Diabetes Prediction System
GlycoCheck

Hrishikesh Balakrishnan

Email: hrishiumb@gmail.com
LinkedIn: www.linkedin.com/in/hrishikeshbalakrishnan
GitHub: Hunt4code

Background and Motivation

Diabetes is a chronic condition affecting millions worldwide, with severe long-term complications such as heart disease, kidney failure, and nerve damage. Early detection and proactive management are crucial in preventing the progression of diabetes and mitigating its health risks. Traditional methods for diagnosing diabetes rely on clinical tests such as fasting blood sugar, HBA1C levels, and glucose tolerance tests. However, these methods often require in-person visits, lab tests, and time-consuming manual assessments.

With the advancement of machine learning (ML) and artificial intelligence (AI), it is now possible to develop predictive models that analyse various health parameters and provide early warnings for diabetes risk. The GlycoCheck Diabetes Prediction System leverages demographic, lifestyle, medical history, and lab test results to build a robust machine-learning model capable of predicting diabetes status with high accuracy.

Objective of the Project

The primary goal of this project is to develop an AI-powered system that predicts an individual's risk of diabetes based on multiple health indicators. The system is designed to:

1. Analyse patient data using Exploratory Data Analysis (EDA) to understand diabetes risk factors.

2. Build and compare multiple machine learning models (Logistic Regression, Random Forest, and XGBoost) to achieve the most accurate prediction model.

3. Identify key features influencing diabetes risk using feature importance techniques.

4. Develop a risk score system to categorize patients based on their likelihood of developing diabetes.

5. Deploy the model via an API and a user-friendly web interface for real-time predictions.

Dataset Overview

The dataset used for this project contains **5,292 patient records** and **27 features**, including:

- **Demographic Information:** Age, Gender, Urban/Rural Living Conditions
- **Lifestyle & Habits:** Smoking Status, Alcohol Intake, Diet Type, Physical Activity
- **Medical History:** Family History of Diabetes, Hypertension, Polycystic Ovary Syndrome (PCOS), Medication for Chronic Conditions
- **Lab Test Results:** BMI, Cholesterol Levels, Fasting Blood Sugar, HBA1C, Glucose Tolerance Test, C-Protein Level

The target variable is Diabetes_Status, which is a binary classification (Yes/No). This serves as the dependent variable for model training and evaluation.

Significance of the Project

The GlycoCheck system has real-world applications in healthcare, particularly for:

- **Primary Care Physicians & Endocrinologists:** Assisting in early-stage diagnosis of diabetes.
- **Public Health Initiatives:** Identifying high-risk individuals in communities and promoting preventive healthcare.
- **HealthTech & Wellness Apps:** Integrating predictive models for proactive health monitoring.
- **Clinical Research & Epidemiology:** Understanding correlations between lifestyle factors and diabetes prevalence.

Methodology

The methodology followed in this project is structured into distinct **phases**, ensuring a systematic approach from **data collection to model deployment**. Each phase plays a crucial role in achieving a **robust and scalable diabetes prediction system**.

1. Data Collection and Preprocessing

1.1 Data Overview

The dataset consists of **5,292 records with 27 features**, including demographic, lifestyle, medical history, and lab test results. The target variable is **Diabetes_Status**, which indicates whether a person has diabetes (**Yes/No**).

1.2 Handling Missing Data

Missing values were analyzed and treated accordingly.

- **Numerical features** were filled with their respective **mean values** to ensure consistency in statistical representation.
- **Categorical features** were replaced with the **most frequent category (mode)** to maintain logical coherence.

1.3 Encoding Categorical Data

Since machine learning models work with numerical data, categorical variables such as **Gender, Smoking Status, and Diet Type** were converted into numerical format using **Label Encoding**. This transformation allows the models to process categorical information effectively.

1.4 Feature Scaling

Features such as **Age, BMI, Cholesterol Levels, and Blood Sugar** exist on different scales. **Standard Scaling** was applied to ensure that all features contribute equally to the model without bias due to varying numerical ranges.

2. Exploratory Data Analysis (EDA)

2.1 Summary Statistics

A statistical summary of the dataset was generated to identify key insights.

