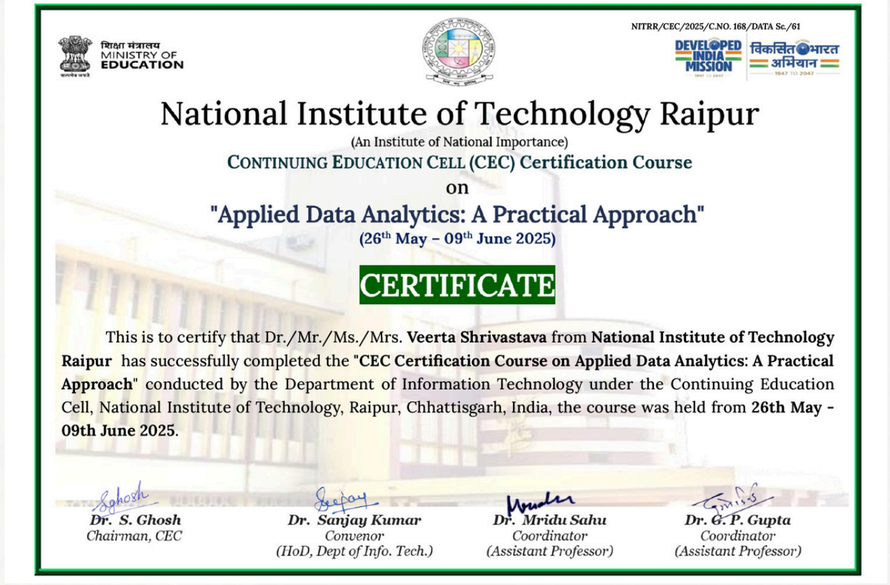
PROJECT REPORT ON

DIABETA

ML-BASED PREDICTION SYSTEM

NAME	ROLL NO.
VEERTA SHRIVASTAVA	23118108
SWARA MANDALE	23118102
NIDHI SAHU	23118070
CHANDRAKANT VERMA	23118024

CERTIFICATE OF COMPLETION:



शिक्षा मंत्रालय
MINISTRY OF EDUCATION

NITRR/CEC/2025/C.NO. 168/DATA Sc./61

DEVELOPED INDIA MISSION | विकसित भारत अभियान

National Institute of Technology Raipur

(An Institute of National Importance)

CONTINUING EDUCATION CELL (CEC) Certification Course

on

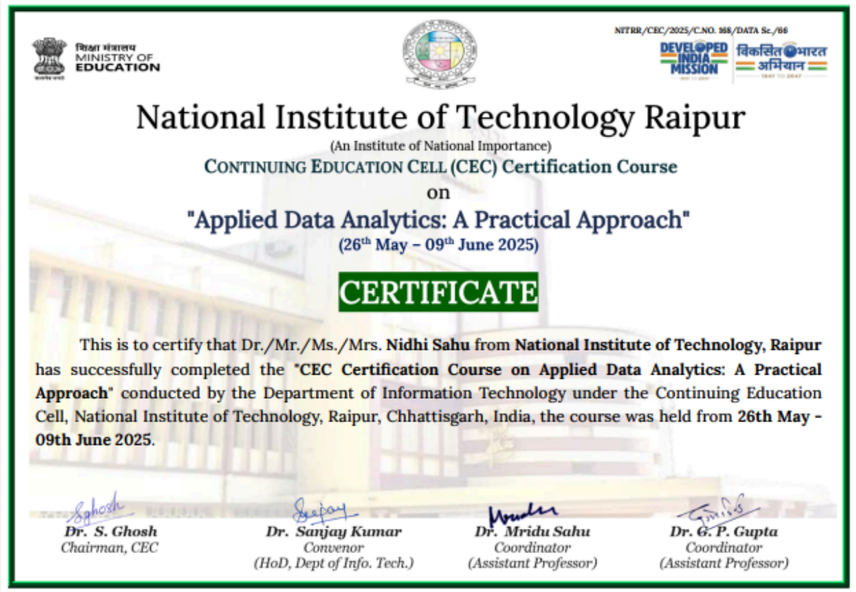
"Applied Data Analytics: A Practical Approach"

(26th May – 09th June 2025)

CERTIFICATE

This is to certify that Dr./Mr./Ms./Mrs. **Veerta Shrivastava** from **National Institute of Technology Raipur** has successfully completed the "**CEC Certification Course on Applied Data Analytics: A Practical Approach**" conducted by the Department of Information Technology under the Continuing Education Cell, National Institute of Technology, Raipur, Chhattisgarh, India, the course was held from **26th May - 09th June 2025**.

Dr. S. Ghosh	Dr. Sanjay Kumar	Dr. Mridu Sahu	Dr. G. P. Gupta
Chairman, CEC	Convenor (HoD, Dept of Info. Tech.)	Coordinator (Assistant Professor)	Coordinator (Assistant Professor)



शिक्षा मंत्रालय
MINISTRY OF EDUCATION

NITRR/CEC/2025/C.NO. 168/DATA Sc./66

DEVELOPED INDIA MISSION | विकसित भारत अभियान

National Institute of Technology Raipur

(An Institute of National Importance)

CONTINUING EDUCATION CELL (CEC) Certification Course

on

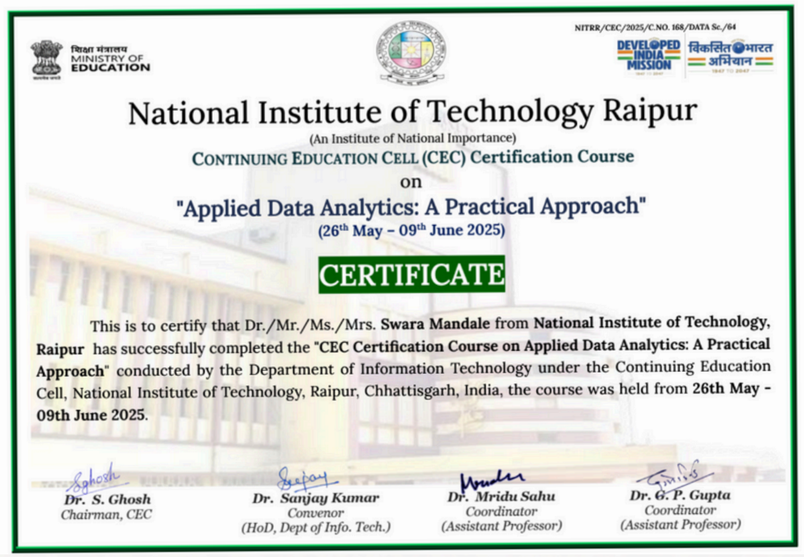
"Applied Data Analytics: A Practical Approach"

(26th May – 09th June 2025)

CERTIFICATE

This is to certify that Dr./Mr./Ms./Mrs. **Nidhi Sahu** from **National Institute of Technology, Raipur** has successfully completed the "**CEC Certification Course on Applied Data Analytics: A Practical Approach**" conducted by the Department of Information Technology under the Continuing Education Cell, National Institute of Technology, Raipur, Chhattisgarh, India, the course was held from **26th May - 09th June 2025**.

