IOSR Journal of Electrical and Electronics Engineering (IOSR-JEEE)
e-ISSN: 2278-1676,p-ISSN: 2320-3331, Volume 13, Issue 6 Ver. I (Nov-Dec. 2018), PP 61-66
www.iosrjournals.org

Machine Learning Approach for the Prediction of Diabetes Mellitus

Anil Kumar Mishra, Sachinanadan Mohany

Abstract: *Diabetes mellitus is a common disease caused by a set of metabolic ailments where the sugar stages over drawn-out period is very high. It touches diverse organs of the human body which therefore harm a huge number of the body's system, in precise the blood strains and nerves. Early prediction in such disease can be exact and save human life. To achieve the goal, this research work mainly discovers numerous factors associated to this disease using machine learning techniques. Machine learning methods provide effectual outcome to extract knowledge by building predicting models from diagnostic medical datasets together from the diabetic patients. Quarrying knowledge from such data can be valuable to predict diabetic patients. In this research, six popular used machine learning techniques, namely Random Forest (RF), Logistic Regression (LR), Naive Bayes (NB), C4.5 Decision Tree (DT), K-Nearest Neighbor (KNN), and Support Vector Machine (SVM) are compared in order to get outstanding machine learning techniques to forecast diabetic mellitus. Our new outcome shows that Support Vector Machine (SVM) achieved higher accuracy compared to other machine learning techniques.*

Key Word*: **(11Bold)** machine learning, C4.5 Decision Tree, Support Vector Machine, Logistic Regression, Naive Bayes, K-Nearest Neighbor, and Random Forest, diabetes.*

I. Introduction

Diabetic is a disease that affects the hormone insulin, follow-on in abnormal metabolism of carbohydrates and advance steps of sugar in the blood. This great blood sugar affects several organs of the human body which in turn complicates many source of the body, in precise the blood strains and nerves. The details of diabetic is not nevertheless totally exposed, many researchers supposed that both hereditary elements and environmental effects are complex therein. As exposed by the International Diabetes Federation, extent of people having diabetes mellitus stretched 382 million out of 2013 [1] that take 6.6% of the world's total adult population. According to the world healthcare medical data it has been probable that diabetic patients will be increased up to 490 billion within the year 2030 [2]. Furthermore, diabetic is imaginably independent causal factor to micro-vascular entanglements. Diabetic patients are maybe more incapable against a hoisted risk of micro-vascular damage, in this way long term difficulty of cardio-vascular disease is the leading reason of death. This micro-vascular harm and hasty cardio vascular disease ultimately quick to retinopathy, nephropathy and neuropathy [3].Early prediction of such disease can be controlled over the diseases and save human life. To accomplish this goal, this research work mainly discovers the early prediction of diabetes by taking into account various risk factors related to this disease. For the willpower of the study we gathered diagnostic dataset having 16 attributes diabetic of 2000 patients. These attributes are age, diet, hyper-tension, problem in vision, genetic etc. In later part, we debate about these attributes with their conforming values. Based on these attributes, we figure prediction model by means of various machine learning techniques to predict diabetes mellitus. Machine learning techniques provide well-organized result to extract knowledge by making predicting models from diagnostic medical datasets composed from the diabetic patients. Haul out knowledge from such data can be beneficial to predict diabetic patients. Innumerable machine learning techniques have the knack to predict diabetes mellitus. Though it is very difficult to select the best technique to predict based on such attributes. Thus for the determination of the study, we deal six popular machine learning algorithms, namely Support Vector Machine (SVM), Naive Bayes (NB), K-Nearest Neighbor (KNN) ,Logistic Regression (LR) ,Random Forest (RF) and C4.5 decision tree (DT), on adult population data to predict diabetic mellitus.

