Journal of Physics: Conference Series

Analysis of machine learning algorithms in diabetes mellitus prediction

To cite this article: R Saxena *et al* 2021 *J. Phys.: Conf. Ser.* **1921** 012073

ECS
The Electrochemical Society
Advancing solid state & electrochemical science & technology
18th
239th ECS Meeting with IMCS18
DIGITAL MEETING • May 30-June 3, 2021
Live events daily • Free to register
Register now!

Analysis of machine learning algorithms in diabetes mellitus prediction

R Saxena[1], S K Sharma[1], M Gupta[1]

[1]School of Information and Communication Technology, Gautam Buddha University, Greater Noida 201312, India

Saxena.roshi@gmail.com

Abstract. Diabetes mellitus is one of the enduring and lethal disease which is caused due to enlarged levels of glucose in the blood. Therefore, it's quite necessary to predict the diabetes in its early stage. In this paper, we have studied machine learning algorithms to predict the diabetes. We have made use of PIMA Indians diabetes dataset and weka (a tool to implement machine learning algorithms) and predicted the precision, recall, accuracy, f-measure and receiver operating characteristics.

1. Introduction

Diabetes mellitus is one of the enduring and lethal disease which is caused due to enlarged levels of glucose in the blood. If diabetes, not predicted on time and is untreated for a longer amount of time, it can be hazardous. Metabolism of the person is highly affected due to the less amount of insulin production and increase in increment of sugar levels in the body. When the amount of insulin in the blood is less, body cells respond to insulin an unfortunate manner. If the ailment does not obtain proper treatment on time, it can have a major influence on kidneys, nervous system, retina of the eyes, heart disease. Sometimes it can also lead to failure of some organs and to even death also. In fig 1, We can see the death rate of different categories due to diabetes. It's very necessary to predict diabetes at an early stage. In this research, our main objective to use machine learning procedures and their important features, and design a framework for forecasting diabetes at the primary phase to give the possible closest result. Factors due to which people are diagnosed with diabetes are increased level of sugar in blood and inappropriate working of beta cells in pancreas. Diabetes can be of four types: Type 1 diabetes generally happens in younger one and grown-up people. In this type of diabetes, pancreas is unable to create the insulin and the individuals are assisted through injected insulin from outer drugs to maintain the sugar levels in the body. Type 2 diabetes generally happens in the age group of 45-60. Main cause of this type of diabetes can be hereditary and metabolism. Gestational diabetes generally happens to the ladies during pregnancy. Various hormones in the pregnancy and increased quantity insulin in the blood can prompt the high blood sugar level.

2. Related Work

In recent years, a good amount of research work has been done to forecast the diabetes using machine learning technique. In [2] authors proposed different dimensionality reduction and cross-validation technique by implementing Linear Discriminant Analysis (LDA) [3], Quadratic Discriminant Analysis (QDA) [4], Naive Bayes (NB) [5], Gaussian Process Classification (GPC) [6], Support Vector Machine (SVM) [7], Artificial Neural Network (ANN) [8], AdaBoost (AB) [9], Logistic Regression (LR) [10], Decision Tree (DT) [11], and Random Forest (RF) [12]. Extensive experiments were also carried out for rejecting the outliers and filling missing values by computing the mean and the median for enhancing

Published under licence by IOP Publishing Ltd

the performance of Machine Learning model and received the highest possible area under the curve of 0.930. Authors employed decision tree, support vector machine and Naïve Bayes classifiers in [13] to calculated the diabetes with maximum accuracy and after running all the classifiers proved that the Naïve Bayes is the best amongst the three with an AUC of 0.819.

The organization of the remaining paper is as follows: "Materials and Methods section represents materials and methods, including dataset description, tool description prediction algorithms and classifiers evaluation. Results section discusses the results of all classifiers applied. Conclusion discusses the summary of current work and future work.

3. Materials and methods

3.1 Dataset Description

The PIMA Indian diabetes dataset have been used in this proposed study, derived from UCI repository and is available online in free. The dataset is of 768 people, of which 268 were diabetic and 500 were non-diabetic. There were some missing values and unusual observation in the dataset which were pre-processed using data pre-processing technique. Detailed description of the dataset is shown in table 1. Weka 3.9.4 has been used to implement machine learning techniques. We have run all classifiers on Weka and noted down the results.

