**Fundamentals of Data Mining**

**Final Report**

A blue and orange logo

Description automatically generated

**Group Members:**

|  |  |
| --- | --- |
| **Student Register Number** | **Name** |
| IT21307294 | Gamage D G A S |
| IT21255106 | Nelligahawaththa A D T B |
| IT21206832 | Pabasara S D |

**Video Reference Link:**

Content

[1. Background 3](#_Toc148637044)

[1.1 Problem 3](#_Toc148637045)

[1.2 Solution 3](#_Toc148637046)

[2. Identify the problem with business goals. 4](#_Toc148637047)

[3. Description of the Dataset 4](#_Toc148637048)

[4. Data Identification 5](#_Toc148637049)

[5. Data Visualization Before Preprocessing 7](#_Toc148637050)

[6. Data Preprocessing 9](#_Toc148637051)

[7. Partitioning 15](#_Toc148637052)

[8. Proposed Data Mining Solutions 16](#_Toc148637053)

[9. Model Selection 27](#_Toc148637054)

[10. Model Optimization 28](#_Toc148637055)

[11. Data Visualization 29](#_Toc148637056)

[12. Prediction 31](#_Toc148637057)

[13. Model Deployment 32](#_Toc148637058)

[14. Deploying Implementation 38](#_Toc148637059)

[15. Test Cases 39](#_Toc148637060)

[16. The Project Structure 43](#_Toc148637061)

[17. Benefits of Proposed Solution 44](#_Toc148637062)

[18. Conclusion 45](#_Toc148637063)

[19. Project Team and Workload 46](#_Toc148637064)

[20. References 47](#_Toc148637065)

**DiabetesPredict: Women's Diabetic Prediction System**

# Background

Diabetes is a widespread chronic disease that presents a significant public health challenge due to its global prevalence, particularly affecting women. Characterized by elevated blood glucose levels, early diagnosis and intervention are crucial for effective management and complication prevention, especially in female diabetic patients. The healthcare industry has increasingly turned to data mining and machine learning techniques to improve disease diagnosis and prediction precision. Our research aims to develop web applications using these techniques, focusing on women's diabetic prediction, to meet the urgent demand for enhanced diabetes care while harnessing vast patient data for academic and healthcare practitioner insights.

## Problem

Diabetes can impact millions of women worldwide, potentially resulting in severe health issues without proper management. Many remain undiagnosed until symptoms or complications arise, leading to delayed treatment and worse outcomes. Furthermore, identifying women at risk in underserved areas can be challenging.

## 1.2 Solution

Our "Women's Diabetic Prediction System" project addresses these concerns by developing a user-friendly web application. Our objective is to equip healthcare providers and individuals with predictive tools to enhance diabetes prevention and management specifically for women.

The proposed platform will analyze publicly available databases with women's demographic details, medical histories, and diabetes-related diagnostics. It will employ data mining and machine learning algorithms to forecast the risk of diabetes in women. Users will access pertinent health information, along with practical self-assessment and prevention recommendations.

Our web application is essential for its accessibility, accuracy, and adherence to privacy standards. It not only predicts diabetes risk in women but also promotes health awareness and informed medical decision-making. Through this initiative, we aspire to advance medical technology and enhance public health outcomes with a particular focus on women's diabetic prediction and care.

**Dataset Link:** [**https://www.kaggle.com/datasets/mathchi/diabetes-data-set**](https://www.kaggle.com/datasets/mathchi/diabetes-data-set)

**Git Hub Link:** [**https://github.com/it21206832/FDM-Project-G29.git**](https://github.com/it21206832/FDM-Project-G29.git)

# Identify the problem with business goals.

The business goals for a women's diabetic prediction system face key challenges. They often lack specificity, which makes measuring success difficult. Biased or inadequate data can lead to inaccurate predictions. Privacy and legal compliance are essential. Ensuring inclusivity, integrating with healthcare providers, defining clear KPIs, and maintaining ethical and sustainable financial models are vital for success in delivering tailored healthcare solutions for women with diabetes.

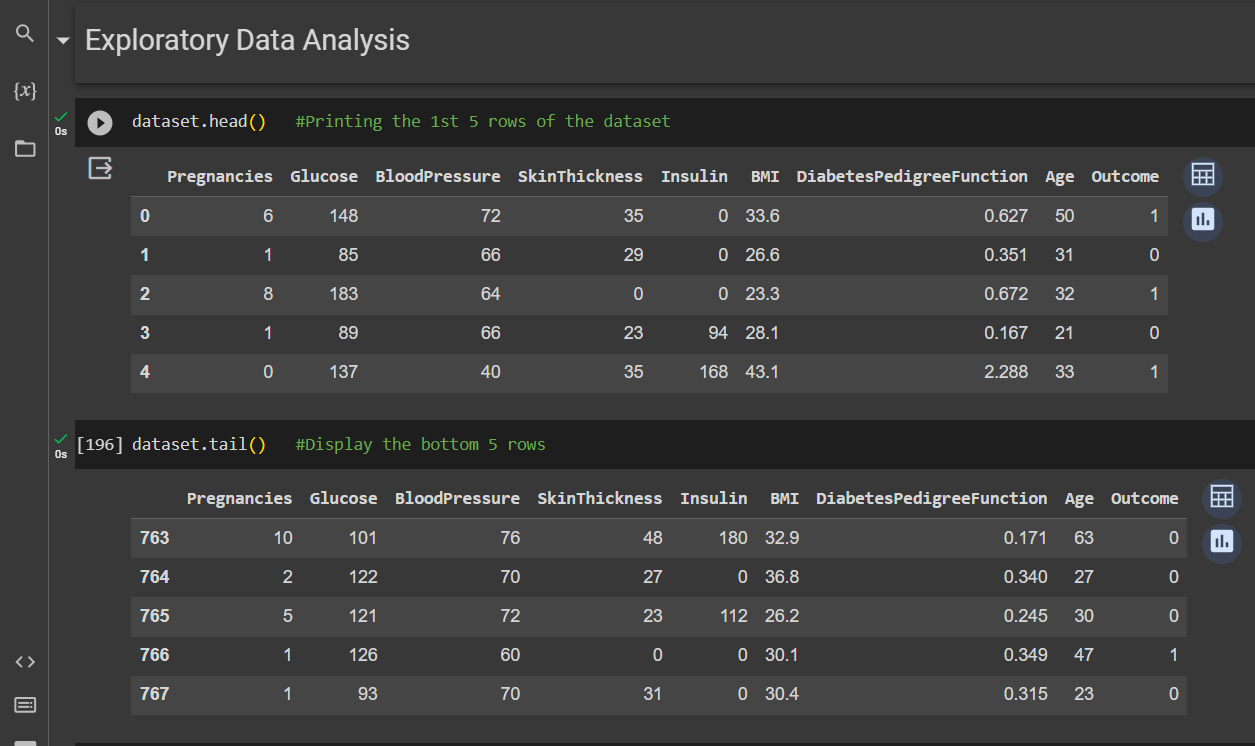
# Description of the Dataset

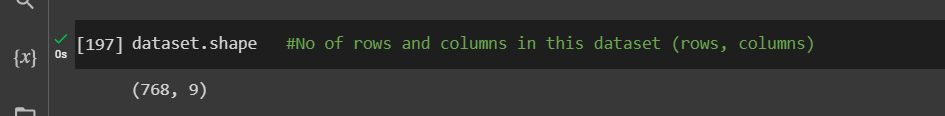
|  |  |  |
| --- | --- | --- |
| **Column Name** | **Description** | **Data Type** |
| Pregnancies | Number of times pregnant -  (0-17) | Int |
| Glucose | Plasma glucose concentration 2 hours in an oral glucose tolerance test - (0-199) | Int |
| Blood Pressure | Diastolic blood pressure (mm Hg) - (0-122) | Int |
| Skin Thickness | Triceps skin fold thickness (mm) – (0-99) | Int |
| Insulin | 2-Hour serum insulin (mu U/ml) – (0-846) | Int |
| BMI | Body mass index (weight in kg/(height in m)^2) – (0-67.1) | Float |
| Diabetes Pedigree Function | Diabetes pedigree function – (0.08-2.42) | Float |
| Age | Age (years) – ( 31-81) | Int |
| Outcome | Class variable (0 or 1) | Int |

