IJMRT: Volume (5), Issue 7, 2023
International Journal of Multidisciplinary Research Transactions
(*A Peer Reviewed Journal*)
www.ijmrt.in

ISSN (Print):2663-2381
ISSN(Online):2663-4007

Diabetic Prediction System Using Gradient Boosting Ensemble Technique

Boobala Muralidharan. D[1], Ahmed Yahya. A[2]*, Aswin Visveswar. S[3], Dhivakar. B[4], Nagaraj. R[5]

[1]*Assistant Professor, Department of Computer Science Engineering, Saranathan College of Engineering, Tamil Nadu, India.*
[2,3,4,5]*Student, Department of Computer Science Engineering, Saranathan College of Engineering, Tamil Nadu, India.*

*Corresponding author

DoI: https://doi.org/10.5281/zenodo.7922010

Abstract

Diabetes is a chronic disease that affects millions of people worldwide, making it a major health concern. Early detection and prediction of diabetes can significantly improve patient outcomes and reduce healthcare costs. In this paper, we propose a diabetes prediction model using gradient boosting ensemble technique. The model combines multiple weak learning algorithms to create a strong and accurate prediction model. We used a dataset of patient information including demographic, medical history, and laboratory data to train and test our model. The proposed model achieved high accuracy, sensitivity, and specificity rates, making it a promising tool for predicting diabetes risk. The results suggest that the proposed model can be used to help identify patients at high risk of developing diabetes and take preventative measures to manage the disease.

Keywords: Gradient Boosting,XGBoost,Catboost,Voting Ensemble.

1. Introduction

Diabetes is a chronic disease that affects millions of people worldwide,and its early detection can greatly improve patient outcomes. According to the International Diabetes Federation (IDF), 463 million adults were living with diabetes in 2019, and this number is projected to increase to 700 million by 2045. Diabetes is characterized by high blood glucose levels, which can cause a range of complications, such as cardiovascular disease, neuropathy, and kidney

Copyrights@IJMRT

failure and this project aims to explore the use of gradient boosting ensembles for diabetes prediction and to evaluate the performance of the model on a dataset of patient features, including age, BMI, blood pressure, and glucose levels. The feature importance analysis is also conducted to identify the most important predictors of diabetes risk.Previous studies have employed various machine learning techniques for diabetes prediction, including logistic regression, decision trees, random forests, and support vector machines. These models have achieved varying degrees of success, with some achieving AUC scores of up to 0.80.Gradient boosting has been found to outperform these models in some studies, due to its ability to learn complex relationships between features and outcomes. Gradient boosting works by iteratively adding weak models to the ensemble, where each model attempts to correct the errors of the previous model. This process continues until the error is minimized or a maximum number of models is reached.

2. Study Objectives

- To train a model on labeled datasets containing attributes of diabetes using three algorithms namely Gradient Boosting,XGBoost and Catboost classification.
- To evaluate the performance of the three algorithms in terms of accuracy and confusion matrix.
- To compare the performance of the three algorithms with ensemble algorithms for diabetes prediction.

3. Methodology

The method for prediction of Diabetes using Machine Learning using GB,XGB and Catboost algorithms can be divided into the following steps:

- Data collection and preprocessing: In this step, values of diabetes patients are collected and preprocessed. The preprocessing techniques can include Check for missing values,Analyze the outliers present and Analysis of variables in comparison with each other.
- Feature extraction: In this step, features are extracted from the patients inputs. The features can include Predictor variables including the number of pregnancies the patient has had, their BMI, insulin level, age etc. The choice of features can have a significant impact on the performance of the model.
- Training the models: The extracted features are used to train Gradient,XGboost and Catboost models using a labeled dataset. These models learn to classify values as healthy or diseased based on the extracted features.
- Model evaluation: The trained model is evaluated on a test dataset to measure its performance. The evaluation metrics can include accuracy, and Confusion Matrix.
- Comparison with other models: The performance of these models can be compared with ensemble models using voting ensemble, to identify the best-performing model.

