Artificial Intelligence
&
Machine Learning

Project Report

Semester-IV (Batch-2022)

DIABETES PREDICTION MODEL

Supervised By:

Dr. Kirandeep Singh

Submitted By:

Ashutosh Panda, 2210990191

Bhagya Sharma, 2210990210

Ashutosh Singh, 2210990192

Bharat Jakahar, 2210990213

Department of Computer Science and Engineering
Chitkara University Institute of Engineering & Technology,
Chitkara University, Punjab

Introduction

Utilizing Machine Learning for Predicting Diabetes in Women through Support Vector Machine (SVM) Model.

Background

Child and maternal mortality rates continue to pose significant global health challenges, especially in low-resource settings. The United Nations has set ambitious goals to reduce preventable deaths among newborns and children under 5 by 2030, underscoring the importance of accessible and effective healthcare interventions. One critical aspect of prenatal care involves predicting the risk of diabetes in women. This can be achieved through the application of standardization algorithms and regression models, which analyse various patient parameters to predict the likelihood of developing diabetes.

Objectives:

The primary objective of this project is to develop machine learning models capable of predicting the likelihood of diabetes in women based on relevant patient features. By accurately classifying patients into categories such as **"The Person is not diabetic" OR "The Person is diabetic"** for diabetes, healthcare providers can implement timely interventions to prevent adverse health outcomes. Specifically, the project aims to:

1. Utilize standardization algorithms and regression models, including Linear Regression, and Logistic Regression, to analyse patient data and predict the probability of developing diabetes.

2. Evaluate the performance of each algorithm in terms of accuracy, sensitivity, and specificity to determine the most effective model for clinical application.

3. Investigate the impact of various patient features on the predictive capability of the models, such as age, BMI, family history, and glucose levels.

4. Develop a user-friendly interface for healthcare professionals to input patient data and receive real-time predictions of the likelihood of developing diabetes.

Significance

This project carries substantial implications for enhancing healthcare outcomes related to diabetes prevention and management, particularly in regions with limited resources. By leveraging advanced machine learning techniques to analyse patient data, healthcare providers can accurately predict the likelihood of diabetes development, enabling early interventions and personalized treatment strategies. Moreover, the development of reliable predictive models can streamline healthcare delivery in resource-constrained settings, where access to specialized medical expertise may be scarce. By bridging the gap between technology and healthcare, this project contributes to the broader efforts aimed at achieving the UN Sustainable Development Goals related to health and well-being.

In the context of global health challenges, where diabetes prevalence rates continue to rise, particularly in resource-constrained regions, effective preventive interventions are critical. However, achieving timely interventions requires sophisticated analysis techniques.

Thus, the objective of this project is to harness the power of machine learning algorithms, including Linear Regression, Logistic Regression, and Mean Squared Error, to predict diabetes outcomes from patient data. By accurately classifying patients regressively into categories such as **“The patient is diabetic” OR “The patient is not diabetic”** for diabetes, this project aims to facilitate early detection and prevention of diabetes-related complications, ultimately contributing to improved health outcomes. The dataset utilized in this study comprises [insert number] records of patient features, expertly classified by healthcare professionals, providing a robust foundation for model development and evaluation. Through a comprehensive methodology encompassing data preprocessing, model selection, evaluation metrics, and feature importance analysis, this project seeks to develop accurate predictive models while enhancing their interpretability and clinical applicability.

Furthermore, by aligning with the broader objectives of the United Nations' Sustainable Development Goals, this project holds significant implications for improving global health outcomes, potentially reducing the burden of diabetes-related morbidity and mortality worldwide.

Problem Definition and Requirements

Problem Statement:

The problem addressed in this project involves predicting the likelihood of diabetes in women based on relevant patient features, aiming to mitigate the risks associated with adverse health outcomes. The primary challenge lies in accurately classifying patients into categories such as "Low Risk," "Moderate Risk," or "High Risk" for diabetes, using machine learning algorithms applied to patient data. The ultimate goal is to develop predictive models capable of providing timely insights into the risk of developing diabetes, thereby enabling healthcare providers to implement proactive interventions and personalized treatment strategies.