Dr. S. Ghosh	Dr. Sanjay Kumar	Dr. Mridu Sahu	Dr. G. P. Gupta
Chairman, CEC	Convenor (HoD, Dept of Info. Tech.)	Coordinator (Assistant Professor)	Coordinator (Assistant Professor)



शिक्षा मंत्रालय
MINISTRY OF EDUCATION

NITRR/CEC/2025/C.NO. 168/DATA Sc./64

DEVELOPED INDIA MISSION | विकसित भारत अभियान

National Institute of Technology Raipur

(An Institute of National Importance)

CONTINUING EDUCATION CELL (CEC) Certification Course

on

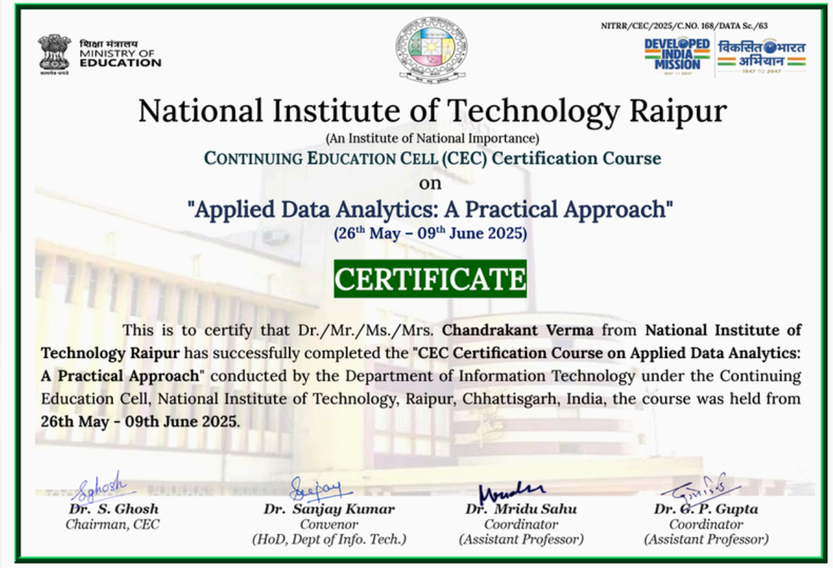
"Applied Data Analytics: A Practical Approach"

(26th May – 09th June 2025)

CERTIFICATE

This is to certify that Dr./Mr./Ms./Mrs. **Swara Mandale** from **National Institute of Technology, Raipur** has successfully completed the "**CEC Certification Course on Applied Data Analytics: A Practical Approach**" conducted by the Department of Information Technology under the Continuing Education Cell, National Institute of Technology, Raipur, Chhattisgarh, India, the course was held from **26th May - 09th June 2025**.

Dr. S. Ghosh	Dr. Sanjay Kumar	Dr. Mridu Sahu	Dr. G. P. Gupta
Chairman, CEC	Convenor (HoD, Dept of Info. Tech.)	Coordinator (Assistant Professor)	Coordinator (Assistant Professor)



शिक्षा मंत्रालय
MINISTRY OF EDUCATION

NITRR/CEC/2025/C.NO. 168/DATA Sc./63

DEVELOPED INDIA MISSION | विकसित भारत अभियान

National Institute of Technology Raipur

(An Institute of National Importance)

CONTINUING EDUCATION CELL (CEC) Certification Course

on

"Applied Data Analytics: A Practical Approach"

(26th May – 09th June 2025)

CERTIFICATE

This is to certify that Dr./Mr./Ms./Mrs. **Chandrakant Verma** from **National Institute of Technology Raipur** has successfully completed the "**CEC Certification Course on Applied Data Analytics: A Practical Approach**" conducted by the Department of Information Technology under the Continuing Education Cell, National Institute of Technology, Raipur, Chhattisgarh, India, the course was held from **26th May - 09th June 2025**.

Dr. S. Ghosh	Dr. Sanjay Kumar	Dr. Mridu Sahu	Dr. G. P. Gupta
Chairman, CEC	Convenor (HoD, Dept of Info. Tech.)	Coordinator (Assistant Professor)	Coordinator (Assistant Professor)

TABLE OF CONTENT

About the Course – Applied Data Analytics

- THE CEC CERTIFICATION COURSE ON “APPLIED DATA ANALYTICS: A PRACTICAL APPROACH” WAS ORGANIZED BY THE DEPARTMENT OF INFORMATION TECHNOLOGY, NIT RAIPUR, UNDER THE CONTINUING EDUCATION CELL (CEC).

- THE COURSE WAS CONDUCTED FROM 26TH MAY TO 9TH JUNE 2025, FOCUSING ON PRACTICAL APPLICATIONS OF DATA ANALYTICS AND MACHINE LEARNING.

- IT PROVIDED PARTICIPANTS WITH HANDS-ON EXPERIENCE IN DATA PREPROCESSING, VISUALIZATION, MODEL BUILDING, AND EVALUATION USING MODERN ANALYTICAL TOOLS.

- THE COURSE EMPHASIZED BRIDGING THEORETICAL CONCEPTS WITH REAL-WORLD PROBLEM SOLVING, ENCOURAGING PARTICIPANTS TO APPLY ANALYTICAL THINKING IN DOMAINS SUCH AS HEALTHCARE, BUSINESS, AND RESEARCH.

- THIS PROJECT, “DIABETES PREDICTION USING MACHINE LEARNING,” WAS DEVELOPED AS PART OF THE COURSE TO DEMONSTRATE PRACTICAL IMPLEMENTATION OF LEARNED TECHNIQUES

Introduction

- Problem Statement:

Diabetes is a chronic disease affecting millions worldwide. Early detection is crucial to prevent complications and manage the condition effectively.