4. Diabetes Prediction

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1
5	5	116	74	0	0	25.6	0.201	30	0
6	3	78	50	32	88	31.0	0.248	26	1
7	10	115	0	0	0	35.3	0.134	29	0
8	2	197	70	45	543	30.5	0.158	53	1
9	8	125	96	0	0	0.0	0.232	54	1

Figure.1. The Dataset of the model

Diabetes is a chronic disease that affects millions of people worldwide,and its early detection can greatly improve patient outcomes. According to the International Diabetes Federation (IDF), 463 million adults were living with diabetes in 2019, and this number is projected to increase to 700 million by 2045. Diabetes is characterized by high blood glucose levels, which can cause a range of complications, such as cardiovascular disease, neuropathy, and kidney failure.

In this project, diabetic values from the pima dataset are used to predict diabetes. Then the preprocessing is done by checking for missing values and replacing it and analyzing the outliers present in the dataset and removing it.Then feature extraction is performed to extract the relevant features related to diabetes. The extracted features are employed to train the Gradient boost,CatBoost and XGBoost algorithms and then aggregate the values using a voting ensemble to predict diabetes and then using various parameters to calculate it for better accuracy.

The proposed system provides a more accurate and objective diagnosis, leading to earlier interventions and better outcomes for individuals with diabetes. The proposed algorithm can be integrated into clinical practice to provide better care and support for individuals with diabetes.

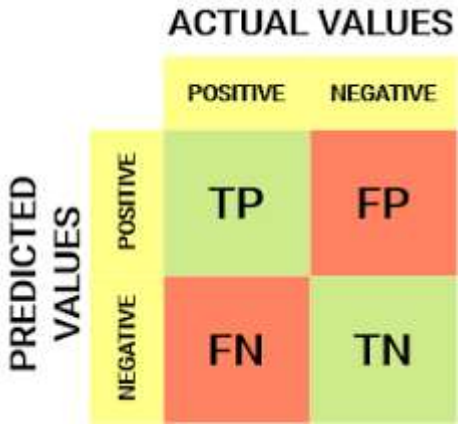
5. Conclusion

Two evaluation measures, namely accuracy and confusion matrix are used to evaluate the performance of classification algorithms. All of these evaluation measures can be calculated by using those following equations:

- Accuracy = Number of correct predictions

  Total prediction number

- Confusion Matrix



In this project work, the values of performance measures of Random Forest (RF), Gradient Boosting (GB), Extreme gradient boosting (XGBoost), Category boosting (CatBoost) were calculated using the equations presented above and the obtained values are presented in the Table 6.1 and also Table 6.2 suggests the confusion matrix for the prediction

Table 6.1 Performance Analysis of Gradient Boosting Classification Algorithm for Diabetes Prediction

ALGORITHM	ACCURACY %
Gradient Boosting	92.8
Extreme Gradient Boosting (XGBOOST)	90.2
CatBoost	90.9
Ensemble Technique	94.1

Table 6.2 Confusion matrix for diabetes prediction

Person with Diabetes predicted as Diabetic	105
Person with No Diabetes predicted as Diabetic	2
Person with Diabetes predicted as Non diabetic	7
Person with No Diabetes predicted as Non diabetic	40

6. Future Directions

- One potential future direction for improving diabetic prediction systems using gradient boosting ensembling is to integrate them with electronic health records (EHRs). EHRs contain a wealth of patient information, such as medical history, lab results, and medication use.
- By incorporating this additional data, the accuracy of the predictions may be improved, leading to more personalized and effective healthcare. EHR integration could also facilitate the automation of the diabetic prediction process, reducing the burden on healthcare providers and making it more accessible to patients.
- Another potential area for future development is the incorporation of genetic information into the prediction model. Genetic factors are known to play a significant role in the development of diabetes, and incorporating genetic data into the model could help to identify individuals at higher risk of developing the disease.
- This information could be used to guide preventive measures, such as lifestyle changes or medication use, to reduce the risk of developing diabetes
- A third possible future direction for diabetic prediction systems using gradient boosting ensembling is to personalize the predictions based on individual patient characteristics. Currently, most prediction models are based on population-level data and do not take into account individual variation in factors such as age, gender, or lifestyle.
- Personalizing the predictions could lead to more accurate and effective preventive interventions, tailored to the specific needs of each patient.