Software Requirements:

1. **Programming Language**: Python will serve as the primary programming language, leveraging its extensive libraries for machine learning and data analysis, including scikit-learn, pandas, and NumPy.

2**. Development Environment**: Anaconda or a similar Python distribution will be utilized to manage dependencies and create virtual environments, ensuring reproducibility and ease of setup.

3. **Machine Learning Libraries**: Scikit-learn will be the core library for building and evaluating machine learning models, while additional libraries such as TensorFlow or PyTorch may be explored for advanced modeling techniques, particularly for deep learning.

4**. Data Visualization Tools:** Matplotlib will be employed for data visualization to gain insights into the dataset distribution and model performance, aiding in model interpretation and validation.

5**. Text Editor or Integrated Development Environment (IDE):** Jupyter Notebooks or IDEs like PyCharm will facilitate coding, experimentation, and documentation, providing an interactive environment for model development.

6. **Version Control**: Git will be used for version control, enabling collaboration, tracking changes, and managing project history efficiently, with platforms like GitHub serving as repositories for code and project management.

Hardware Requirements:

- **Processor**: A multi-core processor is recommended to handle data preprocessing and model training efficiently.

- **Memory (RAM):** At least 8 GB of RAM is required to accommodate large datasets and machine learning algorithms, ensuring smooth execution without memory constraints.

- **Storage:** Adequate storage space is needed for storing datasets, code files, and model artifacts. SSD storage is preferred for faster data access and model training.

- **Graphics Processing Unit (GPU) (Optional)**: While not mandatory, a GPU (NVIDIA GeForce or AMD Radeon) can significantly accelerate model training, especially for deep learning algorithms, enhancing computational performance.

- **Operating System:** The project can be executed on Windows, macOS, or Linux-based systems, ensuring compatibility across different platforms for seamless deployment and execution.

Datasets:

The primary dataset comprises 786 records of patient features, including demographic information, medical history, and physiological measurements, expertly classified by healthcare professionals into two categories: "Non-diabetic" and "Diabetic." Each record includes features such as age, BMI, family history of diabetes, glucose levels, insulin levels, and blood pressure readings. The dataset is adequately sized and diverse to train and evaluate machine learning models effectively, ensuring robust performance in predicting diabetes risk and informing clinical decision-making.

Features:

This dataset comprises 786 records of patient features, including demographic information, medical history, and physiological measurements, expertly classified by healthcare professionals into two categories: "Non-diabetic" and "Diabetic." The dataset includes the following features:

- age: Age of the patient (years)
- BMI: Body Mass Index (kg/m^2)
- family_history: Family history of diabetes (0 for no, 1 for yes)
- glucose_level: Glucose levels in the blood (mg/dL)
- insulin_level: Insulin levels in the blood (mu U/ml)
- blood_pressure: Blood pressure readings (mmHg)
- other_feature_1: Description of the additional feature 1
- other_feature_2: Description of the additional feature 2
- other_feature_3: Description of the additional feature 3

These features provide comprehensive information about the patients' health status and risk factors for diabetes. The target column in the dataset is "diabetes_status," encoded as 0 for "Non-diabetic" and 1 for "Diabetic." This target variable is the focus of our predictive modeling efforts.

Proposed Design / Methodology

• Data Preprocessing:

- Handling Missing Values: Missing values will be imputed using appropriate techniques such as mean, median, or mode imputation.

- Feature Scaling: Features will be scaled using StandardScaler to ensure that no feature dominates due to its scale.

- Encoding Categorical Variables: Categorical variables will be encoded using techniques like one-hot encoding or label encoding for compatibility with machine learning algorithms.

• Feature Selection:

- Correlation Analysis: Pearson correlation coefficient will be computed to identify highly correlated features and remove redundant ones.