- Why We Chose This Topic:

Diabetes is a growing global health concern, and we wanted to explore how data-driven approaches can assist in early detection and intervention using machine learning.

- Objective:

Develop a predictive model to determine the likelihood of a patient having diabetes based on medical diagnostic measurements.

- Dataset:

We used the Pima Indians Diabetes Dataset, a widely used benchmark dataset for binary classification problems in healthcare analytics.

Tech Stack

Frontend:

- **ReactJS, TypeScript, Tailwind CSS**
- Responsive design for user data input

Backend:

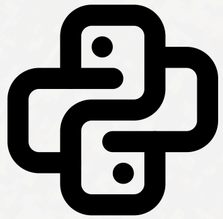
- **FastAPI**
- Scikit-learn (Model Integration)

Machine Learning Libraries:

- Pandas, NumPy, Matplotlib, Seaborn
- Algorithms: Logistic Regression, Decision Tree, Random Forest, SVM

Visualization:

- Power BI (for interactive dashboards and analytics)



Data Overview

- Total Records: 768
- Features: 8
- Target Variable: 0 = Non diabetic,1 = Diabetic

Features	Description
Pregnancies	number of times pregnant
Glucose	plasma glucose concentration
Blood pressure	diastolic blood pressure
Skin thickness	triceps skinfold thickness
Insulin	serum insulin level
BMI	body mass index
Diabetes Pedigree Function	diabetes risk based on family history
Age	In years

Data Preprocessing

- Handling Missing Values:

Certain features had zero values, which are not physiologically plausible. These were treated as missing and imputed with the median of respective columns.

- Features Imputed:
- Glucose
- Blood Pressure
- Skin Thickness
- Insulin
- BMI

- Standardization:

Applied Standard Scaler to normalize feature values, ensuring each has a mean of 0 and standard deviation of 1. This step is crucial for algorithms sensitive to feature scaling.

- Data Splitting:

Divided the dataset into training and testing sets using an 80/20 split:

- Training Set: 614 records
- Testing Set: 154 records

Data Preprocessing

Features	N	Min	Max	Mean	S.D
Pregnancies	768	0	17	3.85	3.370
Glucose	768	44	199	121.66	30.438
Blood pressure	768	24	122	72.39	12.097
Skin thickness	768	7	99	29.11	8.791
Insulin	768	14	846	140.67	86.383
BMI	768	18	67	32.46	6.875
Diabetes Pedigree Function	768	0	2	0.47	0.331
Age	768	21	81	33.24	11.760
Diabetes Prediction	768	0	1	0.27	0.447

Insights

- The dataset is imbalanced in terms of diabetes prediction (only 27%positive).
- Health indicators like glucose, BMI, and insulin show widevariability, suggesting a mix of healthy and high-risk individuals.
- Variables such as insulin and skin thickness may have skeweddistributions due to high maximum values.
- Mean BMI being in the obese range aligns with the relatively highdiabetes rate.

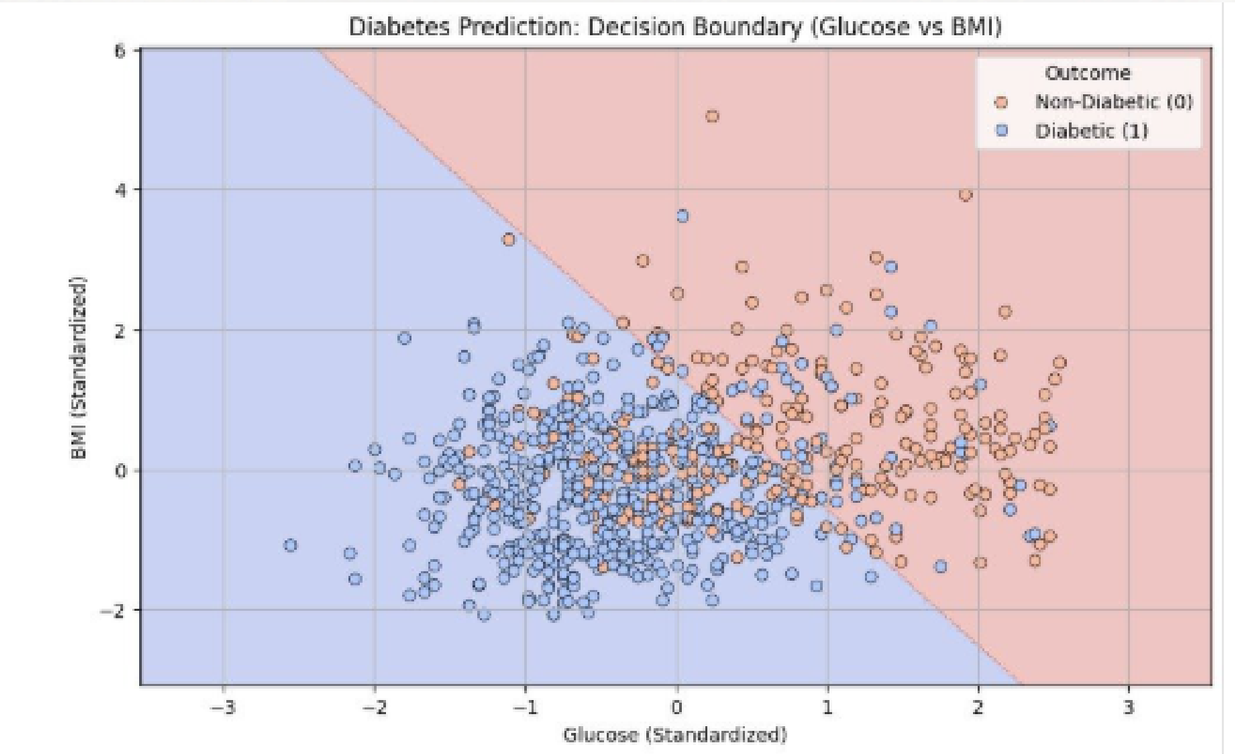
Model Selection

Why Random Forest?

- Random Forest is an ensemble algorithm that combines multiple decision trees for higher accuracy.
- It reduces overfitting and improves generalization across diverse medical data.
- Ideal for binary classification tasks like diabetes prediction (0 = Non-Diabetic, 1 = Diabetic).
- Provides feature importance, helping identify key health factors influencing diabetes.
- Offers better accuracy, robustness, and non-linear handling than Logistic Regression.
- Selected as the final model for its reliable and stable performance

Training the model

- The Random Forest model learns patterns by building multiple decision trees using different subsets of the data and features.
- Each tree makes an independent prediction, and the final output is decided by majority voting among all trees



Correlation Analysis

Pregnancies & Age: 0.544 – Older women tend to have had morepregnancies.

