Contents

Chapter 1

Introduction

Diabetes caused because of obesity or high blood glucose level, and so forth. It affects the hormone insulin, resulting in abnormal metabolism of crabs and improves level of sugar in the blood. Diabetes occurs when body does not make enough insulin. The annual report of World Health Association, add up to the number of individuals experiencing diabetes is 422 million the year. Consistently, there is a significant increment in the number individuals experiencing diabetes in different healing centre.
There are different purposes behind reason like a way of life of a man, the absence of activity, sustenance propensities, heftiness, smoking, high cholesterol (Hyperlipidaemia), high blood pressure (Hyperglycaemia) etc. which fundamentally increment the risk of treating diabetes. It influences a wide range of ages, including youngsters to grown-up and matured people.

Whenever the glucose, or sugar level is high in the circulatory system, Beta cells of pancreas discharges the insulin to the circulation system, to assimilate the exorbitant sugar substance from the blood into liver, later it is changed over into a frame vitality. Similarly, at whatever point the glucose level is low, the creation of insulin is occupied and generating of glucagon by the alpha cells of the pancreas will be started to keep up the glucose level in the blood. The admission sugar in the body likewise assumes an imperative job in diabetes

Diabetes is a long-haul issue, many hazard factors, intricacies, expand passing's rates. It is arranged into four kind's type-1 , type-2 , prediabetes , and gestational diabetes.

Type-1 A serious, incessant illness happens frequently happens in youngsters and grownups. Here pancreas totally stops the creation of the insulin. The individual assaulted by Type 1 is totally subject to insulin from outer drugs to control the sugar levels in the body. The DCCT (Diabetes Control and intricacies trail) assisted the individual through the rundown solutions with being taken after to keep away from the symptoms, extreme difficulties on different organs and live longer better life through the rules and sustenance propensities. A dietary methodology was found through these rules.

Type-2 It is a class of perpetual; non-insulin subordinate sickness regularly happens in grownups. There are a few realities of the events of sort 2 are hereditary and metabolic components, family history, physical dormancy overweight, heftiness, undesirable eating regimen, smoking propensities expands the danger of diabetes.

Prediabetes It is a phase before type 2 diabetes, where glucose level of the individual has been higher than typical yet not to the levels of sort 2. A man with prediabetes condition has more odds of getting compose 2 under specific conditions and measures.

Gestational It is a basic classification influenced for ladies amid pregnancy. A variety of hormones amid pregnancy and expanded insulin substance can prompt the high blood glucose level. The newly conceived babies have the odds of creating diabetes. The dietary propensities to diminish the level of diabetes.

Diabetes is influenced by different parts of the body which incorporates

a. *Loss of vision* Retinopathy retina is a condition where the retina, optic nerve, the focal point is harmed. A result of finish night visual impairment issues, swelling in the region of the retina, lessening the contact the mind may happen. A Diabetic individual should deal with eye vision through a few tests and pharmaceutical at the beginning times. The treatment incorporates visual sharpness testing, tonometry, student enlargement, and optic intelligibility tomography (OCT). Different medicines incorporate Anti-VEGF infusion therapy, focal/lattice macular laser medical procedure, corticosteroid.

b. *Kidney neuropathy* Chronic kidney infection or diabetic neuropathy is where the high sugar level in blood harms the vessels in the kidney. The usefulness of the kidney is to channel the waste and abundant water in the blood. Because of hypertension and sugar level in Kidney endeavours to have overhead to clean the blood this may prompt kidney disappointment or successive dialysis of blood is required. The treatment may incorporate kidney substitution treatment, kidney and pancreas transplant.

c. *Liver problems* Liver assumes an indispensable job in adjusting the blood glucose level in blood through starch digestion by methods neoglucogenesis and glycogenosis's. Sort 2 diabetes expands the danger of liver issues. Fatty liver assumes the stipulate job in creating a liver tumour. The difficulties incorporate renal debilitation, modified metabolism, Insulin opposition and hyperglycaemia, malnutrition. Affect individual needs to experience different anti-toxin drugs and administration of liver incorporates other treatment like the way of life alteration, pharmacological treatment, insulin secretagogues, biguanides, α-glucosidase inhibitors, TZDs, weight to decrease.

d. *Heart problems* Cardiovascular ailment: According to American heart affiliation, 68% of individuals will experience the ill effects of heart issues to driving even to death, heart stroke, atherosclerosis or solidifying of the supply routes, stress and load on the heart make individual to death. Because of high sugar level, blood conveys greater thickness, it adheres to the veins, supply routes and veins put more strain to proceed onward. Persistently it harms the vessels and nerves prompting disappointment of circulatory framework or organ disappointment in person. Hazard for creating cardiovascular illness incorporates hypertension, unusual cholesterol and high triglycerides, corpulence, the absence of physical activity. The effect of different clinical parameters like poor glycaemic control, insulin opposition of diabetes greatly affects heart issues.

e. The different issues may incorporate foot issues and so on.