- **Age**: Most individuals in the dataset fall within the **30–60 age range**.
- **BMI**: The average BMI of **27.46** indicates a general trend toward being overweight.
- **HBA1C & Fasting Blood Sugar**: Many individuals show values suggesting **prediabetes or diabetes**.
- **Cholesterol Levels**: High variance observed, with extreme values above **300 mg/dL**.

2.2 Correlation Analysis

A correlation analysis was conducted to understand the relationships between variables.

- **Fasting Blood Sugar and HBA1C (0.81 correlation)**: A strong positive correlation, as both serve as indicators of diabetes.
- **Glucose Tolerance Test and Fasting Blood Sugar (0.78 correlation)**: Individuals with higher fasting glucose levels often have impaired glucose tolerance.
- **BMI and Waist-Hip Ratio (0.63 correlation)**: These features strongly correlate due to their connection with body fat distribution.

2.3 Outlier Detection

Outliers were identified in several numerical features, particularly in **Blood Sugar, HBA1C, Cholesterol, and C-Protein Levels**. These extreme values suggest possible diabetes cases or erroneous data entries.

3. Machine Learning Model Development

3.1 Data Splitting

The dataset was split into **80% training and 20% testing** to evaluate the model's performance effectively. **Stratified sampling** was used to ensure that both classes (**Diabetic and Non-Diabetic**) were well-represented in both sets.

3.2 Model Training

Three machine learning models were trained and evaluated for diabetes prediction:

- **Logistic Regression**: A simple and interpretable model that performed well but had limitations in handling complex relationships.
- **Random Forest Classifier**: A powerful ensemble model that significantly improved prediction accuracy by capturing non-linear relationships.
- **XGBoost Classifier**: The best-performing model, achieving the highest accuracy and effectively handling complex patterns in the data.

3.3 Model Performance Comparison

The models were evaluated using **accuracy, precision, recall, and F1-score**.

Model	**Accuracy (%)**
Logistic Regression	82.4
Random Forest	87.2
XGBoost	**89.5**

The **XGBoost model outperformed the others** with an accuracy of **89.5%**, making it the best candidate for deployment.

4. Feature Engineering & Clustering

4.1 Risk Score Calculation

A **Diabetes Risk Score** was created using a weighted combination of the most important health indicators:

- **HBA1C** (40%)
- **Fasting Blood Sugar** (30%)
- **BMI** (20%)
- **Cholesterol Levels** (10%)

This risk score helps categorize individuals into **low-risk, medium-risk, and high-risk** groups for diabetes.

4.2 K-Means Clustering for Patient Segmentation

Patients were clustered into **three groups** based on their health attributes:

- **Cluster 0 (Low Risk)**: Individuals with normal glucose levels, healthy BMI, and controlled cholesterol levels.
- **Cluster 1 (Medium Risk)**: Individuals with slightly elevated blood sugar and cholesterol levels, requiring lifestyle interventions.
- **Cluster 2 (High Risk - Likely Diabetic)**: Patients with significantly high HBA1C and fasting blood sugar levels, indicating a strong diabetes diagnosis.

These insights can be useful in healthcare applications for **personalized treatment recommendations**.

5. Model Deployment

5.1 Flask API for Predictions

The trained **Random Forest model** was deployed using **Flask**, allowing real-time predictions via an API. This API accepts **patient health data** and returns a **diabetes risk prediction (Diabetic/Non-Diabetic)**.

5.2 Streamlit Web App for User Interaction

A **Streamlit-based web interface** was developed to allow users to manually input their **health parameters** and receive an instant **diabetes prediction**. The web app provides a **simple and interactive experience**, making it easy for non-technical users to assess their diabetes risk.

Conclusion & Future Enhancements

Final Findings

- The **XGBoost model achieved the highest accuracy (89.5%)**, demonstrating strong predictive capabilities.
- **HBA1C, Fasting Blood Sugar, and BMI** were identified as the most influential factors in diabetes risk assessment.
- **The API and web app were successfully deployed**, enabling real-time diabetes predictions.

Future Enhancements

- **Hyperparameter tuning** to further optimize model performance.
- **Deployment on AWS/GCP** to enhance scalability and accessibility.
- **Integration with SHAP/LIME** for improved model interpretability.
- **Real-time health monitoring features** for continuous tracking and risk assessment.