IJARCCE

ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 12, December 2021

DOI: 10.17148/IJARCCE.2021.101209

Predictive Analysis for the Detection of Diabetes Mellitus (DM) based on Machine Learning Classification Algorithm

Dillip Narayan Sahu[1], Vijay Pal Singh[2*]

[1]Lecturer, Department of MCA, School of Computer Science, Gangadhar Meher University (GMU), Odisha, India

[2]Associate Professor, Department of Computer Science, OPJS University, Rajasthan, India

*Corresponding Author: Dr. Vijay Pal Singh

Abstract: According to the World Health Organization, around 1.5 million people worldwide died due to diabetes in 2019. It is estimated that approximately 462 million people live with diabetes around the globe. According to other sources, about 432 million people worldwide have diabetes, the bulk living in low-and middle-income countries, and 1.5 million deaths are directly related to the disease diabetes annually. The amount of cases, morbidity and mortality rates in a specific time period or over time to time, the diabetes is steadily increasing over the past few decades. No doubt, Diabetes mellitus is a leading cause of deaths world wide and reduced life expectancy. This disease can be curable with early diagnosis and proper treatment. The purpose of this paper is to establish some predictive models using Machine Learning algorithms by taking a real time Diabetes mellitus dataset. In this paper, we have shown some real-time experiments and observations with the help of some Machine Learning algorithms, and also shown a clear picture on the predictive analysis for the detection of the disease Diabetes mellitus in medical science using Machine Learning algorithms using which patients may get accurate data so as to diagnose better for their early treatment.

Keywords: Algorithm, Classifier, Diabetes Mellitus, Machine Learning, Prediction.

I.INTRODUCTION

Machine learning tools are being extensively utilized in all scientific, medical fields and are liable for revolutionizing businesses everywhere. Healthcare systems, on the opposite hand, are very slow in adopting these advancements and are lagging far behind in this[1][2].

Machine learning are often useful within the management of chronic diseases, namely, diabetes[3][4]. In fact, Machine learning is already getting used to predict risk of diabetes supported genomic data, diagnosis of diabetes supported EHR data, to predict risk of complications. Adoption of Machine Learning technologies can significantly increase detection and early treatment of diabetic complications of patients[5][6].

II.EXPERIMENTS AND OBSERVATIONS

Here we have taken Diabetes dataset and Weka knowledge analysis tool to classify and predict the disease. The figure 1 and figure 2 shows the diabetes dataset and data preprocess respectively.

