DEPI Graduation Project

Diabetes Diagnosis

Team Members:

- Omar Mohamed Ahmed Mashaly
- Micheal Mourad
- Abdelrahman Zayn
- Mennatullah Ayman

ALX1_AIS3_S1e IBM Data Science

Eng/Sherif Said

Contents

Project Overview : Methodology and Approach

1. Executive Summary: This project aims to develop a machine learning model for diagnosing diabetes using a comprehensive dataset of health and demographic information. The project encompasses data preprocessing, exploratory data analysis (EDA), feature engineering, model development, and the creation of a user-friendly interface for diabetes prediction.

2. Introduction: Diabetes is a chronic health condition affecting millions worldwide. Early diagnosis is crucial for effective management and prevention of complications. This project leverages machine learning techniques to create a predictive model for diabetes diagnosis based on various health and demographic factors.

3. Dataset: The project utilizes a dataset comprising health and demographic data of 100,000 individuals, including:

- Demographic information (age, gender, race)
- Health metrics (BMI, blood glucose level, HbA1c)
- Medical history (hypertension, heart disease)
- Lifestyle factors (smoking history)
- Source: Kaggle (https://www.kaggle.com/datasets/priyamchoksi/100000-diabetes-clinical-dataset/code)

4. Methodology:

4.1 Data Preprocessing:

- Removal of outliers and duplicate entries as there were outliers in the BMI column and HbA1c column and there were 14 duplicate rows.

- Handling of missing values and there were no missing values in the dataset.

- Feature engineering In which age and BMI categories columns were created to gain deep insights and also removed rows from the smoking history column which had no info as a value as this is not relevant for the analysis, also combined the columns related to race into one column.

- Encoding categorical variables to help in our prediction using machine learning models and in it encoded the gender and smoking history columns as the are relevant for the analysis

4.2 Exploratory Data Analysis:

- Analysis of feature correlations with diabetes diagnosis and found the most correlated features to diabetes which are: Blood glucose level,Age, HbA1c level,BMI,Hypertension and Heart Disease.

- Found imbalance in the dataset as 92% of the individuals in the dataset are Non-Diabetic and only 8% are diabetic which will make the machine learning model biased toward Non-Diabetic Patients and will find difficulty predicting if the individual is diabetic and solved this by taking random Equal sample from both to avoid model bias.

- Found some Key insights which are:

  This dataset includes the year range from 2015-2022 with the median year is 2019 with most data points.

  The age range is from 0 to 80 years, with a mean age of approximately 53 years.The median age is 55 years, indicating that half of the individuals are younger than 55 years and half are older.

  Hypertension: Approximately 16.7% of individuals have hypertension.

  Heart Disease: Approximately 9% of individuals have heart disease.

  The mean BMI is 30.10.The median BMI is approximately 28, indicating that half of the individuals have a BMI below this value,BMI values range from 10.19 to 88.76.

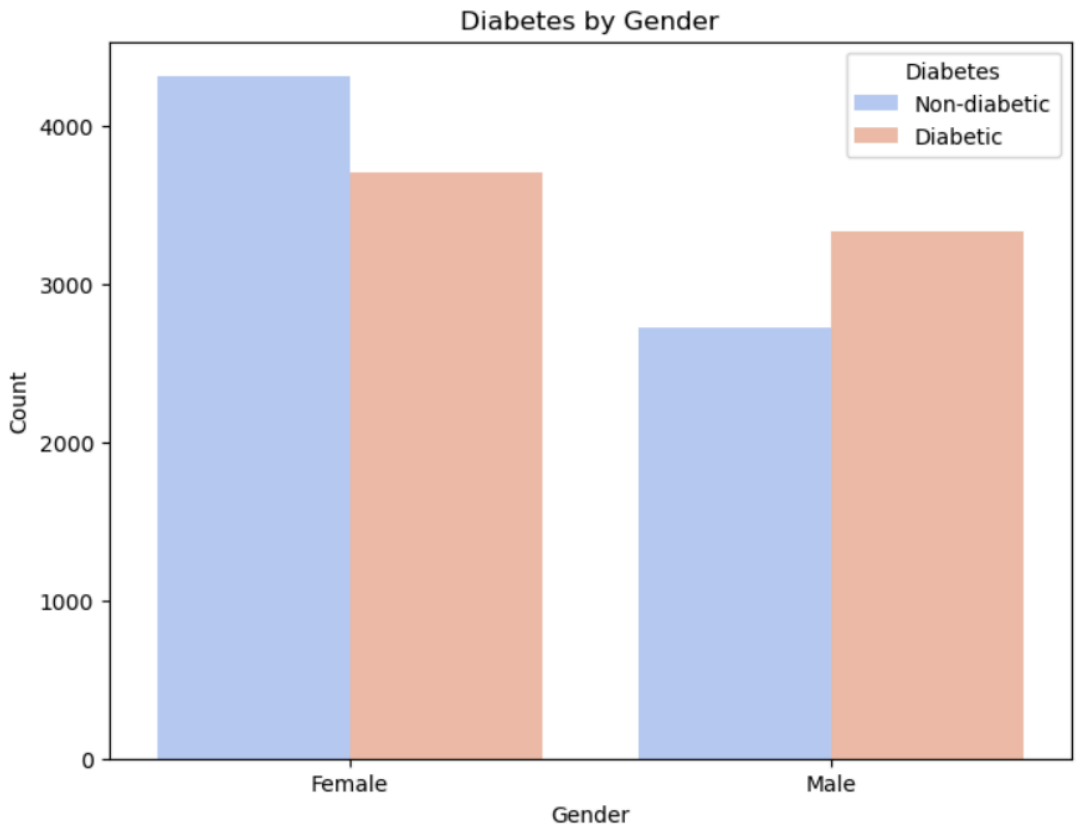
  HbA1c levels range from 3.50 to 9.00.The median HbA1c level is 6.15.

  Blood Glucose Level :The mean blood glucose level is 163.25 mg/dL. The median blood glucose level is 155 mg/dL.Blood glucose levels range from 80 to 300 mg/dL.

