SugarSense

PREDICTING DIABETES RISK

Introduction

The project aims to develop a predictive model for assessing diabetes risk based on user input data.

Leveraging machine learning techniques and a dataset containing relevant health information.

Providing an interactive tool for understanding individual diabetes risk profiles.

Key factors analyzed include BMI, general health status, high blood pressure, high cholesterol levels, and age.

Empowering individuals with personalized risk assessments to promote early intervention and informed decision-making.

Contributing to preventive healthcare and efforts to combat the global diabetes epidemic.

Data Set

- Dataset Overview:
  - Source: Kaggle dataset titled “Diabetes Health Indicators Dataset”
  - Size: 6.2 MB
  - Shape: (70692, 22)
  - Period: 2015
  - Target variable: Diabetes_binary
- Preprocessing Steps:
  - Handling missing/duplicate values
  - Encoding categorical variables
  - Data standardization
  - Feature selection based on relevance to diabetes risk prediction

Exploratory Data Analysis

- Histogram of Features:

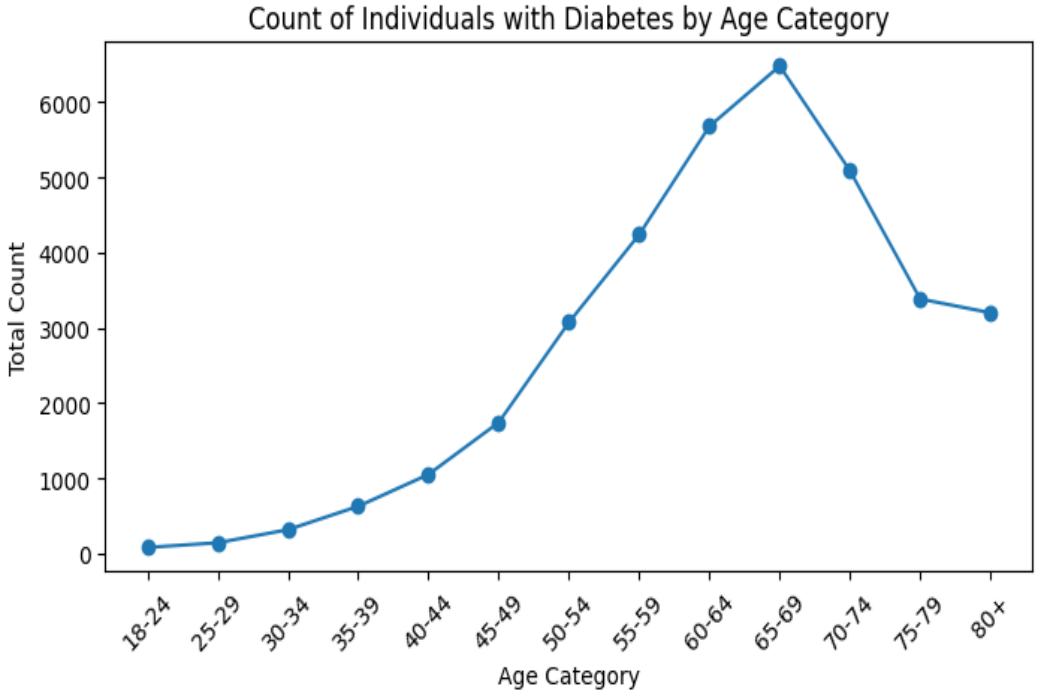
  Visualizes the distribution of each feature, aiding in identifying patterns, outliers, and distribution.