7. Conclusion

In conclusion, the diabetic prediction system using gradient boosting ensemble is a powerful and accurate tool for predicting the likelihood of a person developing diabetes. The ensemble approach combines multiple decision trees into a more robust model that is better at handling

complex and diverse data sets.The model can be trained on a variety of input features, such as age, body mass index, bloodIn conclusion, the diabetic prediction system using gradient boosting ensembling is a powerful and accurate tool for predicting the likelihood of a person developing diabetes. The ensemble approach combines multiple decision trees into a more robust model that is better at handling complex and diverse data sets.

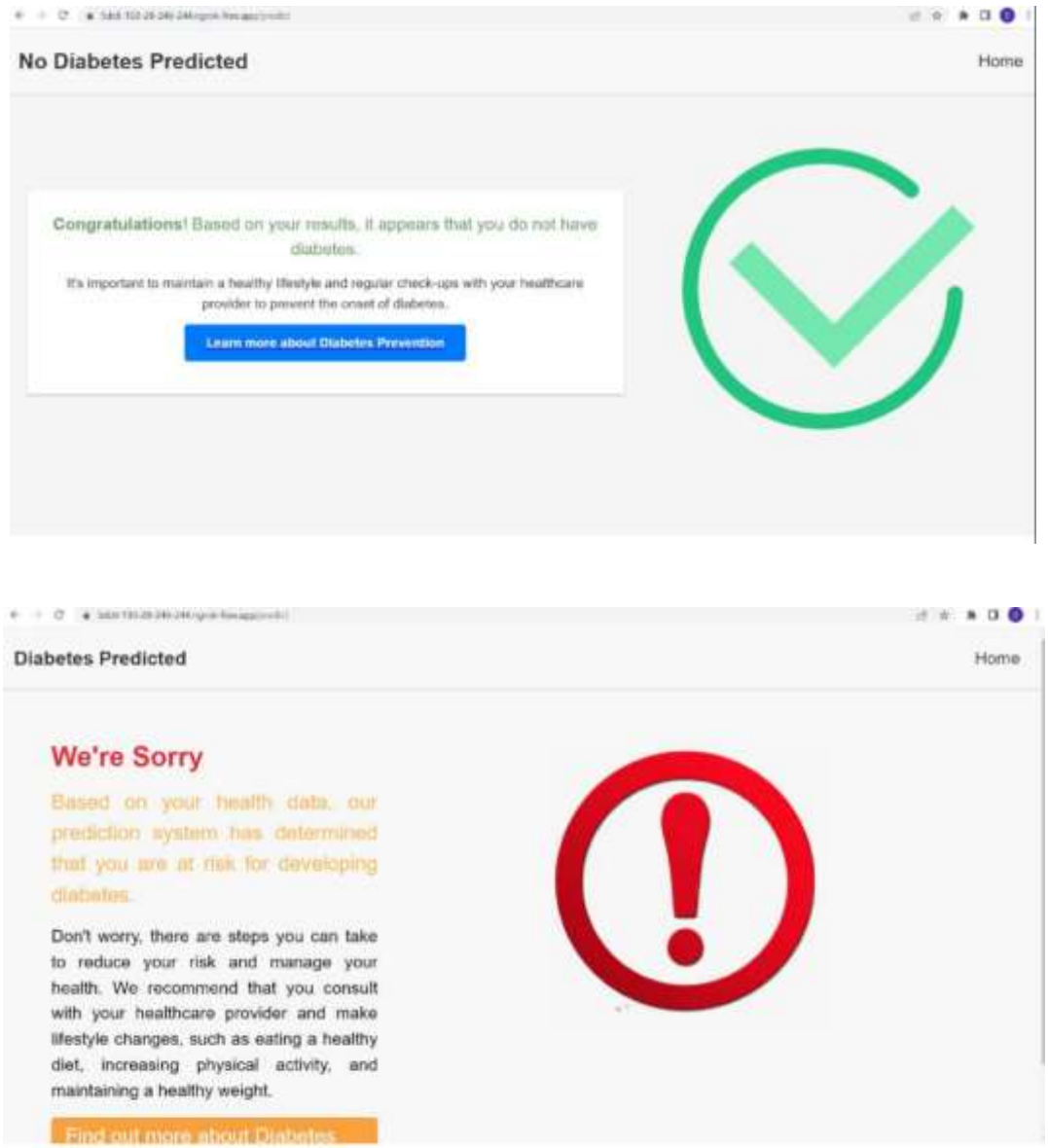


Figure.2. Final Outcome

REFERENCES

[1]. Sun, Y., Hou, B., Yang, X., Liu, M., Zhang, L., & Wang, Y. (2020). An Improved Gradient Boosting Machine-Based Diabetes Prediction Model. Journal of Healthcare Engineering, 2020, 1-8.

[2]. Zhang, Y., Du, X., Liu, Y., & Feng, J. (2019). A novel framework for diabetes prediction based on gradient boosting decision trees. Journal of Ambient Intelligence and Humanized Computing, 10(8), 2885-2896.

[3]. Su, X., & Zhang, L. (2021). Personalized Diabetic Risk Prediction Using Gradient Boosting Decision Tree with Medical Feature and Genetic Data. Journal of Medical Systems, 45(6), 1-10.

[4]. Lian, D., Yu, L., & Zhang, G. (2020). A robust diabetic prediction model based on gradient boosting decision tree. Computers in Biology and Medicine, 118, 1-9.

