WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Understanding the Dataset: A Machine Learning Approach on Diabetes Data

Sarthak Sengar

Student MCA Semester 2
Jagan Institute of Management studies, Rohini Sector-5, Near Rithala Metro Station, New Delhi

ABSTRACT :

This paper presents an in-depth analysis of the Diabetes dataset obtained from Kaggle. The study explores various machine learning classification algorithms, including K-Nearest Neighbors (KNN), Gaussian Naïve Bayes, and Decision Tree Classifier, along with cross-validation techniques to evaluate model performance.

The data was processed and analyzed using Python in Google Colab, utilizing key libraries such as Pandas, NumPy, Matplotlib, and Scikit-learn. Extensive Exploratory Data Analysis (EDA) was conducted using various visualization techniques, including pie charts, bar plots, histograms, violin plots, box plots, count plots, pair plots, and correlation heatmaps.

The research highlights the strengths and limitations of each algorithm and provides a comparative analysis based on various performance metrics such as accuracy, precision, recall, and F1-score. The results offer valuable insights into the effectiveness of different classifiers in predicting diabetes and suggest possible improvements for future research.

Introduction

Machine learning plays a crucial role in predictive analytics, particularly in the healthcare sector, where early detection of diseases can significantly impact patient outcomes. Diabetes is a chronic disease that affects millions worldwide, making it an important area for predictive modeling. This study focuses on analyzing the Diabetes dataset using multiple classification algorithms to predict diabetes occurrence. The primary objectives of this research are:

- To preprocess and visualize the dataset for better understanding.
- To apply and compare multiple machine learning models for classification.
- To evaluate model performance using cross-validation techniques.
- To analyze the strengths and weaknesses of each classifier in predicting diabetes.
- The paper outlines the importance of data preprocessing, model selection, and performance evaluation in building robust predictive models, ensuring accurate and reliable diabetes diagnosis.

Literature Survey

Several studies have explored machine learning techniques in healthcare diagnostics, particularly in diabetes prediction. Prior research has demonstrated the significance of feature selection, hyperparameter tuning, and algorithm optimization in improving classification accuracy.
Some key findings from previous studies include:

- Logistic Regression and Decision Trees are commonly used classifiers in medical diagnosis due to their interpretability.
- K-Nearest Neighbors (KNN) is effective for pattern recognition but may suffer from high computational cost for large datasets.
- Naïve Bayes classifiers perform well in probabilistic decision- making but assume feature independence, which may not always hold true.
- Cross-validation techniques improve model reliability by reducing bias and variance.

Building upon these findings, this paper applies multiple classifiers and evaluates their performance using cross-validation, aiming to contribute to the field of diabetes prediction.

System Architecture

The system follows a structured approach to data analysis, including preprocessing, model training, evaluation, and validation. The architecture consists of the following components:

Methodology

- **Data preprocessing:** Handling missing values, feature scaling, and splitting data into training and test sets.
- **Exploratory Data Analysis (EDA):** Visualizing data distributions, correlations, and trends to understand feature relationships.
- **Model implementation:** Applying KNN, Gaussian Naïve Bayes, and Decision Tree Classifier to classify diabetic and non-diabetic patients.
- **Hyperparameter tuning:** Adjusting parameters like the number of neighbors in KNN, tree depth in Decision Tree, and prior probabilities in Naïve Bayes.
- **Cross-validation:** Using k-fold cross-validation to assess model performance and ensure generalizability.

Data Collection

The Diabetes dataset was sourced from Kaggle and analyzed in Google Colab using Python. The dataset includes features such as gender, age, hypertension,heart_disease,smoking_history,bmi,HbA1c_level,blood_glucos e_level , which are critical indicators for diabetes prediction. Key libraries used for data handling and visualization include:

- **Pandas & NumPy:** For data manipulation and numerical computations.
- **Matplotlib & Seaborn:** For visualizing feature distributions and relationships.
- **Scikit-learn:** For implementing machine learning algorithms and evaluation metrics.

