International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249 – 8958 (Online), Volume-9 Issue-5, June 2020

Effective Prediction of Diabetes Mellitus using Nine different Machine Learning Techniques and their Performances

Shashank Joshi, Vijayendra Gaikwad, Sairam Rathod, Anamika Rathod, Neha Sagar

Abstract: ***Diabetes is a disease where the predominant finding is high blood sugar. The high blood sugar may either be because of deficient insulin production (Type 1) or insulin resistance in peripheral tissue cells (Type 2). Many problems occur if diabetes remains untreated and unidentified. It is additional inventor of various varieties of disorders for example: coronary failure, blindness, urinary organ diseases etc. Nine different machine learning techniques are used in this research work for prediction of diabetes. A dataset of diabetic patient's is taken and nine different machine learning techniques are applied on the dataset. Positive likelihood ratio, Negative likelihood ratio, Positive predictive value, Negative predictive value, Disease prevalence, Specificity, Precision, Recall, F1-Score ,True positive rate, False positive rate of the applied algorithms is discussed and compared. Diabetes is growing at an increasing in the world and it requires continuous monitoring. To check this we use Logical regression, Random forest, Logical regression CV, Support Vector Machine, Artificial Neural Network (ANN), Decision Tree, k-nearest neighbors (KNN), XGB classifier.***

Keywords: Diabetes Prediction, SVM, Random Forest ,Logical Regression, XGB classifier, Accuracy, Precision, Recall, F1-Score, Nine machine learning techniques, Twelve Measures.

I. INTRODUCTION

The annual report of World Health Organization, add up to the number of individuals experiencing diabetes is estimated to be 9.3% (463 million people), rising to 10.2% (578 million) by 2030 and 10.9% (700 million) by 2045. Diabetes affects the ability of the body in producing the hormone insulin or increasing the resistance of body cells to the insulin produced, which in turn makes the metabolism of carbohydrate abnormal and raise the levels of glucose in the blood. In Diabetes a person generally suffers from high blood sugar as an indicator. Intense thirst, Intense hunger and Frequent urination are common the symptoms caused due to high blood sugar. Many complications occur if diabetes remains untreated. This disease causes coronary failure, blindness, urinary organ diseases and hence the early detection will help the medical organization in treatment of it. Type1 diabetes occurs because of the failure of pancreas to supply enough hypoglycemic agent i.e. insulin. This type is labelled as "juvenile diabetes". The type one polygenic disease found in children beneath twenty years old. People suffer throughout their life because of the type one diabetes.

Revised Manuscript Received on May 25, 2020.
* Correspondence Author
Vijayendra Gaikwad*, Dept. of Computer Engg, VIT, Pune, India. E-mail: vij711@gmail.com
Shashank Joshi, Dept. of Computer Engg, VIT, Pune, India. E-mail: sameer.shashank18@vit.edu

Type 2 diabetes is an adult-onset diabetes value occurring in obese individual or those with the family history . In this there is a resistance of muscles, fat and other tissues to insulin. Although insulin production is in normal quantity due to resistance ,glucose cannot be effectively utilized to carry out cells metabolic activity. The usual predisposing cause is extreme weight and there is also family history. Type3 also known as Gestational diabetes occurs when a woman is pregnant and develops the high blood sugar levels without a previous history of diabetes. Therefore, it is found that in total 18% of women in pregnancy have diabetes.

So, in the older age pregnancy there is a risk of emerging the gestational diabetes in pregnancy. For predictive analysis of diabetes Logical regression, Random forest, Logical regression CV, Support Vector Machine, Artificial Neural Network (ANN), Decision Tree, k-nearest neighbors (KNN), XGB Classifier, Naive Bayesian are used. Positive likelihood ratio, Negative likelihood ratio, Positive predictive value, Negative predictive value, Disease prevalence, Specificity, Precision, Recall, True positive rate, F1-Score, False positive rate, Accuracy of the applied algorithms is discussed and compared. From the above twelve measures which machine learning technique is the best for prediction of diabetes is calculated.

