com | ISSN (Online): 2581-5792

Diabetes Prediction Using SVM and Logistic Regression

J. Krishnendhu[1], G. Arnesh[2], B. Harish[3], K. Vidhya[4]
[1,2,3]*Student, Department of Computer science and Engineering, Dr. Mahalingam College of Engineering and Technology, Pollachi, India*
[5]*Assistant Professor, Department of Computer science and Engineering, Dr. Mahalingam College of Engineering and Technology, Pollachi, India*

***Abstract*: Diabetes has become a common metabolic disorder among people now a days. Based on various conditions patients experience there are types of diabetes as Type I, II, gestational and prediabetes. Dataset containing nine attributes including an Outcome field is obtained from an online dataset repository with 2000 instances. In this system, Support Vector Machine and Logistic Regression algorithms are applied to predict diabetes in patients using BMI and Outcome. Accuracy of algorithms are calculated and compared using performance measures.**

***Keywords*: Diabetes, FPR, SVM, TPR, Regression, Accuracy.**

1. Introduction

Diabetes is a syndrome characterized by metabolic disorder and abnormal rise in the concentration of blood sugar caused by insulin deficiency, or low insulin sensitivity of tissues, or both. The hormone insulin carries the sugar or glucose from the blood and stored in the cells for energy. The situation in which pancreas does not produce sufficient insulin hormone that is needed to absorb glucose from the blood leading to increase in blood sugar level is termed as Type-I diabetes. About 10 percent of people are with this type of diabetes. The Type-II diabetes causes resistance to the insulin produced leading to increase in blood sugar level in blood. Gestational diabetes is a metabolic disorder where the blood sugar level is high during pregnancies due to production of insulin blocking hormones by placenta. Dataset obtained contains various features on which diabetes is dependent on. Blood sugar level, Skin thickness, Age, Number of pregnancies, pedigree function, Blood pressure, Insulin, BMI and an Outcome are the features used for the prediction of diabetes. Outcome field helps to predict whether the patient is diabetic or not. F-Measure is used to measure the accuracy of SVM and Logistic Regression algorithms. SVM yields accuracy of 93% than Logistic regression with 75% accuracy.

2. Methods and Materials

A. Related Work

Kumari, V. Anuja, and R. Chitra used datasets for diabetes disease from the machine learning laboratory at University of California, Irvine. All the patients‟ data are trained by using SVM. The choice of best value of parameters for particular kernel is critical for a given amount of data SVM approach can be successfully used. SVM with Radial basis function kernel is used for classification. The performance parameters such as the classification accuracy, sensitivity, and specificity of the SVM and RBF have found to be high thus making it a good option for the classification process. Accuracy is acquired as 76% and sensitivity and specificity are measured.

Deepti Sisodia, Dilip Singh Sisodia discusses a research work focusing on pregnant women suffering from diabetes. The dataset comprises of 768 instances with 8 attributes. In this work, NaïveBayes, SVM and Decision Tree machine learning classification algorithms are used and evaluated on the PIDD dataset to find the prediction of diabetes in a patient. The performances of all the three algorithms are evaluated on various measures like Precision, Recall, Accuracy and F-Measure. Accuracy is measured over correctly and incorrectly classified instances. The results show Naïve Bayes outperforms with highest accuracy of 76.30% comparatively other algorithms. These results are verified using Receiver Operating Characteristic (ROC) curves in a proper and systematic manner.

Tejas N. Joshi, Pramila M. Chawan describes a machine learning approach to predicting diabetes. The technique may also help researchers to develop an accurate and effective tool that will reach at the table of clinicians to help them make better decision about the disease status. Logistic regression was used to predict whether a patient suffer from diabetes, Concerning the complexity and variety of data, the final result is 0.78. SVM showed better performance in accuracy as the best result is around 0.79.

Minyechil Alehegn and Rahul Joshi, Preetimulay, in this study the proposed method provides high accuracy with accuracy value of 90.36% and decision Stump provided less accuracy than other by providing 83.72% accuracy. Therefore, using ensemble method used to provide better prediction performance or accuracy than single one. A total of 768 records, data set from PIDD (Pima Indian Diabetes Data Set) which is access from online source. In the proposed system most known predictive algorithms are applied SVM, Naïve Net,

Decision Stump, and Proposed Ensemble method (PEM). An ensemble hybrid model by combining the individual techniques/methods into one we made Proposed Ensemble method (PEM). The proposed ensemble method (PEM) provides high accuracy of 90.36%

J. Steffi, R. Balasubramanian, K. Aravind Kumar discusses the performance of data mining classification algorithms using Pima Indian Diabetes Dataset with 768 instances and 9 attributes. Naïve Bayes, Logistic Regression, ANN, C5.0 Decision tree and support Vector Machines algorithms are used to model for predicting diabetes using common risk factors. Overall performance analysis of algorithms shows that C5.0 and logistic regression are equally good on their accuracy measures and gradually decreases as naïve bayes followed by ANN and SVM showed lowest accuracy.

Priya B. Patel and Parth P. Shah, Himanshu D. Patel describes prediction of diabetes using different data mining algorithms. Naive bayes, KNN, SVM, Decision tree algorithms are discussed. Errorrates for the prediction of disease using each algorithm are calculated. The analysis shows the order of algorithms as Naïve Bayes, KNN, SVM followed by decision tree. SVM acquired accuracy of 64.173% and error rate of 0.29 in Pima Indians Diabetes Dataset.