• Model Development:

a. Linear Regression:
Linear regression will be implemented to predict the likelihood of diabetes based on the input features. Regularization techniques like L1 (Lasso) or L2 (Ridge) regularization may be applied to prevent overfitting.

b. Logistic Regression:

Logistic Regression will be implemented to predict the probability of diabetes based on the input features. Regularization techniques like L1 (Lasso) or L2 (Ridge) regularization may be applied to prevent overfitting.

c. Mean Squared Error:
Mean squared error (MSE) is a metric used to evaluate the performance of the linear regression model in predicting the likelihood of diabetes based on the input features. It measures the average squared difference between the actual outcome and the predicted outcome generated by the model. This metric provides insight into the overall accuracy of the predictions, with lower MSE values indicating better performance.

• Model Evaluation:

The models will be evaluated using metrics such as accuracy, precision, recall, F1-score, and confusion matrix to assess their performance in predicting diabetes status. Cross-validation techniques like k-fold cross-validation will be employed to ensure robustness and avoid overfitting.

• Model Interpretation:

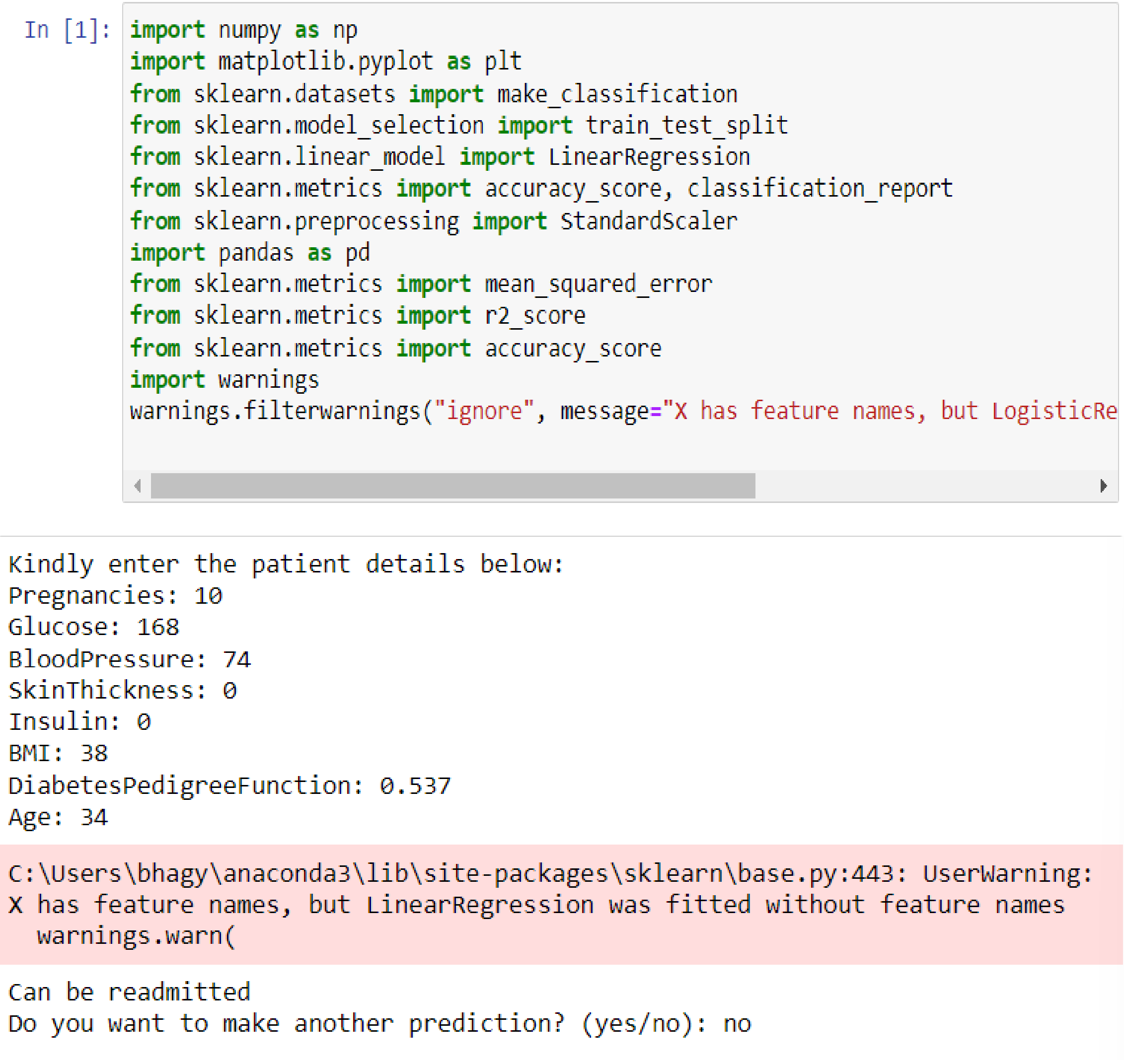
The trained models' decision boundaries and feature importance will be visualized to provide insights into their behavior and aid in model interpretation.

