**Mini Project Report**

**Title:** Prediction of Diabetes Using Machine Learning

**Group Members:**

1. Chetan G. Kotangale - Roll No: 211009
2. Atharva V. Maslekar - Roll No: 211037
3. Atharva R. Joshi - Roll No: 211049
4. Manas G. Kulkarni - Roll No: 211058

**Department:** Electronics and Telecommunication

**Institution:** Vishwakarma Institute of Information Technology, Pune

**Submitted To:** Prof. Vilas Ghonge

**Date:** [29/04/2025]

**1. Abstract**

Diabetes is a major global health concern and particularly prevalent in India. Early diagnosis can significantly improve patient outcomes and reduce the economic burden of treatment. This project focuses on building a machine learning-based predictive model using patient medical data to forecast the probability of diabetes. Using supervised learning algorithms, we analysed key health indicators and evaluated multiple models for optimal performance. The insights gained aim to support early detection, prevention, and personalized healthcare.

**2. Introduction**

Diabetes is a chronic disease that shows clinical symptoms and can be identified with early screening. Machine learning models can aid significantly in predicting diabetes by analysing patterns in patient health records. Our dataset comprises over 765 patients with attributes like age, BMI, glucose levels, insulin levels, blood pressure, and more. By utilizing machine learning, this study attempts not only to predict diabetes onset but also to understand the contributing factors better.

**3. Problem Statement**

To develop a machine learning model that accurately predicts the likelihood of a patient being diabetic, based on medical attributes. Early prediction can help prevent severe complications and promote timely interventions.

**4. Objectives**

* Predict the probability of being diabetic using ML models.
* Identify key features influencing diabetes.
* Utilize health indicators for early diagnosis.
* Offer actionable insights for preventive healthcare.

**5. Data Sources**

* Kaggle Dataset: [Predict Diabetes from Medical Records](https://www.kaggle.com/code/paultimothymooney/predict-diabetes-from-medical-records)

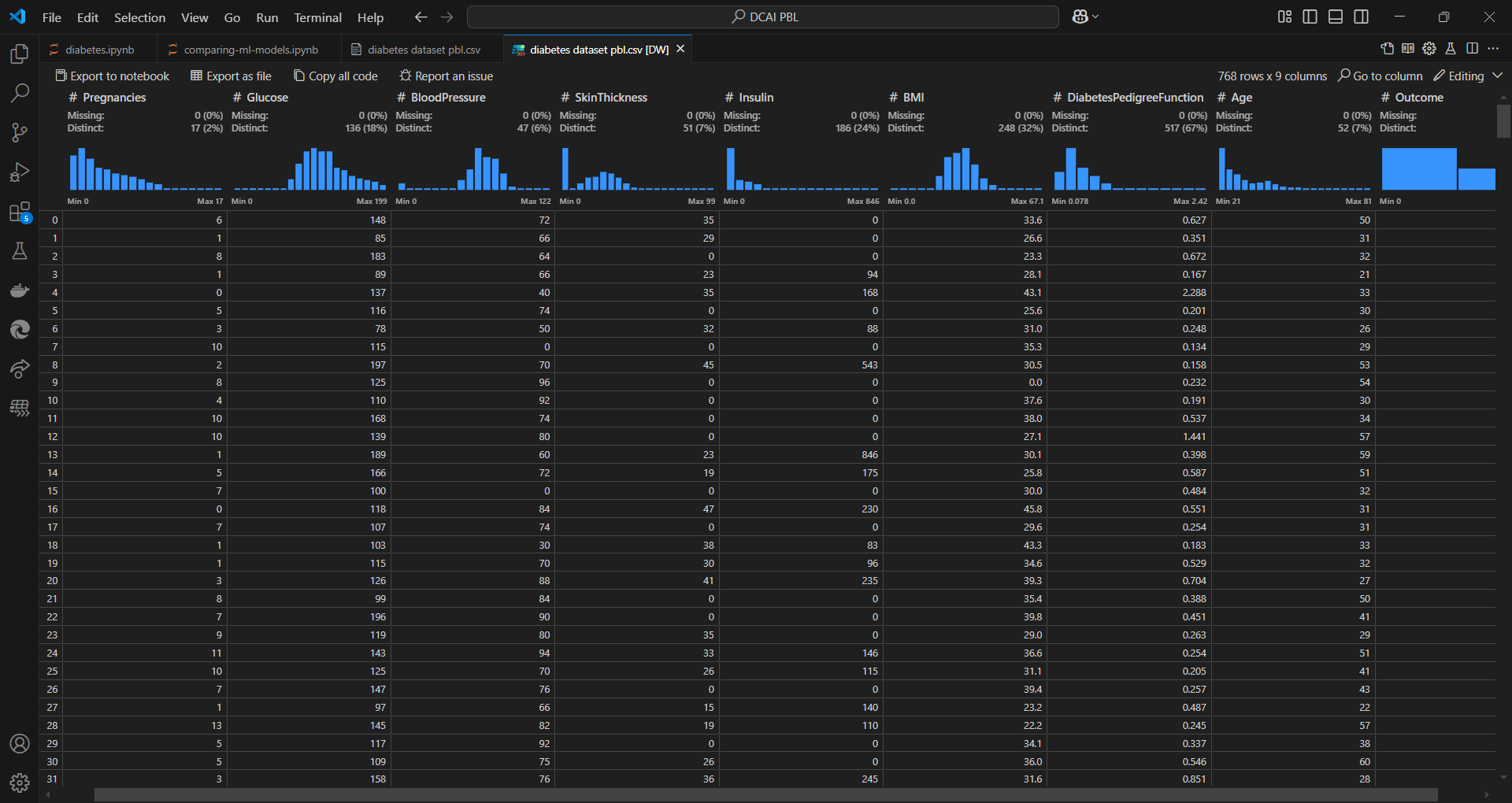


Figure 1 Diabetes Dataset

**6. Methodology**

**6.1 Data Preprocessing**

* Handled missing values with median imputation.
* Outliers removed using the Interquartile Range (IQR) method.
* Normalized numerical data for model stability.

