**CSE 587 – Project Phase 3**

**Diabetes Prediction and Analysis App**

**Bhanu Teja Veeramachineni – 50606694 Venkata Anudeep Bandi – 50606997**

**Sai Sohan Kosaraju – 50560534 Akshay Kumar Dasarla - 50592353**

**Abstract**

This project focuses on developing a diabetes prediction and analysis application using Python, Streamlit, and machine learning. The app enables users to explore diabetes-related datasets, analyze health metrics through visualizations, and predict diabetes risk using key health indicators such as age, BMI, HbA1c level, and blood glucose level. The application features real-time predictions and stores user inputs and results for review, enhancing its potential as a tool for healthcare analysis and decision-making.

**1. Introduction**

The rise in diabetes prevalence necessitates accessible tools for early prediction and analysis. This project aims to bridge the gap by providing a web-based application that simplifies the process of diabetes risk assessment. The application leverages machine learning algorithms to predict diabetes risk, making it a practical tool for individuals, researchers, and healthcare providers. Its intuitive interface allows users to explore datasets, generate visual insights, and make predictions efficiently.

**2. Literature Review**

Several existing diabetes prediction models rely on complex workflows or are inaccessible to non-specialists. While tools like Python's scikit-learn library and Streamlit framework simplify model deployment, many apps lack the integration of real-time data input, visualization, and prediction storage. This app combines these features in a user-friendly platform, addressing a critical gap in usability and functionality.

**3. Features of the Application**

**3.1 Dataset Viewing**

The application provides an interactive view of the diabetes dataset, as shown in **Figure 1**. Users can inspect the data and understand its structure, enabling transparency and ease of exploration. The dataset includes attributes like age, BMI, hypertension, HbA1c level, and blood glucose level.

**3.2 Visualizations**

The app offers visualizations to analyze trends and patterns in the dataset:

* **Age Distribution**  
  Figure 2 illustrates the distribution of ages in the dataset, enabling insights into the age groups affected by diabetes.

A graph with blue lines

Description automatically generated

* **BMI vs. Diabetes**  
  Figure 3 highlights the correlation between BMI and diabetes status, showing distinct patterns for diabetic and non-diabetic individuals.

A graph of a diabetes

Description automatically generated with medium confidence

**3.3 Diabetes Prediction**

The app features a prediction module where users input health parameters such as age, hypertension, heart disease, BMI, HbA1c level, and blood glucose level. Using a trained Random Forest Classifier, the app predicts the likelihood of diabetes (Figure 4).

**3.4 Prediction Storage**

Predictions are stored in a structured table for review (Figure 5). Users can download the stored predictions as a CSV file for further analysis or record-keeping.

**4. Methodology**

**4.1 Data Preprocessing**

The dataset was preprocessed to handle missing values, standardize inputs, and prepare features for training.

**4.2 Machine Learning Model**

The Random Forest Classifier was selected for its robustness and accuracy. The model was trained on an 80/20 train-test split to predict diabetes based on six health indicators.

**4.3 Application Development**

The application was developed using:

* **Python libraries**: Pandas, scikit-learn, Matplotlib.
* **Framework**: Streamlit for rapid deployment and interactivity.

**5. Results**

The app successfully predicts diabetes with high accuracy and provides meaningful insights through visualizations. The user-friendly interface ensures ease of use for non-technical users.

**6. Conclusion and Future Work**

The Diabetes Prediction and Analysis App demonstrates the potential of machine learning and interactive web applications in healthcare. Future enhancements may include:

* Incorporating additional datasets for model improvement.
* Expanding the range of visualizations.
* Adding support for other health conditions.

**References**

1. Breiman, L. (2001). Random forests. *Machine learning, 45*(1), 5-32.
2. Streamlit Documentation: https://docs.streamlit.io/
3. scikit-learn Documentation: <https://scikit-learn.org/>