Chapter 2

Basic Concepts

2.1 Data mining and Classification

Information mining is a ground-breaking procedure with a huge measurement of a data set where the data set is extremely large, huge in terms of type, to remove data that is useful for deciding on company selection or finding new solutions. comparison examples to decide better choice. It is used to find new examples, find comparable links between information, correlate information, it can find answers to problems, create rules from old information, addressing commercial ad placement best choices, finding hidden information designs from datasets, future performance expectations, i.e., practices and models.

2.2 Data Pre-processing-

Data pre-processing is the process of preparing raw data and making it suitable for machine learning models. This is the first critical step in creating a machine learning model.

When creating machine learning projects, you don't always come across clean and formatted data. Also, it is imperative to store the data in a clean and formatted way every time you work with it. For this, we use a data pre-processing task.

2.3 Missing values Removal-

Remove all the instances that have zero (0) as worth. Having zero as worth is not possible. Therefore, this instance is eliminated. Through eliminating irrelevant features/instances we make feature subset and this process is called features subset selection, which reduces dimensionality of data and help to work faster

2.4 Splitting of Data-

Data is standardised for the model's training and testing once it has been cleaned. After the data is spilt, we train the algorithm on the training data set while putting the test data aside. Based on the logic, methods, and values of the feature in the training data, this training process will generate the training model. Basically, the purpose of normalisation is to scale up all the attributes.

2.5 Applying Machine learning techniques

When data has been ready, we apply Machine Learning Technique. We use different classification and ensemble techniques, to predict diabetes.

The methods applied on Pima Indians diabetes dataset.

Main objective to apply Machine Learning Techniques to analyse the performance of these methods and find accuracy of them, and also been able to figure out the responsible/important feature which play a major role in prediction.

Chapter 3

Problem Statement / Requirement Specifications

3.1 Project Planning

our objective is to predict whether the patient has diabetes or not based on various features like *Glucose level, Insulin, Age, BMI*. We will perform all the steps from *Data gathering to Model deployment.* During Model evaluation, we compare various machine learning algorithms on the basis of accuracy_score metric and find the best one. Then we create a web app using Flask which is a python micro framework.

3.2 Project Analysis

We will use the Pima Indians dataset from the UCI Machine learning repository. We will develop this project in six steps which follows data gathering to model deployment.

3.3 System Design

All the standard libraries like numpy, pandas, matplotlib and seaborn are imported in this step. We use numpy for linear algebra operations, pandas for using data frames, matplotlib and seaborn for plotting graphs.

Chapter 4

Implementation Methods:

4.1 Support vector machine

It is a supervised learning, discriminative classification [42, 43] technique. This method can be used for both regression and classification. The logic behind the SVM is finding a hyper line between the dataset, which best divides the dataset into two classes.

It includes 2 steps, identifies the right or optimal hyper line in data space and mapping the objects to the boundaries specified. The SVM training algorithm builds a model that assigns new samples to one of the classes.

4.2 K nearest neighbor (KNN)

It is a classification technique which classifies the new sample based on similarity measure or distance measure. The measure includes 3 distance measures Euclidean distance, Manhattan, Minkowski. The steps for KNN is given below.

1.Training phase of the algorithm consists of only storing the feature sample and class label of training sample.

2.Classification phase: the user has to define a “k” value for the classification of the undefined sample for the k number of the class labels, so the unlabelled sample can be classified into the defined class based on the feature similarity.

3.Majority of voting classification occurs for unlabelled class.

The value of the k can be selected by various techniques like heuristic technique.

4.3 Logistic Regression

Logistic regression is also a supervised learning classification algorithm. It is used to estimate the probability of a binary response based on one or more predictors. They can be continuous or discrete. Logistic regression used when we want to classify or distinguish some data items into categories.

It classifies the data in binary form means only in 0 and 1 which refer case to classify patient that is positive or negative for diabetes.

