International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-9 Issue-3, February 2020

Future Prediction of Diabetics using XG Booster Classifiers

Iyapparaja M, Manivannan S.S, Vinoth Kumar M, Thanapal P, Kamalakannan J

Abstract: *Diabetes is a most common disease that occurs to most of the humans now a day. The predictions for this disease are proposed through machine learning techniques. Through this method the risk factors of this disease are identified and can be prevented from increasing. Early prediction in such disease can be controlled and save human's life. For the early predictions of this disease we collect data set having 8 attributes diabetic of 200 patients. The patients' sugar level in the body is tested by the features of patient's glucose content in the body and according to the age. The main Machine learning algorithms are Support vector machine (SVM), naive bayes (NB), K nearest neighbor (KNN) and Decision Tree (DT). In the exiting the Naive Bayes the accuracy levels are 66% but in the Decision tree the accuracy levels are 70 to 71%. The accuracy levels of the patients are not proper in range. But in XG boost classifiers even after the Naïve Bayes 74 Percentage and in Decision tree the accuracy levels are 89 to 90%. In the proposed system the accuracy ranges are shown properly and this is only used mostly. A dataset of 729 patients can be stored in Mongo DB and in that 129 patients repots are taken for the prediction purpose and the remaining are used for training. The training datasets are used for the prediction purposes.*

Keywords : diabetes mellitus, machine learning, Decision Tree prediction.

I. INTRODUCTION

Earlier it is only affected for aged persons but coming to the earlier generations it is even affected to kids and young generation. The disease is most harmful and it is difficult for young generation and kids to have balance diet and to control from increasing. The Diabetes for all is caused mainly by the intake of food. Food plays a major role for a Diabetic patient. Proper intake of food helps them prevent from the increase of the disease. The disease also affects the human body which thus harm a large number of body's system. It is affected particularly in blood veins and nerves. The disease affects the hormone insulin and improve the sugar levels in the blood. The diabetes is mostly said that it comes hereditary vise but now a day's hormonal impacts also due to food intake can also be caused. The existing technique it was difficult to find the exact accuracy of the sugar levels but the proposed technique is far better and the accuracy levels are correct.

More of sugar content food should be avoided to reduce the insulin level in the blood. Diabetics are a most harmful disease which makes us avoid most of the food products such as chocolate, sweets, and fruits likewise. The Instance based learning is done which improves the performance of several supervised learning algorithms [1]. These includes algorithms that learn Decision tree, classification rules and distributed networks. We describe however storage needs are often keep are often mechanically reduced with, at most, minor sacrifices in learning rate and classification accuracy. Diabetic patients are mostly powerless and, in this way, long term complications effects of cardio-vascular disease are the leading cause of death. Early predictions can control the increase of disease and can prevent the growing insulin levels in the blood cells. We have taken the original patients medical documents based on risk factors using popular machine learning classification to examine the performance for predicting diabetes mellitus[2].

Untreated high glucose from polygenic disorder will harm your nerves, eyes, kidneys and other organs. The general symptoms of diabetics are increased hunger, weight loss and increased thirst. Diabetic's symptoms can be so soft that they are hard to find. Which results from endocrine deficiency or resistance resulting in high blood sugar, conjointly known as glucose[3].

II. RELATED WORK

The diabetic is analyzed in 2002 using some Canadian population records. They analyzed particular aged people based on some attributes such as age, sex, body mass index, blood pressure etc. Diabetic level will be calculated by using FDRSM (Framingham diabetes rick scoring) model technique. It is mainly used for finding the risk score of the patients [6]. Diabetes is most common growing disease so this will be control and prevent by using some technique like data mining technique they all are trying to prevent the disease in starting stage itself by predicting based on the symptoms of the diabetes. In this paper they using some SVM and K-Nearest Neighbors algorithm to analyze the patient's disease level They used some datasets and these techniques for predicting the diabetes level but it doesn't give that such of results [14]. Then they applied some machine learning algorithm on the features that will be sensitivity and specificity model against some algorithms. Applying proposed framework that achieved high performance of the disease. It will be providing the 71 percentage of accuracy only by using this technique
[12].

