Diabetes Prediction System: A Comparative Study of Classification Models

Amiya Ranjan Panda
School of Computer Engineering
KIIT *deemed to be university*
Bhubaneswar, India
amiya.pandafcs@kiit.ac.in

Pratyush Kumar Prasad
School of Computer Engineering
KIIT *deemed to be university*
Bhubaneswar, India
2128035@kiit.ac.in

Kalyanbrata Giri
School of Computer Engineering
KIIT *deemed to be university*
Bhubaneswar, India
2128075@kiit.ac.in

Sriram Nilakantha Padhy
School of Computer Engineering
KIIT *deemed to be university*
Bhubaneswar, India
2128098@kiit.ac.in

Sudeshna Rath
School of Computer Engineering
KIIT *deemed to be university*
Bhubaneswar, India
2128101@kiit.ac.in

***Abstract*—Diabetes is becoming more and more common worldwide, which presents serious problems for healthcare systems, especially in underdeveloped areas. Early detection is frequently not possible with traditional diagnostic techniques, which leads to treatment delays and more Prediction. To fill these gaps, this work introduces a novel Diabetes Prediction System that uses cutting-edge machine learning algorithms to forecast a person's probability of developing diabetes. [3] This system seeks to make diabetes screening more accessible, accurate, and early by combining a user-friendly web-based interface with powerful prediction models including Random Forest, K-Nearest Neighbors (KNN), and CATBoost. The system emphasizes a comprehensive approach to diabetes prevention and management by providing tailored lifestyle suggestions in addition to simple predictions. [8] This study demonstrates how artificial intelligence can revolutionize conventional medical procedures and encourage proactive action.**

***Keywords*- Machine Learning, Diabetes Detection, Classification, CATBoost, Artificial Intelligence, Prediction System, Healthcare.**

I. INTRODUCTION

Diabetes Mellitus is a chronic condition marked by high blood glucose levels. If untreated, it can lead to serious complications such as renal failure, neuropathy, and cardiovascular disorders. The World Health Organization (WHO) estimates that 422 million people worldwide suffer from diabetes, most of whom live in low- and middle-income nations with inadequate access to early diagnostic techniques [2]. Conventional diagnostic methods mostly depend on laboratory testing and clinician visits, which are frequently unavailable to those living in rural or underdeveloped areas. Furthermore, the necessity for individualized preventive interventions is not sufficiently addressed by these approaches. New opportunities for creating predictive healthcare systems that provide prompt and economical interventions have been made possible by recent developments in artificial intelligence and machine learning [9]. The Diabetes Prediction System described in this study fills these gaps by offering full diabetes screening management capabilities through the combination of a web-based user interface and predictive analytics. High-accuracy projections based on health data, easy accessibility, and practical suggestions for lifestyle changes are some of the system's primary advantages.

II. RELATED WORKS

In this, they Explored diabetes classification using the K-Nearest Neighbors (KNN) algorithm, emphasizing its simplicity and effectiveness in predicting diabetes [1].

This paper proposed an accurate diabetes prediction system integrating K-Means clustering for data preprocessing and a novel classification approach for improved results [2].

Utilized various machine learning algorithms to predict diabetes, focusing on comparative performance analysis [3].

This study used the Decision Tree (C4.5) algorithm for diabetes classification, highlighting its interpretability and effectiveness [4].

Developed a data analytics suite for exploratory, predictive, and visual analysis of Type 2 diabetes, enabling more informed decision-making [6].

Reviewed data cleaning techniques to address challenges like missing values and noise in datasets, ensuring higher-quality data for modeling [7].

Employed machine learning techniques to predict diabetes, focusing on the feature importance and model optimization [8].

Investigated the application of classification algorithms in diabetes prediction, emphasizing accuracy improvements [9].

Conducted an exploratory study using a combined Random Forest classifier for diabetes classification, achieving high accuracy through feature importance analysis [10].

