

DIABETES DETECTION

Leveraging Machine Learning for
Early Risk Assessment



<https://depiproject-a7txbipsjwvaawrmsftjef.streamlit.app/>

AGENDA

- Objectives
- Dataset
- Imbalancement
- Solution
- Models used
- Models
- Ethical Considerations
- Technology Stack
- Future Direction



<https://depiproject-a7txbipsjwvaawrmsftjef.streamlit.app/>



PROJECT OVERVIEW



- Developing a machine learning model to be used in the early detection of diabetes
- The system analyzes patient data to identify health trends and predict potential diabetes risks
- Aiming to enhance patient outcomes and optimize healthcare resource management.

<https://depiproject-a7txbipsjwvaawrmsftjef.streamlit.app/>



OBJECTIVES

- Develop a data-driven predictive model for accurate patient diabetes risk assessment.
- Identify trends in health metrics to support early intervention and prevention strategies.
- Assist healthcare professionals and individuals in making informed decisions through accessible, real-time insights.
- Improve operational efficiency in healthcare settings through the application of predictive analytics.

<https://depiproject-a7txbipsjwvaawrmsftjef.streamlit.app/>

STAKEHOLDERS

- Patients
- Healthcare providers
- Healthcare organizations

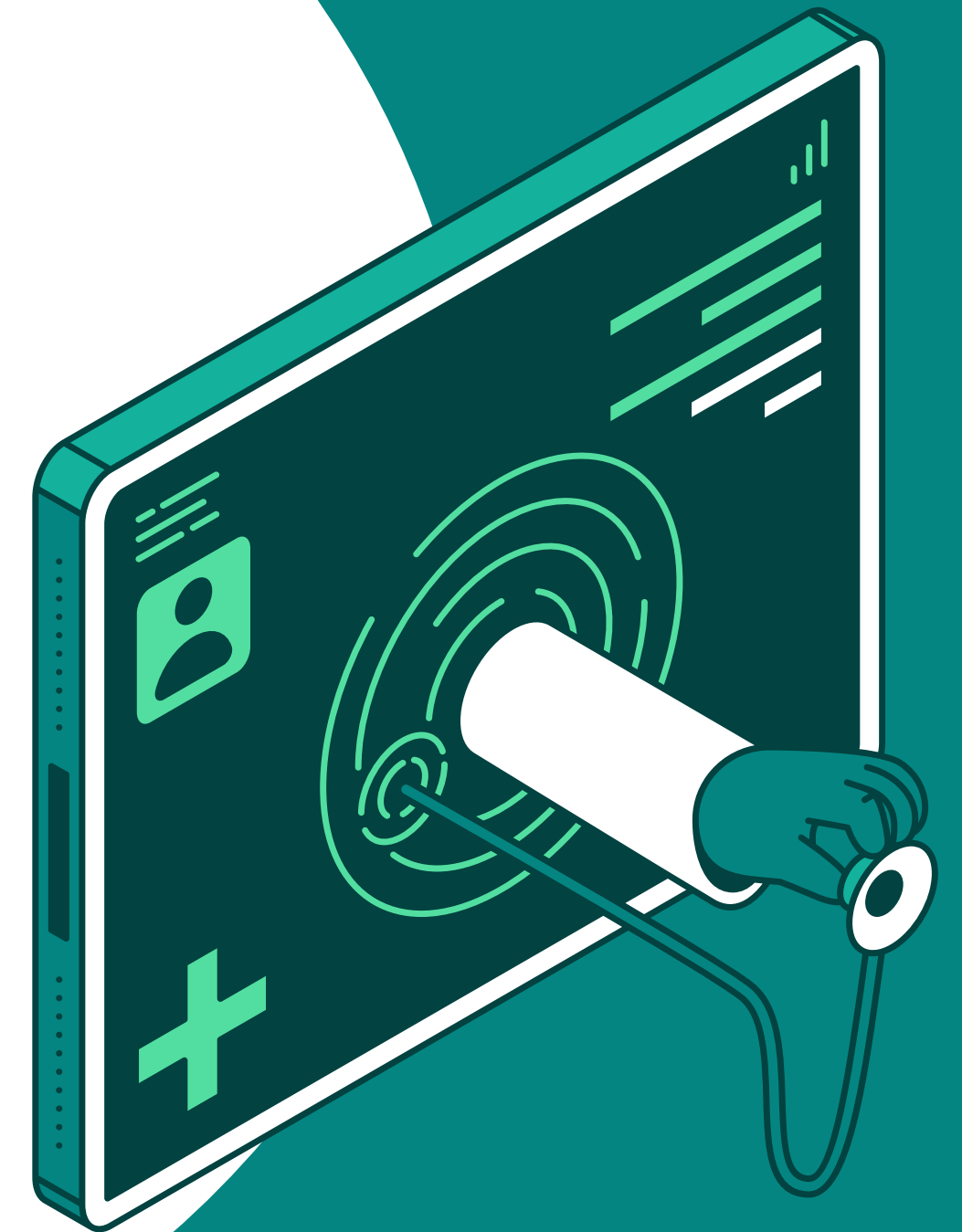


<https://depiproject-a7txbipsjwvaawrmsftjef.streamlit.app/>

DATA