Exploratory Data Analysis (EDA) Visualization

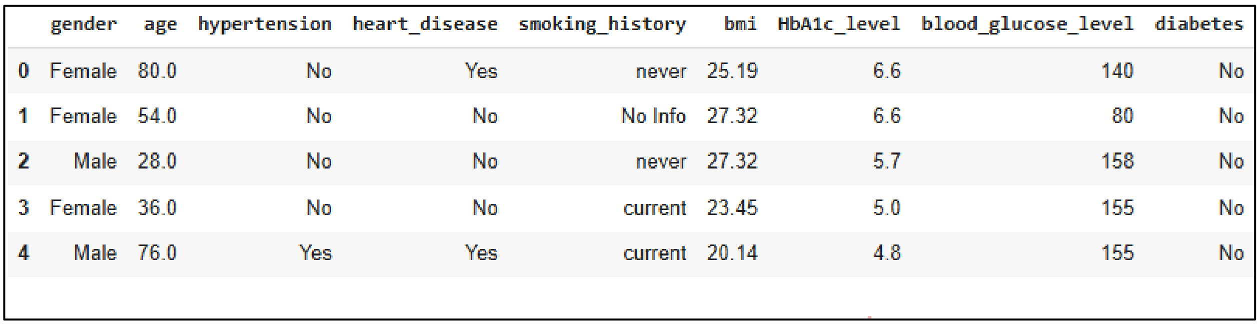
To gain insights into the dataset, multiple visualization techniques were applied:

1. **Pie Chart** – Showed the proportion of diabetic and non-diabetic cases.
2. **Bar Plot** – Represented categorical feature distributions such as the number of patients in different age groups.
3. **Histogram with Internal Distribution** – Analyzed the spread and frequency distribution of numerical features such as glucose levels and BMI.
4. **Violin Plot** – Combined box plots and KDE (Kernel Density Estimation) to visualize feature distributions.
5. **Box Plot** – Highlighted outliers and the spread of numerical attributes like insulin levels and skin thickness.
6. **Count Plot** – Displayed the count of different categorical values, helping in understanding class imbalances.
7. **Pair Plot** – Showed pairwise relationships between numerical features, identifying correlations and trends.
8. **Correlation Heatmap (Pearson Correlation Coefficients)** – Demonstrated the strength and direction of feature correlations, helping in feature selection and multicollinearity analysis.

Feature Importance Analysis

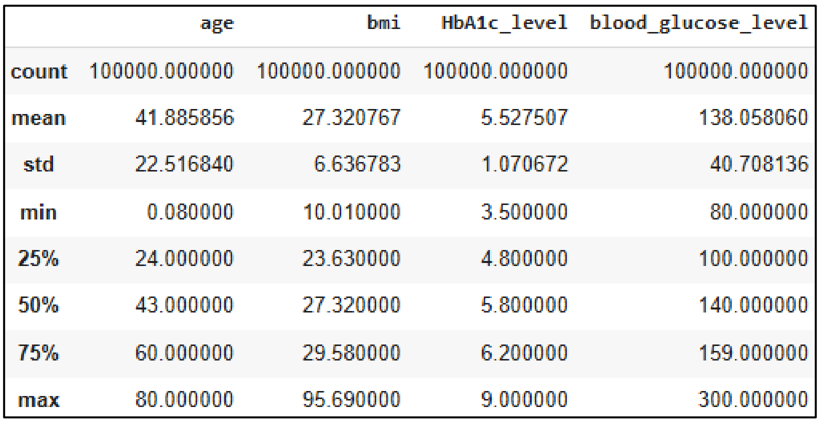
Feature selection was performed to determine the most influential attributes for diabetes prediction. Pearson correlation, mutual information scores, and feature importance from Decision Tree Classifier were used to identify key features.