# Data Identification

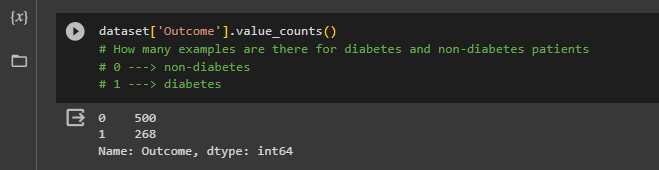
**Exploratory data Analysis:**

Print the first 5 rows of the data set and display bottom 5 rows. Then get the number of rows and columns in the data set by ‘dataset.shape’ command.

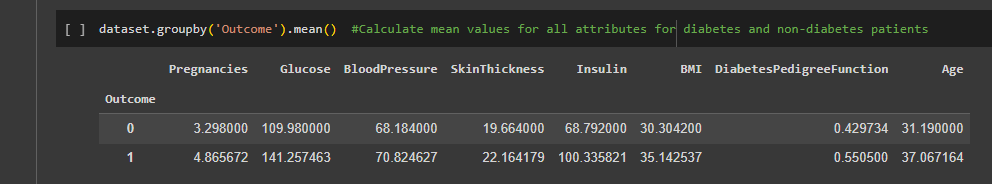




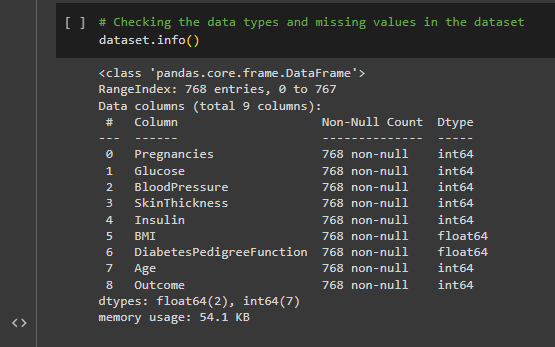
Get How many examples are there for diabetes and non-diabetes patients.



Calculate the mean values for all attributes for the diabetes and non-diabetes patients.

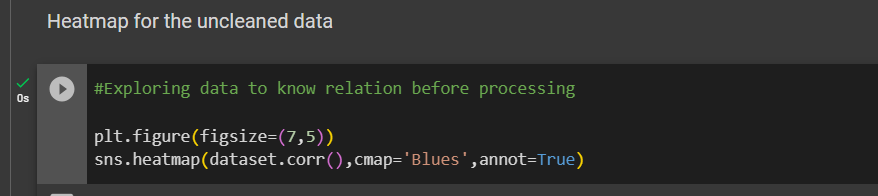


Checking the data types and missing values in the data set.



# Data Visualization Before Preprocessing

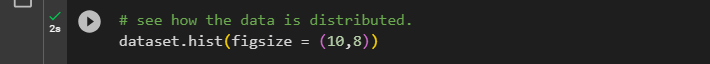
Exploring the data to know the relation before preprocessing using Heatmap.



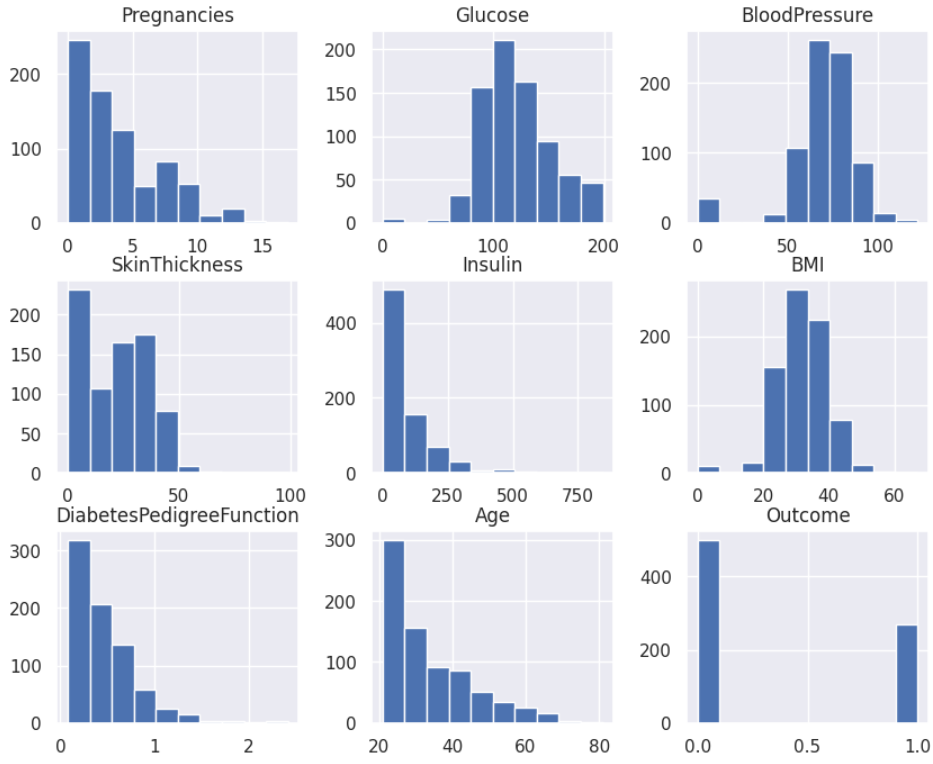
**Heatmap:**



Show the data distribution using Histograms before preprocessing.

