Final Project Report

Diabetes Prediction using Machine Learning

Submitted By:

David Salazar 300356922

Seulah Lee 300361747

Ishwinder Singh 300351938

Introduction and discovery:

Diabetes is a complex metabolic disorder that poses significant health risks if not managed effectively. Its prevalence worldwide has led to a growing need for accurate and early diagnostic methods. The dataset sourced from the National Institute of Diabetes and Digestive and Kidney Diseases provides an opportunity to explore predictive modeling using machine learning techniques to aid in diagnosing diabetes based on diagnostic measurements.

**Medical Importance:** Diabetes, if left undiagnosed or untreated, can lead to severe health complications, including cardiovascular diseases, kidney issues, eye damage, and more. Early detection plays a crucial role in preventing these complications and improving patient outcomes.

**Diagnostic Challenges:** Traditional diagnostic methods may not always be sufficient or timely in identifying diabetes. Machine learning offers a promising avenue to complement existing diagnostic procedures by leveraging a wide range of data points and complex patterns to enhance accuracy and reliability in identifying patients at risk.

**Data-Driven Insights:** The dataset contains valuable diagnostic measurements that, when analyzed using machine learning algorithms, can potentially uncover hidden patterns and correlations that human analysis might overlook. This can lead to the identification of significant predictive factors associated with diabetes.

Framing the Problem:

The analysis aims to address the following questions:

1. Can we accurately predict the presence or absence of diabetes in patients based on diagnostic measurements?
2. What are the key factors or measurements that significantly contribute to this prediction?
3. How can machine learning assist in enhancing the diagnostic process for diabetes, potentially improving early detection and subsequent treatment?

These questions are crucial as accurate early diagnosis of diabetes significantly impacts patient outcomes. Identifying patterns within diagnostic measurements through machine learning can augment conventional diagnostic methods and enable healthcare practitioners to intervene earlier, potentially preventing complications associated with diabetes.

Initial Hypotheses:

**Feature Significance:** Certain diagnostic measurements such as glucose levels, BMI, insulin, and blood pressure may have a substantial impact on predicting diabetes.

**Correlation Between Features:** There might exist correlations among the diagnostic measurements, where some features could have a strong interdependency in predicting the likelihood of diabetes.

**Machine Learning Model Performance:** Hypothesize that employing various machine learning algorithms will yield predictive models capable of accurately identifying diabetes based on the provided dataset.

Data Preparation:

**Data Inventory:**

Source: National Institute of Diabetes and Digestive and Kidney Diseases

Features: The dataset includes features such as Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI (Body Mass Index), Diabetes Pedigree Function, Age, and the target variable indicating diabetes presence or absence.

A screenshot of a computer

Description automatically generatedImporting necessary libraries and the dataset into jupyter file

A screenshot of a computer

Description automatically generated

Checking the corelation of all the features

we can see skin thickness, insulin, pregnancies and age are full independent to each other.

Age and pregnancies have negative correlation.

A screenshot of a computer screen

Description automatically generated

Plotting histogram of all the features

**Data Processing:**

* Summary Statistics: Basic statistical summaries for each feature, including mean, median, standard deviation, minimum, maximum, etc., to understand the distribution and range of values.
* Data Peek: Initial examination of a sample of the dataset to visually understand its structure and check for missing values, outliers, or anomalies.
* Data Transformation: Pre-processing steps may involve handling missing values, normalization of features, addressing outliers.

A computer screen shot of a computer code

Description automatically generated

Removing the outliers

Model Planning and Implementation

The primary goal of the data preparation phase is to gain a comprehensive understanding of the dataset, ensuring its quality and suitability for building a predictive model. This includes both descriptive statistical analysis and necessary data transformations to prepare it for machine learning model development.

The main plan is to predict **the presence or absence of diabetes** based on diagnostic measurements. Logistic Regression, Decision Trees, Random Forest, Support Vector Machines (SVM), Naivye Bayes, and Gradient Boosting are potential models due to their effectiveness in binary classification tasks and suitability for medical diagnostics. The accuracy of different classification models is compared in this project.

**Objective Alignment:** The chosen techniques directly support testing initial hypotheses:

* Logistic Regression can highlight feature significance and assess their impact on diabetes prediction.
* Decision Trees and Random Forest can reveal complex interactions among features, potentially validating correlations between diagnostic measurements.
* SVM and Gradient Boosting aim for higher predictive accuracy, providing insights into which model best fits the data for practical use.

Features and Target variable Selection

* A screenshot of a computer program

  Description automatically generatedWe must predict the presence of diabetes, so the target columns are outcome. The rest of the columns we could select in the features column.
* We use 80:20 ratio of training and test set.
* A cross validation function is defined to input actual and predicted value of the model.

Classifier Models:

1. **Logistic Regression**

A screenshot of a computer program

Description automatically generated

A screenshot of a computer screen

Description automatically generatedLinear regression model was implemented

Linear Regression model gave an accuracy of 78% with Receiver Operating Characteristic (roc) of 0.72

1. **A computer code on a dark background

   Description automatically generatedSVM (Support Vector Machine)**

A screen shot of a graph

Description automatically generatedSvm classifier model was implemented

1. **KNN**

**A screen shot of a computer program

Description automatically generated**

KNN had an accuracy of 87% with roc 0.84

1. **A screen shot of a graph

   Description automatically generatedRandom Forest classifier**

Random Forest classifier had an accuracy of 96% which is the best so far.

An accuracy of 96% achieved by a Random Forest classifier is quite promising and indicates that the model is performing well on our dataset.

A screenshot of a graph

Description automatically generatedResult:

After comparing all the classifiers, we could see The Random Forest gives the best output for the model. With 96% accuracy random forest performs best for the unseen data.