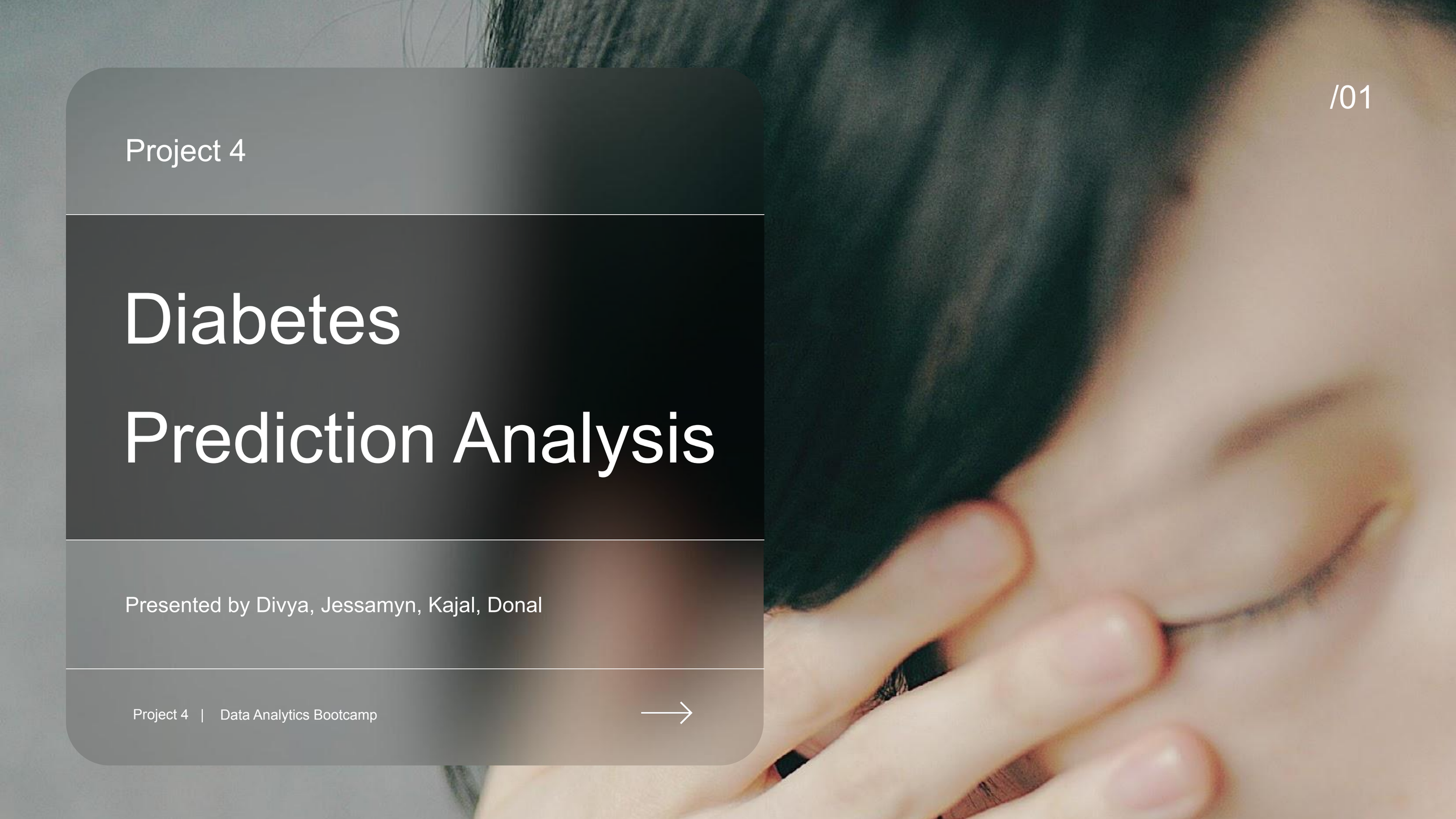
Project 4

Diabetes Prediction Analysis

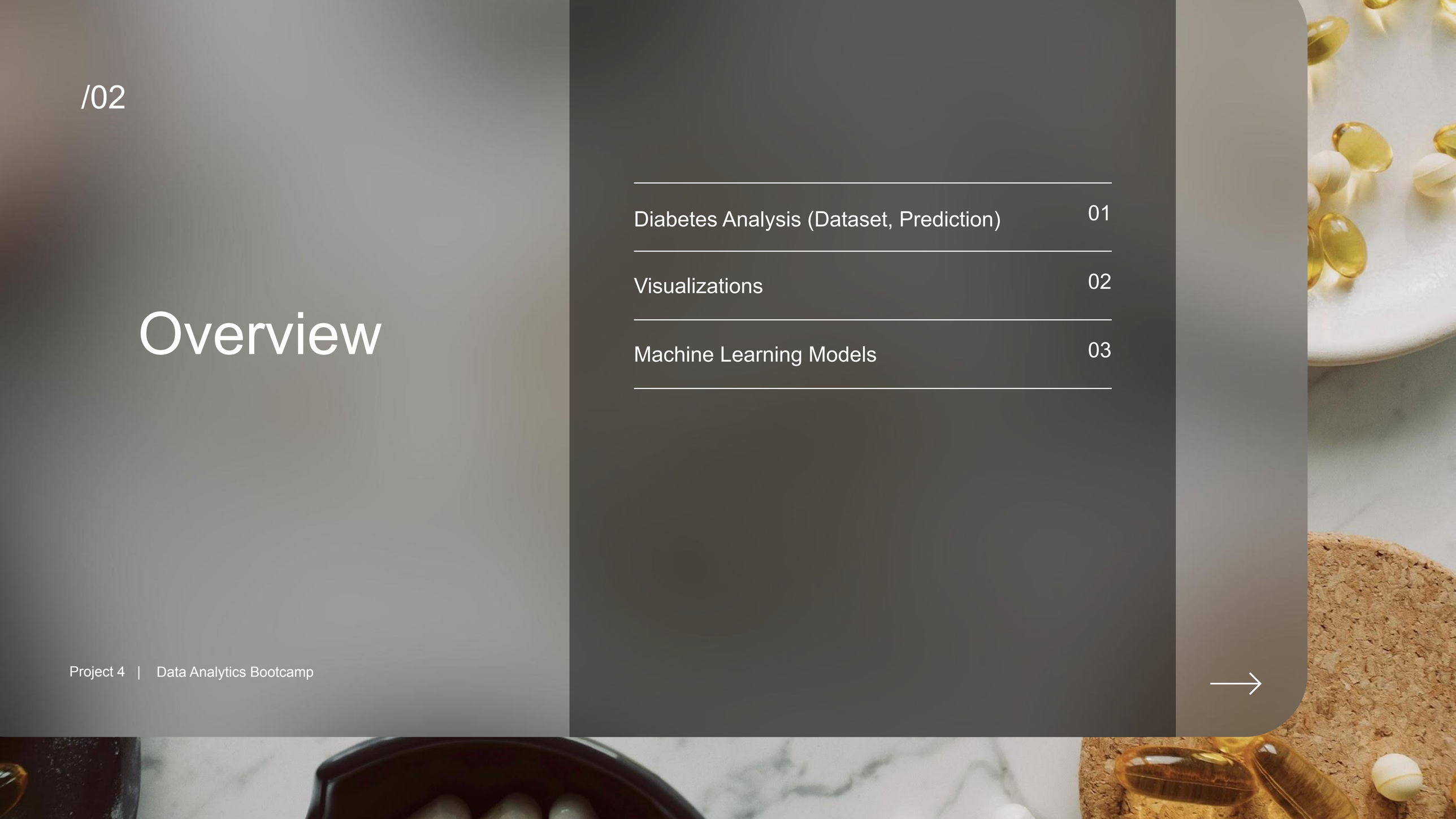
Presented by Divya, Jessamyn, Kajal, Donal



/02

Overview

Diabetes Analysis (Dataset, Prediction)	01
Visualizations	02
Machine Learning Models	03



Diabetes Summary and Prediction

Diabetes is a chronic metabolic disorder characterized by elevated levels of blood glucose (sugar). There are primarily two types of diabetes, Type 1 diabetes & Type 2.