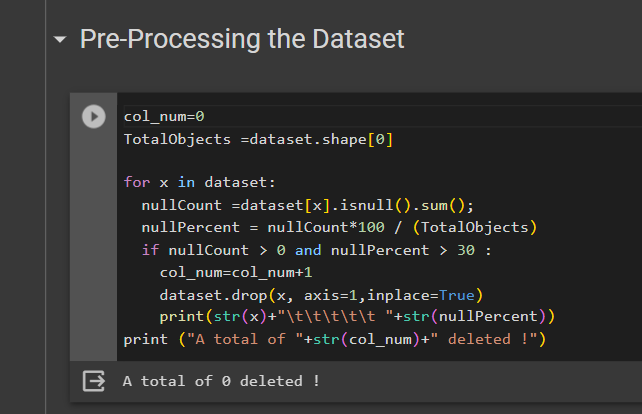


**Histograms:**

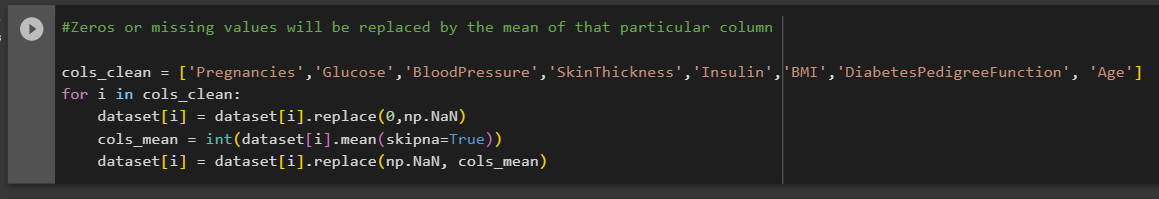


# Data Preprocessing

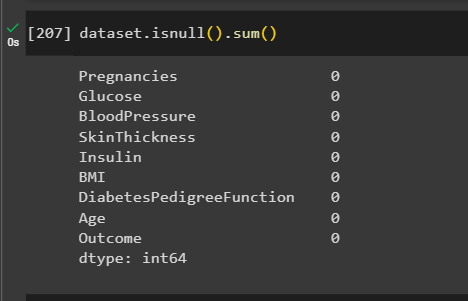
Identify and remove columns from the dataset that have a high percentage (greater than 30%) of null values.



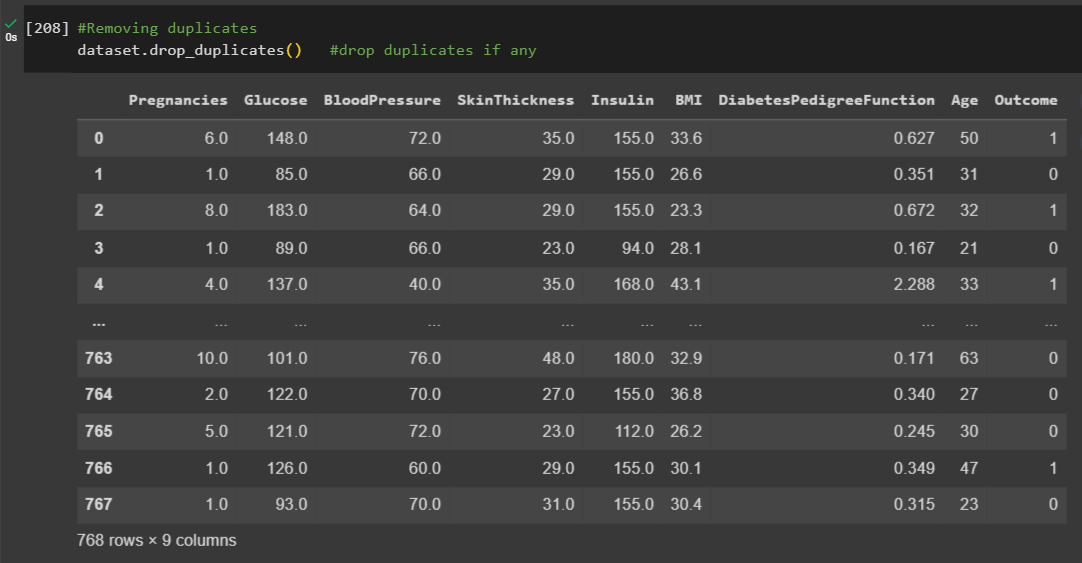
The zeros or missing values will be replaced by the means of that particular column.



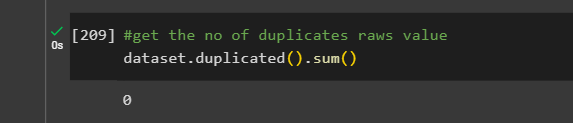
The code, ‘dataset.isnull().sum()’ is used to count the number of missing values (NaN or null values) in each column of a dataset.



Remove duplicate values and drop the columns if needed.

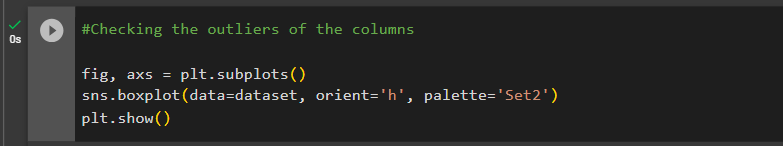


The code, ‘dataset.duplicated().sum()’ is used to count the number of duplicate rows in a dataset.

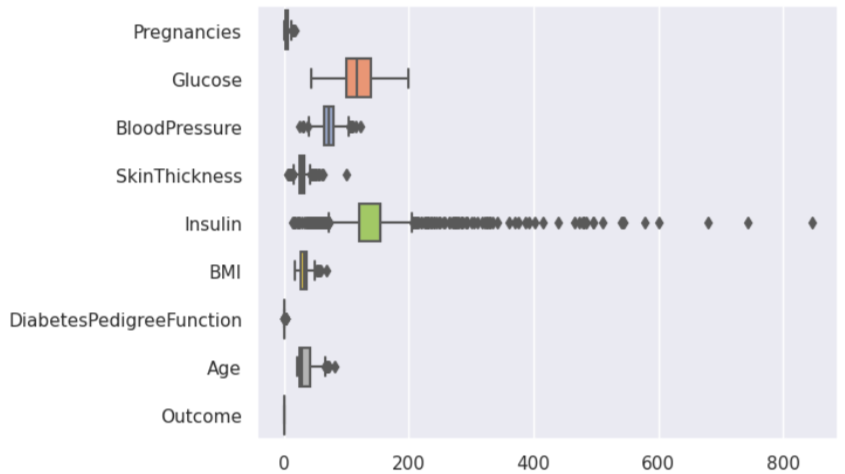


**Outlier detection and treatment**

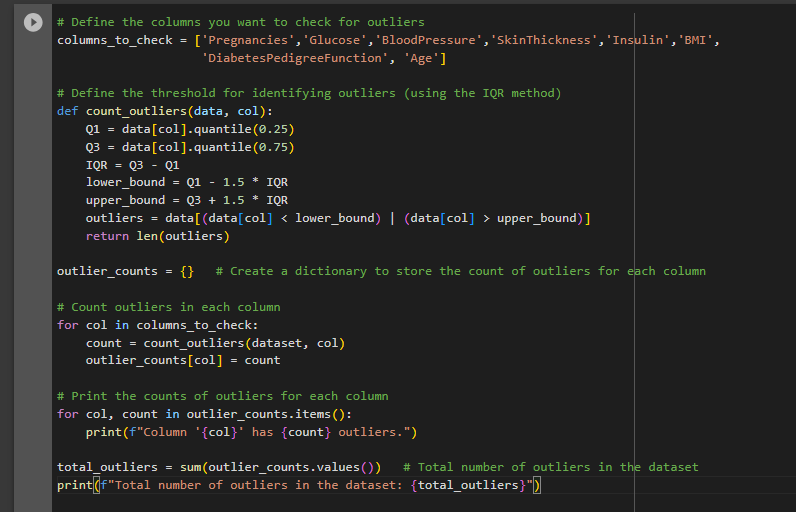
Checking the Outliers of the columns.

