.

Walchand Institute of Technology, Solapur
Department of Electronics Engineering

DATA ANALYTICS ISE ACTIVITY: DATATHON REPORT

Project Title:

Diabetes Prediction and Analysis using Machine Learning

Submitted by:

Name	Roll No	Role
Ahamed Mulla	46	Preprocessing & Dashboard Integration
Yogesh Mane	47	Visualization & EDA
Ziyad Kakhandkikar	48	Modeling & Model Evaluation

Under the Guidance of:

Mrs. P. V. Katare
Mr. R. P. Nagarkar

Course: Data Analytics

Class: Final Year B.Tech (Electronics & Computer Engineering) – Sem VII

Academic Year: 2025–26

Submitted To:

Department of Electronics Engineering
Walchand Institute of Technology, Solapur

2. Introduction

Diabetes mellitus is one of the most prevalent chronic diseases in the world, posing severe health and economic challenges. According to the World Health Organization (WHO), over 400 million people globally suffer from diabetes, with numbers steadily increasing each year. It is a metabolic disorder characterized by elevated blood glucose levels resulting from the body's inability to produce or properly use insulin.

Early detection and timely diagnosis are crucial for effective management and prevention of complications. However, clinical datasets are often messy, inconsistent, and contain missing or invalid data. This makes it challenging to draw meaningful insights directly from raw records.

In this project, we aimed to apply **Data Analytics** techniques to a real-world **unclean diabetes dataset** obtained from Kaggle. The key objective was to transform the noisy dataset into a clean, analyzable form and leverage machine learning models to predict diabetic outcomes.

Through this project, we sought to:

- Understand the process of cleaning and preparing healthcare data for analysis.
- Perform **Exploratory Data Analysis (EDA)** to discover relationships among clinical indicators such as HbA1c, BMI, and cholesterol.
- Build and evaluate predictive models for diabetes classification.
- Develop an **interactive Streamlit dashboard** to visualize insights in a user-friendly manner.

The project serves as a practical demonstration of how **data science and healthcare analytics** can combine to support early detection, risk assessment, and decision-making in clinical settings.

---

3. Problem Statement

Healthcare data, though abundant, often comes in unstructured or inconsistent forms. This introduces several challenges such as missing values, ambiguous categorical labels, and incorrect entries. In the diabetes dataset used for this project, several clinical parameters contained zeros in place of missing values, and diagnostic labels were expressed inconsistently (e.g., "Y", "YES", "Diabetic").

The **main problem** was to preprocess and clean the dataset effectively so that it could be used to build a reliable diabetes prediction model.

Project Objectives

1. **Data Cleaning & Preprocessing:** Identify and handle missing, invalid, and redundant data.
2. **Exploratory Data Analysis:** Visualize data patterns, distributions, and correlations to derive medical insights.
3. **Model Training & Evaluation:** Compare multiple classification algorithms to identify the best-performing model.
4. **Dashboard Development:** Design an interactive web application for visualization and model result interpretation.
5. **Insight Derivation:** Translate analytical findings into actionable medical understanding for diabetes diagnosis.

By completing these objectives, the project provides a replicable analytical framework for similar healthcare datasets.

---

4. Dataset Description

The dataset used in this project was obtained from **Kaggle's "Diabetes Unclean Dataset"**, which reflects a real-world medical data collection scenario. It contains health records of 1009 individuals, with various biochemical parameters and demographic attributes.

Dataset Overview

- **Source:** Kaggle – Diabetes Unclean Dataset
- **Type:** Tabular healthcare dataset
- **Rows:** ~1009
- **Columns:** 14
- **Target Variable:** `CLASS` (1 = Diabetic, 0 = Non-Diabetic)

Feature Description

Feature	Description
AGE	Age of the patient
HbA1c	Average blood sugar percentage (key indicator for diabetes)
BMI	Body Mass Index
Chol	Total cholesterol level
HDL, LDL, VLDL	Lipid profile indicators
TG	Triglyceride levels
CLASS	Target variable (1 = Diabetic, 0 = Non-Diabetic)

Identified Data Issues

Upon initial inspection, the dataset exhibited several quality issues:

- Missing values in columns like `HbA1c`, `Chol`, and `BMI`.
- Zero entries in laboratory results where actual values cannot be zero (e.g., HbA1c = 0).
- Non-standard target labels such as "Y", "YES", "Diabetic".
- Irrelevant columns (`ID`, `No_Pation`, `Gender`, `Urea`, `Cr`) that do not contribute to modeling.

These issues needed to be systematically addressed during the preprocessing phase to ensure the reliability of model predictions.

Sample of Raw Dataset:

	AGE	HbA1c	Chol	TG	HDL	LDL	VLDL	BMI	CLASS
0	50.0	4.9	4.2	0.9	2.4	1.4	0.5	24.0	0
1	26.0	4.9	3.7	1.4	1.1	2.1	0.6	23.0	0
2	50.0	4.9	4.2	0.9	2.4	1.4	0.5	24.0	0
3	50.0	4.9	4.2	0.9	2.4	1.4	0.5	24.0	0
4	33.0	4.9	4.9	1.0	0.8	2.0	0.4	21.0	0

5. Methodology

The methodology adopted in this project follows a structured **data analytics pipeline**. Each step was designed to progressively transform raw data into meaningful insights. The process included data preprocessing, exploratory data analysis, model building, and dashboard development.