**6.2 Exploratory Data Analysis**

* Visualized distributions using histograms and boxplots.
* Analysed correlations through heatmaps.
* Key factors identified: Glucose, BMI, Insulin levels, Age, and Diabetes Pedigree Function.

Figure 2 Prediction Report on patients

**6.3 Feature Engineering**

* Analysed combinations like BMI and Blood Pressure.
* Created derived features such as risk scores.
* Segmented age groups to study risk variation.

**6.4 Modeling**

* Logistic Regression
* Random Forest Classifier
* Support Vector Machine (SVM) *(Model Used for Training)*
* XGBoost Classifier
* LightGBM Classifier
* CatBoost Classifier

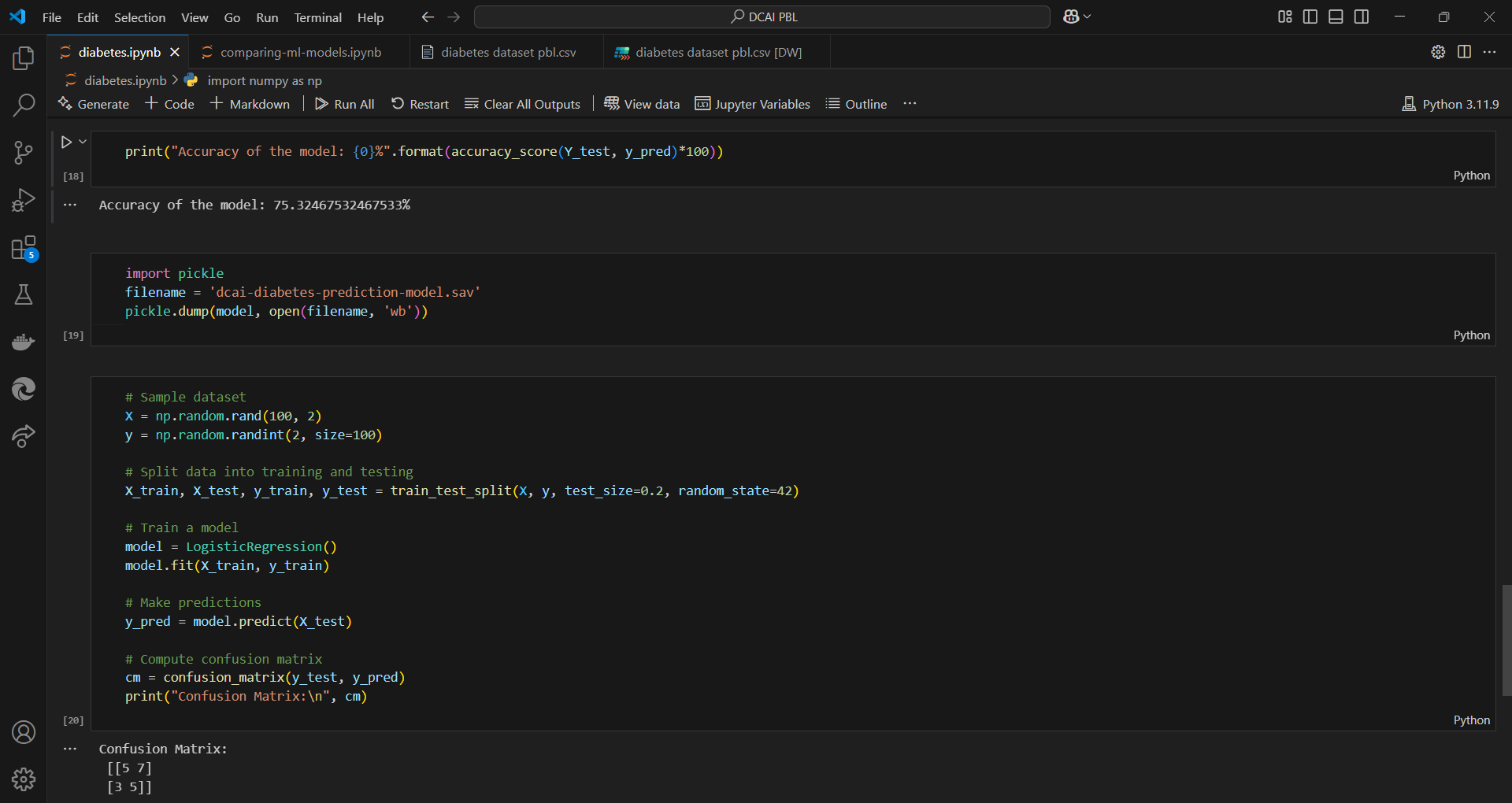


Figure 3 Code for Confusion matrix and classification report of SVM

**6.5 Evaluation Metrics**

* Accuracy, Precision, Recall, F1-Score, ROC-AUC
* Confusion Matrix for classification evaluation