III. Basic concepts

a. *K Nearest Neighbour*

It is a simple, non-parametric algorithm that classifies data points based on the majority class of their kk nearest neighbors in the feature space. It uses distance metrics like Euclidean to find similarities and is effective for smaller datasets due to its simplicity[1].

b. *Decision Tree*

A Decision Tree is a hierarchical model that classifies data by dividing it into branches based on feature thresholds, eventually assigning a class label. Its clear structure makes it well-suited for interpreting how individual features influence the classification process.[4]

c. *Naive Bayes*

Naive Bayes is a probabilistic model based on Bayes' Theorem, which operates under the assumption that features are independent. Despite this simplification, it performs effectively across various classification tasks, including binary problems like diabetes prediction. In this context, Naive Bayes predicts diabetes by calculating the likelihood of feature values for each class[10].

d. *Random Forest*

Random Forest is an ensemble method that constructs multiple Decision Trees and aggregates their predictions to improve accuracy and minimize overfitting. Each tree is trained on a random subset of the data and features, enhancing the model's robustness[11].

e. *XG Boost (Extreme Gradient Boosting)*

XGBoost (Extreme Gradient Boosting) is an advanced machine learning algorithm built on the gradient boosting technique[4]. It creates an ensemble of decision trees in a sequential manner, where each new tree seeks to correct the errors made by the previous ones by minimizing a given loss function. XGBoost is renowned for its speed and high performance, owing to features like regularization (to prevent overfitting), parallel computation, tree pruning, and efficient handling of missing values. It is commonly applied to structured/tabular data for tasks such as classification, regression, and ranking, due to its scalability and ability to model complex relationships.

f. *Ridge Classifier*

The Ridge Classifier is a linear classification model that applies L2 regularization (also known as Ridge regression) to prevent overfitting by penalizing large coefficients[5]. It works by finding the optimal hyperplane that separates classes while minimizing the impact of irrelevant features. The Ridge Classifier is particularly useful when dealing with multicollinearity or when the number of features exceeds the number of observations, offering a more stable solution compared to regular linear classifiers[5].

g. *CATBoost*

CatBoost is a gradient-boosting algorithm optimized for efficiently handling categorical data without requiring extensive preprocessing. It constructs an ensemble of decision trees, where each tree aims to fix the mistakes of the previous one. Known for its speed and accuracy, CatBoost can handle various data types, including missing values, and is designed to reduce overfitting. It is commonly applied in classification, regression, and ranking tasks, especially when dealing with categorical features.

h. *Support Vector Machine (SVM)*

Support Vector Machine (SVM) is a supervised machine learning algorithm used for both classification and regression tasks. It works by identifying the optimal hyperplane that best separates different classes in a high-dimensional space, aiming to maximize the margin between them for better generalization[7]. SVM is effective with both linear and non-linear data, particularly when kernel functions are used to map the data into higher dimensions. It is recognized for its robustness, especially in handling high-dimensional datasets.

IV. Methodology

The methodology involves preprocessing the dataset by normalizing features and encoding categorical variables to ensure compatibility with machine learning models.

1. *Exploratory Data Analysis*

In Fig 1, we plot multiple histograms amongst various attributes of the dataset.

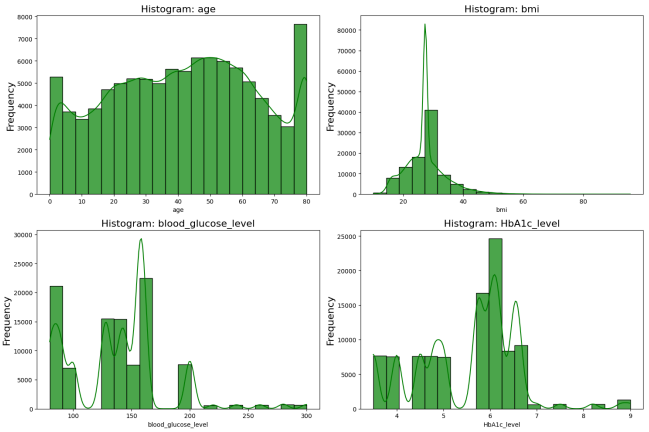


Fig. 1: Attribute histogram

In Fig 2, we plot a scatter plot between BMI & Blood Glucose Levels.

Fig. 2: Scatter plots of Attributes

In Fig 3, A box plot is plotted to establish the relation between BMI and Diabetes Status.