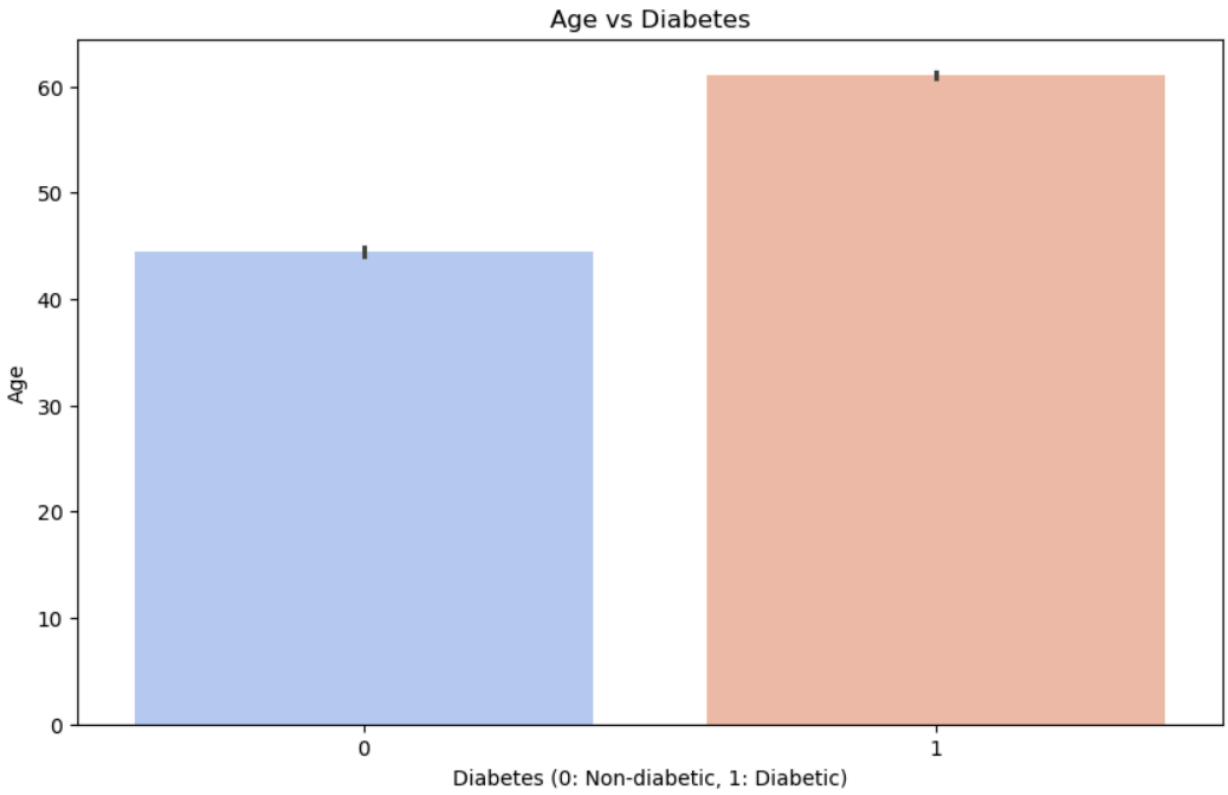
  Diabetes: Approximately half of individuals in the dataset have diabetes.

Data Visualizations

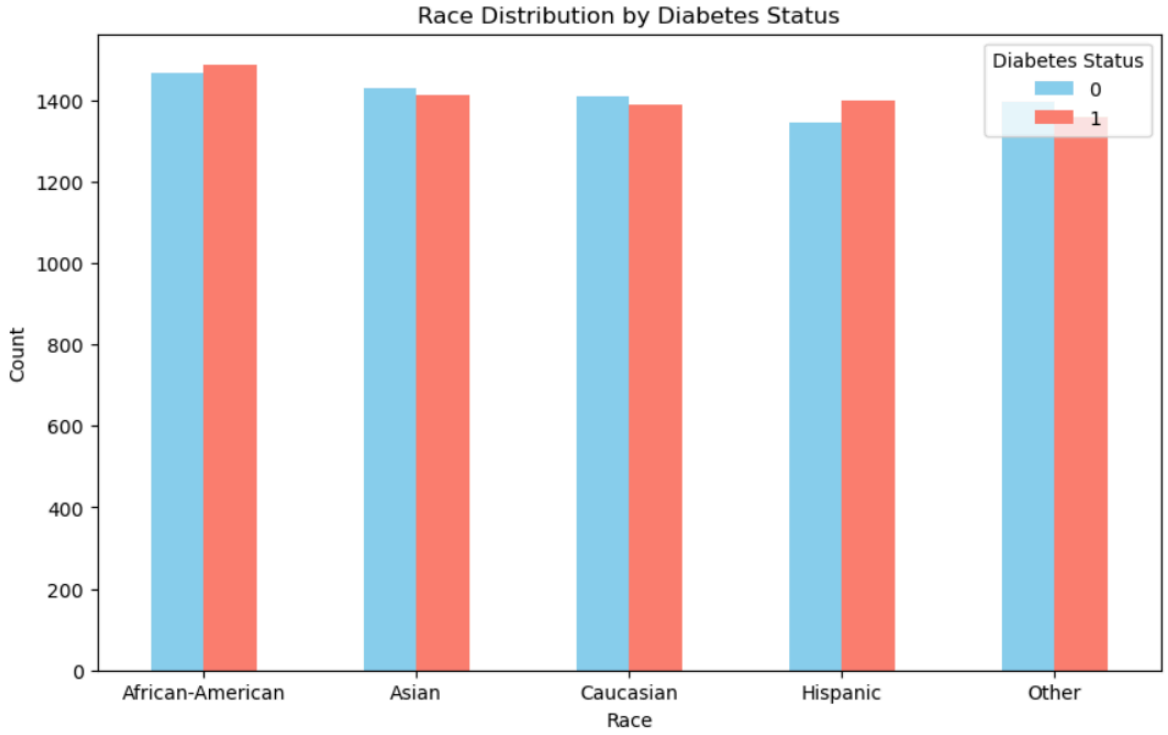
Here are Few Visualizations and the insights gained from them:



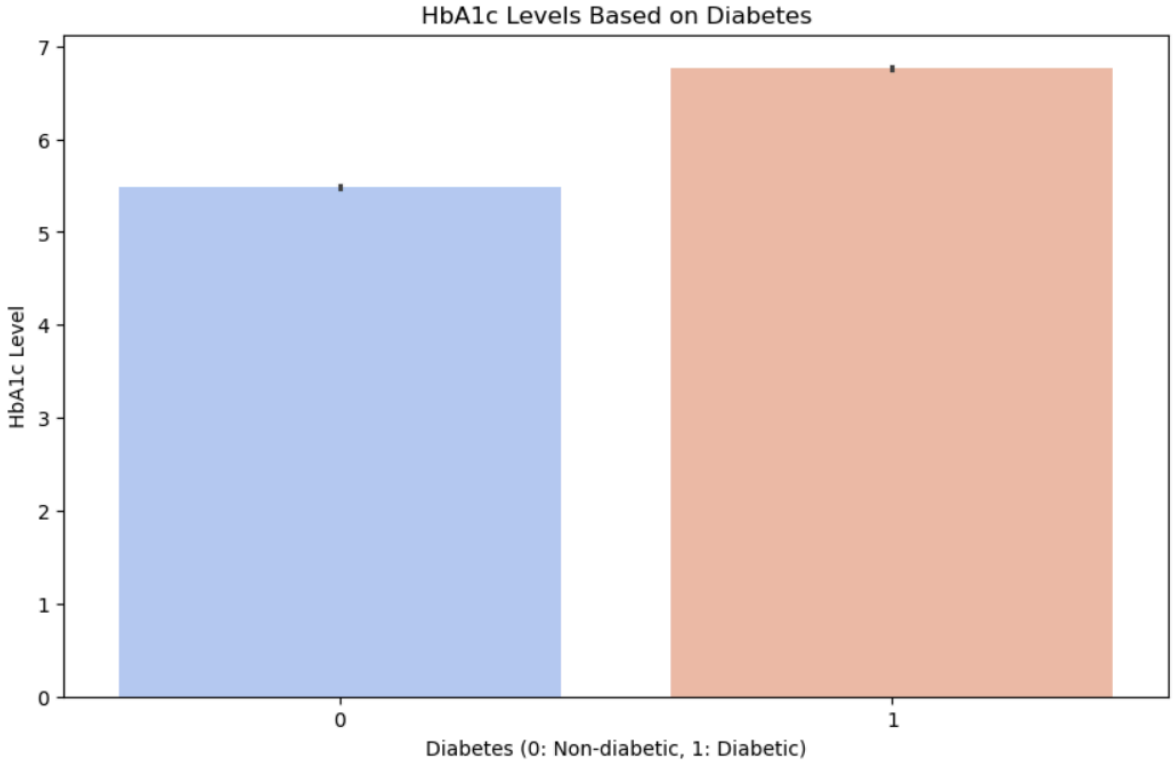
In the **Gender** column There are more females than males in the dataset and the most of them are non-diabetic but there are also a big amount of them are diabetic but in the males it is the opposite as more men are diabetic than non-diabetic.