Results and Analysis Data Set



	gender	age	hypertension	heart_disease	smoking_history	bmi	HbA1c_level	blood_glucose_level	diabetes
0	Female	80.0	No	Yes	never	25.19	6.6	140	No
1	Female	54.0	No	No	No Info	27.32	6.6	80	No
2	Male	28.0	No	No	never	27.32	5.7	158	No
3	Female	36.0	No	No	current	23.45	5.0	155	No
4	Male	76.0	Yes	Yes	current	20.14	4.8	155	No

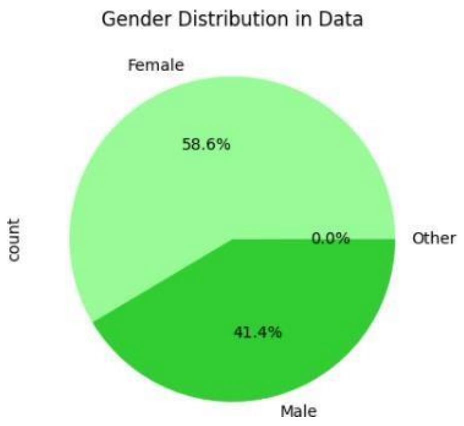
Data Stats



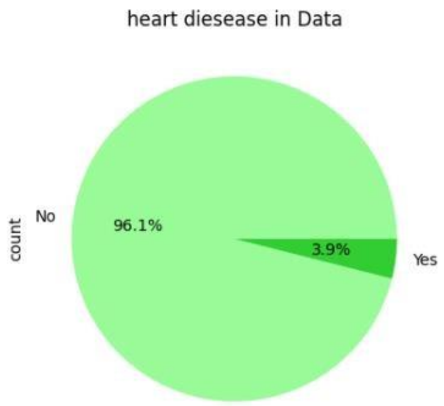
	age	bmi	HbA1c_level	blood_glucose_level
count	100000.000000	100000.000000	100000.000000	100000.000000
mean	41.885856	27.320767	5.527507	138.058060
std	22.516840	6.636783	1.070672	40.708136
min	0.080000	10.010000	3.500000	80.000000
25%	24.000000	23.630000	4.800000	100.000000
50%	43.000000	27.320000	5.800000	140.000000
75%	60.000000	29.580000	6.200000	159.000000
max	80.000000	95.690000	9.000000	300.000000