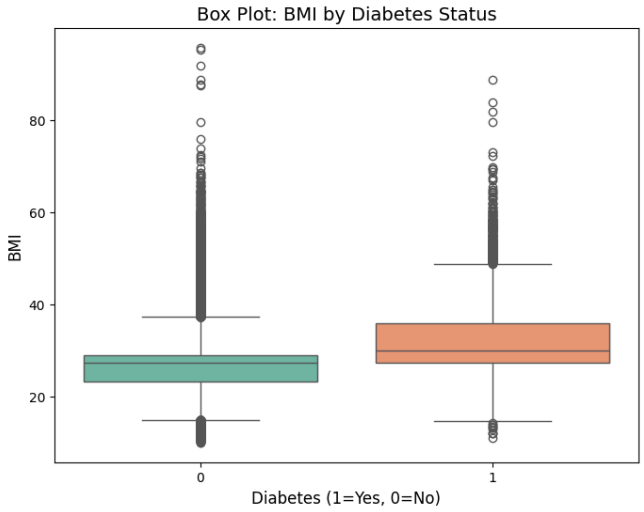
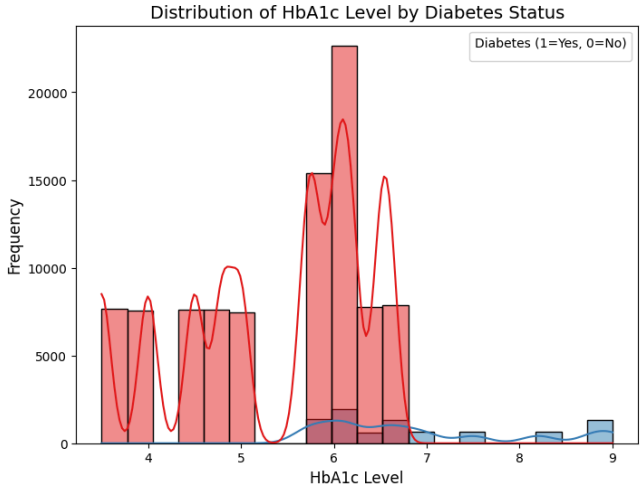


Fig. 3: Box plot

In Fig 4, we establish the correlation between HbA1c Level and frequency of Diabetes Status using a Violin plot.



In Fig 5, Furthermore, we establish the correlation between the attributes and other attributes, using a **Heatmap**

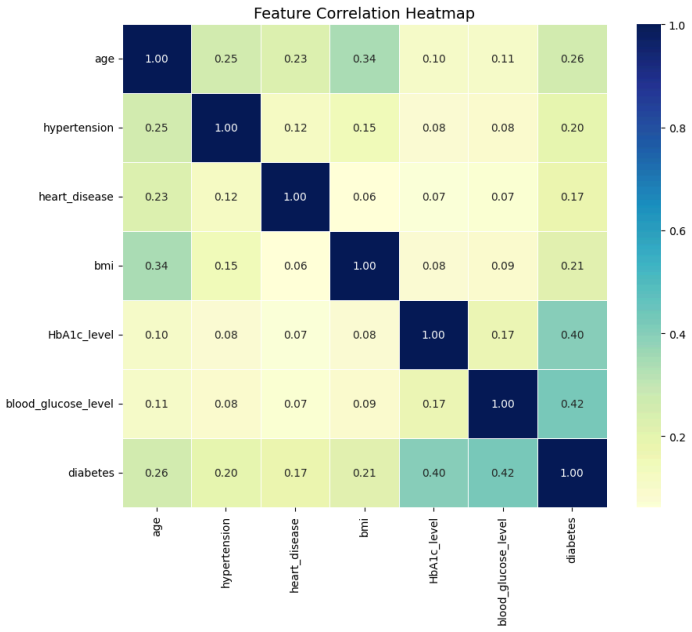


Fig. 5: Heatmap

2. Data Preprocessing

I. Missing Data - in which all the data that are missing are identified and dealt with using various methods like dropping rows with missing / null values, and dropping an entire column with missing values.

Column	Non Null Content	d_type
age	0	int64
hypertension	0	int64
heart_decision	0	int64
bmi	0	int64
HbA1c_level	0	int64
blood_glucose	0	int64
diabetes	0	int64

Fig. 6: missing values

II. Outlier Detection- For the detection of Outliers, we start by plotting Box Plots for each attribute, by which we can visualize the outliers.
We implement the Interquartile Ranges (IQR) for the detection of outliers in the dataset used for analysis.