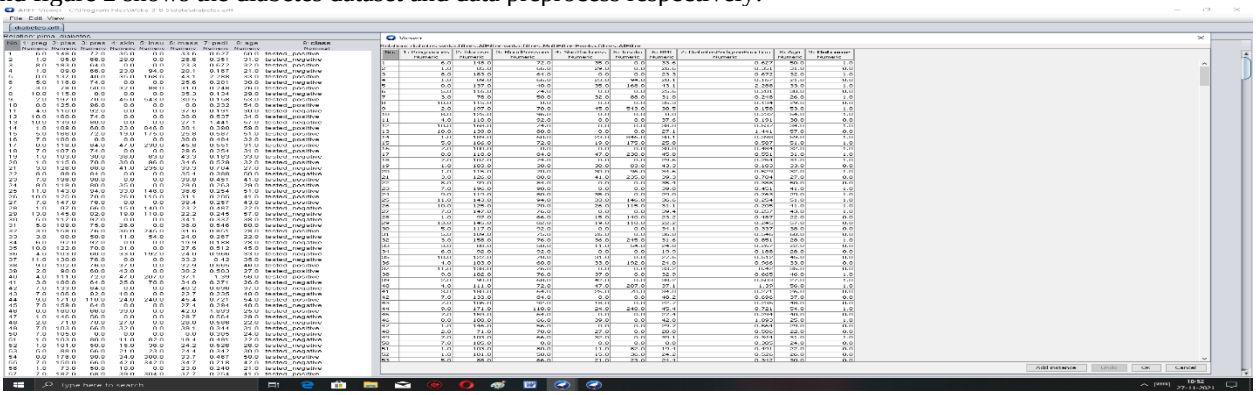

Fig.1. Diabetes Mellitus (DM) Dataset

IJARCCE

ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 12, December 2021

DOI: 10.17148/IJARCCE.2021.101209

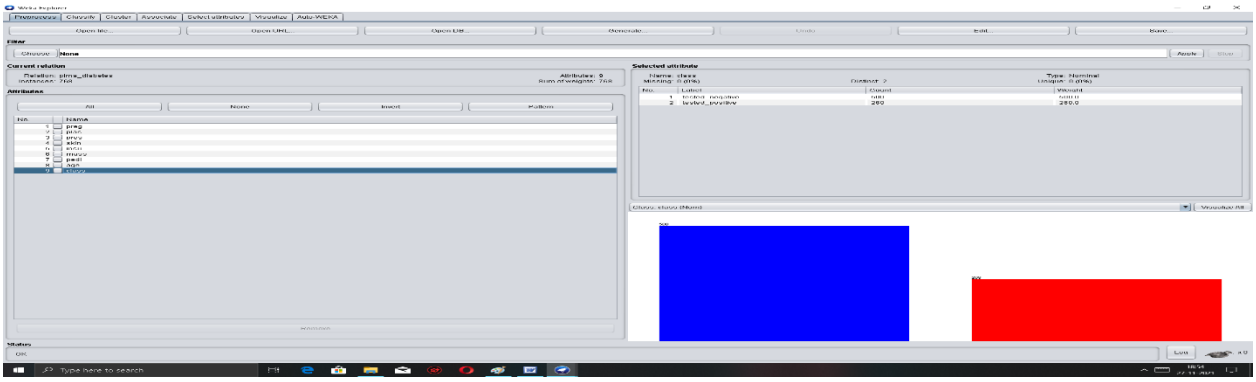

Fig.2. Data Preprocess

Experiments and Observations-1

Classifier Output=== Run information ===
Scheme: weka.classifiers.rules.ZeroR
Relation: pima_diabetes
Instances: 768 Attributes: 9
Preg plas pres skin insu mass pedi age class
Test mode: 10-fold cross-validation
=== Classifier model (full training set) ===
ZeroR predicts class value: tested_negative
Time taken to build model: 0 seconds
=== Stratified cross-validation ====== Summary ===
Correctly Classified Instances 500 65.1042 %
Incorrectly Classified Instances 268 34.8958 %
Kappa statistic 0
Mean absolute error 0.4545
Root mean squared error 0.4766
Relative absolute error 100 %
Root relative squared error 100 %
Total Number of Instances 768
=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	1.000	1.000	0.651	1.000	0.789	?	0.497	0.650	tested_negative
	0.000	0.000	?	0.000	?	?	0.497	0.348	tested_positive
Weighted Avg.	0.651	0.651	?	0.651	?	?	0.497	0.544	