For the purpose of this analysis we will be analyzing relationships of several features including Age, BMI, Blood Glucose levels, Hypertension, HbA1c Levels and Smoking History.

Purpose: predict the factors that may have a more significant impact on the risk of diabetes compared to others.

Keywords

Hypertension

A medical condition in which the blood pressure in the arteries is persistently elevated.

Heart disease

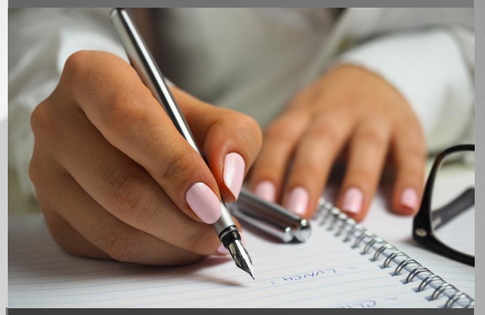
Medical condition that is associated with an increased risk of developing diabetes.

BMI (Body Mass Index)

Measures body fat based on weight and height.

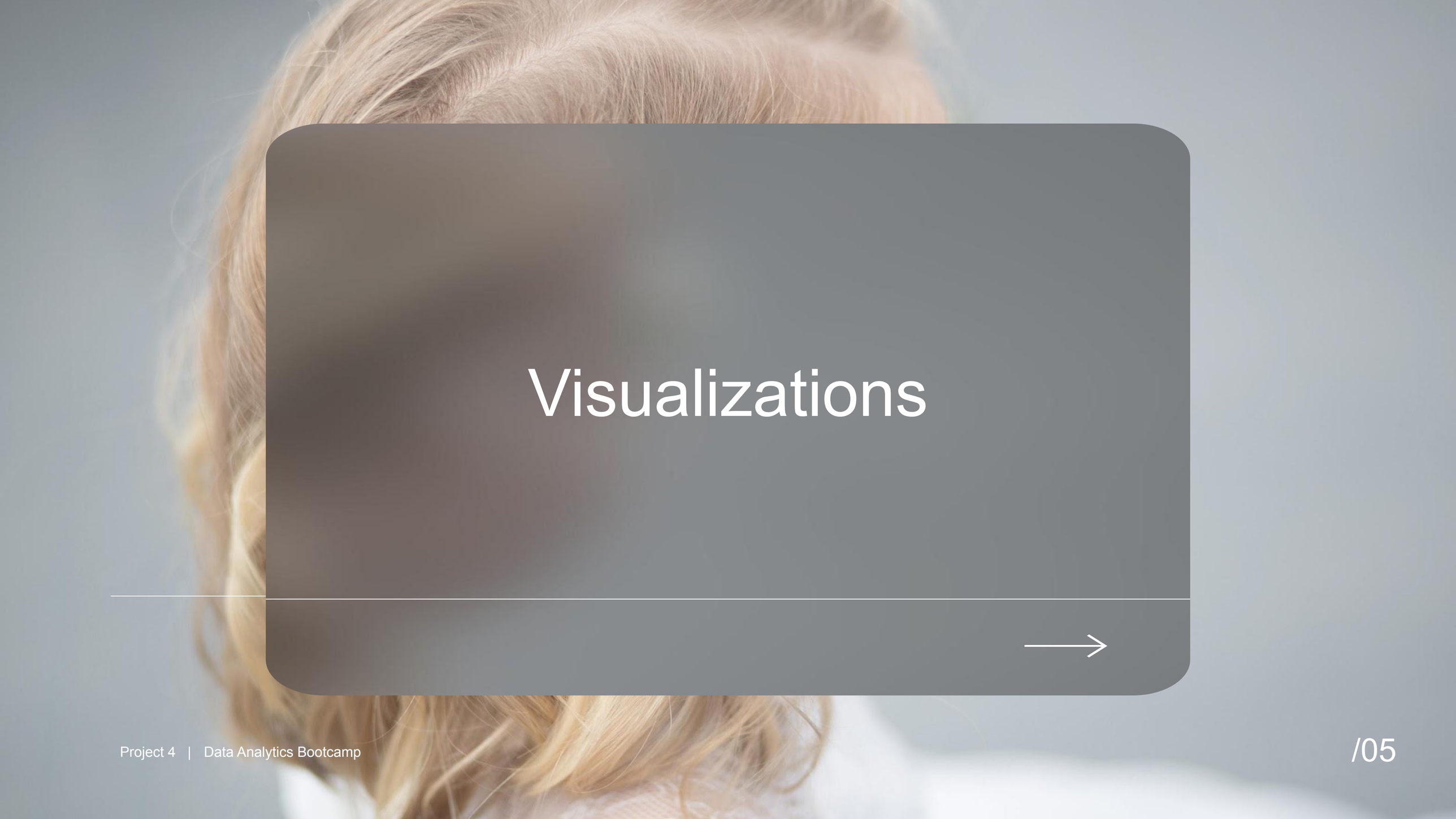
Blood glucose level

refers to the amount of glucose in the bloodstream at a given time.



HbA1c (Hemoglobin A1c) level

A measure of a person's average blood sugar level over the past 2-3 months.

