



VIT[®]
Vellore Institute of Technology
(Deemed to be University under section 3 of UGC Act, 1956)

SCHOOL OF COMPUTER SCIENCE ENGINEERING AND INFORMATION SYSTEMS

Course Code : SWE2011

Course Name : **Big Data Analytics**

FINAL REVIEW

SLOT : B1+TB1

Diabetes data analysis and prediction

TEAM MEMBERS:

1. Sridhar S 21MIS0140
2. Kaushik S 21MIS0362
3. Tharun S 21MIS0408

FACULTY NAME:

Dr. J.Jagannathan,

Assistant Professor Sr.Grade I

CONTENTS

| Chapter No | Title | Page No |
|------------|-----------------------------------|---------|
| 1 | Abstract | 1 |
| 2 | Introduction | 1 |
| 3 | Objective | 2 |
| 4 | Problem statement | 2 |
| 5 | Literature Review | 2 |
| 6 | Data Exploration | 4 |
| 7 | Methods | 5 |
| 8 | Modeling and Result | 6 |
| 9 | Conclusion and Future Development | 11 |
| 11 | References | 12 |

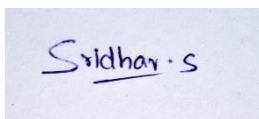
DECLARATION

We hereby declare that the report entitled “**Diabetes data analysis and prediction**” submitted by us, for the **SWE2011-Big Data Analytics** to Vellore Institute of Technology is a record of bonafide work carried out by me under the supervision of **Dr.J.Jagannathan.**

We further declare that the work reported in this report has not been submitted and will not be submitted, either in part or in full, for any other courses in this institute or any other institute or university.

Place : Vellore

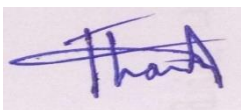
Date : 14/11/2024

A photograph of a handwritten signature in blue ink on a light-colored background. The signature reads "Sridhar S" with a horizontal line under the first name.

Signature of the Candidate 1 (Sridhar S - 21MIS0140)

A photograph of a handwritten signature in blue ink on a light-colored background. The signature reads "S. Kaushik" with a horizontal line under the name.

Signature of the Candidate 2 (Kaushik S - 21MIS0362)

A photograph of a handwritten signature in blue ink on a light-colored background. The signature reads "Tharun S" with a horizontal line under the name.

Signature of the Candidate 3 (Tharun S - 21MIS0408)

Abstract

Diabetes is a serious chronic disease affecting millions of people worldwide. It is a metabolic condition characterized by high blood sugar levels. Untreated diabetes can lead to severe complications, including heart disease, stroke, kidney damage, blindness, and limb amputation. Early prevention and management of diabetes are crucial for reducing the risk of complications. Machine learning can be a powerful tool for predicting an individual's risk of diabetes. This project aims to explore various datasets related to diabetes and develop a model to predict diabetes in individuals using a dataset of patients with and without diabetes. XGboost machine learning algorithm has Accuracy of 97.105% and used to evaluate and accurately predict diabetes in new patients.

Keywords: Medical diagnostics, Real-time predictive tools, Clinical tests, Machine Learning, Diabetes prediction, Diabetes data

Introduction

Diabetes is marked by persistently elevated blood sugar levels. Without proper treatment, it can lead to severe complications, including heart disease, stroke, kidney damage, blindness, and limb amputation. Early detection and effective management are essential to mitigating these risks. Although medical diagnostics have advanced significantly, early detection of diabetes remains challenging due to the limited availability of accessible, real-time predictive tools. Traditional diagnostic methods rely heavily on clinical tests, which may not always be readily available or efficient in identifying individuals at risk before significant symptoms develop.

Machine Learning can aid in analyze large datasets and identify patterns, offers a promising alternative for early diabetes prediction. The goal of this project is to analyze diabetes data and develop a machine learning model capable of accurately predicting an individual's risk of diabetes using a dataset that includes both diabetic and non-diabetic patients. The model will be designed to predict possibility of diabetes, which can aid healthcare professionals in proactive diabetes management and intervention. Constraints includes the model needs to perform well across diverse patient demographics and not just on the dataset it was trained on. Training complex machine learning models can be resource-intensive, requiring adequate computational power and storage. model must adhere to medical data privacy regulations and ensure that predictions are used ethically in clinical settings.

This project seeks to create a machine learning model to predict diabetes risk early, leveraging data for accurate risk assessment and aiding proactive healthcare, with considerations for diverse populations, computational efficiency, and compliance with ethical and privacy guidelines.