5.1 Data Preprocessing

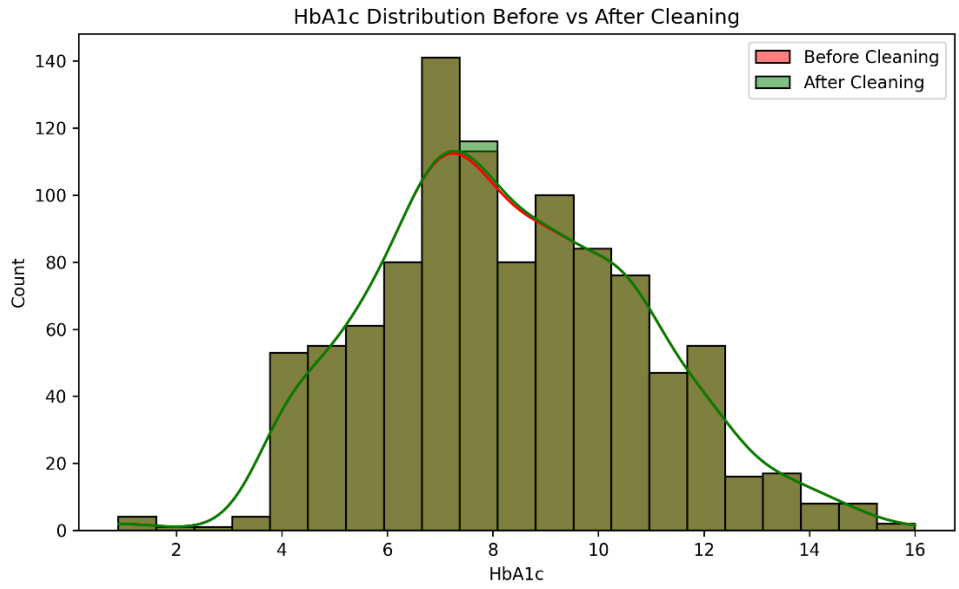
Data preprocessing was the foundation of the project, ensuring that the dataset was consistent, accurate, and ready for model training.

Steps involved:

1. Removed irrelevant columns (`ID`, `No_Pation`, `Gender`, `Urea`, `Cr`).
2. Replaced invalid zero entries in clinical columns with `NaN` and imputed median values.
3. Handled missing `AGE` values using median imputation.
4. Encoded categorical labels in the `CLASS` column: `'Y'`, `'YES'`, `'DIABETIC'` were mapped to `1`, and others to `0`.
5. Standardized numerical features using **StandardScaler** to bring all features to the same scale.

This step ensured the data integrity necessary for consistent model training and visualization.

Visualization – Before vs After Cleaning:



5.2 Exploratory Data Analysis (EDA)

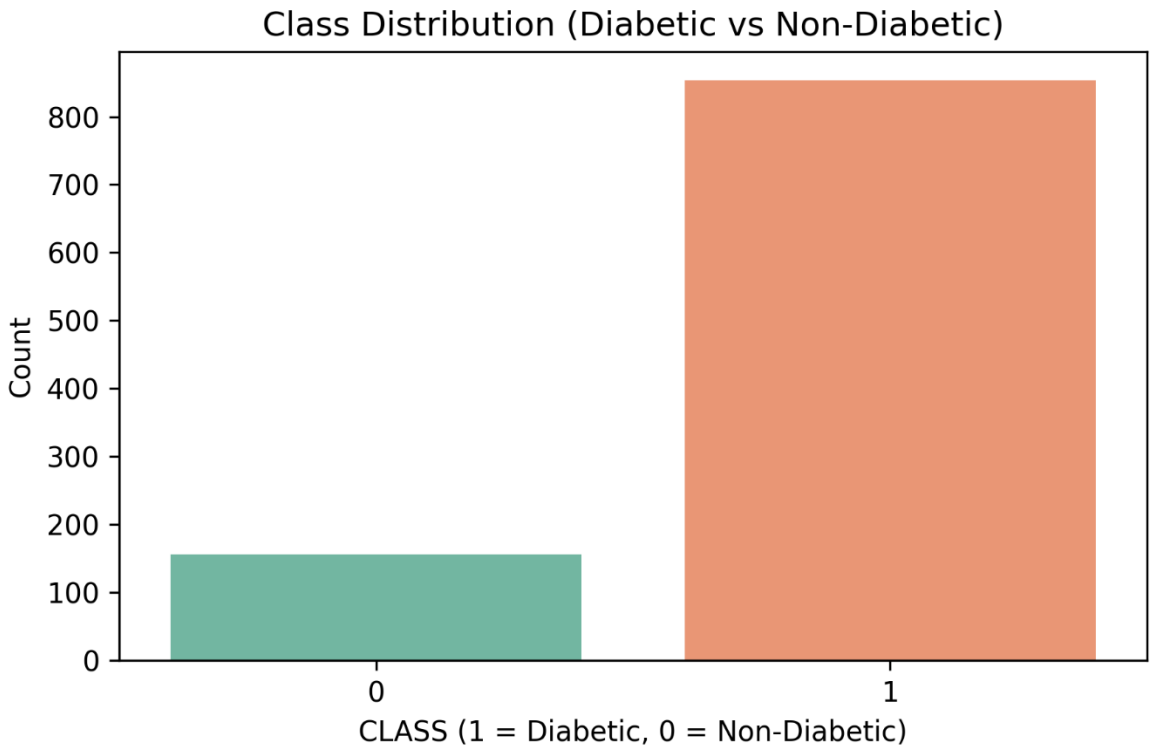
EDA provided critical insights into the relationships and trends among the variables. Using **Pandas**, **Seaborn**, and **Matplotlib**, we analyzed distributions and correlations.