II. LITERATURE REVIEW

Priyanka Sonar, Prof. K. JayaMalini used Decision Tree, ANN, Naive Bayes and SVM algorithms to develop a system which predicts the diabetic risk level of a patient with a better accuracy [1]. Deepti Sisodia , Dilip Singh Sisodia used three machine learning algorithms which are Decision tree, SVM, Naive Bayesian to help diagnose diabetes. They based analysis on Precision, Recall, F-measure and accuracy [2]. Muhammad Azeem Sarwar, Nasir Kamal, Wajeeha Hamid, Munam Ali Shah used six different algorithms for helping doctors in early prediction of diabetes using machine learning techniques [3]. Deeraj Shetty, Kishor Rit , Sohail Shaikh, Nikita Patil used algorithms Naive Bayes and K-Nearest Neighbor (KNN) on diabetic patient's database for predictive analysis of diabetes.[4]. Zhilbert Tafa and Nerxhivane Pervetica have discussed the results of algorithms that are implemented in order to predict the diagnosis reliability [5]. P. Suresh Kumar and V. Umatejaswi has presented the algorithms like Decision Tree, SVM, Naive Bayes for identifying diabetes using data mining techniques [6]. Sadegh Bafandeh Imandoust and Mohammad Bolandraftar have proposed a system using data mining [7].

III. PROPOSED MODEL

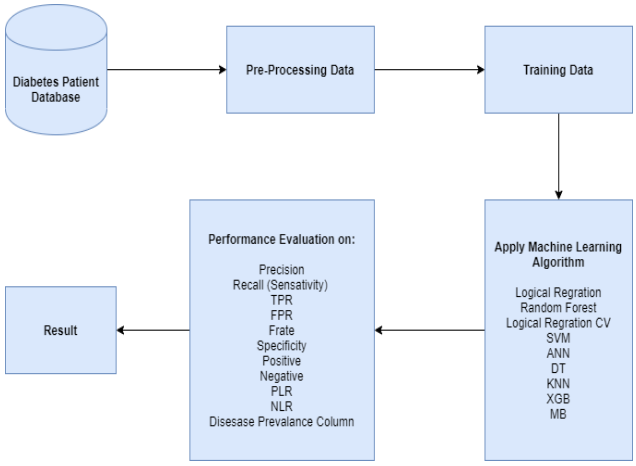

Figure 1: Proposed Model

A) Dataset Collection Global dataset:

The dataset contains 768 instances and 9 features. The dataset features are:Number of pregnancy, Glucose_concentration, Diastolic_bp, Skin fold thickness in mm, insulin, Bmi, Diabetes pedigree function, Age, Class '0' or '1'.

B) Pre-processing Data:

Data Pre-processing technique converts the raw data into an understandable data set. This includes filling missing values, removing duplicate, data conversion. In this system by using Imputer missing values are replaced with mean.

C) Training data:

Out of 768 instances 75 % which is 576 is used for training and 25% which is 192 is used for testing[2].

D) Apply Machine learning algorithm:

We used nine different machine learning algorithms Logical regression, Random forest, Logical regression CV, Support Vector Machine, Artificial Neural Network (ANN), Decision Tree, k-nearest neighbors (KNN), XGB Classifier, Naive Bayesian, for predictive analytics.

1) Decision Tree Classifier:

Table- I: Confusion Matrix

	Predicted: Yes	Predicted: No
Actual: Yes	36	26
Actual: No	28	102

Table- II: Classification Report

	Precision	recall	F1-score	Support
1	0.56	0.58	0.57	62
0	0.80	0.78	0.79	130
Accuracy			0.72	192
Macro avg	0.68	0.68	0.68	192
Weighted avg	0.72	0.72	0.72	192

2) Naive Bayes Classifier:

Table- III: Confusion Matrix

	Predicted: Yes	Predicted: No
Actual: Yes	36	26
Actual: No	13	117

Table- IV: Classification Report

	Precision	Recall	F1-score	Support
1	0.67	0.53	0.59	62
0	0.80	0.88	0.84	130
Accuracy			0.77	192
Macro avg	0.74	0.70	0.71	192
Weighted avg	0.76	0.77	0.76	192

3) Random Forest:

Table- V: Confusion Matrix

	Predicted: Yes	Predicted: No
Actual: Yes	32	30
Actual: No	14	116

Table - VI: Classification Report

	Precision	Recall	F1-score	Support
1	0.70	0.52	0.59	62
0	0.79	0.89	0.84	130
Accuracy			0.77	192
Macro avg	0.75	0.70	0.72	192
Weighted avg	0.76	0.77	0.76	192

