**PROJECT REPORT**

Abstract:

--- Abhijeet Dutta

**Diabetes mellitus is the most common disease worldwide and keeps increasing everyday due to the ever changing lifestyle, unhealthy food habits and over weight problems.**

**Data Science methods have the potential to benefit other scientific fields by shedding new light on common questions.**

**In the proposed system, an efficient way of detecting diabetes is proposed through machine learning.**

**The experiment results show that the prediction is done at a high accuracy.**

Introduction:

**Diabetes is noxious diseases in the world. Diabetes is caused because of obesity or high blood glucose level, and so forth. It affects the hormone insulin, resulting in abnormal metabolism of carbs and increases level of sugar in the blood. Diabetes occurs when body does not make enough insulin. According to (WHO) World Health Organization about 422 million people suffering from diabetes particularly from low or idle income countries. And this could be increased to 490 billion up to the year of 2030. However prevalence of diabetes is found among various Countries like Canada, China, and India etc. Population of India is now more than 100 million so the actual number of diabetics in India is 40 million. Diabetes is the major cause of death in the world. Early prediction of disease like diabetes can be controlled and save the human life. To accomplish this, this work explores prediction of diabetes by taking various attributes related to diabetes disease. For this purpose we use the Pima Indian Diabetes Dataset, we apply various Machine Learning classification and ensemble Techniques to predict diabetes. Machine Learning Is a method that is used to train computers or machines explicitly. Various Machine Learning Techniques provide efficient result to collect Knowledge by building various classification and ensemble models from collected dataset.**

Types of Diabetes :

* Type 1 diabetes means that the immune system is compromised and the cells fail to produce insulin in sufficient amounts. There are no eloquent studies that prove the causes of type 1 diabetes and there are currently no known methods of prevention.
* Type 2 diabetes means that the cells produce a low quantity of insulin or the body can’t use the insulin correctly. This is the most common type of diabetes, thus affecting 90% of persons diagnosed with diabetes. It is caused by both genetic factors and the manner of living.
* Gestational diabetes appears in pregnant women who suddenly develop high blood sugar. In two thirds of the cases, it will reappear during subsequent pregnancies. There is a great chance that type 1 or type 2 diabetes will occur after a pregnancy affected by gestational diabetes.

Symptoms of Diabetes:

• Frequent Urination

• Increased thirst

• Tired/Sleepiness

• Weight loss

• Blurred vision

• Mood swings

• Confusion and difficulty concentrating

• Frequent infections

Causes of Diabetes :

Genetic factors are the main cause of diabetes. It is caused by at least two mutant genes in the chromosome 6, the chromosome that affects the response of the body to various antigens. Viral infection may also influence the occurrence of type 1 and type 2 diabetes. Studies have shown that infection with viruses such as rubella, Coxsackievirus, mumps, hepatitis B virus, and cytomegalovirus increase the risk of developing diabetes.

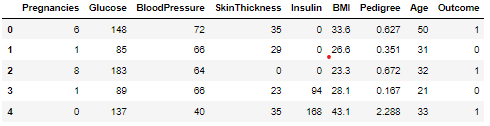
Objective :

To predict diabetic patients using machine learning techniques.

Data Description:

Diabetes dataset containing 768 cases. The objective is to predict based on the measures to predict if the patient is diabetic or not.

|  |  |
| --- | --- |
| Columns | Description |
| Pregnancies | Number of times pregnant |
| Glucose | Level of glucose in the body |
| BloodPressure | Level of blood pressure in the body |
| SkinThickness | Thickness of skin |
| Insulin | Lvel of Insulin in the body |
| BMI | Body mass index |
| Pedigree | Likelihood of diabetes on the basis of family history |
| Age | Age of the patient |
| Outcome | Diabetic or not (1/0) |



Approach:

Diabetes Dataset: Diabetes Dataset is very useful in our system for prediction of more accurate result. Using the Diabetes Dataset the system will automatically predict which patients have diabetes and which do not.

Determine the training and testing data: Typically , here the system separate a dataset into a training set and testing set ,most of the data use for training ,and a smaller portions of data is use for testing. After a system has been processed by using the training set, it makes the prediction against the test set.

Data cleaning and processing: In data cleaning the system detect and correct corrupt or inaccurate records from database and refers to identifying incomplete, incorrect, inaccurate or irrelevant parts of the data and then replacing , modifying or detecting the dirty or coarse data. In Data processing the system convert data from a given form to a much more usable and desired form i.e. make it more meaningful and informative.

PROPOSED ALGORITHM:

The following shows the pseudo code for the proposed loan prediction method

**1. Load the data**

**2. Determine the training and testing data**

**3. Data cleaning and pre-processing.**

a) Fill the missing values with mean values regarding numerical values.

b) Fill the missing values with mode values regarding categorical variables

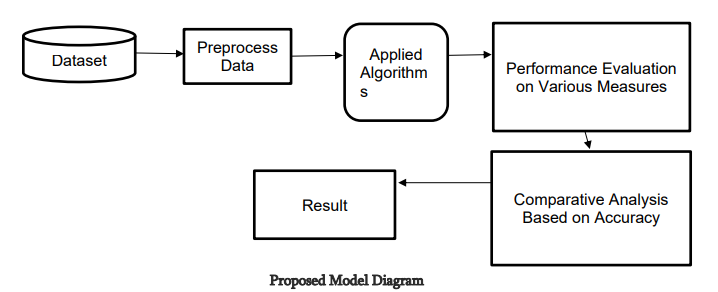
c) Outlier treatment.

