A DIABETES DETECTING EXPERT SYSTEM

Group Members

WHAT IS DIABETES?

According to the National Institute of Health (NIH), "Diabetes is a disease that occurs when your blood glucose, also called blood sugar, is too high. Blood glucose is your main source of energy and comes from the food you eat. Insulin, a hormone made by the pancreas, helps glucose from food get into your cells to be used for energy. Sometimes your body does not make enough—or any—insulin or does not use insulin well. Glucose then stays in your blood and does not reach your cells.

Over time, having too much glucose in your blood can cause health problems. Although diabetes has no cure, you can take steps to manage your diabetes and stay healthy. Sometimes people call diabetes “a touch of sugar” or “borderline diabetes.” These terms suggest that someone does not really have diabetes or has a less serious case, but every case of diabetes is serious.

TYPES OF DIABETES

The most common types of diabetes are type 1, type 2, and gestational diabetes.

* Type 1 diabetes: with type 1 diabetes, your body does not make insulin. Your immune system attacks and destroys the cells in your pancreas that make insulin. Type 1 diabetes is usually diagnosed in children and young adults, although it can appear at any age. People with type 1 diabetes need to take insulin every day to stay alive.
* Type 2 diabetes: with type 2 diabetes, your body does not make or use insulin well. You can develop type 2 diabetes at any age, even during childhood. However, this type of diabetes occurs most often in middle-aged and older people. Type 2 is the most common type of diabetes.
* Gestational diabetes: this develops in some women when they are pregnant. Most of the time, this type of diabetes goes away after the baby is born. However, if you have had gestational diabetes, you have a greater chance of developing type 2 diabetes later in life. Sometimes diabetes diagnosed during pregnancy is usually type 2 diabetes.

This project predicts diabetes among female Pima-Indians.

WHO IS A PIMA INDIAN?

"The Pima (or Akimel O'odham, also spelled Akimel O'otham, "River People", formerly known as Pima) are a group of Native Americans living in an area consisting of what is now central and southern Arizona. The majority population of the surviving two bands of the Akimel O'odham are based in two reservations: the Keli Akimel O'otham on the Gila River Indian Community (GRIC) and the On'k Akimel O'odham on the Salt River Pima-Maricopa Indian Community (SRPMIC)." Wikipedia

**Describe problem (Statement of the problem)**

Diabetes is a common, chronic disease. It dramatically increases the risk of various cardiovascular problems, including coronary artery disease with chest pain (angina), heart attack, stroke and narrowing of arteries (atherosclerosis). If you have diabetes, you are more likely to have heart disease or stroke. Hence, the prediction of diabetes at an early stage is very essential as it can lead to improved treatment.

The aim of this project is to develop an application that can quickly and efficiently determine if a person has diabetes. Rather than going to a hospital, this system can be used to check if someone is likely to have diabetes. This application system is backed up with data that has been analysed using an effective machine learning model.

**Background Information on the Expert System**

The expert system built is a case-based reasoning ES. This is an expert system that solves problems and makes decisions based on the solutions of similar past problems.

An ES system shell consists of:

Knowledge acquisition subsystem

Knowledge base- the data set

Inference mechanism

Explanation subsystem

User interface

**Data Description**

This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases and it was downloaded from Kaggle. It consists of 768 instances and 8 attributes. The objective of the dataset is diagnosis of diabetes of Pima Indians living in America. Out of the 768 instances, 268 are diabetic and 500 are non-diabetic and are described as 1 and 0 respectively in the target attribute (Outcome). In this database, all patients are females and are of age at least 21 years. Attributes in the database are either integers or real numbers.

Based on the various attributes provided in this data set, this paper shows whether a person is diabetes positive or not using Logistic Regression, Decision Tree and Random Forest.