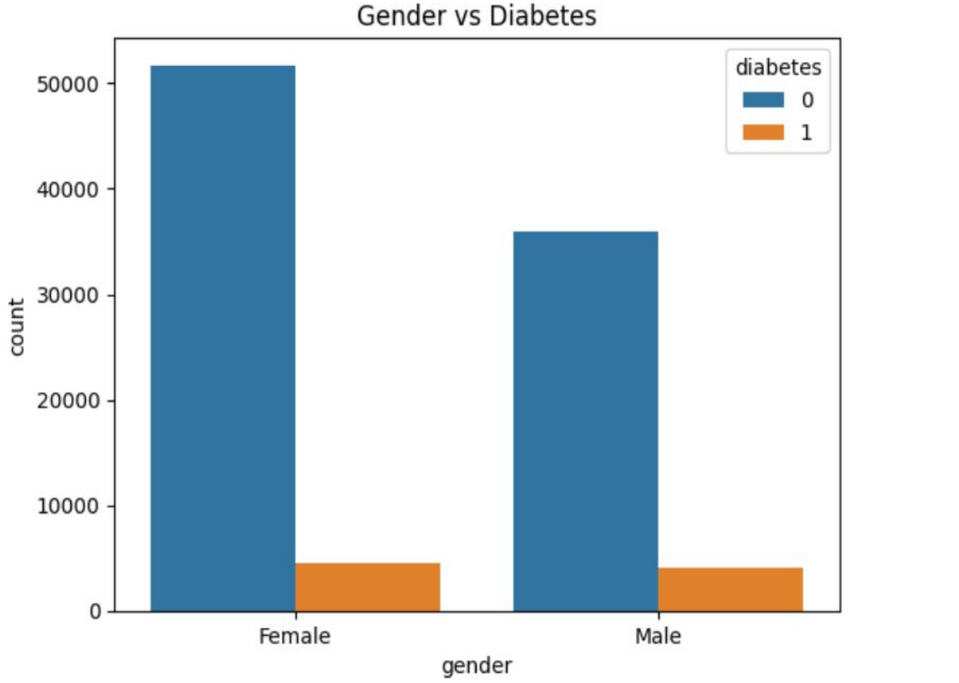


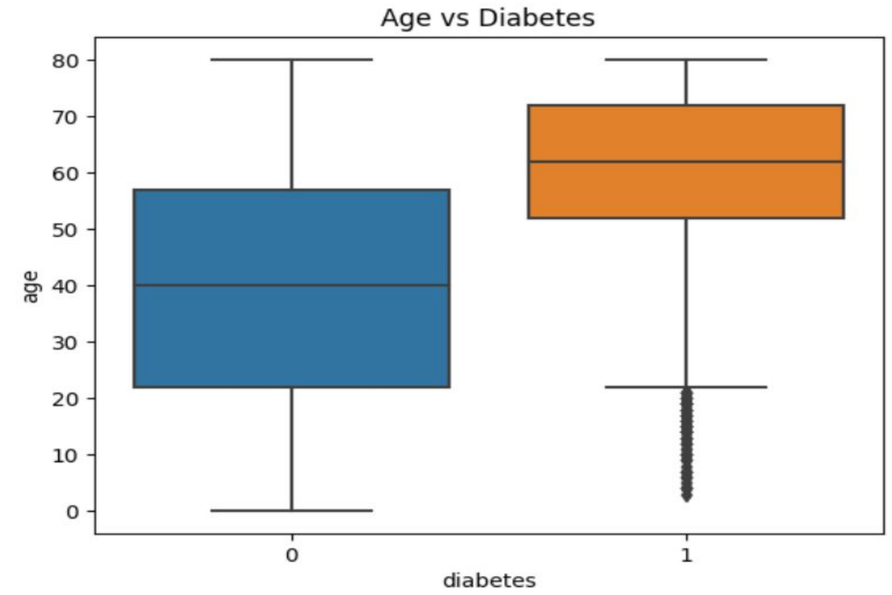
Visualizations

Gender vs. Diabetes Visualization
Age Vs. Diabetes Visualization

To **visualize** the relationship between gender and diabetes status (presence or absence of diabetes), we can use a bar chart to show the distribution of diabetic and non-diabetic individuals across different genders. This visualization will help us understand how diabetes prevalence varies between different gender categories regarding age.

To create a visualization of **Gender vs. Diabetes** and **Age vs. Diabetes** using Python and matplotlib, we can plot a bar chart and Box chart to show the distribution of diabetes cases among different gender categories regarding age.This will help us understand how diabetes prevalence varies between male, female, and other genders in the dataset.





Gender Vs. Diabetes

Interpretation:

- The x-axis represents the two genders (Male, Female).
- The y-axis represents the diabetes prevalence percentage.
- Each bar's height corresponds to the respective diabetes prevalence for males and females.
- Color differentiation (e.g., blue for Male, pink for Female) helps visually distinguish between categories.
- This bar chart clearly shows that diabetes prevalence is slightly higher among males (12%) compared to females (10%).
- The visual representation makes it easy to compare and interpret the differences in diabetes prevalence between genders.

Age Vs. Diabetes

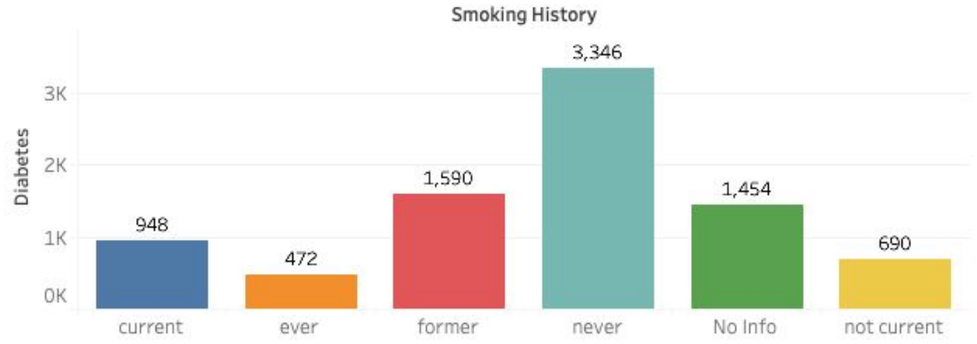
- The box plot allows for a visual comparison of diabetes prevalence across different age groups.
- You can interpret the median (middle line of the box) and the spread of the data (interquartile range) within each age group.
- Outliers can be identified, providing insights into extreme values of diabetes prevalence within specific age categories.

Diabetes Prediction Analysis

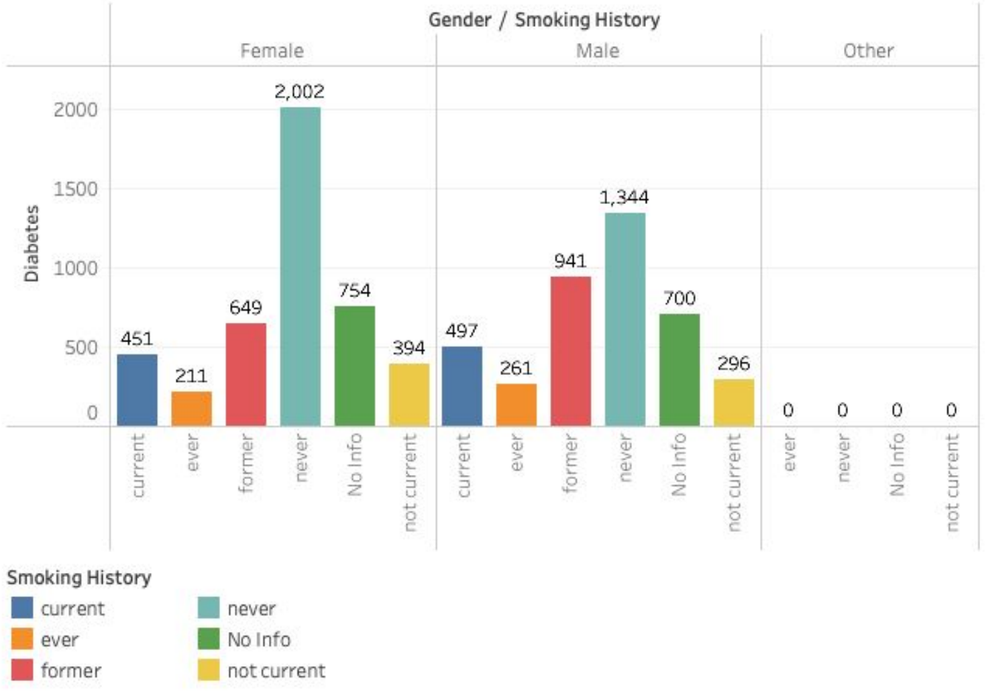
Introduction | Diabetes Analysis | Smoking Habits | BMI & Heartdisease | Other Possible Variables

Diabetes, is a chronic metabolic disorder characterized by elevated levels of blood glucose (sugar). The development of diabetes can be influenced by a myriad of factors, making it a multifaceted disease. The analysis aims to identify and understand how specific features or risk factors contribute to an increased likelihood of developing diabetes. By examining these features, the goal is to predict which factors may have a more significant impact on the risk of diabetes compared to others.

