M.Sc. DATA SCIENCE

Project Report

On

Predicting Diabetes Risk with k-Nearest Neighbors (k-NN)

MDS372 - MACHINE LEARNING

CIA 3 – Modular Coding

Submitted by

Angeline A

M.Sc. Data Science B

2348409

APRIL 2024

CHRIST (Deemed to be University)

Yeshwanthpur Campus

Bangalore, Karnataka

TABLE OF CONTENTS

1. AIM

The aim of our project is to develop a comprehensive Diabetes Risk Prediction tool utilizing advanced machine learning techniques. This tool is meticulously crafted to aid individuals in accurately assessing their susceptibility to developing diabetes by evaluating a spectrum of crucial health parameters. By meticulously analyzing factors such as glucose levels, blood pressure, BMI (Body Mass Index), insulin levels, diabetes pedigree function, and age, our tool endeavors to furnish users with personalized risk assessments. Through this personalized approach, our tool seeks to empower individuals with actionable insights into their diabetes risk profile, thereby fostering early detection, prevention, and effective management of this chronic condition. By harnessing the power of data-driven analysis and predictive modeling, our ultimate goal is to equip users with the knowledge and tools necessary to make informed decisions about their health, ultimately leading to improved overall well-being and a better quality of life.

2. INTRODUCTION

Diabetes has emerged as a formidable health challenge worldwide, with its prevalence escalating to alarming levels in recent years. Addressing this growing epidemic requires innovative approaches that leverage advancements in technology and healthcare. Our project endeavors to tackle this issue by developing a Diabetes Risk Prediction tool powered by machine learning. The primary objective of this tool is to empower individuals to assess their risk of developing diabetes based on a comprehensive analysis of various health parameters. By harnessing the predictive capabilities of machine learning algorithms, our tool aims to provide personalized risk assessments, enabling users to take proactive steps towards managing their health.

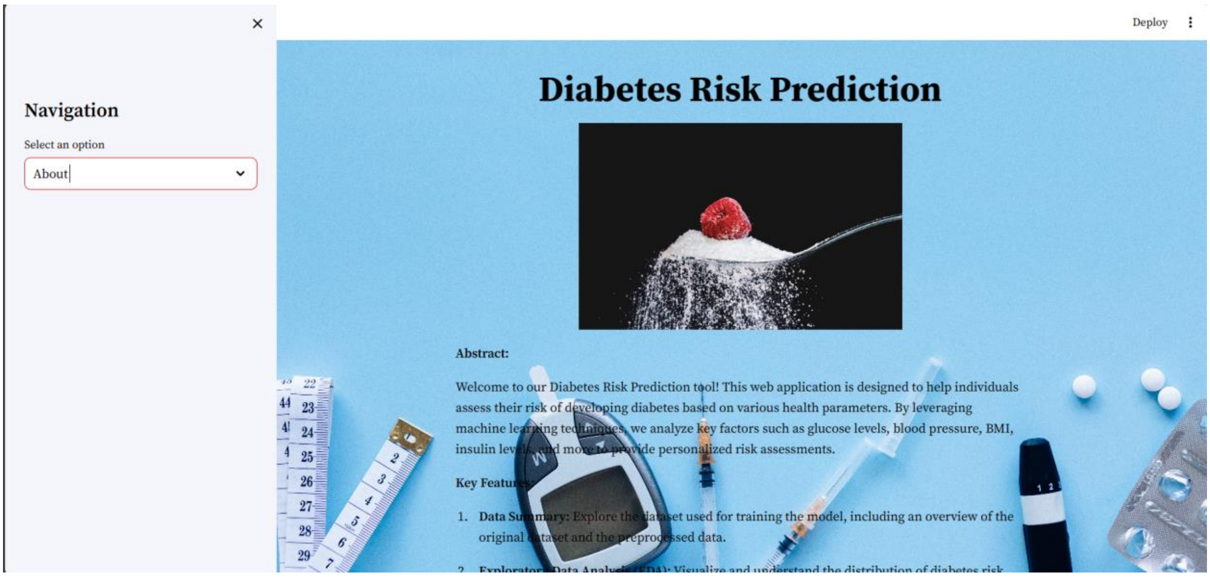
In today's era of personalized medicine, there is a growing recognition of the importance of individualized risk assessment in disease prevention and management. Our Diabetes Risk Prediction tool aligns with this paradigm shift by offering tailored risk assessments that take into account an individual's unique health profile. By analyzing factors such as glucose levels, blood pressure, BMI, insulin levels, diabetes pedigree function, and age, our tool generates insights that are relevant and actionable for each user. Through this project, we aspire to contribute to the global efforts aimed at combating diabetes by empowering individuals with the knowledge and tools they need to make informed decisions about their health and well-being.

3. PROCEDURE

1) **Data Acquisition and Pre-processing:**
    - Load patient data from a CSV file.
    - Handle missing values in the dataset.
    - Clean the dataset by removing outliers.
    - Preprocess the data by handling missing values, cleaning, and encoding categorical variables.
2) **Exploratory Data Analysis (EDA):**
    - Visualize the distribution of diabetes risk levels.
    - Analyze correlations between features and diabetes risk.
    - Explore relationships between features through interactive plots.
    - Provide insights into diabetes risk factors.
3) **Model Building and Evaluation:**
    - Split the data into training and testing sets.
    - Train a k-NN model to predict diabetes risk.
    - Evaluate model performance using metrics such as accuracy, precision, recall, and F1-score.
    - Display the classification report to interpret model performance.
4) **Model Deployment and Monitoring:**
    - Save the trained model for future use.
    - Load the trained model from file.
    - Predict diabetes risk based on user input.
    - Provide personalized suggestions based on input values.
    - Visualize prediction results and suggestions.
5) **Update Based on Feature Importance:**
    - Identify the impact of different features on model accuracy.
    - Monitor changes in model accuracy based on features added or removed.
    - Optimize the model by selecting the most relevant features for prediction.

4. OUTPUT (Refer the Code in Zipped file)

SCREENSHOTS



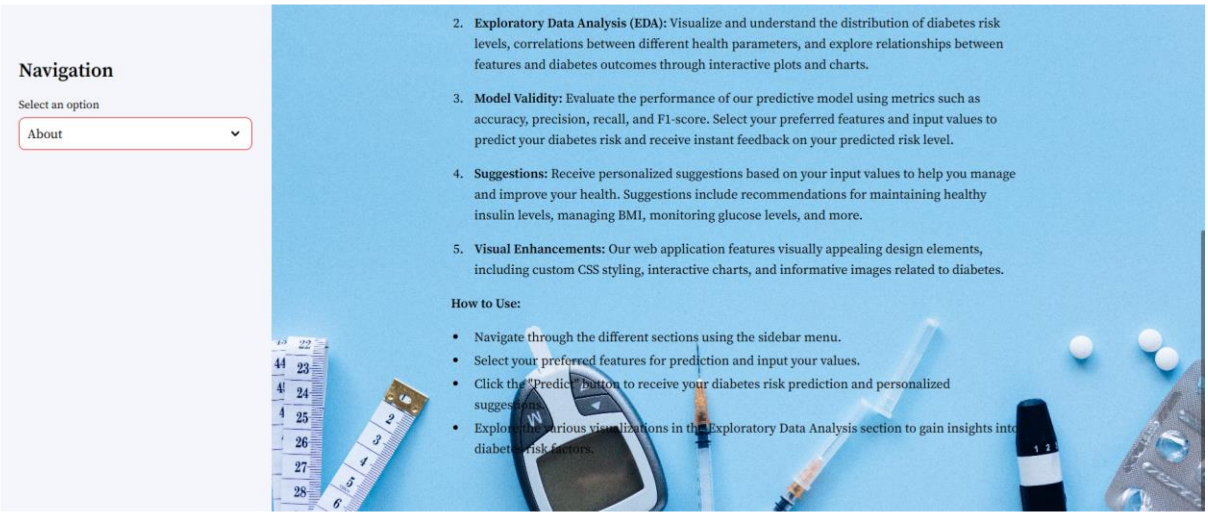
Diabetes Risk Prediction

Abstract:

Welcome to our Diabetes Risk Prediction tool! This web application is designed to help individuals assess their risk of developing diabetes based on various health parameters. By leveraging machine learning techniques, we analyze key factors such as glucose levels, blood pressure, BMI, insulin levels, and more to provide personalized risk assessments.