Utilized a public available dataset from kaggle: <https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset/data> which includes the following features:

- diabetes_binary
- HighBP
- HighChol
- CholCheck
- BMI
- Smoker
- Stroke
- PhysActivity
- Fruits
- Veggies
- HvyAlcoholConsump
- AnyHealthcare
- NoDocbcCost
- Sex
- Age
- Education
- Income



<https://depiproject-a7txbipsjwvaawrmsftjef.streamlit.app/>



DATA PREPROCESSING

- **Initial Exploration:** Checked for missing values, data types, and basic statistics.
- **Data Cleaning:** Addressed inconsistencies, such as zero values in physiological measurements
- **Feature Scaling:** Applied techniques like StandardScaler to normalize feature ranges, ensuring fair contribution from all features to the model.
- **Train-Test Split:** Divided the data into training and testing sets to evaluate model performance on unseen data.

DATASET IMBALANCE

Problem

- A significant challenge encountered was the imbalanced nature of the diabetes dataset.

Initial Attempts

- Class Weighting
- Neural Networks

```
# 7. Evaluate
y_pred = (model.predict(X_test_scaled) > 0.5).astype(int)
print("\nClassification Report:")
print(classification_report(y_test, y_pred))
print("\nConfusion Matrix:")
print(confusion_matrix(y_test, y_pred))
```

648/648 ————— 1s 776us/step

Classification Report:

	precision	recall	f1-score	support
0	0.70	0.70	0.70	10188
1	0.71	0.71	0.71	10530
accuracy			0.71	20718
macro avg	0.71	0.71	0.71	20718
weighted avg	0.71	0.71	0.71	20718