=== Confusion Matrix ===

a	b	<-- classified as
500	0	a = tested_negative
268	0	b = tested_positive

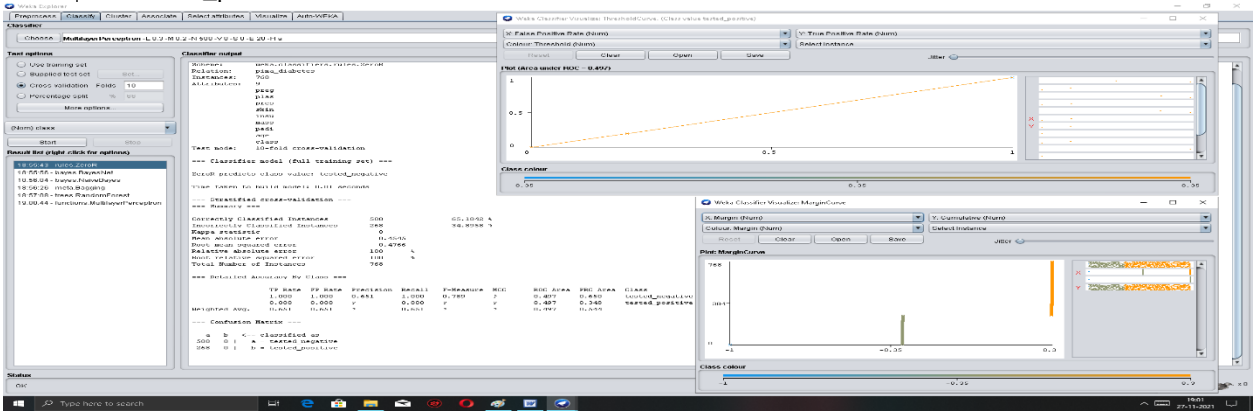

Fig.3. ZeroR Classifier with Visualize curve

IJARCCE

ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 12, December 2021

DOI: 10.17148/IJARCCE.2021.101209

Experiments and Observations-2

Classifier Output=== Run information ===
Scheme: weka.classifiers.bayes.BayesNet -D -Q weka.classifiers.bayes.net.search.local.K2 -- -P 1 -S BAYES -E weka.classifiers.bayes.net.estimate.SimpleEstimator -- -A 0.5
Relation: pima_diabetes Instances: 768 Attributes: 9
Test mode: 10-fold cross-validation
=== Summary ===

Correctly Classified Instances	571	74.349 %
Incorrectly Classified Instances	197	25.651 %
Kappa statistic	0.429	
Mean absolute error	0.2987	
Root mean squared error	0.4208	
Relative absolute error	65.7116 %	
Root relative squared error	88.28 %	
Total Number of Instances	768	

=== Confusion Matrix ===

a	b	<-- classified as
408	92	a = tested_negative
105	163	b = tested_positive

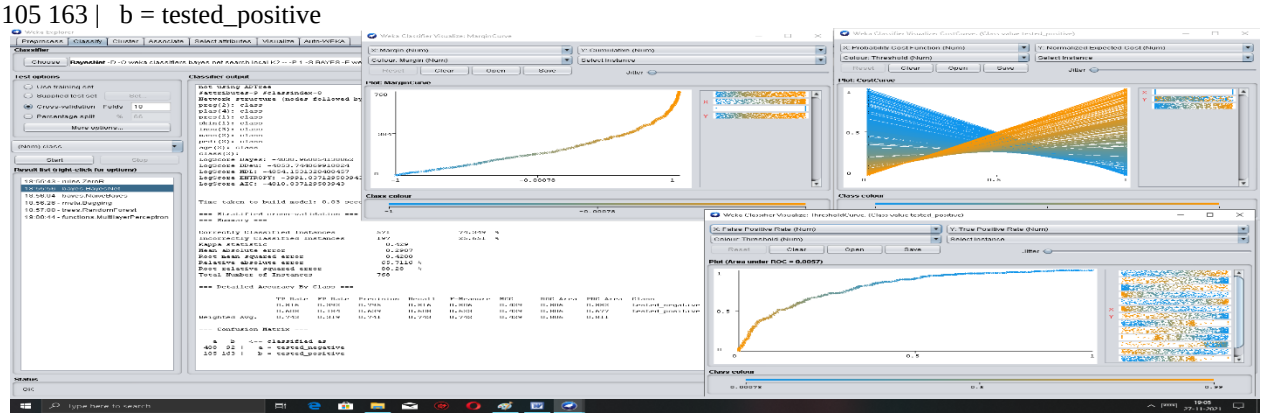

Fig.4. BayesNet Classifier with Visualize different cases

Experiments and Observations-3

Classifier Output === Run information ===
Scheme: weka.classifiers.bayes.NaiveBayesMultinomial
=== Classifier model (full training set) ===
The independent probability of a class
tested_negative 0.65 tested_positive 0.35
=== Summary ===

Correctly Classified Instances	460	59.8958 %
Incorrectly Classified Instances	308	40.1042 %
Kappa statistic	0.1279	
Mean absolute error	0.4009	
Root mean squared error	0.6152	
Relative absolute error	88.2029 %	
Root relative squared error	129.0792 %	
Total Number of Instances	768	