Key Features:

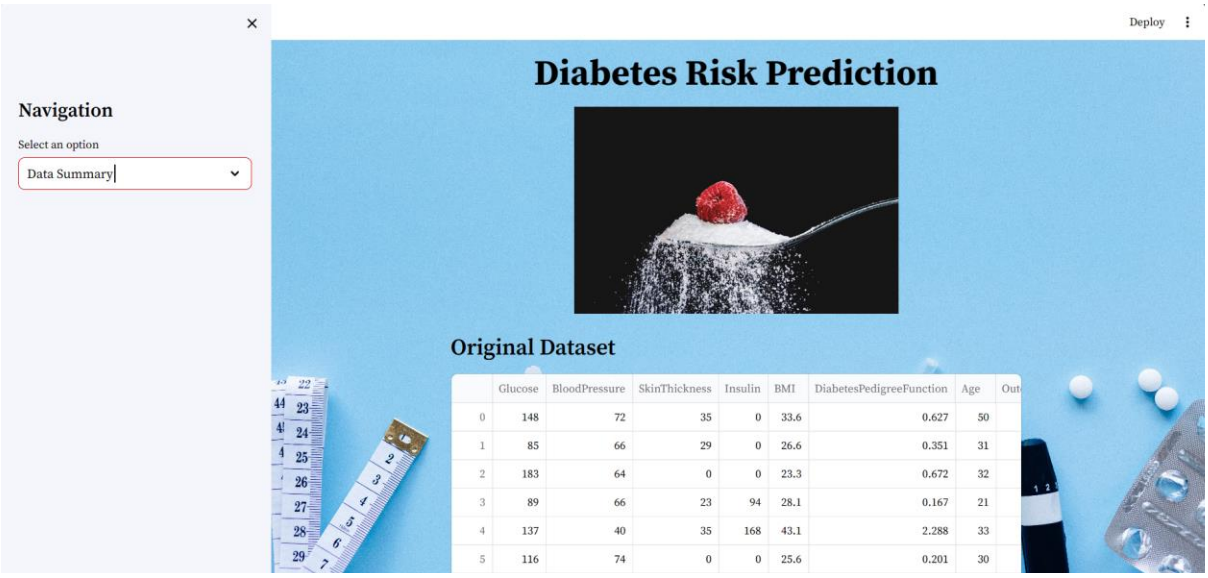
1. **Data Summary:** Explore the dataset used for training the model, including an overview of the original dataset and the preprocessed data.
2. **Exploratory Data Analysis (EDA):** Visualize and understand the distribution of diabetes risk



2. **Exploratory Data Analysis (EDA):** Visualize and understand the distribution of diabetes risk levels, correlations between different health parameters, and explore relationships between features and diabetes outcomes through interactive plots and charts.
3. **Model Validity:** Evaluate the performance of our predictive model using metrics such as accuracy, precision, recall, and F1-score. Select your preferred features and input values to predict your diabetes risk and receive instant feedback on your predicted risk level.
4. **Suggestions:** Receive personalized suggestions based on your input values to help you manage and improve your health. Suggestions include recommendations for maintaining healthy insulin levels, managing BMI, monitoring glucose levels, and more.
5. **Visual Enhancements:** Our web application features visually appealing design elements, including custom CSS styling, interactive charts, and informative images related to diabetes.

How to Use:

- Navigate through the different sections using the sidebar menu.
- Select your preferred features for prediction and input your values.
- Click the "Predict" button to receive your diabetes risk prediction and personalized suggestions.
- Explore the various visualizations in the Exploratory Data Analysis section to gain insights into diabetes risk factors.



Navigation

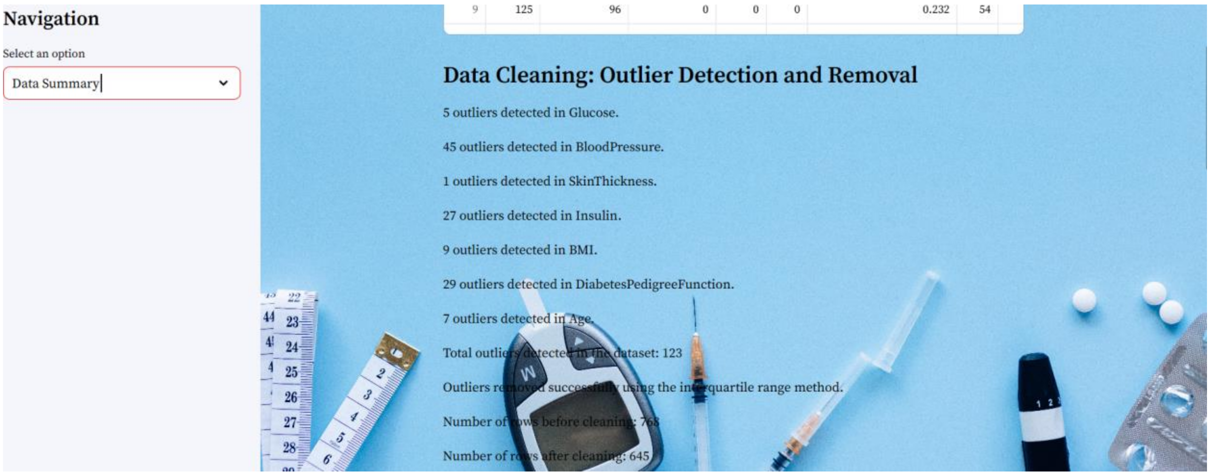
Select an option

Data Summary

Diabetes Risk Prediction

Original Dataset

	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Out
0	148	72	35	0	33.6	0.627	50	
1	85	66	29	0	26.6	0.351	31	
2	183	64	0	0	23.3	0.672	32	
3	89	66	23	94	28.1	0.167	21	
4	137	40	35	168	43.1	2.288	33	
5	116	74	0	0	25.6	0.201	30	



Navigation

Select an option

Data Summary

9	125	96	0	0	0	0.232	54	

Data Cleaning: Outlier Detection and Removal

5 outliers detected in Glucose.

45 outliers detected in BloodPressure.

1 outliers detected in SkinThickness.

27 outliers detected in Insulin.

9 outliers detected in BMI.

29 outliers detected in DiabetesPedigreeFunction.

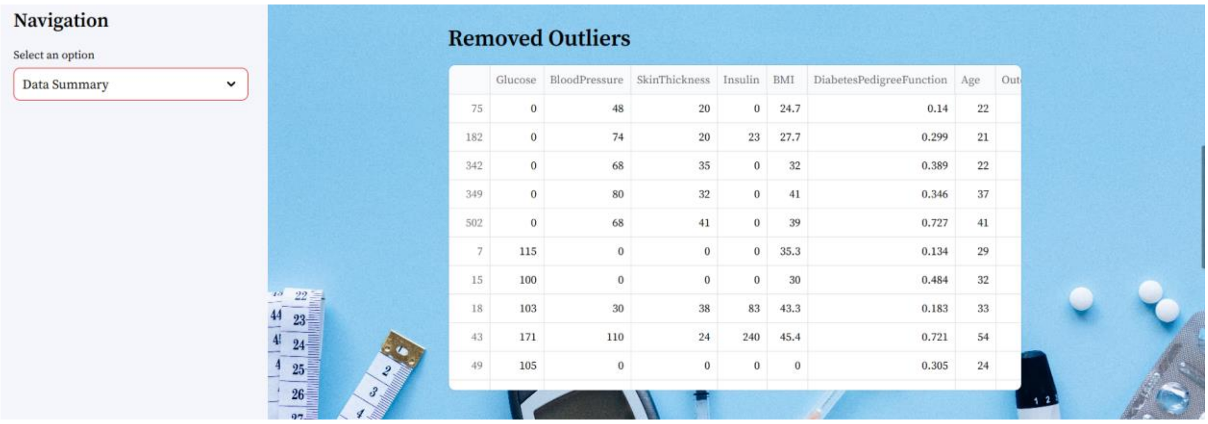
7 outliers detected in Age.

Total outliers detected in the dataset: 123

Outliers removed successfully using the interquartile range method.

Number of rows before cleaning: 768

Number of rows after cleaning: 645



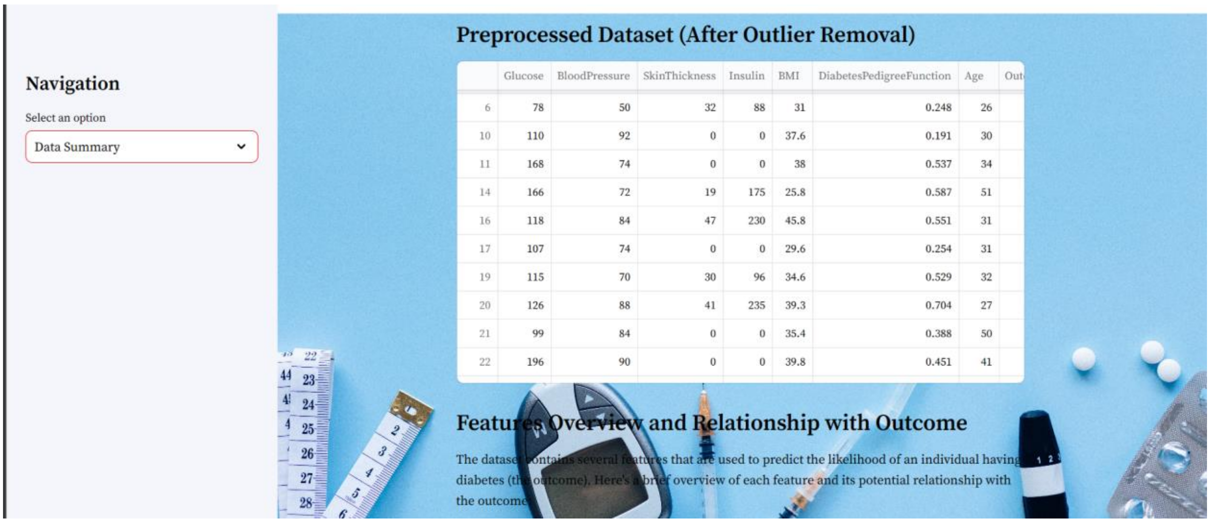
Navigation

Select an option

Data Summary

Removed Outliers

	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Out
75	0	48	20	0	24.7	0.14	22	
182	0	74	20	23	27.7	0.299	21	
342	0	68	35	0	32	0.389	22	
349	0	80	32	0	41	0.346	37	
502	0	68	41	0	39	0.727	41	
7	115	0	0	0	35.3	0.134	29	
15	100	0	0	0	30	0.484	32	
18	103	30	38	83	43.3	0.183	33	
43	171	110	24	240	45.4	0.721	54	
49	105	0	0	0	0	0.305	24	