II. Related Work

Various researchers have been shown revisions in the area of diabetic by using machine learning techniques to extract knowledge from existing medical data. For illustration, ALjumah et. al. [4] established a predictive analysis model using support vector machine algorithm. In [5], Kavakiotis et al. used 10 fold cross validation as evaluation method in three different algorithms, including Logistic regression, Naive Bayes, and SVM, where SVM on condition that better performance and accuracy of 84 % than other algorithms. In [6], Zheng et al. applied Random Forest, KNN, Naive Bayes, SVM, decision tree and logistic regression to predict diabetes mellitus at early stage, where cleaning criteria can be improved. Swarupa et al. [7] applied J48, ANN, KNN, ZeroR and NB on various diabetes dataset. Pradeep et al.[8] applied Random Forest, KNN, SVM

and J48 where J48 shows better performance than others. The classification algorithms did not assess using cross validation method. To predict and control diabetes mellitus Huang et al [9] conversed three data mining methods, including IB1, Naïve Bayes and C4.5 in the year of 2000 to 2004. By smearing feature selection technique, the performance of IB1 and Naive Bayes provided better result. In [10], Xue-Hui Meng et al. used three different data mining techniques ANN, Logistic regression, and J48 to predict the diabetic diseases using real world data sets. Finally it was concluded as J48 performs better accuracy than others. In this work, we examine real diagnostic medical data based on numerous risk factors using popular machine learning classification techniques to assess their performance for predicting diabetes mellitus.

III. Methodology

In order to accomplish our goal study methodology includes of few stages, which are accrual of diabetes dataset with the applicable attributes of the patients, preprocessing the numeric value attributes, to smear several machine learning techniques and conforming predictive analysis employing such data. In the subsequent, we fleetingly confer these stages.

3.1. Dataset and Attributes

In this work, datasets have been collected among the Pima Indian female population near Phoenix, Arizona. This particular dataset has been widely used in machine learning experiments and is currently available through the UCI repository of standard datasets. This population has been studied continuously by the National Institute of Diabetes, Digestive and Kidney. UCI repository contains 768 instances of observations and total 9 attributes with no missing values reported. Data sets contains 8 particular variables which were considered high risk factors for the occurrence of diabetes, like number of times pregnant, plasma glucose concentration at 2 hour in an oral glucose tolerance test (OGTT), diastolic blood pressure, 2 hour serum insulin, body mass index, diabetes pedigree. All the patients in this datasets are female at least 21 years old living near Phoenix, Arizona. All attributes are numeric values except class is nominal type. Attributes name and types are shown in table 1

Table 1 Dataset Description

No	Name of attributes	Type
1	Number of times pregnant	Numeric
2	Plasma glucose concentration a 2 hours in an oral glucose tolerance test Numeric	Numeric
3	Diastolic blood pressure	Numeric
4	Triceps skin fold thickness	Numeric
5	2 hour serum insulin	Numeric
6	Body mass index	Numeric
7	Diabetes pedigree function	Numeric

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	2	138	62	35	0	33.6	0.127	47	1
1	0	84	82	31	125	38.2	0.233	23	0
2	0	145	0	0	0	44.2	0.630	31	1
3	0	135	68	42	250	42.3	0.365	24	1
4	1	139	62	41	480	40.7	0.536	21	0
5	0	173	78	32	265	46.5	1.159	58	0
6	4	99	72	17	0	25.6	0.294	28	0

Figure 1 Sample dataset

It is good to check the correlations between the attributes. From the output graph below, the red around the diagonal suggests that attributes are correlated with each other. The yellow and green patches suggest some moderate correlation and the blue boxes show negative correlations as shown below fig. 2.