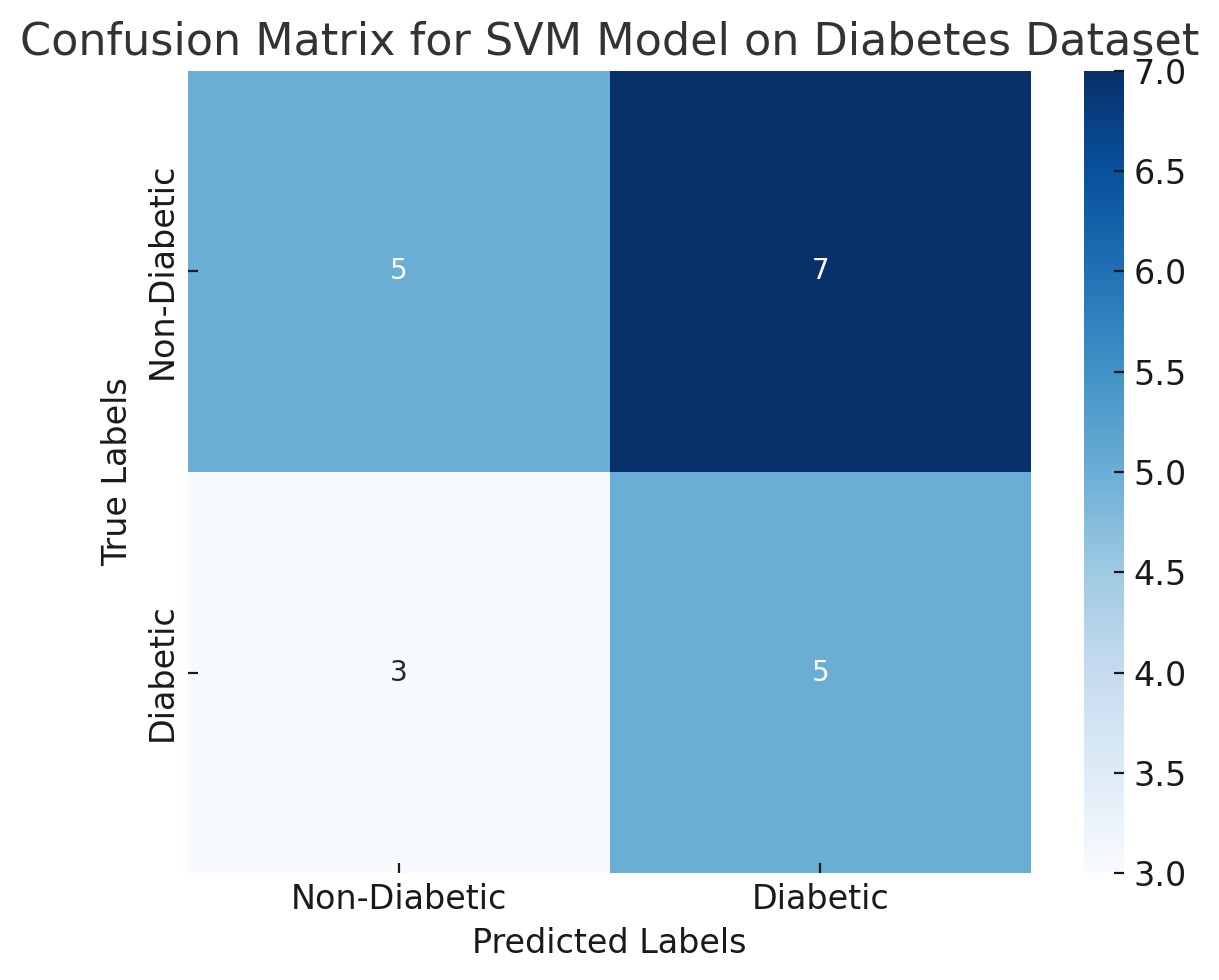


Figure 4 Confusion Matrix for SVM Classification Model

* Cross-validation with 5 folds (CV=5)

**7. Results & Discussion**

**Model Comparison**

|  |  |  |
| --- | --- | --- |
| **Model** | **Accuracy (Test Set)** | **Cross-Val Accuracy (CV=5)** |
| CatBoost | 75.32% | 77.09% |
| Logistic Regression | 75.32% | 76.96% |
| Random Forest | 74.68% | 76.44% |
| SVM *(Trained Model)* | 73.38% | 75.91% |
| LightGBM | 70.78% | 74.74% |
| XGBoost | 72.08% | 74.10% |

* **Best Performing Model:** CatBoost (Highest Cross-Val Accuracy: 77.09%)
* **Model Used:** SVM (Accuracy: 73.38%)

**Visualizations:**

* Average insulin and glucose levels by age

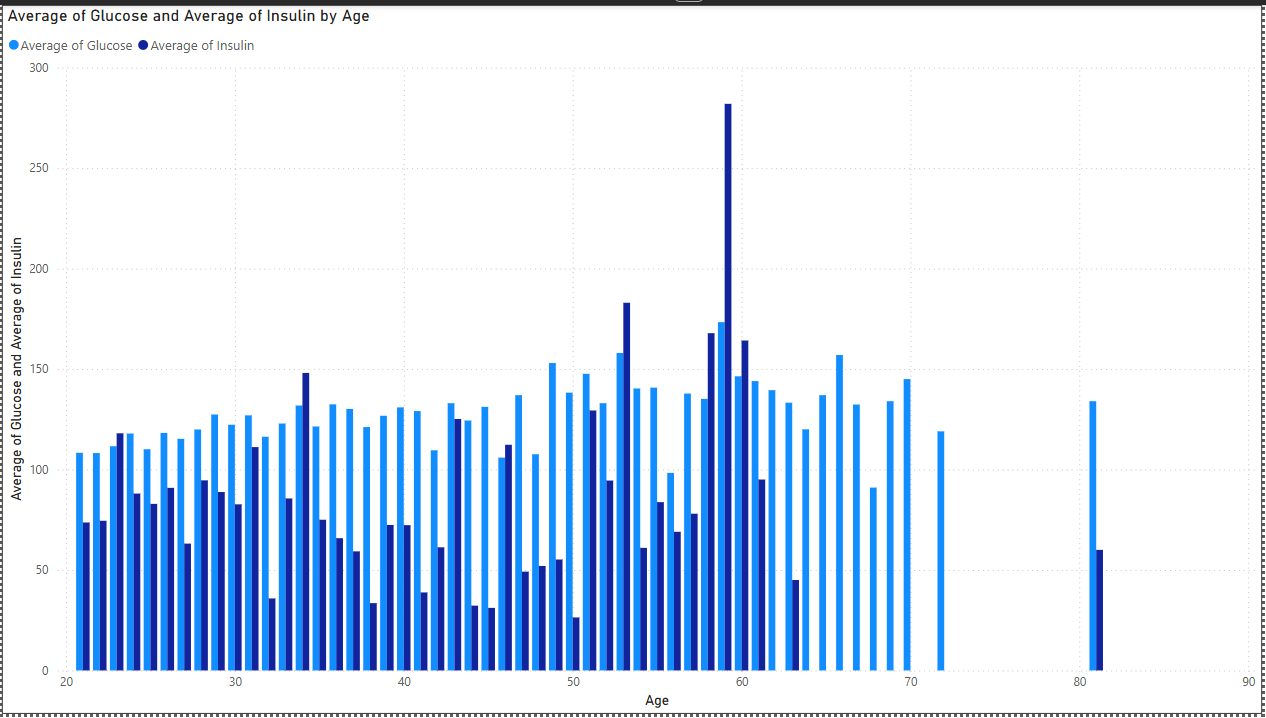


Figure 5 Average of glucose and insulin with age

The graph represents the trends of glucose and insulin levels according to the age of the patients. The conclusion made that patients generally 40 to 60 have a higher level of insulin and glucose. This made conclude that they are more prone to diabetes and higher likelihood of getting one.