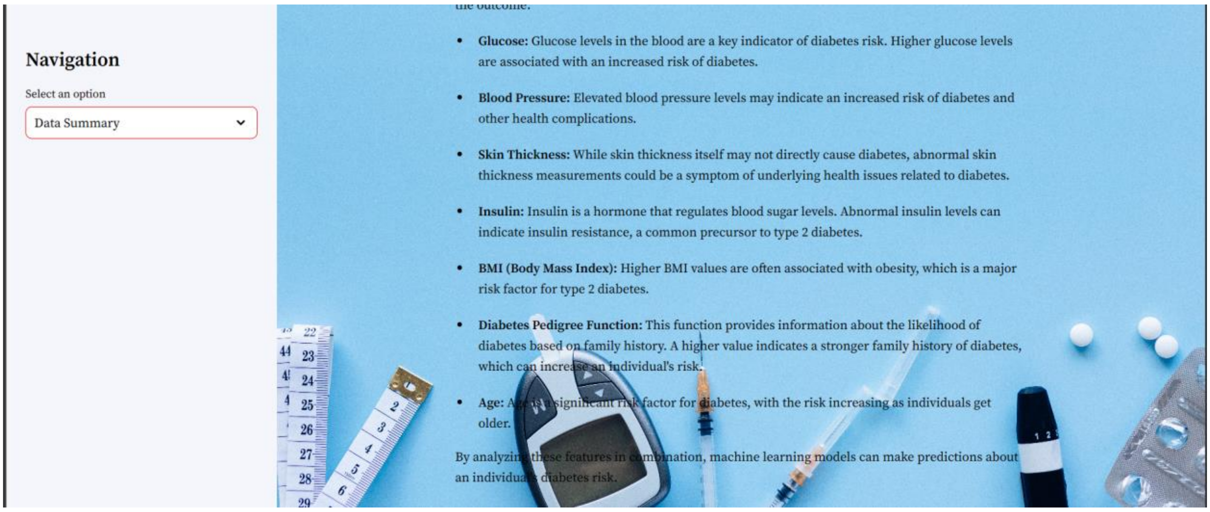


Preprocessed Dataset (After Outlier Removal)

	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Out
6	78	50	32	88	31	0.248	26	
10	110	92	0	0	37.6	0.191	30	
11	168	74	0	0	38	0.537	34	
14	166	72	19	175	25.8	0.587	51	
16	118	84	47	230	45.8	0.551	31	
17	107	74	0	0	29.6	0.254	31	
19	115	70	30	96	34.6	0.529	32	
20	126	88	41	235	39.3	0.704	27	
21	99	84	0	0	35.4	0.388	50	
22	196	90	0	0	39.8	0.451	41	

Features Overview and Relationship with Outcome

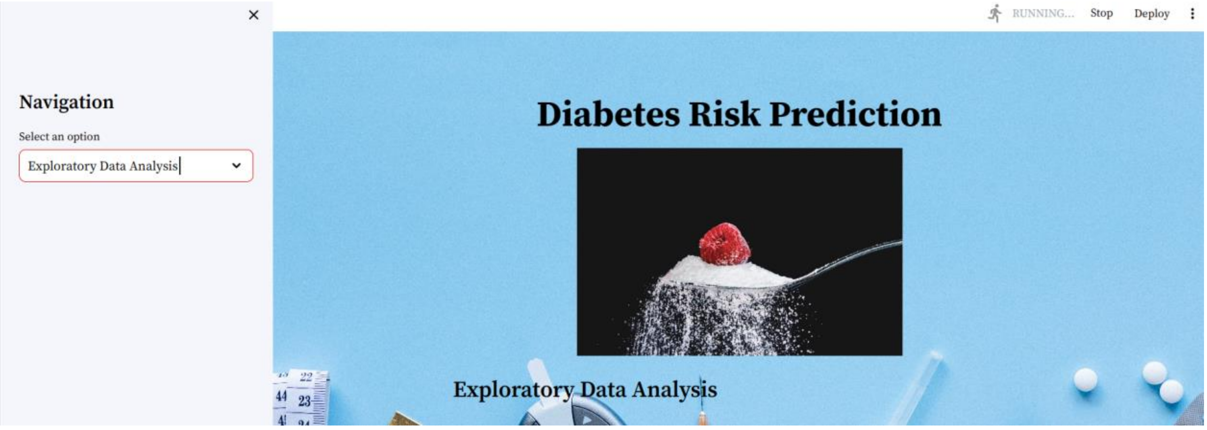
The dataset contains several features that are used to predict the likelihood of an individual having diabetes (the outcome). Here's a brief overview of each feature and its potential relationship with the outcome.



the outcome.

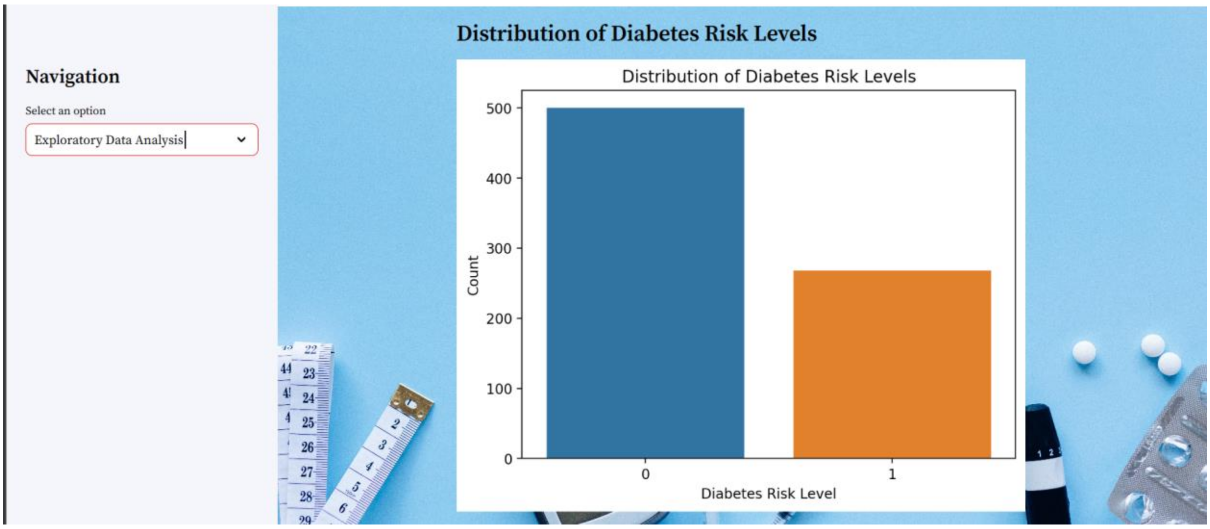
- **Glucose:** Glucose levels in the blood are a key indicator of diabetes risk. Higher glucose levels are associated with an increased risk of diabetes.
- **Blood Pressure:** Elevated blood pressure levels may indicate an increased risk of diabetes and other health complications.
- **Skin Thickness:** While skin thickness itself may not directly cause diabetes, abnormal skin thickness measurements could be a symptom of underlying health issues related to diabetes.
- **Insulin:** Insulin is a hormone that regulates blood sugar levels. Abnormal insulin levels can indicate insulin resistance, a common precursor to type 2 diabetes.
- **BMI (Body Mass Index):** Higher BMI values are often associated with obesity, which is a major risk factor for type 2 diabetes.
- **Diabetes Pedigree Function:** This function provides information about the likelihood of diabetes based on family history. A higher value indicates a stronger family history of diabetes, which can increase an individual's risk.
- **Age:** Age is a significant risk factor for diabetes, with the risk increasing as individuals get older.

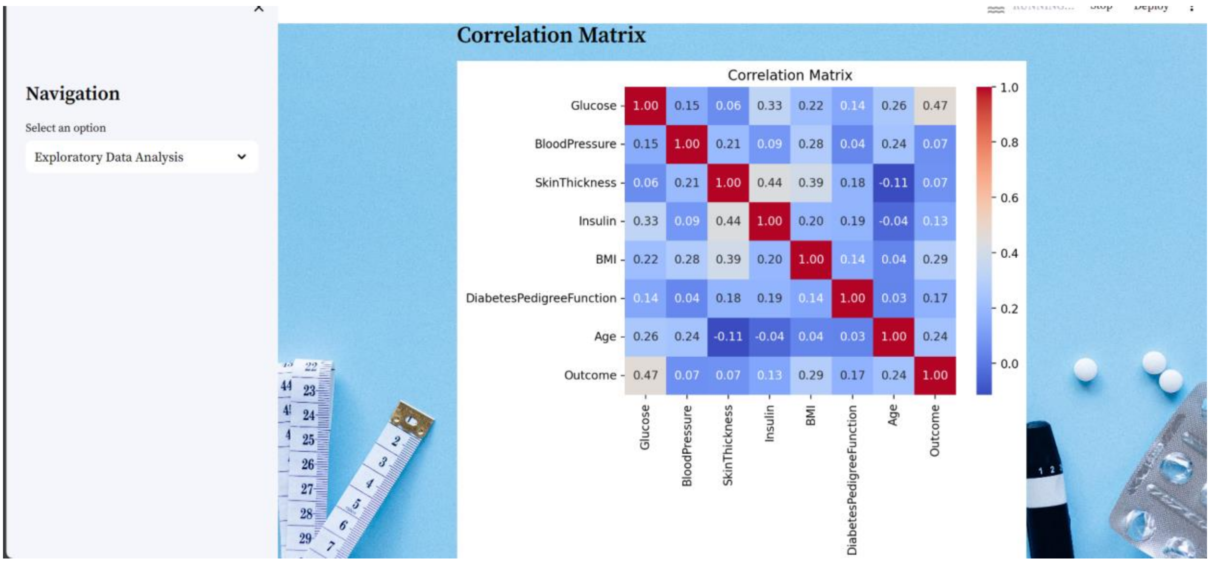
By analyzing these features in combination, machine learning models can make predictions about an individual's diabetes risk.