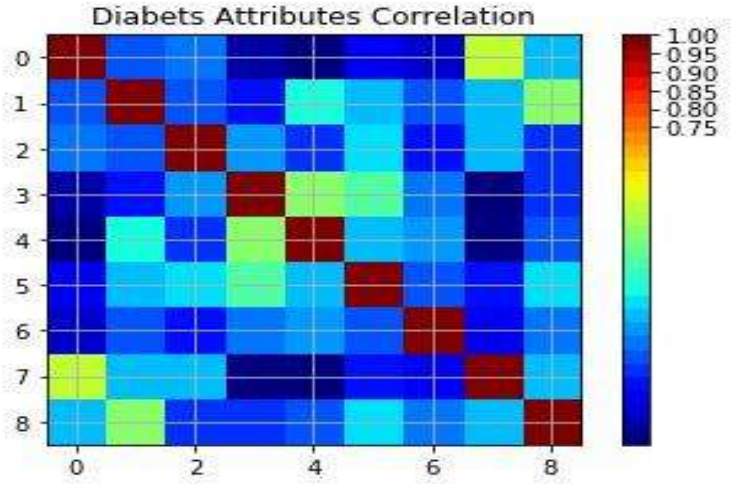

Figure 2 Attributes Correlation

Next, we visualize the data using density plots to get a sense of the data distribution. From the outputs below, you can see the data shows a general Gaussian distribution below fig 3.

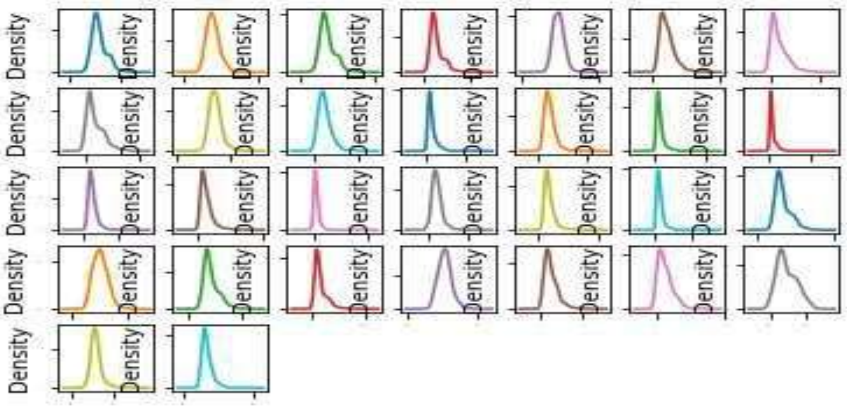

Figure 3 Density plots

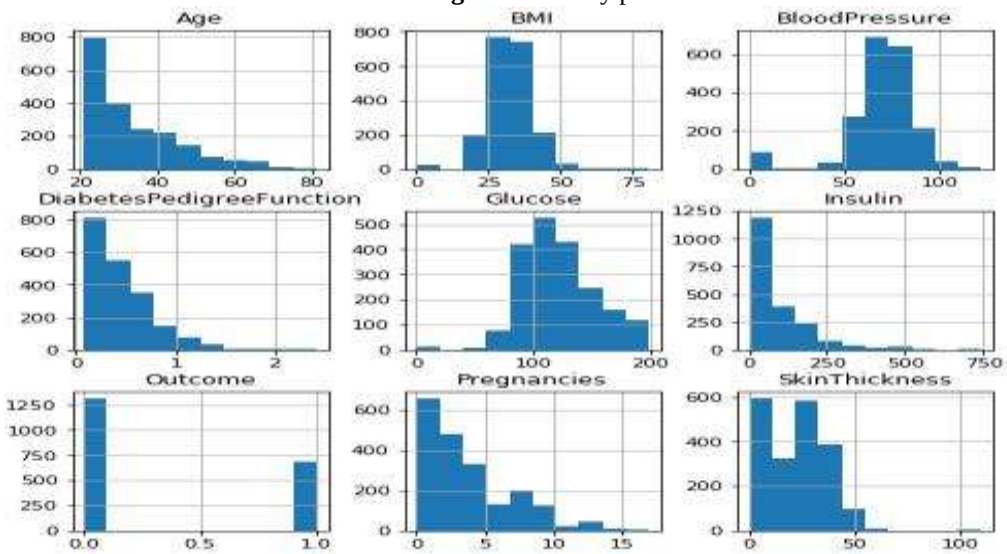

Figure 4 Histogram

According to Fig. 5 below, after defining the problem we collect the relevant data from the Diagnostic Data Storage. We then preprocess the data for the purpose of building the prediction model. After that we apply various machine learning techniques discussed above on the training dataset. Finally, test dataset is used to measure the performance of the techniques in order to choose the best classifier for predicting diabetes mellitus.

