PREDICTIVE ANALYSIS

PROJECT REPORT

(Project Semester Nov-Dec 2025)

Predictive Modeling for Diabetes Using NHANES Data

Submitted by-

Abhishek kashyap

Registration No. **12311363**

Programme: B.Tech CSE

Section: K23DH

Course Code: INT 234

Under the Guidance of

Dr. Baljinder kaur

UID: 27592

Assistant professor

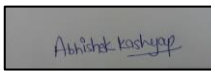
Discipline of CSE/IT

Lovely School of Computer science and Engineering

Lovely Professional University, Phagwara

DECLARATION

I, **Abhishek kashyap** student of **Computer science and Engineering** under CSE/IT Discipline at, Lovely Professional University, Punjab, hereby declare that all the information furnished in this project report is based on my own intensive work and is genuine.

Date: 20-12-2025 Signature: 

Registration No.: 12311363

Name of the student: Abhishek kashyap

CERTIFICATE

This is to certify that Abhishek kashyap bearing Registration no. 12311363 has completed INT 234 project titled, “Machine Learning–Based Prediction of Onion Modal Prices in Nashik District, Maharashtra” under my guidance and supervision. To the best of my knowledge, the present work is the result of his original development, effort and study.

Signature and Name of the Supervisor
Designation of the Supervisor
School of Computer science and Engineering
Lovely Professional University
Phagwara, Punjab.

Date: 20-12-2025

Acknowledgement

I would like to express my sincere gratitude to my faculty, Dr. Baljinder kaur, for her continuous guidance, valuable feedback, and consistent support throughout this project.

I would also like to thank Lovely Professional University for providing the resources and academic environment that allowed me to complete this project successfully.

Last but not least, I extend my thanks to my friends for their encouragement and support during the course of this work.

Abhishek kashyap

12311363

Table of Contents

1. **Introduction**
2. **Source of Dataset**
3. **Dataset Preprocessing**
   - 3.1 Handling Missing Values
   - 3.2 Data Cleaning
   - 3.3 Feature Engineering
   - 3.4 Standardization and Scaling
   - 3.5 Data Splitting for Predictive Modelling
4. **Analysis on Dataset**
   - 4.1 Distribution of Diabetes Cases
   - 4.2 Age Distribution By Diabetes Status
   - 4.3 BMI Distribution by Diabetes Status
   - 4.4 Blood Pressure Analysis (SBP and DBP)
   - 4.5 Haemoglobin Level Comparison
   - 4.6 Gender-wise Diabetes Distribution
   - 4.7 Correlation Analysis of Clinical Features
5. **GitHub & Linkeldn Links**
6. **Conclusion**
7. **Future Scope**
8. **References**

1.Introduction

The Healthcare data analytics plays a critical role in early disease detection, risk assessment, and informed clinical decision-making. Chronic diseases such as diabetes pose a significant public health challenge worldwide due to their long-term complications, increasing prevalence, and economic burden on healthcare systems. Early identification of individuals at risk is essential for timely intervention, prevention strategies, and improved patient outcomes.

This project focuses on the **prediction of diabetes risk using clinical and physiological data** derived from the **National Health and Nutrition Examination Survey (NHANES)**, a large-scale, government-conducted health survey in the United States. NHANES provides comprehensive and reliable data collected through interviews, physical examinations, and laboratory tests, covering diverse demographic groups. The dataset includes critical health indicators such as age, body mass index (BMI), blood pressure measurements, haemoglobin levels, and diabetes status, making it highly suitable for exploratory analysis and predictive modeling in the healthcare domain.

The primary objective of this study is to develop a robust machine learning pipeline that transforms raw clinical data into accurate and interpretable diabetes risk predictions. This involves systematic data preprocessing, handling missing values, feature engineering from repeated clinical measurements, exploratory data analysis to uncover meaningful health patterns, and comparative evaluation of multiple classification algorithms. Several machine learning models—including Logistic Regression, K-Nearest Neighbors, Decision Trees, Random Forests, Gradient Boosting, and Support Vector Machines—are trained and evaluated using appropriate healthcare-focused performance metrics such as accuracy, precision, recall, F1-score, and ROC-AUC.

By identifying key clinical factors associated with diabetes and selecting the most effective predictive model, this project demonstrates an end-to-end healthcare analytics workflow. The outcomes highlight the practical application of machine learning techniques in medical data analysis and emphasize the importance of data-driven approaches for disease risk prediction, clinical research, and public health planning.

2.Source of Dataset

The dataset used in this project is derived from the **National Health and Nutrition Examination Survey (NHANES)**, a large-scale, government-conducted health survey administered by the **Centers for Disease Control and Prevention (CDC)** in the United States. NHANES is designed to assess the health and nutritional status of adults and children through a combination of interviews, physical examinations, and laboratory tests, making it one of the most reliable public healthcare datasets available for research.

The data utilized in this study was compiled from a single NHANES survey cycle to maintain consistency and avoid cross-cycle bias. Multiple component files—including **demographic data**, **body measurements**, **blood pressure readings**, **laboratory test results**, and **health questionnaires**—were merged using a unique participant identifier. After preprocessing and integration, the final dataset contains **4,950 participant records**, each representing an individual respondent.

Dataset Attributes

The final cleaned dataset includes the following key clinical and physiological attributes:

- **Age** (in years)
- **Gender** (encoded as a binary variable)
- **Body Mass Index (BMI)**
- **Weight** (in kilograms)
- **Height** (in centimeters)
- **Mean Systolic Blood Pressure (SBP_MEAN)**
- **Mean Diastolic Blood Pressure (DBP_MEAN)**
- **Haemoglobin level**, representing a key laboratory measurement
- **Diabetes status**, used as the target variable (binary classification: diabetic or non-diabetic)

Repeated blood pressure measurements were aggregated using mean values to reduce measurement variability, and missing laboratory values were handled using appropriate statistical imputation techniques to preserve dataset integrity.