Confusion Matrix:

```
[[7140 3048]
 [3048 7482]]
```

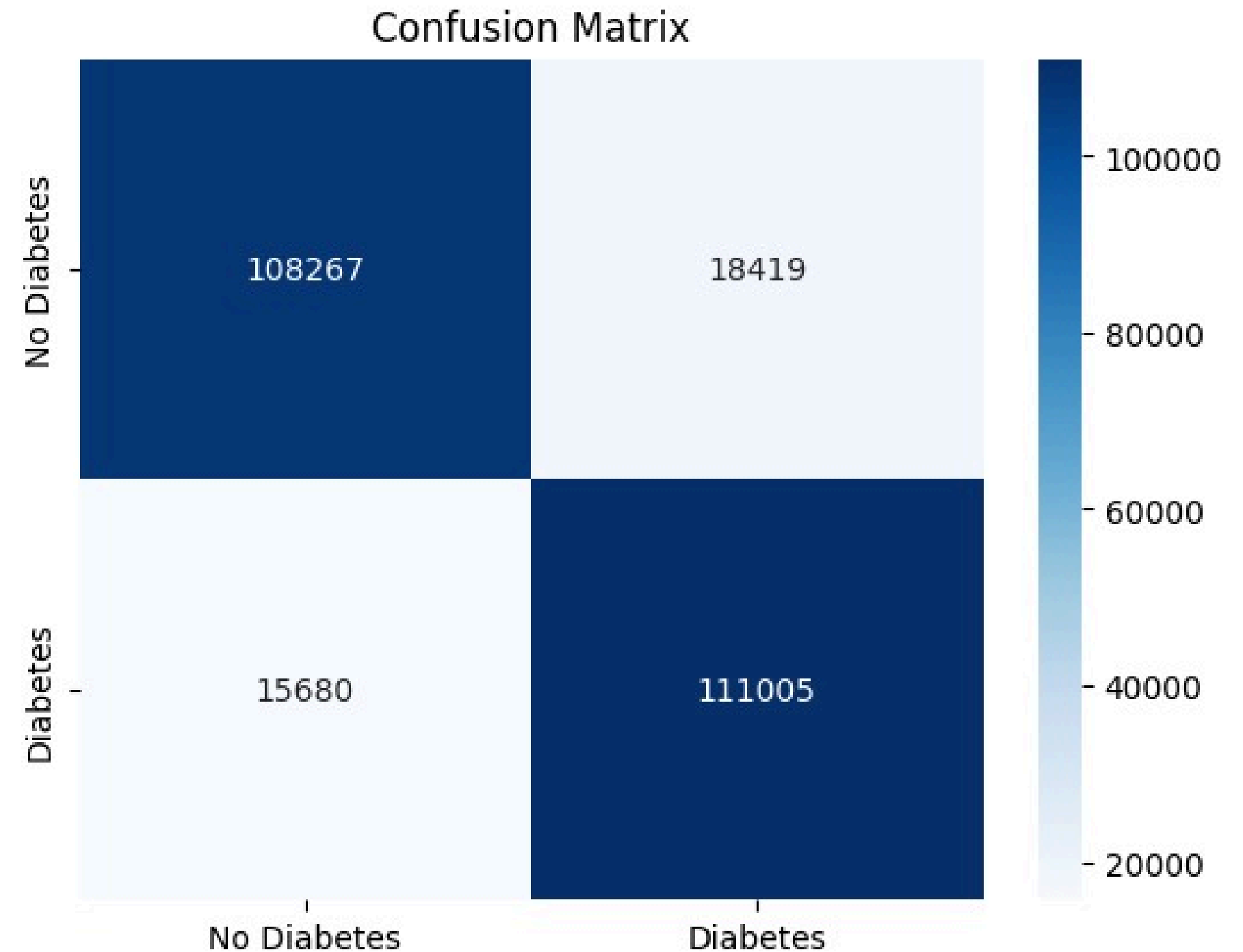

SOLUTION

SMOTE Algorithm

- SMOTE is an oversampling technique that creates synthetic samples of the minority class (diabetic patients).

Impact

- The application of SMOTE led to a more balanced dataset for training the model.



<https://depiproject-a7txbipsjwvaawrmsftjef.streamlit.app/>

MODELS USED



- **At first, Random Forest Algorithm was used but due to poor accuracy achieved**
- **Random Forest Accuracy: 70%**
- **XGB Model Accuracy: 84%**