Md. Aminul Islam and Nusrat Jahan investigated different types of machine learning classification algorithms and shown their comparative analysis. The purpose of this study is to detect the diabetic patient's onset from the outcomes generated by machine learning classification algorithms. With the default configuration, logistic regression had the highest accuracy (78.01%) and AUC (0.833). The second highest accuracy (77.08%) obtained from SVM classifiers but obtained comparatively lower AUC (0.716). Sensitivity and specificity are quite satisfactory. Thus, the SVM classifier appears to perform second best among all 10 classifiers.

B. Dataset Description

Dataset is obtained from Kaggle Online dataset repository containing 2000 instances with 9 attributes. The attributes are namely Number of Pregnancies, Glucose, Blood pressure, Skin thickness, insulin, BMI, Pedigree Function, Age and an outcome field.

3. Methodology Used

A. Preprocessing

The values in the dataset are checked for null values, duplication and data cleaning methods such as missing values, binning are used to clean and preprocess the data for prediction.

B. Support Vector Machines

Support Vector Machine algorithm finds an optimum hyperplane between two classes to classify all the data points which are called support vectors. This is a discriminative classifier[6].In the training phase, numerical values of the features are used to plot the points on a N-dimensional hypercube which is drawn representing each feature as the dimensions. In the testing phase, based on the side of the decision boundary the data falls on SVM determines 1/0 outcome about diabetes. According to SVM, distance between the line and support vectors is computed and is called as margin. An optimal hyperplane has the maximum margin.

1) Tuning Parameters-Gamma

Gamma is a parameter of the RBF kernel that defines how the influence of the single training example reaches. In case of SVC classifier, the decision boundary will be curvy when gamma value is "high" and becomes linear with gradual decrease in gamma value.

2) C is penalty parameter for misclassification of points.

The classifier is clearly tolerant of misclassified data points if C value is "low" and gradually becomes intolerant with increase in gamma value. In this paper, SVM is used to predict diabetes using 8 given attributes and yields a better accuracy of about 0.93.

C. Logistic Regression

Logistic regression can be used if, the dependent variable is categorical. The algorithm can be used for predicting the probability of success or failure of a given process [3]. The sigmoid function or logistic function

$$F(x)=1 / (1 + e ^ -x) \quad (1)$$

Is derived using real valued numbers and mapped into values between 0 and 1. Input values(x) are linearly combined using coefficient values to predict an output(y) value. The logistic regression equation is

$$Y = e^(b0 + b1 * x) / (1 + e^(b0 +b1 * x)) \quad (2)$$

Outcome field has binary values in the dataset representing the person to be diabetic or not. Logistic regression was used to predict whether a person is diabetic or not using the eight attributes. In this paper, Logistic regression yields an accuracy of about 0.76.

D. Accuracy Calculation

Various performance measures are used to determine the accuracy of the algorithms.

Accuracy= TP+TN /TP+TN+FP+FN

Precision= TP / TP + FP

Recall= TP / TP + FN

F=2 * Precision * Recall / Precision + Recall

Where TP defines True Positive, TN defines True Negative, FP defines False Positive and FN defines False Negative[2].The corresponding classifiers performance on the basis of classified instances are shown in the table below.

Table 1
Accuracy of algorithms

Algorithms	Correctly Classified Instance
Support Vector Machine (SVM)	93%
Logistic Regression	76%

Accuracy of SVM is found to be higher than Logistic Regression.

4. Conclusion

In this paper, algorithms Support Vector Machine and Logistic Regression has been used to predict the diabetes for the data containing 2000 instances factors depending diabetes. Algorithms are evaluated based on various performance measures, SVM yields accuracy of about 93% whereas Logistic Regression yields 76% accuracy.

References

[1] Kumari, V. Anuja, and R.Chitra, “Classification of diabetes disease using support vector machine,” International Journal of Engineering Research and Applications, vol. 3, no. 2, pp. 1797-1801, (IJERA). Mar-Apr, 2013.

[2] Deepti Singh Sisodia, Dilip Singh Sisodia, “Prediction of Diabetes using Classification Algorithms,” in ICCIDS., 2018.

[3] Tejas N. Joshi, Pramila M. Chawan, "Logistic Regression and SVM based Diabetes Prediction system" International Journal for Technological Research in Engineering, vol. 5, No. 11, July 2018.

[4] Minyechil Alehegn and Rahul Joshi, Preetimulay" Analysis and Prediction of Diabetes using Machine Learning Algorithm", International Journal of Pure and Applied Mathematics, vol. 118, No. 9, pp.871-878, 2018.

[5] J. Steffi, R. Balasubramanian, K. Aravind Kumar, “Predicting Diabetes mellitus using Data Mining Techniques,” International Journal of Engineering Development and Research, vol. 6, no. 2, 2018.

[6] Priya B. Patel, Parth P. Shah, Himanshu D. Patel “Analysis of Data Mining Algorithms for Prediction of Diabetes,” International Journal of Engineering Development and Research, vol. 5, no. 3, 2017.

[7] Md. Aminul Islam, Nusrat Jahan, “Prediction of Onset Diabetes using Machine Learning Techniques,” International Journal of Computer Applications, vol. 180, no. 5, December 2017.