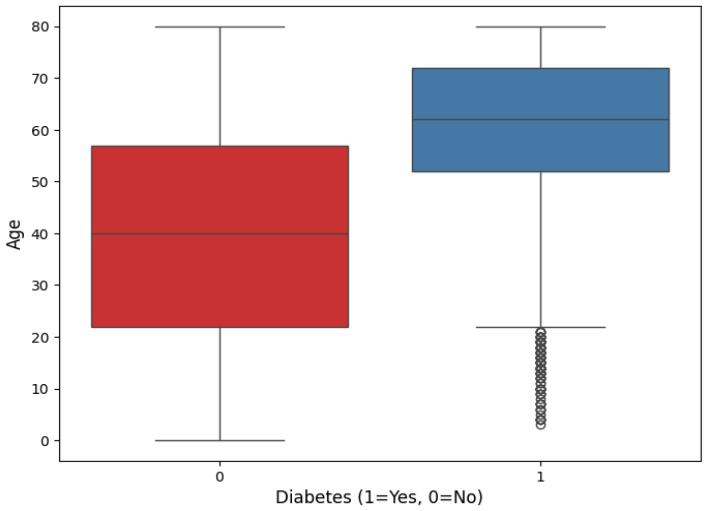


Fig. 7: Box plot for detecting outlier

3. Defining Function

I. Pearson Correlation: In a regression model, this measures the strength of the relationship between two attributes in the dataset.
II. Coefficient of Determination (R-squared): This metric evaluates model performance by calculating the proportion of variance in the dependent variable that is explained by the independent variables.
III. Mean Squared Error (MSE): This metric assesses model performance by averaging the squared differences between predicted and actual values.
IV. Root Mean Squared Relative Error (RMSRE): This measures the model's performance by taking the square root of the squared errors and normalizing it by dividing it by the total squared error.

4. Models -

Finishing all of the above-stated methods, now we move on to training the Machine using various models. We will apply classification models to our data set.

I. KNN Classification
II. Decision Tree
III. Naive Bayes
IV. Random Forest
V. XG Boost
VI. CATBoost
VII. Support Vector Machine
VIII. Ridge Classification

We train our machine through all the models, one by one, and find out the performance of the model for the dataset, using various evaluation metrics for regression.

V. Result Analysis

Classification metrics are qualitative measures used to evaluate the performance of a classification model. These metrics help assess how accurately the model assigns data points to the correct class and distinguishes between different categories. Here are the implemented classification metrics.

A. Error Metrics:

- ***Pearson correlation coefficient(R)***: The strength of the linear relationship between variables increases as the absolute value of the r score approaches 1 (either positively or negatively). Conversely, the linear relationship weakens as the r score moves closer to 0.
- ***R Square Score***: This metric indicates the proportion of variance in the independent variables (features) that explains the variance in the dependent variable (target). A higher value on the scale from 0 to 1 signifies a stronger correlation and a better fit.
- **Root Mean Squared Error (RMSE):** This is the square root of the Mean Squared Error (MSE). It uses the same units as the target variable, making it easier to interpret. A lower RMSE indicates a better model fit.
- **Mean Absolute Error (MAE):** This statistic calculates the average absolute difference between the predicted and actual values. Unlike MSE, it is less influenced by outliers. A lower MAE indicates a closer fit between the predicted and actual values.

Metric	KNN	Decision Tree	Naive Bayes	Random Forest
Pearson R	0.6565816141924140	0.7107311940629520	0.4698219688731770	0.7678542457669300
R2 Score	0.3982039396579090	0.42515938819406500	-0.2612642431336330	0.5718471779024500
MSE	0.048	0.04585	0.1006	0.03415
RMSE	0.21908902300206600	0.21412613105363900	0.3171750305430740	0.1847971861257630
Accuracy	0.952	0.95415	0.8994	0.96585
Recall	0.5251716247139590	0.7305491990846680	0.6155606407322660	0.6796338672768880
Precision	0.8759541984732830	0.7411491584445730	0.445364238410596	0.9061784897025170
F1 Score	0.6566523605150210	0.7358110054739270	0.5168107588856870	0.7767244197450150
Metric	**XGBoost**	**Ridge Classifier**	**CatBoost**	**SVM**
Pearson R	0.7929255711461000	0.5235615459544330	0.7932775128335900	0.7932775128335900
R2 Score	0.6169818824281070	0.22518757230955800	0.6176087533242960	0.6176087533242960
MSE	0.03055	0.0618	0.0305	0.0305
RMSE	0.17478558292948500	0.24859605789312100	0.1746424919657300	0.1746424919657300
Accuracy	0.96945	0.9382	0.9695	0.9695
Recall	0.6773455377574370	0.29462242562929100	0.6756292906178490	0.6756292906178490
Precision	0.9618196588139720	0.9942084942084940	0.9648692810457520	0.9648692810457520
F1 Score	0.7948976166498830	0.45454545454545500	0.7947510094212650	0.7947510094212650