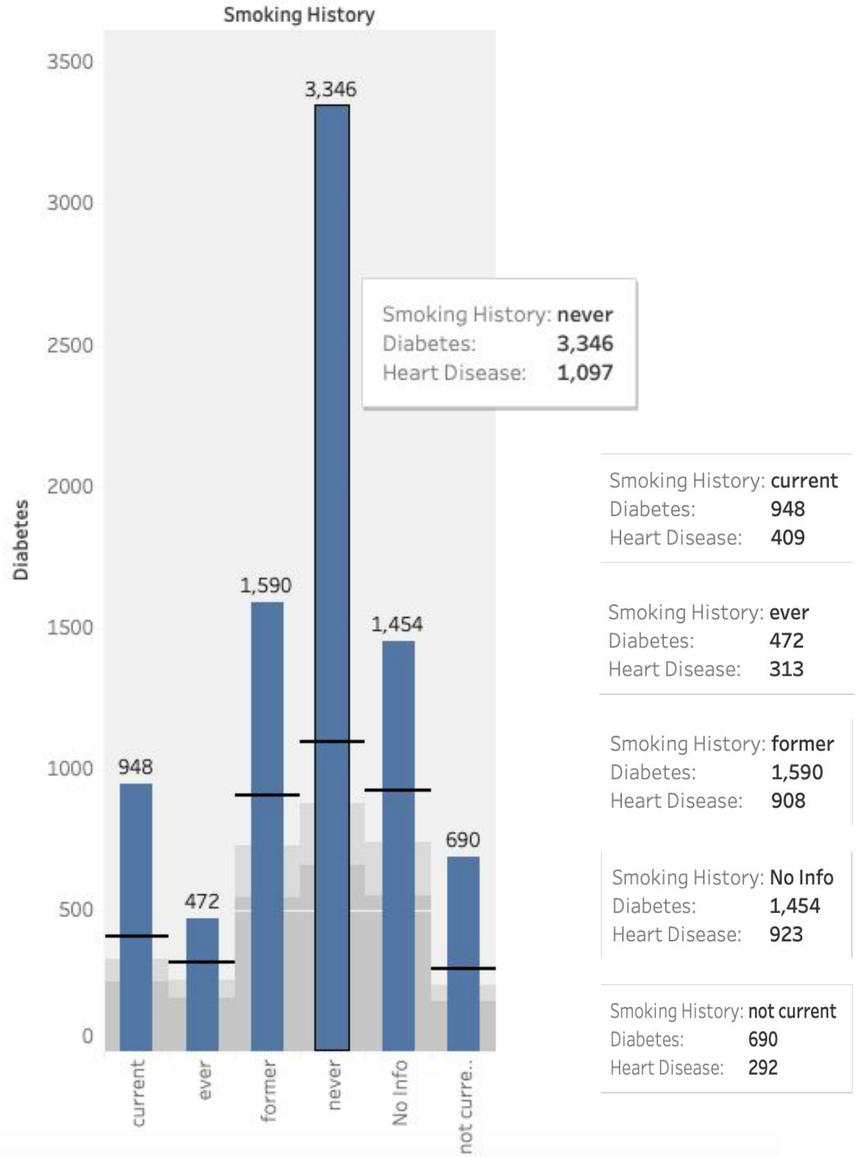
Smoking History vs. Diabetes



Smoking History vs. Diabetes by Gender



Smoking History with underlying Heart Disease



Smoking History: **never**
Diabetes: **3,346**
Heart Disease: **1,097**

Smoking History: **current**
Diabetes: **948**
Heart Disease: **409**

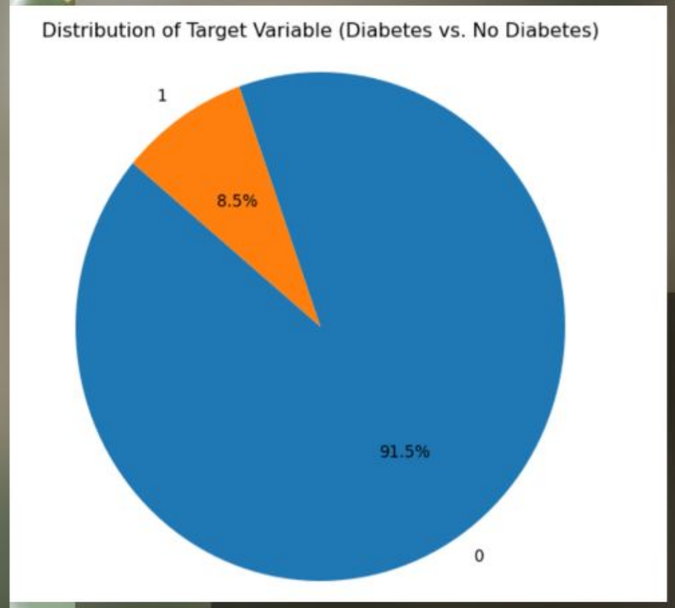
Smoking History: **ever**
Diabetes: **472**
Heart Disease: **313**