Objective

The objective is to explore medical dataset related to diabetes and develop a model that can accurately predict an individual's risk of diabetes based on a dataset containing both diabetic and non-diabetic patients. This predictive tool aims to support healthcare professionals in proactive diabetes management and early intervention by offering an accessible, real-time alternative to traditional diagnostic methods. The model should perform effectively across diverse patient demographics, comply with medical data privacy regulations, and be used ethically in clinical settings.

This work focuses on creating a diabetes risk prediction model from medical data to aid healthcare professionals in proactive management and early intervention, ensuring accessibility, demographic inclusivity, and adherence to privacy and ethical standards.

Problem statement

Diabetes is a major public health issue characterized by high blood sugar levels, leading to various complications such as cardiovascular diseases, kidney failure, and neuropathy. Despite the prevalence of diabetes, many individuals remain undiagnosed until they experience severe health issues. This project focuses on developing a machine learning-based predictive model that identifies individuals at risk of developing diabetes using a dataset containing crucial health indicators. By analyzing features such as age, BMI, blood pressure, glucose levels, and lifestyle factors, the model aims to provide timely and accurate predictions of diabetes risk. The project will explore various machine learning algorithms, including logistic regression, decision trees, and ensemble methods, to determine the most effective approach for early detection. Ultimately, this predictive tool aims to assist healthcare providers in implementing preventative strategies, thereby reducing the incidence of diabetes and improving patient outcomes.

This project aims to create a machine learning model that analyzes health indicators to predict diabetes risk early, supporting healthcare providers in implementing preventive strategies to improve patient outcomes and reduce diabetes incidence.

Literature Review

This study focuses on early diabetes prediction using a variety of machine learning algorithms, including Decision Trees, Support Vector Machines (SVM), and Random Forests. The authors emphasize the importance of feature selection and ensemble approaches to improve prediction accuracy. Their findings suggest that a combination of these techniques can significantly enhance the early detection of diabetes, which is crucial for timely intervention and management[1]. This research introduces a web-based interface for real-time diabetes prediction using machine learning models like Logistic Regression, Naive Bayes, and k-Nearest Neighbors (k-NN). The authors

highlight the system's user-friendliness and its potential to make healthcare more accessible by providing instant predictions based on user input. The study emphasizes the integration of technology into healthcare to improve patient outcomes[2]. This paper explores the potential of machine learning in revolutionizing healthcare, particularly in diabetes prediction. The authors compare various algorithms, including Neural Networks and Gradient Boosting Machines (GBM), and discuss their potential to provide more accurate and timely predictions. The study also touches on the broader implications of machine learning in healthcare, including ethical considerations and the future of predictive analytics[3]. This paper presents a comparative analysis between traditional machine learning algorithms, such as Random Forest and SVM, and deep learning models like Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks for diabetes prediction. The study evaluates these models based on accuracy, precision, and computational efficiency, finding that deep learning models often outperform traditional ones but at the cost of higher computational requirements[4]. This study focuses on predicting the early onset of diabetes using machine learning algorithms, such as Decision Trees, SVM, and k-NN. The research highlights the importance of preprocessing steps, like feature scaling and normalization, to enhance the performance of these models. The study concludes that early detection is vital in managing diabetes, and machine learning techniques can significantly contribute to this goal[5]. This paper introduces a long-term diabetes prediction model based on machine learning techniques and general medical check-up data collected over several years. The author uses algorithms like Random Forest, XGBoost, and Support Vector Regression (SVR) to predict diabetes progression over time. The study emphasizes the importance of continuous monitoring and longitudinal analysis to improve healthcare management and delay the onset of diabetes[6]. This research provides a comparative analysis of classification techniques used for diabetes diagnosis, specifically using the PIMA Indian diabetes dataset. The study compares algorithms such as Decision Trees, SVM, Naive Bayes, and Logistic Regression, focusing on their accuracy, sensitivity, and specificity. The authors highlight the effectiveness of data mining in improving the accuracy of diabetes diagnosis and patient outcomes[7]. This paper explores the use of machine learning and data analytics in the detection of Parkinson's disease. The authors apply techniques such as Decision Trees, SVM, and Neural Networks to patient data to identify patterns indicative of Parkinson's disease. The study emphasizes the potential of machine learning to improve early detection and diagnosis, thereby enhancing patient management and outcomes. The paper also discusses the broader applicability of these techniques in other neurodegenerative diseases[8]. This research focuses on the real-time analysis of health data collected from IoT-connected wearable devices. The authors discuss how machine learning algorithms can be integrated with wearable technology to provide continuous monitoring and predictive analytics for various health conditions, including diabetes. The study highlights the potential of IoT in transforming healthcare by enabling real-time data collection and analysis, which can lead to timely interventions and improved