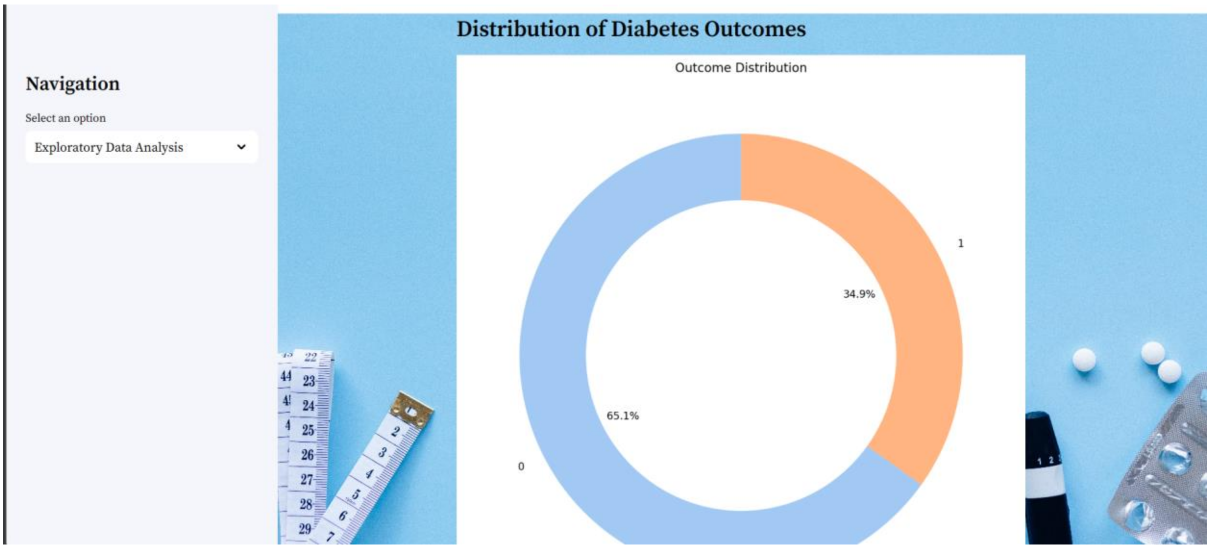
Diabetes Risk Prediction

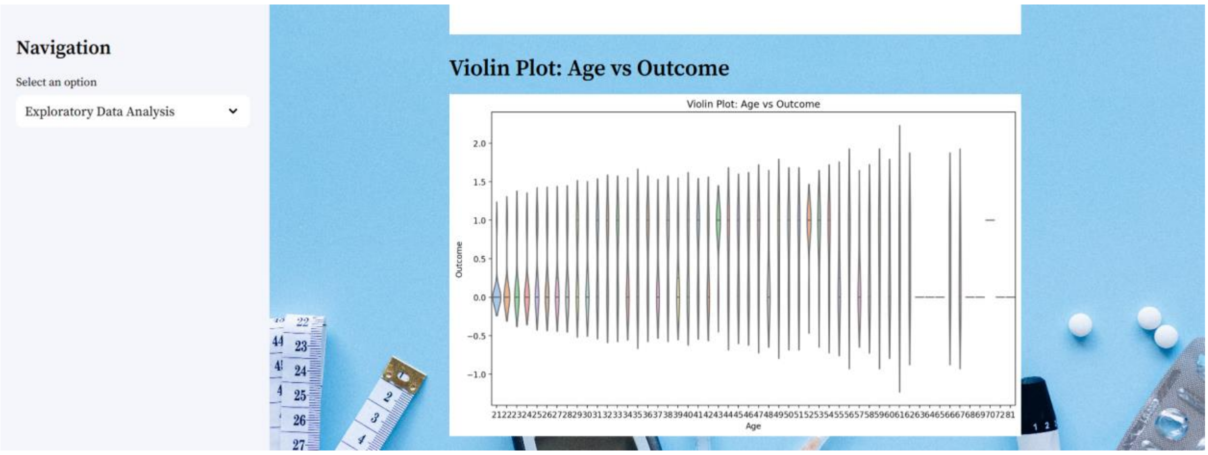
Exploratory Data Analysis


Navigation
Select an option
Exploratory Data Analysis
Distribution of Diabetes Risk Levels
Distribution of Diabetes Risk Levels
Count
500
400
300
200
100
0
0
1
Diabetes Risk Level



	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
Glucose	1.00	0.15	0.06	0.33	0.22	0.14	0.26	0.47
BloodPressure	0.15	1.00	0.21	0.09	0.28	0.04	0.24	0.07
SkinThickness	0.06	0.21	1.00	0.44	0.39	0.18	-0.11	0.07
Insulin	0.33	0.09	0.44	1.00	0.20	0.19	-0.04	0.13
BMI	0.22	0.28	0.39	0.20	1.00	0.14	0.04	0.29
DiabetesPedigreeFunction	0.14	0.04	0.18	0.19	0.14	1.00	0.03	0.17
Age	0.26	0.24	-0.11	-0.04	0.04	0.03	1.00	0.24
Outcome	0.47	0.07	0.07	0.13	0.29	0.17	0.24	1.00