- Skin Thickness & BMI: 0.543 – Logical, as higher BMI often corresponds tomore fat under the skin.
- Glucose & Insulin: 0.419– Elevated glucose often corresponds with elevatedinsulin due to insulin resistance.
- BMI & Blood Pressure: 0.281 – Obesity increases blood pressure.

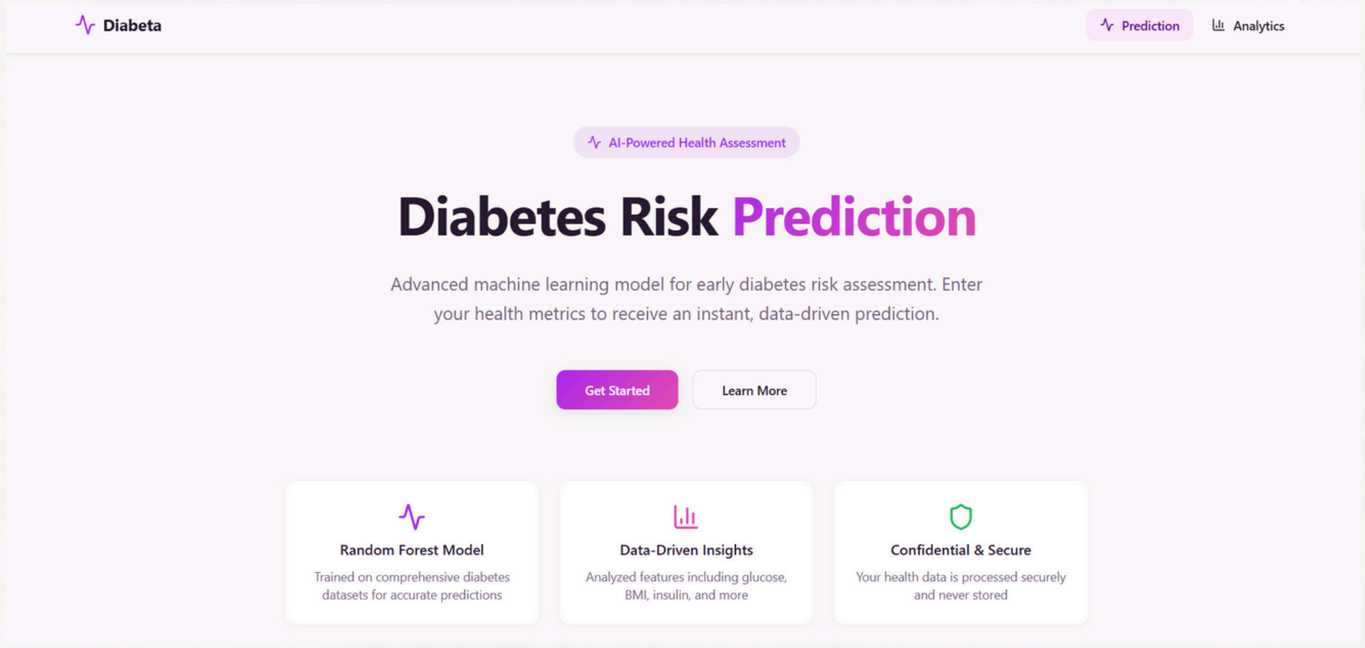
Weak or Non-significant Correlations:

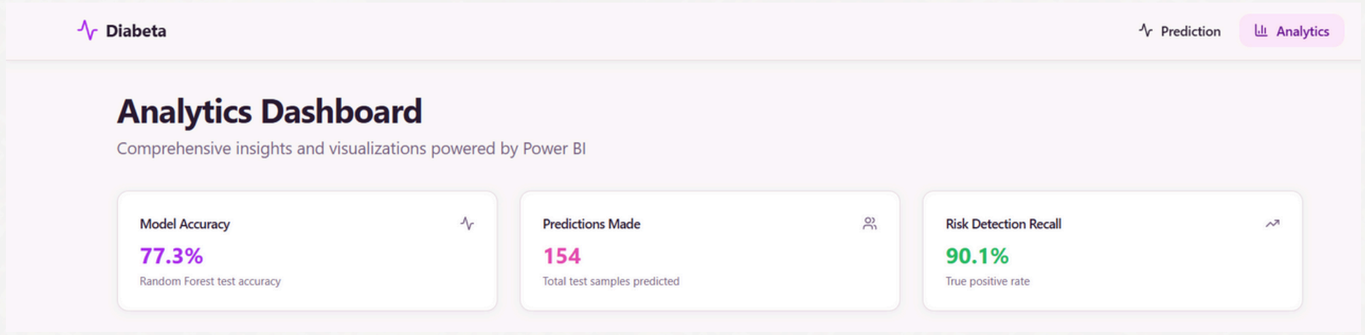
- Insulin & Blood Pressure: 0.045 – Weak and not significant.
- Pregnancies & Insulin: 0.025 – Very weak.
- Age & BMI: 0.026 – Virtually no relationship.
- Diabetes Pedigree Function & Age: 0.034 – Insignificant.