Main aim of logistic regression is to best fit which is responsible for describing the relationship between target and predictor variable. Logistic regression is a based on Linear regression model. Logistic regression model uses sigmoid function to predict probability of positive and negative class

4.4 Decision Tree

Decision Tree is a supervised learning method that may be used to classification and regression issues; however, it is most frequently used to address classification issues. It is a tree-structured classifier, where internal nodes stand in for the dataset's characteristics, branches for the rules of classification, and each leaf node for the result.

The Decision Node and Leaf Node are the two nodes of a decision tree. While Leaf nodes are the results of decisions and do not have any more branches, Decision nodes are used to create decisions and have numerous branches.

It is known as a decision tree because, like a tree, it begins with the root node and expands on successive branches to form a network like a tree.

4.5 Random Forest

A random forest is a classifier that takes a set of decision trees over different subsets of a given dataset and takes an average to improve the prediction accuracy of that dataset. Instead of relying on decision trees, random forests get predictions from each tree. Predict the final output based on the majority vote of the predictions.

The greater the number of trees in the forest, the improved the accuracy and the avoidance of overfitting problems.

4.6 Naive Bayes

The naive Bayes algorithm is a supervised learning algorithm based on Bayes' theorem and is used to solve classification problems.

It is primarily used in text classification with high-dimensional training datasets.

Naive Bayes Classifier is one of the simplest and most effective classification algorithms that help you build fast machine learning models that can make fast predictions.

It's a probabilistic classifier, meaning it makes predictions based on object probabilities.

Chapter 5

Standards Adopted

5.1 Design Standards

In all the engineering streams, there are predefined design standards are present such as IEEE, ISO etc. List all the recommended practices for project design. In software the UML diagrams or database design standards also can be followed.

5.2 Coding Standards

Used proper naming and formatting (indentations, braces, one statement per line, line wrapping wherever needed etc.) standards in variable, methods, interfaces, class and packages naming: Used CamelCase while declaring variables, beginning names with lowercase letters, variable name gives proper insights, what it is used for.

- Used underscores while naming lengthy variable names.
- Avoided unnecessary and redundant initializations of variables, methods, objects etc.
- Avoided memory leaks.
- Took proper actions if encountered any exceptions.

5.3 Testing Standards

Software testing standards are equally important as it holds validation for its user as well as the company to make sure software is meeting certain criteria. The basic goals of these standards are to make sure that user is happy and testing company is getting positive remarks.

The following standards were followed during quality assurance and testing of products:

➢ ISO / IEC / IEEE 29119 - 1

➢ ISO / IEC / IEEE 29119 - 2

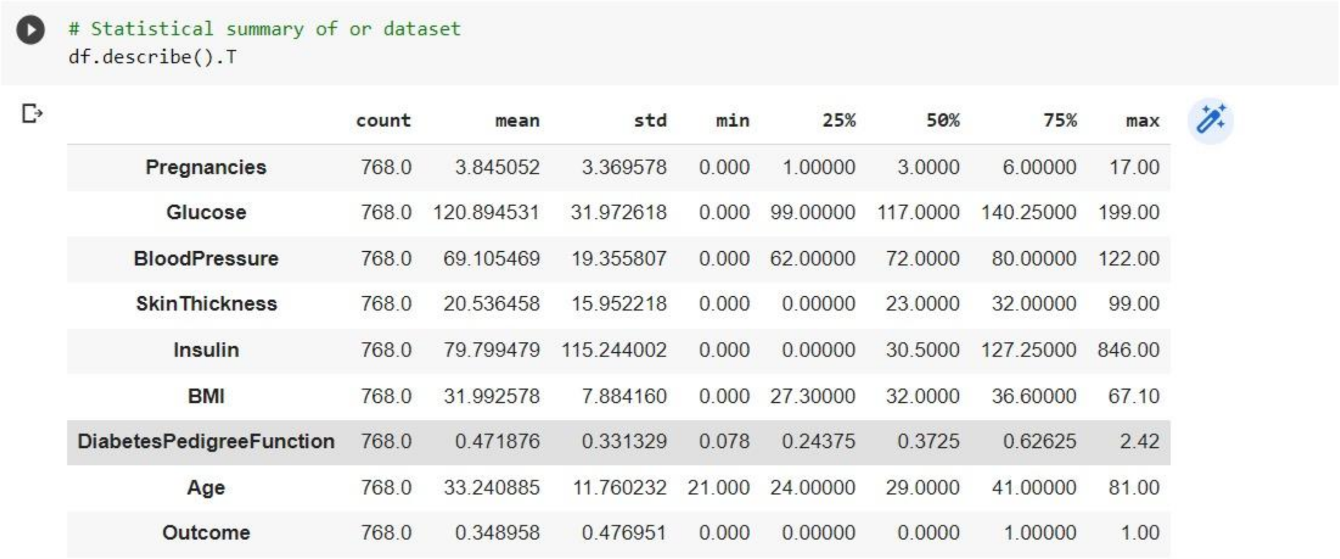
➢ ISO / IEC / IEEE 29119 - 3

➢ ISO / IEC / IEEE 29119 - 4

➢ ISO / IEC / IEEE 29119 - 5

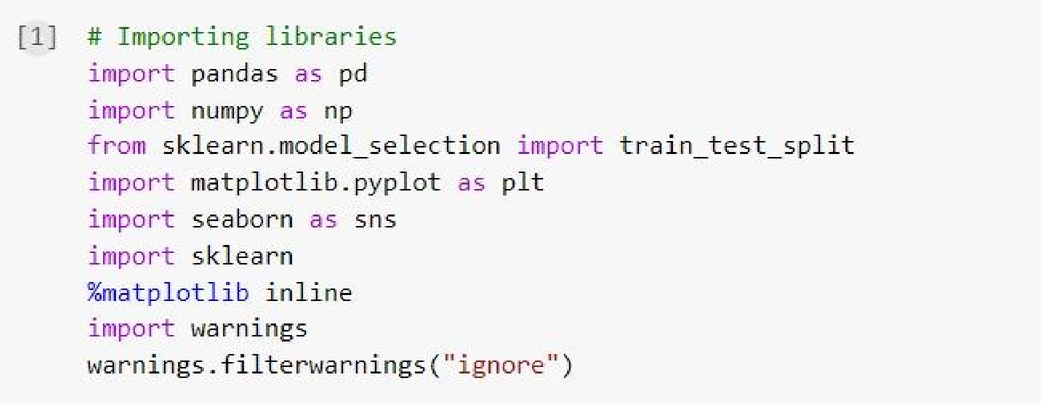
Chapter 6

Code Snapshot summary

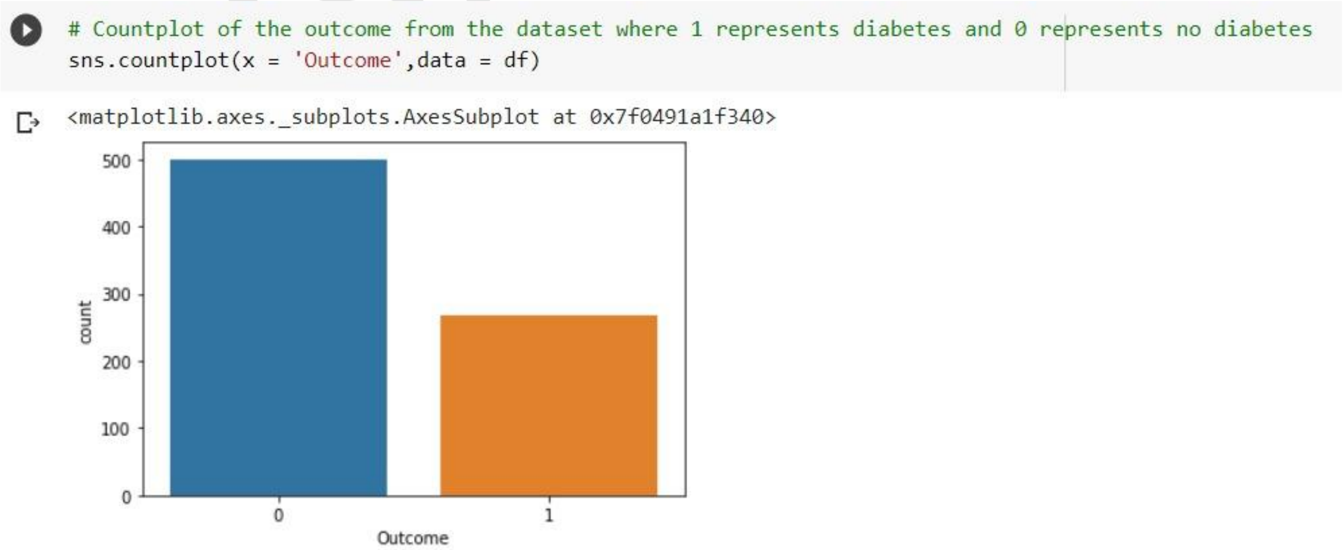


```
# Statistical summary of or dataset
df.describe().T
```