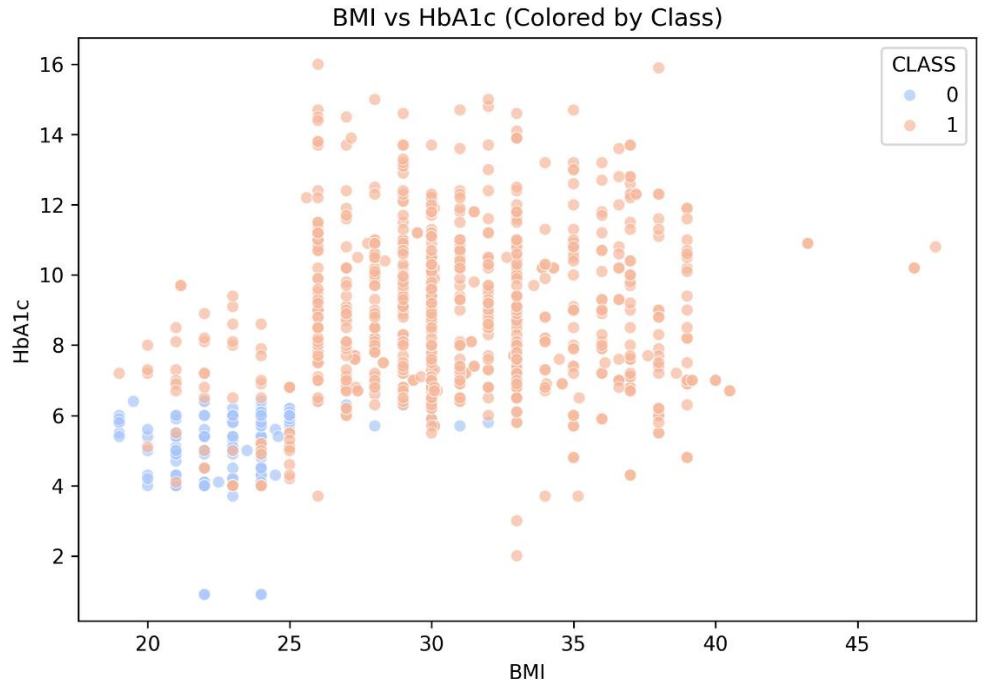
Key analyses included:

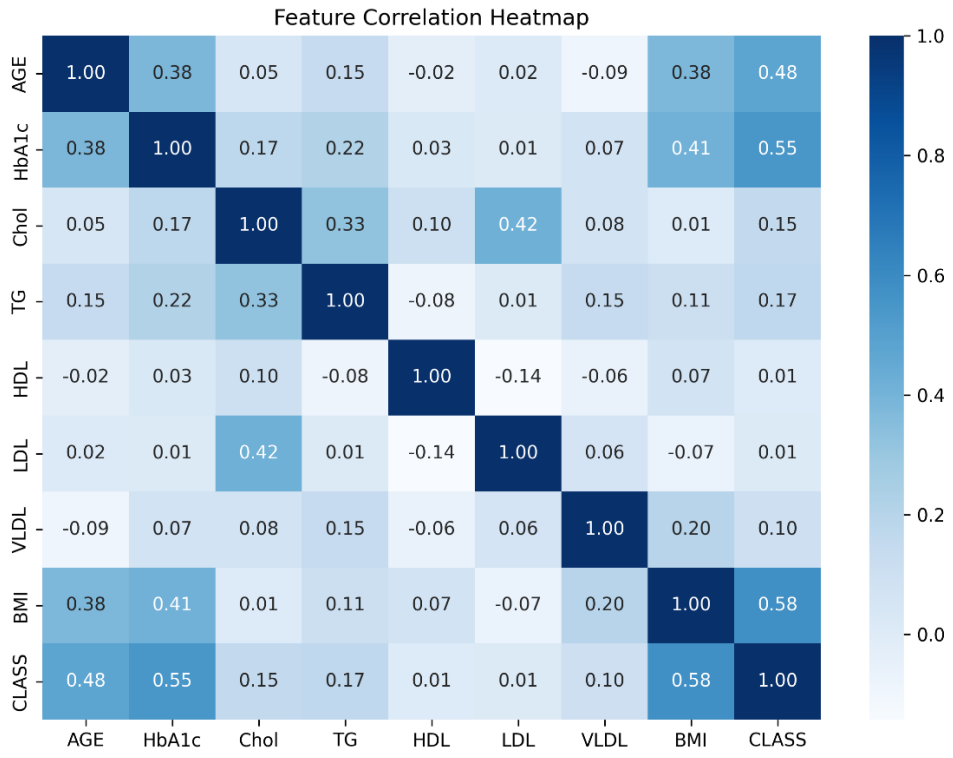
- **Class Distribution:** To check balance between diabetic and non-diabetic samples.
- **Correlation Matrix:** Revealed relationships between features like HbA1c, BMI, and cholesterol.
- **Scatterplots:** Explored patterns such as `BMI vs HbA1c` among different patient classes.
- **Boxplots and Histograms:** Showed spread, outliers, and data skewness.