Table 1: Error metrics of all implemented models

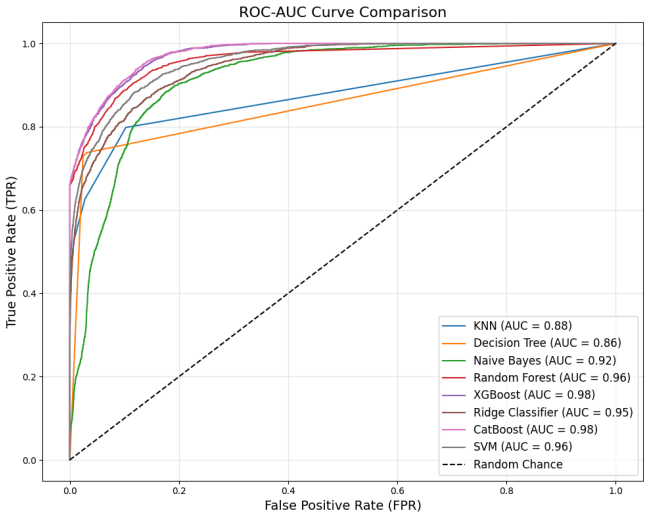


Fig 8: ROC-AUC curve comparison for all models.

VI. Conclusion & Future Scope

The Diabetes Prediction System effectively showcases the use of machine learning for early diabetes detection, with CatBoost proving to be the most accurate and precise model. However, challenges such as optimizing model performance and ensuring its adaptability to real-world data still exist. Future improvements could include incorporating a wider range of datasets, exploring advanced techniques like deep learning, and implementing continuous learning to sustain accuracy. Moreover, integrating real-time health monitoring and ensuring compliance with regulatory standards could make the system more practical and accessible for clinical use, supporting proactive diabetes prevention and expanding healthcare accessibility.

REFERENCES

[1] AMEER Ali, MOHAMMED Alrubei, LF Mohammed Hassan, M Al-
Ja'afari, and Saif Abdulwahed. Diabetes classification based on knn.

IIUM Engineering Journal, 21(1):175–181, 2020.

[2] Mustafa S Kadhm, Ikhlas Watan Ghindawi, and Duaa Enteesha Mhawi.
An accurate diabetes prediction system based on k-means clustering
and proposed classification approach. International Journal of Applied
Engineering Research, 13(6):4038–4041, 2018.

[3] Aishwarya Mujumdar and Vb Vaidehi. Diabetes prediction using
machine learning algorithms. Procedia Computer Science, 165:292–
299, 2019.

[4] B A C Permana, R Ahmad, H Bahtiar, A Sudianto, and I Gunawan.
Classification of diabetes disease using decision tree algorithm (c4.5).
Journal of Physics: Conference Series, 1869(1):012082, apr 2021.

[5] Nada Y Philip, Manzoor Razaak, John Chang, Maurice O'Kane, Bar-
bara K Pierscionek, et al. A data analytics suite for exploratory
predictive, and visual analysis of type 2 diabetes. IEEE Access,
10:13460–13471, 2022.

[6] Richard R Picard and Kenneth N Berk. Data splitting. The American
Statistician, 44(2):140–147, 1990.

[7] Erhard Rahm, Hong Hai Do, et al. Data cleaning: Problems and current
approaches. IEEE Data Eng. Bull., 23(4):3–13, 2000.

[8] KJ Rani. Diabetes prediction using machine learning. International
Journal of Scientific Research in Computer Science, Engineering and
Information Technology, 6:294–305, 2020.

[9] Deepti Sisodia and Dilip Singh Sisodia. Prediction of diabetes using
classification algorithms. Procedia computer science, 132:1578–1585,
2018.

[10] Deepti Sisodia and Dilip Singh Sisodia. Prediction of diabetes using
classification algorithms. Procedia computer science, 132:1578–1585,
2018.

[11] Xuchun Wang, Mengmeng Zhai, Zeping Ren, Hao Ren, Meichen Li,
Dichen Quan, Limin Chen, and Lixia Qiu. Exploratory study on
classification of diabetes mellitus through a combined random forest
classifier. BMC medical informatics and decision making, 21:1–14,
2021.