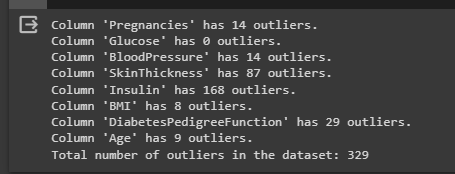


**Boxplot:**

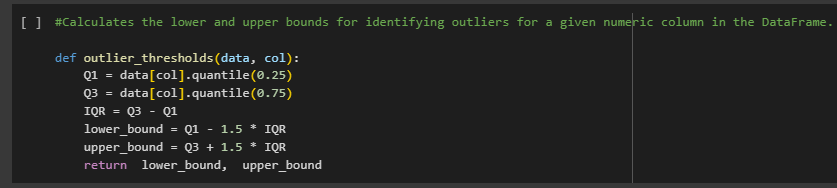


Print the outliers of each column.

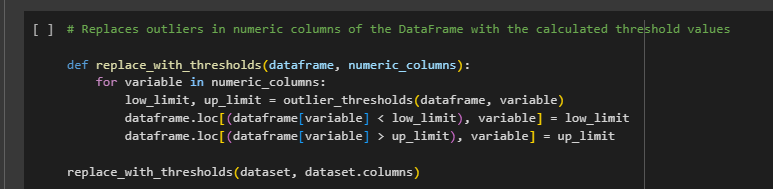




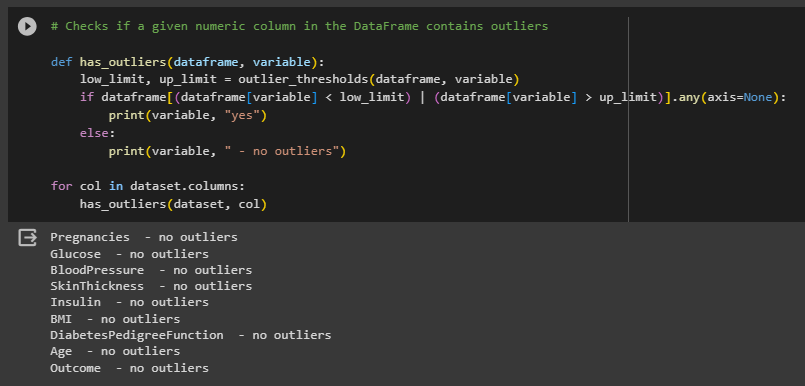
Calculates the lower and upper bounds for identifying outliers for a given numeric column in the DataFrame.



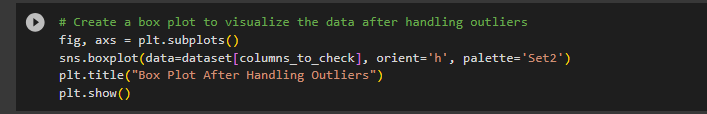
Handling outliers replacing outliers with calculated threshold values.



Checks if a given numeric column in the DataFrame contains outliers.



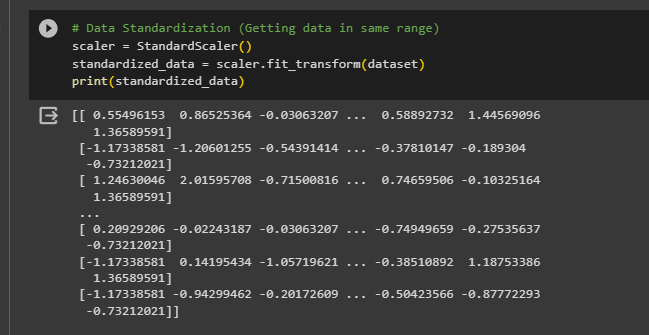
Box plot after handling the outliers.



**Boxplot:**

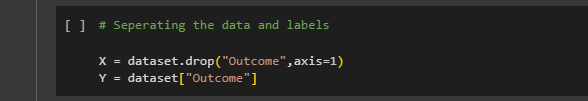


**Data Standardization** – Getting data in the same range.

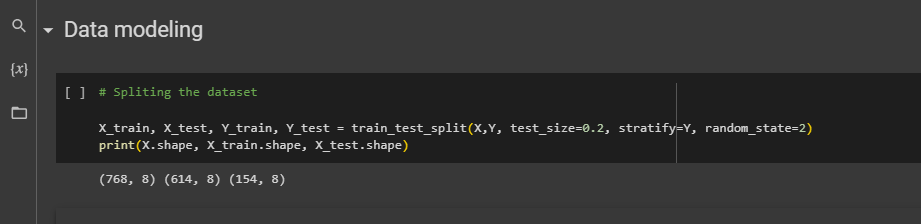


# Partitioning

Separating the data and labels.



Splitting the dataset into training and testing data.



# Proposed Data Mining Solutions

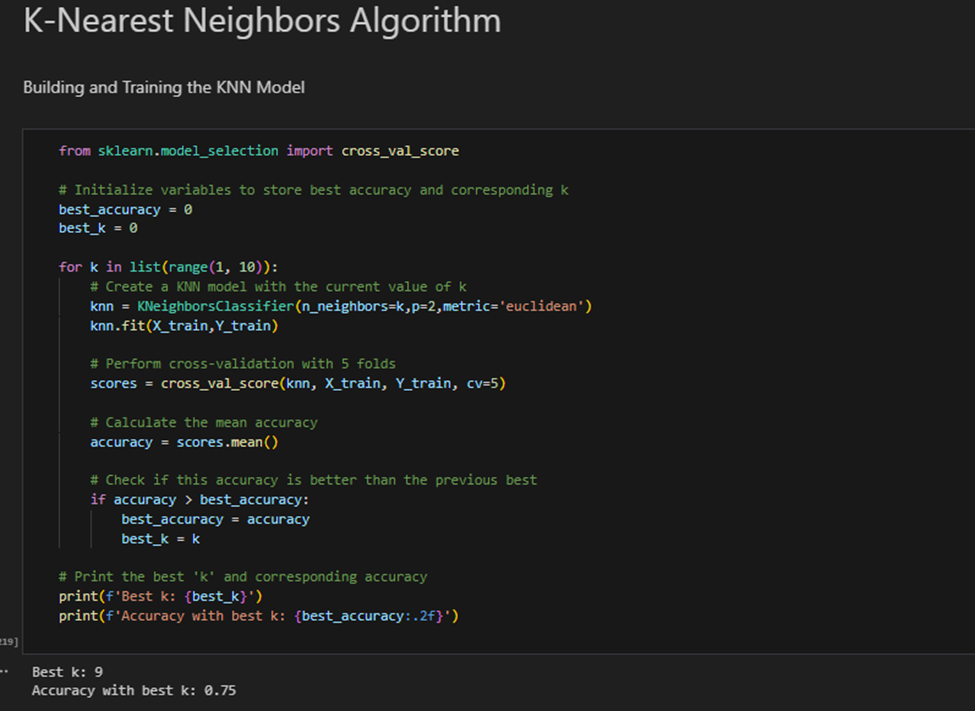
We employed K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Random Forest, and Logistic Regression models for the "Women's Diabetic Prediction System" dataset. Classification prediction modeling assigns class labels to input data, enabling the accurate prediction of target classes for each case. This choice of classification prediction aims to precisely forecast women's diabetic risk categories for improved healthcare outcomes.