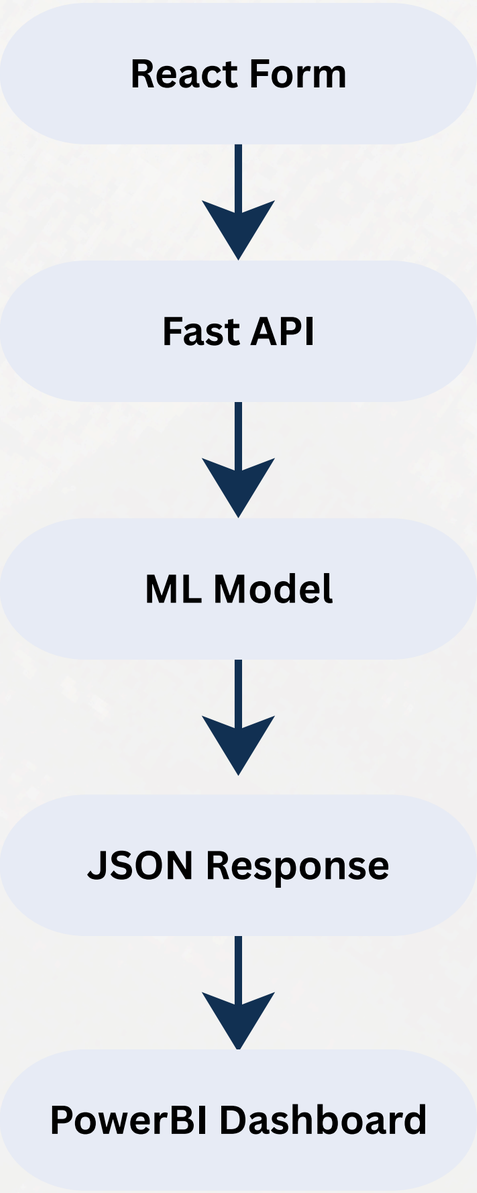
Correlation Analysis Conclusion

- Glucose is the strongest individual predictor of diabetes in this dataset.
- BMI, age, insulin, and skin thickness also play moderate roles.
- All predictors except insulin and pedigree function show positive andstatistically significant correlations with diabetes.
- The predictors are not highly collinear, which is good for regressionmodelling.