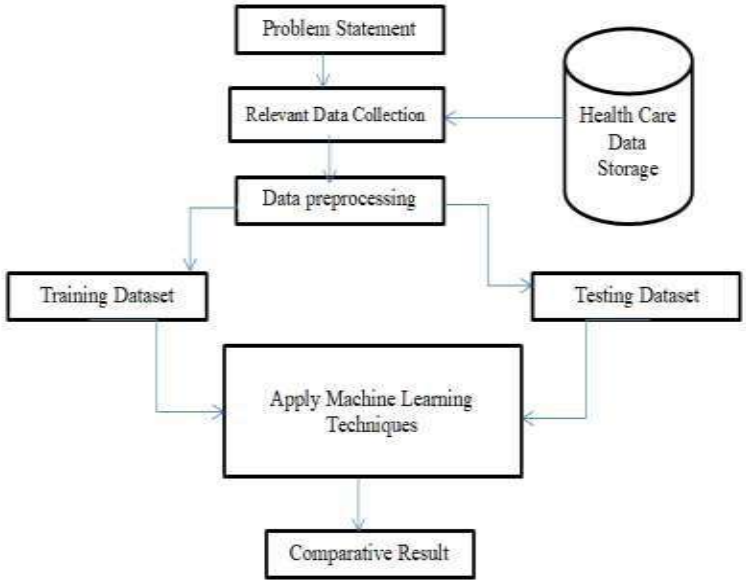

Figure 5 Shows an overview of the overall process of our work.

3.2. Apply Machine Learning Techniques

After the data has been ready for demonstrating, we use six popular machine learning techniques to predict diabetes mellitus. Later we offer an overview of these techniques.

Support Vector Machines: is widely classification technique proposed by J. Platt et. al. [11]. A Support Vector Machine is known by characterizing the data by separating a hyper plane. SVM detaches entities in specified classes. It can also recognize and classify instances which are not supported by data. SVM is not caring in the distribution of acquiring data of each class. The one delay of this algorithm is to execute regression analysis to produce a linear function and another extension is learning to rank elements to yield classification for individual elements.

Naive Bayes: is a popular probabilistic classification technique proposed by John et. al. [12]. Naive Bayes also called Bayesian theorem is a simple, effective and commonly used machine learning classifier. The algorithm calculates probabilistic results by counting the regularity and combines the value given in data set. By using Bayesian theorem, it adopts that all attributes are independent and based on variable values of classes. In real world solicitation, the conditional independence hypothesis rarely holds true and gives well and more sophisticate classifier results.

K-Nearest Neighbor Algorithm: K-nearest neighbor is simple regression and classification algorithm that used non parametric method proposed by Aha et. al. [13]. The algorithm trains all usable attributes and classifies new query based on their likeness measure. To determine the distance from point of interest to points in training data set it uses tree like data structure. The attribute is classified by its neighbors.

Decision Tree: is a tree that provides powerful classification techniques to predict diabetes mellitus. Every discrete area and feature of the domain is called a class. An input feature of the class attribute is labeled with the internal node in a tree. The leaf node of the tree is labeled by attribute and each attribute associated with a target value. There are some popular decision tree algorithms are accessible to classify diabetic data in machine learning techniques, including ID3, C4.5 , J48, C5, CART and CHAID. C4.5 provides extended features of ID3 decision tree algorithm proposed by Ross Quinlan et. al. [14]. C4.5 decision tree uses same training data as ID3, in which learned function is introduced. The learning method can be used to diagnose medical data to predict the value of the decision attribute.