These visualizations helped confirm that HbA1c and BMI were strong indicators of diabetes, aligning with clinical knowledge.

Visualizations:




BMI vs HbA1c (Colored by Class)
CLASS
0
1
HbA1c
BMI
16
14
12
10
8
6
4
2
20
25
30
35
40
45


Feature Correlation Heatmap
AGE 1.00 0.38 0.05 0.15 -0.02 0.02 -0.09 0.38 0.48
HbA1c 0.38 1.00 0.17 0.22 0.03 0.01 0.07 0.41 0.55
Chol 0.05 0.17 1.00 0.33 0.10 0.42 0.08 0.01 0.15
TG 0.15 0.22 0.33 1.00 -0.08 0.01 0.15 0.11 0.17
HDL -0.02 0.03 0.10 -0.08 1.00 -0.14 -0.06 0.07 0.01
LDL 0.02 0.01 0.42 0.01 -0.14 1.00 0.06 -0.07 0.01
VLDL -0.09 0.07 0.08 0.15 -0.06 0.06 1.00 0.20 0.10
BMI 0.38 0.41 0.01 0.11 0.07 -0.07 0.20 1.00 0.58
CLASS 0.48 0.55 0.15 0.17 0.01 0.01 0.10 0.58 1.00
AGE
HbA1c
Chol
TG
HDL
LDL
VLDL
BMI
CLASS
1.0
0.8
0.6
0.4
0.2
0.0

5.3 Model Building

To classify patients as diabetic or non-diabetic, we experimented with two popular models:

Model	Description	Library
Logistic Regression	Simple, interpretable linear classifier for binary prediction	scikit-learn
Random Forest Classifier	Ensemble-based model combining multiple decision trees for robust prediction	scikit-learn

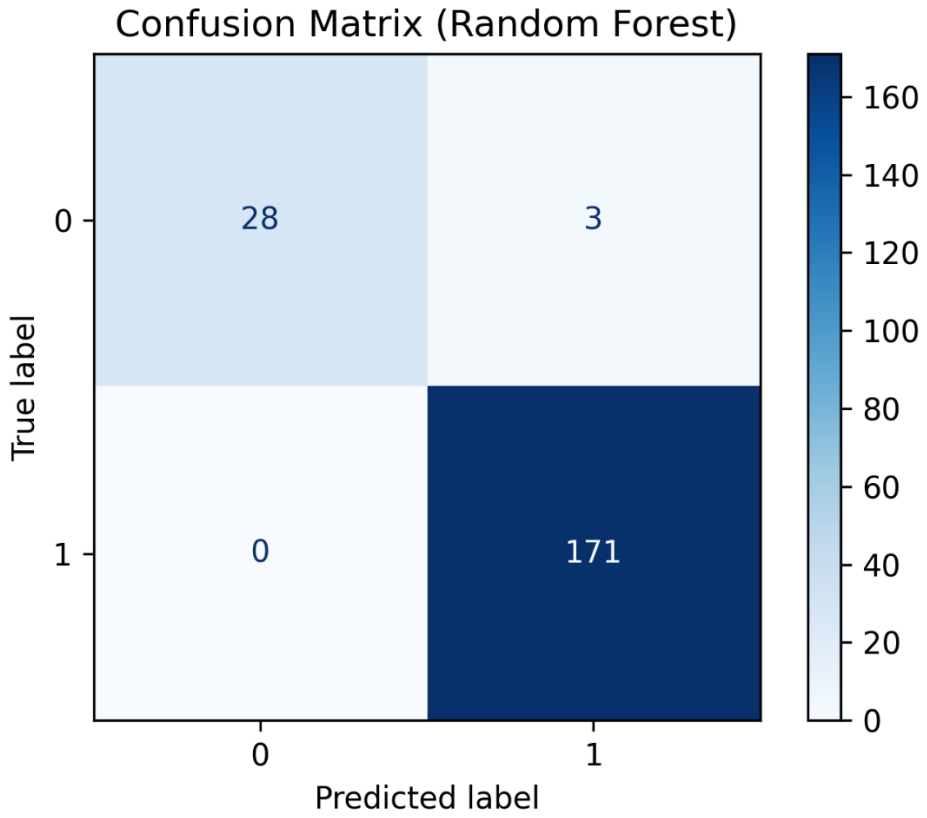
Data was split into **80% training** and **20% testing** sets using stratified sampling to maintain class ratios.