4) Logistic Regression:

Table- VII: Confusion Matrix

	Predicted: Yes	Predicted: No
Actual: Yes	36	26
Actual: No	13	117

Table - VIII: Classification Report

	Precision	recall	F1-score	Support
1	0.73	0.58	0.65	62
0	0.82	0.90	0.86	130
Accuracy			0.80	192
Macro avg	0.78	0.74	0.75	192
Weighted avg	0.79	0.80	0.79	192

5) Support Vector Machine*:*

Table- IX: Confusion Matrix

	Predicted: Yes	Predicted: No
Actual: Yes	37	25
Actual: No	13	117

Table- X: Classification Report

	Precision	Recall	F1-score	Support
1	0.74	0.6	0.66	62
0	0.82	0.9	0.86	130

Accuracy			0.8	192
Macro avg	0.78	0.75	0.76	192
Weighted avg	0.8	0.8	0.8	192

6) Artificial Neural Network:

Table- XI: Confusion Matrix

	Predicted: Yes	Predicted: No
Actual: Yes	30	32
Actual: No	23	107

Table- XII: Classification Report

	Precision	recall	F1-score	Support
1	0.57	0.48	0.52	62
0	0.77	0.82	0.80	130
Accuracy			0.71	192
Macro avg	0.67	0.65	0.66	192
Weighted avg	0.70	0.71	0.71	192

7) k-NN:

Table- XIII: Confusion Matrix

	Predicted: No	Predicted: Yes
Actual: No	113	17
Actual: Yes	24	38

Table- XIV: Classification Report

	Precision	Recall	F1-score	Support
0	0.82	0.87	0.85	130
1	0.69	0.61	0.65	62
Accuracy			0.79	192
Macro avg	0.76	0.74	0.75	192
Weighted avg	0.78	0.79	0.78	192

8) Logical regression CV:

Table- XV: Confusion Matrix

	Predicted: Yes	Predicted: No
Actual: Yes	42	20
Actual: No	28	102

Table- XVI: Classification Report

	Precision	recall	F1-score	Support
1	0.6	0.68	0.64	62
0	0.84	0.78	0.81	130
Accuracy			0.75	192
Macro avg	0.72	0.73	0.72	192
Weighted avg	0.76	0.75	0.75	192

9) XGB Classifier:

Table- XVII: Confusion Matrix

	Predicted: No	Predicted: Yes
Actual: No	114	16
Actual: Yes	23	39

Table- XVIII: Classification Report

	Precision	Recall	F1-Score	Support
0	0.83	0.88	0.85	130
1	0.71	0.63	0.67	62
Accuracy			0.8	192
Macro avg	0.77	0.75	0.76	192
Weighted avg	0.79	0.8	0.79	192

E) Performance evaluation on:

Positive likelihood ratio, Negative likelihood ratio, Positive predictive value(PV), Negative predictive value(NV), Disease prevalence, Specificity, Precision, Recall, True positive rate, False positive rate,F1-score,Accuracy of the applied algorithms is discussed and compared. From the above we are going to calculate which algorithm is the best for early prediction of diabetes.

Table- XIX: Performance evaluation on

Precision	Tp/Tp+Fp *100
Recall (Sensitivity)	Tp/Tp+Fn *100
TPR	Tp/Tp+Fn *100
FPR	Fp/Fp+TN *100
F1-Score	2*precision*recall/precision+recall
Specificity	Tn/TN+Fp *100
Positive	Tp/Tp+Fp *100
Negative	Tn/TN+Fn *100
PLR	Sensitivity/100-specificity
NLR	100-Sensitivity/Specificity
Disease prevalence column	Tp+Fn/Total
Accuracy	Tp+TN/Tp+TN+Fp+Fn

Here, TP defines True Positive, TN defines True Negative, FP defines False positive, FN defines False Negative.

IV. RESULTS

A) Comparative analysis of various measures on which machine learning techniques are evaluated:

Table- XX: Comparative analysis of various measures on which machine learning techniques are evaluated .