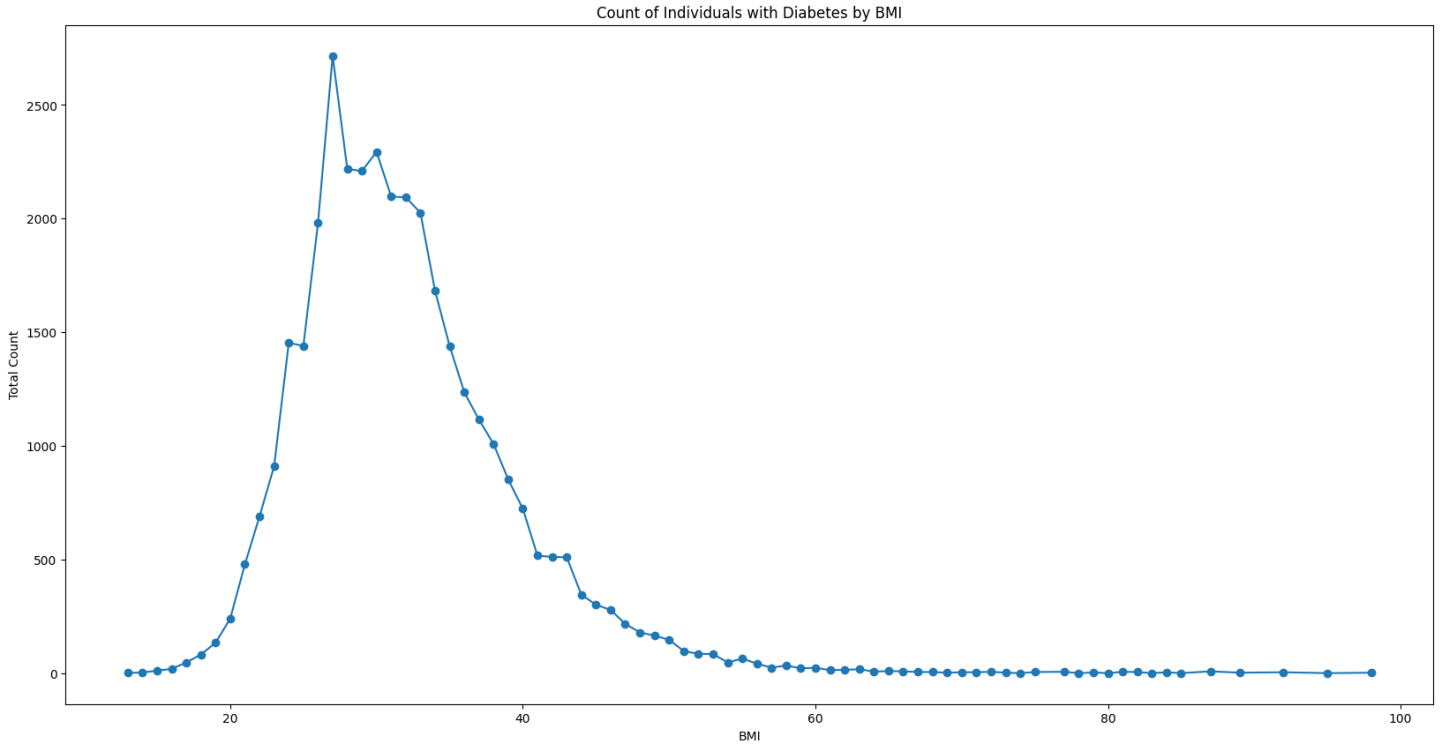
- Age and BMI Analysis:

  Illustrates the relationship between age, BMI, and diabetes prevalence, showing increased likelihood with advancing age and higher BMI values.

- Effect of High Blood Pressure and High Cholesterol:

  Examines the impact of these factors on diabetes risk, suggesting a notable association between their presence and increased likelihood of diabetes.





Exploratory Data Analysis

Exploratory Data Analysis

- General Health and Diabetes:

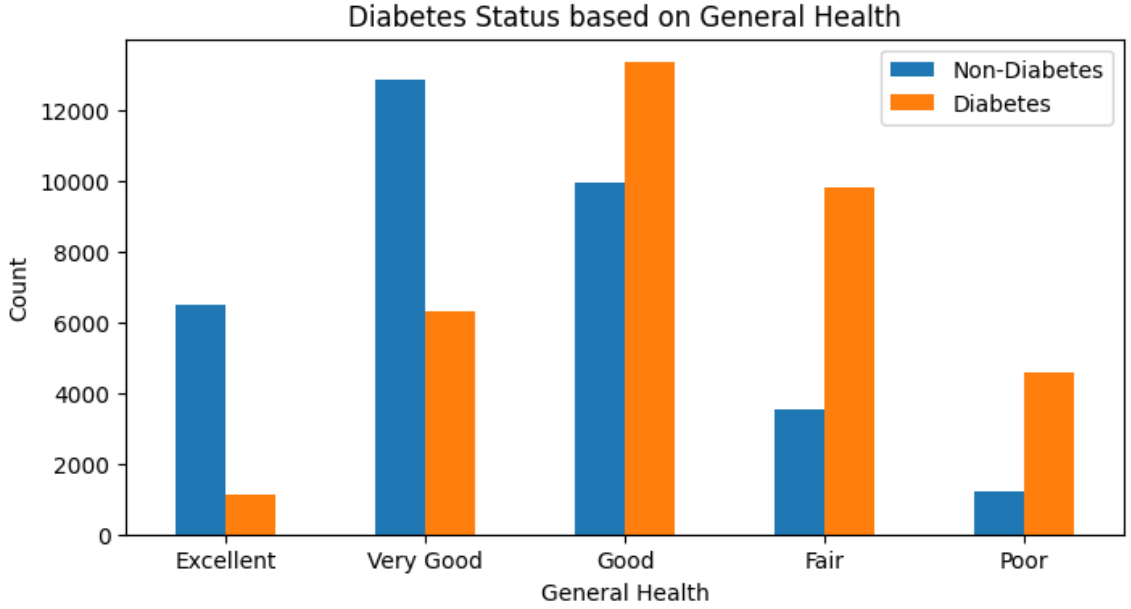
  Highlights the correlation between declining general health and higher diabetes prevalence.

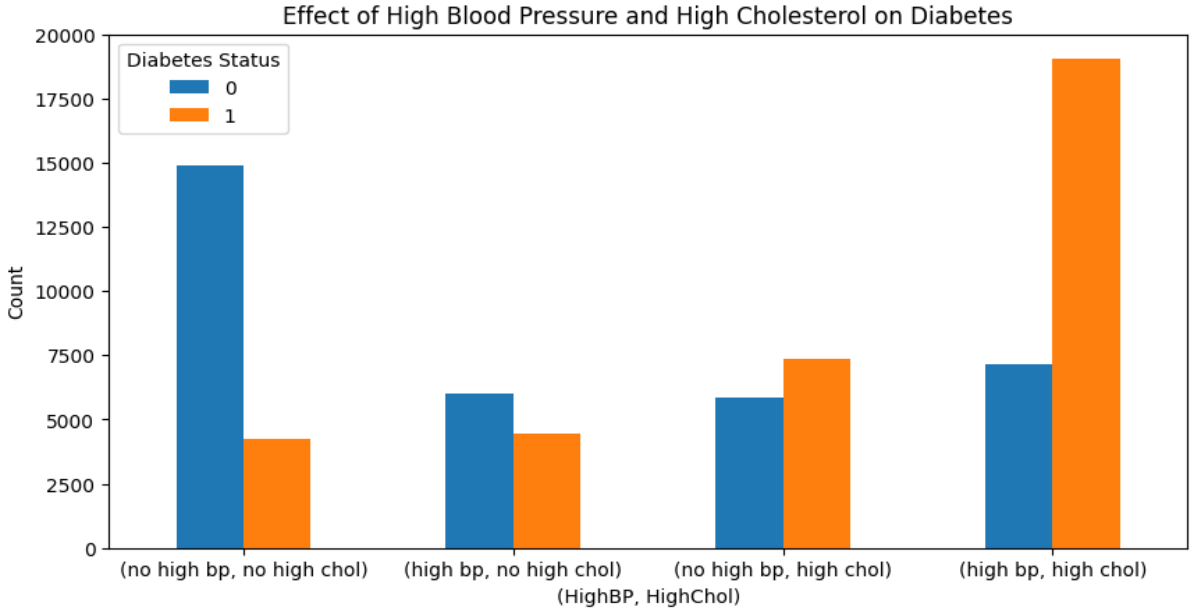
- Stroke and Heart Disease Influence:

  Indicates a potential correlation between stroke, heart disease, and diabetes.