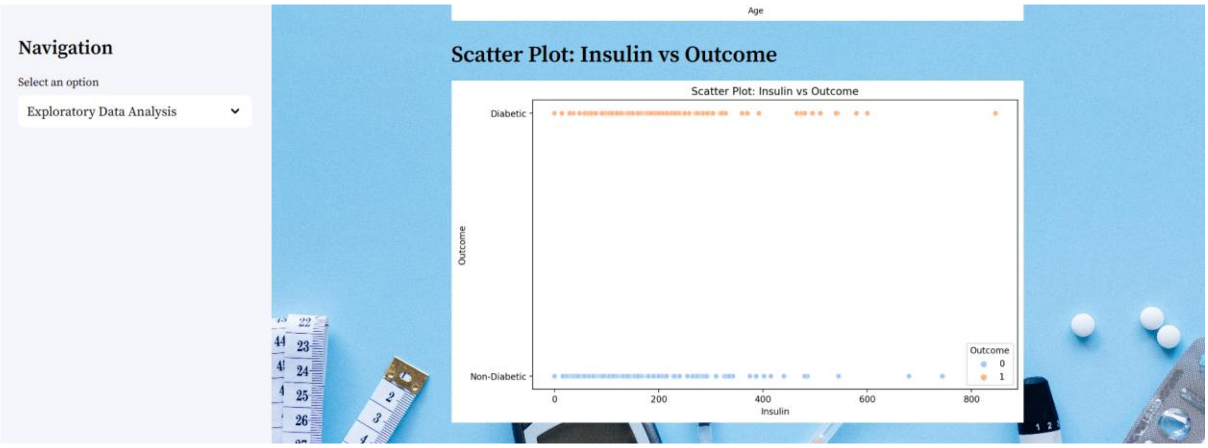


Navigation

Select an option

Exploratory Data Analysis

Violin Plot: Age vs Outcome

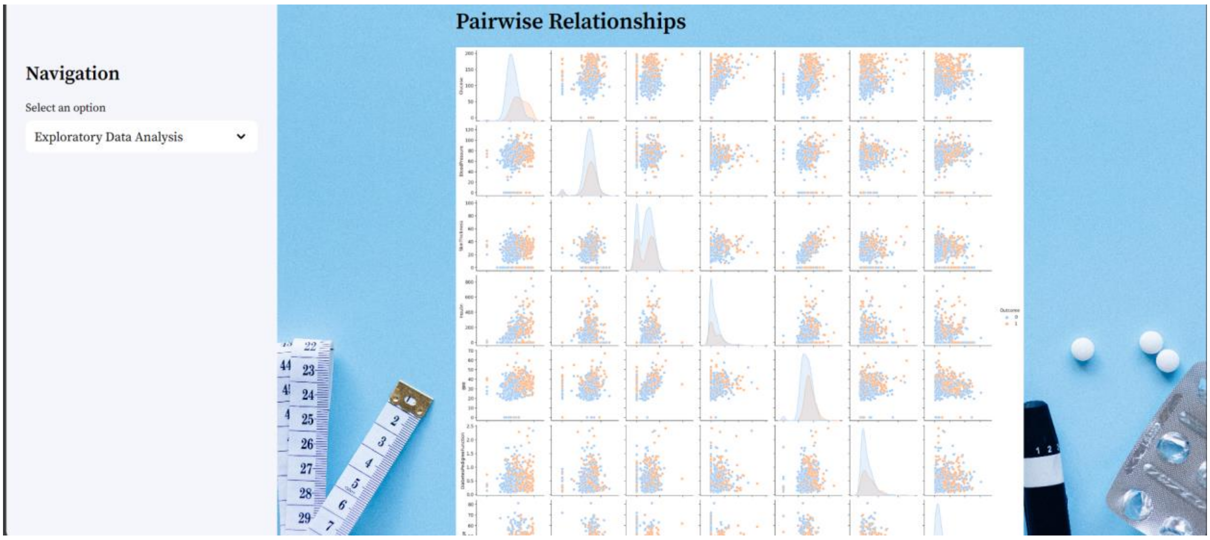


Navigation

Select an option

Exploratory Data Analysis

Scatter Plot: Insulin vs Outcome

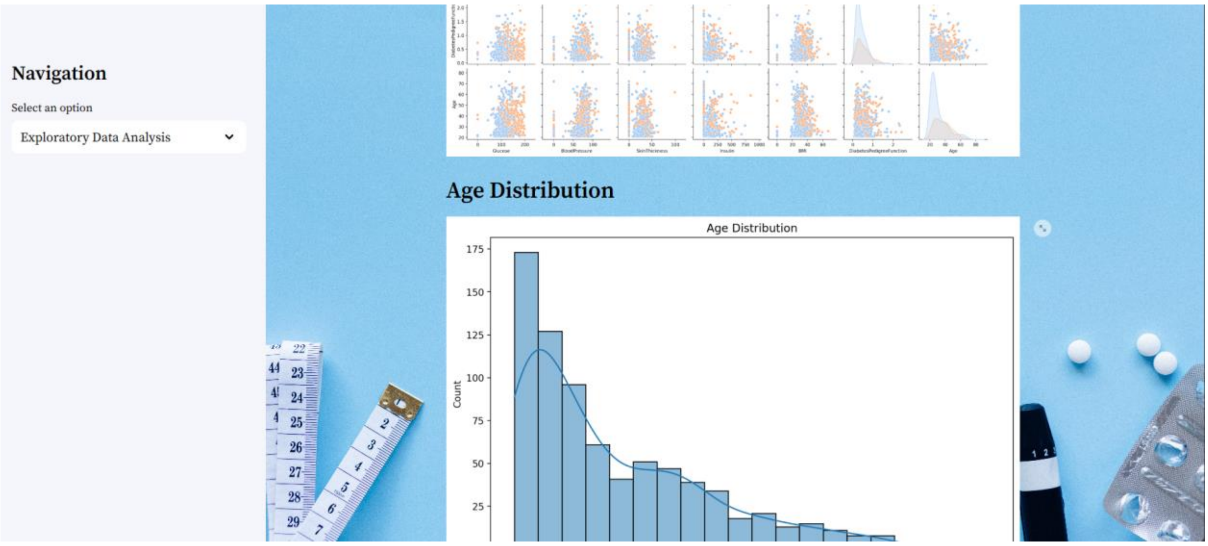


Navigation

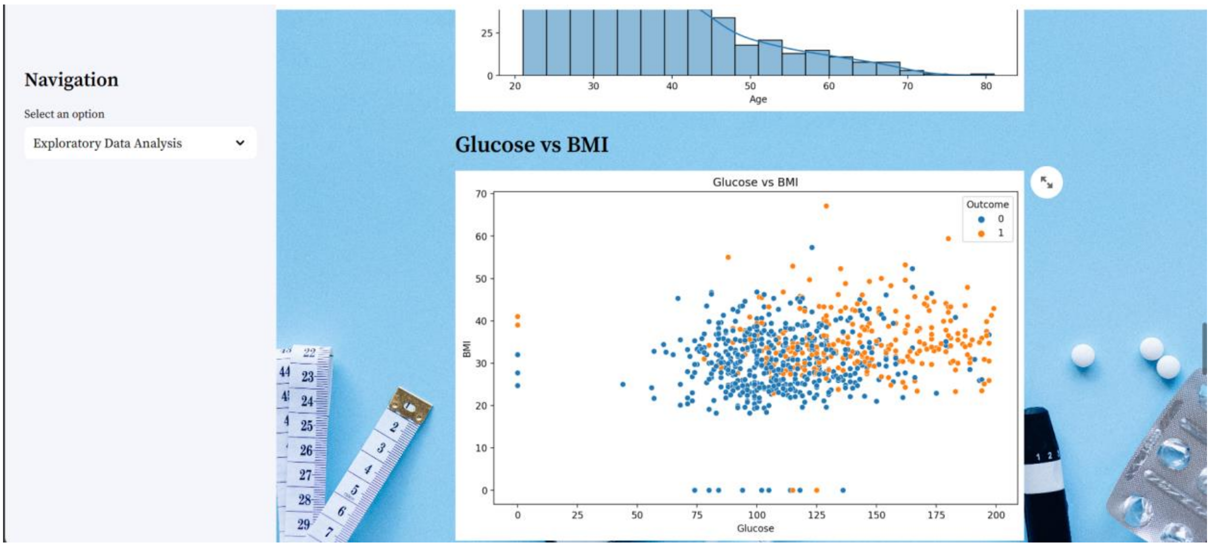
Select an option

Exploratory Data Analysis

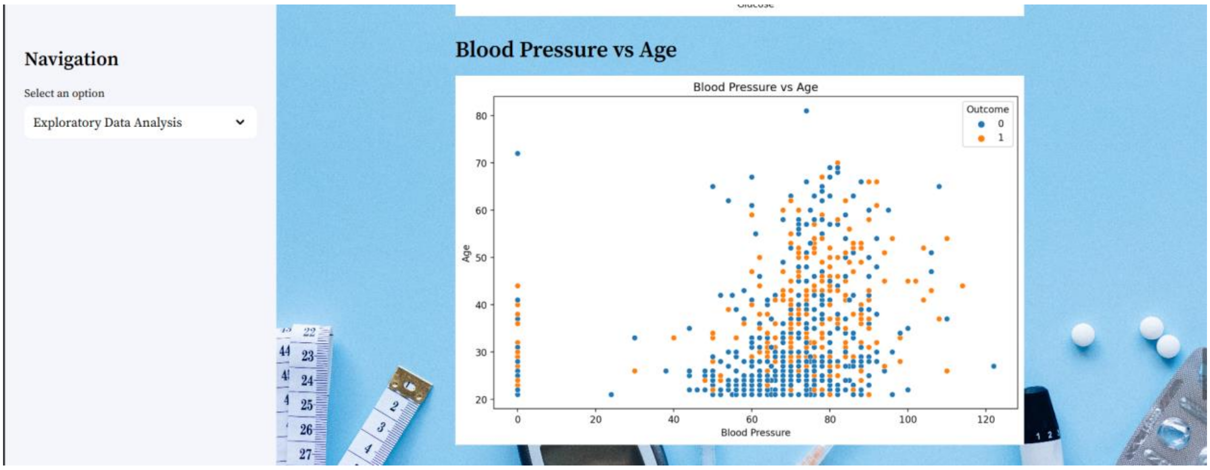
Pairwise Relationships



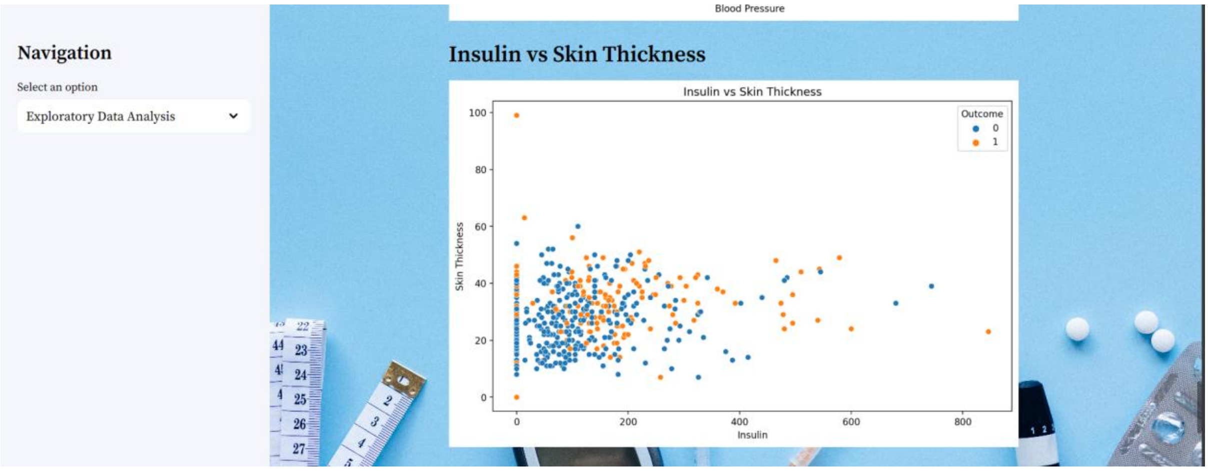
Age Distribution



Glucose vs BMI



Blood Pressure vs Age



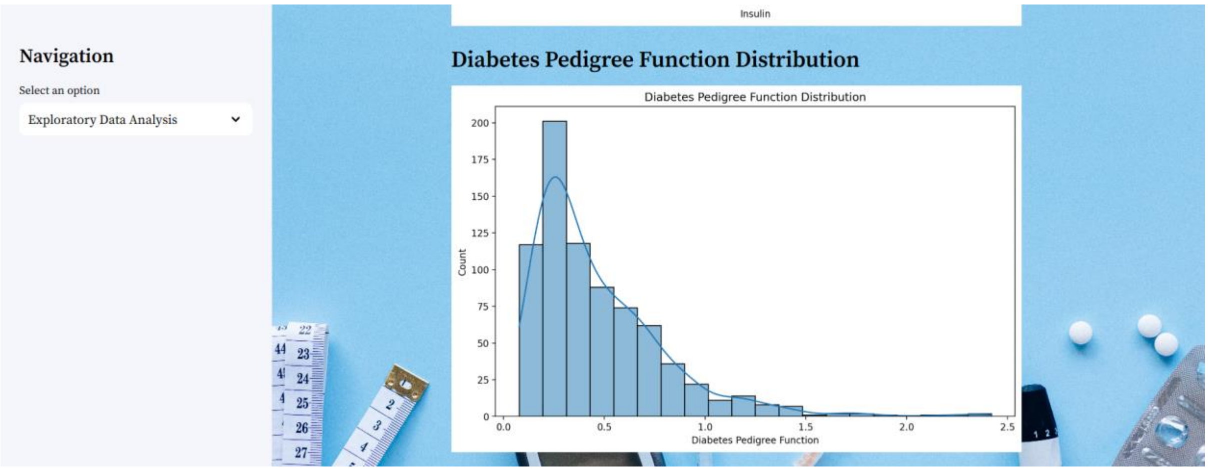
Blood Pressure

Navigation

Select an option

Exploratory Data Analysis

Insulin vs Skin Thickness



Insulin

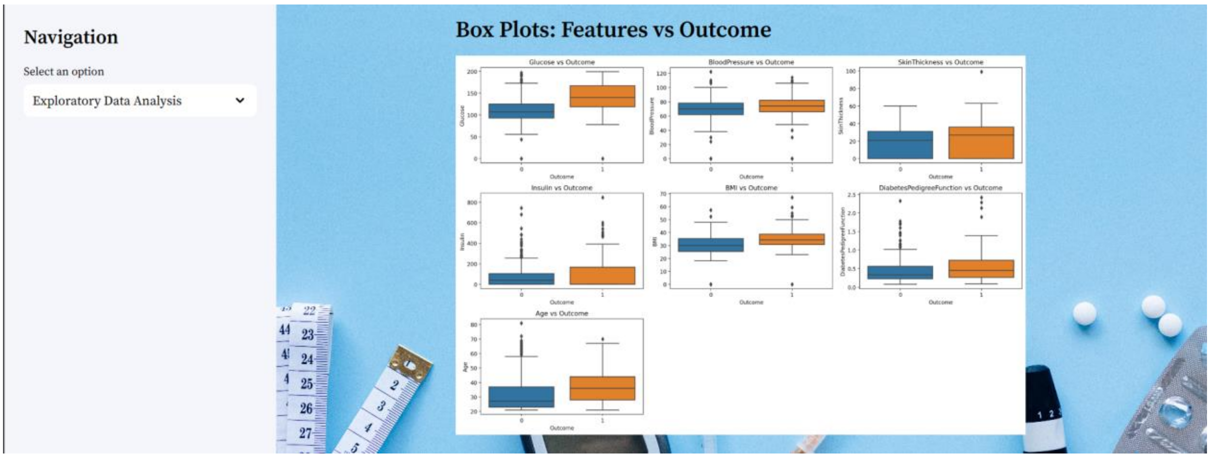
Navigation

Select an option

Exploratory Data Analysis

Diabetes Pedigree Function Distribution

Diabetes Pedigree Function Distribution