TECHNOLOGY STACK



Data Analysis

- Pandas
- NumPy

Data Visualization

- matplotlib
- seaborn

ML Model

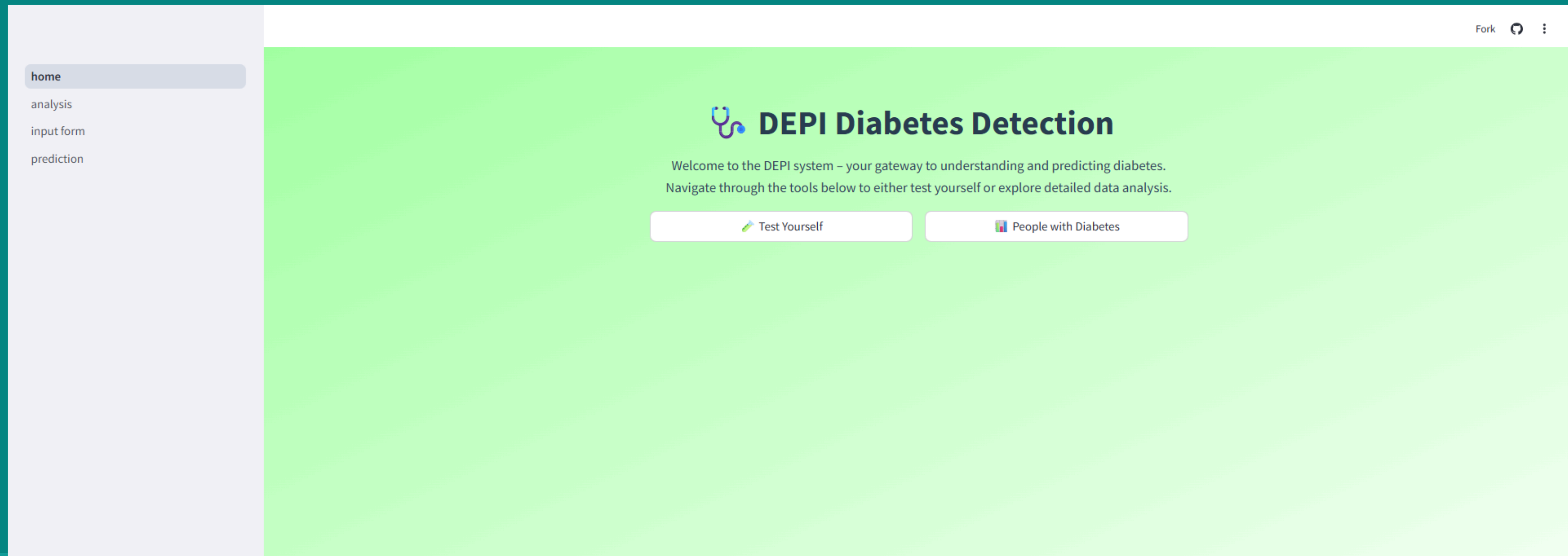
- sckicit-learn
- XGBboost
- SMOTE

Web App

- Streamlit

<https://depiproject-a7txbipsjwvaawrmsftjef.streamlit.app/>

UI- HOME PAGE



<https://depiproject-a7txbipsjwvaawrmsftjef.streamlit.app/>

UI- INPUT FORM



[home](#)
[analysis](#)
[input form](#)
[prediction](#)

Model loaded successfully!

Diabetes Risk Assessment Tool

Please fill in your health information below to assess your diabetes risk. All fields are required.

Personal Information

Age Group

50-54

Sex

☒ Female
☐ Male

Education Level

High school graduate

Income Level

\$25,000-\$35,000

BMI (Body Mass Index)

25.00

General Health

Excellent Good Poor

Number of days with physical health issues in past 30 days

0 30

Number of days with mental health issues in past 30 days

0 30

Health Conditions

☐ High blood pressure

☐ High cholesterol

☒ Had cholesterol check in past 5 years

☐ Smoked at least 100 cigarettes in life

☐ Ever had a stroke

☐ Heart disease or heart attack

☐ Difficulty walking or climbing stairs

☒ Physical activity in past 30 days

☐ Consume fruit daily

☐ Consume vegetables daily

Healthcare Access

☐ Heavy alcohol consumption (adult men >14 drinks/week, women >7)