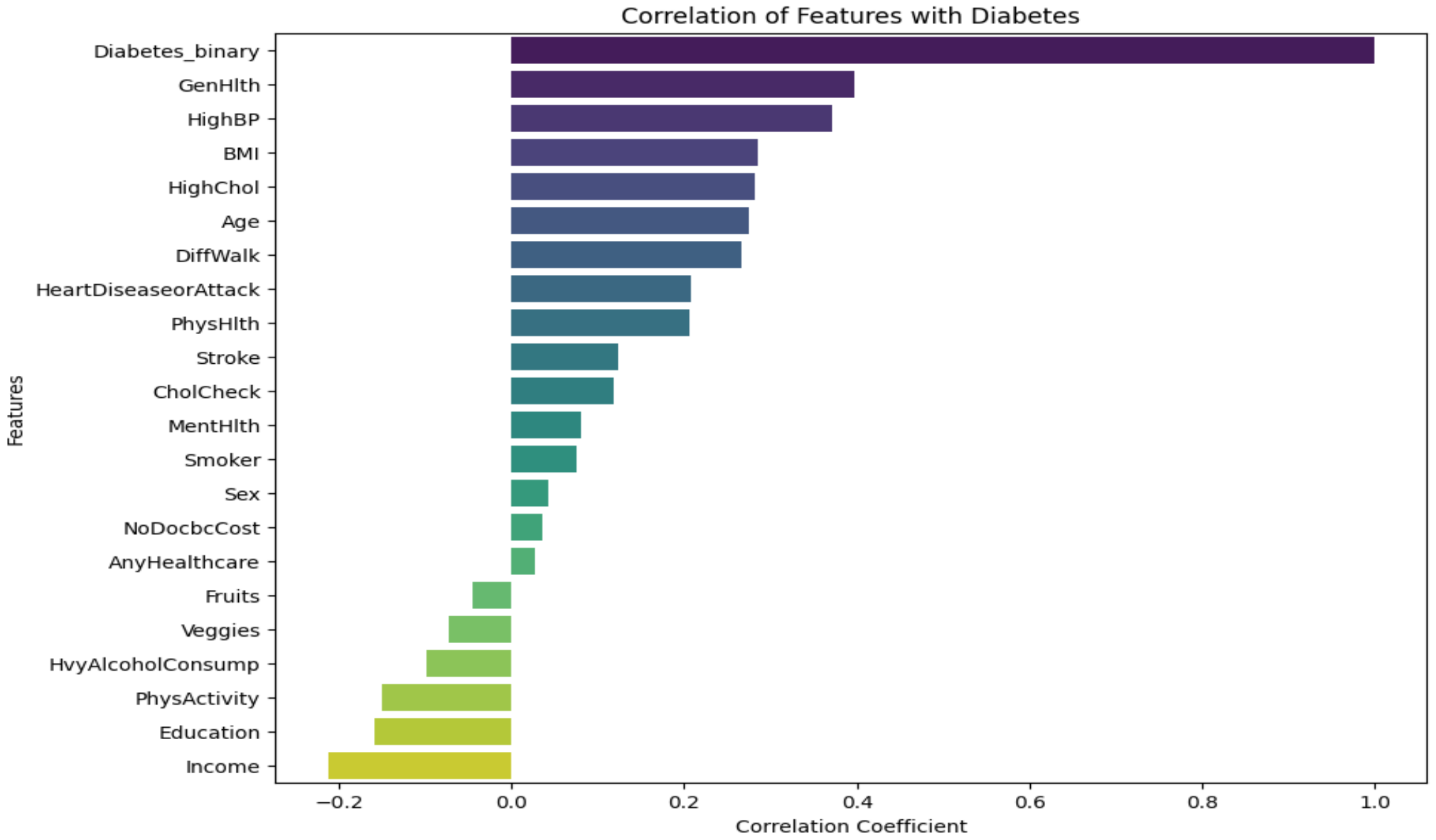
- Correlation Analysis:

  The correlation heatmap visualizes the strength and direction of correlations between features, aiding in identifying relationships and potential predictors of diabetes.