	count	mean	std	min	25%	50%	75%	max
Pregnancies	768.0	3.845052	3.369578	0.000	1.00000	3.0000	6.00000	17.00
Glucose	768.0	120.894531	31.972618	0.000	99.00000	117.0000	140.25000	199.00
BloodPressure	768.0	69.105469	19.355807	0.000	62.00000	72.0000	80.00000	122.00
SkinThickness	768.0	20.536458	15.952218	0.000	0.00000	23.0000	32.00000	99.00
Insulin	768.0	79.799479	115.244002	0.000	0.00000	30.5000	127.25000	846.00
BMI	768.0	31.992578	7.884160	0.000	27.30000	32.0000	36.60000	67.10
DiabetesPedigreeFunction	768.0	0.471876	0.331329	0.078	0.24375	0.3725	0.62625	2.42
Age	768.0	33.240885	11.760232	21.000	24.00000	29.0000	41.00000	81.00
Outcome	768.0	0.348958	0.476951	0.000	0.00000	0.0000	1.00000	1.00

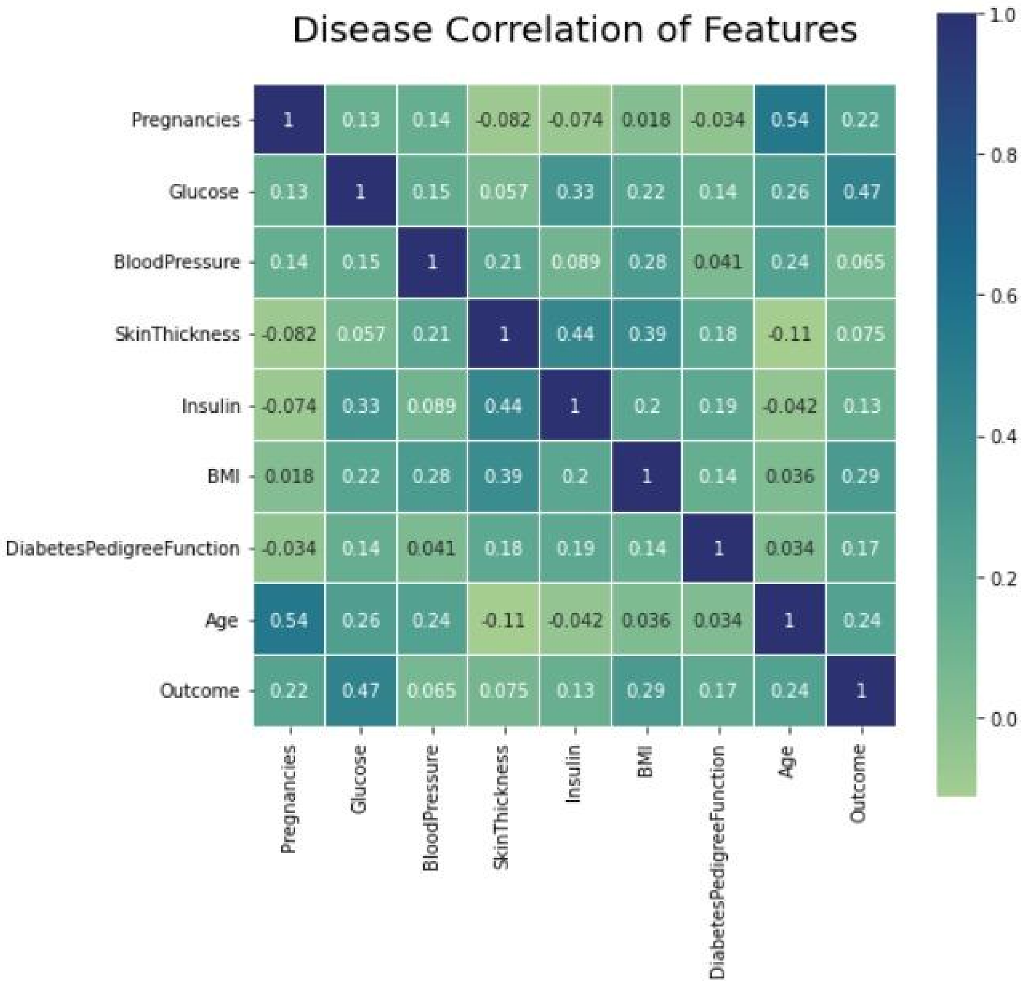


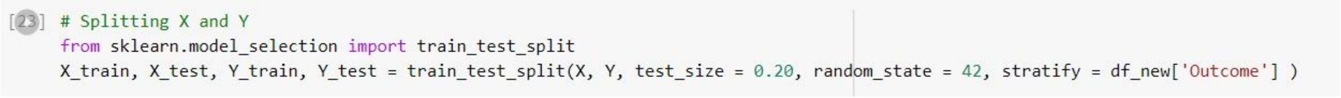
```
[1] # Importing libraries
    import pandas as pd
    import numpy as np
    from sklearn.model_selection import train_test_split
    import matplotlib.pyplot as plt
    import seaborn as sns
    import sklearn
    %matplotlib inline
    import warnings
    warnings.filterwarnings("ignore")
```



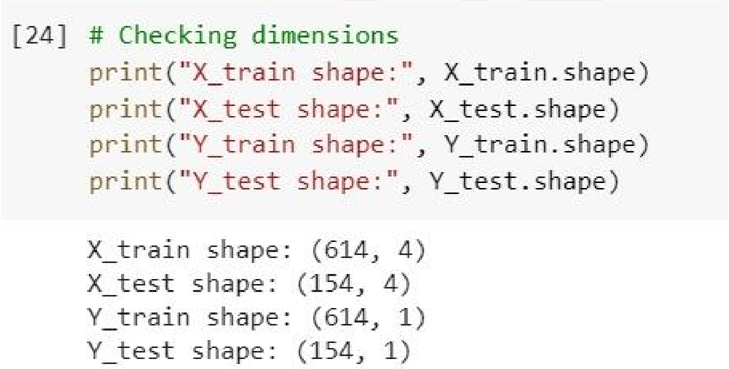
```
# Countplot of the outcome from the dataset where 1 represents diabetes and 0 represents no diabetes
sns.countplot(x = 'Outcome',data = df)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f0491a1f340>
```