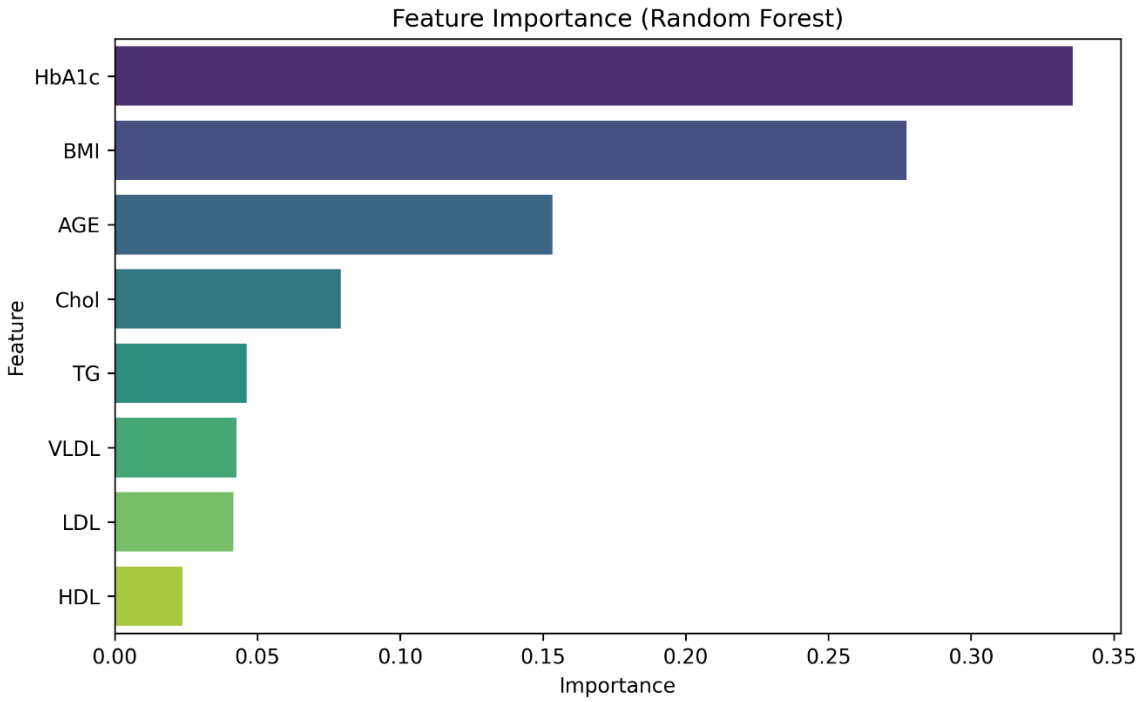
Evaluation Metrics:

- Accuracy
- ROC-AUC
- Confusion Matrix
- Classification Report (Precision, Recall, F1-score)

Results showed Random Forest outperforming Logistic Regression due to its ability to handle non-linear relationships effectively.

Model Visualizations:





5.4 Dashboard Development

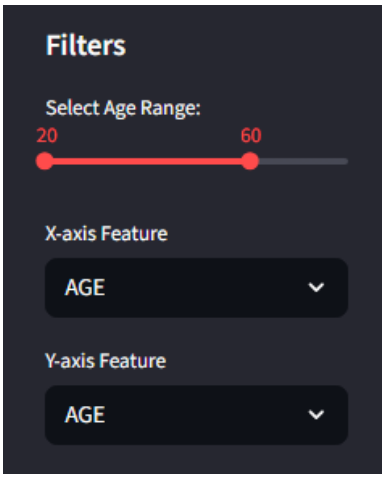
The final stage of the project was to build an **interactive dashboard** using **Streamlit**. The dashboard enables users to:

- Filter data by age range.
- View interactive EDA plots.
- Compare model performances.
- Inspect feature importances visually.