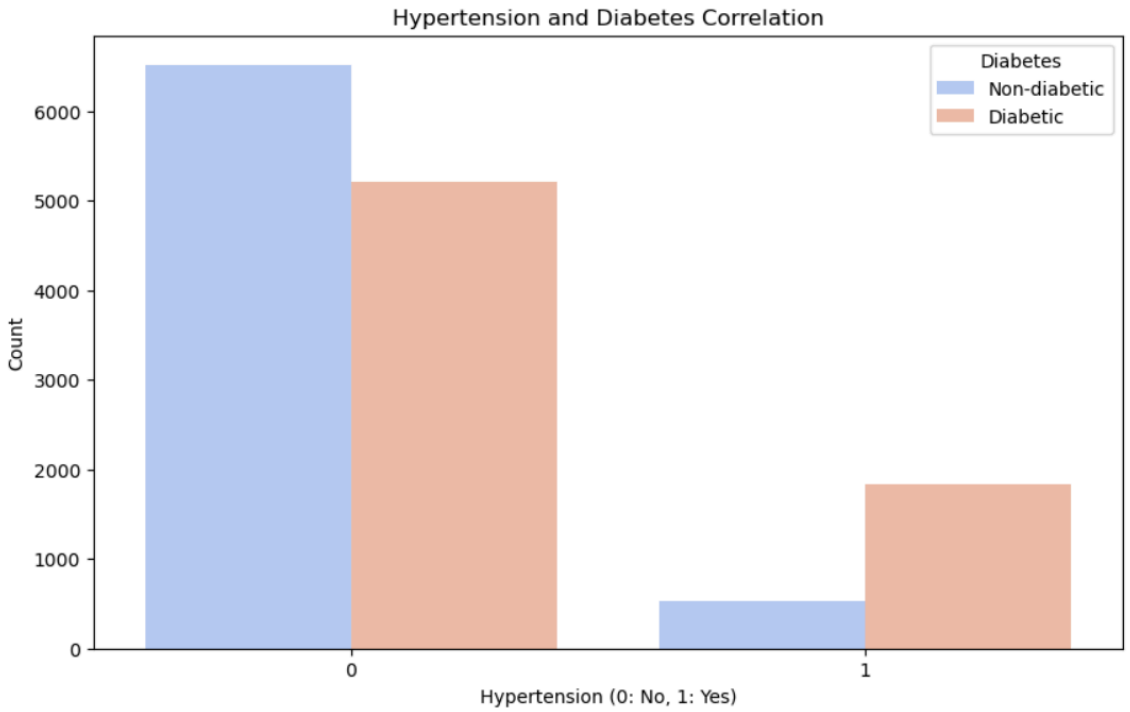
The average age of non-diabetic individuals (represented by the blue bar) is around 45 years.The average age of diabetic individuals (represented by the orange bar) is significantly higher, around 60 years.There is a clear trend indicating that diabetes is more prevalent in older individuals, as evidenced by the higher average age in the diabetic group.



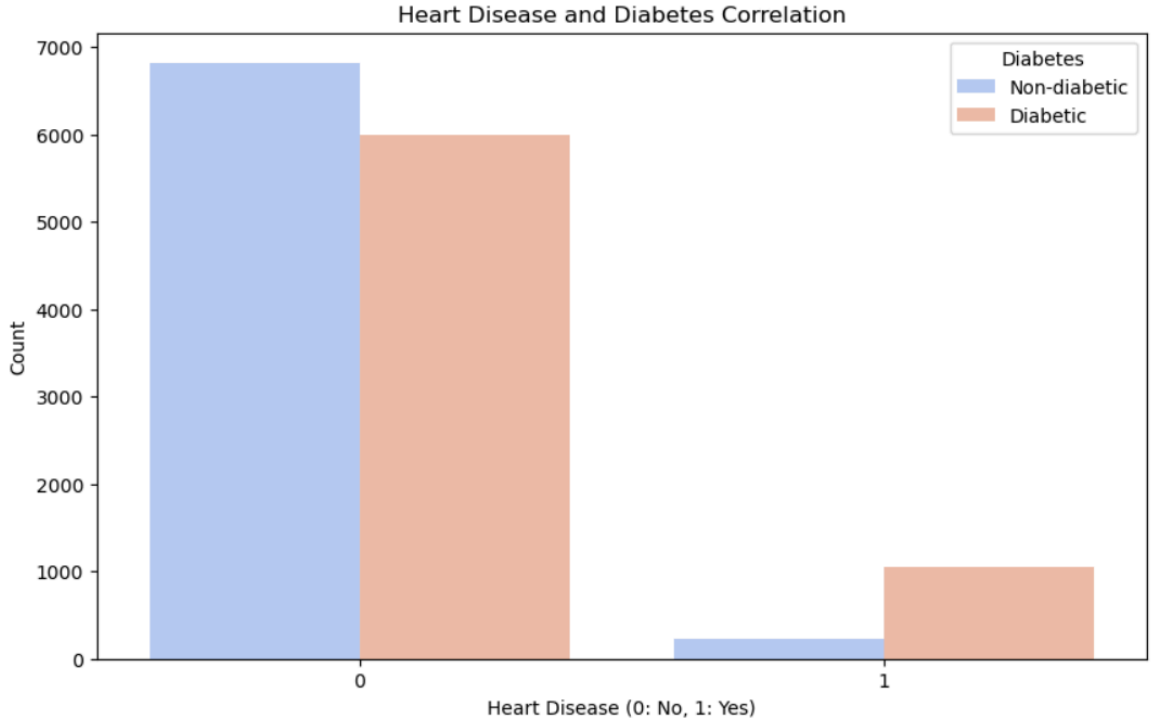
The counts for each racial group (African-American, Asian, Caucasian, Hispanic, and Other) are relatively similar, indicating a balanced representation across these categories.The data suggests that across all racial categories, there are more individuals without diabetes compared to those with diabetes. However, the differences are not drastic, indicating that diabetes prevalence may be relatively consistent across these racial groups.