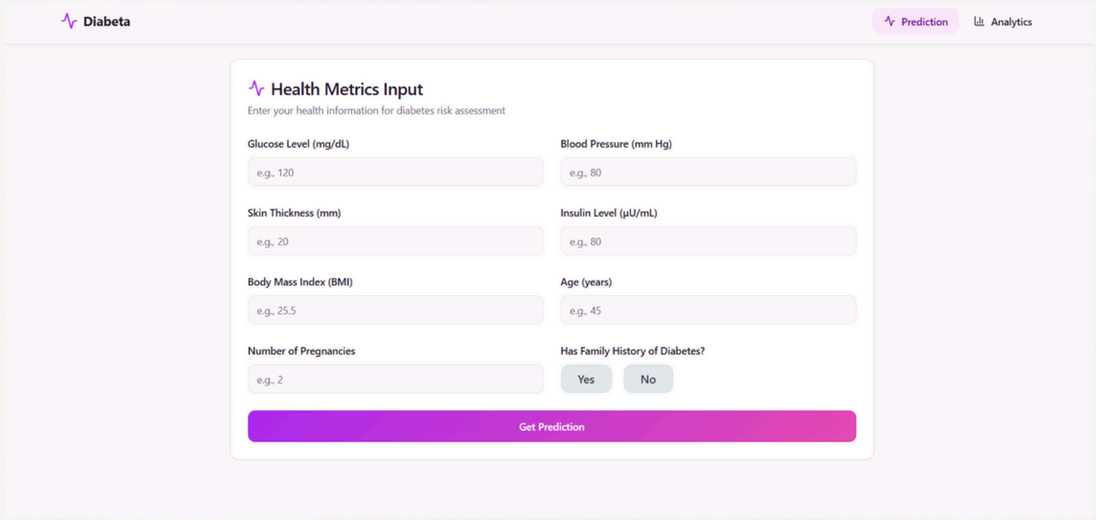
Frontend Implementation

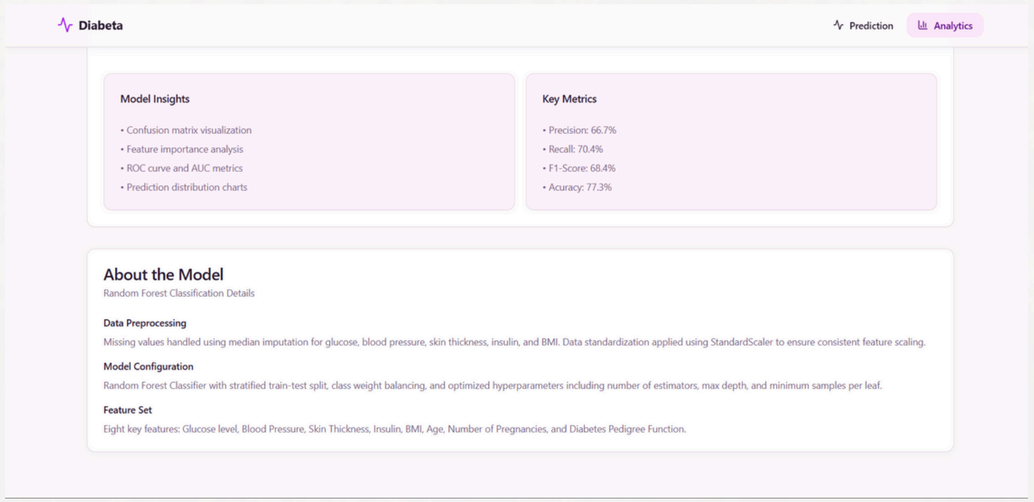






Backend Implementation





Backend powered by FastAPI for lightweight, high-performance REST APIs

Model Evaluation

1.Basic Metrics

- TP = 38
- TN = 81
- FP = 19
- FN = 16

Total samples = 88 + 12 + 28 + 26 = 154

2. Accuracy

Accuracy = (TP + TN) / Total = (38 + 81) / 154 = 77.3%

3. Precision (for class 1)

Precision = TP / (TP + FP) = 38/ (38 + 19) = 66.7%

4. Recall (Sensitivity, for class 1)

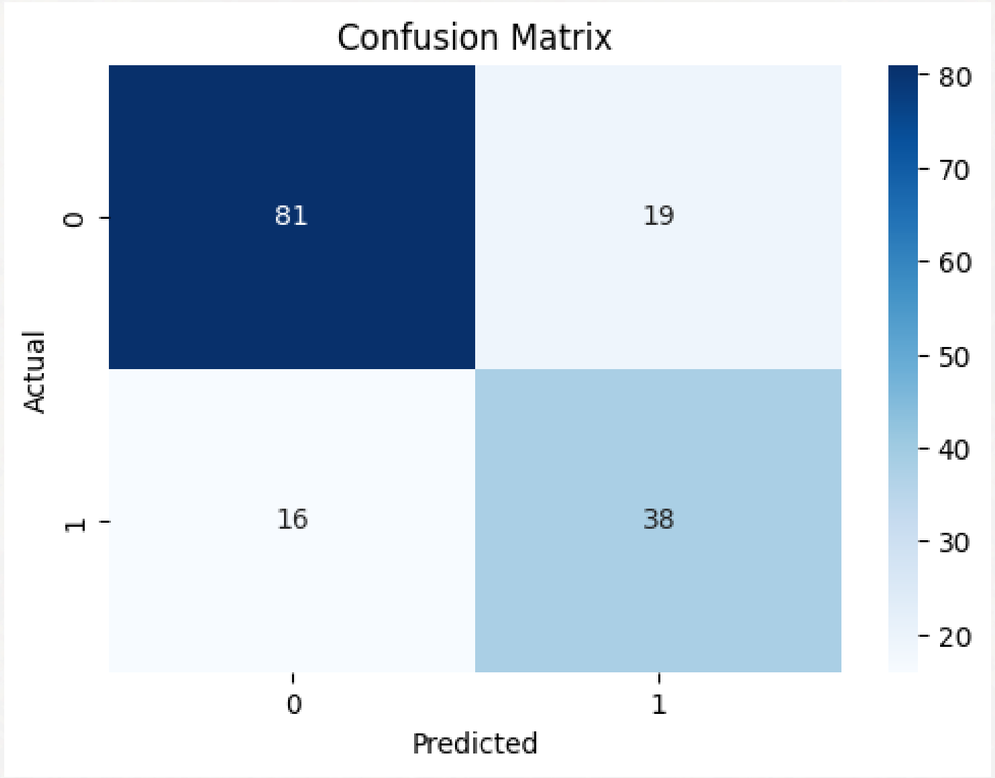
Recall = TP / (TP + FN) = 38/ (38 + 16) = 70.4%

5. F1 Score (for class 1)

F1 = (2 * Precision * Recall) / (Precision + Recall) = 68.4%

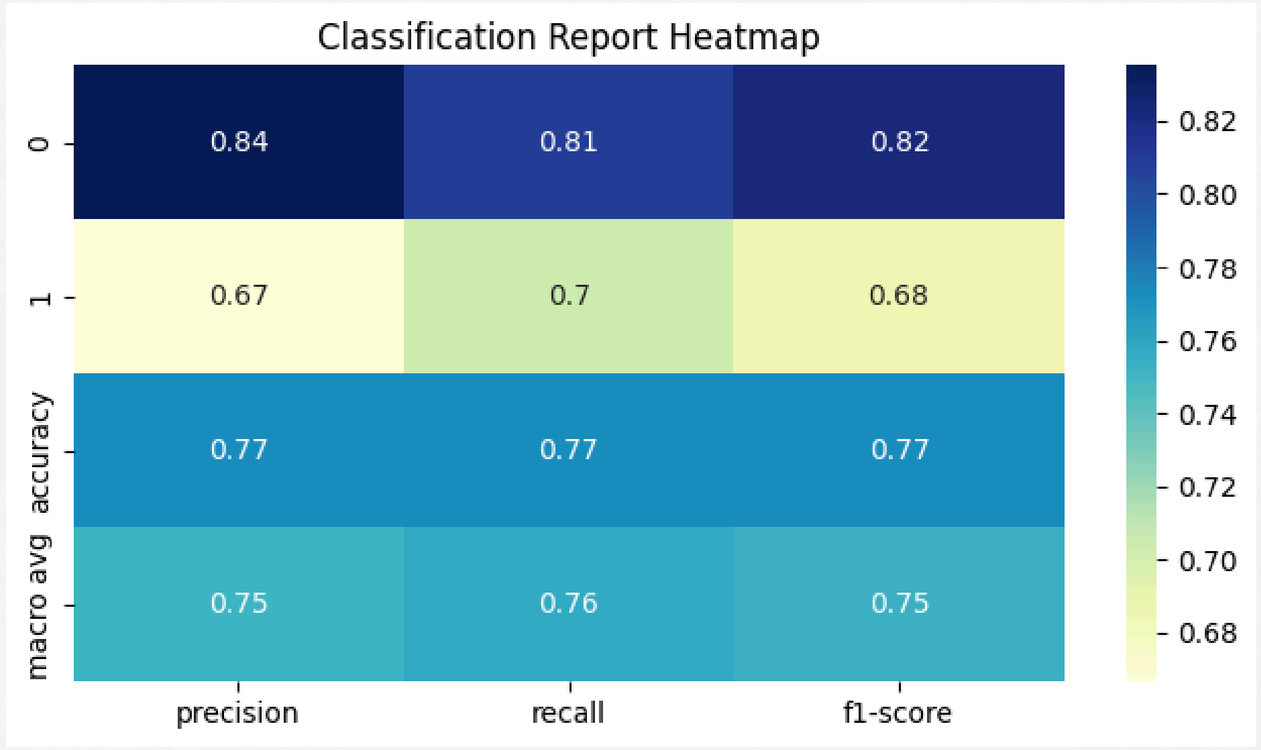
Model Evaluation

Confusion Matrix



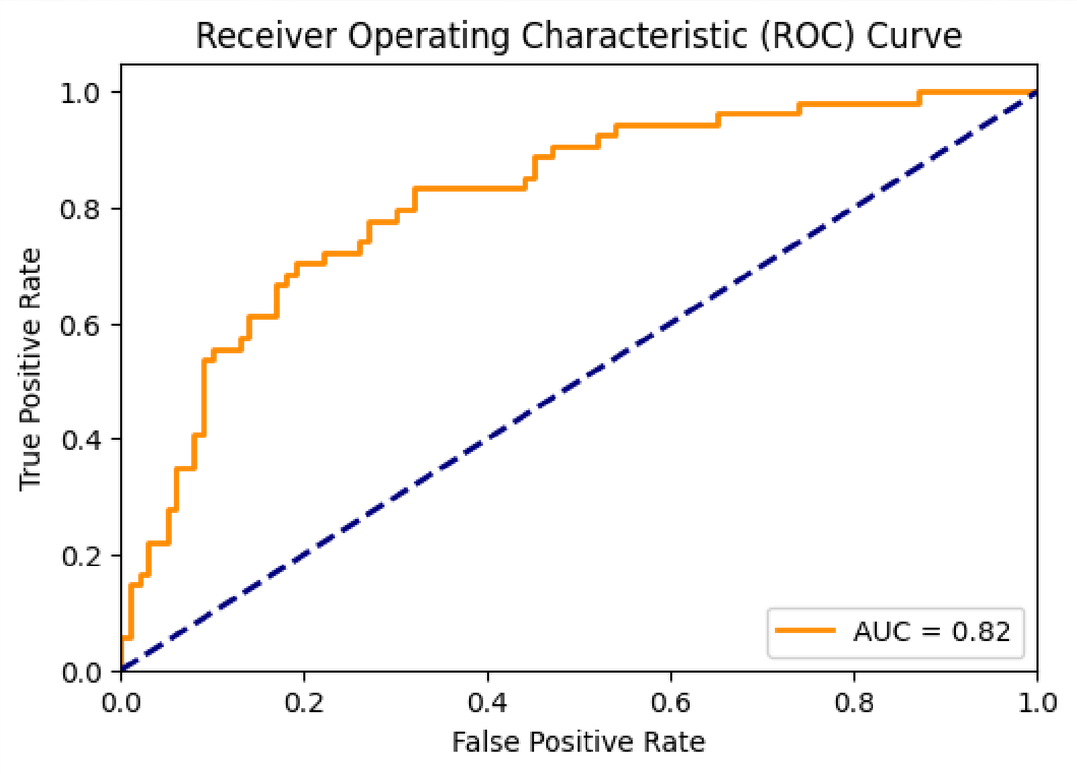
Class 0: True Negatives(TN) and False Positives(FP)
Class 1: False Negatives (FN) and True Positives (TP)