Dataset Suitability

This dataset provides sufficient **depth, diversity, and clinical relevance** for predictive modeling in the healthcare domain. The combination of demographic, anthropometric, physiological, and laboratory variables enables meaningful **exploratory data analysis (EDA)** and supports the development of robust **classification models** for diabetes risk prediction.

The presence of both continuous and categorical features makes the dataset well-suited for **comparative evaluation of multiple machine learning classifiers**, while the binary diabetes outcome aligns naturally with healthcare decision-support scenarios. The dataset's quality, scale, and standardized data collection methodology ensure reliable model training and evaluation.

Source

National Health and Nutrition Examination Survey (NHANES)
Centers for Disease Control and Prevention (CDC), USA
Official Data Portal: https://www.cdc.gov/nchs/nhanes/

3.Dataset Preprocessing

Dataset preprocessing is a critical step in building a reliable and accurate predictive model for diabetes risk classification. The raw NHANES data is distributed across multiple component files and initially contained missing values, repeated clinical measurements, heterogeneous formats, and variables not directly relevant to predictive modeling. A systematic preprocessing pipeline was implemented to ensure data consistency, clinical validity, and suitability for classification-based machine learning analysis.

3.1 Handling Missing Values

- All variables were examined to assess the extent and pattern of missing values.
- Non-essential attributes with excessive missing values or limited analytical relevance were excluded from the study.
- Missing values in key clinical features—such as **haemoglobin levels** and blood pressure readings—were carefully evaluated.
- Minor gaps in numerical laboratory measurements were handled using **median imputation**, a robust approach suitable for skewed clinical data.
- Records with incomplete target labels (diabetes status marked as "refused" or "unknown") were removed to preserve label integrity.

3.2 Data Cleaning

- Redundant variables and non-analytical identifiers, including repeated measurement fields and participant IDs, were removed after data integration.
- Multiple blood pressure readings were aggregated by computing **mean systolic and diastolic blood pressure**, reducing measurement noise.
- Inconsistent formats across merged NHANES components were standardized to ensure uniform data representation.
- Outliers caused by measurement errors or physiologically implausible values were identified and filtered to prevent distortion of model learning.

3.3 Feature Engineering

To improve predictive performance and clinical interpretability, several derived features were created:

- **Mean systolic and diastolic blood pressure** values were computed from repeated readings to represent stable physiological indicators.
- Anthropometric measures such as **BMI** were retained as core predictors due to their strong association with diabetes risk.
- The diabetes target variable was transformed into a **binary classification label** (diabetic / non-diabetic) suitable for supervised learning.
- Feature selection was guided by exploratory data analysis to retain clinically meaningful variables while reducing redundancy.

These engineered features enabled the models to better capture underlying health patterns associated with diabetes risk.

3.4 Standardization and Formatting

- All numerical features were converted to appropriate **integer or floating-point data types** based on their clinical meaning.
- Feature scaling was applied using **standardization techniques** to ensure uniform value ranges, particularly for distance-based and linear classification models.
- Consistent naming conventions were applied across all variables to improve readability, interpretability, and reproducibility of the modeling pipeline.

3.5 Data Splitting for Predictive Modeling

- The dataset was divided into **training and testing sets** using a stratified split to preserve the class distribution of the diabetes target variable.
- Data leakage was prevented by ensuring that preprocessing steps such as scaling were fitted only on the training data.
- The test set was reserved exclusively for final model evaluation to obtain an unbiased estimate of predictive performance.

4.Analysis on Dataset

This section presents the exploratory data analysis conducted to understand the distribution, relationships, and patterns within the NHANES clinical dataset. Multiple analytical objectives were achieved through statistical summaries and visualizations to identify key risk factors associated with diabetes and to guide subsequent predictive modeling.

4.1 Distribution of Diabetes Cases

i. General Description

This analysis examines the overall distribution of diabetic and non-diabetic individuals in the dataset. Understanding class balance is essential for selecting appropriate classification models and evaluation metrics.