Smoking History: **former**
Diabetes: **1,590**
Heart Disease: **908**

Smoking History: **No Info**
Diabetes: **1,454**
Heart Disease: **923**

Smoking History: **not current**
Diabetes: **690**
Heart Disease: **292**

Machine Learning Models



Logistic Regression

- Simplicity
- Interpretability
- Effectiveness for binary classification tasks

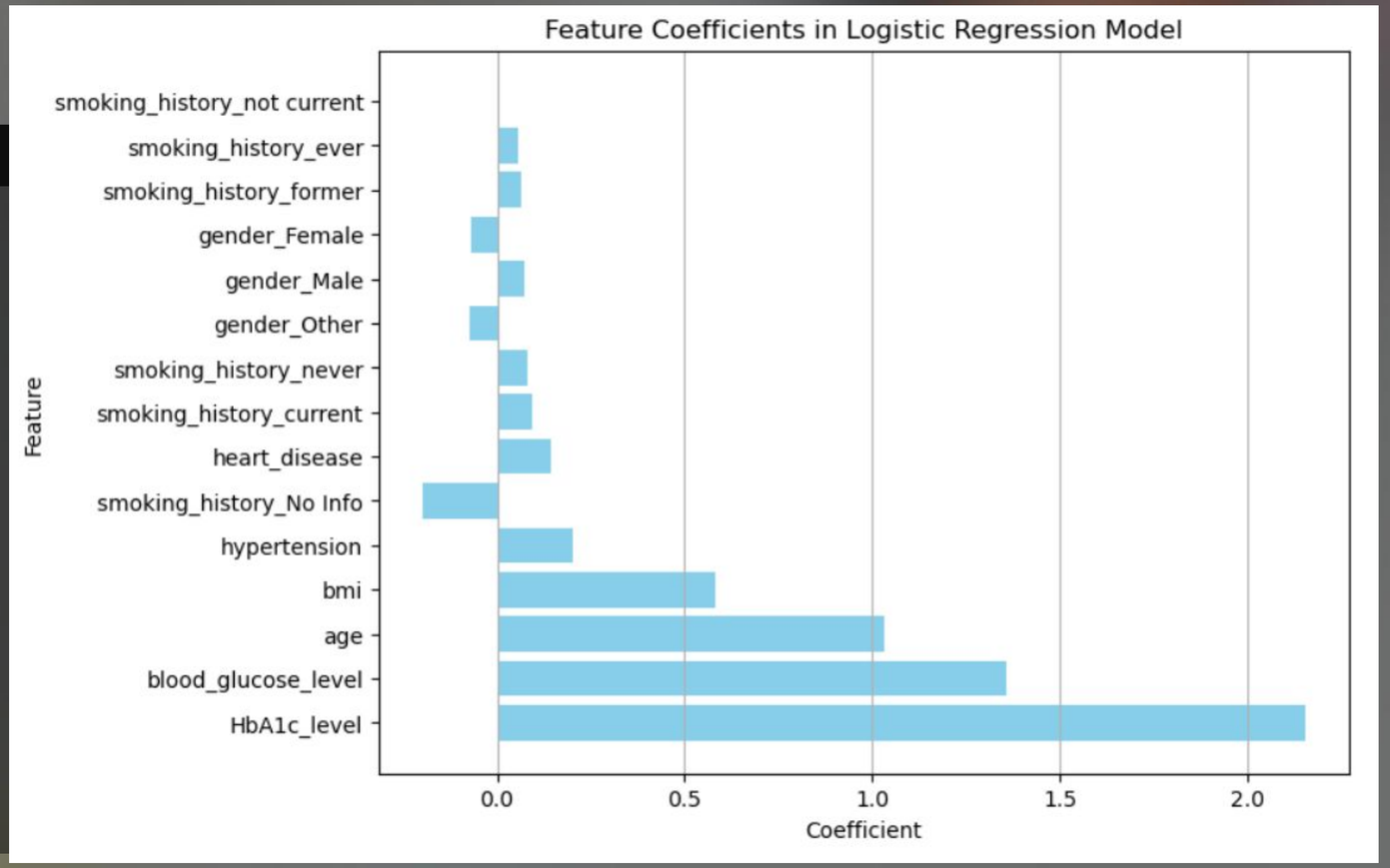
Random Forest Classifier

- Ability to handle complex data
- Feature importance analysis
- Resistance to overfitting

Logistic Regression

Coefficients to Identify Important Features

- 'HbA1c_level', 'blood_glucose_level', and 'age' have the largest absolute coefficients
- Positive coefficients - 'hypertension', 'heart_disease', and certain smoking history categories.
- Negative coefficients - 'smoking_history_No Info', 'gender_Other', and 'gender_Female'



Logistic Regression - Confusion Matrix

Accuracy: 95.78%

Correctly Classified

It has a high number of true positives and true negatives, indicating effective classification performance

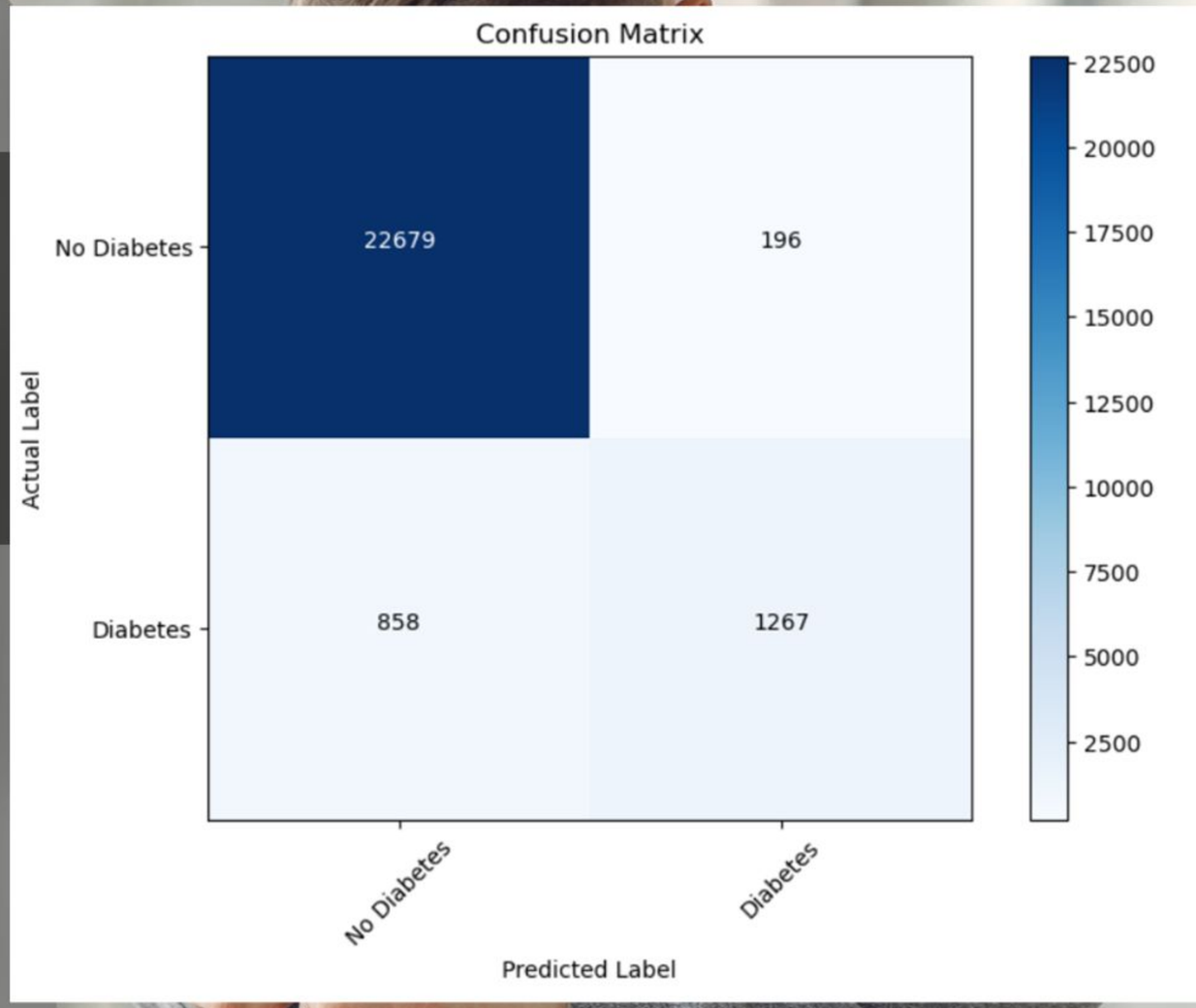
Incorrectly Classified

The presence of false positives (FP = 196) indicates instances incorrectly classified as positive, while false negatives (FN = 858) indicate instances incorrectly classified as negative.

Model’s Precision and Recall

The proportion of true positive predictions out of all positive predictions indicates that Precision is reasonably high.

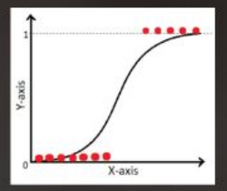
The proportion of true positive predictions out of all actual positives indicates a lower Recall.