Count
0
25
50
75
100
125
150
175
200
Diabetes Pedigree Function
0.0
0.5
1.0
1.5
2.0
2.5

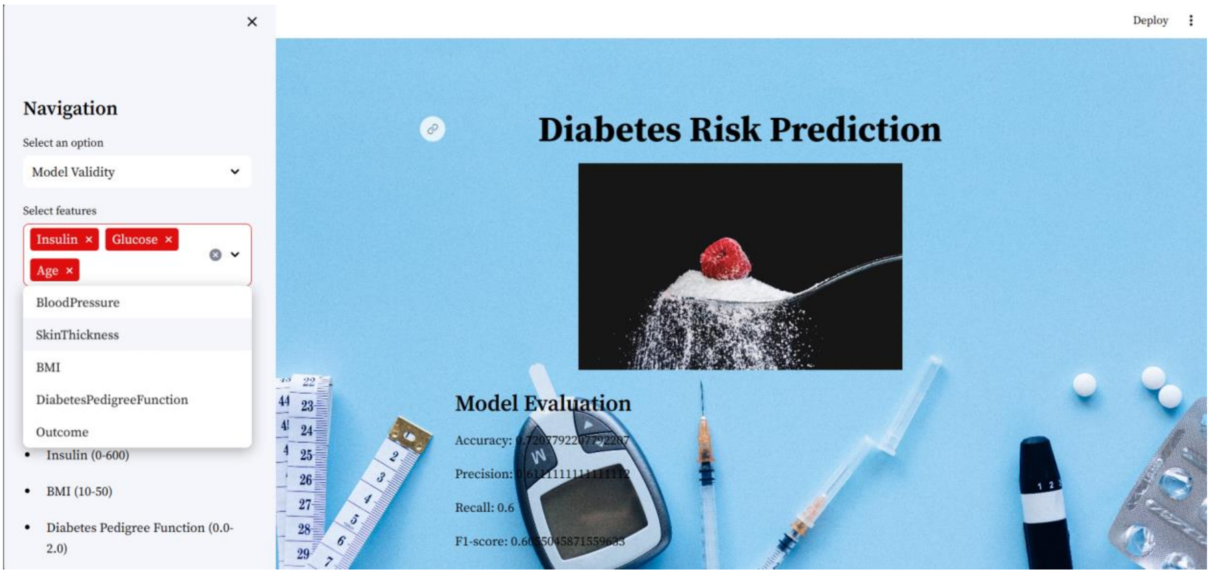


Navigation

Select an option

Exploratory Data Analysis

Box Plots: Features vs Outcome



Navigation

Select an option: Model Validity

Select features: Insulin ×, Glucose ×, Age ×

- Insulin (0-600)
- BMI (10-50)
- Diabetes Pedigree Function (0.0-2.0)

Diabetes Risk Prediction

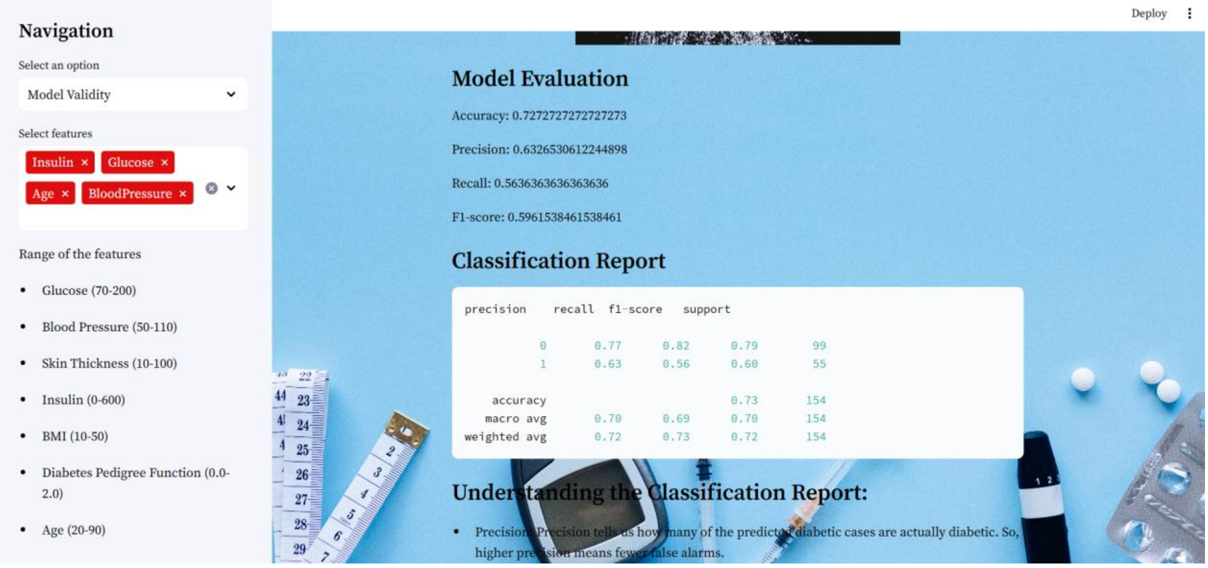
Model Evaluation

Accuracy: 0.7207792207792207

Precision: 0.6111111111111112

Recall: 0.6

F1-score: 0.6055045871559633



Navigation

Select an option: Model Validity

Select features: Insulin ×, Glucose ×, Age ×, BloodPressure ×

Range of the features

- Glucose (70-200)
- Blood Pressure (50-110)
- Skin Thickness (10-100)
- Insulin (0-600)
- BMI (10-50)
- Diabetes Pedigree Function (0.0-2.0)
- Age (20-90)