patient outcomes[9]. This paper provides a systematic review and analysis of recent advancements in artificial intelligence (AI) algorithms for diabetes prediction. The authors explore various AI techniques, including deep learning and reinforcement learning, and their applications in healthcare .emphasizes the increasing accuracy and precision of AI-driven models in predicting diabetes, discussing their potential to transform healthcare by enabling more personalized and effective treatments. The paper also examines the challenges and future directions for AI in diabetes prediction and healthcare analytics[10].

Data Exploration

Dataset :

The diabetes_prediction_dataset.csv file contains medical and demographic data of patients along with their diabetes status, whether positive or negative. It consists of various features such as age, gender, body mass index (BMI), hypertension, heart disease, smoking history, HbA1c level, and blood glucose level. The Dataset can be utilized for constructing machine learning models that can predict the likelihood of diabetes in patients based on their medical history and demographic details.

| | |
|-----------------|---|
| gender | the biological sex of the individual. (male, female, other) |
| age | age of the individual,which ranges from 0 to 80. |
| hypertension | a medical condition in which the blood pressure in the arteries is persistently elevated. (0: no hypertension 1: has hypertension) |
| heart_disease | a medical condition that is associated with an increased risk of developing diabetes. (0: no heart disease 1: has heart disease) |
| smoking_history | a risk factor for diabetes and can exacerbate the complications associated with diabetes. (not current, former, No Info, current, never, and ever) |
| bmi | BMI (Body Mass Index) is a measure of body fat based on weight and height. |

| | |
|---------------------|--|
| HbA1c_level | a measure of a person's average blood sugar level over the past 2-3 months. Mostly more than 6.5% of HbA1c Level indicates diabetes. |
| blood_glucose_level | the amount of glucose in the bloodstream at a given time. |
| diabetes (Target) | 1 indicating the presence of diabetes 0 indicating the absence of diabetes. |

Methods

The methodology explores the diabetes dataset, trains and compares various models, leverages best algorithm to predict diabetes using the given features from the diabetes_prediction_dataset.csv file, which contains both medical and demographic data such as gender, age, BMI, hypertension, heart disease, smoking history, HbA1c levels, and blood glucose levels.

i. Data Collection and Preprocessing:

Load the dataset diabetes_prediction_dataset.csv which contains various features relevant to diabetes diagnosis.

Data Cleaning: Handle missing values in the dataset, especially in the smoking_history feature, which includes "No Info" . Standardize the feature names and ensure consistency in categorical data

Categorical Encoding: Convert categorical variables like gender (male, female, other) and smoking_history (not current, former, current, never, ever) into numerical form using techniques such as one-hot encoding.

Scaling: Normalize numerical features like age, BMI, HbA1c_level, and blood_glucose_level to ensure they are on the same scale, which helps the model in optimization.

Train-Test Split: Split the dataset into a training set (80%) and a testing set (20%) for model evaluation.

ii. Feature Engineering:

Create interaction terms between relevant features, such as combining age and HbA1c_level, or heart_disease and hypertension, to potentially improve model accuracy.

Feature Selection: Use feature importance techniques or correlation analysis to filter out less relevant features and focus on the ones that contribute most to the prediction of diabetes.

iii. Model Selection and Training:

Use XGBoost, which is known for its high efficiency and accuracy in handling large datasets, complex relationships, and feature importance evaluation.

Hyperparameter Tuning:

- `learning_rate`: Controls the update step size.
- `n_estimators`: The number of trees in the ensemble.
- `max_depth`: Maximum depth of each tree.
- `gamma`, `lambda`, and `alpha`: Regularization parameters to prevent overfitting.
- Cross-Validation: Apply cross-validation techniques like k-fold cross-validation to validate model performance and avoid overfitting.

iv. Model Evaluation:

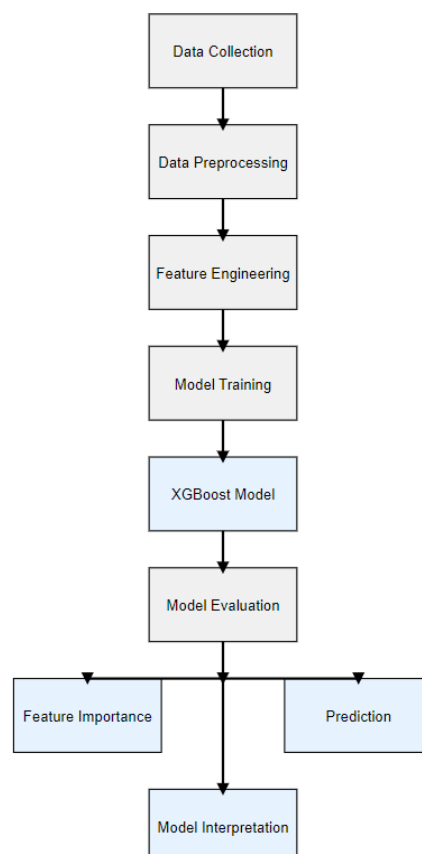
Evaluation Metrics:

- Accuracy: Overall performance of the model.
- Precision, Recall, F1-Score: Especially important for dealing with imbalanced datasets where the number of diabetic and non-diabetic patients might differ.
- AUC-ROC Curve: To evaluate the model's ability to discriminate between diabetic and non-diabetic patients.
- Confusion Matrix: Use a confusion matrix to visualize model performance in terms of true positives, true negatives, false positives, and false negatives.

v. Real-Time Predictions:

Integrate the model with patient data to continuously monitor and predict the likelihood of diabetes based on updated medical information.

This approach aims to build a robust, efficient, and interpretable diabetes prediction system that leverages demographic and medical data for early diagnosis and personalized treatment strategies.

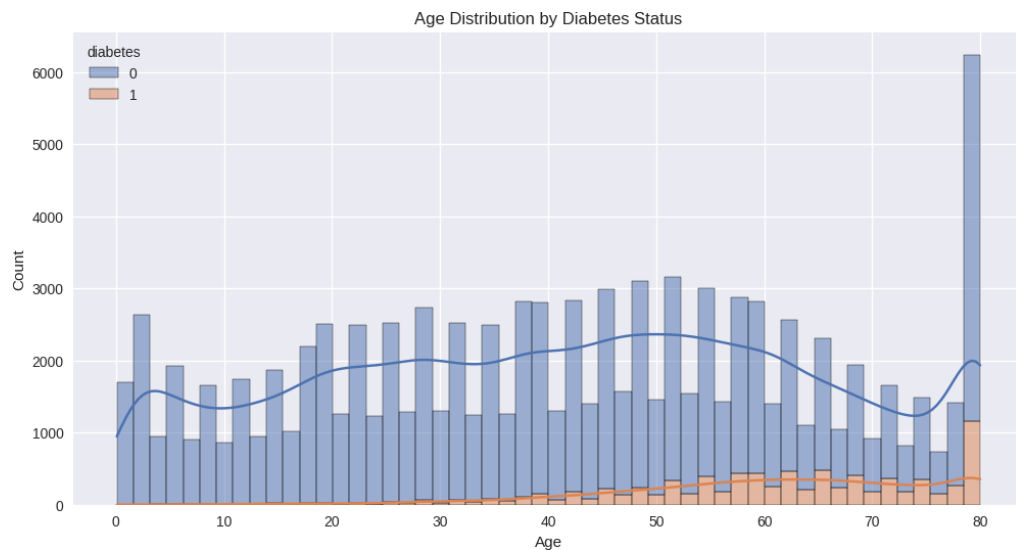


Modeling and Result

1. Age Distribution

Diabetes is more prevalent among older age groups, with a significant increase in cases starting around age 40.

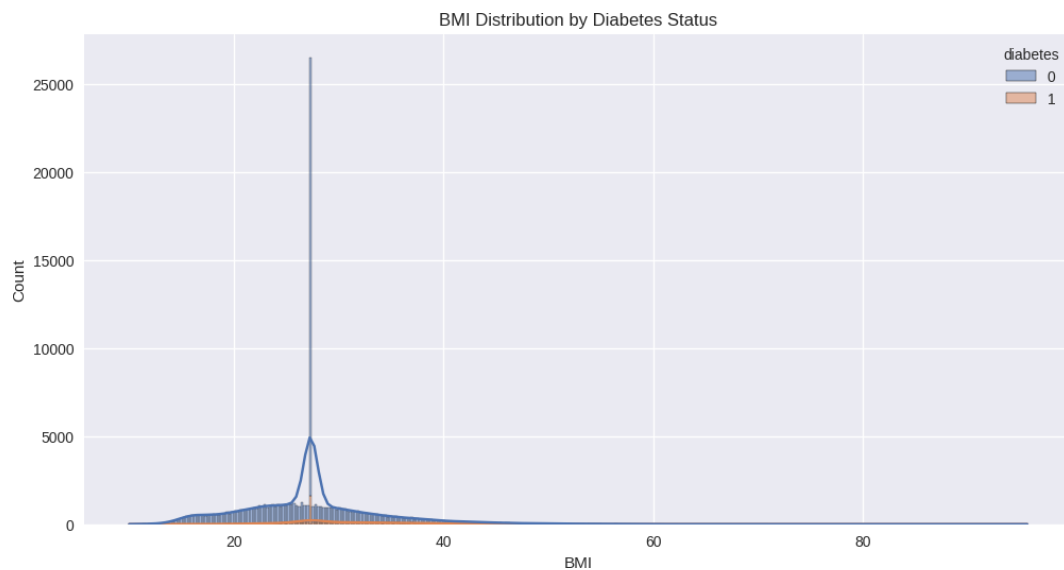
This trend suggests age as an important factor to consider when assessing diabetes risk.