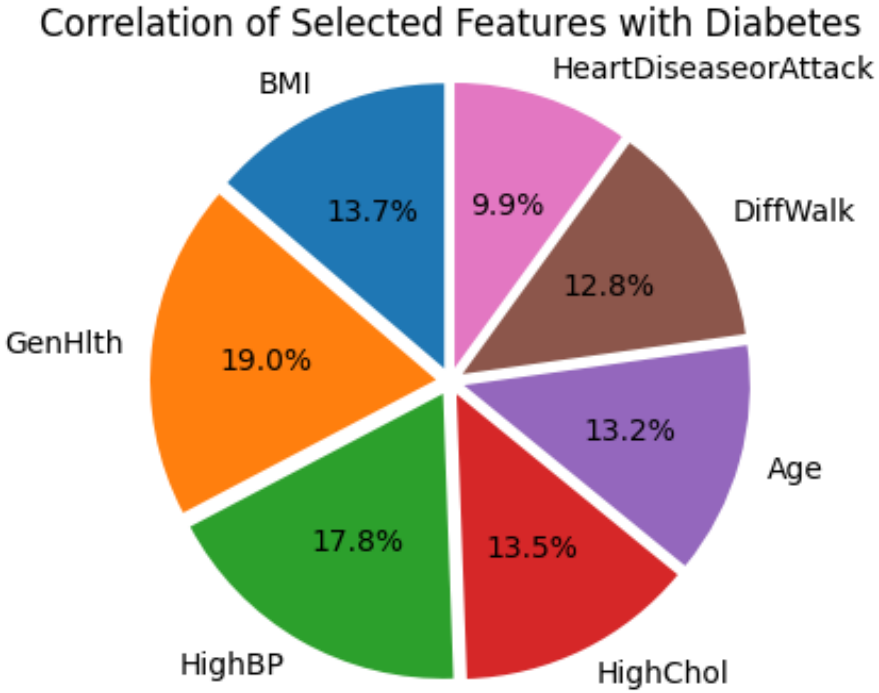
Exploratory Data Analysis

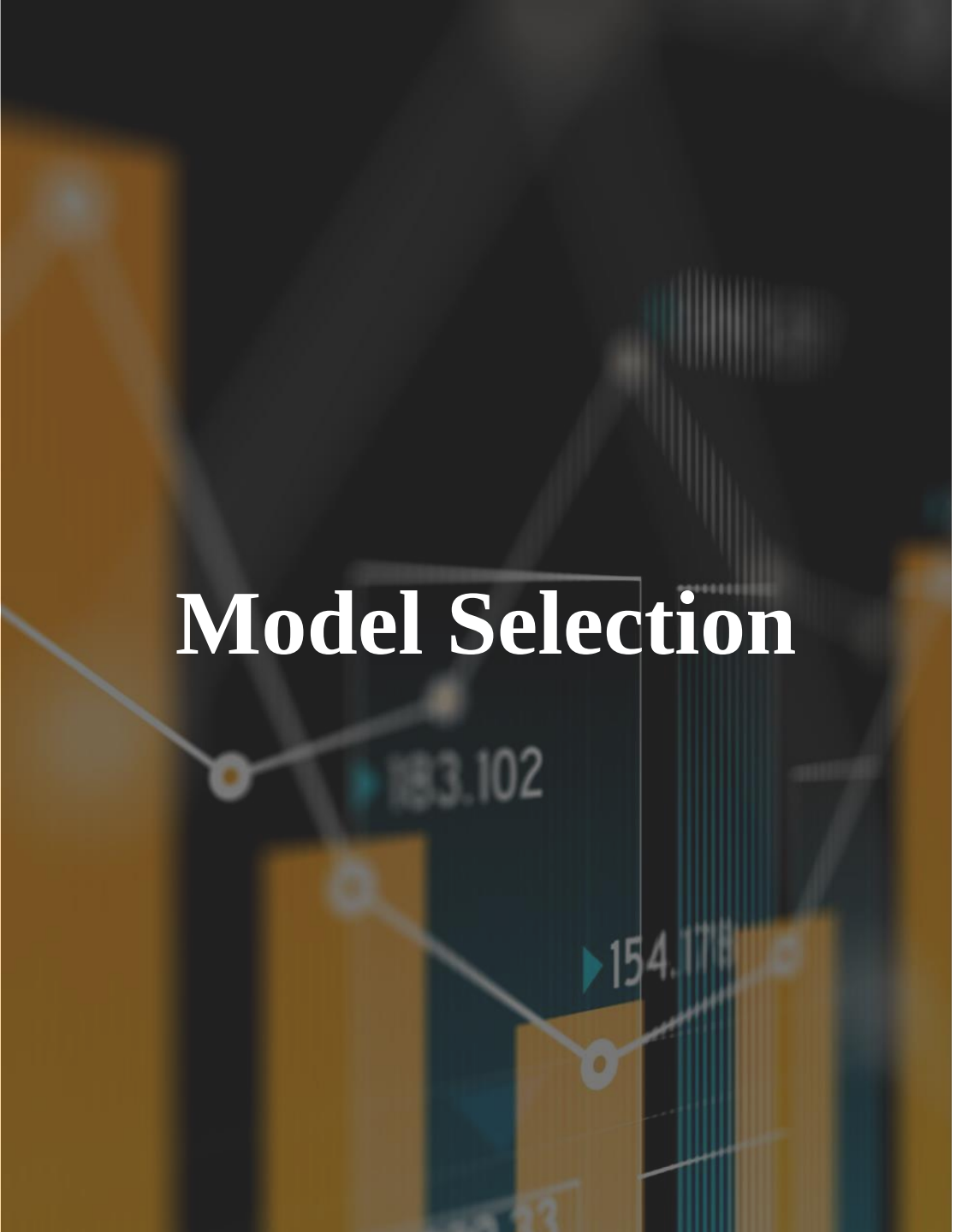


Exploratory Data Analysis

Feature Analysis

- Selected features for analysis:
  - BMI
  - General Health
  - High Blood Pressure
  - High Cholesterol
  - Age
  - Difficulty Walking
  - Heart Disease or Attack



