# Glucose Guru

**Submitted for**

**Statistical Machine Learning CSET211**


Submitted by:

**E23CSEU0920 Isha Kaushik**


Submitted to

**DR. ASHIMA YADAV**

**July-Dec 2024**

**SCHOOL OF COMPUTER SCIENCE AND ENGINEERING**

## 1. Abstract

**Glucose Guru** is a machine learning-based project aimed at predicting the likelihood of diabetes in individuals based on health indicators like glucose levels, BMI, blood pressure, insulin levels, age, and others. This tool helps in early detection and prevention, offering personalized health insights. By using various machine learning algorithms (such as Random Forest, Naive Bayes, and others), it delivers accurate predictions that assist healthcare professionals and individuals in making informed decisions.

## 2. Introduction

Diabetes, particularly Type 2 diabetes, is one of the leading causes of health complications globally. Early detection is crucial for effective management and prevention. **Glucose Guru** uses machine learning techniques to predict the risk of diabetes based on the Pima Indians Diabetes dataset, which includes health attributes like glucose, BMI, age, and more. This project seeks to provide an easy-to-use platform that allows users to input their data and receive predictions, enabling them to make informed decisions about their health.

## 3. Related Work

Several notable projects have been developed to predict diabetes risk, each contributing to the field in different ways:

**1. Diabetes Prediction API:** This project uses logistic regression and random forest to predict the likelihood of diabetes based on health parameters. It is designed for healthcare applications, utilizing common machine learning models to deliver predictions.

**2. IBM Watson:** IBM Watson is a large-scale healthcare AI system that leverages neural networks and extensive datasets to provide medical insights and predictions, including diabetes risk assessments. It operates on a much larger scale with diverse healthcare data.

**3. Microsoft Azure Model:** This model offers seamless integration into cloud systems, enabling deployment in various healthcare settings. It supports predictive analytics and aims for scalable deployment within enterprise-level applications.

**4. Google Health:** Google Health employs deep learning techniques to achieve high prediction accuracy in diagnosing various medical conditions, including diabetes, by analyzing a vast amount of healthcare data.

**5. Ada Health:** Ada Health uses a questionnaire-based approach to assess an individual's risk of diabetes. By collecting detailed health history from users, it provides a personalized risk assessment.

These projects are all valuable contributions to healthcare analytics and prediction, but **Glucose Guru** focuses on delivering an accessible, user-friendly, and interpretable solution specifically designed for diabetes risk prediction, aiming for precision and small-scale deployment in healthcare settings.

## 4. Methodology

The **Glucose Guru** system follows a structured methodology for predicting diabetes:

- **Data Collection**: The Pima Indians Diabetes dataset was sourced from Kaggle. The dataset contains 768 records of female patients and includes 8 features: Number of Pregnancies, Glucose Level, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function, and Age.
- **Data Pre-processing**: To ensure the quality of predictions, data pre-processing steps were employed:
    - Feature scaling was applied to normalize the data.
    - SMOTE (Synthetic Minority Over-sampling Technique) was used to address class imbalance, ensuring the dataset had a balanced number of positive and negative diabetes instances.
- **Model Building**: Various machine learning algorithms were trained and tested, including:
    - **Random Forest**: An ensemble method that uses multiple decision trees to improve predictive performance.
    - **Naive Bayes**: A probabilistic classifier based on Bayes' Theorem, assuming independence between features.
    - **XGBoost**: A powerful gradient boosting technique known for its accuracy and efficiency.
    - **SVM**: A powerful model that finds the optimal hyperplane to separate classes.
    - **Decision Trees**: A simple and interpretable model that splits data into subsets based on feature thresholds to make prediction.

    - **KNN:** A non-parametric model that classifies based on the nearest neighbours.
    - **Logistic Regression:** A classic model for binary classification tasks.

- **Model Evaluation**: Each model was evaluated using metrics like accuracy, precision, recall, and F1-score. Random Forest showed the best performance and was selected as the final model.
- **Deployment**: A Flask web application was created to make **Glucose Guru** accessible to users. The backend processes user inputs (health indicators) and provides predictions using the trained model. Users receive immediate feedback regarding their diabetes risk.

## 5. Hardware/Software Used

- **Hardware**:
  - A computer with at least 4GB of RAM and 1GB of available storage.
  - Internet access for installing necessary libraries and running the web application.
- **Software**:
  - **Python 3.x**: The programming language used for building the machine learning model and the web application.
  - **Libraries**:
    - Flask: For creating the web framework.
    - Scikit-learn: For implementing machine learning algorithms like Random Forest, Naive Bayes, and XGBoost.
    - Pandas and Numpy: For data manipulation and preprocessing.
    - Pickle: To save and load the trained machine learning model.
    - Matplotlib and Seaborn: For data visualization and model evaluation metrics.
  - **IDE**: IDLE and text editor Sublime Text

## 6. Experimental Results

| | Accuracy | F1 score | Precision | Recall |
|---|---|---|---|---|
| Logistic Regression | 68% | 0.62 | 0.54 | 0.72 |
| Decision Trees | 68% | 0.60 | 0.55 | 0.67 |
| Random Forest | 70% | 0.65 | 0.56 | 0.76 |
| SVM | 67% | 0.61 | 0.53 | 0.72 |
| XGBoost | 68% | 0.62 | 0.54 | 0.72 |

Among the models tested, **Random Forest** performed the best, making it the final choice for the backend. The model was trained using 5-fold cross-validation to ensure generalization and prevent overfitting.

## 7. Conclusions

The **Glucose Guru** application offers a simple yet powerful solution for predicting the likelihood of diabetes based on a user's health data. By using advanced machine learning models such as Random Forest, the system provides reliable predictions that can assist in early diabetes detection. Although the model is already effective, the project can be further enhanced with additional data and more sophisticated algorithms. The web application makes this tool accessible to non-technical users, empowering them to monitor their health proactively. The project has shown that machine learning can be an invaluable tool in healthcare, providing a foundation for future advancements in predictive health analytics.

## 8. Future Scope

The **Glucose Guru** project has significant potential for expansion and improvement:

- **Integration with Health Devices**: Future versions of **Glucose Guru** could sync with wearable health devices like smartwatches to collect real-time data on vital signs (e.g., heart rate, step count, activity levels), further improving the accuracy of predictions.
- **Mobile Application**: Creating a mobile version of **Glucose Guru** would make it even more accessible to a larger audience, offering features like push notifications for regular check-ups and reminders to monitor health.
- **Cross-disease Predictions**: The model could be adapted to predict other diseases, such as hypertension, cardiovascular disease, and kidney disease, using similar health indicators.
- **Longitudinal Data**: Incorporating longitudinal data (data collected over time) would help improve the prediction model's performance by accounting for the progression of health over months or years.
- **AI-driven Personalization**: With further development, the platform could provide personalized health suggestions based on a user’s data, helping to lower their risk of developing diabetes or other related conditions.

## 9. GitHub Link of Your Complete Project

You can access the full project, including the code, dataset, README file, and presentation, through the following GitHub repository:

https://github.com/IshaKaushik31/GlucoseGuru