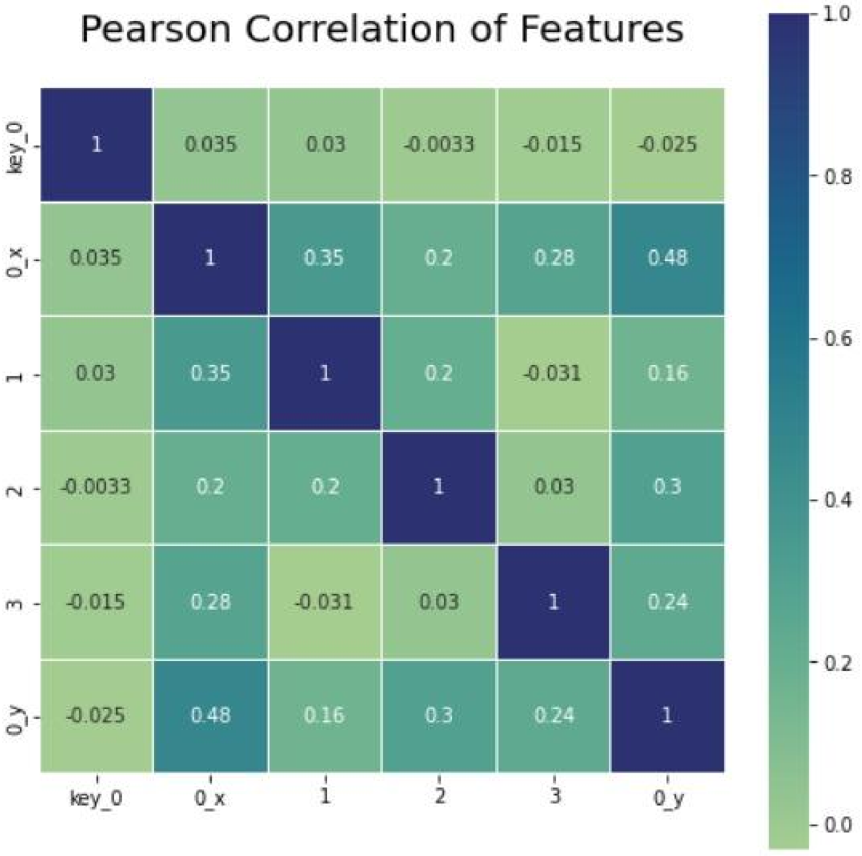


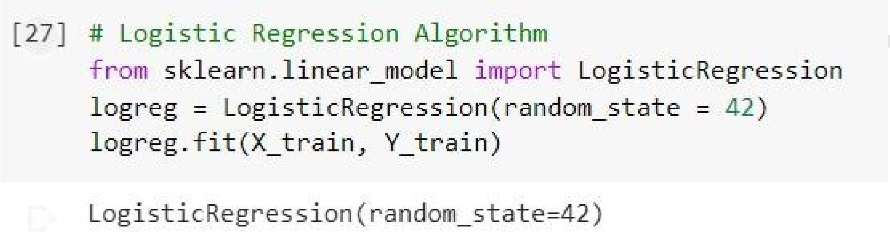


```
[23] # Splitting X and Y
     from sklearn.model_selection import train_test_split
     X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size = 0.20, random_state = 42, stratify = df_new['Outcome'] )
```

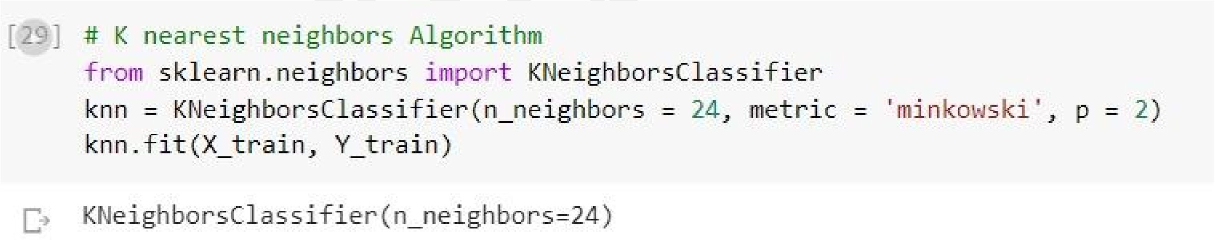


```
[24] # Checking dimensions
     print("X_train shape:", X_train.shape)
     print("X_test shape:", X_test.shape)
     print("Y_train shape:", Y_train.shape)
     print("Y_test shape:", Y_test.shape)

     X_train shape: (614, 4)
     X_test shape: (154, 4)
     Y_train shape: (614, 1)