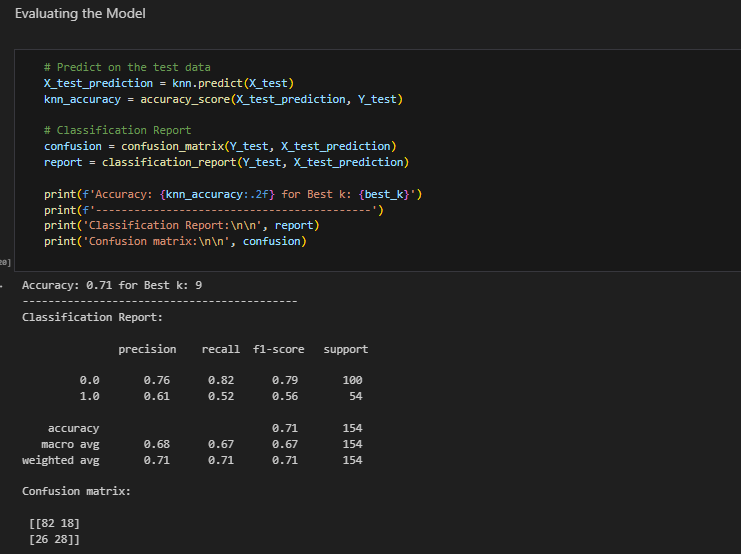
**Classification:**

In general, a classification algorithm within the Women's Diabetic Prediction System is a computational function that assesses input features to distinguish between various health conditions. This algorithm assigns positive and negative values to predict whether a woman is at risk of developing diabetes or not. Like other classification tasks, it relies on labeled data to accurately categorize individuals and facilitate the identification of those at risk for diabetes.

**Classification Model:**

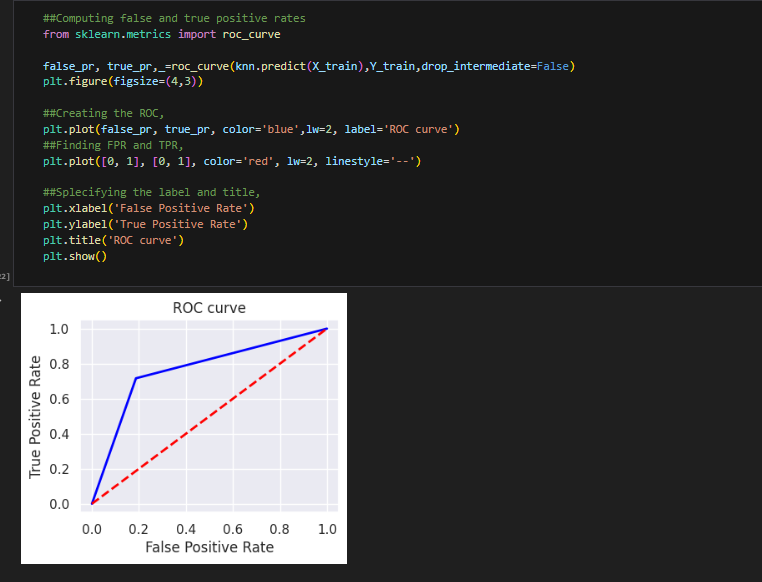
**1. K-Nearest Neighbors Algorithm (KNN)**





A screenshot of a computer

Description automatically generated



**2. Logistic Regression**

A screen shot of a computer program

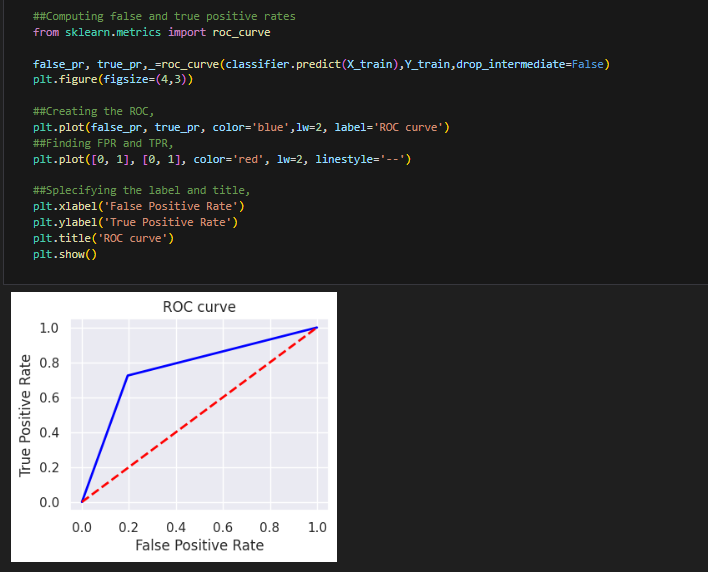
Description automatically generated

A screenshot of a computer program

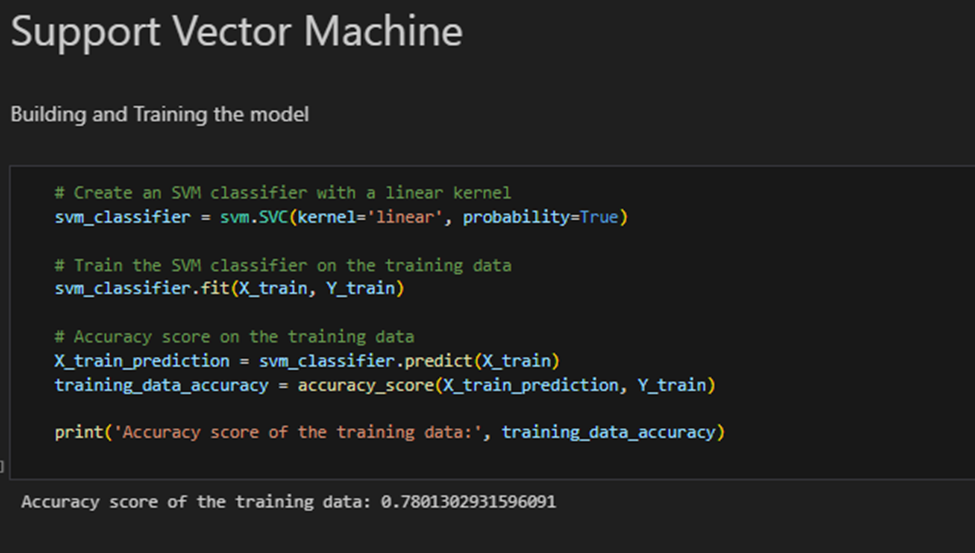
Description automatically generated

A screenshot of a computer

Description automatically generated



**3. Support Vector Machine**

****

A screenshot of a computer

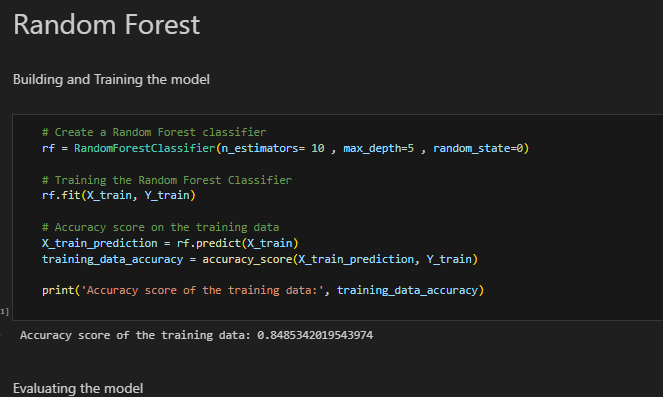
Description automatically generated

A screenshot of a computer

Description automatically generated



**4. Random Forest**



A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screen shot of a computer screen

