Diabetes Prediction in Teenagers using Machine Learning Algorithms

Jayavrinda Vrindavanam
Department of Computer Science &
Engineering (AI & ML)
School of Engineering
Dayananda Sagar University
Bengaluru, India
0000-0003-3527-3001

Raye Haarika
Department of Computer Science &
Engineering (AI & ML)
School of Engineering
Dayananda Sagar University
Bengaluru, India
0000-0001-7938-8359

Sindhu MG
Department of Computer Science &
Engineering (AI & ML)
School of Engineering
Dayananda Sagar University
Bengaluru, India
0000-0001-9368-7590

Kilari Sumanth Kumar
Department of Computer Science & Engineering (AI & ML)
School of Engineering, Dayananda Sagar University
Bengaluru, India
0000-0001-9662-8334

***Abstract*—Diabetes Mellitus (DM) is a hazardous condition that can lead to worldwide health problems owing to an increase in blood glucose levels, age, lack of movement, high blood pressure, poor nutrition, and other factors. The aim of this study is to develop a system that can predict diabetes in individuals aged 10 to 30 by merging the results of different machine-learning algorithms. Machine learning is used for increasing performance and forecasting accuracy. Some of the methods used to detect diabetes early include Logistic Regression (LR), Random Forest (RF), Support Vector Machine (SVM), K Nearest Neighbours (KNN), and XGBoost.In comparison to the other methods, the random forest and XGBoost Classifiers algorithm outperformed all other ML algorithms, with a maximum accuracy of 86**

Keywords—Diabetes Mellitus, Logistic Regression, K- Nearest Neighbours, Random Forest, Support vector machine, XG-Boost

I. Introduction

Type 1 and type 2 diabetes are two different types of diabetes, both characterized by high blood glucose levels. Type 1 diabetes is an autoimmune disease that occurs when the body's immune system attacks and destroys the insulin-producing beta cells in the pancreas. As a result, the body cannot produce enough insulin to regulate blood sugar levels, and people with type 1 diabetes require lifelong insulin injections.

Type 2 diabetes, on the other hand, occurs when the body becomes resistant to the effects of insulin, or when the pancreas cannot produce enough insulin to meet the body's demands. This type of diabetes is often linked to lifestyle factors such as obesity, physical inactivity, and poor diet. It can often be managed with lifestyle changes, medications, and sometimes insulin injections. Early diagnosis and effective management of both types of diabetes are essential to prevent complications such as heart disease, stroke, kidney damage, and blindness. [4]

Diabetes sometimes goes unreported because persons who have it lack awareness about it or are asymptomatic; roughly one-third of diabetic patients are unaware of their condition. Diabetes causes substantial long-term damage to multiple organs and bodily systems, including the kidneys, heart, nerves, blood vessels, and eyes, if it is not treated. Thus, early identification of the condition allows persons at risk to adopt preventative measures to slow disease development and enhance the quality of life.

Glucose levels in a normal individual range from 70 to 99 mg/dL. A person is diagnosed with diabetes if their fasting glucose level exceeds 126 mg/dL. Over time, it has been discovered that people with BMI more than 25 and cholesterol less than 40 mg/dl and a lazy lifestyle have more chance of getting diabetes.

When a doctor diagnoses a patient with prediabetes, they urge them to make lifestyle changes. Adopting an exercise regimen and healthy eating habits can aid in the prevention of diabetes. The goal of this study is to determine the risk of developing diabetes in people aged 10 to 30. Section 3 describes the methodology, while Section 4 contains the results. The conclusion is summed up in Section 5.

II. Literature Survey

Machine learning's rapid advancement has improved its possibility for accurate predictive analysis. As a result, a variety of businesses, most of the healthcare industry, have embraced its use for many uses. The predictive capability of machine learning has increased the precision and speed with which diabetes is recognised and medicated [2].