Table 1. Description of PIMA Indian diabetes Dataset

S.No	Attributes	Mean	Standard Deviation	Min/Max Value
1	No. of times pregnant	3.8	3.4	1/17
2	Plasma glucose concentration	120.9	32	56/197
3	Diastolic Blood Pressure	69.1	19.4	24/110
4	Triceps skin fold thickness(mm)	20.5	16	7/52
5	2-hour serum insulin	79.8	115.2	15/846
6	Body mass index(kg/m2)	32	7.9	18.2/57.3
7	Diabetes pedigree function	0.5	0.3	0.0850/2.32
8	Age	33.2	11.8	21/81
9	Class		Tested Positive:	Diabetic
			Tested Negative:	Non-Diabetic

3.2 Methodology

In our methodology, we have used few machine learning algorithms such as Naïve Bayes, Support vector Machine, Random Forest, Neural Network, logistic regression and a brief study of all the algorithms are presented below. After the brief study of all algorithms, we have presented a confusion matrix of above said classifiers. Table 2 to 6 shows the confusion matrix of Naïve Bayes, random forest, logistic regression, neural network and support vector machines.

3.2.1 Naïve Bayes. Naïve Bayes is a machine learning algorithm which considers all attributes are independent and not related to each other. It assumes that the status of particular attribute remains unaffected by the status of another attribute. Naïve Bayes [1] classification technique is based on conditional probability and it makes use of Bayes theorem which tells us that the posterior probability of target class i.e. $P(C|X)$ can be calculated from class C 's probability being true i.e. P(C), predictor's prior probability i.e. P(X) and predictor class's probability i.e. $P(X|C)$ [2].
So, $\mathrm{P}(C|X) = (P(X|C)\ \mathrm{P(C)})/P(X)$. Confusion matrix is shown in table 2.

Table 2. Confusion Matrix of Naïve Bayes algorithm

	Diabetic	Non-diabetic
Diabetic	164	104
Non-diabetic	78	422

Classification has classified the patients into:

- 164 (True positive): are predicted as diabetic patient
- 104 (False Negative): are predicted to have no diabetes but are actually suffering from diabetes.
- 78(False positive): are predicted to have diabetes but are not suffering from diabetes. (False positive)
- 422(true negative): are predicted as non-diabetic patient

3.2.2 Random Forest. Random forest is a collection of large number of decision trees. Each tree in the forest makes a prediction and the class which has got the most votes becomes the algorithm's prediction. In the random forest classifier, multiple number of decision trees are used in which some of the trees may predict the wrong information while some of the trees predict right information. So, as a group, they will be able to move in right direction and make correct predictions. Predictions made by individual trees should have less co-relation amongst themselves. Table 3 discusses the result after application of Random forest classifier.

Table 3: Confusion Matrix of Random Forest algorithm

	Diabetic	Non-diabetic
Diabetic	106	162
Non-diabetic	82	418

3.2.3 Logistic Regression. Logistic regression works with probability and works on discrete set of classes. Observations are assigned to discrete classes. By making use of this classifier, we can predict whether a person has certain disease or not. Output is transformed to a sigmoid function to return a probability value. Confusion Matrix of logistic regression is being shown in table 4.

Table 4: Confusion Matrix of Logistic Regression algorithm

	Diabetic	Non-diabetic
Diabetic	153	115
Non-diabetic	60	440

3.2.4 Neural Network. Neural Network is a supervised machine learning algorithm used for backpropagation. It is a feedforward artificial neural network which consists of input layer, a hidden layer and an output layer. Hidden layer nodes and output nodes are neurons which makes use of non-linear activation function. It can easily differentiate data which is not linearly separable. Results after application of neural network classifier are shown in table 5.

Table 5: Confusion Matrix of Neural Network algorithm

	Diabetic	Non-diabetic
Diabetic	164	104
Non-diabetic	87	413

3.2.5 Support Vector Machines: Support vector machine is used to separate the hyperplane between the two classes, classes used here are diabetic and non-diabetic. In this, hyperplane should be far from the data points that belong to the other class. Support vectors are points which are closed to the margin of the classifier. It maximizes the distance between decision boundaries and calculate the optimal separation hyperplane. Mathematically, maximized distance between the hyperplane is defined as wT $x + b = -1$ and the hyperplane will be defined by wT $x + b = 1$ The distance between the hyperplanes is equal to 2 _w_. i.e we have to solve max 2 _w_. Parallelly, we have to evaluate min _w_| 2. All x(i) should also be correctly classified by the SVM which means *yi* (*wT xi* + *b*) >= 1, _*i* _ {1, ¢¢, *N*}. Application of support vector machine on PIMA Indians diabetes dataset is shown in table 6.