2. BMI Distribution

Individuals with higher BMI levels have a greater prevalence of diabetes.

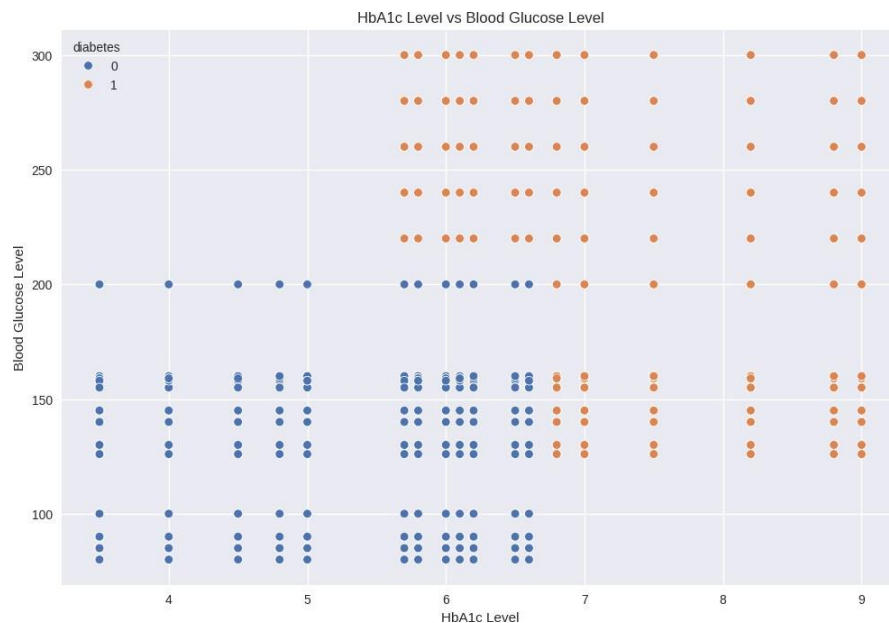
The distribution for diabetic individuals is shifted to the right compared to non-diabetic individuals, indicating a positive correlation between BMI and diabetes.



3. HbA1c Level vs Blood Glucose Level

There is a strong positive correlation between HbA1c levels and blood glucose levels.

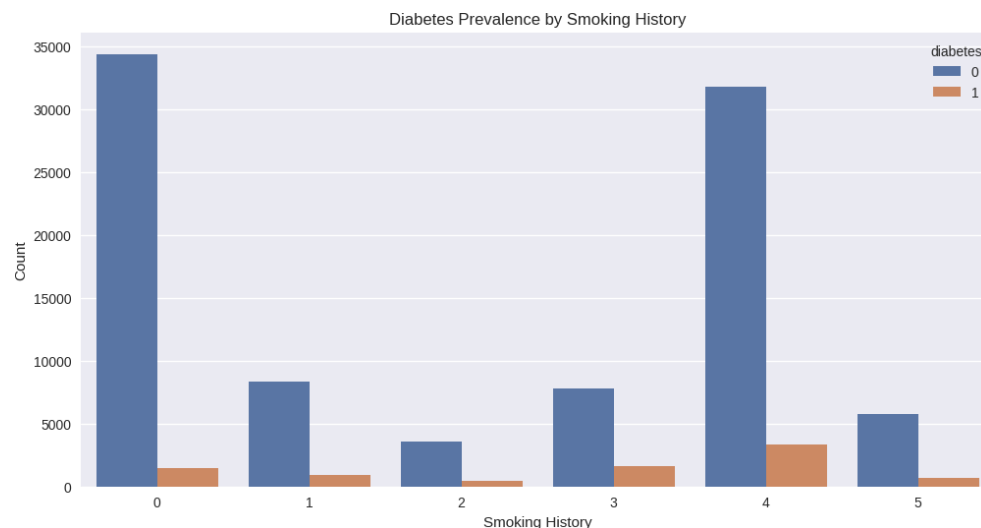
Diabetic individuals tend to have higher values for both measures, creating a clear distinction between diabetic and non-diabetic groups.