**Data Attributes**

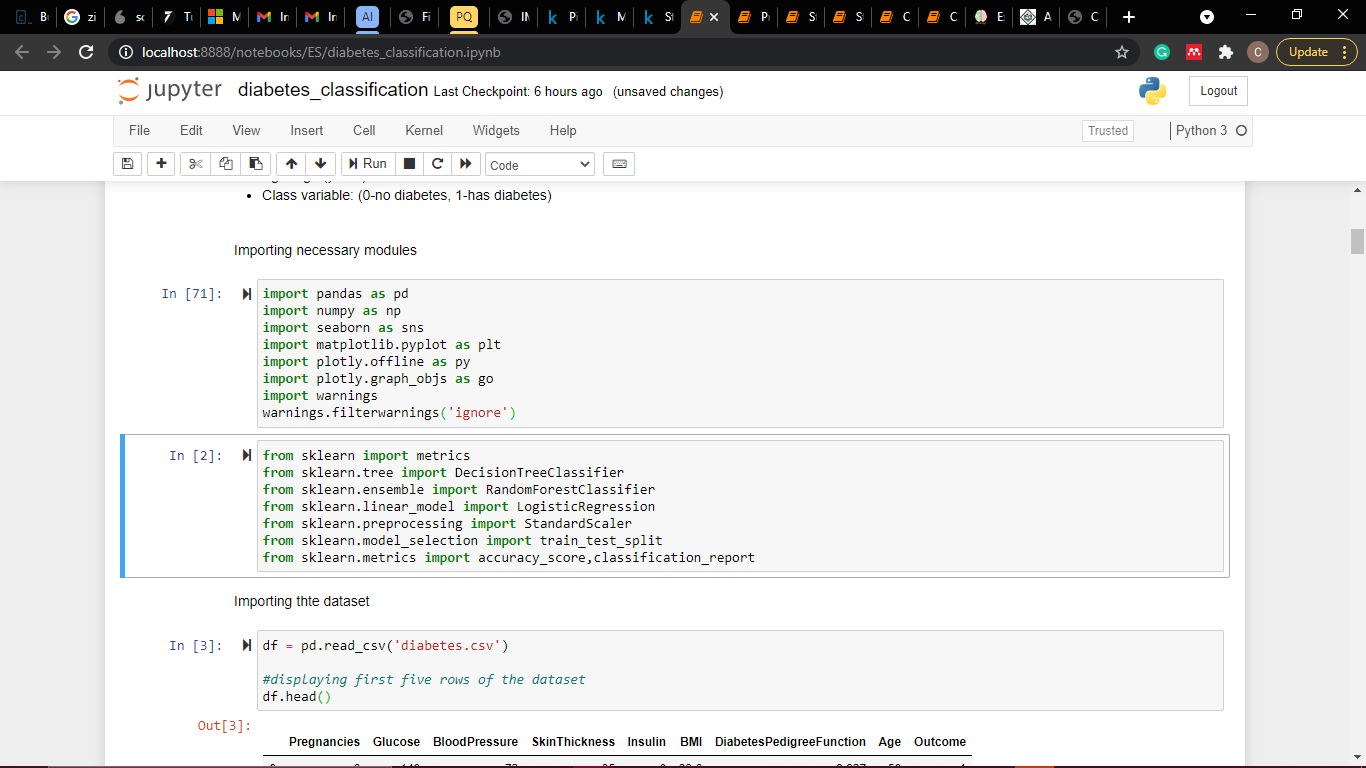
The following shows the available attributes in the dataset.

1. **Pregnancies:** The number of times the person was pregnant
2. **Glucose:** Plasma glucose concentration
3. **Blood Pressure:** Diastolic blood pressure (mm Hg)
4. **SkinThickness**: Triceps skin fold thickness (mm)
5. **Insulin:** 2-Hour serum insulin (mu U/ml)
6. **BMI:** Body mass index (weight in kg/ (height in m²))
7. **DiabetesPedigreeFunction:** Diabetes pedigree function
8. **Age:** Age (years)
9. **Outcome (target variable):** 0-no diabetes, 1-has diabetes.

**DATA PRE-PROCESSING STEPS**

Data pre-processing includes all necessary steps to transform the dataset into a format that is acceptable by the machine model that will be built. This could include missing value detection and imputation/removal, outlier detection, or scaling the variables. In this project, all pre-processing steps were carried out using Python programming language.

The first step was to import all necessary modules that would be used in building the model.



Next, the dataset is loaded, and the first five rows are displayed.