Regularization Techniques – Lasso (l1) and Ridge (l2)

Logistic Regression

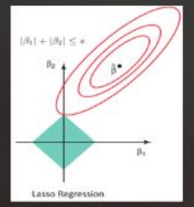
Logistic Regression



Classification Report:

	precision	recall	f1-score	support
0	0.96	0.99	0.98	22875
1	0.87	0.60	0.71	2125
accuracy			0.96	25000
macro avg	0.91	0.79	0.84	25000
weighted avg	0.96	0.96	0.95	25000

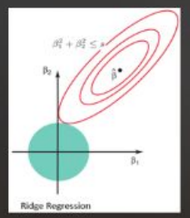
Lasso (L1) Regularization



Classification Report (L1 Regularization):

	precision	recall	f1-score	support
0	0.96	0.99	0.98	22875
1	0.87	0.60	0.71	2125
accuracy			0.96	25000
macro avg	0.91	0.79	0.84	25000
weighted avg	0.96	0.96	0.95	25000

Ridge (L2) Regularization



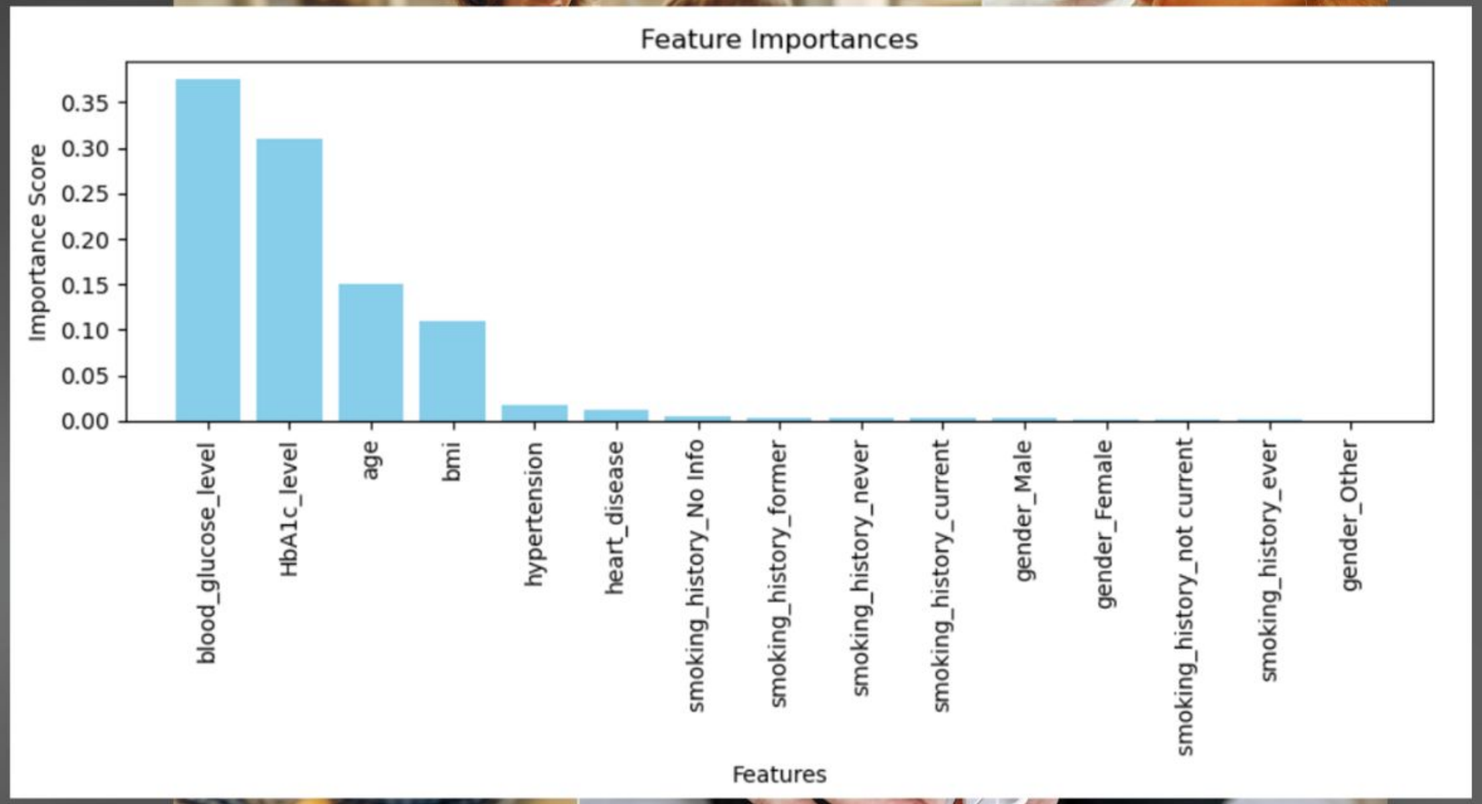
Classification Report (L2 Regularization):

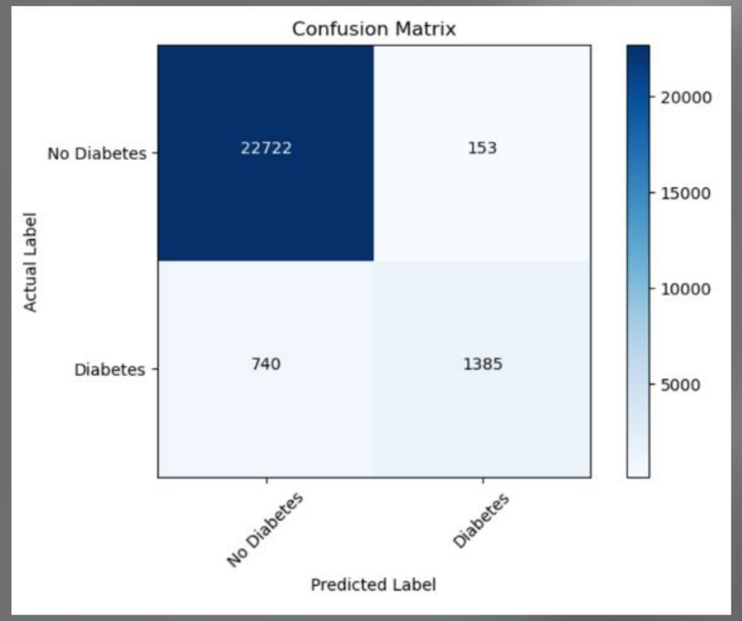
	precision	recall	f1-score	support
0	0.96	0.99	0.98	22875
1	0.87	0.60	0.71	2125
accuracy			0.96	25000
macro avg	0.91	0.79	0.84	25000
weighted avg	0.96	0.96	0.95	25000

Random Forest Classifier Model

Top 3 features contributing to the target variable diabetes :

- blood_glucose_level – highest importance score suggesting key indicator.
- HbA1c_level – Hemoglobin A1c (HbA1c) level is another crucial predictor associated with poorly controlled diabetes.
- Age - significant predictor of diabetes risk, associated with changes in metabolism, hormone levels, and lifestyle factors.