Revised Manuscript Received on February 22, 2020.
* Correspondence Author

Iyapparaja M, Associate Professor, SITE, Vellore Institute of Technology, Vellore, Tamilnadu, INDIA.
Manivannan S.S, Associate Professor, SITE, Vellore Institute of Technology, Vellore, Tamilnadu, INDIA.
Vinoth Kumar M, Associate professor at the Faculty of Information Science and Engineering, Dayananda sagar Academy of Technology and Management, Bangalore, Karnataka, INDIA.
Iyapparaja.m@vit.ac.in
Thanapal P, Associate Professor, SITE, Vellore Institute of Technology, Vellore, Tamilnadu, INDIA.
Kamalakannan J, Associate Professor, SITE, Vellore Institute of Technology, Vellore, Tamilnadu, INDIA.

In generally they are characterized the data by using the supervised learning and unsupervised learning algorithm but the SVM is most successfully used algorithm based on the collective clinical datasets. In proposed they will be predicting by using the datasets only but it was also not providing accurate results [11].

III. EXISTING SYSTEM

In existing system they analyze the patient's details using the dataset. They used the SVM (Support vector Machine) algorithm but they didn't get this much accuracy about the diabetes Accuracy is very low and not effective also.

IV. PROPOSED SYSTEM

In proposed system they collecting the data and that will be stored in the mongo DB then they using two algorithms for checking which one is providing exact accuracy. First we implementing the naïve bayes algorithms it will providing 79 percentage only and then implementing the decision tree algorithm splitting the data's into trained and testing. Applying the algorithm on the testing data it will giving the 80 percent accuracy then trying on the trained data that result should be in 100 percent accuracy so this decision tree algorithm is providing the exact accuracy about diabetes[4]. Using this technique, it will be providing more accuracy prediction level of diabetes and more effective also. They will be using the gradient boosting for boosting all algorithms and this classification technique is more relevant.

V. MODULES

A. DATA COLLECTION

Diabetes data's will be collected and stored in mongo DB. The data's will be like the patients age, sex, blood pressure etc. That will be the raw data and it will be cleaning and preprocessing by using some technique then it will be stored in mongo DB for predicting the diabetes level of patients[5].

B.DATA PROCESSING

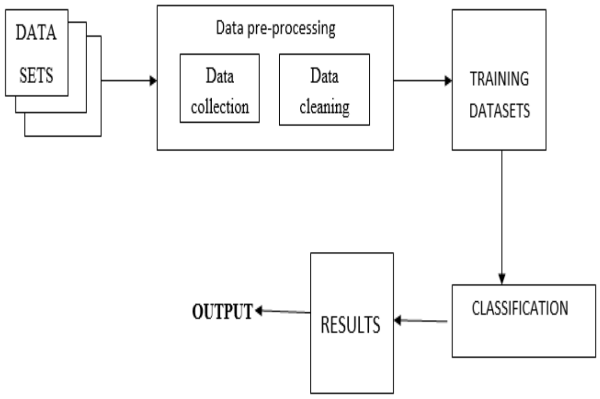


Fig.1: Data Processing

C. DATA CLEANING AND PROCESSING

Data cleaning is the process of ensuring that your data is correct, consistent and useable by identifying any errors and corruptions in the data and deleting them. We convert the numeric data values into nominal. It is a time consuming process. We will remove the null values from the original datasets and make it as a useful data. The values are collected from various datasets and the null values have been placed which has to be converted.