* Outcome and Variation of Diabetic Pedigree function based on age of the person

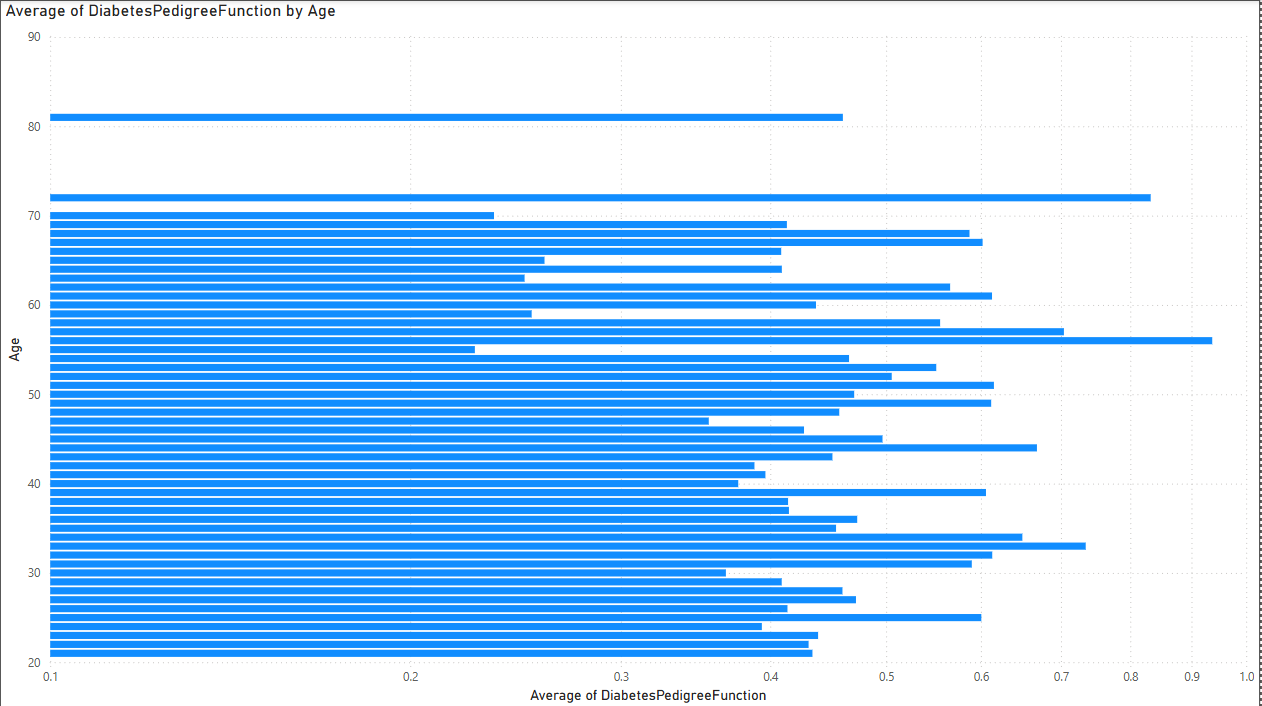


Figure 6 Variation of Diabetic Pedigree Function by age

The graph is the variation of Diabetic Pedigree Function that affects the patient status due to its genetics. This function is the measure of likelihood of the person being diabetic based of his/ her genetics. This is the variation with respect to the age of the person. The conclusion can be made that most susceptible diabetic patients are above 40 years of age.