A picture containing text, screenshot, computer

Description automatically generated

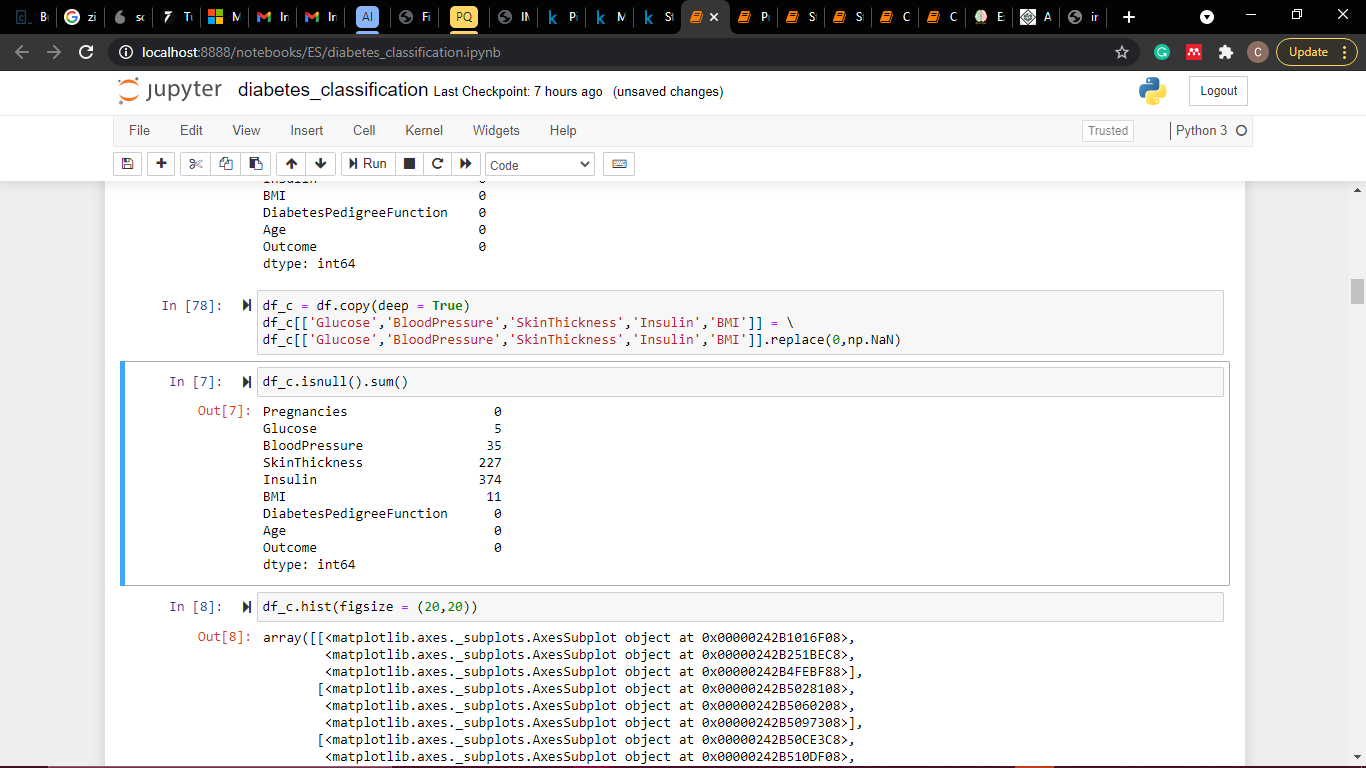
DataFrame.describe() method is called on the data frame to generate descriptive statistics that summarize the central tendency, dispersion, and shape of a dataset’s distribution, excluding NaN values. The describe() method deals only with numeric values and does not work with any categorical values. So if there are any categorical values in a column the describe() method will ignore it and display summary for the other columns unless parameter include="all" is passed. The statistics generated by the describe() method include:

A picture containing text, screenshot, computer, indoor

Description automatically generated

Count depicts the number of non-empty rows in a feature, mean shows the mean value of that feature, std gives the Standard Deviation Value of that feature, min shows the minimum value of that feature, max shows the maximum value of that feature, and 25%, 50%, and 75% are the percentile/quartile of each features. This quartile information helps to detect Outliers.

The next pre-processing step is to detect missing values. In this dataset, missing values were represented as 0 in some attributes (Glucose, BloodPressure, SkinThickness, Insulin, and BMI). To handle this, 0 was replaced with NaN then the missing values were imputed using either mean or median.



The image below shows the histogram distribution of the dataset before replacing null values.