4. Smoking History and Diabetes

Diabetes prevalence varies across different smoking history categories.

However, the relationship between smoking history and diabetes is not straightforward, suggesting that other factors may have a more significant impact.



5. Age vs. BMI with Diabetes

A general trend of increasing BMI with age is observed.

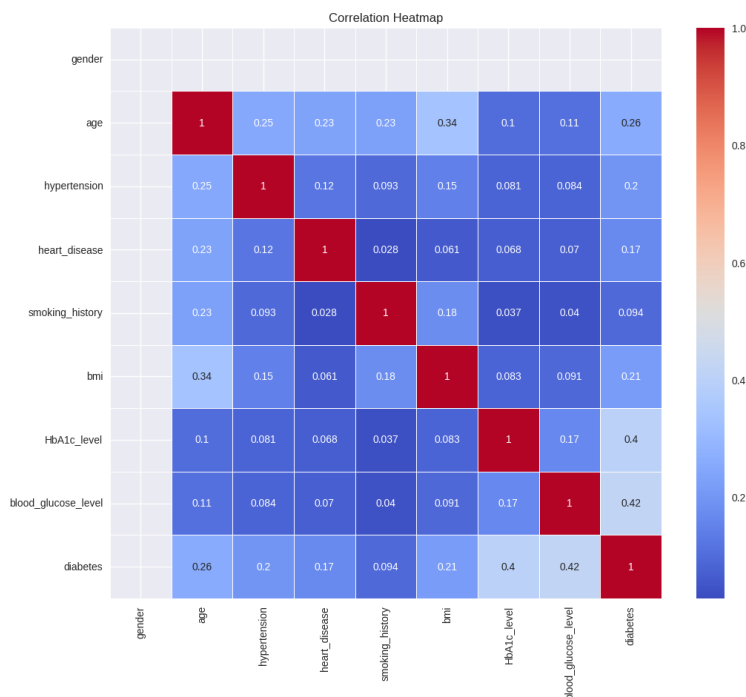
Diabetic individuals are clustered in regions with higher BMI and older age, with larger data points (indicating higher blood glucose levels) more common among diabetic individuals.



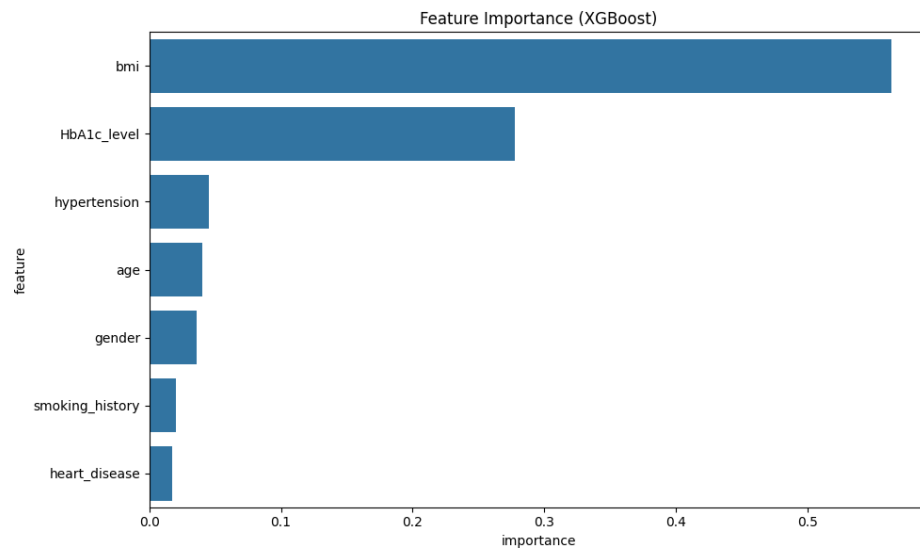
6. Correlation Heatmap

Blood glucose level and HbA1c level show the strongest positive correlation with diabetes, followed by age and BMI, which have moderate positive correlations.