     Y_test shape: (154, 1)
```

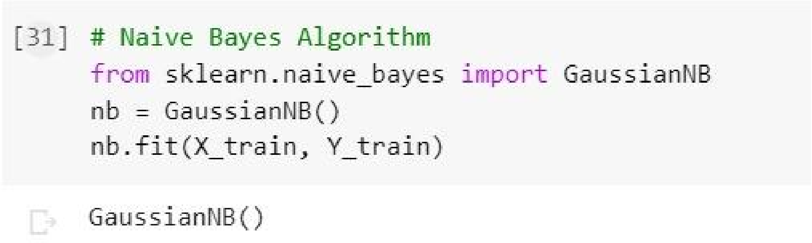




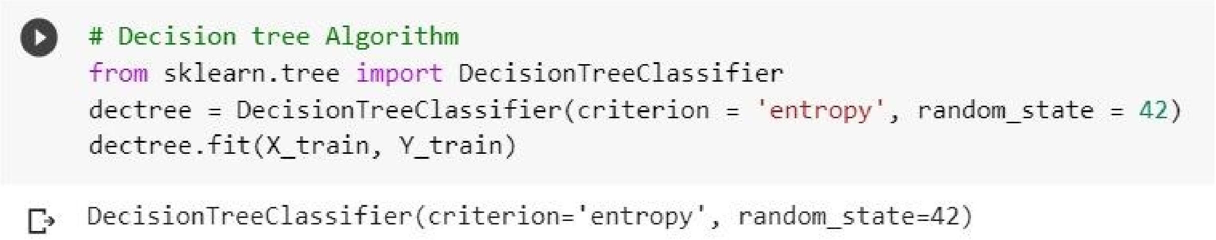
```
[27] # Logistic Regression Algorithm
     from sklearn.linear_model import LogisticRegression
     logreg = LogisticRegression(random_state = 42)
     logreg.fit(X_train, Y_train)

     LogisticRegression(random_state=42)
```



```
[29] # K nearest neighbors Algorithm
     from sklearn.neighbors import KNeighborsClassifier
     knn = KNeighborsClassifier(n_neighbors = 24, metric = 'minkowski', p = 2)
     knn.fit(X_train, Y_train)

     KNeighborsClassifier(n_neighbors=24)
```