Heatmap



Model Evaluation

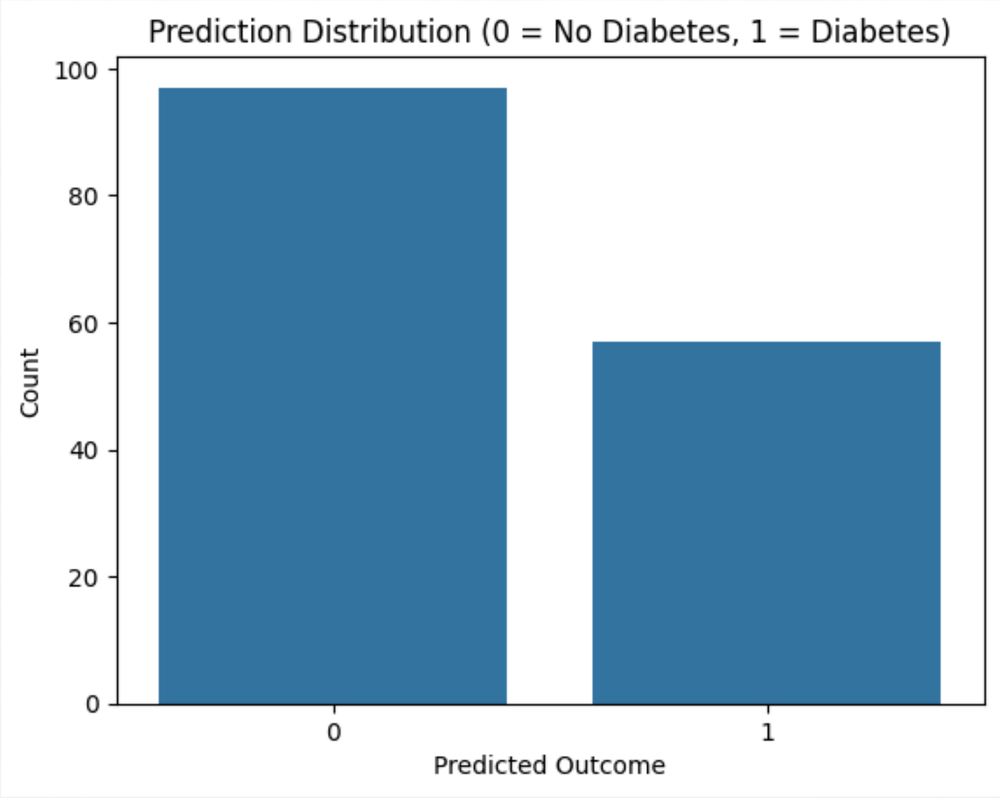
ROC Curve



AUC (Area Under the Curve) = 0.82

AUC values range from 0.5 (random) to 1.0 (perfect).
An AUC of 0.81 is considered good, indicating the model has a strongability to distinguish between the two classes

Prediction Distribution



Conclusion:

Most Predictive Features

- Glucose, BMI, and Age were the most influential factors in predicting diabetes.
- Higher glucose levels significantly increased the likelihood of being diabetic.
- The model also showed meaningful contributions from Insulin and Blood Pressure in certain cases.

Model Characteristics

- It effectively captured non-linear relationships between medical features and diabetes outcomes.
- Achieved an AUC of 0.81, indicating strong separability between diabetic and non-diabetic patients.
- The model offered high interpretability through feature importance rankings and consistent prediction performance.

DIABETES PREDICTOR

Average of Glucose
121.66

Average of BMI
32.46

Average of Insulin
140.67

Average of BloodPressure
72.39

Average of SkinThickness
29.11

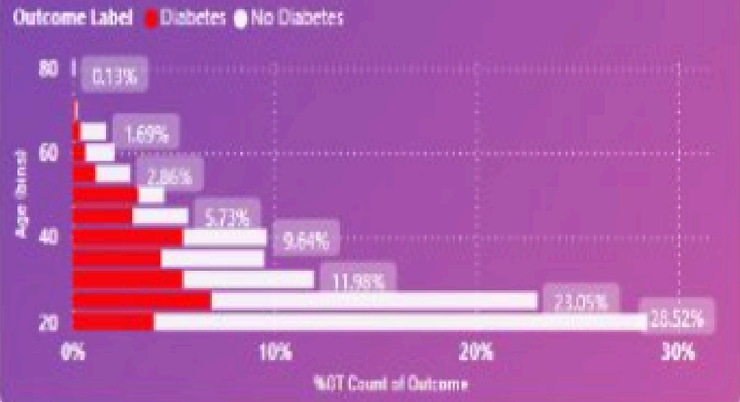
Average of Pregnancies
3.85

Distribution of Actual Diabetes Cases

Diabetes cases: 35% yes, 65% no.

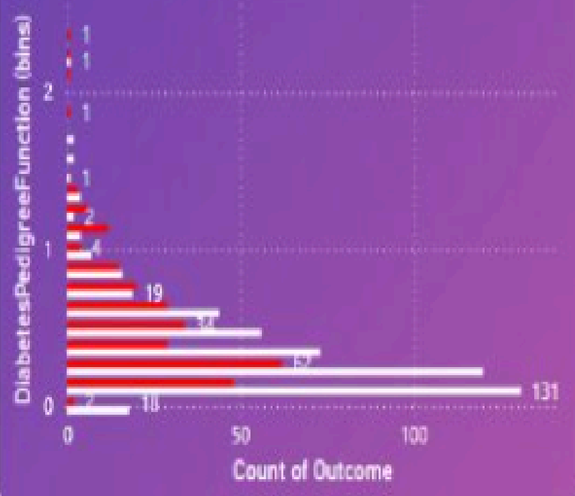
Age-wise Diabetes Distribution (% of Total)

Young adults (20–30) have the highest share of both diabetic and non-diabetic cases.