☒ Have any healthcare coverage

☐ Couldn't see doctor in past year due to cost

Assess Diabetes Risk

<https://depiproject-a7txbipsjwvaawrmsftjef.streamlit.app/>

UI- PREDICTION



Diabetes Risk Assessment Results

Your diabetes risk is: LOW

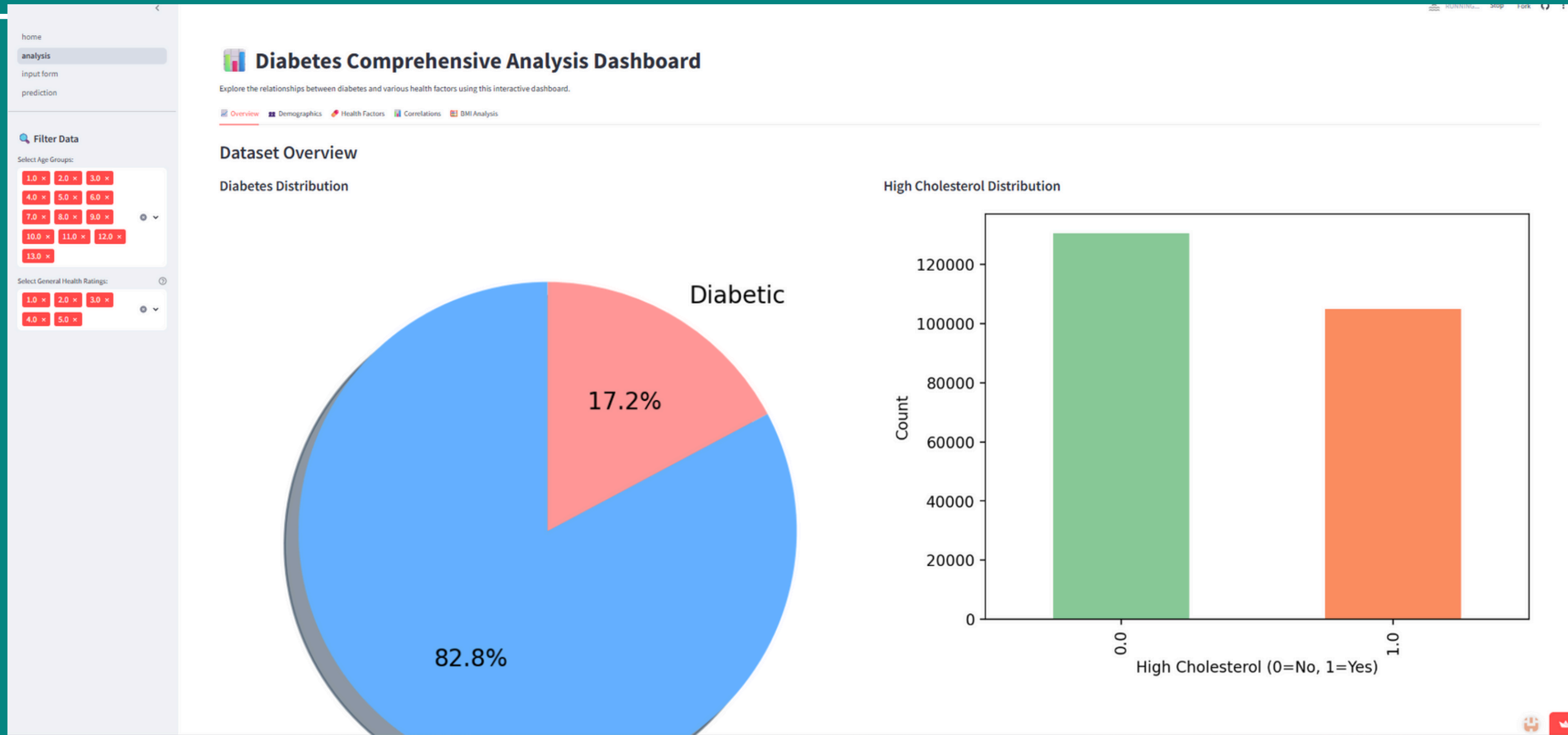
Confidence: 92.4%

- Maintain healthy habits
- Get regular check-ups
- Continue monitoring

Perform New Assessment

<https://depiproject-a7txbipsjwvaawrmsftjef.streamlit.app/>

UI- ANALYSIS



<https://depiproject-a7txbipsjwvaawrmsftjef.streamlit.app/>

THANK YOU

QUESTIONS?