=== Confusion Matrix ===

a	b	<-- classified as
339	161	a = tested_negative
147	121	b = tested_positive

IJARCCE

ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 12, December 2021

DOI: 10.17148/IJARCCE.2021.101209

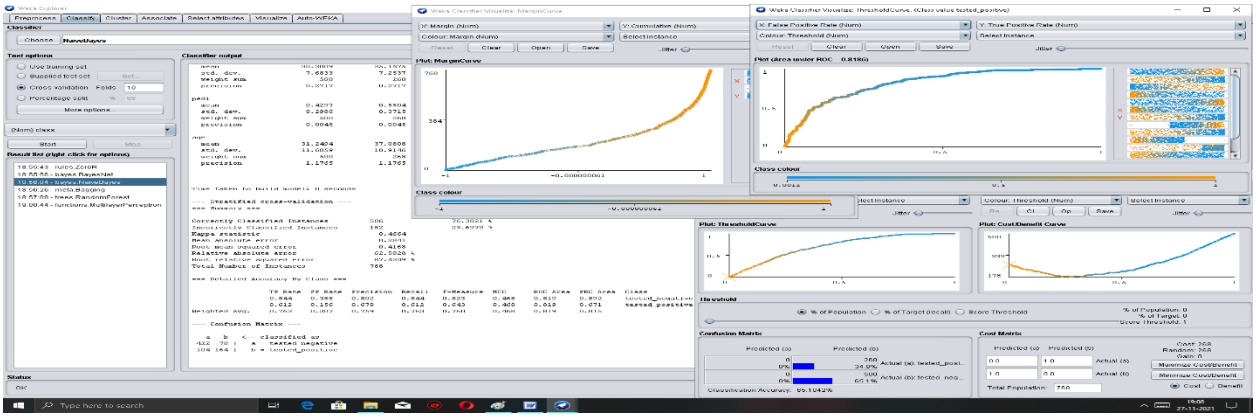

Fig.4. NaiveBayes Classifier with Visualize different cases

Experiments and Observations-4

Classifier Output === Run information ===

Scheme: weka.classifiers.meta.Bagging -P 100 -S 1 -num-slots 1 -I 10 -W weka.classifiers.trees.REPTree -- -M 2 -V 0.001 -N 3 -S 1 -L -1 -I 0.0

Bagging with 10 iterations and base learner

weka.classifiers.trees.REPTree -M 2 -V 0.001 -N 3 -S 1 -L -1 -I 0.0

=== Summary ===

Correctly Classified Instances	582	75.7813 %
Incorrectly Classified Instances	186	24.2188 %
Kappa statistic	0.4498	
Mean absolute error	0.315	
Root mean squared error	0.4063	
Relative absolute error	69.3049 %	
Root relative squared error	85.2474 %	
Total Number of Instances	768	

=== Confusion Matrix ===

```
   a   b   <-- classified as
 425  75 |   a = tested_negative
 111 157 |   b = tested_positive
```