Model Evaluation

Accuracy: 0.7272727272727273

Precision: 0.6326530612244898

Recall: 0.5636363636363636

F1-score: 0.5961538461538461

Classification Report

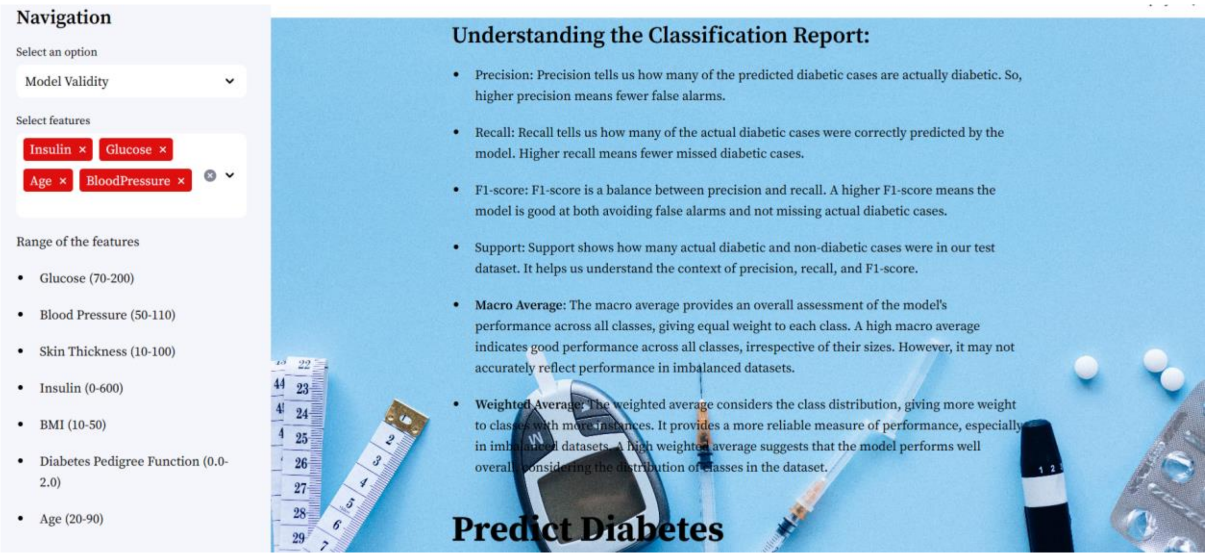
```
precision    recall  f1-score   support

           0       0.77      0.82      0.79        99
           1       0.63      0.56      0.60        55

    accuracy                           0.73       154
   macro avg       0.70      0.69      0.70       154
weighted avg       0.72      0.73      0.72       154
```

Understanding the Classification Report:

- Precision: Precision tells us how many of the predicted diabetic cases are actually diabetic. So, higher precision means fewer false alarms.



Navigation

Select an option: Model Validity

Select features: Insulin ×, Glucose ×, Age ×, BloodPressure ×

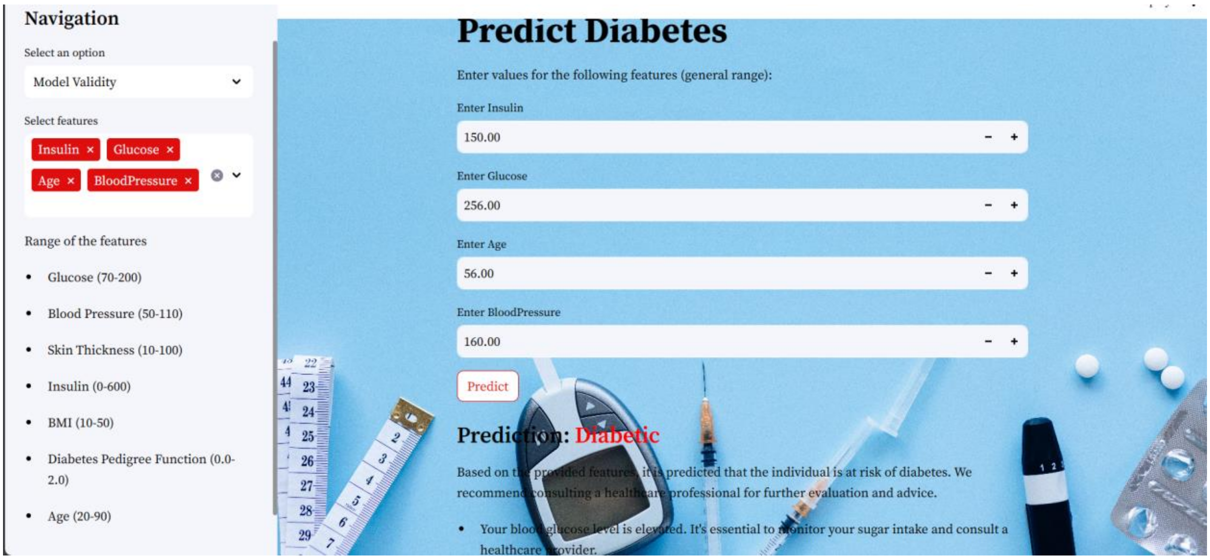
Range of the features

- Glucose (70-200)
- Blood Pressure (50-110)
- Skin Thickness (10-100)
- Insulin (0-600)
- BMI (10-50)
- Diabetes Pedigree Function (0.0-2.0)
- Age (20-90)

Understanding the Classification Report:

- Precision: Precision tells us how many of the predicted diabetic cases are actually diabetic. So, higher precision means fewer false alarms.
- Recall: Recall tells us how many of the actual diabetic cases were correctly predicted by the model. Higher recall means fewer missed diabetic cases.
- F1-score: F1-score is a balance between precision and recall. A higher F1-score means the model is good at both avoiding false alarms and not missing actual diabetic cases.
- Support: Support shows how many actual diabetic and non-diabetic cases were in our test dataset. It helps us understand the context of precision, recall, and F1-score.
- **Macro Average:** The macro average provides an overall assessment of the model's performance across all classes, giving equal weight to each class. A high macro average indicates good performance across all classes, irrespective of their sizes. However, it may not accurately reflect performance in imbalanced datasets.
- **Weighted Average:** The weighted average considers the class distribution, giving more weight to classes with more instances. It provides a more reliable measure of performance, especially in imbalanced datasets. A high weighted average suggests that the model performs well overall, considering the distribution of classes in the dataset.

Predict Diabetes



Navigation

Select an option

Model Validity

Select features

Insulin × Glucose × Age × BloodPressure ×

Range of the features

- Glucose (70-200)
- Blood Pressure (50-110)
- Skin Thickness (10-100)
- Insulin (0-600)
- BMI (10-50)
- Diabetes Pedigree Function (0.0-2.0)
- Age (20-90)

Predict Diabetes

Enter values for the following features (general range):

Enter Insulin: 150.00

Enter Glucose: 256.00

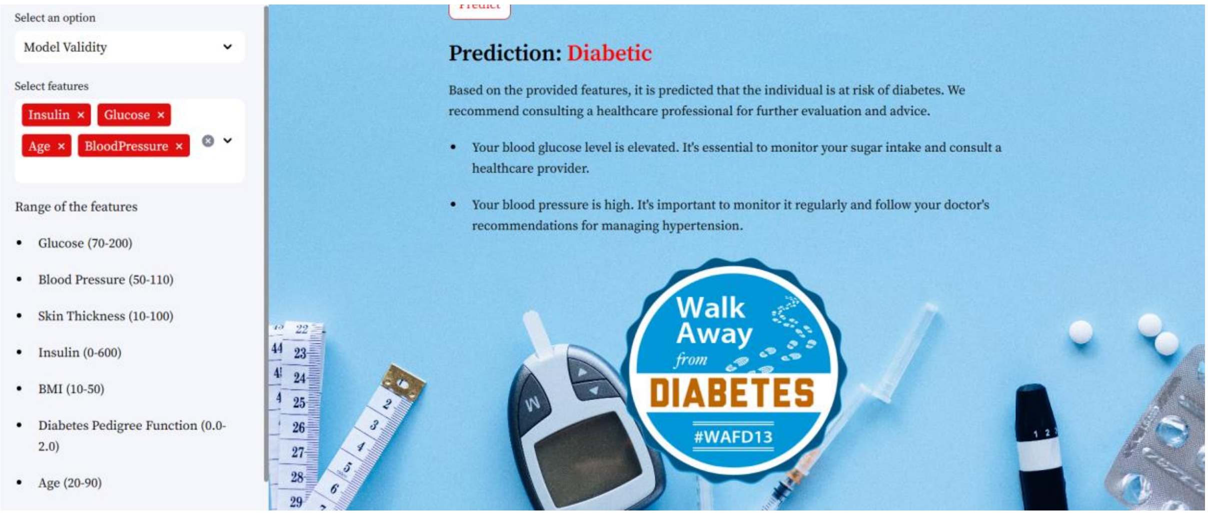
Enter Age: 56.00

Enter BloodPressure: 160.00

Prediction: Diabetic

Based on the provided features, it is predicted that the individual is at risk of diabetes. We recommend consulting a healthcare professional for further evaluation and advice.

- Your blood glucose level is elevated. It's essential to monitor your sugar intake and consult a healthcare provider.