In order to determine whether or not a person has diabetes, Yasodha et al [5]. classify many sorts of datasets. The hospital's data warehouse, which has 200 instances with nine attributes, was used to create the data set for the diabetic patient. Both blood tests and urine tests are mentioned in these instances of the dataset. The results showed that among the others, J48 has the highest accuracy (60.2%).

The data set from [1] was obtained at the CPCCSSN. Gradient Boost Method, Logistic Regression, Random Forest, and Rpart classification algorithms are used. The appropriate hyperparameters are used to maximise the area under the ROC curve. Random Forest has the best accuracy.

Diabetes prediction makes use of a variety of categorising algorithms. [6] performs classification utilising three classification algorithms, namely NB, SVM, and DT, and evaluates them using a variety of measures. With a seventy-three per cent accuracy, Naive Bayes was shown to be the most efficient of the three.

This study [7] uses the KNN and LR classification methods to classify. Gradient Boosting feature selection is performed, resulting in higher accuracy. Across feature numbers, the gradient boost scores are compared.

In [8], Decision trees, random forest, and neural network classification methods are employed for model prediction, and performance assessment metrics are investigated. Following feature selection, the findings were confirmed using all features as well as single features such as glucose or skin thickness.

Though several studies have been conducted to predict diabetics in the elderly, few studies have been conducted for teens. We collected data from 200 youngsters and analysed and predicted the results.

III. METHODOLOGY

Figure 1 depicts the algorithm procedure presented in this paper. First, the data set is fed into the prediction algorithm, and then the evaluation model is used to validate the system's classification accuracy by adding a confusion matrix. [4] Finally, the algorithm with the best accuracy in predicting diabetes is obtained.

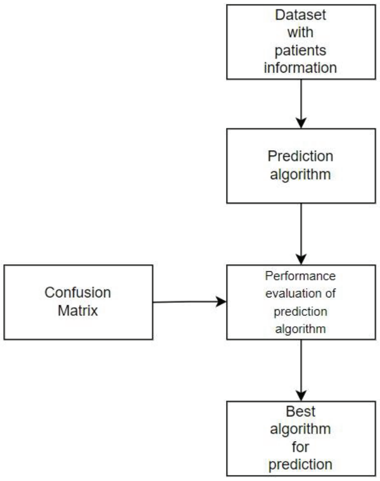
Dataset with patients information
Prediction algorithm
Confusion Matrix
Performance evaluation of prediction algorithm
Best algorithm for prediction

Fig. 1. Methodology

A. Dataset

The data set was gathered by direct surveys distributed to 150 students at Dayananda Sagar University and was validated by physicians. The data set is separated into ten attributes: age, gender, BMI, diet type, blood pressure, exercise routine, parental history, smoking or drinking habits, and so forth. The tenth attribute is the class variable of each data point. This class variable indicates the diabetic outcomes 0 and 1, indicating whether the outcome is favourable or unfavourable for diabetes.

TABLE I. DATASET INFORMATION

Features	Datatype
Gender	Object
Age	Integer
BMI	Float
Vegetarian	Object
Blood Pressure	Integer
Diabetes	Integer
Exercise	Object
Parental History	Object
smoke/drink	Object
Outcome	Integer

B. Data Pre-processing

The most critical stage is to prepare the data. The majority of healthcare-related data has missing values and other impurities, which may limit data efficacy. Data preparation is performed to increase the quality and effectiveness of the mining process's findings. When applying Machine Learning Techniques to a dataset, this strategy is required for accurate findings and effective prediction. Two phases of pre-processing are required for the dataset. [14]

- Missing Values removal- Missing values removal is a data pre-processing technique used in machine learning to handle missing data points. This technique involves removing rows or columns that contain missing values, or replacing the missing values with an estimate such as the mean or median of the feature values. The goal is to improve the quality of the data and prevent errors in the analysis.