ii. Requirements

- Count individuals by diabetes status
- Assess class imbalance

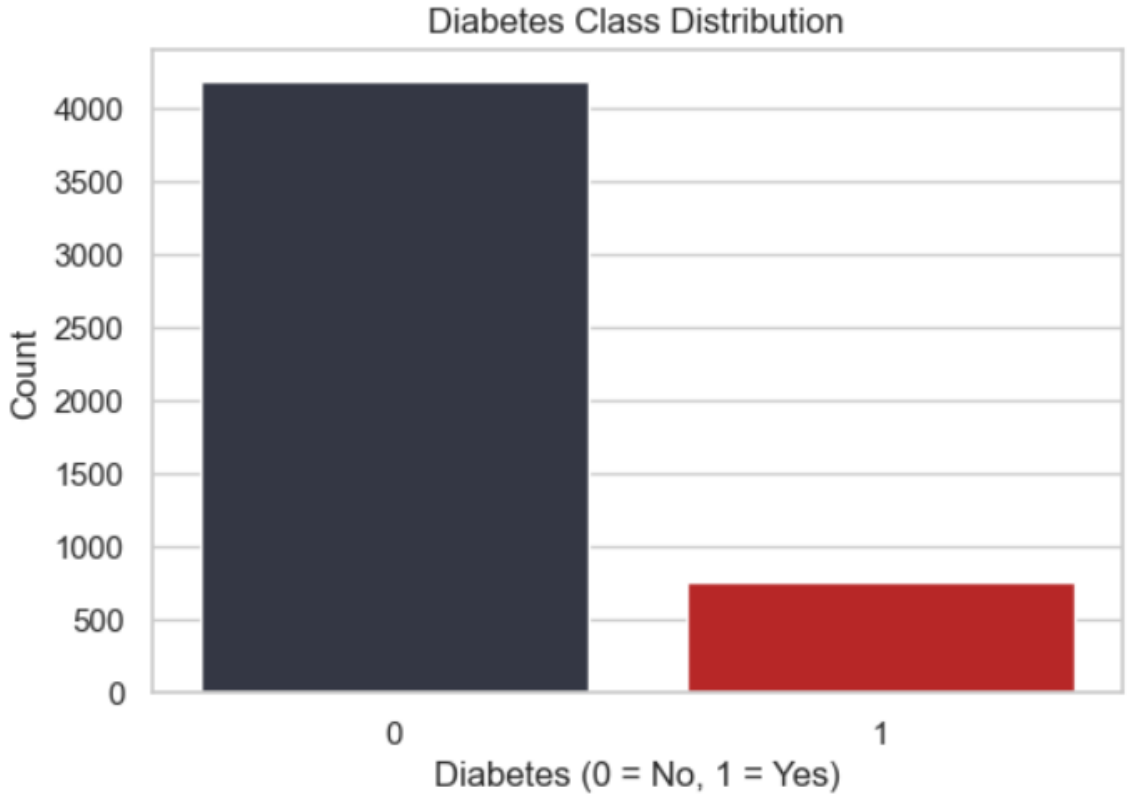
iii. Analysis Results

The dataset contains both diabetic and non-diabetic individuals, with non-diabetic cases forming the majority. However, the proportion of diabetic cases is sufficiently large to support reliable classification modeling.

iv. Visualization

Bar chart showing the distribution of diabetes classes.

Interpretation:
The presence of both classes in adequate proportions makes the dataset suitable for supervised classification.



4.2 Age Distribution by Diabetes Status

i. General Description

This analysis explores how diabetes prevalence varies across different age groups.

ii. Requirements

- Analyze age distribution
- Compare age patterns between diabetic and non-diabetic individuals

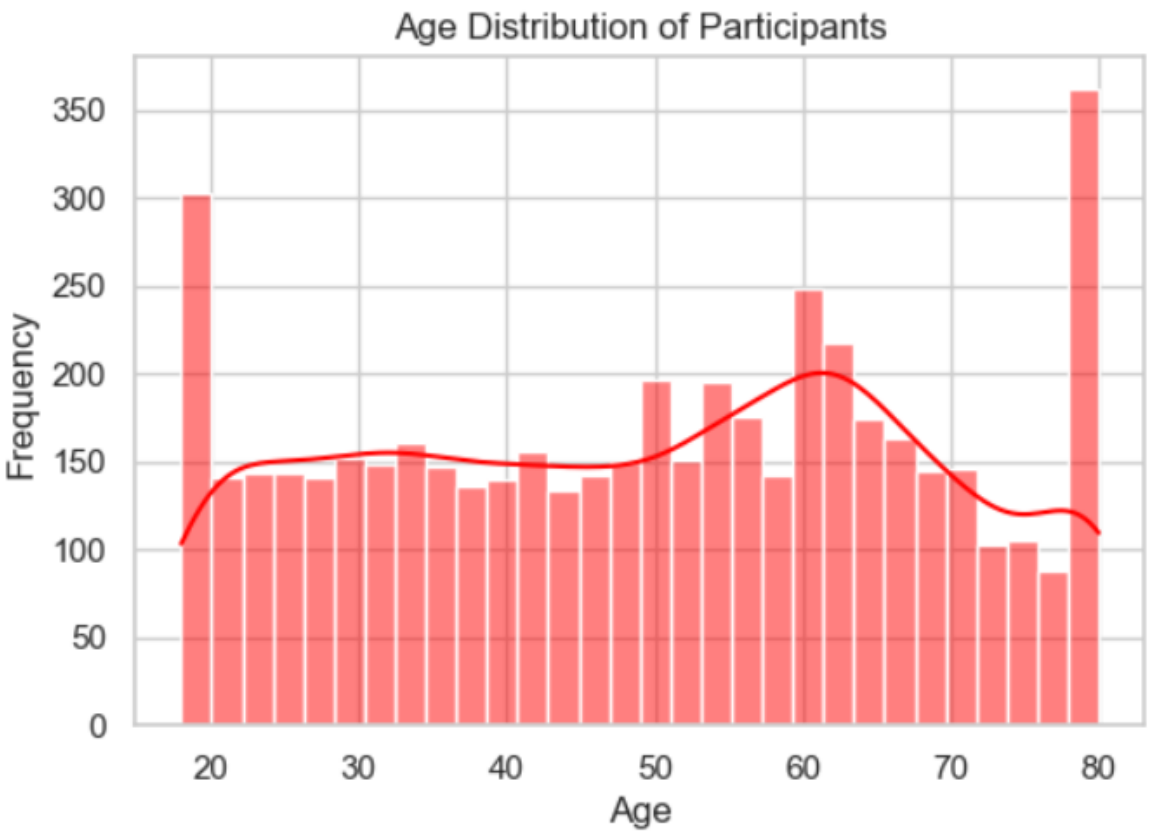
iii. Analysis Results

Diabetes prevalence increases noticeably with age. Older age groups show a significantly higher proportion of diabetic cases compared to younger individuals.

iv. Visualization

Histogram or boxplot showing age distribution by diabetes status.

Interpretation:
Age is a strong demographic risk factor for diabetes and an important predictor in the modeling process.