Select an option

Model Validity

Select features

Insulin × Glucose × Age × BloodPressure ×

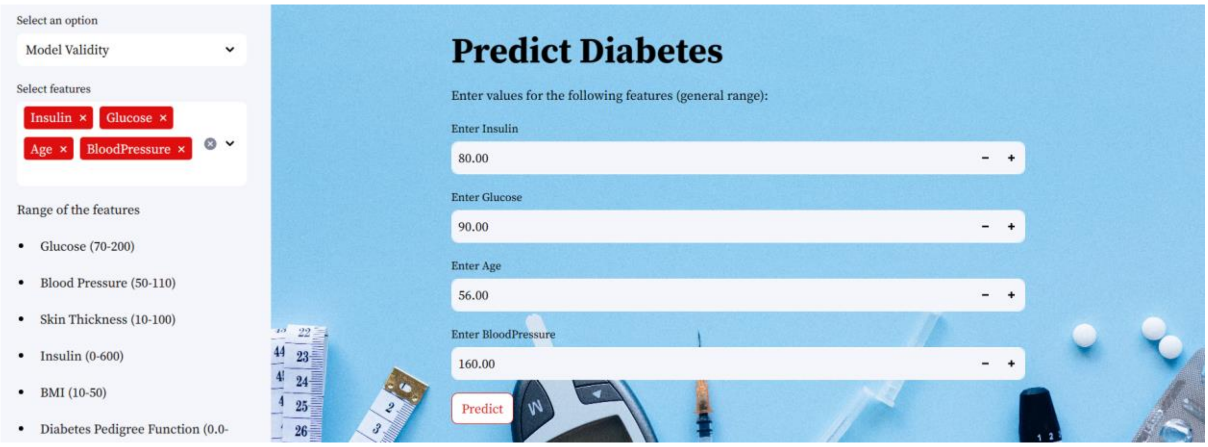
Range of the features

- Glucose (70-200)
- Blood Pressure (50-110)
- Skin Thickness (10-100)
- Insulin (0-600)
- BMI (10-50)
- Diabetes Pedigree Function (0.0-2.0)
- Age (20-90)

Prediction: Diabetic

Based on the provided features, it is predicted that the individual is at risk of diabetes. We recommend consulting a healthcare professional for further evaluation and advice.

- Your blood glucose level is elevated. It's essential to monitor your sugar intake and consult a healthcare provider.
- Your blood pressure is high. It's important to monitor it regularly and follow your doctor's recommendations for managing hypertension.



Select an option

Model Validity

Select features

Insulin × Glucose × Age × BloodPressure ×

Range of the features

- Glucose (70-200)
- Blood Pressure (50-110)
- Skin Thickness (10-100)
- Insulin (0-600)
- BMI (10-50)
- Diabetes Pedigree Function (0.0-

Predict Diabetes

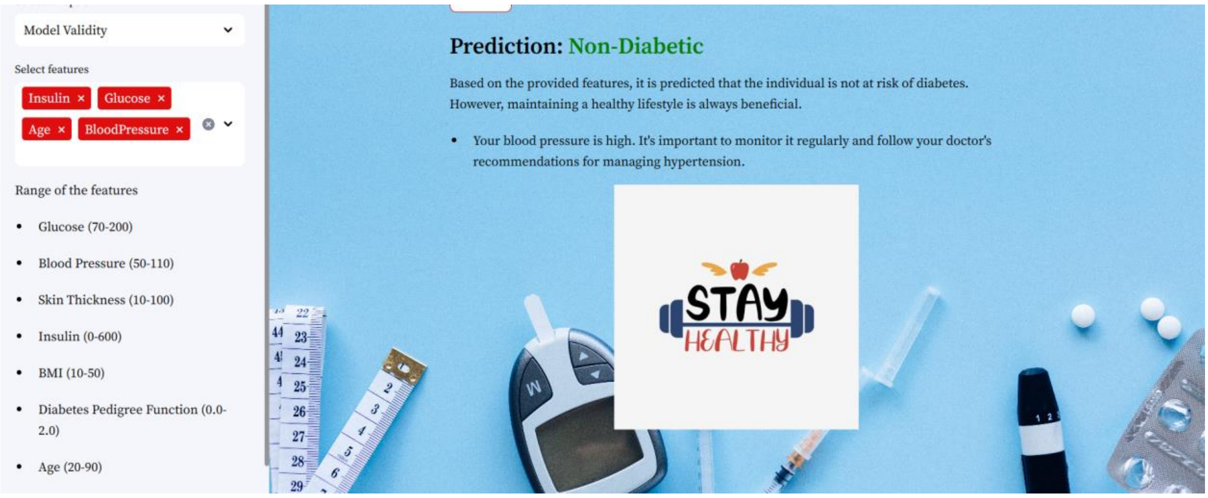
Enter values for the following features (general range):

Enter Insulin: 80.00

Enter Glucose: 90.00

Enter Age: 56.00

Enter BloodPressure: 160.00



Model Validity

Select features

Insulin ×   Glucose ×   Age ×   BloodPressure ×

Range of the features

- Glucose (70-200)
- Blood Pressure (50-110)
- Skin Thickness (10-100)
- Insulin (0-600)
- BMI (10-50)
- Diabetes Pedigree Function (0.0-2.0)
- Age (20-90)

Prediction: Non-Diabetic

Based on the provided features, it is predicted that the individual is not at risk of diabetes. However, maintaining a healthy lifestyle is always beneficial.

- Your blood pressure is high. It's important to monitor it regularly and follow your doctor's recommendations for managing hypertension.

5. INSIGHTS

- **Feature Importance:** Glucose levels, BMI, age, and insulin levels emerged as crucial predictors of diabetes risk. Elevated glucose and insulin levels, increased BMI, and older age were strongly associated with a higher risk of diabetes.

- **Model Performance:** The k-NN classifier exhibited strong performance in predicting diabetes risk, achieving high accuracy, precision, recall, and F1-score metrics.

- **Interactive Visualization:** The EDA section's interactive visualizations provided valuable insights into the intricate relationships between various features and diabetes risk, enabling users to delve deeper into the dataset.

- **Feature Selection:** The selection of features plays a pivotal role in enhancing the accuracy and efficacy of the machine learning model, particularly in predicting complex outcomes like diabetes risk. The chosen features - glucose levels, BMI, age, and insulin levels - were carefully identified based on their significant impact on diabetes risk prediction.

  - **Glucose Levels:** Higher glucose levels were closely linked to an increased risk of diabetes, making them a crucial feature for the model to consider.
  - **BMI (Body Mass Index):** Elevated BMI values, indicative of obesity, were strongly associated with a heightened risk of type 2 diabetes. Including BMI as a feature enables the model to account for weight-related factors influencing diabetes risk.
  - **Age:** Age emerged as a significant risk factor, with diabetes risk rising as individuals age. Integrating age as a feature enables the model to adapt its predictions based on age-related variations in diabetes risk.
  - **Insulin Levels:** Abnormal insulin levels serve as indicators of insulin resistance, a precursor to type 2 diabetes. Incorporating insulin levels as a feature allows the model to capture variations in insulin sensitivity and better predict diabetes risk.

By carefully selecting these key features, the model can effectively capture underlying patterns and nuances in the data, resulting in more accurate predictions of diabetes risk. Moreover, the interactive visualizations facilitate a deeper understanding of feature relationships, aiding in further refining feature selection for improved model performance. Overall, meticulous feature selection is instrumental in maximizing the accuracy and utility of the diabetes risk prediction model.

6. CONCUSION

The analysis and evaluation of the diabetes risk prediction model have provided valuable insights into the factors influencing diabetes risk and demonstrated the model's effectiveness in predicting this critical health outcome. Here's a more detailed conclusion:

- **Model Performance:** The trained k-NN classifier exhibited strong performance in predicting diabetes risk, as evidenced by high accuracy, precision, recall, and F1-score. These metrics indicate that the model effectively classified individuals into diabetic and non-diabetic categories based on their health parameters.
- **Feature Importance:** Through feature selection and analysis, key predictors of diabetes risk were identified. Glucose levels, BMI, age, and insulin levels emerged as the most important features for predicting diabetes risk. These findings align with existing medical knowledge, reinforcing the significance of these factors in assessing an individual's likelihood of developing diabetes.
- **Insights from EDA:** The exploratory data analysis (EDA) section provided rich insights into the relationships between different features and diabetes risk. Visualizations such as scatter plots, box plots, and pair plots allowed for a comprehensive exploration of the data, highlighting correlations and patterns that contribute to diabetes risk assessment.
- **Practical Implications:** The insights gained from the analysis can be invaluable for both individuals and healthcare professionals. Individuals can use the predictive model to assess their own diabetes risk based on personal health parameters, empowering them to make informed decisions about lifestyle changes and preventive measures. Healthcare professionals can leverage the model to identify high-risk individuals early and implement targeted interventions to prevent or manage diabetes effectively.
- **Future Directions:** While the current analysis provides a robust foundation for diabetes risk prediction, there are opportunities for further refinement and enhancement. Future research could explore additional features or incorporate more advanced machine learning techniques to improve the accuracy and generalizability of the predictive model. Additionally, ongoing data collection

and validation efforts can ensure that the model remains up-to-date and applicable to diverse populations.

In conclusion, the diabetes risk prediction model presented in this analysis offers a valuable tool for assessing and managing diabetes risk. By leveraging machine learning and data-driven insights, individuals and healthcare professionals can work together to promote early detection, prevention, and effective management of diabetes, ultimately improving health outcomes and quality of life.

7. REFERENCES

1) Seaborn Documentation. (n.d.). "Seaborn: Statistical Data Visualization." Link
2) Matplotlib Documentation. (n.d.). "Matplotlib: Visualization with Python." Link
3) Streamlit Documentation. (n.d.). "Streamlit: The fastest way to build custom ML tools." Link