Accuracy Score : 0.96428
Classification Report

	precision	recall	f1-score	support
0	0.97	0.99	0.98	22875
1	0.90	0.65	0.76	2125
accuracy			0.96	25000
macro avg	0.93	0.82	0.87	25000
weighted avg	0.96	0.96	0.96	25000

The model correctly identified:

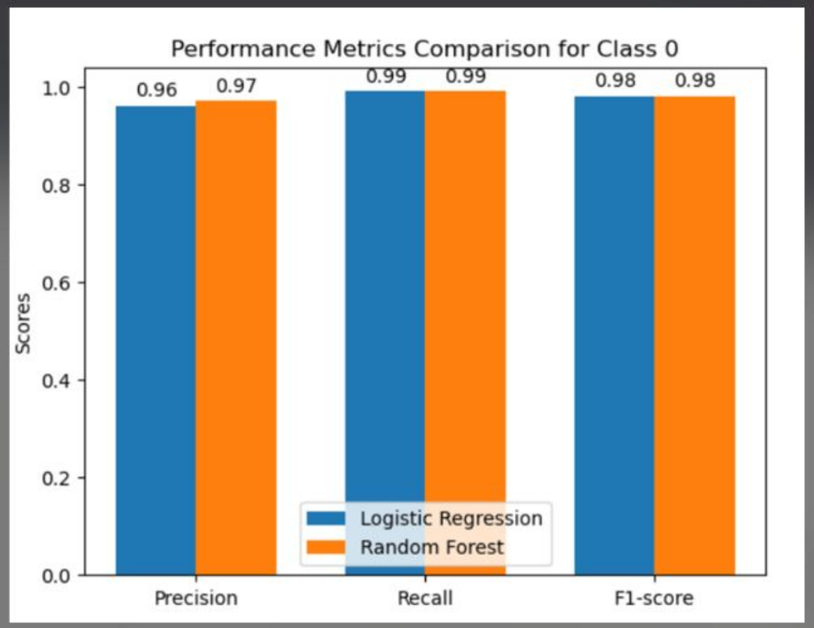
- 22,722 instances as negative (class 0) and labeled them correctly (True Negatives).
- 1,385 instances as positive (class 1) and labeled them correctly (True Positives).

The model incorrectly classified:

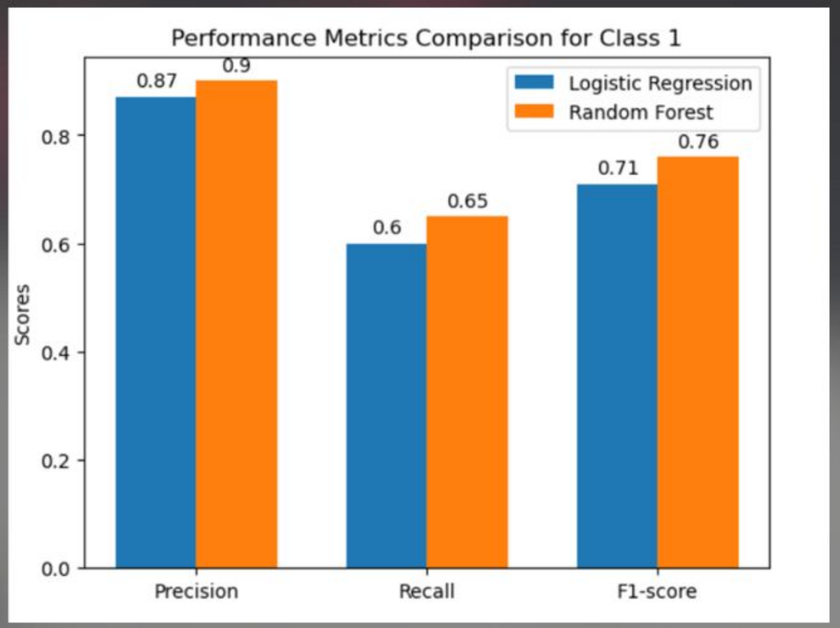
- 153 instances as positive (class 1) when they were actually negative (False Positives).
- 740 instances as negative (class 0) when they were actually positive (False Negatives).

Comparison of Logistic Regression vs Random Forest Model Performance

Performance Metrics of Class 0



Performance Metrics of Class 1



Neural Network

- **Dataset Overview:**

- Contains 100,000 entries.
- Consists 0f 9 columns in total.

- **Dataset Types:**

- Three columns are of type float64 : Age, BMI(Body Mass Index),HbA1c_level (Hemoglobin A1c level)
- Four columns are of type int64:Hypertension, Heart disease, Blood glucose level, Diabetes
- Two columns are of type object:Gender, Smoking history

- **Dataset Preprocessing:**

- Categorical Varible Encoding
- Standardization
- Train-Test-Split

Visualization of Neural Network Architecture

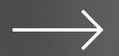
Model: "sequential"

Layer (type)	Output Shape	Param #
dense (Dense)	(None, 64)	576
dense_1 (Dense)	(None, 32)	2,080
dense_2 (Dense)	(None, 1)	33

Total params: 2,689 (10.50 KB)

Trainable params: 2,689 (10.50 KB)

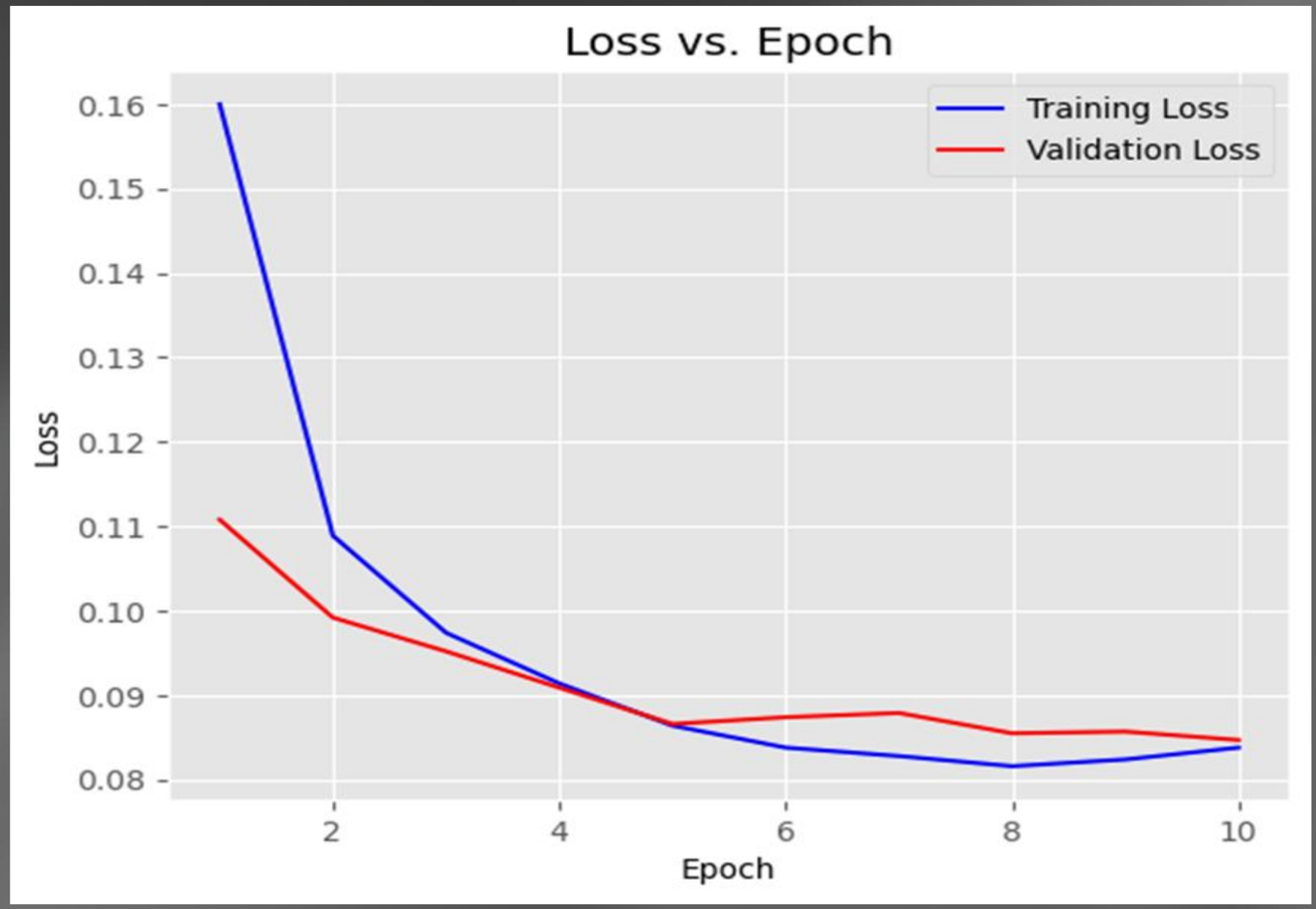
Non-trainable params: 0 (0.00 B)