Sr. no	Algor-ithm	Precision	Recall	TPR	FPR	F1-Score	Specificity	PV	NV	PLR	NLR	Accuracy	Disease Prevalence
1	LR	73.0416	58.064	58.064	10	64.844	90	73.416	81.8181	5.826	0.465	79.69	32.291
2	RF	69.565	51.612	51.612	10.76	59.259	89.23	69.565	79.452	4.792	0.208	77.08	32.291
3	LRCV	60	67.741	67.741	21.538	63.635	78.461	60	83.606	3.145	0.4111	75	32.291
4	SVM	74	59.677	59.677	10	66.071	90	74	82.394	5.967	0.448	80.21	32.291
5	ANN	56.603	48.387	48.387	17.692	52.173	82.03	56.603	76.978	2.734	0.627	71.35	32.291
6	DT	56.25	58.064	58.064	21.538	57.142	78.461	56.25	79.687	2.695	0.534	71.88	32.291
7	KNN	70.909	62.903	62.9032	12.308	66.666	87.692	70.909	83.211	5.1108	0.423	78.645	32.291
8	XGB	69.0909	61.29	61.2903	13.071	64.9572	86.923	69.09	82.48	4.6868	0.4453	79.69	32.291
9	NB	67.346	53.225	53.225	12.3	59.459	87.692	67.346	79.72	4.3245	0.2312	76.56	32.291

- LR(Logical Regression)
- RF(Random Forest)
- LRCV(Logical Regression CV)
- SVM(Support Vector Machine)
- ANN (Artificial Neural Network)
- DT (Decision Tree)
- KNN (K-Nearest Neighbors)
- XGB(Extreme Gradient Boosting)
- NB(Naïve Bayes)