Logistic regression: is a probabilistic statistical model for investigative a dataset in which there are one or more independent variables that govern a result. In logistic regression, the dependent variable is binary that means 1 as (TRUE, patient, etc.) or 0 as (FALSE, healthy, etc.). Logistic regression produces the coefficients of a procedure to predict a log it adjustment of the probability of occurrence of the characteristic of nosiness.

Random forest: is collaborative classification scheme based on Decision Tree. At the training stage, it produces a massive number of trees and creates a forest of Decision Trees. At the testing stage, each tree of the forest predicts a class label for each data. When each tree predicts a class label, then the final decision for each test data depends on widely held voting. Which class label gets the majority of votes this label accepts to be the correct label allotted to the test data. This process is continual for each of data in the dataset.

IV. Experimental Results and Discussion

To conduct the experiment six popular used machine learning techniques, namely Random Forest (RF), Logistic Regression (LR), Naive Bayes (NB), C4.5 Decision Tree (DT), K-Nearest Neighbor (KNN), and Support Vector Machine (SVM) are were used. Machine learning techniques were implemented in pycharm 3.6. An experimental result shows that the performance of SVM is significantly superior to other machine learning techniques for the classification of diabetic data. The experimental results could assist health care to take early prevention and make better clinical decisions to control diabetes and thus save human life. To take into account additional attributes and evaluation for further analysis is our future work.

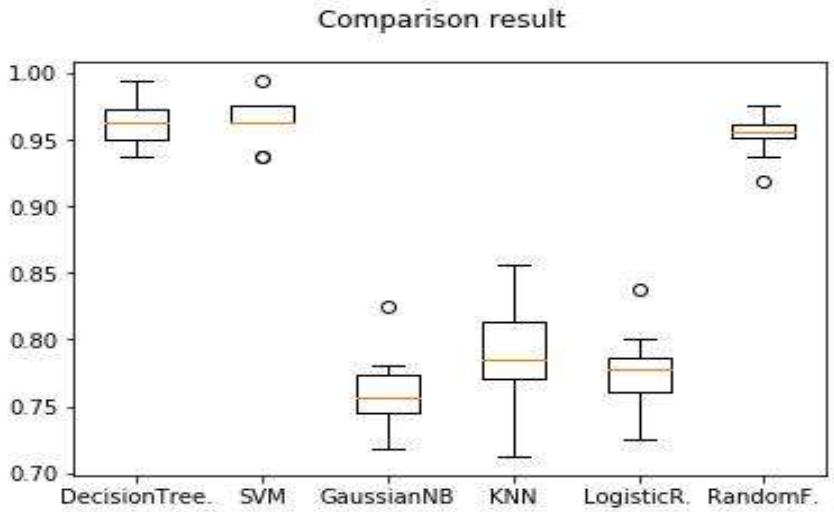

Figure 6 Comparison result of Various Machine Learning Techniques

In order to assess the performance of various machine learning techniques, we have showed the prediction results in Fig. 6 on the basis of accuracy. The figure shows the results of studies such as LR, SVM, NB, KNN, RF and C4.5 and SVM accomplishes better results than other classifiers to predict diabetes mellitus. According to Fig. 6, SVM achieves 96.4% on this dataset, which is greater than other learning techniques. This experimental result provides indication that SVM performs acceptable on medical datasets for the determination of predicting diabetes mellitus based on numerous risk factors, deliberated in the earlier section.