- Splitting of data- Splitting of data is a crucial step in machine learning where the available dataset is divided into three subsets - training, validation, and testing sets. The training set is used to train the model, the validation set is used for hyperparameter tuning, and the testing set is used for evaluating the performance of the final model. The split is typically done randomly, with a common split being 80% training data and 20% testing data. Proper splitting of data ensures that the model is trained and evaluated on different sets of data, which helps to avoid overfitting and to obtain a more accurate and reliable model.

C. Machine learning Algorithms and Classification

- K-Nearest Neighbours: Possibly the simplest machine learning algorithm is the k-NN algorithm. The only step in creating the model is saving the training data set. The technique locates a new data point's" nearest neighbours" or the data points in the training data set that are the close to it. [11] [15] [16]

- Logistic Regression: This supervised learning method forecasts a target variable that is category-dependent. It is a sort of machine learning classification issue in which the experimental variable can be binominal, ordinal, interval, or ratio-level, while the dependent variable can be 0 or 1, - 1 or +1, true or false.

- Support Vector Machine: This classifier attempts to generate a hyper plane that can separate the classes as much as possible by altering the distance between the data points and the hyper plane. A number of kernels are used to select the hyperplane.

- Random Forest: The Random Forest classifier builds a large number of decision trees from a random subset of the training dataset. [10]
- XGBoost: Extreme Gradient boosting(XGBoost) is gradient-boosted trees approach. With XG boost models, we employed the entire scikit-learn package. XGB Classifier can generate and fit the XG Boost model for classification to our training dataset. XGBoost is ideally suited for classification issues, particularly fraud detection and customer churn prediction.

IV. Performance Evaluation Measures

There are various performance evaluation measures. The results of the analysis were analysed based on several statistical metrics given below

A. *Accuracy*

Accuracy is a metric used in machine learning to measure the correctness of a model's predictions. It is calculated as the number of correct predictions divided by the total number of predictions made. Accuracy is a widely used metric in classification problems where the goal is to correctly identify the class of a given input. [12]

$$\text{Accuracy} = \frac{T_p + T_n}{T_p + T_n + F_p + F_n} \tag{1}$$

B. *Sensitivity*

The sensitivity of a machine learning model assesses how effectively it can recognise positive examples. A high sensitivity means the model is correctly identifying the majority of the positive findings, whereas a low sensitivity means the model is significantly underreporting the positive findings. [12]

$$SEN = \frac{T_p}{P} \tag{2}$$

C. *Specificity*

Specificity is the percentage of real negatives that were projected as true negative. While a model with low specificity can wrongly categorise many negative results as positive, one with high specificity will correctly identify the bulk of the negative outcomes.

$$SPE = \frac{T_n}{F_p + T_n} \tag{3}$$

D. *F1 Score*

The F1-score is used as a measure when selecting between the two scores might result in the model producing a lot of false positives and false negatives. It is the harmonic mean of the accuracy and recall scores.It also goes by the term FMeasure. [12].

$$FM = 2 \times \frac{PRE \times REC}{PRE + REC} = \frac{2 \times T_p}{2 \times T_p + F_p + F_n} \tag{4}$$

E. *Precision*

In machine learning, precision is a metric that measures the proportion of true positive predictions among all the positive predictions made by a model. In other words, it measures how often a model correctly identifies positive cases. The precision score is calculated as the number of true positives divided by the sum of true positives and false positives. A high precision score indicates that the model has a low rate of false positives and is a reliable indicator of positive cases. [12].

$$PRE = \frac{T_p}{T_p + F_p} \tag{5}$$

F. *ROC Curve*

ROC (Receiver Operating Characteristic) is a performance evaluation metric used in machine learning to assess the tradeoff between the true positive rate and the false positive rate. It plots the true positive rate against the false positive rate to visualize the performance of a binary classification model.

V. Experimental Evaluation

The data obtained from students aged 10 to 30 is utilised to train the model in this research project.80% of the data is utilized for training the model and 20% of the data is used for testing the model. Algorithms such as LR, KNN, SVM, RF and XGBoost algorithms are used and a comparative study is thru to verify and analyse the accuracy of the algorithm. Based on this comparison, it is determined that Random Forest and XGBoost Classifiers offer the highest accuracy of 86% of all algorithms tested and give the desired outcomes. The suggested model's performance evaluation metrics are shown in the table below.