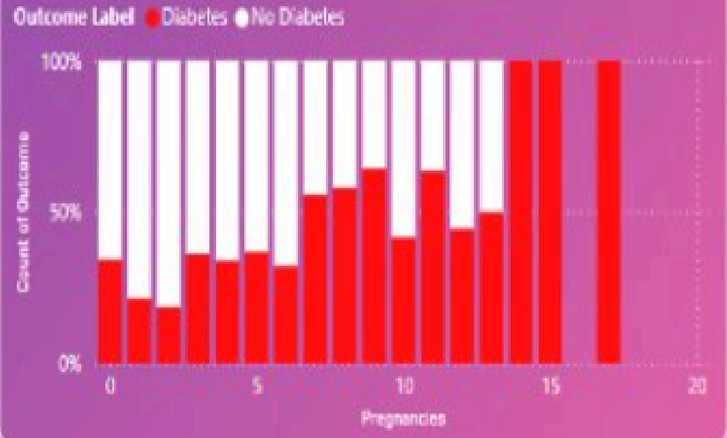
Impact of Heredity Risk on Diabetes Outcome

Higher DPF is linked to more diabetes cases.



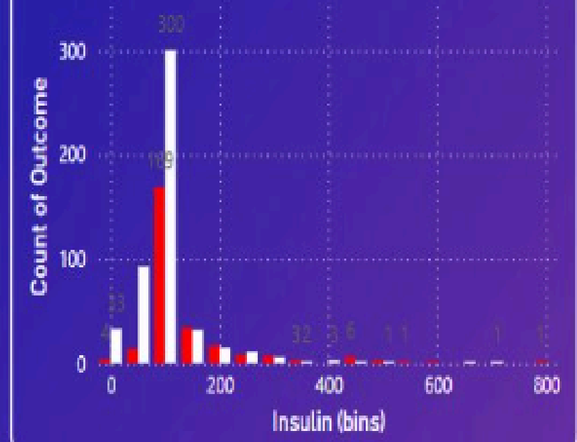
Pregnancy Count vs Diabetes Likelihood

More pregnancies raise diabetes risk.



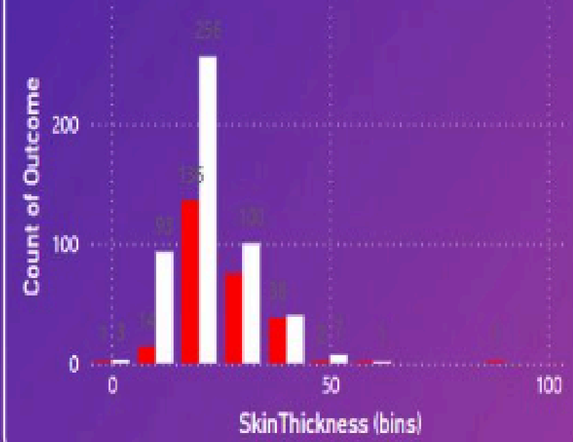
Influence of Insulin on Diabetes Diagnosis

Diabetes is common at insulin levels below 200.



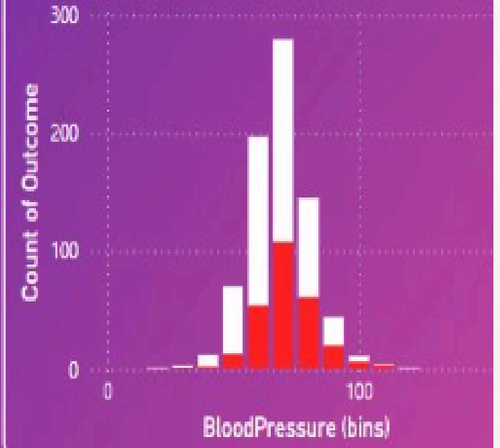
Skin Thickness Impact on Diabetes Diagnosis

Most diabetes cases occur with skin thickness between 20–40.



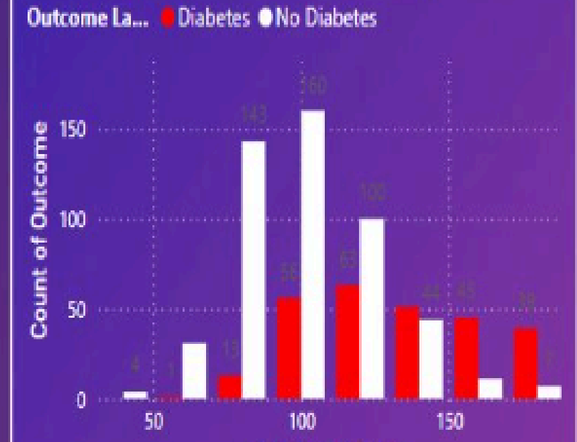
Diabetes Cases by BP Category

Most diabetes cases are in the 60–80 BP range.



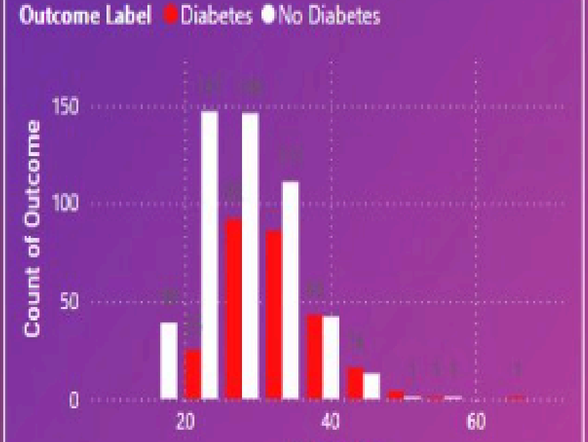
Glucose Level vs Diabetes Diagnosis

Higher glucose links to more diabetes cases.



BMI Influence on Diabetes Outcome

Diabetes rises with BMI, especially above 30.



Prediction Accuracy by Actual Diagnosis

Model predicted 595 of 768 cases accurately.