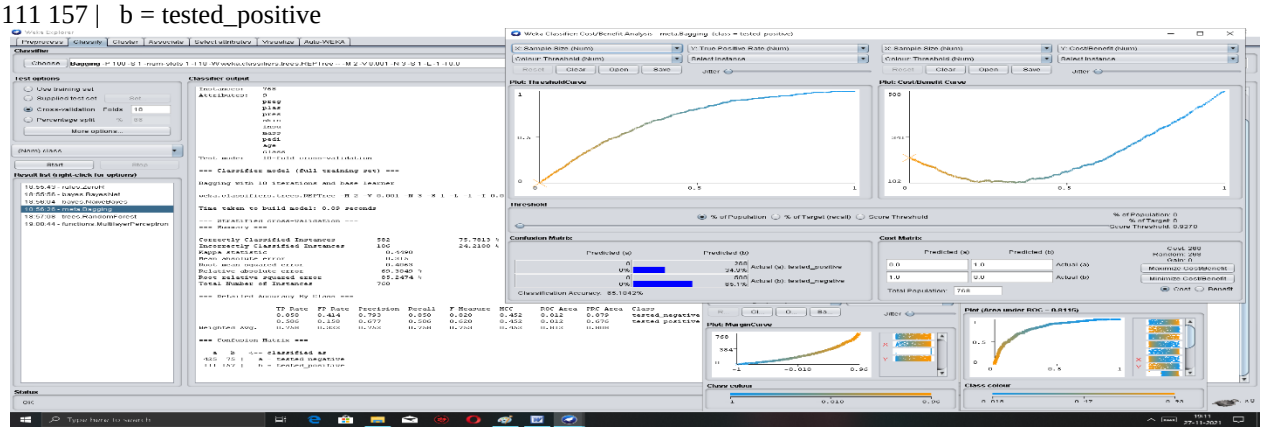

Fig.5. Bagging Classifier with Visualize different cases

Experiments and Observations-5

Classifier Output === Run information ===

Scheme: weka.classifiers.trees.RandomForest -P 100 -I 100 -num-slots 1 -K 0 -M 1.0 -V 0.001 -S 1 RandomForest weka.classifiers.trees.RandomTree -K 0 -M 1.0 -V 0.001 -S 1 -do-not-check-capabilities

=== Summary ===

Correctly Classified Instances	582	75.7813 %
Incorrectly Classified Instances	186	24.2188 %
Kappa statistic	0.4566	
Mean absolute error	0.3106	

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 12, December 2021

DOI: 10.17148/IJARCCE.2021.101209

Root mean squared error 0.4031
Relative absolute error 68.3405 %
Root relative squared error 84.5604 %
Total Number of Instances 768
=== Confusion Matrix ===
a b <-- classified as
418 82 | a = tested_negative
104 164 | b = tested_positive

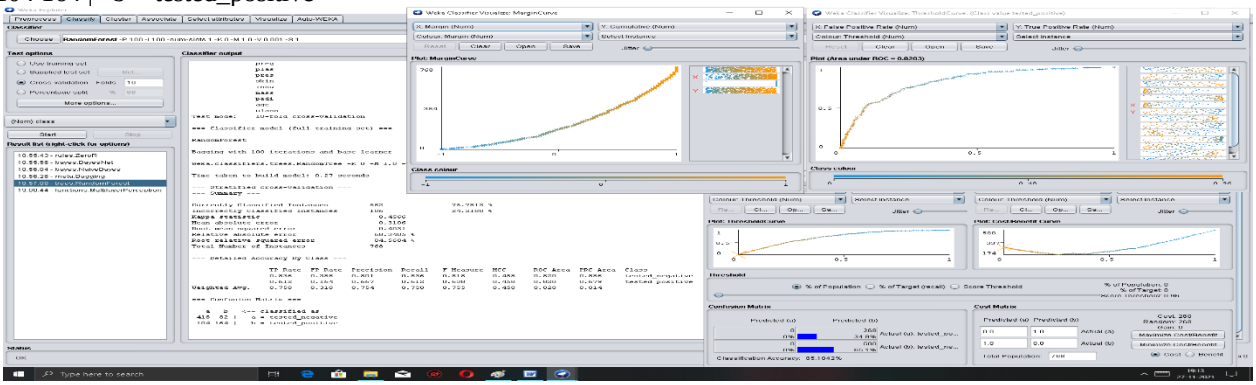

Fig.6. RandomForest Classifier with Visualize different cases

III.DISCUSSION

We have taken 5 different experimental observations using the machine learning tool to clearly analyze, detect and predict for the Diabetes Disease. In the study of the above experimental observations, it is found that, machine learning tools are no doubt an excellent way to predict and detect the Diabetes disease at an early stage prior to the satisfiability of the conditions of the early stage patient. It is found that the accuracy level using different algorithm in Machine Learning is an excellent option for detection and prediction of Diabetes disease, having good accuracy rate and so will be efficient and acceptable.

IV.CONCLUSION

In 2019, diabetes was the ninth leading cause of death with an estimated 1.5 million deaths directly caused by diabetes. Diabetes mellitus is a disease, which can cause many complications. How to exactly predict and diagnose this disease by using machine learning is worthy studying. According to the all above experiments, we found good accuracy using Random Forest and Bagging classifier so that will be acceptable.