* Effect of BMI on the Blood Pressure

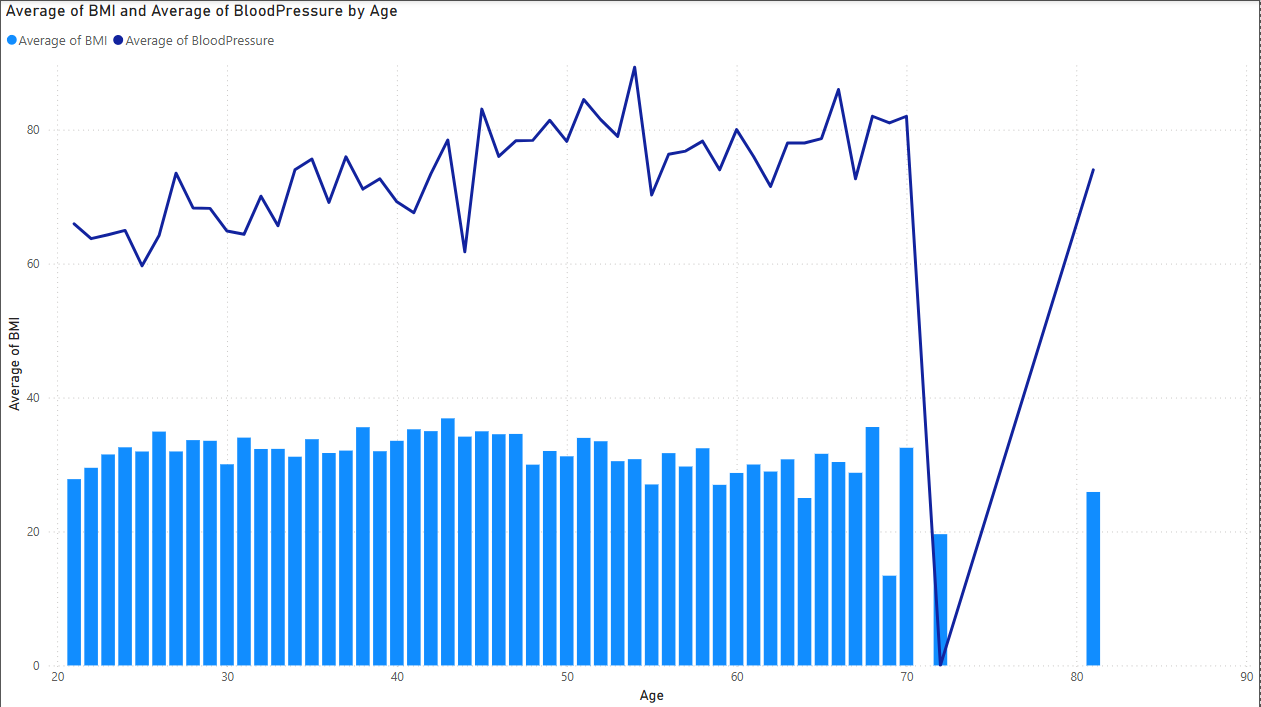


Figure 7 Blood pressure variation with BMI

The above graph represents the effect Body Mass Index (BMI) on the blood pressure of the person. The graph suggests that the more balanced the BMI is the more balanced the blood pressure would be.

* Graph of outcome and glucose by pregnancies

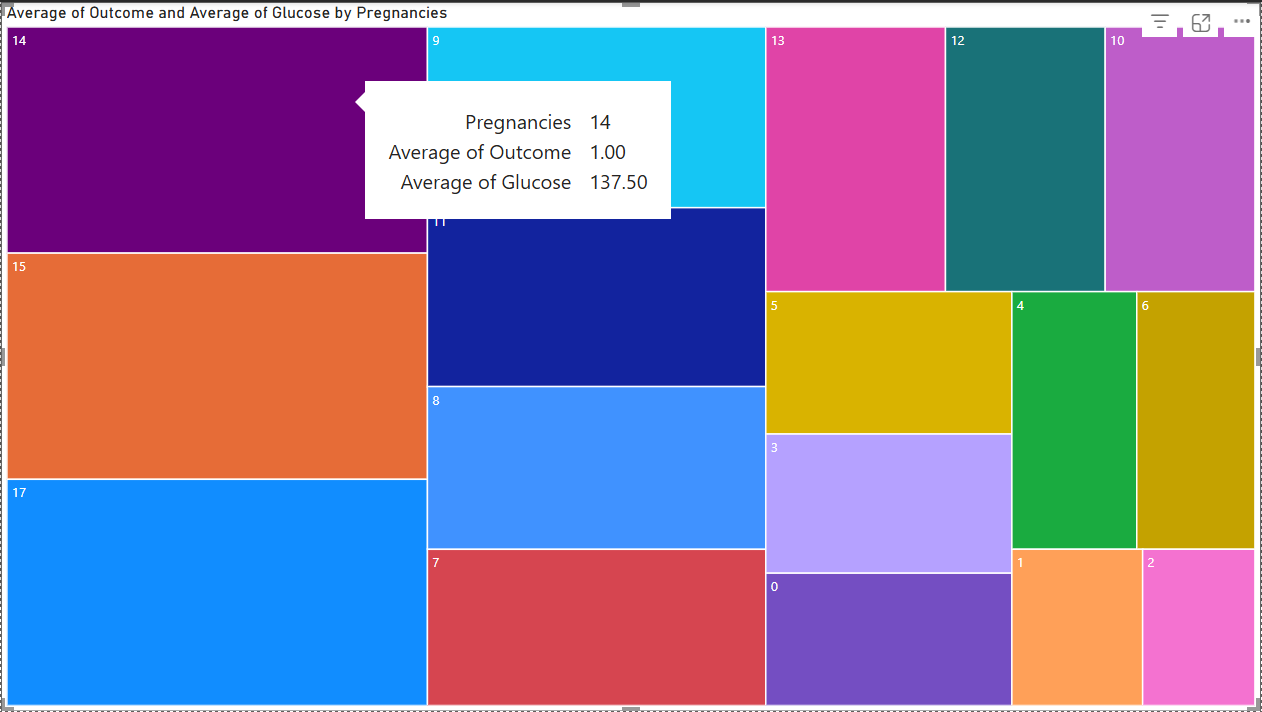


Figure 8 Graph of outcome and glucose by pregnancies

This a tree map view of the effect of number of pregnancies to the individual’s outcome of being a diabetic as well as the glucose levels. The higher number of pregnancies shows higher likelihood of being diabetic.

**8. Conclusion**

This project demonstrates the power of machine learning models in the early prediction of diabetes using health indicators. Analyzing features such as BMI, glucose, and insulin levels reveals critical patterns associated with diabetes risk. Although CatBoost achieved the highest accuracy, SVM was used for model training and provided robust performance. Early identification through such predictive modeling can enable better disease management and preventive care strategies.

**9. Future Scope**

* Integration of the model into mobile fitness and health apps.
* Real-time tracking using smart fitness bands.
* Personalized healthcare recommendations using AI.
* Expansion of the dataset with longitudinal patient tracking.

**10. References**

1. Kaggle Dataset: [Predict Diabetes from Medical Records](https://www.kaggle.com/code/paultimothymooney/predict-diabetes-from-medical-records)
2. Scikit-learn Documentation: <https://scikit-learn.org/stable/documentation.html>
3. CatBoost Documentation: <https://catboost.ai/en/docs/>
4. World Health Organization (WHO) Reports on Diabetes: <https://www.who.int/news-room/fact-sheets/detail/diabetes>
5. XGBoost Documentation: <https://xgboost.readthedocs.io/en/stable/>

**11. Appendix**

* **Sample Code Snippets:** Data preprocessing, model training, evaluation code.