D. PROPOSED SYSTEM ARCHITECTURE

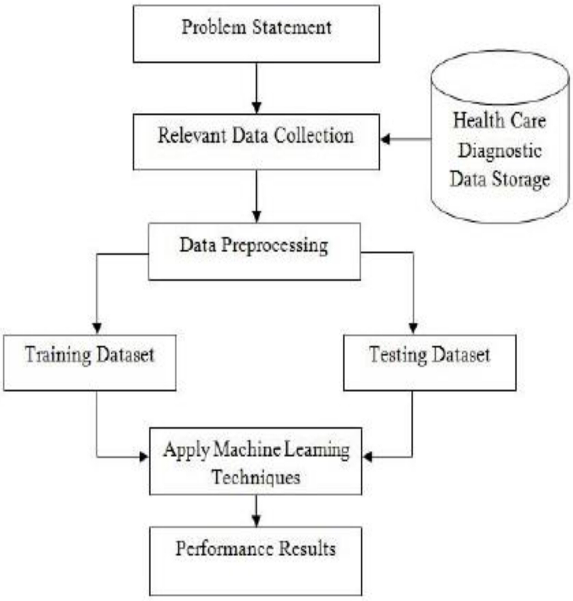


Fig.2 System Architecture

E. DATA ANALYSIS ANDS TRANSFORMING

The data are taken in the form of pie charts, bar diagrams and graphs. The data are transformed in the statistical logistics form and the meaningful data is taken from the graph. This type of analyzing data is more useful to check the drastic change in the growth of the data[6]. The patients report is tested to check the level of sugar content in the blood to take the overall growth percentage.

F. DATA SPLITTING AND TRAINING

The datasets will be collected and storing in the mongo DB. Then they using the gradient boosting for boosting the algorithm for predicting the level of diabetes based on some features. Data's will be classified using the decision tree algorithm. We applied the naive bayes algorithm for predicting the level of diabetes Naïve bayes will be applied on the features of data sets were they collect for predicting and it will be storing in the mongo DB. These techniques will be mainly used for knowing about the level of the diabetes for getting the treatment for the diabetes Now a day's new born child also affected by the diabetes it was so harmful so that will be reduced by implementing these techniques.

G. TECHNIQURE PROCESSING

Data Collecting and storing in mongo DB then data sets will be classified by cleaning the data then preprocessing the data with the use of gradient boosting. Training will be done by using train, test and split method then we separated the data after completing the training process finally we applying the algorithm on the data. We using the anaconda navigator, jupyter and visual studio tools for running the project.

Done the graphs part using the anaconda software than we hosting the designing part in local host using visual studio. In that designing part you can entering the parameters which placing in the datasets then you can see the final result as 0 or 1 based on this 0 or 1 we can deciding the patients having the diabetes or not.

H. ALGORITHM PROCESS

Data predicting will be done by using two algorithms

Step 1:
First we used the naïve bayes algorithm automatically that will provide 68 percent accuracy only it will be providing but that will not enough for us for providing treatment for the patients so we boosting that algorithm using gradient boosting then the naïve bayes algorithm will provide 79 percent accuracy only.

Step 2:
They used another algorithm is decision tree algorithm while using that algorithm first they got 74 percent accuracy then there boosting the algorithm using the gradient boosting finally they got 80 percent accuracy so comparing this both algorithm decision tree algorithm will be providing exact accuracy level

VI. RESULTS AND DISCUSSION

Based on dataset it predicts how many of them has diabetes or no (0 or 1). Note: Here 0 predicts who doesn't diabetes and 1 predicts who have diabetes

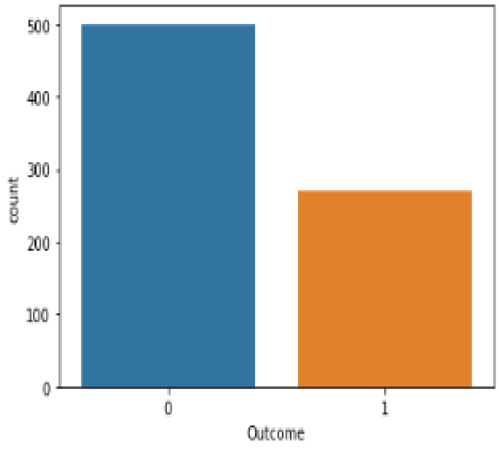


Fig.3: Outcome of Using Diabetes Datasets

It is a graphical representation estimate of probability distribution of continuous variable that divides the value with numerical that contains rows and columns.