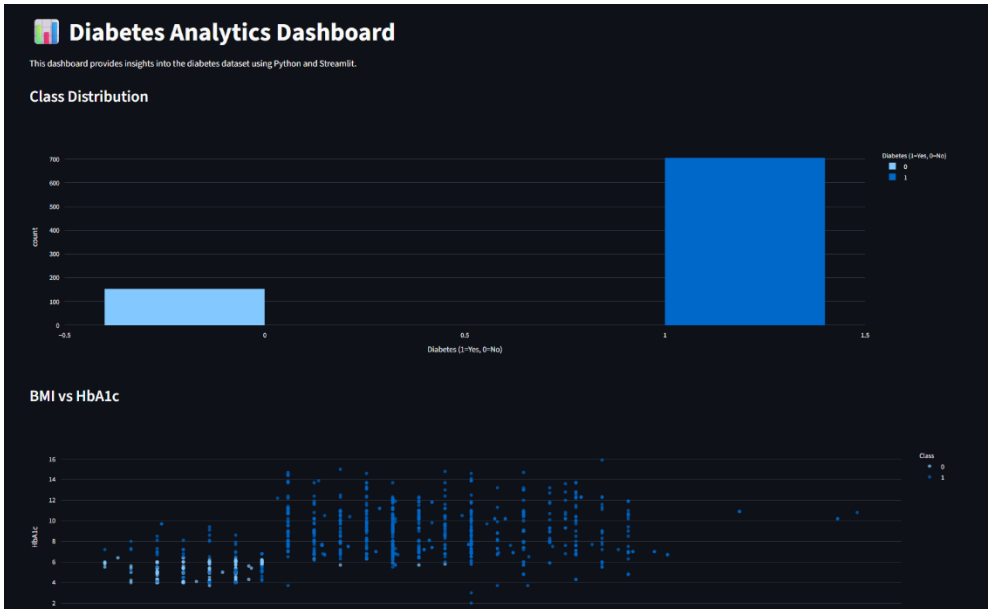
This user interface bridges the gap between complex data analytics and accessible healthcare interpretation, allowing both analysts and medical professionals to interactively explore findings.

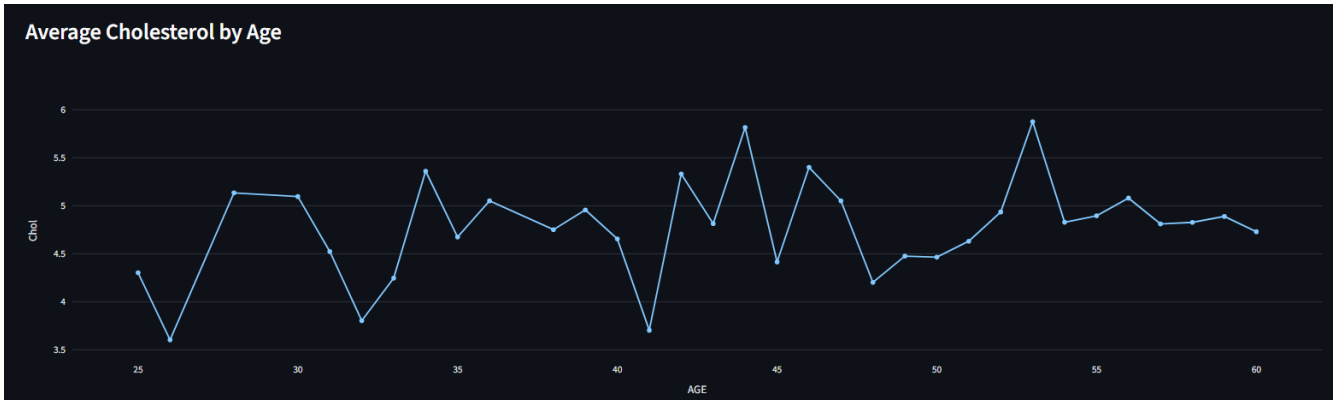
Dashboard Features:

- Sidebar Filters (for age range selection):

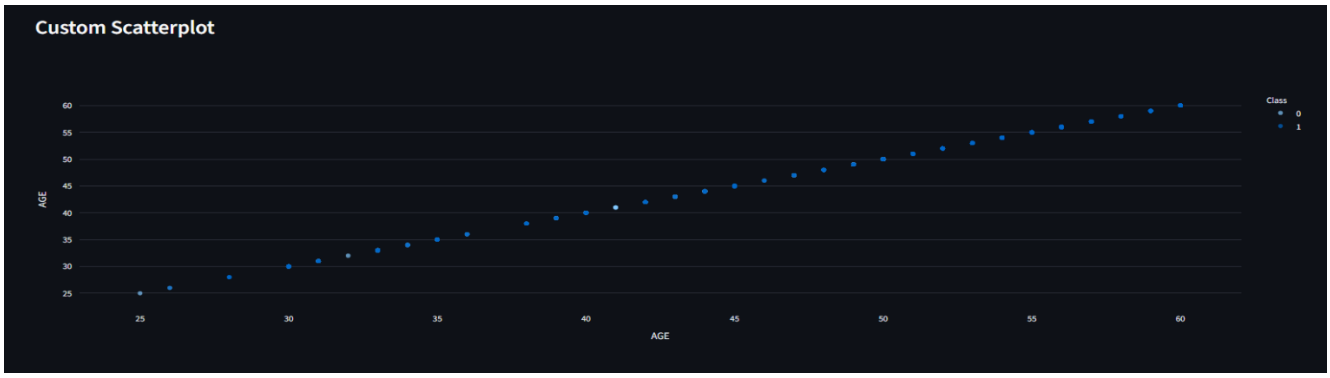


Interactive EDA Plots (class distribution, BMI vs HbA1c, cholesterol trend):