REFERENCES

[1] Alghamdi, M., Al-Mallah, M., Keteyian, S., Brawner, C., Ehrman, J., and Sakr, S. (2017). Predicting diabetes mellitus using SMOTE and ensemble machine learning approach: the henry ford exercise testing (FIT) project. PLoS One 12:e0179805. doi: 10.1371/journal.pone.0179805
[2] American Diabetes Association (2012). Diagnosis and classification of diabetes mellitus. Diabetes Care 35(Suppl. 1), S64–S71. doi: 10.2337/dc12-s064
[3] Bengio, Y., and Grandvalet, Y. (2005). Bias in Estimating the Variance of K -Fold Cross-Validation. New York, NY: Springer, 75–95. doi: 10.1007/0-387-24555-3_5
[4] Breiman, L. (2001). Random forest. Mach. Learn. 45, 5–32. doi: 10.1023/A:1010933404324
[5] Chen, X. X., Tang, H., Li, W. C., Wu, H., Chen, W., Ding, H., et al. (2016). Identification of bacterial cell wall lyases via pseudo amino acid composition. Biomed. Res. Int. 2016:1654623. doi: 10.1155/2016/1654623
[6] Cox, M. E., and Edelman, D. (2009). Tests for screening and diagnosis of type 2 diabetes. Clin. Diabetes 27, 132–138. doi: 10.2337/diaclin.27.4.132
[7] Duygu,ç., and Esin, D. (2011). An automatic diabetes diagnosis system based on LDA-wavelet support vector machine classifier. Expert Syst. Appl. 38, 8311–8315.
[8] Friedl, M. A., and Brodley, C. E. (1997). Decision tree classification of land cover from remotely sensed data. Remote Sens. Environ. 61, 399–409.
[9] Georga, E. I., Protopappas, V. C., Ardigo, D., Marina, M., Zavaroni, I., Polyzos, D., et al. (2013). Multivariate prediction of subcutaneous glucose concentration in type 1 diabetes patients based on support vector regression. IEEE J. Biomed. Health Inform. 17, 71–81. doi: 10.1109/TITB.2012.2219876
[10]Habibi, S., Ahmadi, M., and Alizadeh, S. (2015). Type 2 diabetes mellitus screening and risk factors using decision tree: results of data mining. Glob. J. Health Sci. 7, 304–310. doi: 10.5539/gjhs.v7n5p304
[11]Han, L., Luo, S., Yu, J., Pan, L., and Chen, S. (2015). Rule extraction from support vector machines using ensemble learning approach: an application for diagnosis of diabetes. IEEE J. Biomed. Health Inform. 19, 728–734. doi: 10.1109/JBHI.2014.2325615
[12]Iancu, I., Mota, M., and Iancu, E. (2008). "Method for the analysing of blood glucose dynamics in diabetes mellitus patients," in Proceedings of the 2008 IEEE International Conference on Automation, Quality and Testing, Robotics, Cluj-Napoca. doi: 10.1109/AQTR.2008.4588883

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 12, December 2021

DOI: 10.17148/IJARCCE.2021.101209

[13]Jackson, D. A. (1993). Stopping rules in principal components analysis: a comparison of heuristical and statistical approaches. Ecology 74, 2204–2214. doi: 10.2307/1939574

[14]Jegan, C. (2014). Classification of diabetes disease using support vector machine. Microcomput. Dev. 3, 1797–1801.

[15]Jia, C., Zuo, Y., and Zou, Q. (2018). O-GlcNAcPRED-II: an integrated classification algorithm for identifying O-GlcNAcylation sites based on fuzzy undersampling and a K-means PCA oversampling technique. Bioinformatics 34, 2029–2036. doi: 10.1093/bioinformatics/bty039

[16]Jiang, Y., and Zhou, Z. H. (2004). Editing training data for kNN classifiers with neural network ensemble. Lect. Notes Comput. Sci. 3173, 356–361. doi: 10.1007/978-3-540-28647-9_60

[17]Jolliffe, I. T. (1998). "Principal components analysis," in Proceedings of the International Conference on Document Analysis and Recognition (Heidelberg: Springer).