4.3 BMI Distribution by Diabetes Status

i. General Description

Body Mass Index (BMI) is a key clinical indicator associated with metabolic disorders. This analysis evaluates its relationship with diabetes.

ii. Requirements

- Compare BMI distributions across diabetes classes

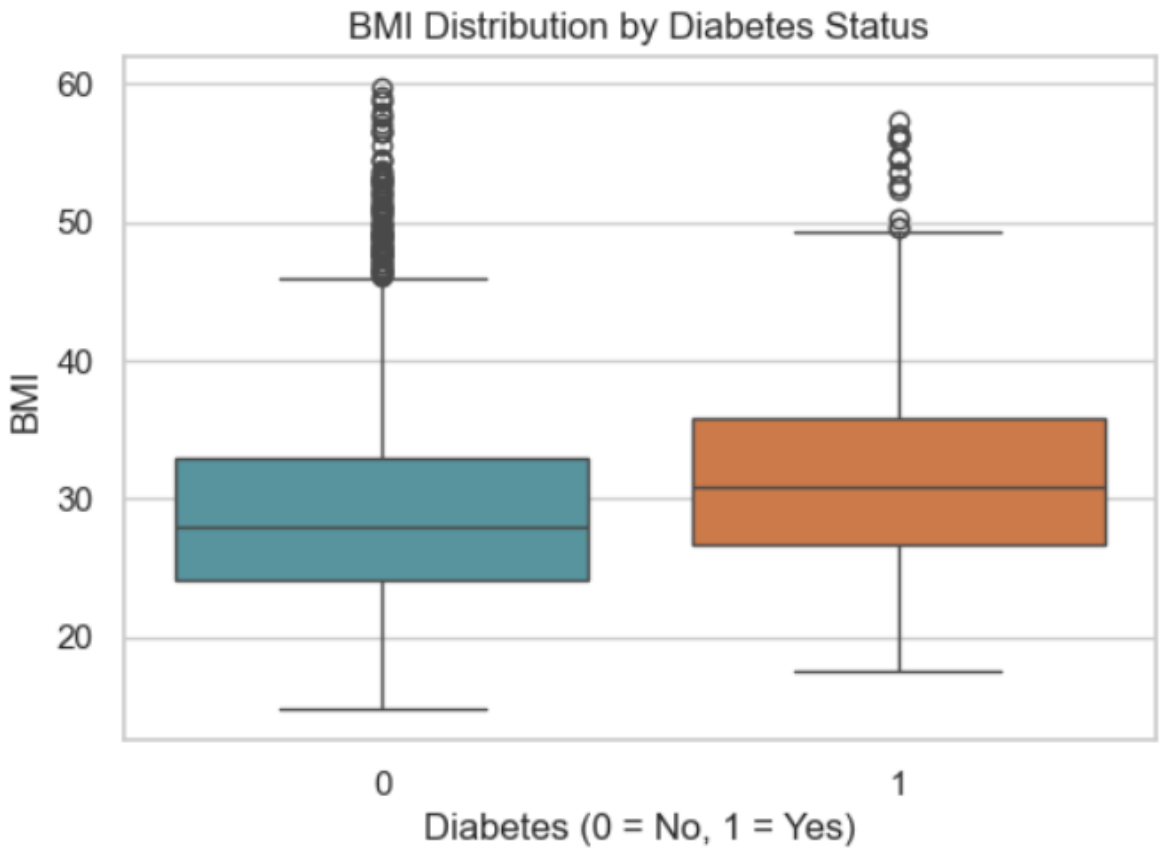
iii. Analysis Results

Individuals with diabetes exhibit higher median BMI values compared to non-diabetic individuals, indicating a strong association between obesity and diabetes.

iv. Visualization

Boxplot comparing BMI for diabetic and non-diabetic groups.

Interpretation:
Elevated BMI is a significant contributor to diabetes risk.



4.4 Blood Pressure Analysis

i. General Description

This analysis evaluates systolic and diastolic blood pressure patterns across diabetes status.

ii. Requirements

- Analyze mean systolic and diastolic blood pressure
- Compare across diabetes classes

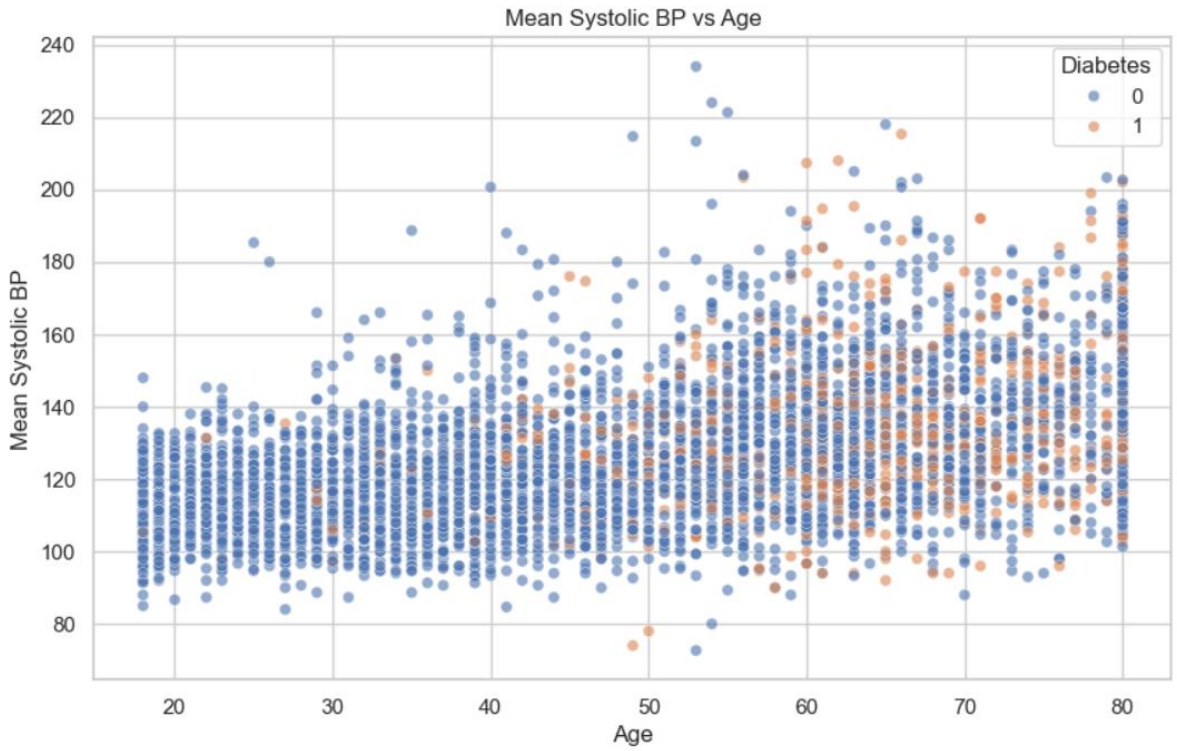
iii. Analysis Results

Diabetic individuals generally show higher mean systolic and diastolic blood pressure values, reflecting the coexistence of hypertension and diabetes.

iv. Visualization

Scatterplot showing blood pressure distributions.