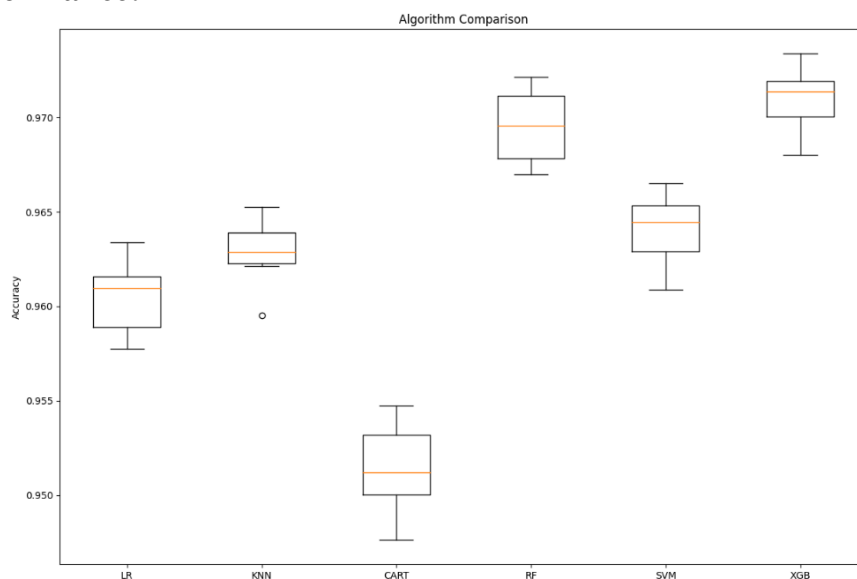
Hypertension and heart disease show weak positive correlations, indicating they may have less impact on diabetes compared to other factors.



The visualizations highlight that bmi, H1bA1c level, hypertension, age are key factors associated with diabetes risk. The complex relationship between these variables underscores the need to consider multiple factors for effective diabetes prediction and management.



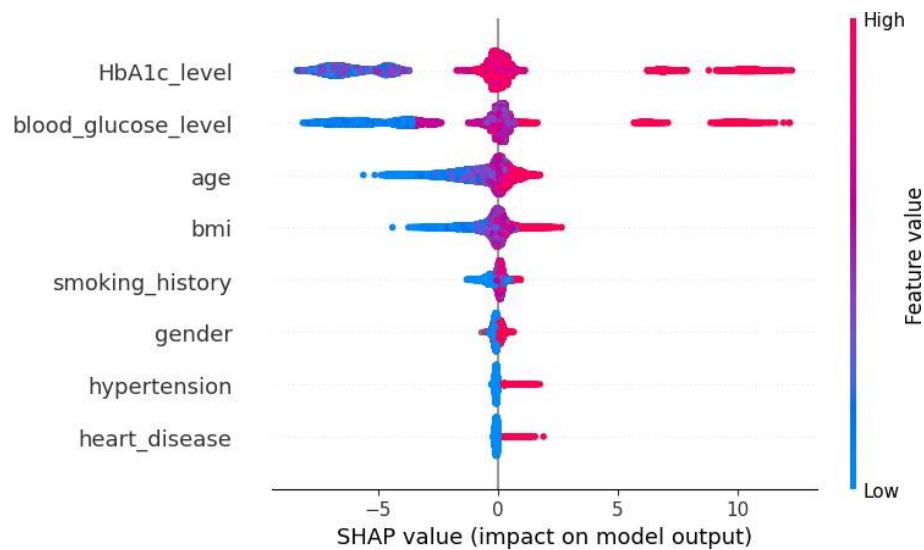
Model Performance:



- Gradient Boosting Classifier: Accuracy of 97.255%
- Decision Tree Classifier: Accuracy of 95.025%
- XGBoost Classifier: Accuracy of 97.105%

All three models (Gradient Boosting, Decision Tree, and XGBoost) performed well, with accuracies above 95%. Gradient Boosting slightly outperformed XGBoost in this case, but the difference is minimal (0.15%). The Decision Tree model, while still accurate, performed slightly worse than the ensemble methods. Model Interpretation

demonstrates the use of various model interpretation techniques, including ELI5 and SHAP, to explain individual predictions and overall model decision-making process.



Conclusion and Future Development

Diabetes prevalence increases with age, particularly from age 40 onward, and is notably higher among individuals with elevated BMI, indicating both as key risk factors. Strong positive correlations are observed between HbA1c and blood glucose levels, clearly distinguishing diabetic from non-diabetic individuals, while smoking history shows a less direct relationship with diabetes risk. Additionally, diabetic individuals tend to cluster in regions of higher age and BMI. The correlation heatmap highlights blood glucose and HbA1c levels as the strongest predictors of diabetes, followed by moderate correlations with age and BMI, with hypertension and heart disease having a weaker impact.