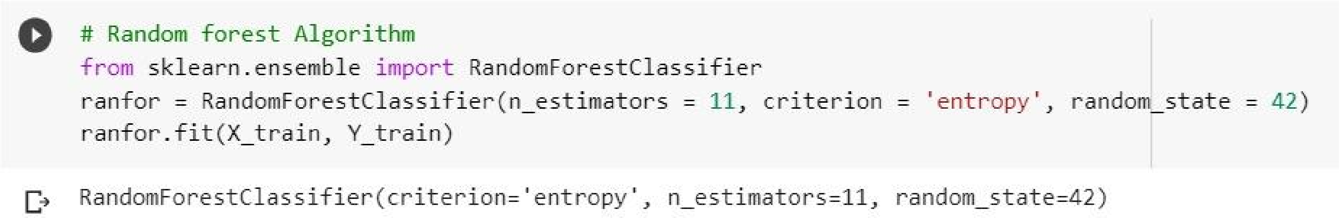


```
[31] # Naive Bayes Algorithm
     from sklearn.naive_bayes import GaussianNB
     nb = GaussianNB()
     nb.fit(X_train, Y_train)

     GaussianNB()
```



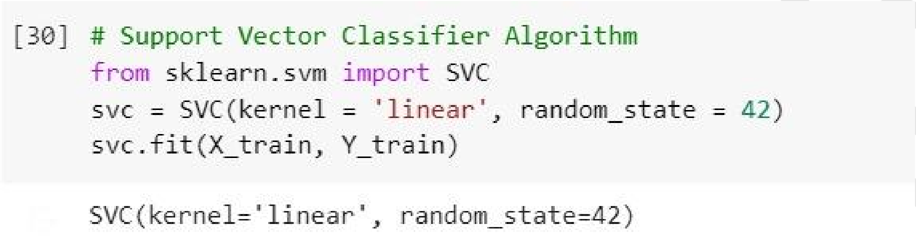
```
# Decision tree Algorithm
from sklearn.tree import DecisionTreeClassifier
dectree = DecisionTreeClassifier(criterion = 'entropy', random_state = 42)
dectree.fit(X_train, Y_train)
```

```
DecisionTreeClassifier(criterion='entropy', random_state=42)
```



```
# Random forest Algorithm
from sklearn.ensemble import RandomForestClassifier
ranfor = RandomForestClassifier(n_estimators = 11, criterion = 'entropy', random_state = 42)
ranfor.fit(X_train, Y_train)
```

```
RandomForestClassifier(criterion='entropy', n_estimators=11, random_state=42)
```



```
[30] # Support Vector Classifier Algorithm
     from sklearn.svm import SVC
     svc = SVC(kernel = 'linear', random_state = 42)
     svc.fit(X_train, Y_train)
```

```
SVC(kernel='linear', random_state=42)
```

Chapter 7

Conclusion & Future Scope

7.1 Conclusion

The project entitled Diabetes Prediction was completed successfully.The system has been developed with much care and free of errors and at the same time it is efficient and less time consuming as well as handy to use. The purpose of this project was to develop a ML application to predict Diabetes using the currently available vast dataset.This project helped us in gaining valuable information and practical knowledge on several topics like Google Co-laboratory, Python, different algorithms such as Linear regression, SVM, KNN, Random Forest, Decision Tree and Naive Bayes algorithm.The entire system is secured. Also the project helped us understand about the different development phases of a project. We did also learn how to test different features of a project.This project has overall given us great satisfaction in having designed an application which can be implemented to any nearby medical shops or branded shops or even big companies or products by simple modifications or using more technologies like HTML, CSS, React etc to make a full fledged app, with this prediction technique in the back-end.

7.2 Future Scope

This model could be further modified so that it works for an average citizen as well by improvising it as an application based software which can be accessed easily via playstore, appstore, etc. A B2C model can serve a large segment of the population. Many surveys will have to be conducted in order to understand the basic needs and desires of the common man and accordingly we can make modifications to it so as to meet their demands this can make it a potential market product that can be very helpful in medical and research industries. The aim is to take this application forward and keep on upgrading it as time goes by, to increase its performance and make it even more robust and powerful so as to make it reliable while handling large volume of global data.

The End