Interpretation:
Blood pressure measurements are important physiological indicators for diabetes risk classification.



4.5 Haemoglobin Level Analysis

i. General Description

Haemoglobin levels provide insight into metabolic and physiological health. This analysis examines their relationship with diabetes.

ii. Requirements

- Compare haemoglobin levels across diabetes classes

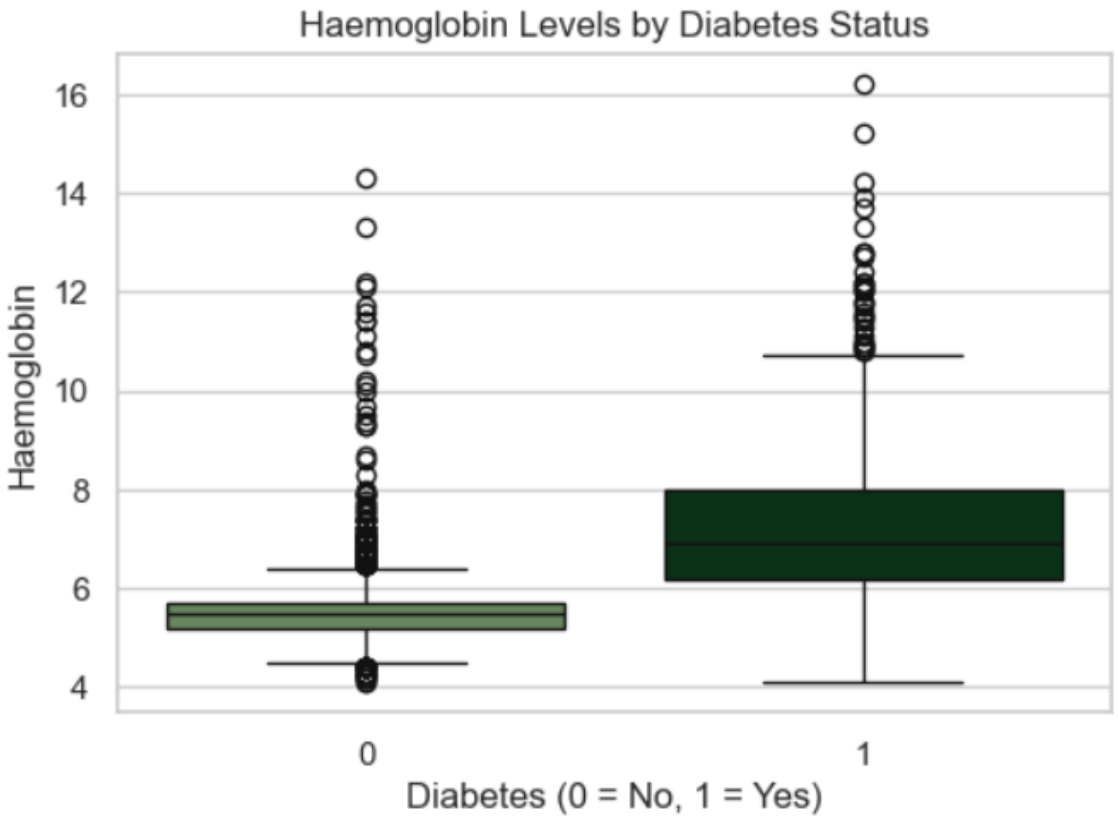
iii. Analysis Results

Distinct differences in haemoglobin distributions are observed between diabetic and non-diabetic individuals, indicating its relevance as a predictive feature.

iv. Visualization

Boxplot showing haemoglobin levels by diabetes status.

Interpretation:
Laboratory indicators contribute significantly to diabetes risk prediction.



4.6 Gender-wise Diabetes Distribution

i. General Description

This analysis evaluates diabetes prevalence across gender groups.