Description automatically generated

# Model Selection

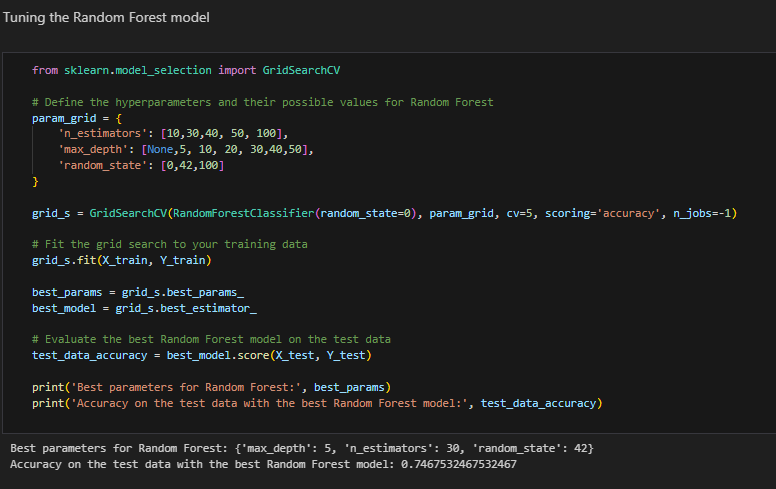
A screenshot of a computer program

Description automatically generated

After calculating the accuracy of each model, it has been determined that the 'Random Forest algorithm' for classification consistently achieved the highest level of accuracy among the alternatives. Therefore, Random Forest algorithm was chosen for further activities.

# Model Optimization

Tuning the model, which gave the best accuracy.



# Data Visualization

Feature importance’s bar plot for the selected Random Forest Model.

A screen shot of a computer program

Description automatically generated

**Bar plot:**

A graph with different colored bars

Description automatically generated

**Heatmap for the cleaned data:**

**A chart of different colored squares

Description automatically generated**

# Prediction

predict the risk that a patient can have diabetes based on input values.

.A screenshot of a computer program

Description automatically generated

The Output:

A screenshot of a computer

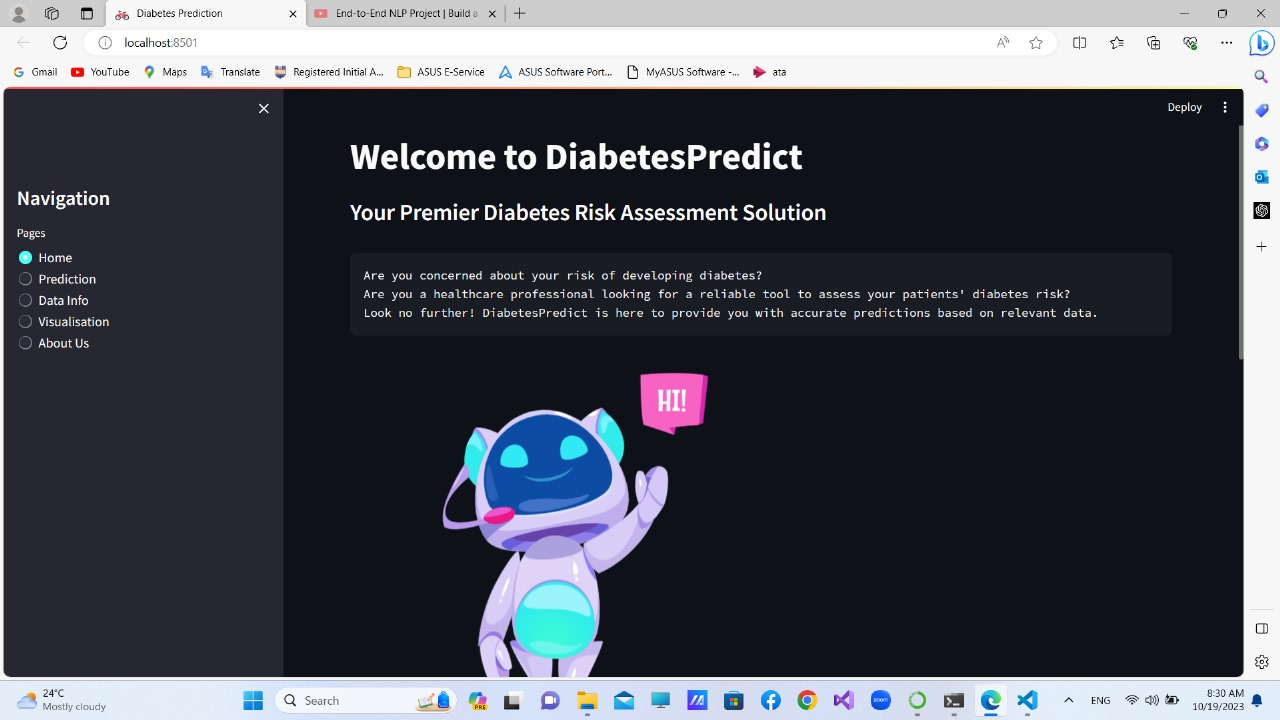
Description automatically generated

# Model Deployment

In our model development, we utilize a web-based platform, streamlining data processing, modeling, and collaboration. This web application enhances accessibility, efficiency, and collaborative efforts across our team.

**Technologies Used**: Python, Streamlit, scikit-learn, TensorFlow, Keras, Pandas and NumPy

**Home Page:**



**Prediction Page**: User can use both slider and text input as they prefer.

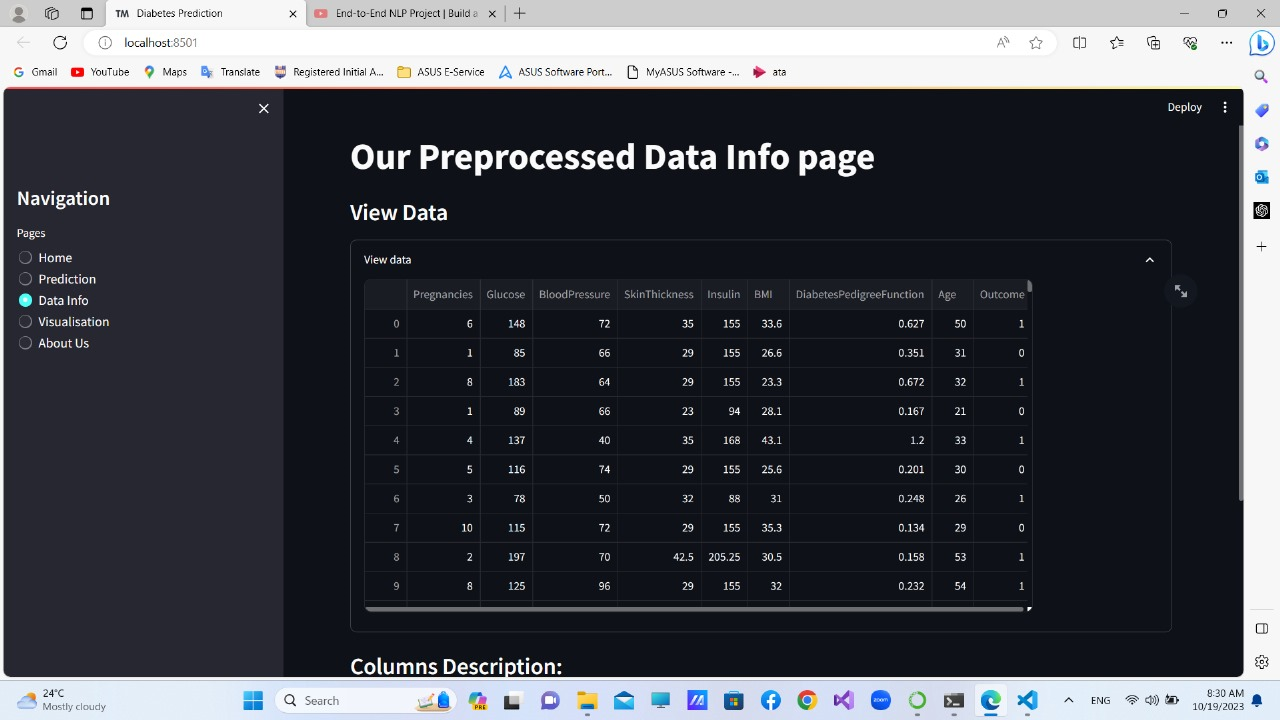
A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

