1. **Introduction**
   1. **Problem Statement**

Diabetes has become very prevalent today causing serious harm to people’s health. Its protection and prevention have been the center of medical research for decades now. To ensure a healthy lifestyle, it becomes important that diabetes is diagnosed as early as possible. In this machine learning-based program, we are detecting diabetes in patients based on predefined characteristics.

* 1. **Objective**
* The objective of the study is to classify the Indian PIMA dataset for diabetes.
* Predict if a person is a diabetes patient or not
* Find the most indicative features of diabetes

1. **Background Study**

**2.1 Data and preprocessing**

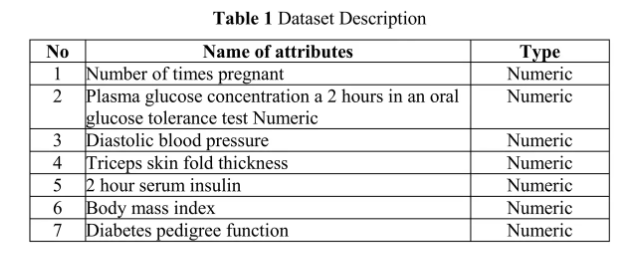
Data preprocessing is the most important process. Most healthcare-related data contains missing vale and other impurities that can cause the effectiveness of data. To improve the quality and effectiveness obtained after the mining process, Data preprocessing is done. To use Machine Learning Techniques on the dataset effectively this process is essential for accurate results and successful prediction. For Pima Indian diabetes dataset we need to perform pre-processing in two steps.

* 1. 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 a feature subset and this process is called features subset selection, which reduces the dimensionality of data and helps to work faster.
  2. Splitting of data- After cleaning the data, data is normalized in training and testing the model. When data is spitted then we train the algorithm on the training data set and keep the test data set aside. This training process will produce the training model based on logic and algorithms and values of the feature in training data. Basically aim of normalization is to bring all the attributes under the same scale.

1. **Detailed Design**

**3.1 Dataset used**

In this work, datasets have been collected among the Pima Indian female population near Phoenix, Arizona. This particular dataset has been widely used in machine learning experiments and is currently available through the UCI repository of standard datasets. This population has been studied continuously by the National Institute of Diabetes, Digestive, and Kidney. UCI repository contains 768 instances of observations and a total of 9 attributes with no missing values reported. Data sets contain 8 particular variables which were considered high-risk factors for the occurrence of diabetes, like the number of times pregnant, plasma glucose concentration at 2 hours in an oral glucose tolerance test (OGTT), diastolic blood pressure, 2-hour serum insulin, body mass index, diabetes pedigree. All the patients in this dataset are females at least 21 years old living near Phoenix, Arizona. All attributes are numeric values except class is a nominal type. Attributes name and types are shown in table 1.

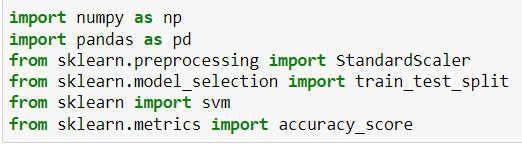


**3.2 Algorithm**

Support Vector Machine also known as SVM is a supervised machine learning algorithm. SVM is the most popular classification technique. SVM creates a hyperplane that separates two classes. It can create a hyperplane or set of hyperplanes in high-dimensional space. This hyperplane can be used for classification or regression also. SVM differentiates instances in specific classes and can also classify the entities which are not supported by data. Separation is done through hyperplane performs the separation to the closest training point of any class.