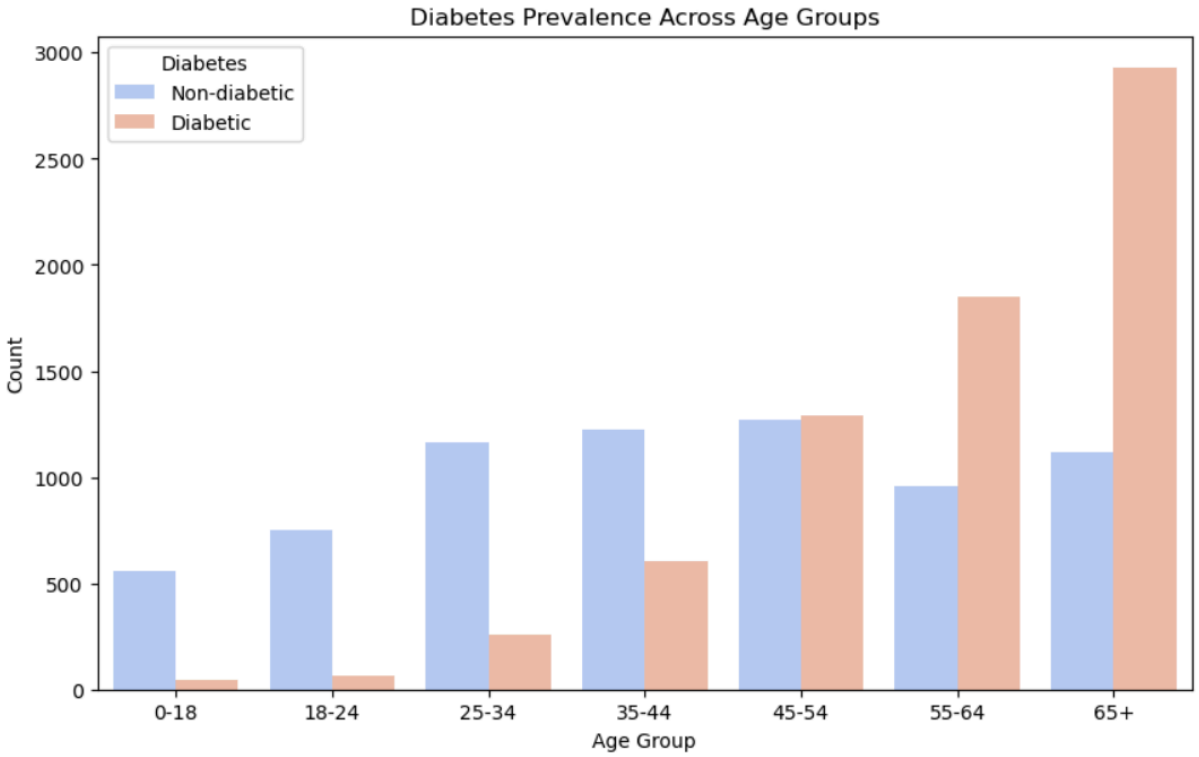
The average HbA1c level for non-diabetic individuals is significantly lower than that for diabetic individuals. This suggests that diabetes is associated with higher blood sugar levels over time. The non-diabetic group shows an HbA1c level around 5.5, while the diabetic group has an HbA1c level closer to 7. This indicates a clear distinction in glycemic control between the two groups.Higher HbA1c levels in diabetics can indicate poor long-term glucose control, which is associated with increased risk of diabetes-related complications.



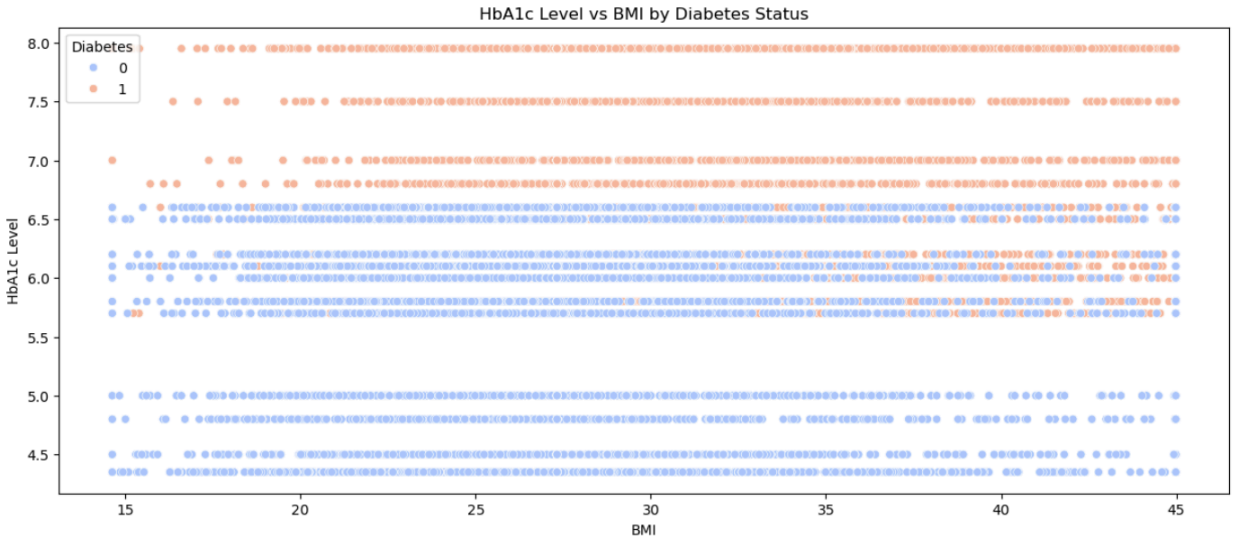
Hypertension and Diabetes Correlation: The first chart shows that most non-diabetic individuals do not have hypertension, while diabetic individuals are more likely to have hypertension compared to non-diabetics as the count is still relatively high.



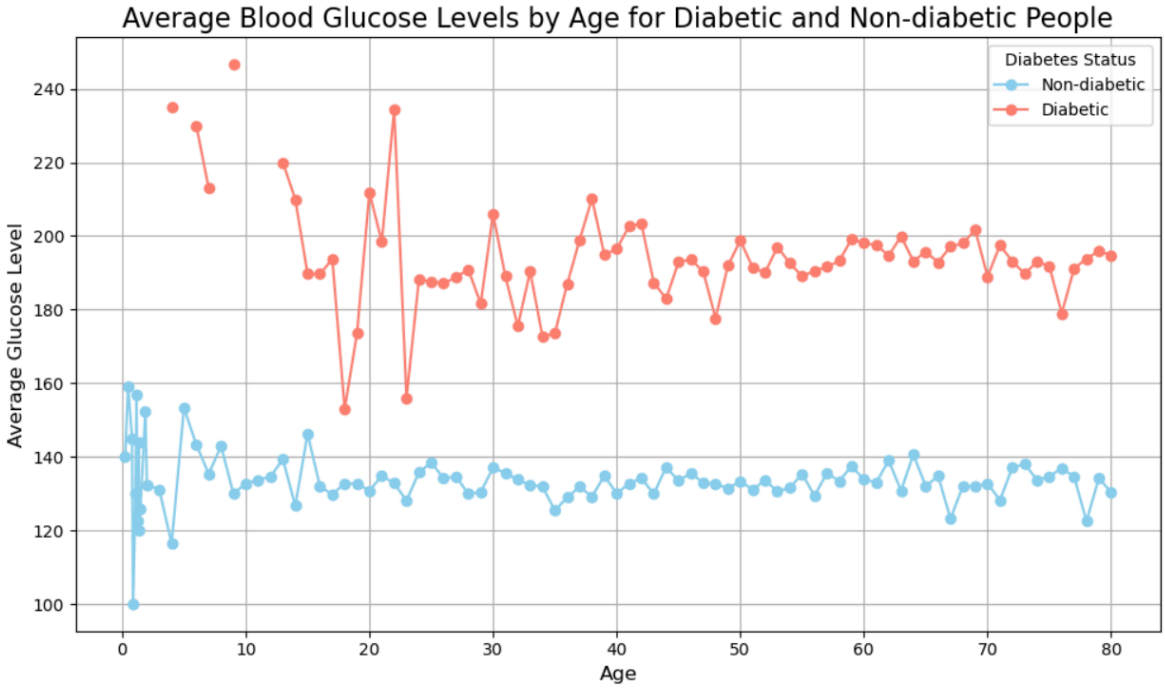
Heart Disease and Diabetes Correlation: The chart reveals a similar pattern, where most non-diabetic individuals do not have heart disease, but diabetic individuals have a higher occurrence of heart disease compared to non-diabetics.



The diabetic population is notably low in the younger age groups (0-18 and 18-24) but increases gradually from the 25-34 age group, indicating a potential rise in diabetes prevalence as age increases.The age group 65+ has the highest count of diabetic individuals, suggesting that this age range may be critical for diabetes management and prevention strategies.The 65+ age group shows a decrease in non-diabetic individuals, while the diabetic count increases as age increase, indicating that older adults may have a higher risk of diabetesThe data suggests that diabetes prevalence increases with age, particularly after 40, highlighting the importance of monitoring and preventive measures in older populations.The non-diabetic count is nearly stable across age groups.