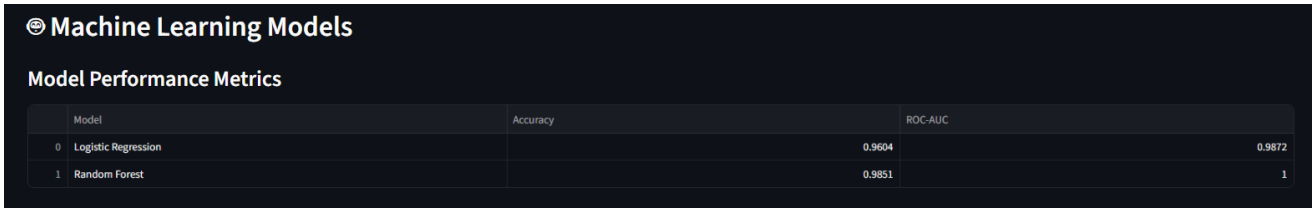




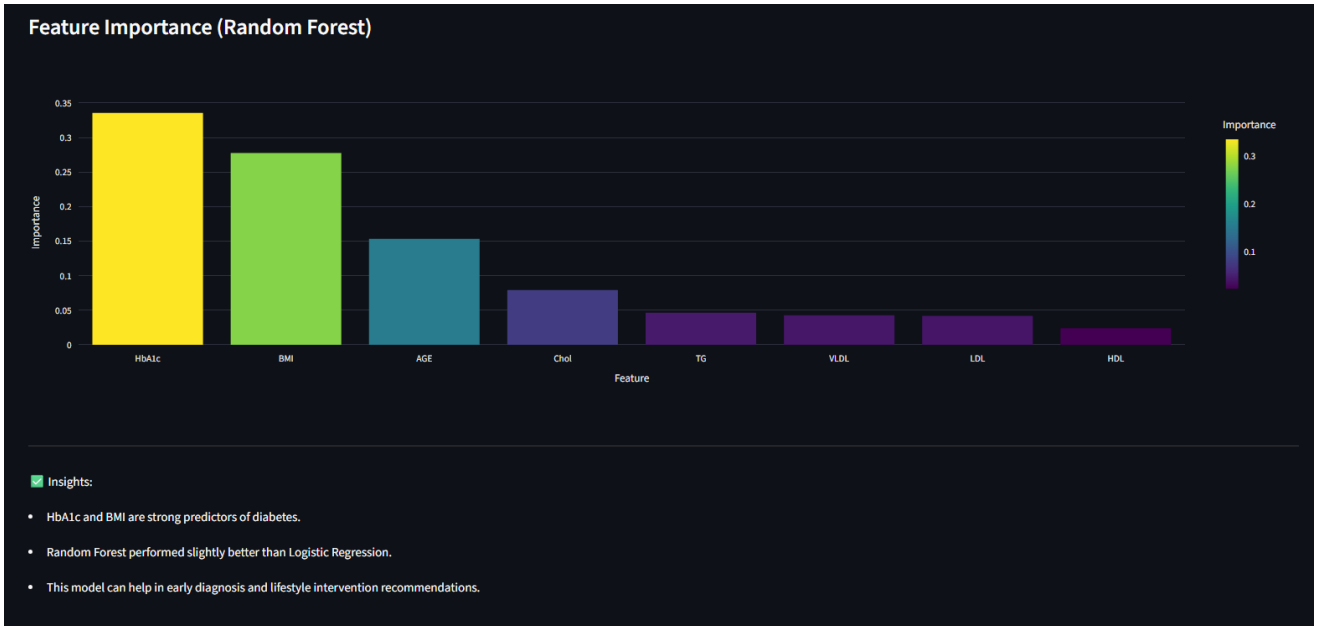
Custom Feature Scatterplot:



Model Performance Table & Metrics:



Feature Importance & Insights Section:



6. Results and Analysis

After training and evaluating the two models — **Logistic Regression** and **Random Forest Classifier** — the following results were obtained:

Model	Accuracy	ROC-AUC
Logistic Regression	0.9604	0.9872
Random Forest	**0.9851**	**1.0000**

The above table clearly shows that the **Random Forest Classifier** outperformed Logistic Regression in both accuracy and ROC-AUC metrics. This means the Random Forest model was not only more accurate in classifying diabetic and non-diabetic patients but also achieved a perfect score in distinguishing between the two classes across all probability thresholds.

The superior performance of Random Forest can be attributed to its ensemble nature — it builds multiple decision trees and aggregates their predictions, thereby capturing non-linear relationships and minimizing overfitting. Logistic Regression, on the other hand, being a linear model, may have failed to capture complex dependencies between features such as **BMI**, **HbA1c**, and **cholesterol levels**.