Chart, box and whisker chart

Description automatically generated

Null values were then replaced with either mean or median. Then another histogram distribution was plotted.

Graphical user interface

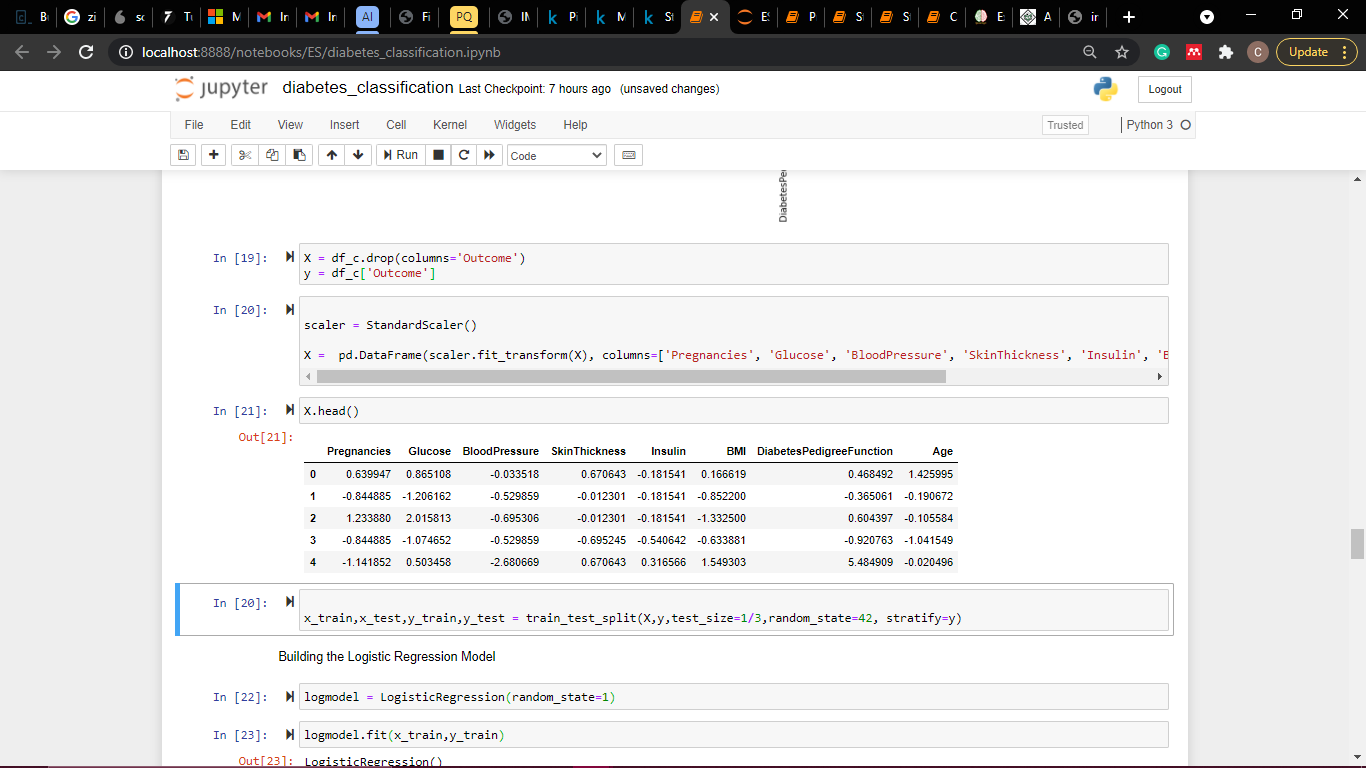
Description automatically generated

Chart, box and whisker chart

Description automatically generated

Finally, the dataset is divided into two data frames: X for the predictor variables and Y for the target variable. The predictor variables are then scaled using Scikit-learn’s StandardScaler() method The Standard Scaler assumes data is normally distributed within each feature and scales them such that the distribution centred around 0, with a standard deviation of 1. This is necessary to prevent higher-ranged values from dominating when calculating distances.

After scaling the dataset is split into a training set of 70% and a testing set of 30% using the train\_test\_split library from sklearn.