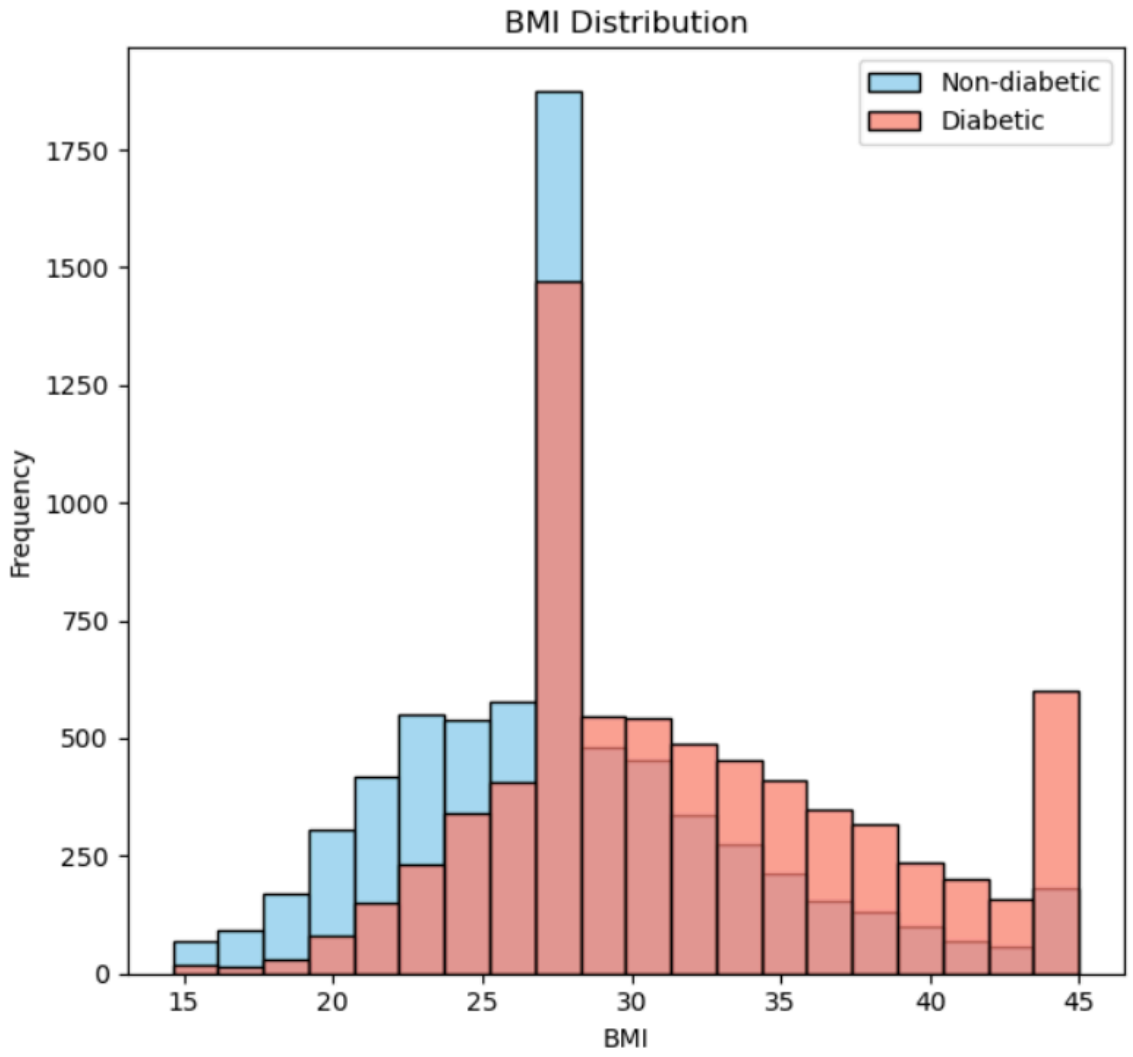
There is some overlap in HbA1c levels between the two groups, particularly at lower BMI values. This indicates that not all individuals with higher HbA1c levels are diabetic, and some non-diabetics may also have elevated levels, as BMI increases, there is a noticeable trend where HbA1c levels tend to be higher for those with diabetes. This could imply that higher BMI is associated with worse glycemic control in diabetic individuals.



Diabetic individuals consistently show higher average blood glucose levels compared to non-diabetic individuals across all age groups.As age increases, the average glucose levels for

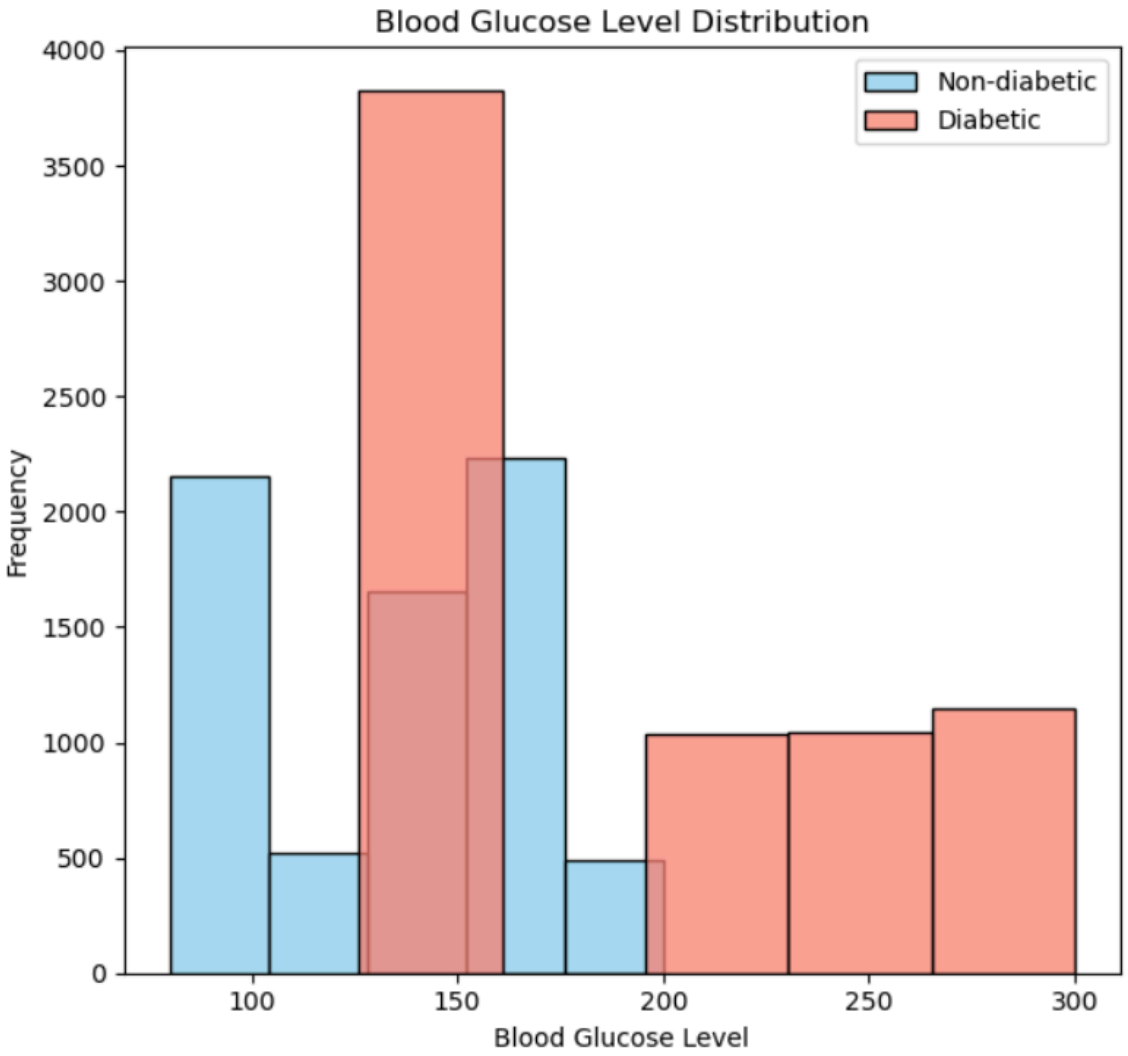
both groups tend to stabilize, but the diabetic group remains significantly higher than the non-diabetic group.The average glucose levels for non-diabetic individuals remain relatively stable and low, suggesting better glucose regulation.