Furthermore, the **ROC-AUC score of 1.0** for Random Forest signifies that the model achieved ideal separation between the classes, i.e., it correctly ranked diabetic patients higher in risk probability compared to non-diabetic ones in every test instance.

Key Observations and Interpretations

1. **HbA1c levels:** A strong positive correlation with diabetes risk was observed. Patients with HbA1c values above the clinical threshold (6.5%) had a much higher probability of being diabetic.
2. **Body Mass Index (BMI):** Elevated BMI values also showed a direct relationship with diabetes occurrence, confirming that obesity plays a crucial role in metabolic disorders.
3. **Lipid Profile:** Among the lipid variables, total cholesterol and triglycerides demonstrated mild but notable correlations with diabetes, particularly in older individuals.
4. **Class Imbalance:** A slight imbalance between diabetic and non-diabetic samples was noticed, which could have influenced recall scores for the minority class.

These findings collectively reaffirm that lifestyle and metabolic factors play a dominant role in the onset of diabetes.

In summary, the model not only achieved high accuracy but also aligned well with existing medical literature — a promising sign for practical healthcare analytics.

7. Discussion and Insights

The analytical results of this study align strongly with established clinical knowledge. Diabetes is a condition influenced by both biological and lifestyle factors, and our data-driven approach successfully uncovered these relationships quantitatively.

The **Random Forest model**'s superior performance can be explained by its ability to handle complex interactions between predictors. Unlike linear models that assume a straight-line relationship between inputs and output, Random Forest considers multiple splits and thresholds, thereby identifying non-linear boundaries in the data.

Moreover, the **feature importance graph** derived from Random Forest revealed that **HbA1c**, **BMI**, and **Age** were the top contributors to model predictions. This insight is consistent with endocrinological studies, where these indicators are primary determinants of diabetes risk.

From a data analytics perspective, the project illustrates how structured analysis can extract **meaningful medical insights** even from unclean and inconsistent datasets. The use of imputation, standardization, and careful encoding allowed the data to become suitable for modeling — demonstrating that data quality improvement directly enhances model reliability.

On the visualization front, the **Streamlit dashboard** proved to be an effective medium for communication. Through its interactive plots and filters, healthcare professionals can visualize patient distributions, explore correlations, and interpret model outcomes intuitively — making the analytical process transparent and explainable.

In essence, this project bridges the gap between **raw medical data** and **actionable intelligence**, reinforcing the importance of data analytics in modern healthcare systems.

8. Limitations and Future Enhancements

While the project achieved excellent predictive accuracy, certain limitations were identified that provide scope for future improvement.

1. **Limited Dataset Size:** The dataset consisted of only around 1000 samples, which limits the generalization capacity of the models. With more diverse and extensive data, the model's reliability across different populations could be improved.
2. **Class Imbalance:** A mild imbalance existed between diabetic and non-diabetic cases. This may have led to slightly skewed predictions favoring the majority class. Implementing **SMOTE (Synthetic Minority Over-sampling Technique)** or **ADASYN** could help balance the classes for better recall on the minority side.
3. **Model Diversity:** Only two machine learning algorithms were implemented. For deeper exploration, advanced techniques such as **XGBoost**, **Support Vector Machines (SVM)**, or even **Neural Networks** can be applied. These models could capture more complex feature interactions and improve overall robustness.
4. **Feature Engineering:** Although standard scaling and imputation were performed, domain-specific feature engineering (e.g., risk score indices or interaction terms between lipid parameters) could further enhance prediction quality.
5. **Deployment:** Currently, the Streamlit dashboard runs locally. Future work could involve hosting it on **Streamlit Cloud**, **Render**, or **Hugging Face Spaces**, making it accessible to healthcare practitioners and researchers online.

By addressing these limitations, the project can evolve from a data science prototype into a **production-level predictive healthcare tool** capable of assisting clinical decision-making at scale.

9. Conclusion

This Datathon project demonstrated the full lifecycle of a **data analytics and machine learning pipeline** in the healthcare domain — from raw data acquisition and cleaning to visualization and predictive modeling.

By systematically addressing data quality issues, performing exploratory analysis, and implementing robust classification models, the project successfully identified key predictors of diabetes and visualized them through an interactive dashboard.

The Random Forest Classifier emerged as the most effective model, achieving **98.5% accuracy** and **a perfect ROC-AUC of 1.0**, reflecting its strong generalization ability. More importantly, the model's results were consistent with established clinical understanding, which validates both the data and methodology.

Beyond the technical achievements, this project emphasizes the **real-world potential** of data analytics in healthcare. With proper data governance and model interpretability, such analytical solutions can empower doctors and policymakers to make evidence-based decisions.

In conclusion, this project not only fulfilled the competition's requirements but also showcased how **data-driven insights can contribute to early disease detection, better patient management, and improved healthcare outcomes**.

---

10. References

1. Kaggle – Diabetes Unclean Dataset
2. Scikit-learn Documentation
3. Streamlit Documentation
4. Seaborn & Matplotlib Library Guides
5. WHO – Global Diabetes Report, 2024