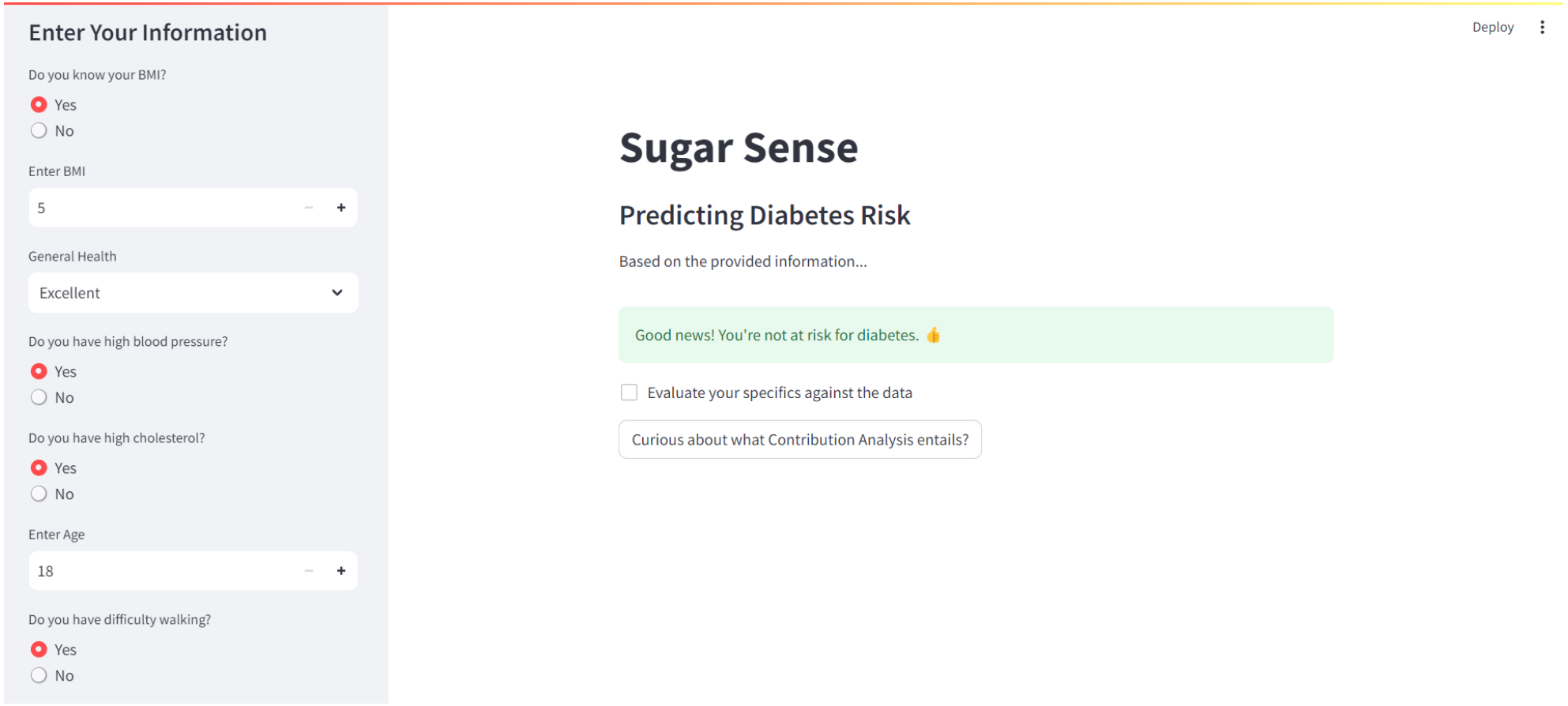


Model Selection

- Trained with multiple models such as random forest regression, support vector machine, logistic regression, decision trees, etc., and compared the accuracy.
- Logistic regression was selected because it performed the best among the models evaluated.
- The model is trained and tested with 80% of the data used for training and 20% for testing.
- The trained model was then evaluated on the testing set to assess its performance in predicting diabetes risk.

Streamlit

- Streamlit app utilizes logistic regression to predict diabetes risk based on user input data.
- Features such as BMI, general health, high blood pressure, high cholesterol, age, difficulty walking, and heart disease or attack are considered in the prediction.
- The app allows users to input their information through a sidebar interface.
- Live updates are provided, giving instant feedback on the predicted diabetes risk.
- Users can evaluate how their specifics compare to the dataset's distribution across various categories such as BMI, general health, high blood pressure, high cholesterol, and age.
- An option is provided to explain Contribution Analysis, showing the relative importance of each feature in shaping the prediction result.