EDA

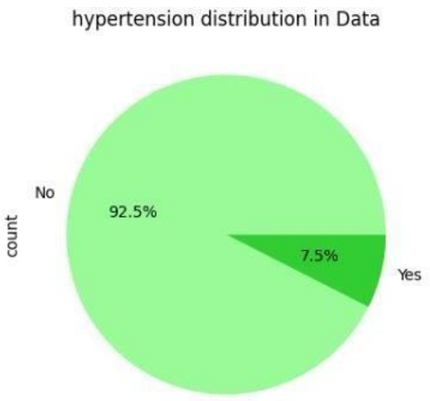
1. **pie chart**



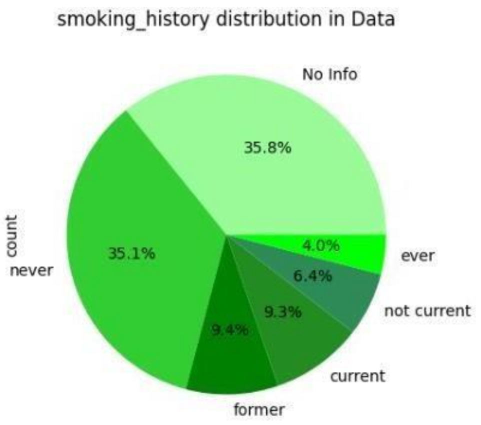
Gender Distribution in Data



heart diesease in Data

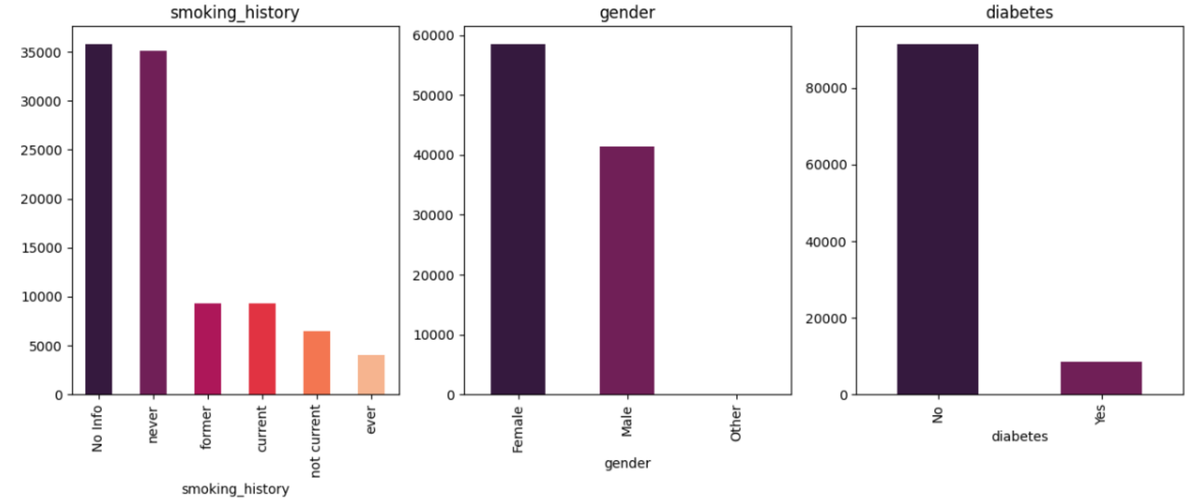


hypertension distribution in Data