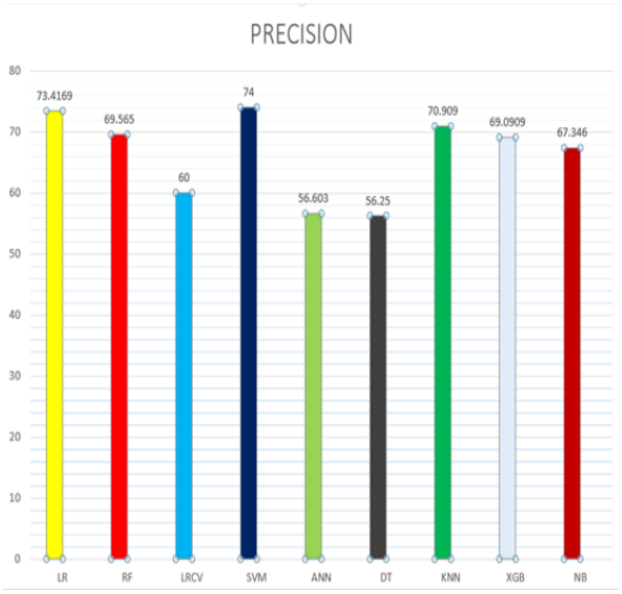

Figure 2: Precision

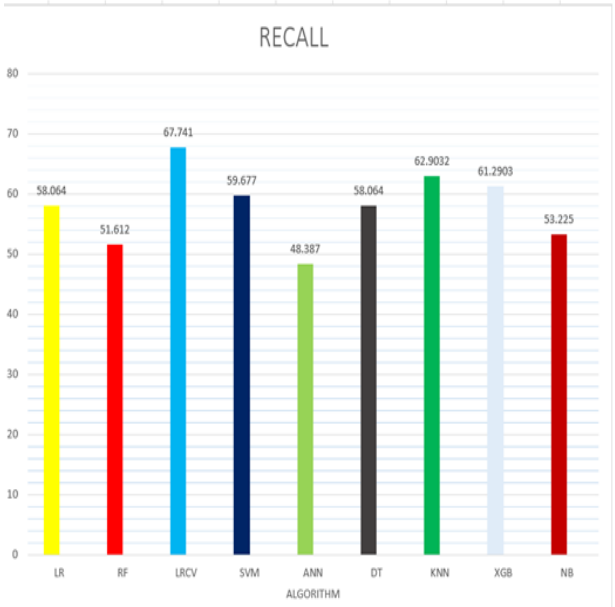

Figure 3: Recall

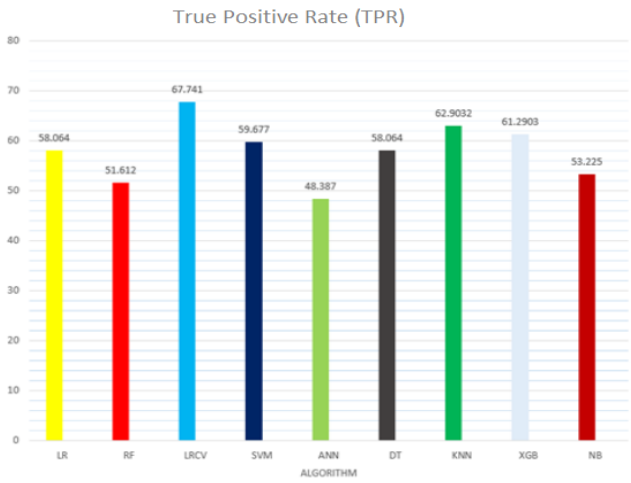

Figure 4: True Positive Rate(TPR)

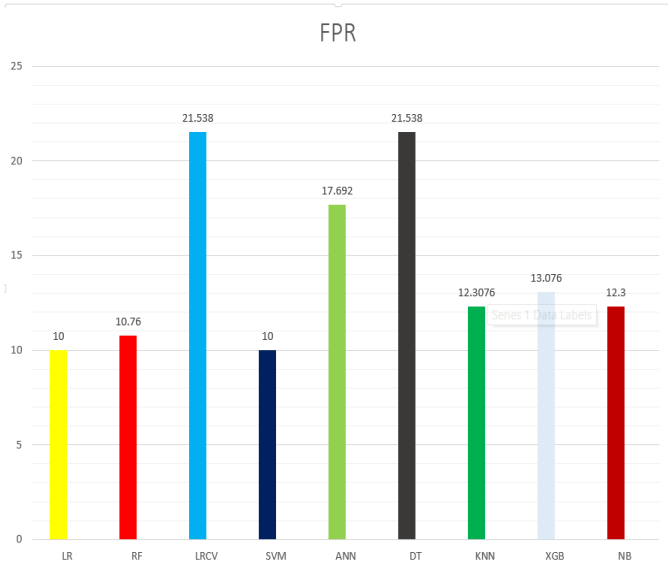

Figure 5: False Positive Rate(FPR)