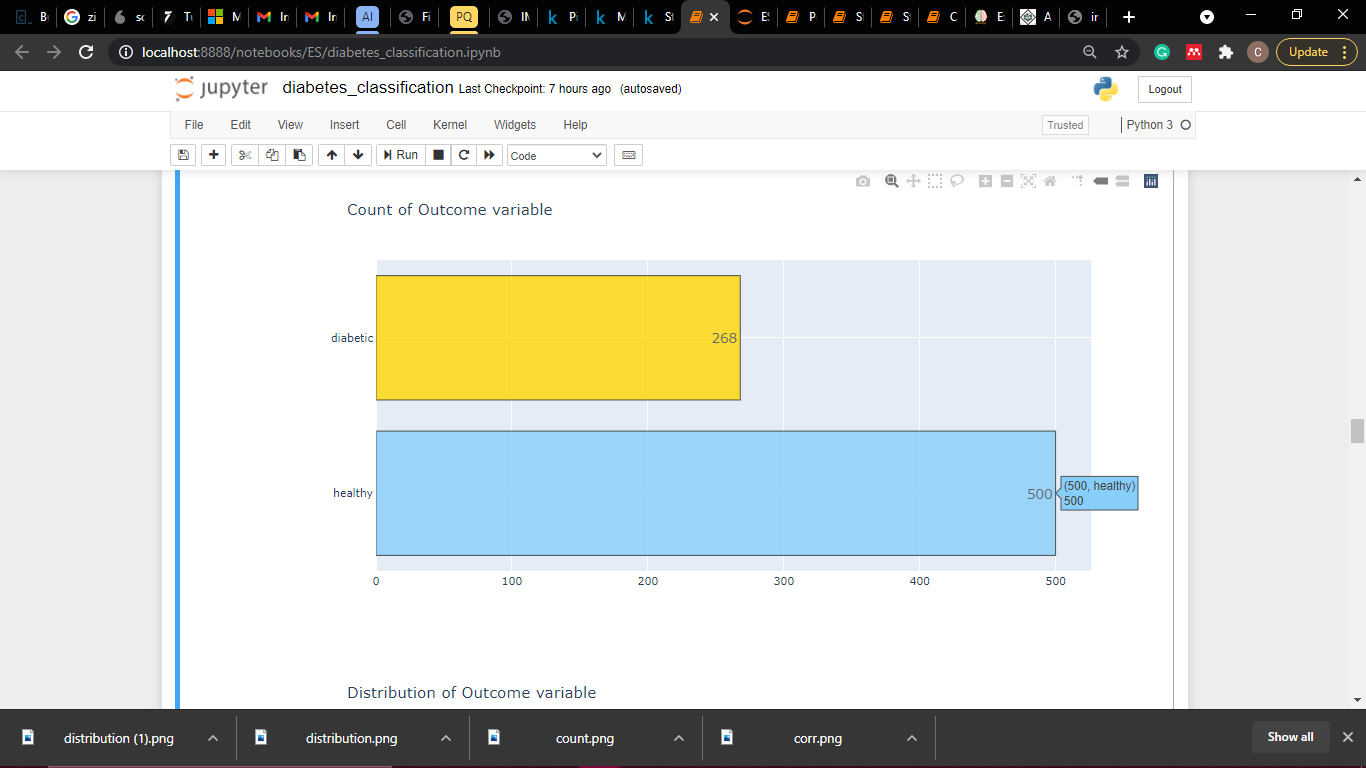
EXPLORATORY DATA ANALYSIS (EDA)

Exploratory Data Analysis refers to the process of performing initial investigations on data to discover patterns, to spot anomalies, to test hypothesis and to check assumptions with the help of summary statistics and graphical representations. It is a good practice to understand the data first and try to gather insights from it. EDA is all about making sense of data in hand, before building the model.

EDA was carried out using three visualization tools.

1. **Bar Chart**

This was used to show the number of people that are diabetic or healthy. Visualization revealed 500 healthy people and 268 diabetics.



1. **Pie Chart**

This was used to show the population of diabetic and healthy people. Visualization revealed 65.1% of the population are healthy and the remaining 34.9% are diabetic.

**Chart, pie chart

Description automatically generated**

1. Correlation Matrix

A correlation matrix is used to show the correlation coefficients. Each cell in the table shows the correlation between two variables. A correlation coefficient of 0.5 to 1 is considered high. From the plot below, glucose, BMI, age, pregnancy, and skin thickness have the highest correlation to the outcome variable.