ii. Requirements

- Compare diabetes counts by gender

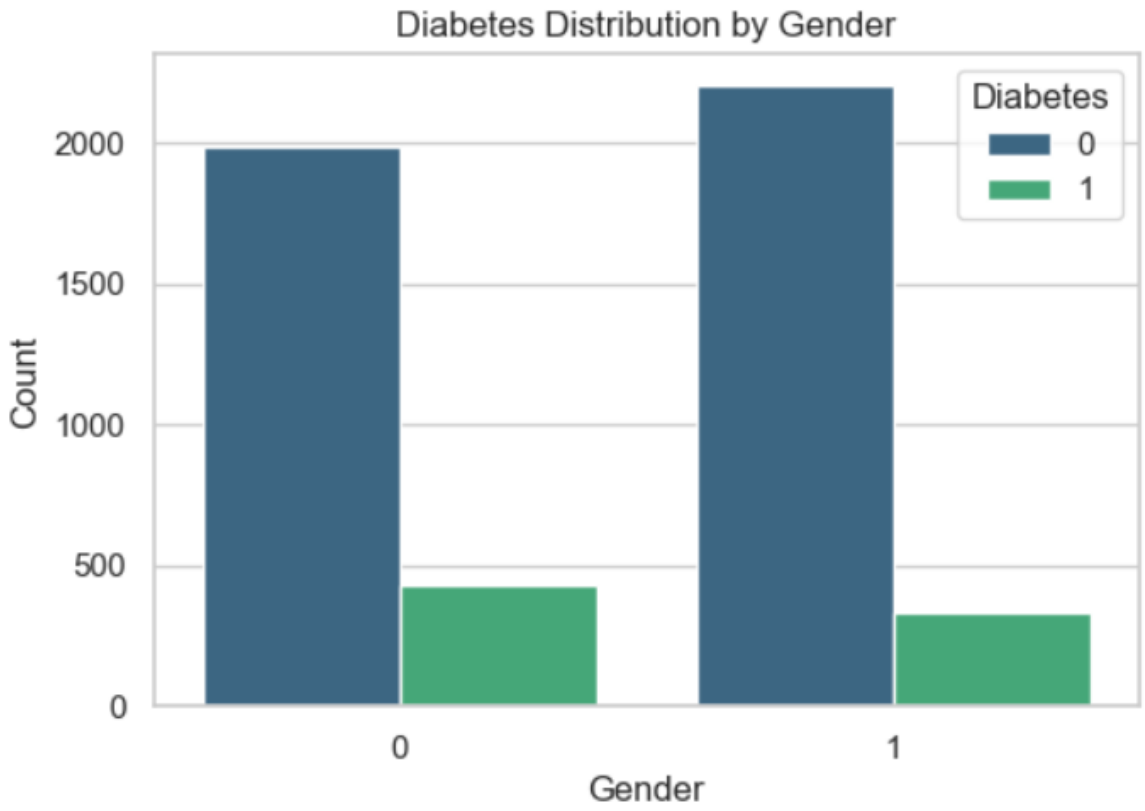
iii. Analysis Results

Both genders show cases of diabetes, with slight variations in prevalence depending on gender.

iv. Visualization

Grouped bar chart showing diabetes distribution by gender.

Interpretation:
Gender-based differences exist and should be considered during model training.



4.7 Correlation Analysis of Clinical Features

i. General Description

Correlation analysis helps identify relationships between numerical variables and detect multicollinearity.

ii. Requirements

- Compute correlation matrix
- Visualize feature relationships

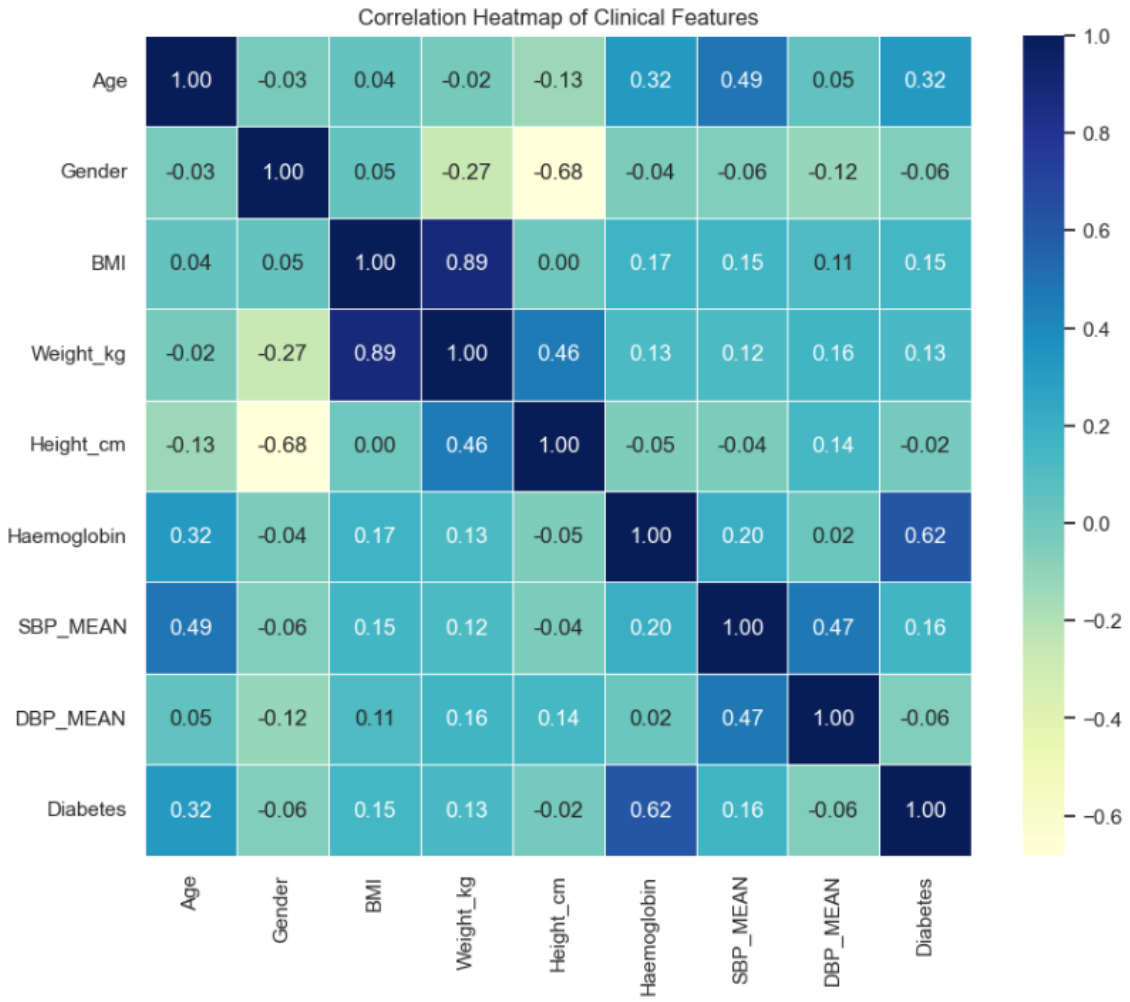
iii. Analysis Results

Strong correlations are observed between BMI, blood pressure, and diabetes status. Some features show moderate inter-correlation but remain clinically meaningful.

iv. Visualization

Correlation heatmap of numerical features.