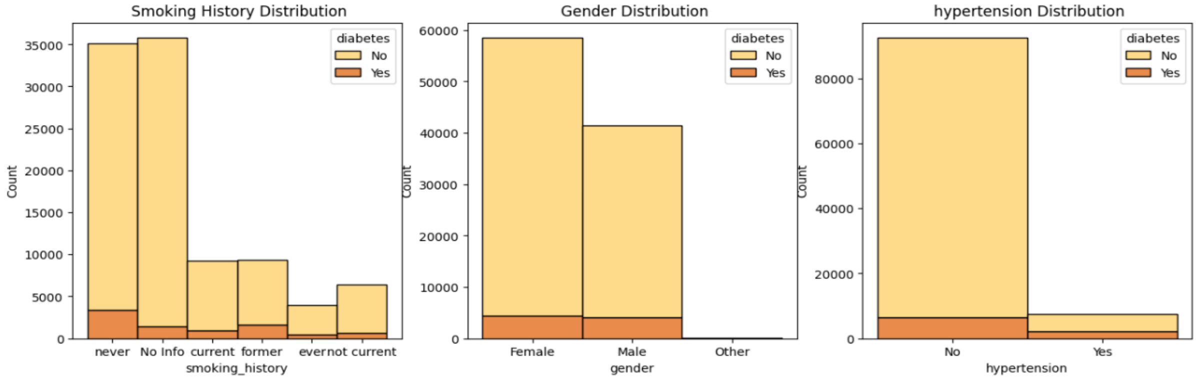


smoking_history distribution in Data

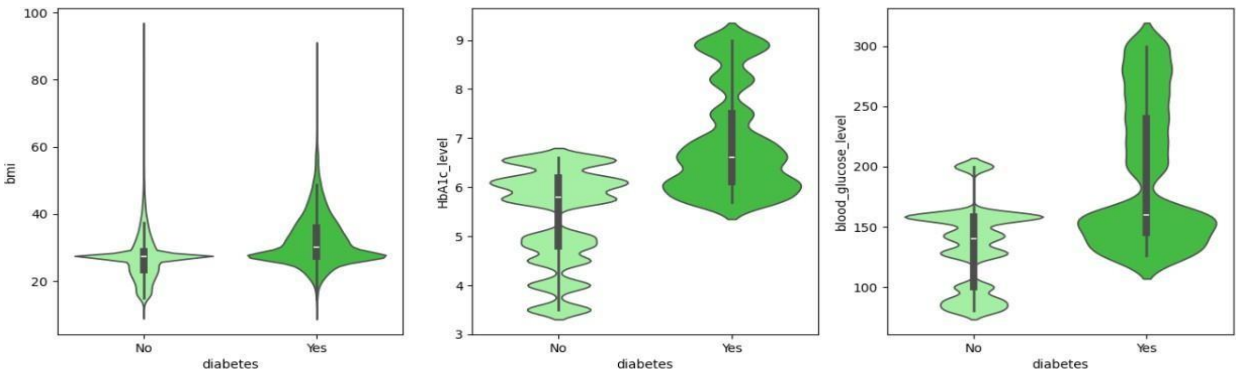
2. **bar plot**



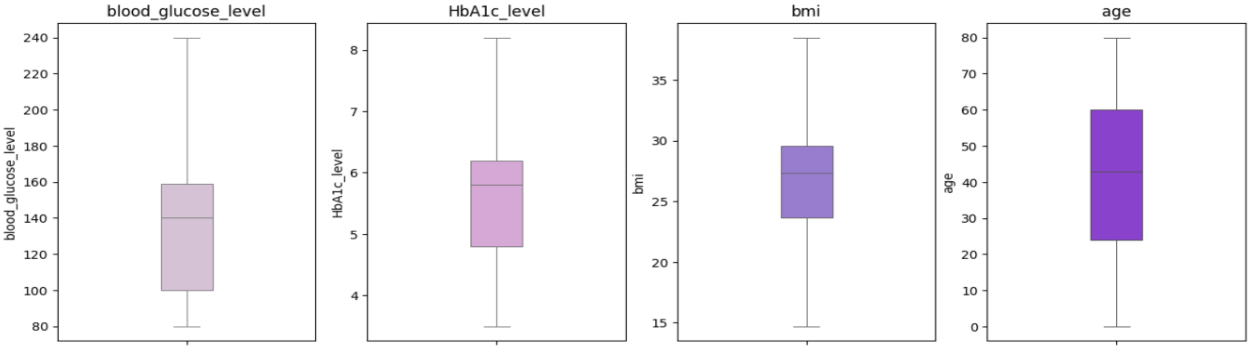
3. **histogram with internal distribution**