A picture containing chart

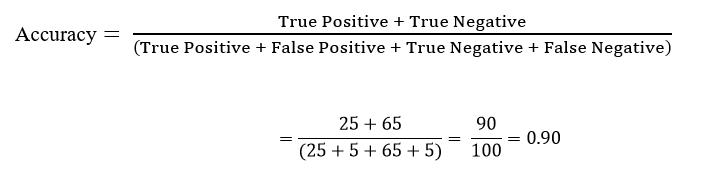
Description automatically generated

METHODS/TECHNIQUES USED

MODEL DEVELOPMENT

To build the model, three machine learning classifiers are trained, tested, and evaluated on the dataset. The best one is chosen based on accuracy and f1-score and deployed on the expert system application system.

The accuracy score is a measure of all the correctly identified cases. It is mostly used when all the classes are equally important. It is measures as shown below.



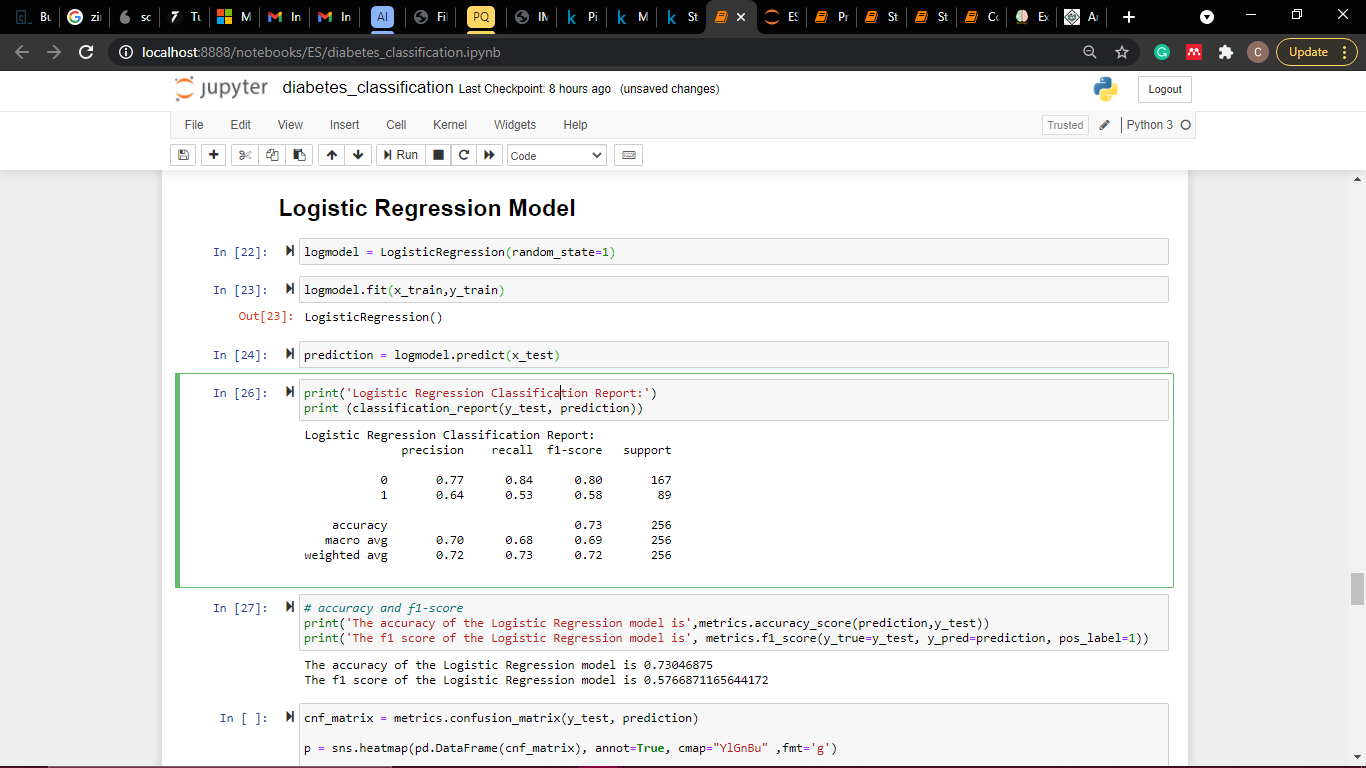
F1-score is the harmonic mean of precision and recall and gives a better measure of the incorrectly classified cases than the accuracy metric. It is measures as shown below.