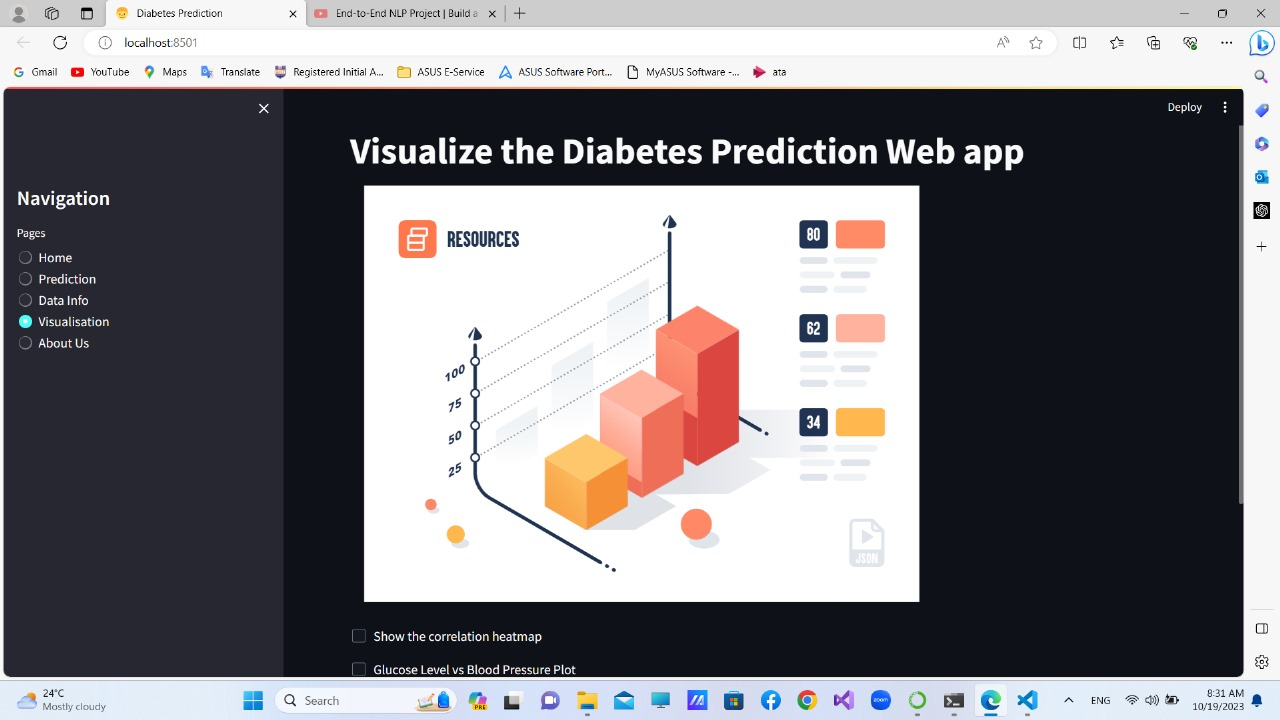
**Data Info Page**: View the Preprocessed dataset, summary, and other column details.



A screenshot of a computer

Description automatically generated

**Data Visualization Page**: Visualize the diabetes prediction web app with various graphs. Such as, Correlation Heatmap, Scatter Plots, Bar Plots etc.



**Correlation Heatmap:**

A screenshot of a computer

Description automatically generated

**Scatter Plots:**

Skin Thickness VS Insulin

A screenshot of a computer

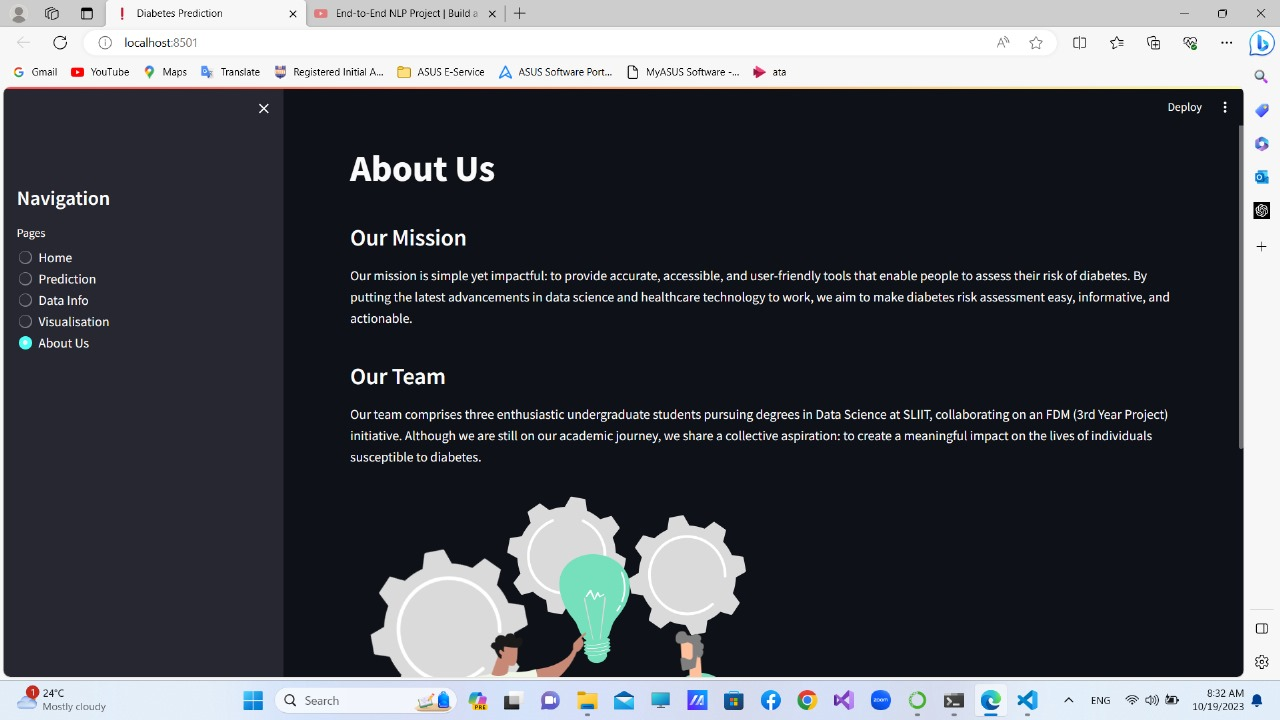
Description automatically generated

Glucose Level VS Blood Pressure

A screen shot of a computer

Description automatically generated

**About Us Page:**



# Deploying Implementation

Using Python, we've developed and trained machine learning models. We've saved these trained model objects as pickle files for serialization. Now, we've created a Streamlit environment with an API endpoint, enabling our trained models to receive feature inputs via HTTP/HTTPS POST requests. After deserializing the models in Streamlit, we can provide users with the measured output.