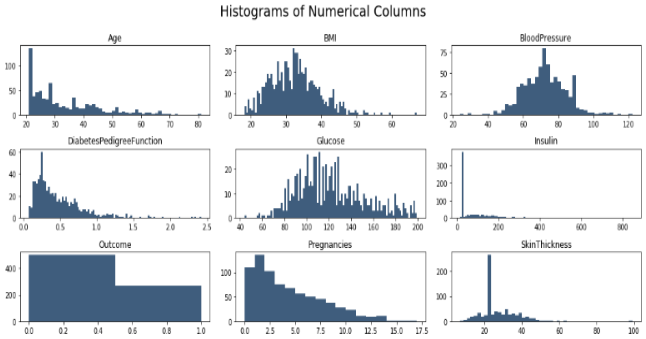


Fig.4 Histogram of Numeric Columns

This graphical plot is that based on age it predicts that how many amounts of ranges are in diabetes.

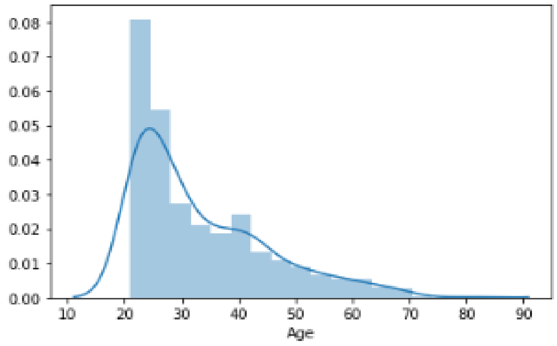
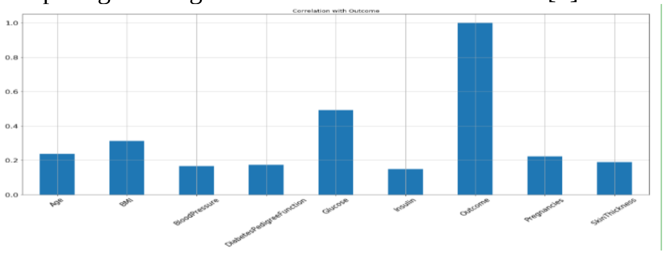
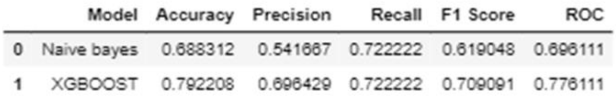


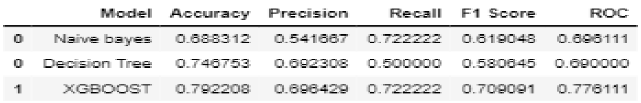
Fig.5: Outcome of Age

Correlation with outcome for all the features to extract the output and accuracy prediction for each algorithm and comparing which gives the best for diabetes dataset[7]





	Model	Accuracy	Precision	Recall	F1 Score	ROC
0	Naive bayes	0.688312	0.541667	0.722222	0.619048	0.696111
1	XGBOOST	0.792208	0.696429	0.722222	0.709091	0.776111



	Model	Accuracy	Precision	Recall	F1 Score	ROC
0	Naive bayes	0.688312	0.541667	0.722222	0.619048	0.696111
0	Decision Tree	0.746753	0.692308	0.500000	0.580645	0.690000
1	XGBOOST	0.792208	0.696429	0.722222	0.709091	0.776111

Fig.6: Correlation with Outcome and Predictions of Diabetes

It's a correlation plot which attribute gives maximum (darkness) as well as minimum (lightness) through the plotting.

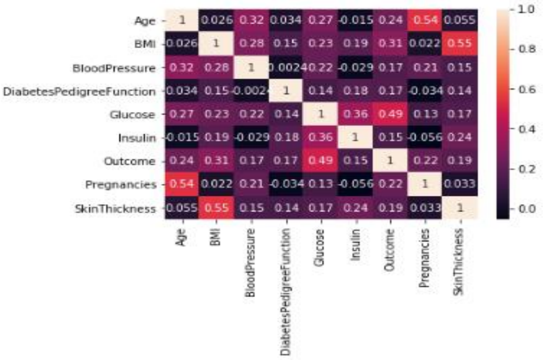


Fig.7: Correlation of Matrix

It's a confusion matrix through Gaussian NB classification report to predict accuracy.