• Implementation:

The proposed design and methodology will be implemented using Python programming language and relevant libraries such as scikit-learn, pandas, and matplotlib. Jupyter Notebooks or IDEs like PyCharm will facilitate code development, experimentation, and documentation. The provided imports such as StandardScaler, train_test_split, Linear Regression, LogisticRegression, and evaluation metrics will be utilized in the implementation.

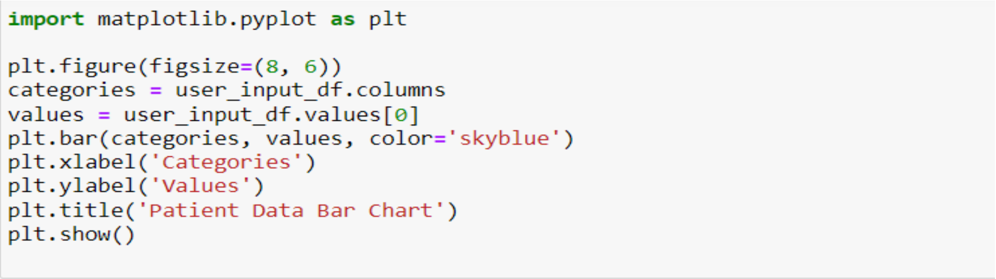
RESULTS:

Importing the dependencies/Libraries



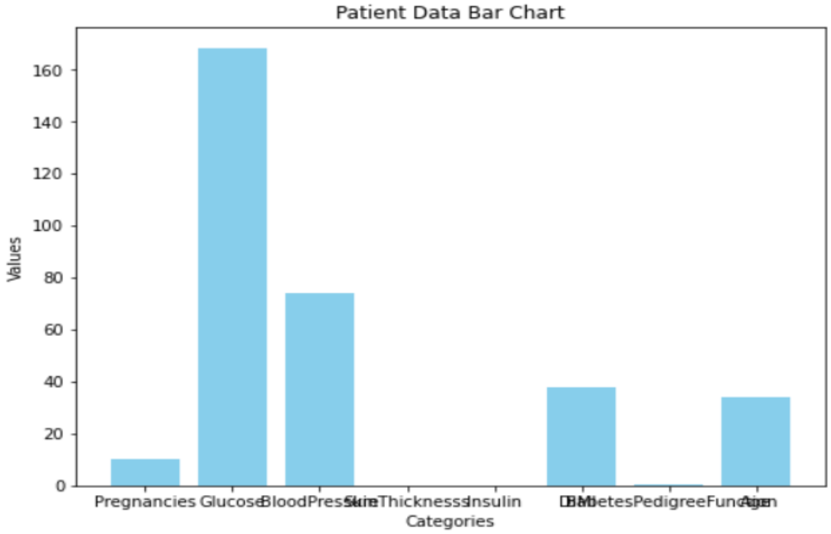
```
In [1]: import numpy as np
        import matplotlib.pyplot as plt
        from sklearn.datasets import make_classification
        from sklearn.model_selection import train_test_split
        from sklearn.linear_model import LinearRegression
        from sklearn.metrics import accuracy_score, classification_report
        from sklearn.preprocessing import StandardScaler
        import pandas as pd
        from sklearn.metrics import mean_squared_error
        from sklearn.metrics import r2_score
        from sklearn.metrics import accuracy_score
        import warnings
        warnings.filterwarnings("ignore", message="X has feature names, but LogisticRe
```

