****

**DEPT: SOFTWARE ENGINEERING**

**FUNDAMENTAL OF MACHINE LEARNING**

**NAME: FUAD SEID**

**ID: DBU1401283**

**SUBMITTED TO TEACHER DERBEW**

**SUBMISSION DATE: 02/05/2017**

**Diabetes Prediction Model Report**

**1. Problem Definition, Data Source, and Description**

**Problem Definition:**

Diabetes is one of the most common chronic diseases in the world, and its early detection plays a significant role in preventing complications. The goal of this project is to develop a machine learning model capable of predicting whether an individual is at risk of developing diabetes based on a set of health-related features. Early intervention and management are crucial for diabetic patients, and this model could serve as a useful tool in healthcare systems to identify at-risk individuals and assist in timely medical intervention.

I started this project because my mother is a diabetic patient, and over time I have realized the need to use technology and data-driven approaches to improve her health management. Given the increasing prevalence of diabetes worldwide, this is a problem that is close to my heart. Through this project, I aim to help individuals detect diabetes early, improving quality of life and preventing long-term complications.

**Data Source:**

The dataset used for this project is the Pima Indians Diabetes Database, a widely used dataset in healthcare machine learning tasks. It is available on GitHub at: https://github.com/praveenbharti1/diabetes\_prediction/blob/main/diabetes.csv

This dataset contains medical diagnostic data collected from female patients of Pima Indian heritage, a population with a relatively high rate of diabetes. It includes various health-related features that can be used to predict diabetes, such as glucose levels, BMI, age, and more.

**Dataset Description:**

The dataset consists of 768 observations (instances) and 9 attributes (features). Here is a breakdown of the features:

* **Pregnancies**: Number of pregnancies the patient has had.
* **Glucose**: Plasma glucose concentration after a 2-hour oral glucose tolerance test.
* **BloodPressure**: Diastolic blood pressure measured in mm Hg.
* **SkinThickness**: Triceps skinfold thickness in mm.
* **Insulin**: 2-hour serum insulin concentration in µU/mL.
* **BMI**: Body Mass Index (weight-to-height ratio, used as an indicator of obesity).
* **DiabetesPedigreeFunction**: A function that scores the likelihood of diabetes based on family history.
* **Age**: The age of the individual.
* **Outcome**: The target variable, where 1 indicates that the patient has diabetes and 0 indicates they do not.

The data is primarily numerical, with the Outcome column being binary. The dataset is relatively clean but has some class imbalance, with fewer diabetic cases than non-diabetic cases.

**2. Exploratory Data Analysis (EDA) Findings and Visualizations**

**EDA Overview:**

The purpose of exploratory data analysis (EDA) is to understand the underlying structure of the data, detect outliers or anomalies, identify trends, and select important features for the predictive model.

* **Missing Values**: No missing values were found, making preprocessing straightforward.
* **Feature Distribution**: Histograms and box plots revealed skewed distributions in features like Glucose, BMI, and Age, indicating the need for scaling.
* **Class Imbalance**: More non-diabetic cases (Outcome = 0) than diabetic cases (Outcome = 1) were observed, potentially impacting model bias.
* **Correlations**: A heatmap showed that Glucose, BMI, and Age had strong correlations with the Outcome variable.
* **Outliers**: Outliers in SkinThickness and Insulin were detected and analyzed.

**3. Preprocessing Steps and Choices**

**Data Cleaning:**

* **Scaling Features**: Standardization (zero-mean, unit-variance scaling) was applied to ensure uniform feature ranges.
* **Train/Test Split**: Data was split into 80% training and 20% test sets.
* **Handling Class Imbalance**: Techniques like oversampling and F1-score evaluation were used to address class imbalance.

**4. Model Selection and Training Details**

The project uses **K-Nearest Neighbors (KNN)** to predict diabetes:

* **KNN Classifier**: A non-parametric algorithm used for classification, where the class of a sample is determined by the majority class of its nearest neighbors. It is a simple yet effective model for classification tasks.
  + **K** represents the number of nearest neighbors considered.
  + **Distance Metric**: Typically Euclidean distance is used, but it can be customized.

You can experiment with different values of **K** to find the best-performing model for this dataset.

**5. Model Evaluation Metrics and Discussion**

We use several metrics to evaluate the model’s performance:

* **Accuracy**: The proportion of correct predictions.
* **Precision**: The proportion of positive predictions that are actually correct.
* **Recall**: The proportion of actual positive cases that are correctly identified.
* **F1-score**: The harmonic mean of precision and recall.
* **Confusion Matrix**: A table used to describe the performance of a classification model.

**6. Interpretation of Results**

* **Feature Importance**: Glucose, BMI, and Age were the top three most important predictors.
* **Model Insights**: The Random Forest model performed well but lacked interpretability.

**7. Deployment Details and Instructions**

**Cloud Deployment:**

The model has been deployed on **Render Cloud**, making it accessible via API at: https://machine-learning-1-yl7v.onrender.com/predict

**API Usage:**

You can send a POST request to the API with patient data to receive a diabetes prediction.

**Using Postman:**

1.  Open Postman.

2.  Select POST as the request type.

3.  Enter the URL: https://machine-learning-1-yl7v.onrender.com/predict

4.  Navigate to the Body tab and select raw > JSON format.

5.  Enter the following JSON payload:

6.  {

 "Pregnancies": 2,

 "Glucose": 120,

 "BloodPressure": 70,

"SkinThickness": 20,

"Insulin": 85,

"BMI": 25.3,

"DiabetesPedigreeFunction": 0.5,

"Age": 30

}

7. Click Send, and you will receive a JSON response with the predicted outcome.

Using cURL:

curl -X POST "https://machine-learning-1-yl7v.onrender.com/predict" \

-H "Content-Type: application/json" \

-d '{"Pregnancies": 2, "Glucose": 120, "BloodPressure": 70, "SkinThickness": 20, "Insulin": 85, "BMI": 25.3, "DiabetesPedigreeFunction": 0.5, "Age": 30}'

**8. Potential Limitations and Future Improvements**

* **Class Imbalance**: Further exploration of SMOTE for synthetic minority data generation.
* **Feature Engineering**: Adding lifestyle factors like physical activity and diet.
* **Model Complexity**: Exploring interpretable AI techniques.
* **Model Generalization**: Testing on different datasets to improve generalizability.

This report provides a comprehensive overview of the diabetes prediction model, from problem definition to deployment, offering insights into its strengths, limitations, and future enhancements.