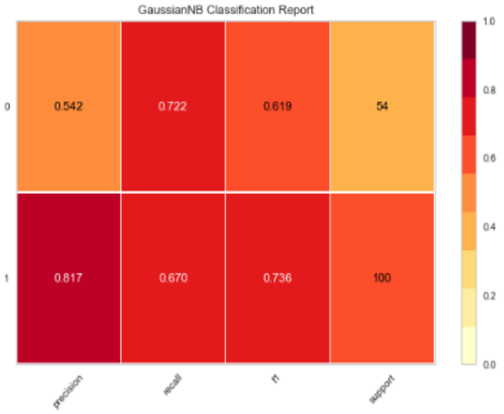
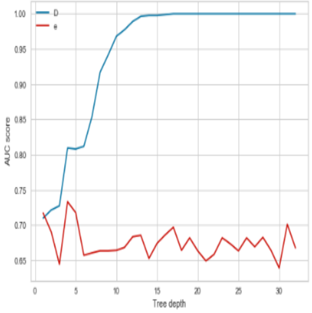


Fig.8: GaussianNB Classification Report

After Optimization

This is a statistical approach for decision tree algorithm that give from the range of 70% to 100% that each and every time the ratio range will keep on changing the value till it gets predicted high accuracy in the tree depth of 30th range and in tree depth of 10th it gives near to 80% [8].



```
[0.709929906542056, 0.7214602803738317, 0.7272196261682242, 0.8093107476635513, 0.8079088785046729, 0.8117757009345794, 0.85
39953271028038, 0.9170210280373832, 0.9417990654205609, 0.9679906542056075, 0.9772313084112151, 0.9889135514018693, 0.996250
0000000001, 0.9974999999999999, 0.9974999999999999, 0.99875, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.
0, 1.0, 1.0, 1.0]
```

Fig.9: Accuracy Prediction in Tree Depth of 30th

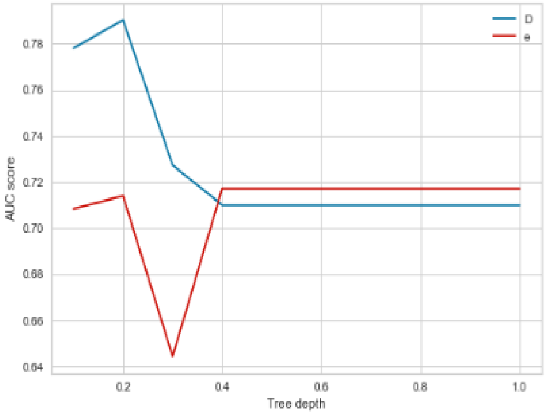


Fig.10: Accuracy Prediction in Tree Depth of 10th

VII. CONCLUSION

This concept is used to predict early predictions of diabetes by using the machine learning techniques. This is done by using various algorithms and the exact data is shown. Through this paper we can know the level of blood sugar level which helps lot of people to avoid from dangerous health problems and death. This helps to prevent from the spreading of the sugar in the blood cells. Most of the patients will be benefited to use the precautions and to avoid other problems in the body. This technique will 00.796be providing prediction of exact level of the diabetes. it is useful for providing the exact treatment for the concern patients. The trained dataset will provide 100 percent accuracy level in predicting first they didn’t get this much of accuracy so they optimizing the parameters values from the database using decision tree algorithm in 30 depth level then the trained data’s will provide the 100 percent accuracy in diabetes mellitus. Then the testing data providing the 80 percent accuracy level the output will be displayed as 0 or 1 based on this result only we deciding the patient having diabetes or not. If we got a output as 0 means the patient will not have diabetes and 1 means patient having the diabetes so they wants to get treatment from the doctor. It is useful to give treatment for concern person.