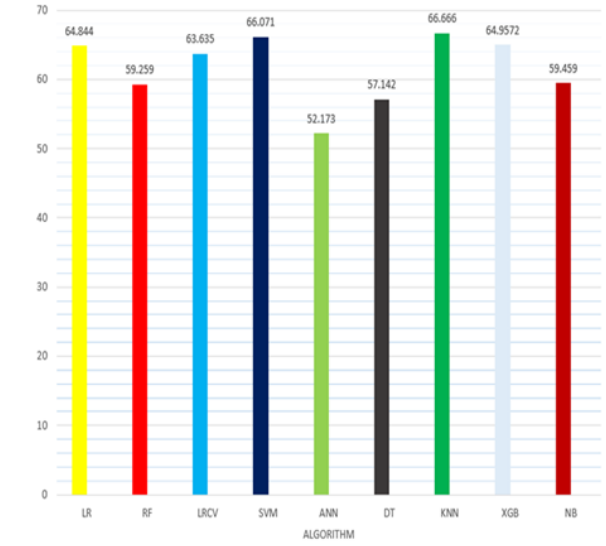

Figure 6: F1-Score

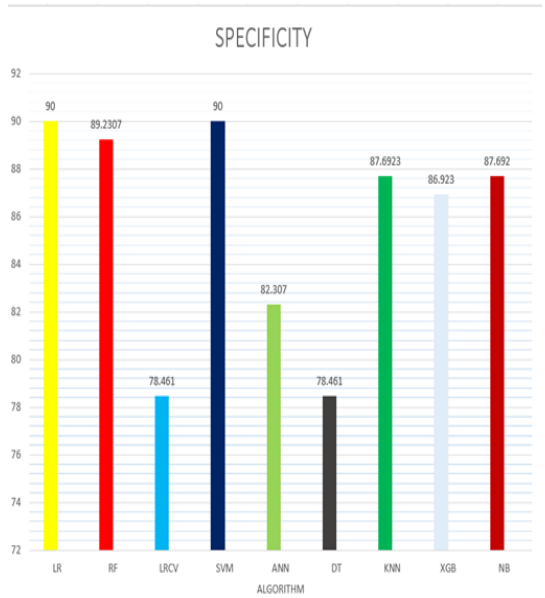

Figure 7: Specificity

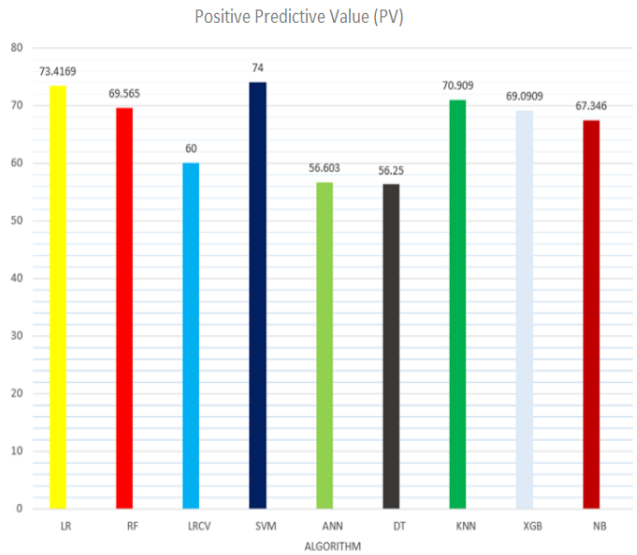

Figure 8: Positive Predictive Value (PV)

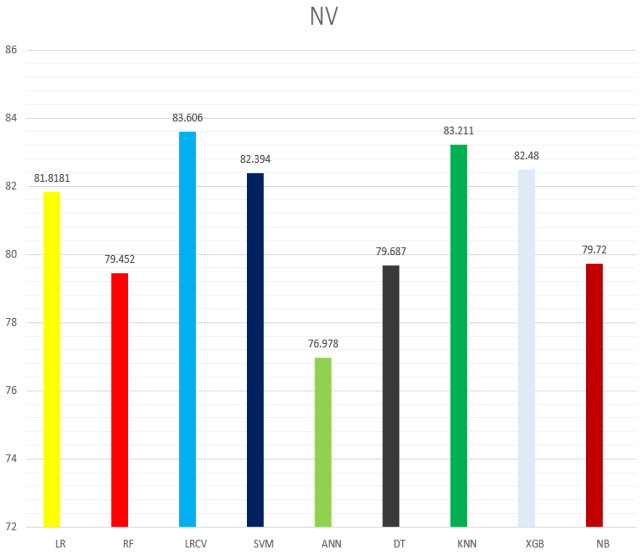

Figure 9: Negative Predictive Value(NV)

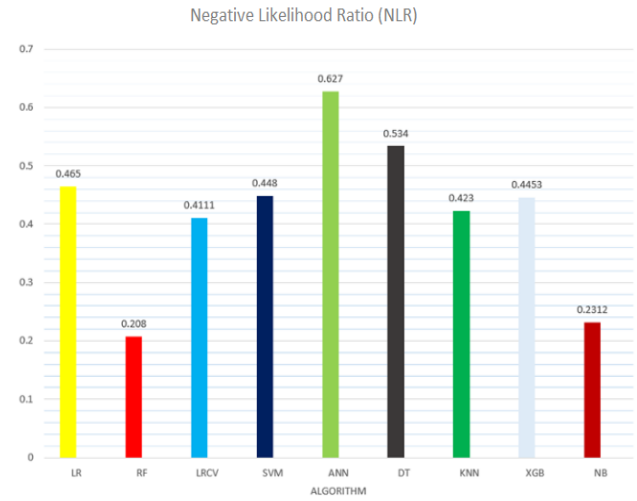

Figure 10: Negative Likelihood Ratio(NLR)