Interpretation:
Correlation analysis supports feature selection and confirms clinically relevant associations.



---

5. GitHub & Linkeldn Links

GitHub: https://github.com/usabhishek/Predictive-Modeling-for-Diabetes-Using-NHANES-Data

Linkeldn: https://www.linkedin.com/posts/abhishekkashyap14_ds-machinelearning-healthcare-activity-7408913445531611137-WniN?utm_source=social_share_send&utm_medium=member_desktop_web&rcm=ACoAAEfp3S0Ba3IvL0fleFhRRq8j_I2LqrBbOCI

Google Drive:
https://drive.google.com/drive/folders/1zUzHHuqmZHMWpsIDQbfms3uhx4GeVUPs?usp=sharing

6.Conclusion

This project successfully developed an end-to-end **machine learning–based diabetes prediction system** using clinical and physiological data from the **NHANES dataset**. Through systematic data preprocessing, exploratory data analysis, and comparative model evaluation, the study demonstrated how real-world healthcare data can be transformed into reliable and interpretable predictive insights.

Exploratory Data Analysis (EDA) revealed strong and clinically meaningful relationships between diabetes status and key health indicators such as **age, body mass index (BMI), blood pressure, and haemoglobin levels**. Visual analyses confirmed that diabetic individuals tend to exhibit higher BMI and blood pressure values, reinforcing established medical knowledge and validating the relevance of the selected features. Correlation analysis further supported feature selection while highlighting manageable levels of interdependence among predictors.

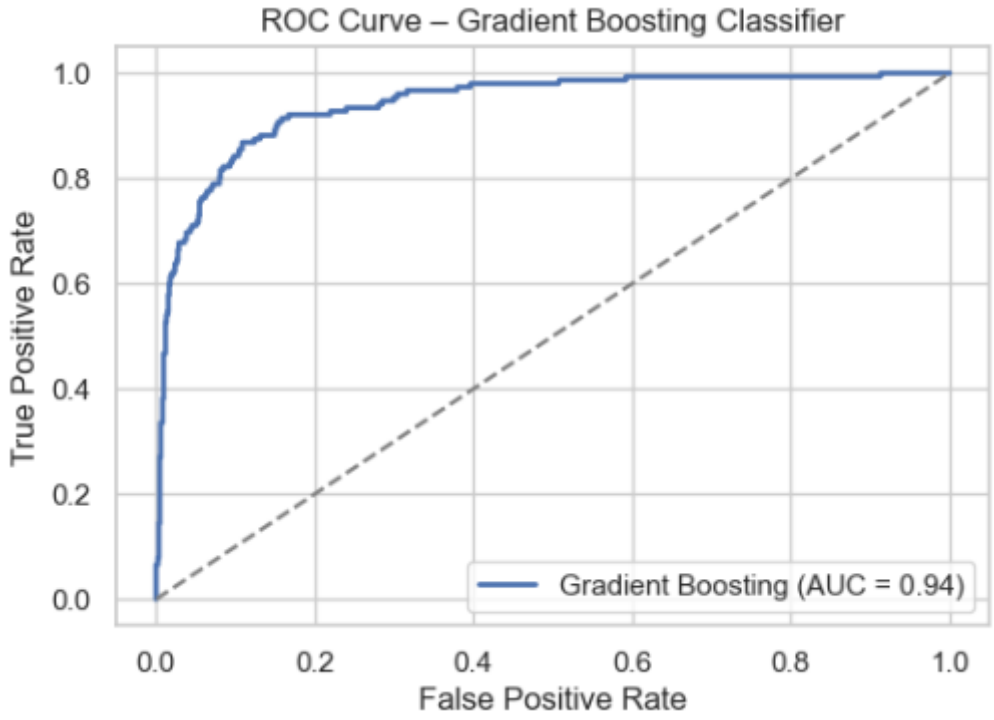
Multiple classification models—including **Logistic Regression, K-Nearest Neighbors, Decision Tree, Random Forest, Gradient Boosting, and Support Vector Machine**—were trained and evaluated to identify the most effective predictive approach. Model performance was compared using standard healthcare-appropriate metrics such as **accuracy, precision, recall, F1-score, and ROC-AUC**, ensuring that evaluation focused not only on overall correctness but also on the model's ability to correctly identify diabetic cases.

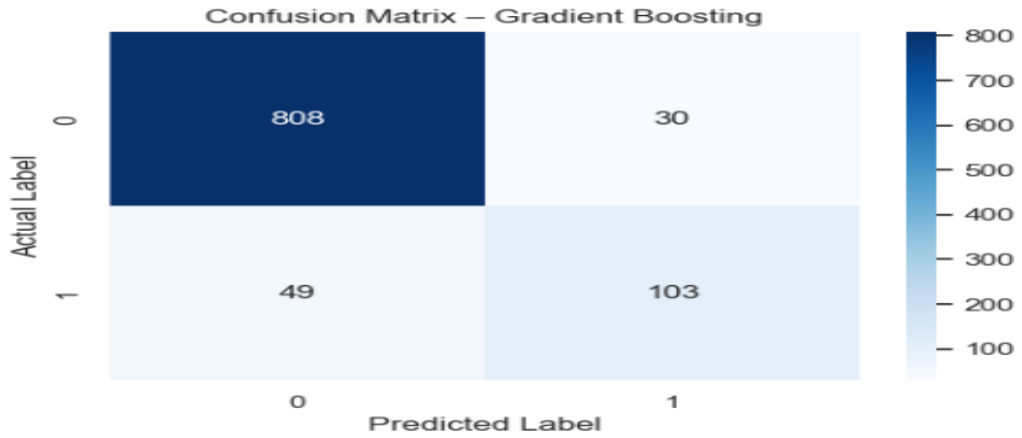
Among all evaluated models, **Gradient Boosting** demonstrated the strongest overall performance. The confusion matrix showed a high number of correct classifications for both diabetic and non-diabetic individuals, with a favorable balance between false positives and false negatives. The ROC curve further confirmed the model's effectiveness, achieving a **high ROC-AUC score (~0.94)**, indicating excellent discriminative capability. Feature importance analysis revealed **haemoglobin** as the most influential predictor, followed by age and blood pressure measures, aligning well with clinical expectations.