REFERENCES

1. Morteza, M., Franklyn, P., Bharat, S., Linying, D., Karim, K. and Aziz G. 2015. Evaluating the Performance of the Framingham Diabetes Risk Scoring Model in Canadian Electronic Medical Records. Canadian journal of diabetes 39, 30(April. 2015), 152-156.
2. Kavakiotis, Ioannis, Olga Tsave, AthanasiosSalifoglou, NicosMaglaveras, IoannisVlahavas, and IoannaChouvarda. "Machine learning and data mining methods in diabetes research." Computational and structural biotechnology journal (2017).
3. 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.
4. Ross Quinlan (1993). C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, San Mateo, CA.
5. Witten, I. H. et al. (1999). Weka: Practical machine learning tools and techniques with Java implementations.
6. 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.
7. 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.
8. C Gopalakrishnan, M Iyapparaja , Active contour with modified Otsu method for automatic detection of polycystic ovary syndrome from ultrasound image of ovary, Multimedia Tools and Applications, 2019.
9. M. Iyapparaja, P. Sivakumar. Metrics Based Evaluation for Disease Affection in Distinct Cities. Research J. Pharm. and Tech. 2017; 10(8): 2487-2491.
10. Iyapparaja M et.al. 2012 Coupling and Cohesion Metrics in Java for Adaptive Reusability Risk Reduction IET Chennai 3rd International Conference on Sustainable Energy and Intelligent Systems (SEISCON 2012),52-57.
11. Iyapparaja M, Tiwari. M, Security policy speculation of user uploaded images on content sharing sites , IOP Conf. Series: Materials Science and Engineering 263 (2017) 042018 doi:10.1088/1757-899X/263/4/042019,pp-1-8.

AUTHORS PROFILE

Dr. Iyapparaja Meenakshisundaram, He received Ph.D in Anna University, Chennai, BE degree from Anna University, Chennai and ME degree from Anna University of Technology, Coimbatore. Presently, He is an Associate Professor in School of Information Technology and Engineering, VIT, Vellore. He has 11 years of experience in Teaching and Big data, Software testing and software Engineering field. He received University Rank holder award for his ME degree. His research interests include Software Testing, Software Engineering, Big data, Networking and Agile Testing. He is life time member of ISTE

Dr.S.S.Manivannan, is currently working as Associate Professor in the School of Information Technology and Engineering in Vellore Institute of Technology (VIT), Vellore. He holds B.E Computer Science and M.E.Computer Science Engineering from University of Madras and College of Engineering (CEG) Anna University, Chennai. He has obtained his Ph.D in the area of Network and Information Security in the Information Technology domain in VIT, Vellore. He has more than 25 standard Scopus Indexed Publications and 5 referred Impact Factor Journals. He is the reviewer in many Scopus indexed and Science Citation Indexed Journals. He has also served as Technical Session Chair in many IEEE International Conferences.

Dr. M.Vinoth Kumar, obtained his Bachelor's degree in Computer Science and Engineering from Periyar University, Salem, Tamilnadu, India. He obtained his Master's degree in Computer Science and Engineering and PhD in Computer Science majoring in Agent Programming from Anna University, Chennai, Tamilnadu, India. Currently, he is an Associate professor at the Faculty of Information Science and Engineering, Dayananda sagar Academy of Technology and Management, Bangalore, Karnataka, India. His specializations includes Artificial Intelligence, Machine learning and Big Data Computing. His current research interests are convolutional neural network and medical image processing.

Dr. Thanapal P, received his B.E degree in Computer Science and Engineering from Madurai Kamaraj University, Madurai India in 1998, M.E degree in Computer Science and Engineering from Anna University Chennai, India in 2005 and Ph.D degree from Vellore Institute of Technology University Vellore, India. He has published more than 25 research paper in reputed international journals and conferences. His main research interests include cloud computing, mobile cloud computing, wireless network and IoT

Dr. Kamalakannan.J, He is a faculty member at School of Information Technology and Engineering, VIT University, Vellore, India. He has completed his Bachelors in Electronics and Communication Engineering and Masters in Computer Science and Engineering from Madras University, pursued Ph.D in computer science and engineering at School of Computing Science and Engineering,VIT University, India . He has 17 years of experience in Teaching His research interests include Image Processing, medical Imaging, Cloud Computing, Big data, Machine Learning, Deep Learning. He is life time member of CSI.