The relationship between BMI and the chance of getting diabetes can be seen in Figure 2. 0 in this graph indicates that the person has no chance of getting diabetes and 1 indicated

TABLE II. PERFORMANCE EVALUATION MEASURES

Performance Measures	**KNN**	**LR**	**SVM**	**RF**	**XGBOOST**
Accuracy	58	79	82	96.49	96.49
Sensitivity	0.69	0.84	0.84	0.92	0.92
Specificity	0.50	0.75	0.81	0.81	0.81
F-Score	0.60	0.78	0.81	0.85	0.85
AUC	0.59	0.86	0.82	0.86	0.86

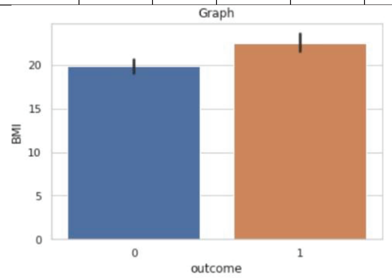

Fig. 2. Relationship between BMI and the chance of getting diabetes that there is a probability or likelihood that an individual may develop diabetes.

The relationship between parental history and the chance of getting diabetes can be seen in figure 3. 0 in this graph indicates that the person has no chance of getting diabetes and 1 indicated that the person has a chance of getting diabetes.

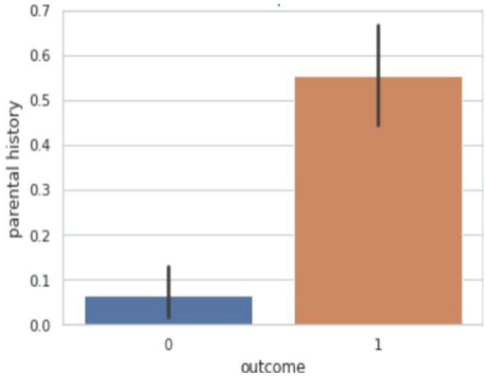

Fig. 3. Relationship between Parental History and the chance of getting diabetes

The relationship between the feature smoke/drink and the chance of getting diabetes can be seen in this figure 4 . 0 in this graph indicates that the person has no chance of getting diabetes and 1 indicated that the person has a chance of getting diabetes.By this graph we can understand that the people who smoke or drink have more chances of getting diabetes. [13]

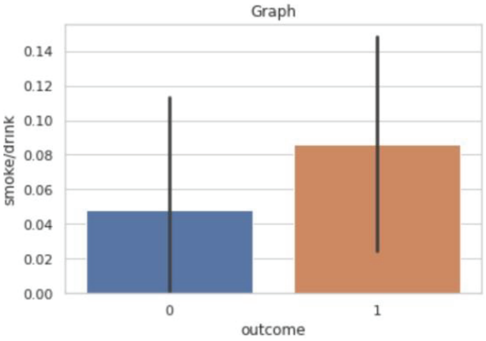

Fig. 4. Relationship between people who smoke or drink and the chance of getting diabetes

The ROC curve for the algorithms used in the proposed model can be seen in the following figures.

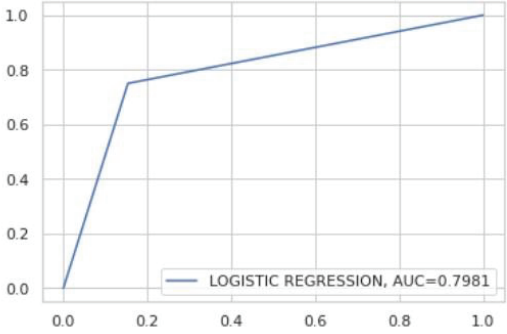

Fig. 5. The ROC Curve of Logistic Regression

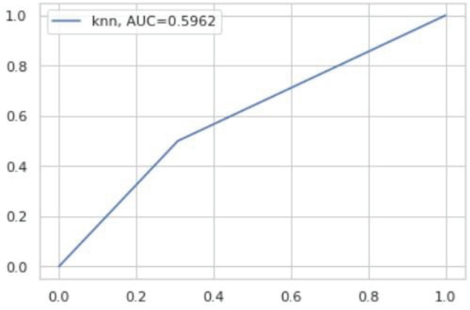

Fig. 6. The ROC Curve of KNN

The graph of the accuracy for the proposed model can be seen in figure 11.