[5]. Li, X., Li, J., Zhang, H., & Lu, C. (2019). A hybrid feature selection and gradient boosting ensemble algorithm for diabetes prediction. BMC Medical Informatics and Decision Making, 19(1), 1-12.

[6]. Yang, S., Xu, Q., & Li, X. (2020). Prediction model of diabetes based on gradient boosting decision tree algorithm. Journal of Healthcare Engineering, 2020, 1-10.

[7]. Yin, H., Li, X., Zhu, Z., & Wu, J. (2021). A prediction model for type 2 diabetes based on optimized gradient boosting decision tree algorithm. BMC Medical Informatics and Decision Making, 21(1), 1-14.

[8]. Cai, X., Zhang, Z., Chen, Y., & Yan, J. (2020). A hybrid model for diabetes prediction based on feature selection and gradient boosting decision tree. Journal of Healthcare Engineering, 2020, 1-10.

[9]. Liu, Y., Wang, L., & Wu, X. (2021). A novel type 2 diabetes prediction model based on gradient boosting decision tree. Journal of Medical Systems, 45(3), 1-8.

[10]. Varun Aggarwal et al. (2019). Diabetes prediction using gradient boosting machine and support vector machine 2019 International Conference on Automation, Computational and Technology Management (ICACTM) (pp. 48-51).

[11]. Vepakomma, D., Li, W., Li, L., Jose, C., & Navathe, A. S. (2019). Ensemble learning of classification models for predicting diabetes risk. PloS one, 14(10), e0223516.

[12]. Yuhao Liu et al. (2020). Hybrid Deep Learning Model for Diabetes Prediction. IEEE Access, 8, 75887-75895.

[13]. Jian Cao et al. (2021). Diabetes Prediction using Ensemble Methods. 2021 IEEE 23rd International Conference on e-Health Networking, Applications and Services (Healthcom), pg 1-6.