The XGBoost-based diabetes prediction model demonstrates high accuracy in classifying diabetes cases. With an accuracy of over 97%, the model shows promise as a tool for early diabetes detection or risk assessment. The use of advanced interpretation techniques like SHAP values adds transparency to the model's decision-making process, which is crucial for building trust in healthcare applications.

Future work could focus on:

1. Fine-tuning hyperparameters to potentially improve model performance further.
2. Investigating the most important features identified by the model to gain medical insights.
3. Validating the model on external datasets to ensure generalizability.
4. Exploring the potential for deploying this model in clinical settings as a supportive tool for healthcare professionals.

References

- [1] Sinha, R., Vennela, B. S., & Babu, S. (2024). Early diabetes prediction using machine learning algorithms. In 2024 3rd International Conference on Applied Artificial Intelligence and Computing (ICAAIC) (pp. 705-708). IEEE. <https://doi.org/10.1109/ICAAIC60222.2024.10575581>
- [2] Krishna Rao, E. V., Pamula, N., Battula, S., & Guntaka, M. (2024). Web-interfaced diagnosis system of diabetes prediction using machine learning algorithms. In 2024 International Conference on Smart Systems for Applications in Electrical Sciences (ICSSSES) (pp. 1-6). IEEE. <https://doi.org/10.1109/ICSSSES62373.2024.10561359>
- [3] Gupta, P., Verma, B., Pawar, M., & Gupta, A. (2024). Diabetes prediction using machine learning: A game-changer for healthcare. In 2024 5th International Conference on Recent Trends in Computer Science and Technology (ICRTCST) (pp. 122-127). IEEE. <https://doi.org/10.1109/ICRTCST61793.2024.10578448>
- [4] Tripathy, N., Moharana, B., Balabantaray, S. K., Nayak, S. K., Pati, A., & Panigrahi, A. (2024). A comparative analysis of diabetes prediction using machine learning and deep learning algorithms in healthcare. In 2024 International Conference on Advancements in Smart, Secure and Intelligent Computing (ASSIC) (pp. 1-6). IEEE. <https://doi.org/10.1109/ASSIC60049.2024.10508008>
- [5] Reza, M. S., Fakir, A. R. S., Islam, M. R., Alom, A. M. T., Sen, T., & Islam, M. S. (2023). Early stage diabetes prediction using machine learning techniques. In 2023 26th International Conference on Computer and Information Technology (ICCIT) (pp. 1-6). IEEE. <https://doi.org/10.1109/ICCIT60459.2023.10441427>
- [6] Gozali, A. A. (2023). Multi-years diabetes prediction using machine learning and general check-up dataset. In 2023 11th International Conference on Information and Communication Technology (ICoICT) (pp. 98-103). IEEE. <https://doi.org/10.1109/ICoICT58202.2023.10262699>
- [7] Verma, P., & Khatoon, A. (2024). Data mining applications in healthcare: A comparative analysis of classification techniques for diabetes diagnosis using the PIMA Indian diabetes dataset. In 2024 4th International Conference on Innovative Practices in Technology and Management (ICIPTM) (pp. 1-5). IEEE. <https://doi.org/10.1109/ICIPTM59628.2024.10563296>
- [8] Shetty, M., Shetty, S. B., Jambha, H. V. G., & Hrithvika. (2024). Application of machine learning and data analytics in detection of Parkinson's disease. In 2024 Second International Conference on Data Science and Information System (ICDSIS) (pp. 1-5). IEEE. <https://doi.org/10.1109/ICDSIS61070.2024.10594328>
- [9] G, J. D., Venkatakirana, S., Prasad, V. N., M, S. P., Prasad, K., & Kawale, S. R. (2024). Real-time health data analytics in IoT-connected wearable devices. In 2024

5th International Conference for Emerging Technology (INCET) (pp. 1-8). IEEE.
<https://doi.org/10.1109/INCET61516.2024.10593176>

[10] Zakizadeh, M., & Goundar, S. (2024). Advancements in artificial intelligence algorithms for precise diabetes prediction and analysis in the healthcare landscape: A systematic and analytical investigation. In 2024 20th CSI International Symposium on Artificial Intelligence and Signal Processing (AISP) (pp. 1-8). IEEE.
<https://doi.org/10.1109/AISP61396.2024.10475207>