Diagram

Description automatically generated

The classifiers used include:

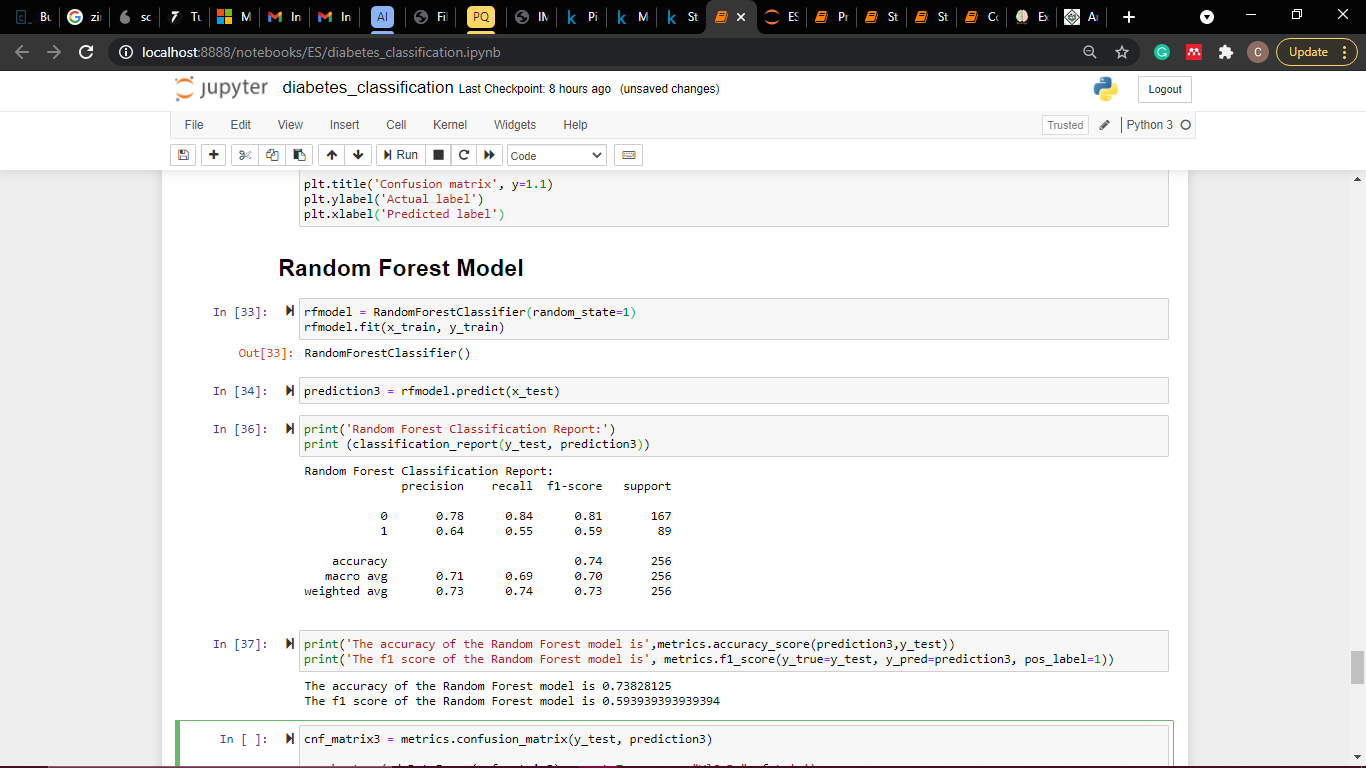
1. **Logistic Regression:** is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval, or ratio-level independent variables.



1. **Decision Tree:** aims to create a training model that can use to predict the class or value of the target variable by learning simple decision rules inferred from prior data (training data).

**Graphical user interface, text, application

Description automatically generated**

1. **Random Forest:** a classifier that contains several decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset. ****

The best classifier score is the Random Forest classifier with accuracy score of 73.8% and f1-score of 59.4%.

Graphical user interface, text, application

Description automatically generated

Conclusion

Graphical user interface

Description automatically generated