**Table of Contents**

1. Abstract
2. Data Collection
3. Data Pre-Processing
4. Exploratory Data Analysis
5. Feature Engineering
6. Model Development
7. Model Deployment
8. Libraries Used
9. Conclusion

**Abstract**

This project involves the development of a diabetes prediction application, leveraging data analysis and machine learning techniques to create a predictive model. The application is implemented using Streamlit, a Python library that allows for the rapid development of interactive web applications. The process begins with thorough data analysis to understand the features and their correlation with diabetes. A suitable machine learning model is then developed and trained using this data to achieve high accuracy in predicting the likelihood of diabetes. The final application prompts users to input their health data and utilizes the trained model to predict their risk of diabetes. This tool aims to provide users with an easy and accessible way to assess their diabetes risk, enabling early intervention and informed health decisions.

Top of Form

Bottom of For

**Data Collection**

I downloaded a dataset from Kaggle containing 100,000 rows and 9 columns. This dataset is used for predicting diabetes.

A screenshot of a computer screen

Description automatically generated

**Data Pre-Processing**

A chart and diagram of a distribution

Description automatically generated with medium confidence

Upon examining the data, I observed a significant gender imbalance, with a notably higher number of females compared to males. To address this and ensure an accurate analysis of the correlation between gender and diabetes, I balanced the dataset by randomly sampling an equal number of female and male participants.

**Exploratory Data Analysis**

To facilitate the analysis of gender correlations, I converted the gender variable into numeric values: 0 for male and 1 for female. This conversion helps in accurately identifying correlations within the dataset.

After this transformation, I plotted a correlation heatmap to visually represent the relationships between different features. The heatmap revealed a notable positive correlation between diabetes and two key indicators: HbA1c levels and Blood-Glucose levels.

A diagram of a heatmap

Description automatically generated

**Feature Engineering**

After identifying the correlations, I combined the HbA1c and Blood-Glucose levels to create a new feature. This engineered feature demonstrated a stronger correlation with diabetes compared to the individual features alone, enhancing the predictive power of the model.



A diagram of a heatmap

Description automatically generated

**Model Development**

To develop a robust model, I utilized pipelines to test different algorithms using 20% of the data for training and 80% for testing. The models tested and their respective accuracies were as follows:

* **Linear Regression:** 95%
* **Support Vector Machine (SVM):** 96%
* **Decision Tree:** 97%

The Decision Tree model proved to be the most accurate and efficient, achieving the highest accuracy of 97%.

**Model Deployment**

To deploy the model, I exported it using joblib and then loaded it in a script named deployment.py. In this script, I take HbA1c and Blood Glucose levels as input to predict whether a patient is likely to have diabetes.

For the user interface, I utilized Streamlit, which allows for a simple and interactive web application to input the data and display the prediction results.

A screenshot of a black screen

Description automatically generated

**Libraries Used**

* **Pandas:** 1.3.3
* **NumPy:** 1.21.2
* **scikit-learn:** 0.24.2
* **matplotlib:** 3.4.3
* **seaborn:** 0.11.2
* **Streamlit:** 0.88.0
* **joblib:** 1.1.0

**Conclusion**

This project successfully developed a diabetes prediction application using machine learning techniques and data analysis. By carefully preprocessing the data, engineering relevant features, and selecting an optimal model, the application achieved high accuracy in predicting diabetes. The use of Streamlit for deployment ensures an easy-to-use interface for users to input their health data and receive a prediction. This tool can help in early detection and management of diabetes, potentially improving health outcomes through timely intervention.