VI. Conclusion

In this work, concerted attempts are made to develop a system that can predict diabetes. Five machine learning classification algorithms are examined and assessed in this paper using a variety of metrics.

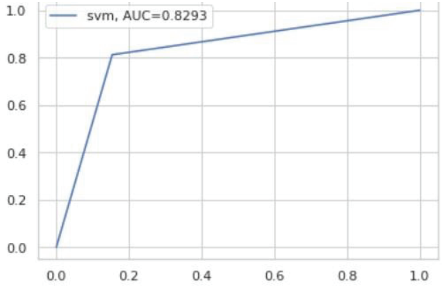

Fig. 7. The ROC Curve of Support Vector Machine

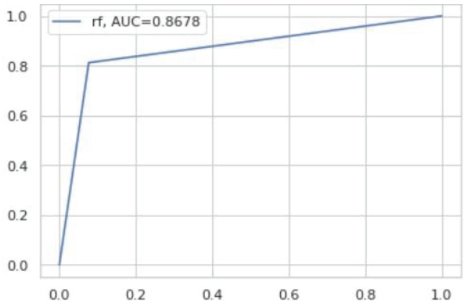

Fig. 8. The ROC Curve of Random Forest

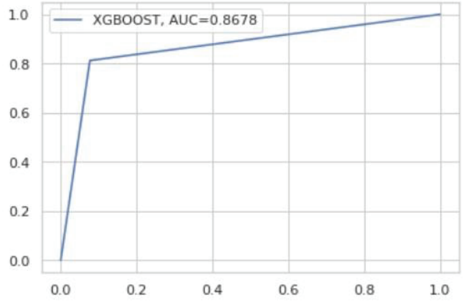

Fig. 9. The ROC Curve of XGBoost

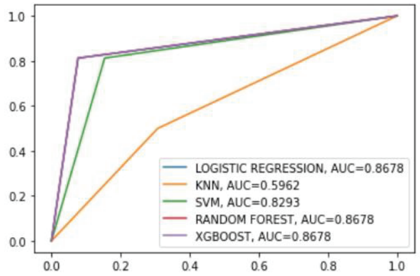

Fig. 10. The ROC Curve of all the algorithms

The Diabetes Database, which is compiled from student data ranging in age from 10 to 30, is the subject of experiments. With an accuracy of 86%, the experimental findings reveal that the proposed system is adequate. Random Forest and the XGBoost algorithms were used.Our aim for the future is to improve the accuracy of predictions by implementing an autonomous deep feature extraction strategy during the feature extraction stage and creating a more fitting model.

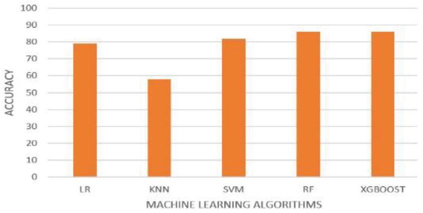

Fig. 11. Comparision of Accuracy of the proposed model

VII. LIMITATIONS

More data would have helped to get more accurate results and would have played a major role in training the model. More features and more samples will help in increasing the accuracy and reliability of the model.

VIII. FUTURE DIRECTION

Future directions for diabetes prediction in teenagers using ML algorithms include the integration of wearable devices, genetic data, electronic health records, deep learning algorithms, and socioeconomic factors to improve accuracy

REFERENCES

[1] Lai, H., Huang, H., Keshavjee, K. et al. Predictive models for diabetes mellitus using machine learning techniques. BMC Endocr Disord 19, 101 (2019).