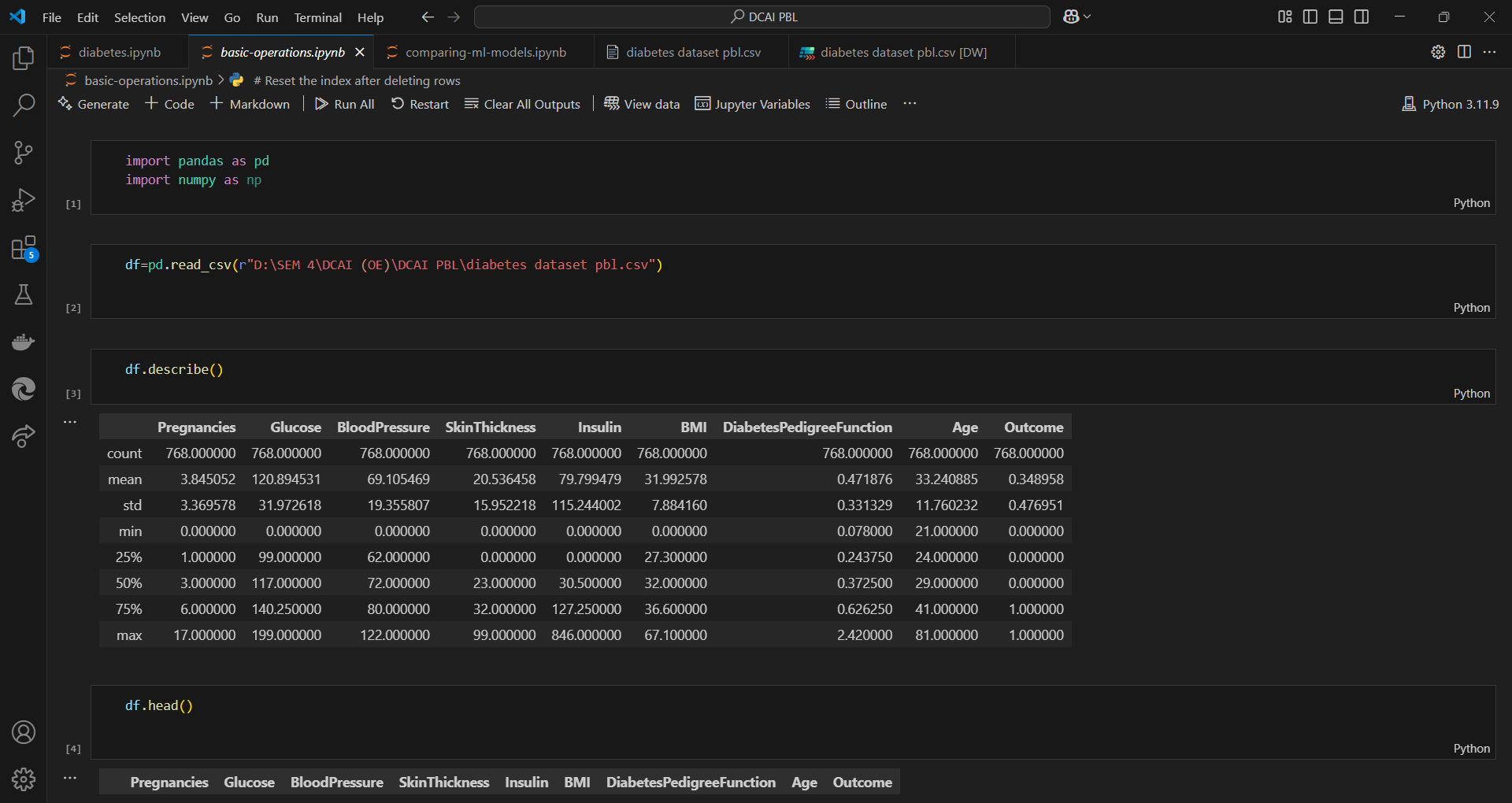


Figure 9 Code for Basic Operations

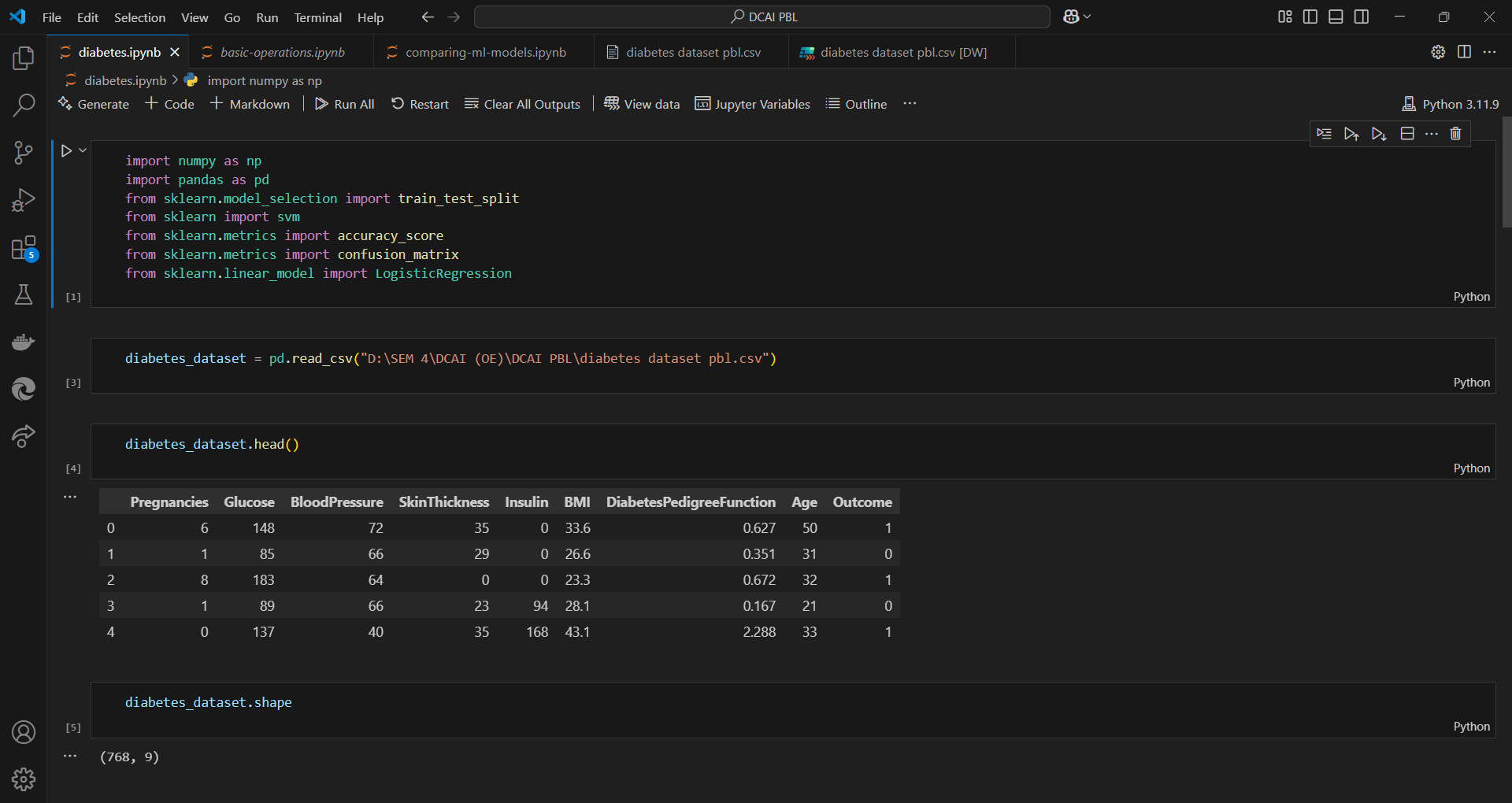


Figure 9 Code for Training Dataset using SVM model

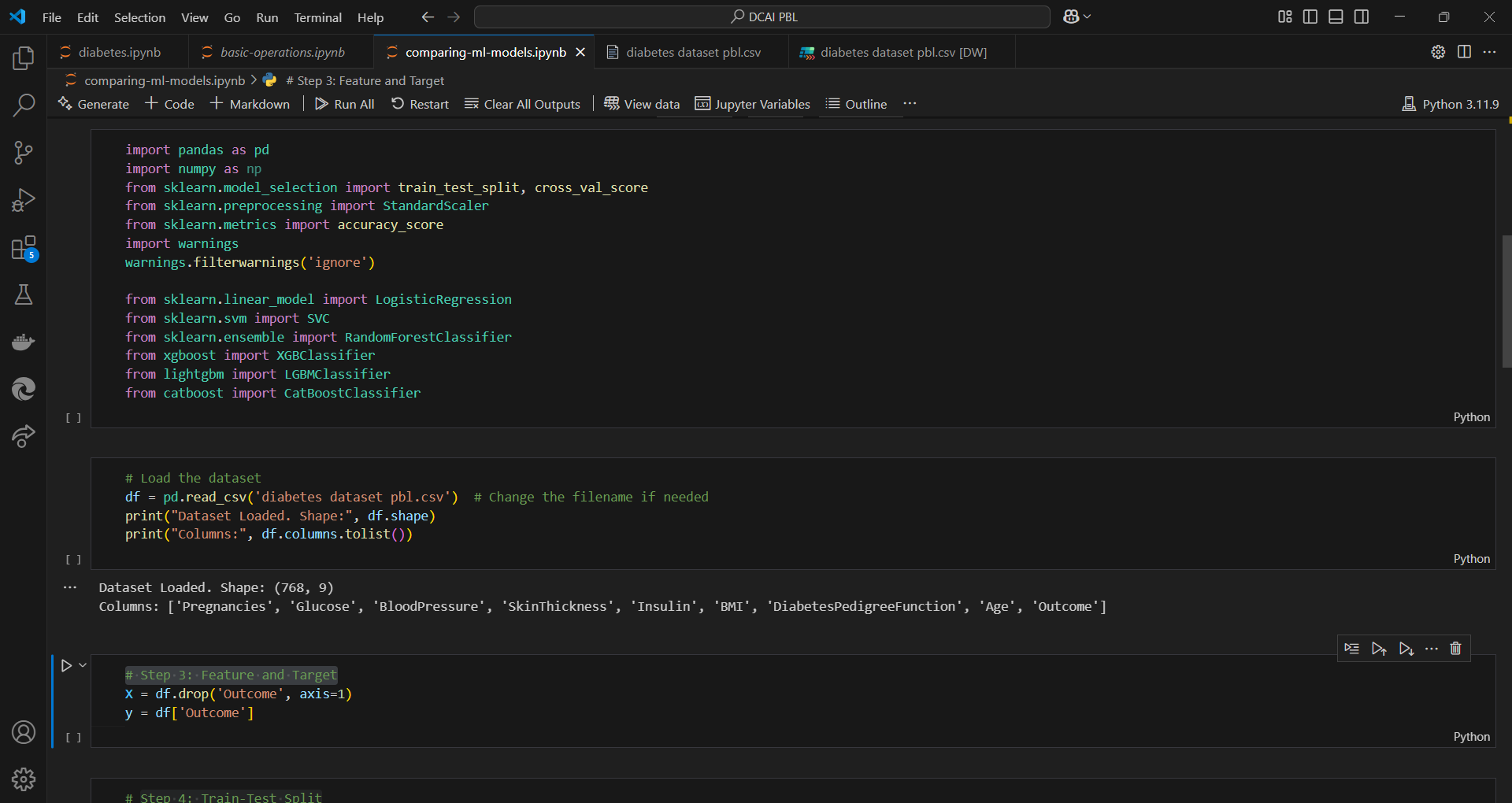


Figure 10 Code for Comparison of Different ML Models

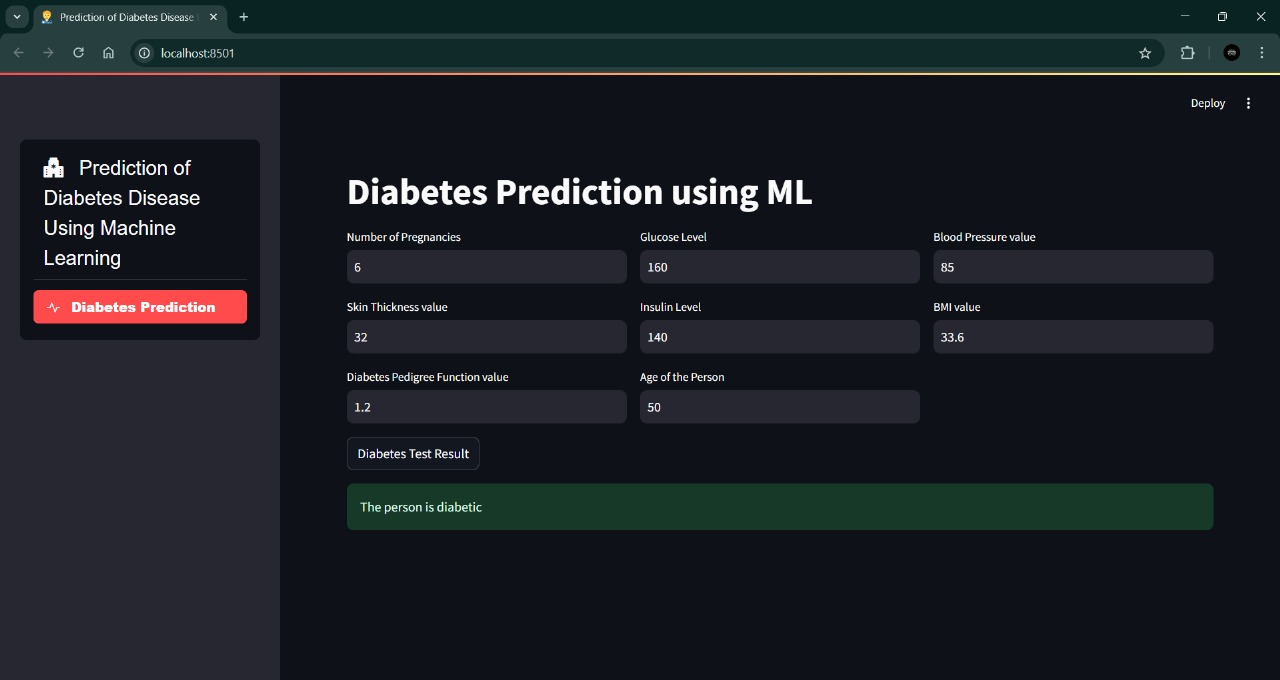