This highlights the importance of monitoring blood glucose levels, especially in older adults and those diagnosed with diabetes, to manage health risks associated with high glucose levels.



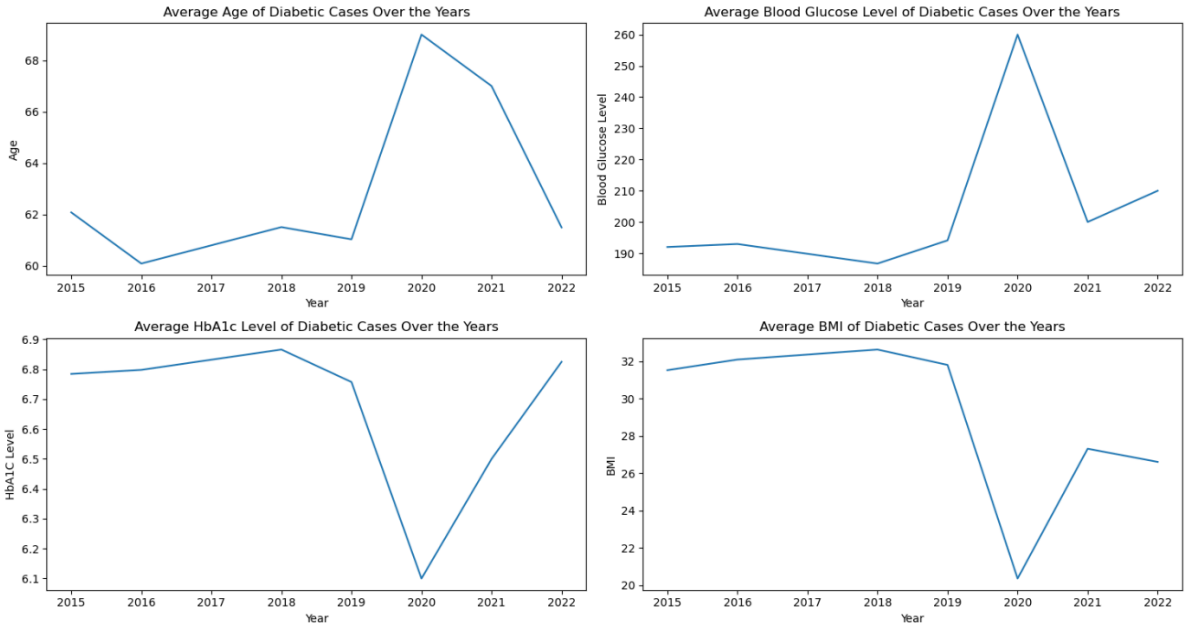
The BMI values range from approximately 15 to 45, with the majority of data concentrated between 20 and 35.The peak frequency for non-diabetic individuals occurs around a BMI of 25, indicating a higher concentration of non-diabetics in this range.Diabetic individuals show a peak frequency around a BMI of 30, suggesting that a higher BMI is associated with diabetes.There are significantly more non-diabetic individuals than diabetic individuals in the lower BMI ranges

(15-25).As BMI increases, the number of diabetic individuals rises, particularly noticeable in the 25-35 rangeThere is considerable overlap in the BMI distributions of both groups, especially around the BMI of 25-30, indicating that both non-diabetic and diabetic individuals can have similar BMI values



There is a significant peak in frequency for both diabetic and non-diabetic individuals at around 150 mg/dL, indicating that this glucose level is common among both groups.The red bars (diabetic) show a higher frequency than the blue bars (non-diabetic) at the 150 mg/dL mark, suggesting that more diabetic individuals have glucose levels around this value compared to non-diabetics.The non-diabetic group has a more spread-out distribution, with noticeable frequencies at lower glucose levels (around 100 mg/dL) and fewer individuals at higher levels (above 200 mg/dL). In contrast, the diabetic group has a more concentrated distribution around

the higher glucose levels.Both groups show lower frequencies at higher glucose levels (above 200 mg/dL), but the diabetic group has a slightly higher count compared to non-diabetics in this range.The data suggests that diabetic individuals tend to have higher blood glucose levels compared to non-diabetics, particularly around the 150 mg/dL mark, which may indicate a threshold for glucose levels that are more common in diabetics.



Average Age of Diabetic Cases

Trend: The average age of diabetic cases shows some fluctuations over the years. The average age started around 62 years in 2015, dropped to about 60 years in 2016, then gradually increased, peaking around more than 68 years between 2020 and 2021, before dropping again.

Average Blood Glucose Level of Diabetic Cases

Trend: The average blood glucose level of diabetic cases shows a noticeable spike.From 2015 to 2018, the average blood glucose level remained relatively stable around 180-190 mg/dL. However, there was a sharp increase in 2019, peaking at around 260 mg/dL, followed by a decrease in 2020 and 2021, then a slight increase again in 2022.

Average HbA1c Level of Diabetic Cases

Trend: The average HbA1c level also shows fluctuations over the years.The average HbA1c level was relatively stable from 2015 to 2017 around 6.8%. It then showed some variations, with a peak in 2018 and a dip in 2020, followed by another peak in 2022.