[2] J. A. M. Sidey-Gibbons and C. J.Sidey-Gibbons, “Machine learning in medicine: a practical introduction,” BMC Med. Res. Methodol., vol. 19, no. 1, p. 64, Mar. 2019, doi: 10.1186/s12874-019-0681-4.

[3] S. Sivaranjani, S. Ananya, J. Aravinth and R. Karthika, ”Diabetes Prediction using Machine Learning Algorithms with Feature Selection and Dimensionality Reduction,” 2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS), 2021, pp. 141-146, doi: 10.1109/ICACCS51430.2021.9441935.

[4] Rani, KM. (2020). Diabetes Prediction Using Machine Learning. International Journal of Scientific Research in Computer Science, Engineering and Information Technology. 294-305. 10.32628/CSEIT206463.

[5] ljumah, A.A., Ahamad, M.G., Siddiqui, M.K., 2013. Application of data mining: Diabetes health care in young and old patients. Journal of King Saud University - Computer and Information Sciences 25, 127–136. doi:10.1016/j.jksuci.2012.10.003.

[6] Deepti Sisodia, Dilip Singh Sisodia, Prediction of Diabetes using Classification Algorithms, Procedia Computer Science, Volume 132,2018

[7] Kayal Vizhi, Aman Dash, “Diabetes Prediction Using Machine Learning”, IJAST, vol. 29, no. 06, pp. 2842 - 2852, May 2020.

[8] Zou Quan, Qu Kaiyang, Luo Yamei, Yin Dehui, Ju Ying, Tang Hua, Predicting Diabetes Mellitus With Machine Learning Techniques,Frontiers in Genetics, volume-9, 2018, 515, ISSN 1664-8021.

[9] Naveen Kishore G, V.Rajesh, A.Vamsi Akki Reddy, K.Sumedh, T.Rajesh Sai Reddy, Prediction Of Diabetes Using Machine Learning Classification Algorithms, International Journal of Scientific & Technology Research Volume 9,Issue 01, January 2020 ISSN 2277-8616.

[10] A. S. Alanazi and M. A. Mezher, ”Using Machine Learning Algorithms For Prediction Of Diabetes Mellitus,” 2020 International Conference on Computing and Information Technology (ICCIT-1441), 2020, pp. 1-3, doi: 10.1109/ICCIT-144147971.2020.9213708.

[11] Neha Prerna Tigga, Shruti Garg, Prediction of Type 2 Diabetes using Machine Learning Classification Methods,Procedia Computer Science,Volume 167,2020,Pages 706-716,ISSN 1877-0509,

[12] Kabiraj et al., ”Breast Cancer Risk Prediction using XGBoost and Random Forest Algorithm,” 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT), 2020, pp. 1-4, DOI: 10.1109/ICCCNT49239.2020.9225451.

[13] Singh, Kuldeep, Pawan et al,. (2018). ”Support vector machine classifierbased detection of fungal rust disease in Pea Plant (Pisam sativam)”. International Journal of Information Technology. 11. 10.1007/s41870018-0134-z.

[14] Nidhi, N., Lobiyal, D.K. ”Traffic flow prediction using support vector regression”. Int. j. inf. tecnol. 14, 619–626 (2022). https://doi.org/10.1007/s41870-021-00852-2

[15] , N. Use of complexity based features in diagnosis of mild Alzheimer disease using EEG signals. Int. j. inf. tecnol. 10, 59–64 (2018). https://doi.org/10.1007/s41870-017-0057-0

[16] , Jameel, R., Shobitha, M. Mourya, A.K. Predictive modeling and cognition to cardio-vascular reactivity through machine learning in Indian adults with a sedentary and physically active lifestyle. Int. j. inf. tecnol. 14, 2129–2140 (2022). https://doi.org/10.1007/s41870-02100721-y