1. **Implementation and Details**

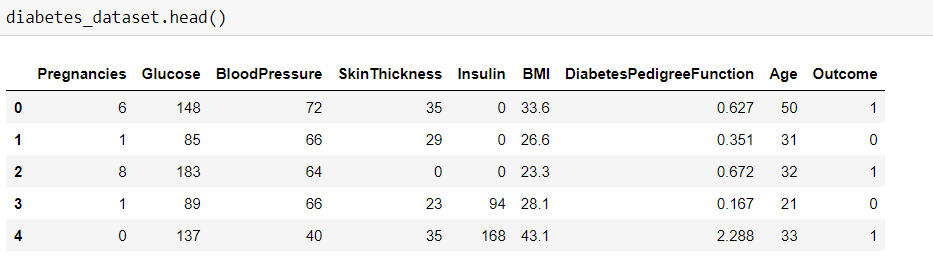
**Importing libraries:**

****

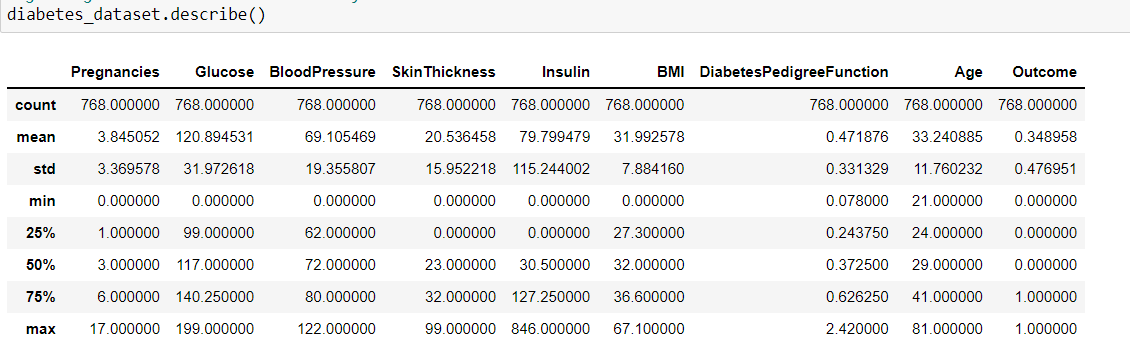
**Reading csv file:**

****

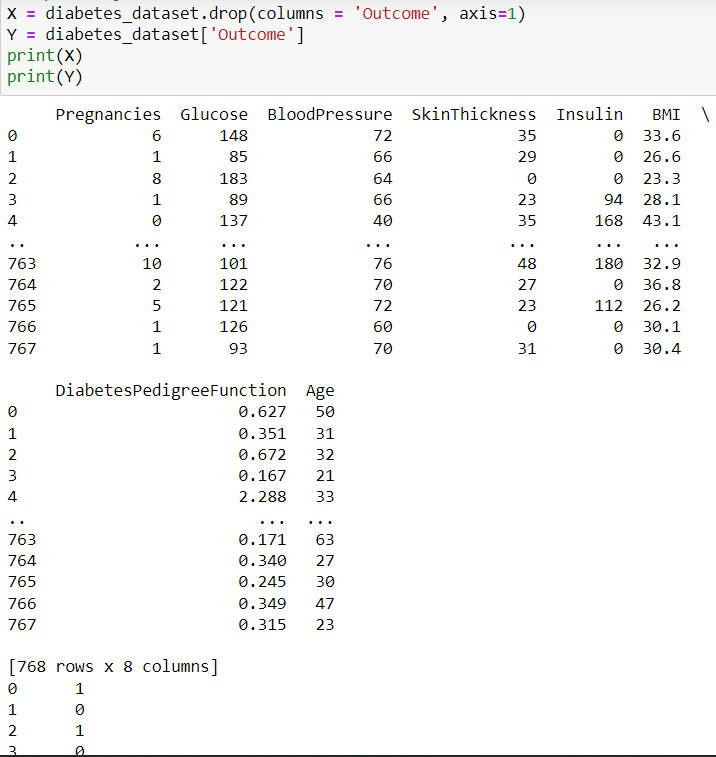
**Printing first 5 data:**

****

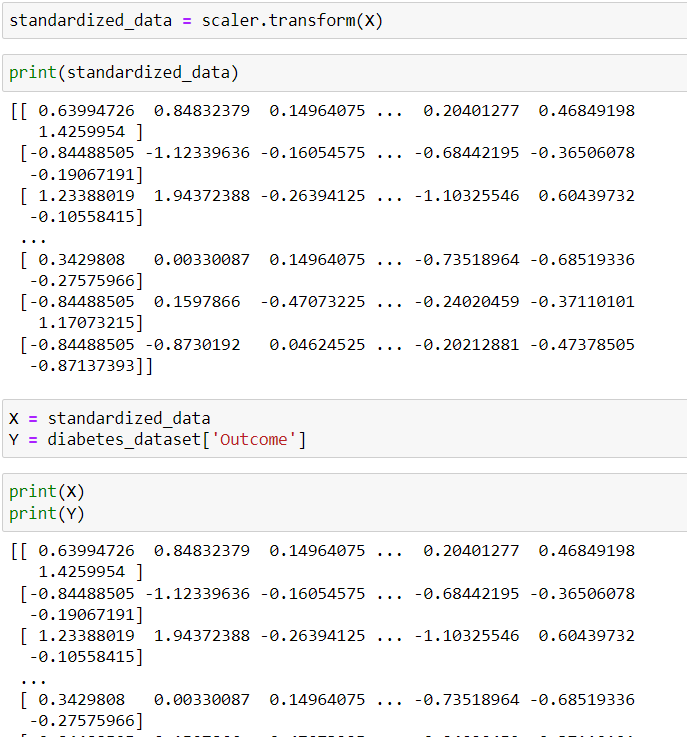
**Getting the statistical measure of the data:**