Figure 11 Person with Diabetes

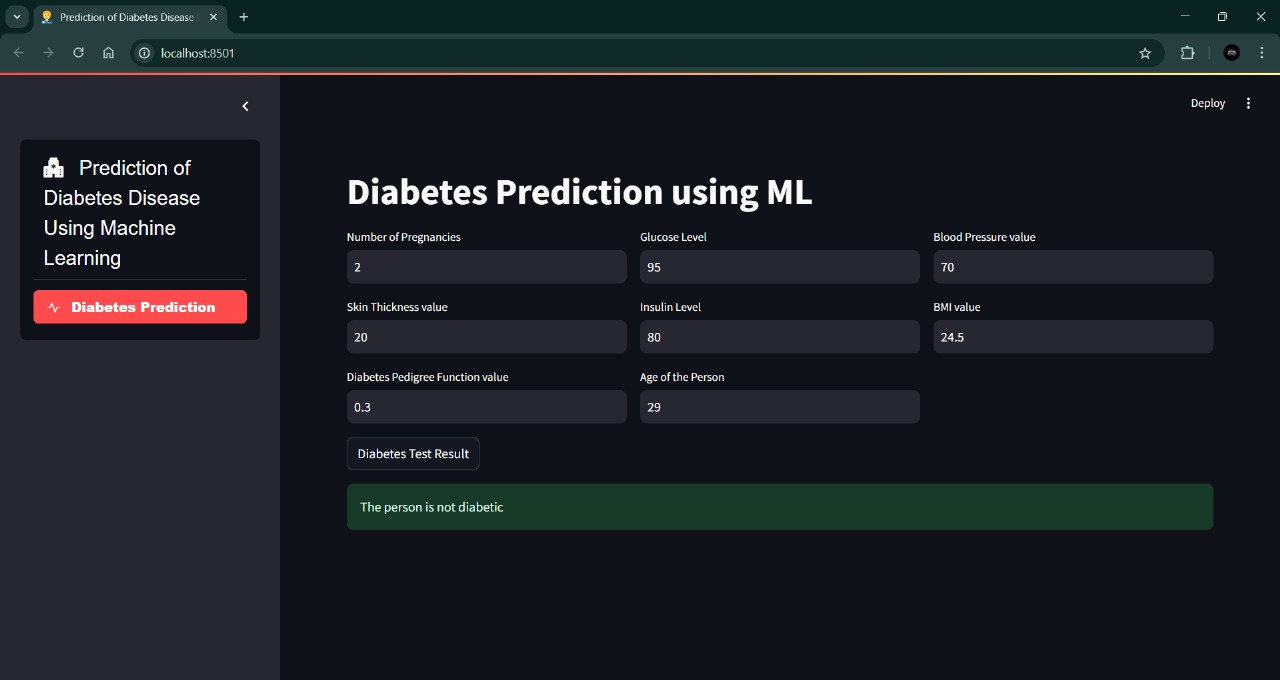


Figure 12 Person without Diabetes

* **Extended Visualizations:** EDA graphs, feature importance charts, cluster plots.
* **Data Dictionary:**
  + Pregnancies: Number of times pregnant
  + Glucose: Plasma glucose concentration
  + Blood Pressure: Diastolic blood pressure (mm Hg)
  + Skin Thickness: Triceps skinfold thickness (mm)
  + Insulin: 2-Hour serum insulin (mu U/ml)
  + BMI: Body Mass Index
  + Diabetes Pedigree Function: Likelihood of diabetes based on family history
  + Age: Age in years
  + Outcome: Diabetes status (1: diabetic, 0: non-diabetic)