Table 6: Confusion Matrix of Support Vector Machine algorithm

	Diabetic	Non-diabetic
Diabetic	164	162
Non-diabetic	82	418

4. Results and discussion

We have made use of Naïve Bayes, random forest, logistic regression, neural network and support vector machines in this research work. We have performed experiments using 10-folds cross validation technique. Parameters such as precision, recall, accuracy, f-measure, receiver operating curve (ROC) and area under the curve (AUC) are used for the classification. Results are discussed in table 7.

Table 7: Classification Results

Classifiers	Precision	Recall	Accuracy	F-	ROC
			measure		
Logistic Regression	0.767	0.772	77.2	0.765	0.832
Support Vector Machine	0.766	0.771	77.08	0.761	0.717
Naïve Bayes	0.759	0.763	76.30	0.760	0.819
Random Forest	0.751	0.755	75.5	0.752	0.820
Neural Networks	0.748	0.751	75.1	0.749	0.791

Comparative approach for the following parameters for various classifiers is shown in figure 1-5.

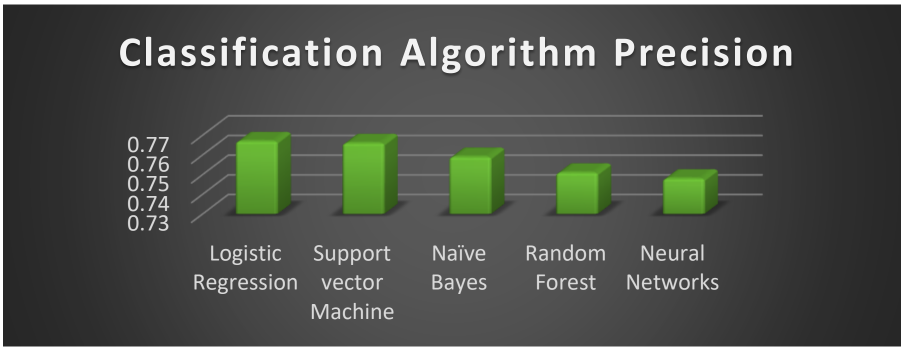

Figure 1: Precision comparison

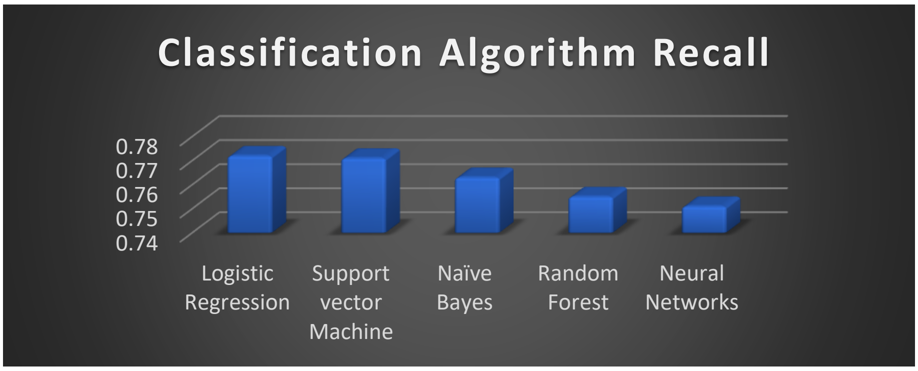

Figure 2: Recall comparison

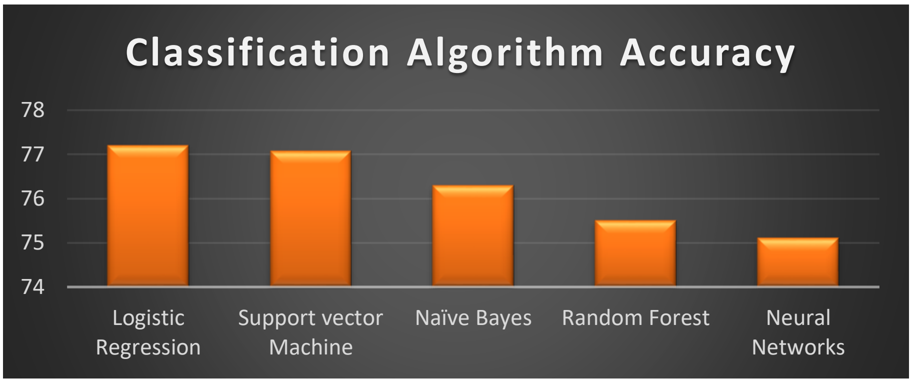

Figure 3: Accuracy comparison

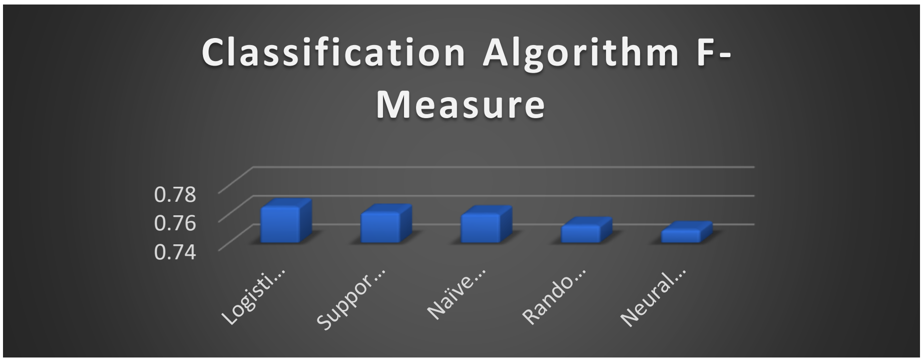

Figure 4: F-measure comparison

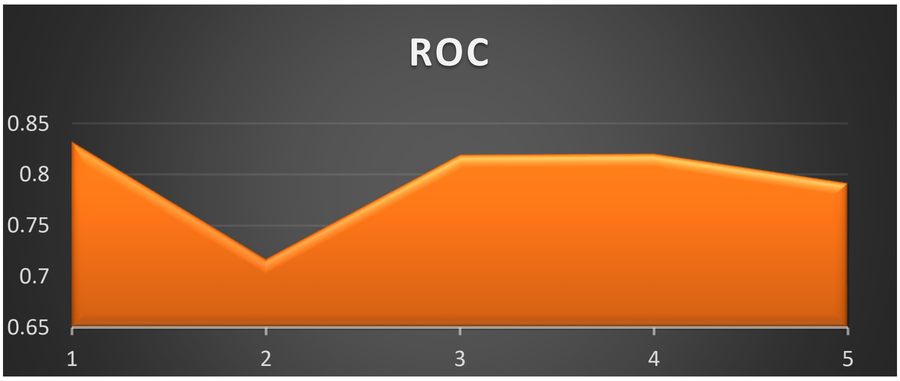

Figure 5: ROC for various classifiers

Table 8 shows the different classifier's correct classified instances and incorrect classified instances. Figure 6 shows the performance of various classifiers and by looking at the table 8 and figure 6 we can say that logistic regression classifier can predict the chances of diabetes with more accuracy as compared with other classifiers.

Table 8: Classification results where the total instances: 768

Classifiers	Correctly Classified Instances	Incorrectly Classified Instances
Logistic Regression	593	175
Support Vector Machine	592	176
Naïve Bayes	586	182
Random Forest	580	188
Neural Networks	577	191