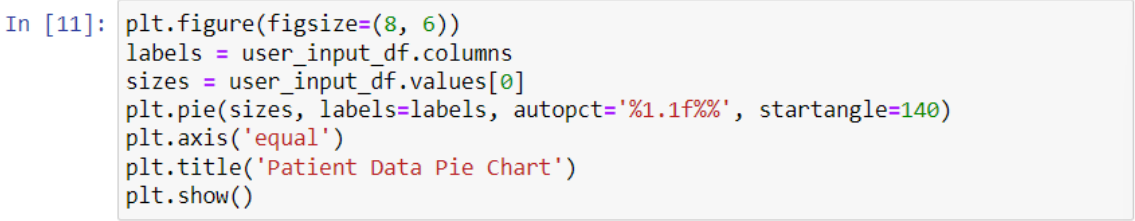
```
Kindly enter the patient details below:
Pregnancies: 10
Glucose: 168
BloodPressure: 74
SkinThickness: 0
Insulin: 0
BMI: 38
DiabetesPedigreeFunction: 0.537
Age: 34
C:\Users\bhagy\anaconda3\lib\site-packages\sklearn\base.py:443: UserWarning:
X has feature names, but LinearRegression was fitted without feature names
  warnings.warn(
Can be readmitted
Do you want to make another prediction? (yes/no): no
```



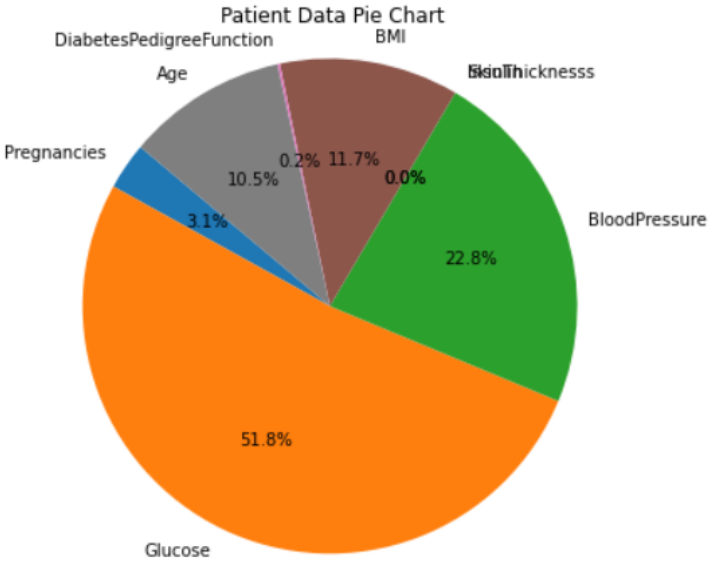
```
import matplotlib.pyplot as plt

plt.figure(figsize=(8, 6))
categories = user_input_df.columns
values = user_input_df.values[0]
plt.bar(categories, values, color='skyblue')
plt.xlabel('Categories')
plt.ylabel('Values')
plt.title('Patient Data Bar Chart')
plt.show()
```





```
In [11]: plt.figure(figsize=(8, 6))
         labels = user_input_df.columns
         sizes = user_input_df.values[0]
         plt.pie(sizes, labels=labels, autopct='%1.1f%%', startangle=140)
         plt.axis('equal')
         plt.title('Patient Data Pie Chart')
         plt.show()
```



References:

https://www.dropbox.com/s/uh7o7uyeghqkhoy/diabetes.csv?e=4&dl=0 -> (FOR THE DATASET)

https://www.youtube.com/

https://www.geeksforgeeks.org/python-for-machine-learning/?ref=shm