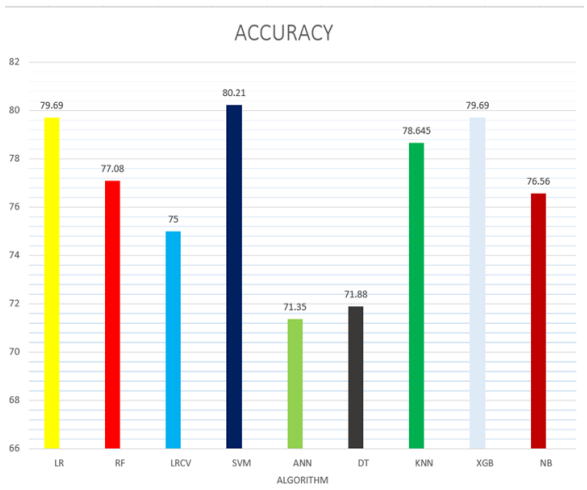

Figure 11: Accuracy

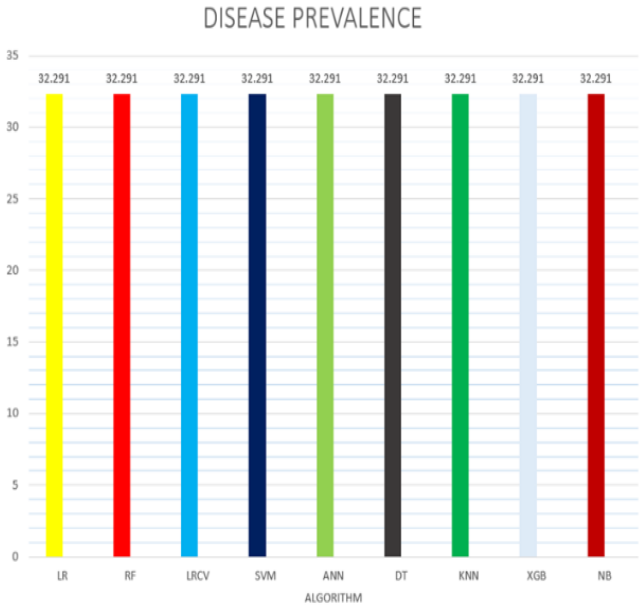

Figure 12: Disease Prevalence

B) Correctly and incorrectly classified instances by machine learning algorithm:

Total No of instances	Classification Algorithms	Correctly Classified Instances	Incorrectly Classified Instances
	Logical Regression	153	39
	Random forest	148	44
	Logical Regression CV	144	48
	SVM	154	38
768	ANN	137	55
	DT	138	54
	KNN	151	41
	XGB	153	39
	NB	153	39

V. CONCLUSION

Diabetes is known as one of the critical and chronic diseases which causes an increase in blood sugar. Diabetes if not diagnosed on time can increase the risk of cardiac stroke, diabetic nephropathy, brokenness and causes coronary failure, blindness, urinary organ diseases. Therefore, the detection of diabetes at its early stage is necessary. Use of Predictive analytics in the healthcare system can change the way how medical researchers and practitioners gain insights from medical data and take decisions. In this paper, we used nine different machine learning algorithms Logical regression, Random forest, Logical regression CV, Support Vector Machine, Artificial Neural Network (ANN), Decision Tree, k-nearest neighbors (KNN), XGB Classifier, Naive Bayesian for predictive analytics. From experimental result it is seen that Support Vector Machine (SVM) has highest accuracy of 80.21%, followed by XGB classifier and logical regression having accuracy of 79.69%.SVM has highest precision of 74%. Logical regression CV has highest recall of 67.741%. Logical regression CV has highest True positive rate of 67.741%. Logical regression CV and Decision tress have highest false positive rate of 21.538%.KNN has highest F1-Score of 66.666%. Logical regression and SVM have highest specificity of 90%.SVM has highest positive predictive value of 74%. Logical regression CV has highest negative predictive value of 83.606%.SVM has highest positive likelihood ratio of 5.967%.ANN has highest negative likelihood ratio of 0.627%. Disease prevalence is 32.291%. Predication and diagnosis of diabetics is possible with the help of the proposed model.

VI. FUTURE SCOPE:

Proposed model can help early prediction of diabetes. By doing so save a lot of lives. Working on some more attributes so to tackle diabetes effectively. Improvising the algorithm to enhance the efficiency and working of the system.