Streamlit Interface

Conclusion

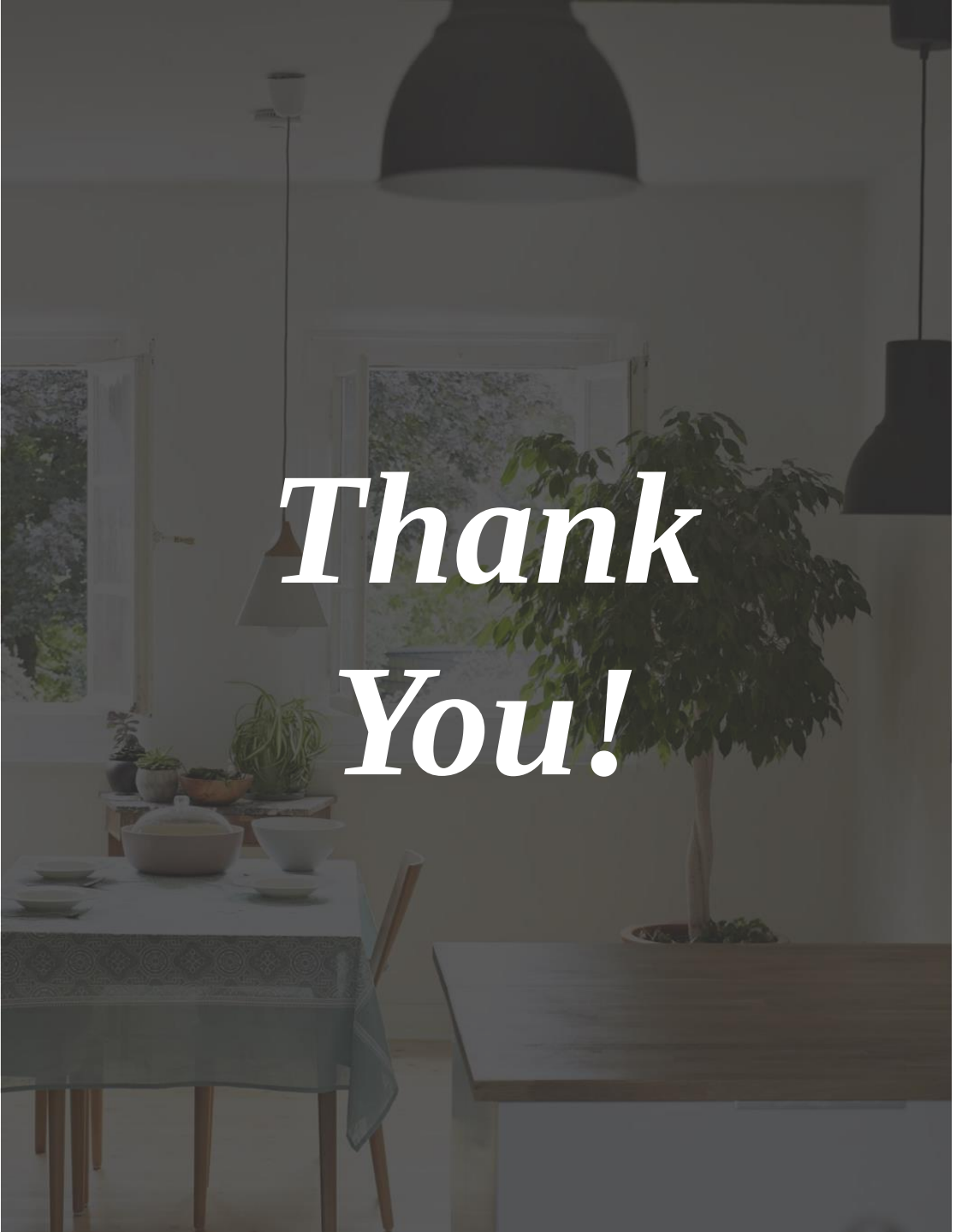
- The predictive model developed using logistic regression provides a valuable tool for assessing diabetes risk based on user input data.
- Through the Streamlit app, individuals can receive instant feedback on their potential risk of diabetes, enabling them to take proactive measures for preventive healthcare.

Future Scope

- Integration of additional features: Future iterations of the app could incorporate more features related to lifestyle, diet, and medical history to enhance the accuracy of diabetes risk prediction.
- Machine learning model enhancements: Exploring advanced machine learning algorithms and techniques could further improve the predictive accuracy of the model.
- Personalized recommendations: Incorporating personalized recommendations for lifestyle modifications and healthcare interventions based on individual risk profiles could empower users to make informed decisions about their health.
- Mobile application development: Developing a mobile version of the app could enhance accessibility and reach a wider audience, facilitating widespread adoption and usage.

References

- "Diabetes Health Indicators Dataset" on Kaggle: https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset
- Research paper: Smith, A., Jones, B., & Patel, C. (2020). Predictive Modeling of Diabetes Risk Factors Using Machine Learning Techniques. Journal of Health Informatics Research, 10(3), 215-228.
- Book: Brown, S. J., & Smith, R. L. (2018). Machine Learning in Healthcare: A Practical Approach. Chapman and Hall/CRC.
- Research paper: Yousuf, A., Khan, N. A., Bokhari, R. H., & Siddiqi, M. H. (2020). Prediction of Diabetes using Machine Learning Algorithms. Pakistan Journal of Medical Sciences, 36(S4), S48–S53. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10107388/



Thank You!

Remember, early detection is key to managing this risk effectively!!!

Under the guidance:
Dr. Chaojie (Jay) Wang

By:
Varshita Adavi
OC83614