Through a REST API developed in Streamlit, our machine learning models are accessible to users. Streamlit, known for its user-friendly data app creation, allows us to present the models and capture user inputs effectively, offering an interactive interface for predictions and insights.

GitHub Link: <https://github.com/it21206832/FDM-Project-G29.git>

Hosted URL: <https://fdm-project-g29-diabetes-prediction.streamlit.app/>

A group of icons with text

Description automatically generated

# Test Cases

|  |  |  |  |
| --- | --- | --- | --- |
| **Test Scenario ID** | 01 | | |
| **Test Case Description** | Test the patient has diabetes or not with Slider Inputs | | |
| **Pre-Requisite** | Dataset was preprocessed | | |
| **Action** | Enter all the inputs and click predict | | |
| **Inputs** | **Expected Output** | **Actual Output** | **Result** |
| Pregnancies - 2  Glucose - 76  Blood Pressure - 62  Skin Thickness - 23  Insulin - 114  BMI - 28.62  DiabetesPedigreFunction-0.59  Age - 28 | A message displays as “The person is predicted to be free from diabetes. “ | A message displays as “The person is predicted to be free from diabetes. “ | **pass** |

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

|  |  |  |  |
| --- | --- | --- | --- |
| **Test Scenario ID** | 02 | | |
| **Test Case Description** | Test the patient has diabetes or not with Slider Inputs | | |
| **Pre-Requisite** | Dataset was preprocessed | | |
| **Action** | Enter all the inputs and click predict | | |
| **Inputs** | **Expected Output** | **Actual Output** | **Result** |
| Pregnancies - 5  Glucose - 100  Blood Pressure - 88  Skin Thickness - 35  Insulin - 154  BMI - 40.66  DiabetesPedigreFunction-0.94  Age - 51 | An error message displays as “The person is predicted to have a high risk of diabetes mellitus. “ | An error message displays as “The person is predicted to have a high risk of diabetes mellitus. “ | **pass** |

A screenshot of a computer

Description automatically generated

A screenshot of a computer

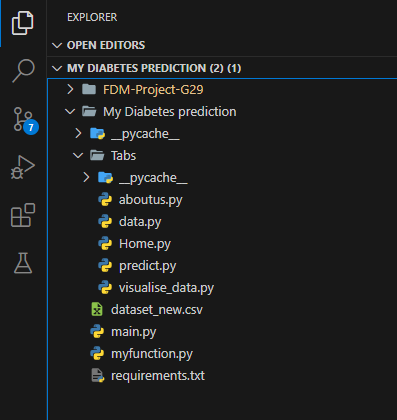
Description automatically generated

|  |  |  |  |
| --- | --- | --- | --- |
| **Test Scenario ID** | 03 | | |
| **Test Case Description** | Test the patient has diabetes or not with Text Input | | |
| **Pre-Requisite** | Dataset was preprocessed | | |
| **Action** | Enter all the inputs and click predict | | |
| **Inputs** | **Expected Output** | **Actual Output** | **Result** |
| Pregnancies - 3  Glucose - 90  Blood Pressure - 88  Skin Thickness - 30  Insulin - 10  BMI - 20.5  DiabetesPedigreFunction-  Age - 20 | An Error message displays as “Please enter valid numeric values for all input fields. “ | An Error message displays as “Please enter valid numeric values for all input fields. “ | **pass** |

A screenshot of a computer

Description automatically generated

# The Project Structure



* \_\_**pyacache\_\_ folder** - It's a folder automatically generated by Python to store compiled bytecode files (.pyc) to improve the execution speed of Python scripts. These files are specific to local development environment and are not needed when deploying or sharing a Streamlit app.
* **main.py** - This Python script is a Streamlit app with tabbed navigation, offering features to load and predict diabetes data, visualize data, and provide information about the project and its creators.
* **myfunction.py -** This Python script appears to define functions for loading, splitting, and training a Random Forest Classifier on diabetes data, along with an evaluation function.
* **predict.py –** This Streamlit app predicts diabetes using a Random Forest Classifier. Users input data via sliders or text, and predictions are displayed.

# Benefits of Proposed Solution

The proposed women's diabetic prediction solution offers several key benefits. It empowers early diabetes risk assessment, improving health outcomes. Tailored for women, it addresses gender-specific health concerns. The user-friendly interface enhances accessibility. Privacy measures ensure data security. This solution not only predicts diabetes risk but also promotes health awareness. By advancing medical technology, it contributes to overall public health improvement and individual well-being.

* Early Intervention
* Enhanced Personal Health Records
* Streamlined Healthcare Processes
* Tailored Preventive Care
* Data-Driven Decision-Making

# Conclusion

The Women's Diabetic Prediction System leverages advanced predictive modeling techniques to assess the risk of diabetes in women. With an accuracy rate of 78%, this system equips healthcare professionals and individuals with a valuable tool for early intervention and effective disease management.

Its user-friendly interface ensures accessibility, making it an efficient and reliable solution for health assessment. Without this system, identifying diabetes risk early would be a challenging task for healthcare providers.

In conclusion, the successful implementation, testing, and integration of this technology signify a significant advancement in women's health management, promising a brighter future in the prevention and treatment of diabetes among women.

# Project Team and Workload

|  |  |  |
| --- | --- | --- |
| **Registration Number** | **Name** | **Responsibility** |
| IT21307294 | Gamage D G A S | Scope Planning  Implement the Classification Model Select the best model.  Data Preprocessing  Evaluate Model Documentation  UI Design and Integrate |
| IT21255106 | Nelligahawaththa A D T B | Scope Planning  Implement the Classification Model Select the best model.  Data Preprocessing  Evaluate Model Documentation  UI Design and Integrate |
| IT21206832 | Pabasara S D | Scope Planning  Implement the Classification Model Select the best model.  Data Preprocessing  Evaluate Model Documentation  UI Design and Integrate |

# References

* [**https://www.kaggle.com/datasets/mathchi/diabetes-data-set**](https://www.kaggle.com/datasets/mathchi/diabetes-data-set)
* [**https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10107388/#:~:text=This%20work%20used%20four%20types,the%20highest%20accuracy%20of%200.82.**](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10107388/%23:~:text=This%20work%20used%20four%20types,the%20highest%20accuracy%20of%200.82.)
* [**https://www.analyticsvidhya.com/blog/2022/01/diabetes-prediction-using-machine-learning/**](https://www.analyticsvidhya.com/blog/2022/01/diabetes-prediction-using-machine-learning/)
* [**https://ieeexplore.ieee.org/document/8748992**](https://ieeexplore.ieee.org/document/8748992)
* [**https://www.frontiersin.org/articles/10.3389/fgene.2018.00515/full**](https://www.frontiersin.org/articles/10.3389/fgene.2018.00515/full)