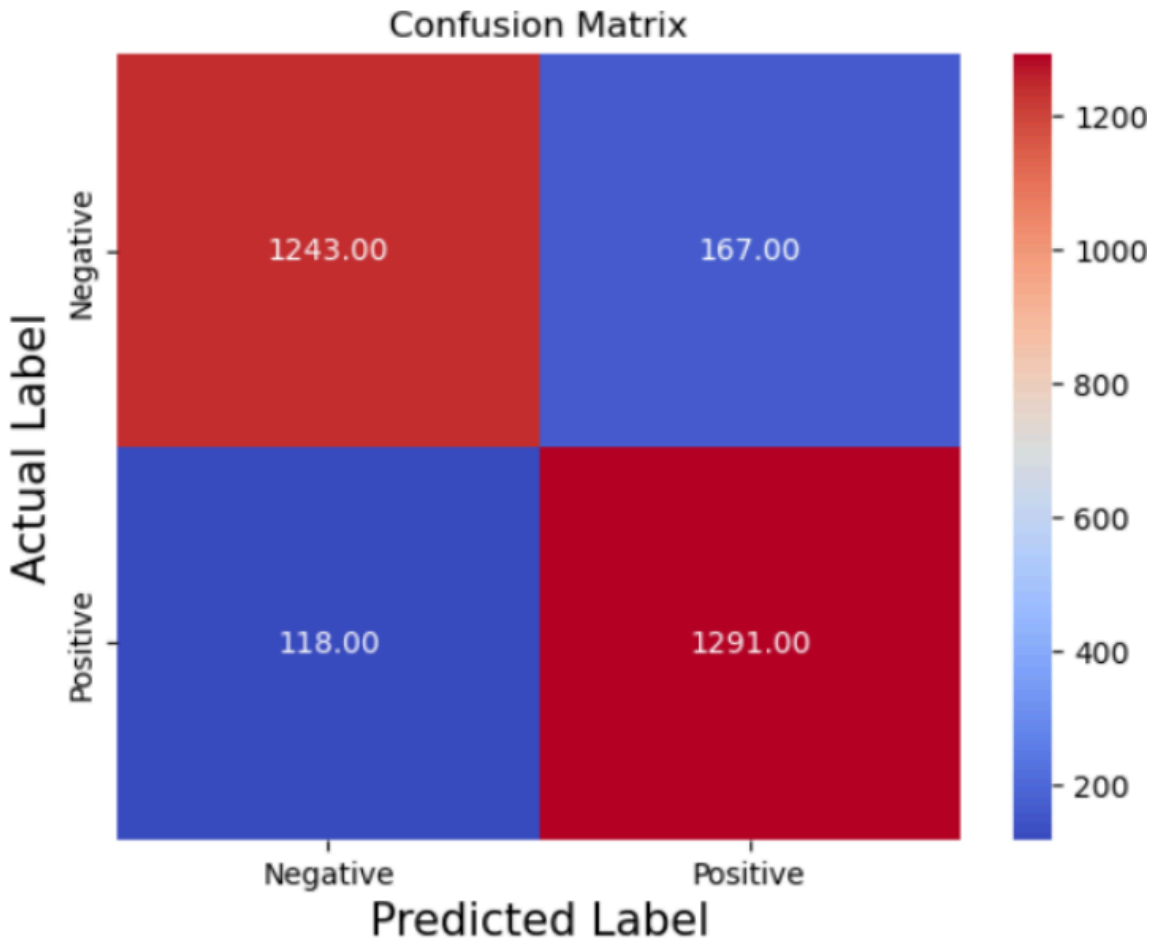
Average BMI of Diabetic Cases

Trend: the average BMI of diabetic individuals remained relatively stable, around 31, until 2019, after which it began to steadily decline, dropping significantly between 2020 and 2022. This decrease could suggest that diabetic patients are becoming more health-conscious or that there are other factors leading to lower BMI in recent years.

Machine Learning and Deployment

Model Selection

After exploring various algorithms, Gradient Boosting was selected due to its strong performance on the data and its ability to classify diabetic and non diabetic people with 92% Recall. The model was trained to predict whether an individual has diabetes based on the provided features and here is its performance using confusion Matrix:



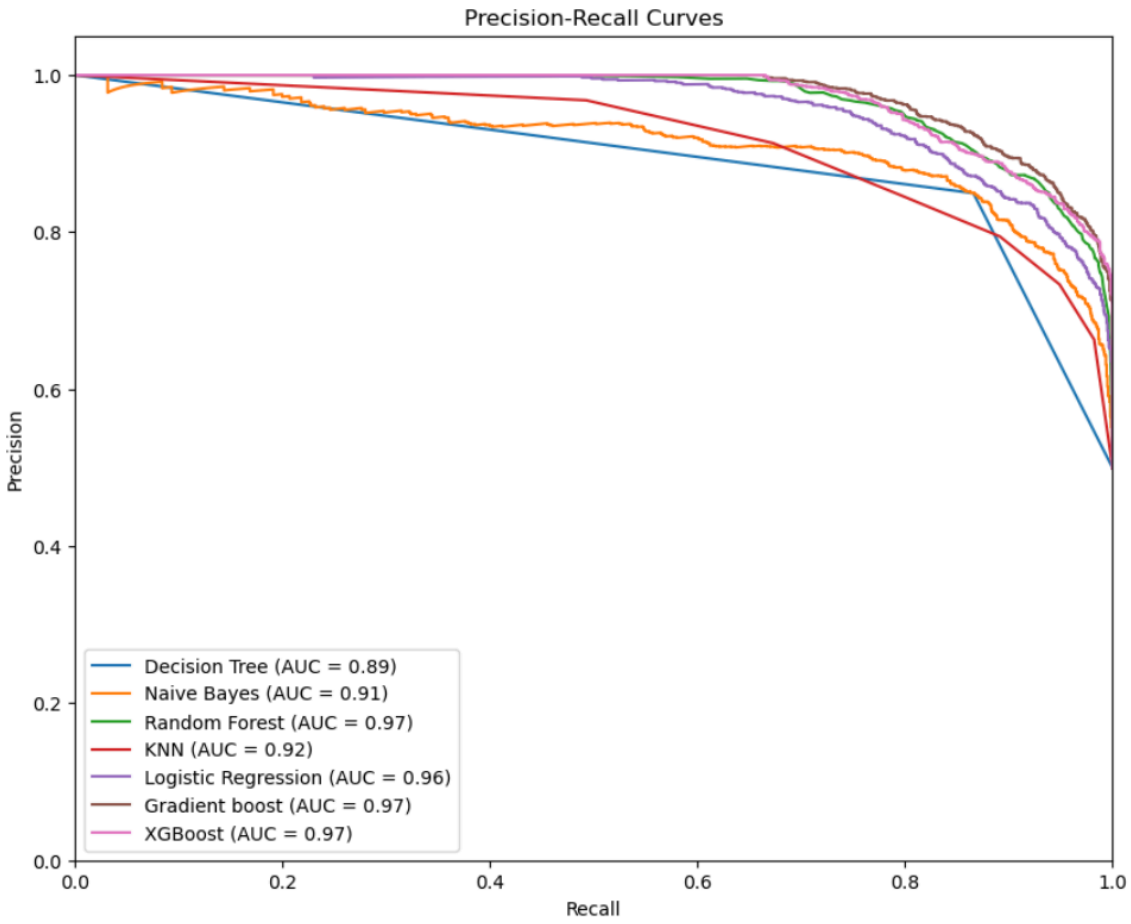
```
Model Accuracy: 89.89%
AUC-PR Score: 97.47%
Classification Report:
              precision    recall  f1-score   support

           0       0.91      0.88      0.90      1410
           1       0.89      0.92      0.90      1409

    accuracy                           0.90      2819
   macro avg       0.90      0.90      0.90      2819
weighted avg       0.90      0.90      0.90      2819
```

Model Evaluation Metrics:

- **Accuracy**: Percentage of correctly classified instances.
- **Precision and Recall**: Evaluated to ensure the model is reliable in both detecting diabetes (recall) and avoiding false positives (precision).
- **AUC-PR :** is a measure of the overall performance of a classification model. It is the area under the precision-recall curve, which represents the average precision at different recall levels. A higher AUC-PR indicates better performance.The AUC-PR is a more suitable metric than the AUC-ROC when dealing with imbalanced datasets, where the number of positive and negative examples is significantly different. This is because AUC-ROC can be misleading in such cases, as it may give a high score even if the model performs poorly on the minority class,There is all used models performance based on this metric



There were also two models that also performed well which are:

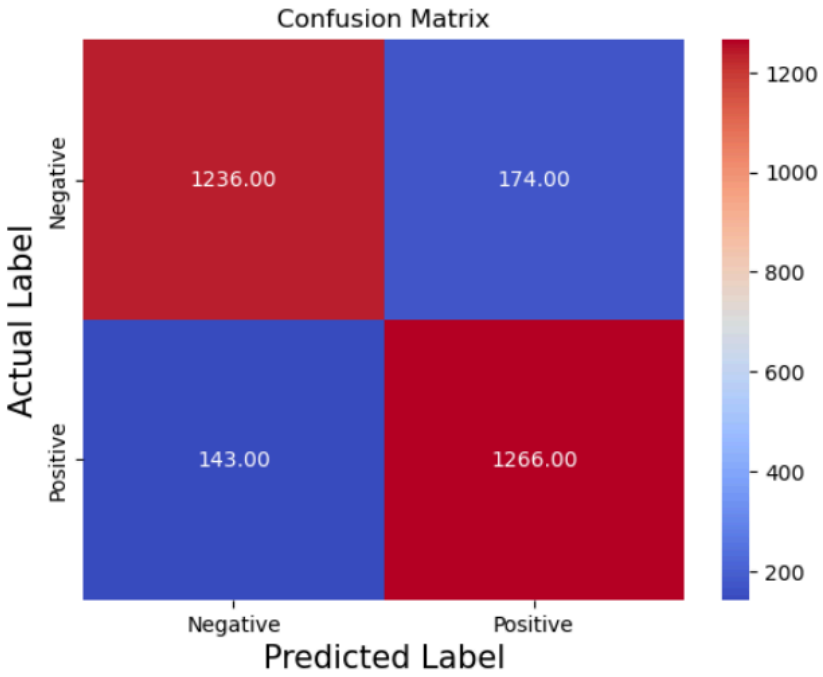
- XGboost Model
- Random Forest classifier

Here is the XG boost model performance:

```
Model Accuracy: 88.75%
AUC-PR Score: 97.06%
Classification Report:
              precision    recall  f1-score   support

           0       0.90      0.88      0.89      1410
           1       0.88      0.90      0.89      1409

    accuracy                           0.89      2819
   macro avg       0.89      0.89      0.89      2819
weighted avg       0.89      0.89      0.89      2819
```

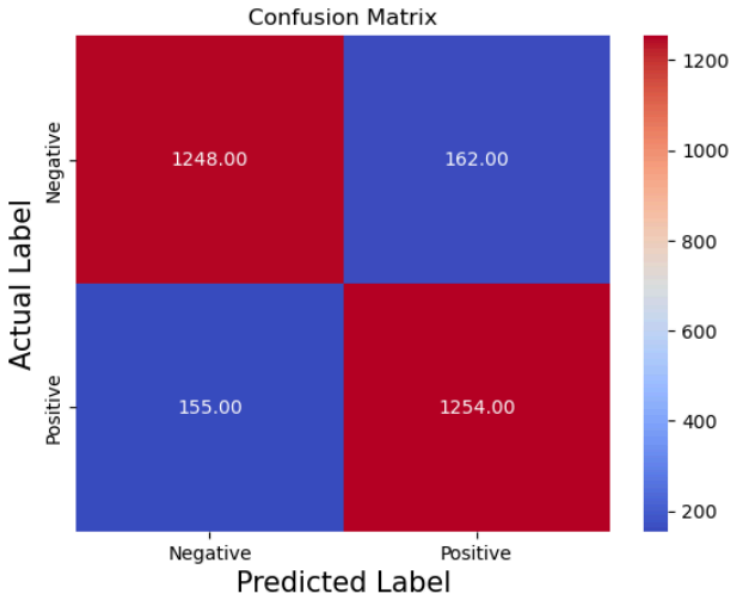


Here is the Random Forest Classifier model performance:

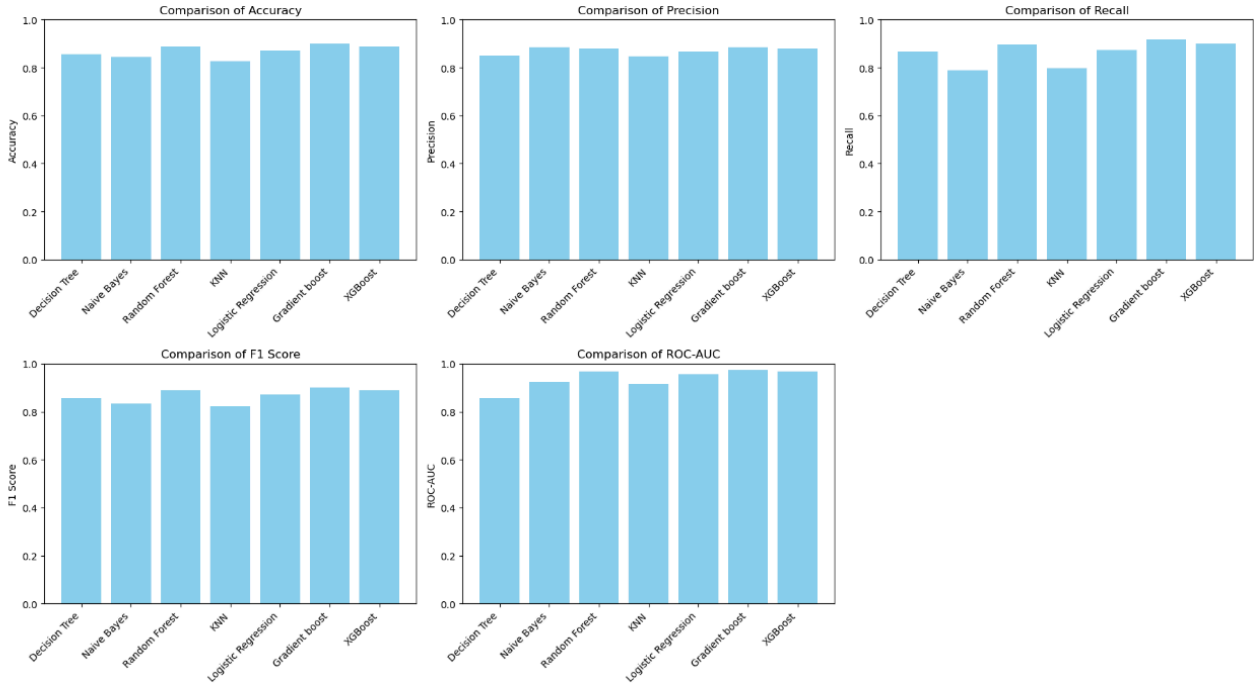
```
Model Accuracy: 88.75%
AUC-PR Score: 96.85%
Classification Report:
              precision    recall  f1-score   support

           0       0.89      0.89      0.89      1410
           1       0.89      0.89      0.89      1409

    accuracy                           0.89      2819
   macro avg       0.89      0.89      0.89      2819
weighted avg       0.89      0.89      0.89      2819
```



There is all models performance with different metrics



Application Deployment

The final model was deployed using Streamlit, a Python-based framework for building interactive web applications. The application features:

- Diagnosis Page: Allows users to input health data and receive a prediction on whether they are at risk of diabetes or not.
- Visualization Page: Provides visual insights into the dataset, showing distributions, correlations, and feature importance.

This app makes the project accessible to users with no coding background, demonstrating the practical value of machine learning in healthcare**.**

Conclusion

This project demonstrates the application of machine learning to a critical healthcare problem. The developed model and deployed application provide a useful tool for identifying individuals at risk of diabetes. Future improvements may include integrating more complex models, using larger datasets, and incorporating additional health metrics.

Project Github Repo: Omar10lfc/DEPI-Graduation-project-Diabetes-Diagnosis: Predictive Analytics for Diabetes Diagnosis and classification Using Machine Learning