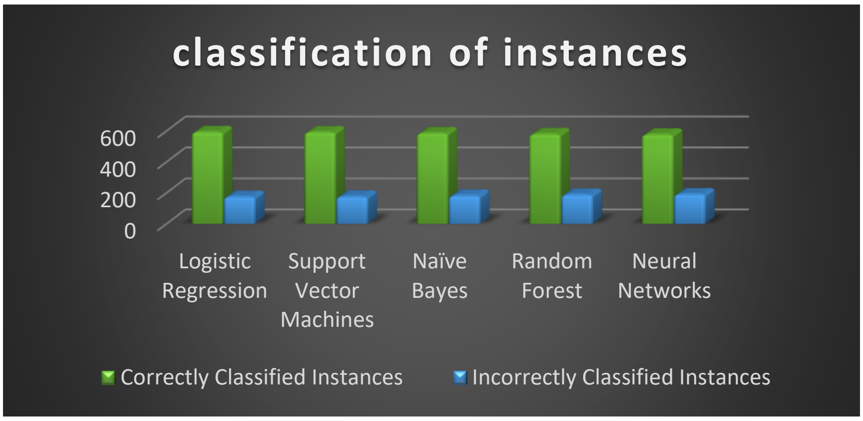

Figure 6: Classified Instance Comparison

Table-8 provides performance of various machine learning classifiers by identifying correctly classified instances and incorrectly classified instances whereas the total number of instances was 768. Performance was calculated on the basis of precision, recall accuracy, f-measure and receiver operating characteristics. Various parameters were analyzed and results were recorded in a table 6 and 7. Graph 1-6 shows the performance of various classification algorithms on the basis of various parameters. we can conclude from table-7 and table-8 that logistic regression classification techniques outperform every other classification technique as it gives the highest accuracy.

5. Conclusion and future work

In this paper, we have tried to identified the best machine learning classification technique to detect the diabetes at an early stage. During our research work, we have studied five classification techniques and evaluate them on various performance measures. Experiments were carried on PIMA Indian diabetes dataset and tool used was Weka. Experimental results determine that logistic regression classification technique gives the best accuracy to detect diabetes. Future work will include the feature selection and extraction of data and pre-processing of data by outlier rejection, filling missing values by mean filter which will further enhance accuracy.

References

[1]. Hasan K, Alam A, Das D, Hussain E and Hasan M, April 2020 “Diabetes prediction using ensembling of different machine learning classifiers”, *IEEE Acess*, vol. **8**, p 76516-31

[2]. Maniruzzaman M, Rahman M J, Hasan M, Suri H S, Abedin M M, El-Baz A, and Suri J S, Jan 2020 “Classification and prediction of diabetes disease using machine learning paradigm," *J. health information science and system.*, vol. **42**, no. 5, p. 92 -103**.**

[3]. Kamadi V S, Varma P, Rao A A, Mahalakshmi T S and Rao P V N , July 2014 “A Computational Intelligence approach for a better diagnosis of diabetic patients” , *J. Computers and Electrical Engineering* , Vol. **40** Issue 5, , p 1758-65.

[4]. Maniruzzaman M, Rahman M J, Hasan M, Suri H S, Abedin M M, El-Baz A, and Suri J S, May 2018. “Accurate diabetes risk stratification using machine learning: Role of missing value and outliers," *J. Medical System*, vol. **42**, no. 5, p. 92

[5]. Cover T M, Jun. 1965, “Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition”, *IEEE Trans. Electronic Computers*, vol. **EC_14**, no. 3, pp. 326-34.

[6]. McLachlan G, Jun. 2005, “Discriminant analysis and statistical pattern recognition” *J. Roy. Stat. Soc., Ser. A, Statist. Soc.*, vol. **168**, no. 3, pp. 635-36.

[7]. Webb G I, Boughton J R, and Wang Z, Jan. 2005. “Not so naive bayes: Aggregating one-dependence estimators," *J. Machine Learning*, vol. **58**, no. 1, pp. 5-24.

[8]. Tabaei B P and Herman W H, Nov. 2002, “A multivariate logistic regression equation to screen for diabetes: Development and validation," *Diabetes Care*, vol. **25**, no. 11, pp. 1999_2003.

[9]. Reinhardt A and Hubbard T, May 1998 “Using neural networks for prediction of the subcellular location of proteins," *Nucleic Acids Res.*, vol. **26**, no. 9, pp. 2230-36.

[10]. Cortes C and V. Vapnik, Sep. 1995 “Support-vector networks," *Mach. Learn.*, vol. 20, pp. 37-297.

[11]. Breiman L, Oct. 2001 “Random forests," *Mach. Learn.*, vol. **45**, no. 1, pp. 5-32.

[12]. Belhouari S B and Bermak A, Nov. 2004, “Gaussian process for nonstationary time series prediction," *Comput. Statist. Data Anal.*, vol. **47**, no. 4, pp. 705-12.

[13]. Jenhani I, Amor N B, and Elouedi Z, Aug. 2008, “Decision trees as possibilistic classifiers," *Int. J. Approx. Reasoning*, vol. **48**, no. 3, pp. 784-807.

[14]. Sisodia D and Sisodia S, Jan. 2018, “Prediction of diabetes using classification algorithms," *Procedia Comput. Sci.*, vol. **132**, pp. 1578-1585