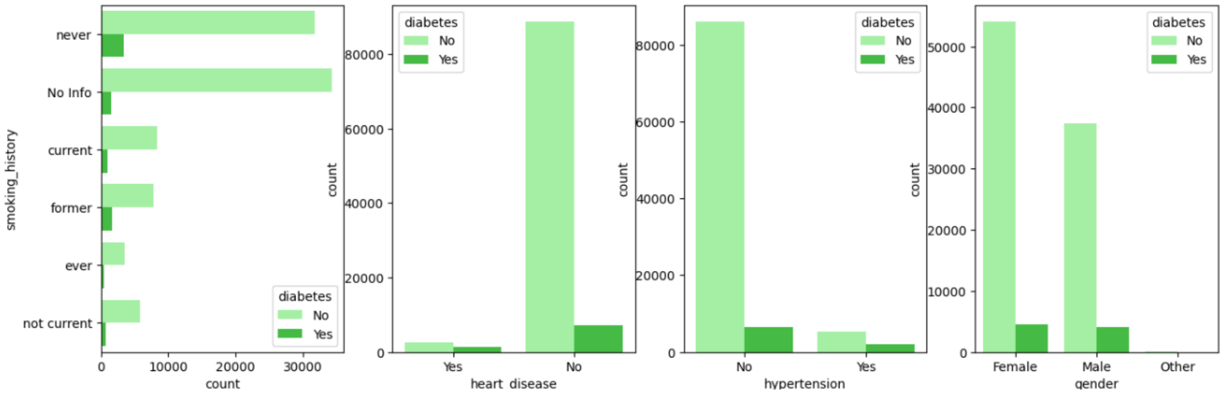
4. **voilin plot**



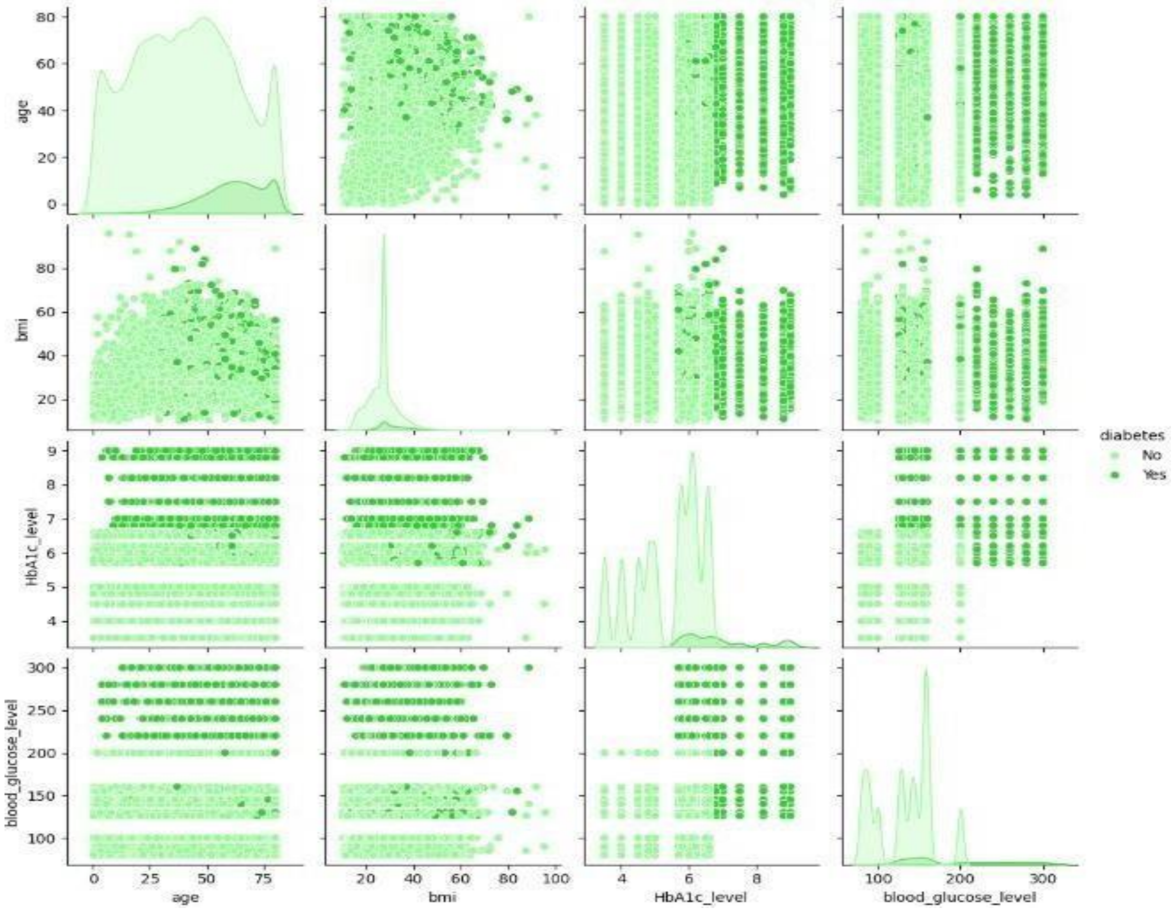
5. **boxplot**



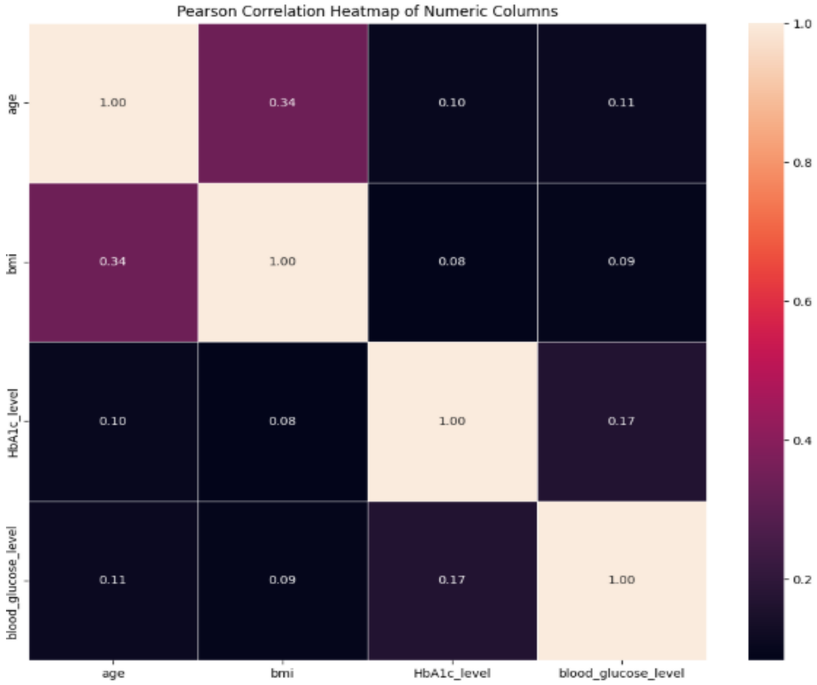
6. **countplot**



7. **pairplot**



8. **correlation heatmap (pearson Coorelation coefficients)**

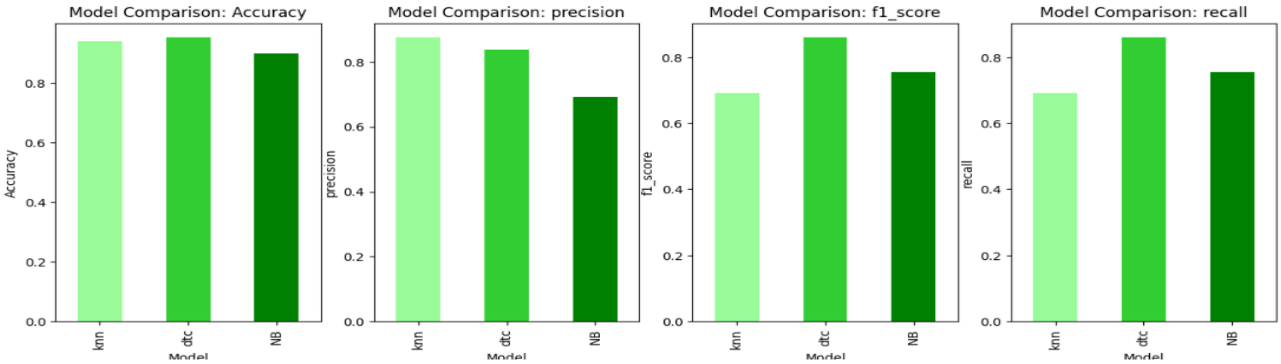


Model Accuracies and performance : -

The study evaluates the classifiers based on their accuracy and other performance metrics. Key findings include:

- **K-Nearest Neighbors (KNN):** Achieved an accuracy of **94.056%**
- **Gaussian Naïve Bayes:** Performed efficiently with an accuracy of **89.87%**,
- **Decision Tree Classifier:** Achieved an accuracy of **95.18%**,

Additionally, k-fold cross-validation was performed to ensure the stability of the results. The confusion matrices and ROC curves provided further insights into model strengths and weaknesses.



Conclusion and Future Recommendations

This research demonstrates the effectiveness of machine learning classifiers in diabetes prediction. The Decision Tree model showed high interpretability, while KNN provided strong classification performance. However, the Gaussian Naïve Bayes model, despite its probabilistic strengths, struggled with correlated features.

Future work could explore:

- The integration of deep learning techniques such as neural networks for enhanced predictive accuracy.
- Feature engineering to derive new attributes from existing data for improved model performance.

- Implementation of ensemble learning techniques like Random Forest to reduce overfitting and improve generalization.
- Expanding the dataset with additional medical attributes to enhance predictive power.

REFERENCES

[List all the sources and Kaggle dataset link used in the study following a standard citation format. Example:]

1. Kaggle Diabetes Dataset: **https://www.kaggle.com/datasets/iammustafatz/diabetes- prediction- dataset**

2. **Introduction to Data Mining**, Pang-Ning Tan, Michael Steinbach, Vipin Kumar, Pearson

3. **Data Mining: Concepts and Techniques**, 3nd edition,Jiawei Han and Micheline Kamber

4. 2006**https://youtube.com/playlist?list=PLKnIA16_Rmvbr7zKYQuBfsVkj oLcJgxHH&si=q0oqOXF3arJrTgXD**

5. Documentation for Scikit-learn, Pandas, NumPy, and Matplotlib