Table 2 Accuracy Results of Various Machine Learning Techniques

Classification Algorithms	Accuracy
Support Vector Machine (SVM)	96.4%
Decision Tree (DT)	95.8%
Random Forest (RF)	95.3%
Naïve Bayes (NB)	76.0%
Logistic Regression (LR)	77.6%
K-Nearest Neighbors (KNN)	78.8%

Overall, we have selected the best machine learning technique to predict diabetes mellitus to achieve high performance, based on the evaluation criteria discuss above. All the techniques mentioned over are estimated on an unseen testing diabetic dataset. The technique which accomplishes the highest performance in terms of accuracy is considered to be the best choice. Based on Fig. 6, it can be observed that SVM achieved the better accuracy of 96.4 % to predict diabetes mellitus utilizing a given medical dataset.

V. Conclusion

In this work, we have investigated the early prediction of diabetes by taking into account several risk factors related to this disease using machine learning techniques To predict diabetes mellitus efficiently, we have done our investigation using six popular machine learning algorithms, namely Support Vector Machine (SVM), Naive Bayes (NB), K-Nearest Neighbor (KNN), Logistic Regression (LR.), Random Forest (RF.) and C4.5 decision tree, on adult population data to predict diabetes mellitus. The technique which accomplishes the highest performance in terms of accuracy is considered to be the best choice. Based on Fig.6, it can be observed that SVM achieved the better accuracy of 96.4 % to predict diabetes mellitus utilizing a given medical dataset.

References

[1]. V., A. K. and R., C. 2013. Classification of Diabetes Disease Using Support Vector Machine. International Journal of Engineering Research and Applications. 3, (April. 2013), 1797-1801.
[2]. Carlo, B G., Valeria, M. and Jesús, D. C. 2011. The impact of diabetes mellitus on healthcare costs in Italy. Expert review of pharmacoeconomics & outcomes research. 11, (Dec. 2011),709-19.
[3]. Nahla B., Andrew et al. 2010. Intelligible support vector machines for diagnosis of diabetes mellitus. Information Technology in Biomedicine, IEEE Transactions. 14, (July. 2010), 1114-20.
[4]. Abdullah A. Aljumah et al., Application of data mining: Diabetes health care in young and old patients, Journal of King Saud University - Computer and Information Sciences, Volume 25, Issue 2,July 2013, Pages 127-136
[5]. Kavakiotis, Ioannis, Olga Tsave, AthanasiosSalifoglou, NicosMaglaveras, IoannisVlahavas, and IoannaChouvarda. "Machine learning and data mining methods in diabetes research." Computational and structural biotechnology journal (2017).
[6]. Zheng, Tao et al. "A machine learning-based framework to identify type 2 diabetes through electronic health records." International journal of medical informatics 97 (2017): 120- 127.
[7]. Rani, A. Swarupa, and S. Jyothi. "Performance analysis of classification algorithms under different datasets." In Computing for Sustainable Global Development (INDIACom), 2016 3rd International Conference on, pp. 1584-1589. IEEE, 2016.
[8]. Kandhasamy, J. Pradeep, and S. Balamurali. "Performance analysis of classifier models to predict diabetes mellitus." Procedia Computer Science 47 (2015): 45-51.
[9]. Y. Huang, P. McCullagh, N. Black, R. Harper, Feature selection and classification model construction on type 2 diabetic patients'data, Artificial Intelligence in Medicine 41 (3)(2015) 251–262.
[10]. Meng, X. H., Huang, Y. X., Rao, D. P., Zhang, Q., & Liu, Q. (2013). Comparison of three data mining models for predicting diabetes or prediabetes by risk factors. The Kaohsiung journal of medical sciences, 29(2), 93-99.
[11]. Platt, John C. "12 fast training of support vector machines using sequential minimal optimization." Advances in kernel methods (1999): 185-208.
[12]. John, George H., and Pat Langley. "Estimating continuous distributions in Bayesian classifiers." Proceedings of the Eleventh conference on Uncertainty in artificial intelligence. Morgan Kaufmann Publishers Inc., 1995.
[13]. Aha, David W., Dennis Kibler, and Marc K. Albert. "Instance-based learning algorithms." Machine learning 6.1 (1991): 37-66.
[14]. Ross Quinlan (1993). C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, San Mateo, CA