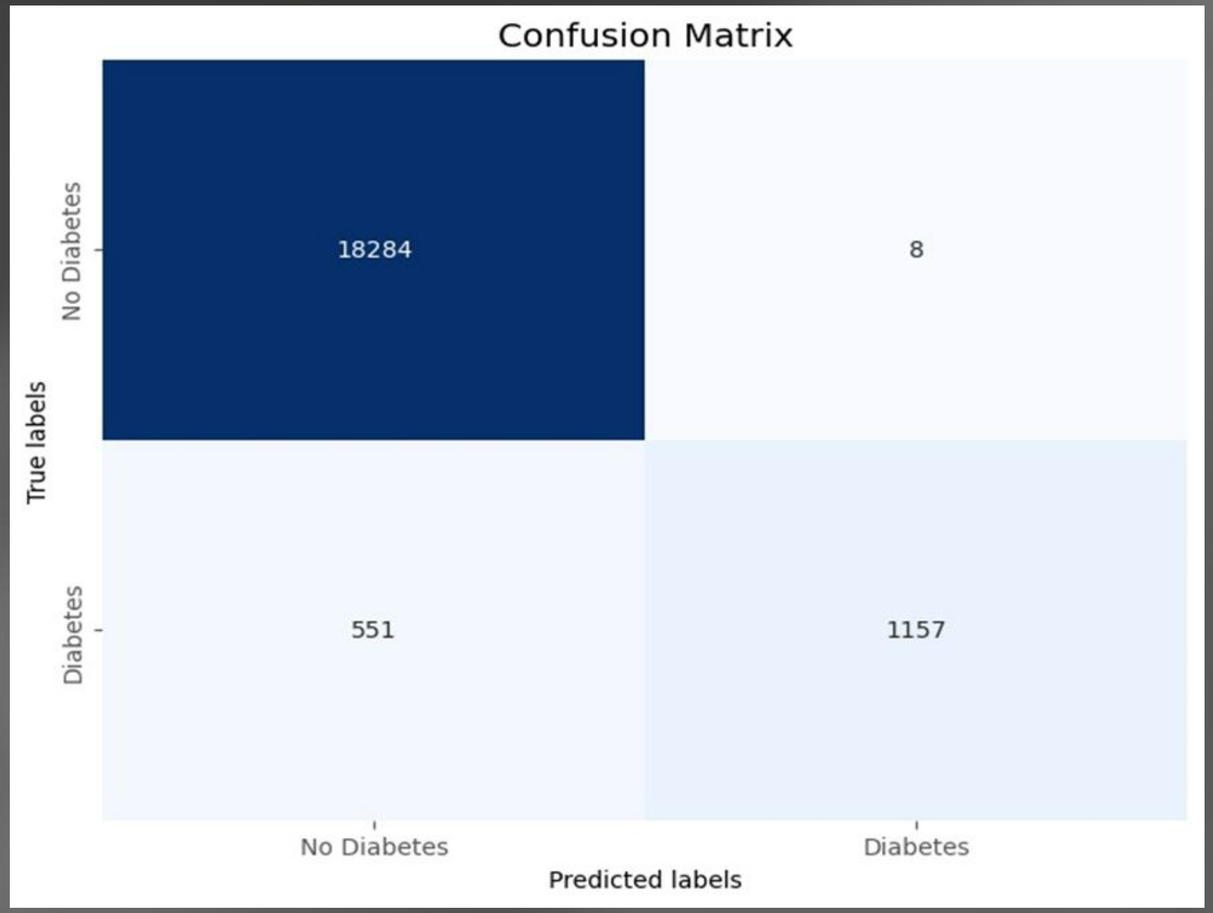
Model Performance


```
Epoch 1/10
2000/2000 ━━━━━━━━━━━━━━━━━━━━ 3s 1ms/step - accuracy: 0.9718 - loss: 0.0810 - val_accuracy: 0.9711 - val_loss: 0.0848
Epoch 2/10
2000/2000 ━━━━━━━━━━━━━━━━━━━━ 3s 1ms/step - accuracy: 0.9722 - loss: 0.0812 - val_accuracy: 0.9686 - val_loss: 0.0901
Epoch 3/10
2000/2000 ━━━━━━━━━━━━━━━━━━━━ 3s 1ms/step - accuracy: 0.9722 - loss: 0.0802 - val_accuracy: 0.9709 - val_loss: 0.0861
Epoch 4/10
2000/2000 ━━━━━━━━━━━━━━━━━━━━ 3s 1ms/step - accuracy: 0.9721 - loss: 0.0819 - val_accuracy: 0.9711 - val_loss: 0.0852
Epoch 5/10
2000/2000 ━━━━━━━━━━━━━━━━━━━━ 3s 1ms/step - accuracy: 0.9729 - loss: 0.0771 - val_accuracy: 0.9710 - val_loss: 0.0856
Epoch 6/10
2000/2000 ━━━━━━━━━━━━━━━━━━━━ 3s 1ms/step - accuracy: 0.9714 - loss: 0.0821 - val_accuracy: 0.9698 - val_loss: 0.0863
Epoch 7/10
2000/2000 ━━━━━━━━━━━━━━━━━━━━ 3s 1ms/step - accuracy: 0.9710 - loss: 0.0814 - val_accuracy: 0.9711 - val_loss: 0.0846
Epoch 8/10
2000/2000 ━━━━━━━━━━━━━━━━━━━━ 3s 1ms/step - accuracy: 0.9703 - loss: 0.0836 - val_accuracy: 0.9709 - val_loss: 0.0864
Epoch 9/10
2000/2000 ━━━━━━━━━━━━━━━━━━━━ 3s 1ms/step - accuracy: 0.9722 - loss: 0.0790 - val_accuracy: 0.9694 - val_loss: 0.0872
Epoch 10/10
2000/2000 ━━━━━━━━━━━━━━━━━━━━ 3s 1ms/step - accuracy: 0.9727 - loss: 0.0795 - val_accuracy: 0.9709 - val_loss: 0.0851
```

Model Performance Visualization



Visualization of Model Predictions



Accuracy 97.21%

Model Evaluation Metrics

Primary Metric Evaluation:
Accuracy measures correct classifications out of total instances.
Additional metrics like precision, recall, and F1-score provide deeper insights.

Performance on Test Dataset:
Achieved 97% accuracy on the test dataset.
Other metrics like precision, recall, and F1-score offer comprehensive evaluation.

Insights from Model Evaluation:
High accuracy suggests effective learning of data patterns.
Further analysis with precision, recall, and F1-score aids in understanding model behavior.
Model's strong performance indicates potential for predicting diabetes onset.

Thank you for listening!



References:

https://www.kaggle.com/code/tumpanjawat/diabetes-eda-random-forest-hp

https://public.tableau.com/app/profile/jessamyn.bacani/viz/DiabetesPredictionAnalysis_17137381253720/Story1?publish=yes