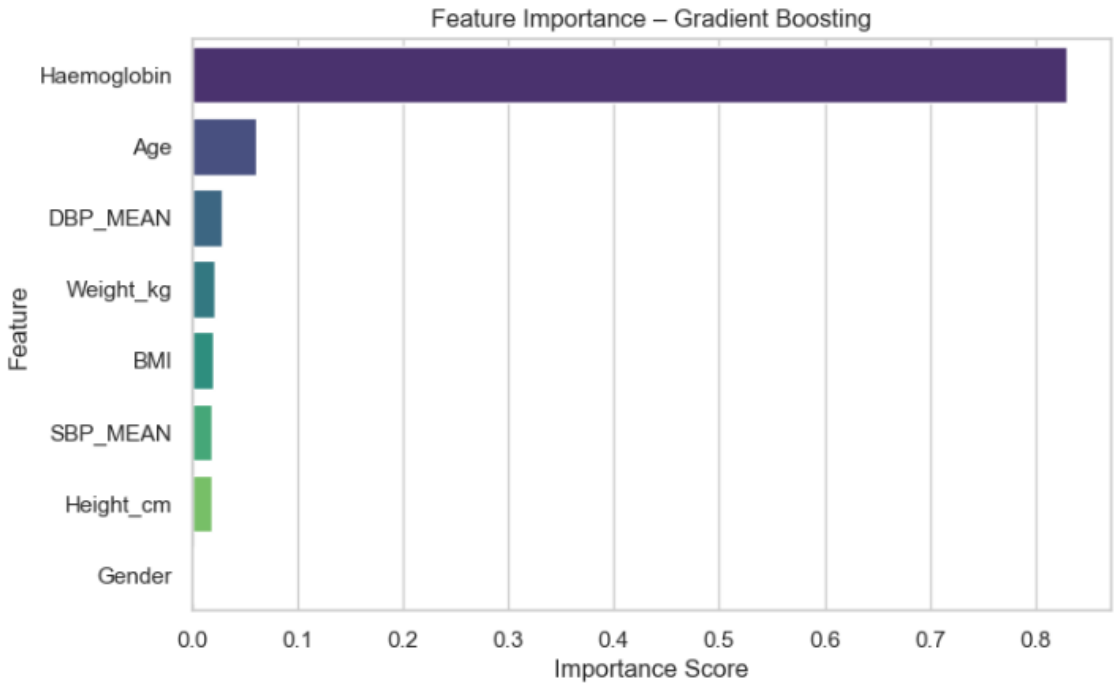
The comparative analysis highlights that ensemble-based models outperform simpler classifiers for this dataset, particularly in handling non-linear relationships and complex feature interactions common in healthcare data. By prioritizing ROC-AUC and recall alongside accuracy, the study ensured that the selected model is suitable for risk-prediction scenarios where missing positive cases can have serious implications.

Overall, this project demonstrates the practical application of machine learning in healthcare analytics, from raw data preprocessing to model interpretation and evaluation. The results confirm that **Gradient Boosting is a robust and reliable choice for diabetes risk prediction using NHANES data**, and the methodology established in this study provides a strong foundation for future extensions such as risk-scoring systems, integration of additional laboratory variables, and deployment within clinical decision-support frameworks.

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Gradient Boosting	0.920202	0.774436	0.677632	0.722807	0.941563
Random Forest	0.921212	0.789062	0.664474	0.721429	0.937394
Logistic Regression	0.923232	0.851852	0.605263	0.707692	0.932782
Support Vector Machine	0.926263	0.862385	0.618421	0.720307	0.911537
K-Nearest Neighbors	0.912121	0.821782	0.546053	0.656126	0.896154
Decision Tree	0.890909	0.641026	0.657895	0.649351	0.795534







---

7. Future Scope

The scope of this project can be extended further through the following enhancements:

• **Advanced Risk Prediction and Forecasting:**
Future work can involve developing more advanced predictive models to estimate an individual's **long-term diabetes risk** or progression probability by incorporating longitudinal health data and follow-up survey cycles.

• **Incorporation of Lifestyle and Behavioral Factors:**
Additional NHANES variables such as dietary intake, physical activity levels, smoking status, alcohol consumption, and sleep patterns can be integrated to improve predictive accuracy and provide a more holistic assessment of diabetes risk.

• **Temporal and Multi-Cycle Analysis:**
The model can be extended to include data from multiple NHANES survey cycles to study **temporal trends** in diabetes prevalence and risk factors, enabling more generalized and robust predictions across different population cohorts.

• **Explainable AI and Clinical Interpretability:**
Advanced explainability techniques such as SHAP or LIME can be applied to enhance model transparency, allowing clinicians and researchers to better understand individual-level predictions and feature contributions.

• **Deployment as a Decision Support Tool:**
The trained model can be deployed as a **web-based or clinical decision support system**, enabling healthcare professionals to assess diabetes risk in real time using patient input data.

• **Integration with Public Health Planning:**
The predictive framework can be adapted for population-level analysis to support **public health policy-making**, early screening initiatives, and targeted intervention programs aimed at reducing diabetes incidence.

References

Skicit-learn documentation : https://scikit-learn.org/stable/supervised_learning.html

Nhanes dataset documentation:
https://wwwn.cdc.gov/nchs/nhanes/continuousnhanes/overview.aspx?BeginYear=2017

Machine learning tutorial : https://www.geeksforgeeks.org/machine-learning/machine-learning/