****

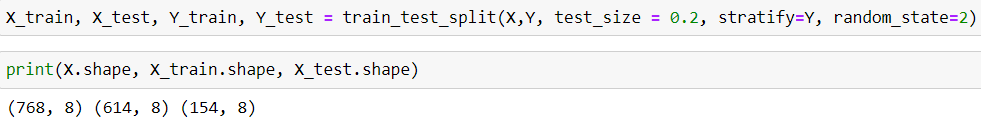
**Separating data and labels:**

****

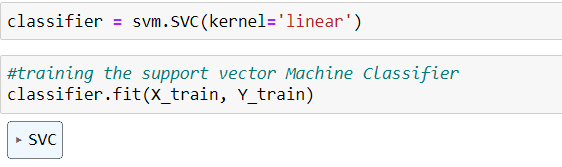
**Data Standardization:**

****

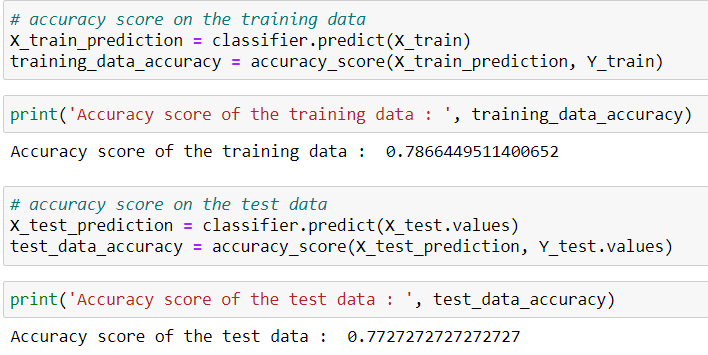
**Splitting training and test data:**

****

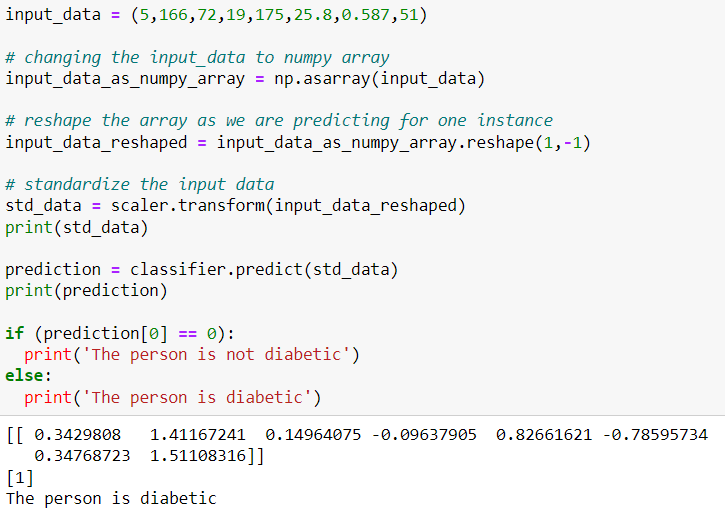
**Training the model:**

****

**Accuracy score of training and test data:**

****

**Making a prediction model:**

****

1. **Conclusion**

* Diabetes is one of the risks during Pregnancy. It has to be treated to avoid complications.
* BMI index can help to avoid complications of diabetes away before
* Diabetes start showing at the age of 35 – 40 and increase with a person’s age.

1. **References**

<https://www.slideshare.net/dataalcott/diabetes-prediction-using-machine-learning-250054453>

<https://medium.com/geekculture/diabetes-prediction-using-machine-learning-python-23fc98125d8>

<https://www.ijert.org/diabetes-prediction-using-machine-learning-techniques>