REFERENCES

1. Priyanka Sonar,Prof. K. JayaMalini, "DIABETES PREDICTION USING DIFFERENT MACHINE LEARNING APPROACHES," Proceedings of the Third International Conference on Computing Methodologies and Communication (ICCMC 2019).
2. Deepti Sisodiaa , Dilip Singh Sisodia, "Prediction of Diabetes using Classification Algorithms,"International Conference on Computational Intelligence and Data Science (ICCIDS 2018).
3. Muhammad Azeem Sarwar, Nasir Kamal, Wajeeha Hamid, Munam Ali Shah, "Prediction of Diabetes Using Machine Learning Algorithms in Healthcare," Proceedings of the 24th International Conference on Automation & Computing, Newcastle University, Newcastle upon Tyne, UK, 6-7 September 2018.
4. Deeraj Shetty Kishor Rit Sohail Shaikh Nikita Patil, " Diabetes Disease Prediction Using Data Mining,"2017 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS).

5. Zhilbert Tafa and Nerxhivan Pervetica, “An Intelligent System for Diabetes Prediction,”4th Mediterranean Conference on Embedded Computing MECO – 2015 Budva, Montenegro.
6. P. Suresh Kumar and V. Umatejaswi, “ Diagnosing Diabetes using Data Mining Techniques,”International Journal of Scientific and Research Publications, June 2017.
7. Sadegh Bafandeh Imandoust And Mohammad Bolandraftar, “Application of K-Nearest Neighbor (KNN) Approach for Predicting Economic Events: Theoretical Background,”S B Imandoust et al. Int. Journal of Engineering Research and Applications Vol. 3, Issue 5, Sep-Oct 2013.
8. M Panda and M Patra, “Network Intrusion Detection Using Naïve Bayes,” at IJCSNS International Journal of Computer Science and Network Security, VOL7, 2007.
9. Berina Alic, Lejla Gurbeta and Almir Badnjevic, “Machine Learning Techniques for Classification of Diabetes and Cardiovascular Diseases,” 2017 6th Mediterranean Conference on Embeded Computing (MECO), 11-15 JUNE 2017, BAR, MONTENEGRO.
10. K Beyer, J Goldstein,R Ramakrishnan and U Shaft, “When is ‘Nearest neighbor’ Meaningful?,” 2014.
11. Milan Kumari, Sunila Godara, “Comparative Study of Data Mining Classification Methods in cardiovascular Disease Prediction, ” IJCST, Vol. 2, Issue 2, 2011, pp. 304-308
12. Vinod Chandra S.S,Anand Hareendran S, “Artificial intelligence and machine learning, ” Private Limited ,Delhi 110092,2014.
13. Sumi Alice Saji and Balachandran K, “Performance Analysis of Training Algorithms in Diabetes Prediction,” International Conference on Advances in Computer Engineering and Applications (ICACEA) IMS Engineering College, Ghaziabad, India 2015.

AUTHORS PROFILE

Mr. Shashank S. Joshi is currently pursuing his B.E. from Dept. Computer Engineering at VIT, Pune, India. His areas of interest are machine learning, Artificial Intelligence and data Analytics.

Mr. Vijayendra S. Gaikwad is currently pursuing his Ph.D. from Dept. Computer Science & Engineering at SGBAU, Amravati, India. He has completed his B.E. in Information Technology from Savitribai Phule Pune University, Pune, India in 2014 and his M.E. in Computer Science and Engineering from Dr. Babasaheb Ambedkar Marathwada University, Aurangabad, India in 2016. His areas of research interest are privacy preserved data mining, information security and privacy in cloud data processing and machine learning.

Mr. Sairam Rathod is currently pursuing his B.Tech. from Dept. Computer Engineering at VIT,Pune, India. His areas of interest are machine learning, Artificial Intelligence and data analytics.

Ms. Anamika Rathod is currently pursuing her B.Tech from Dept. Computer Engineering at VIT,Pune, India. Her areas of interest are machine learning, Artificial Intelligence and data analytics.

Ms. Neha Sagar is currently pursuing her B.Tech from Dept. Computer Engineering at VIT,Pune, India. Her areas of interest are machine learning, Artificial Intelligence and data analytics.