**4**. **Apply the modelling for prediction**

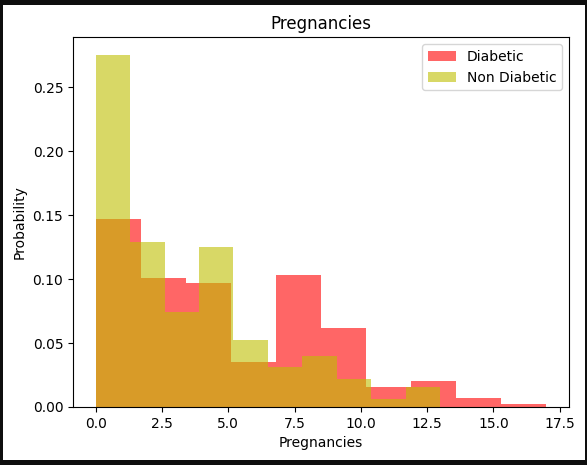
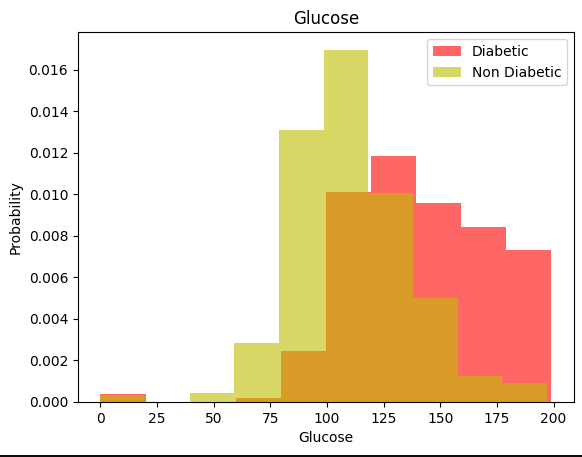
b) Create the target variable (based on the requirement). In this approach, target variable is the outcome

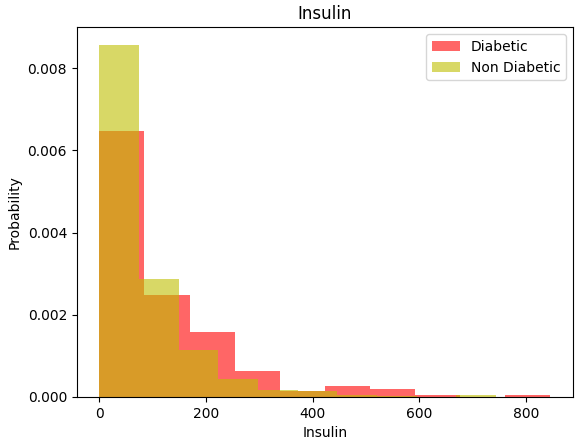
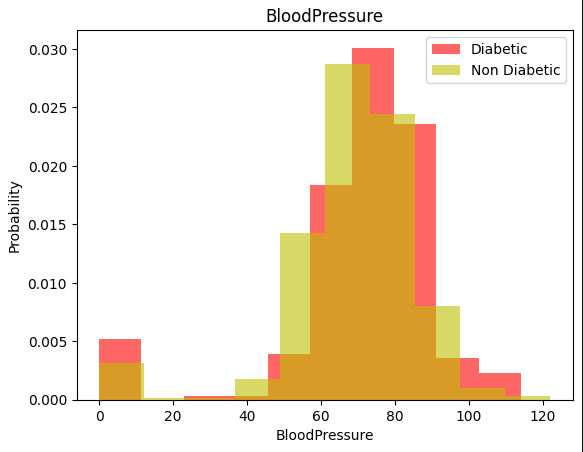
c) Create a dummy variable for categorical variable (if required) and split the training and testing data for validation.

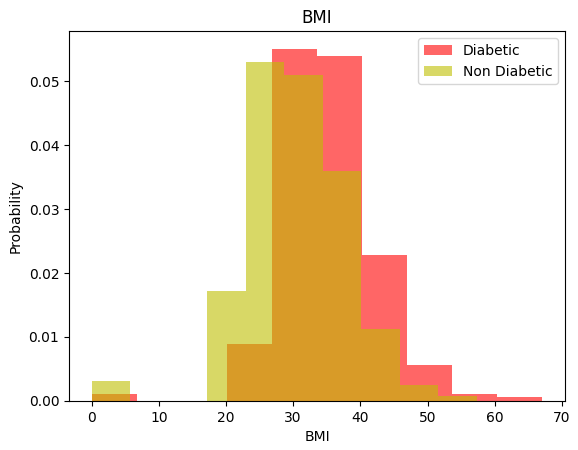
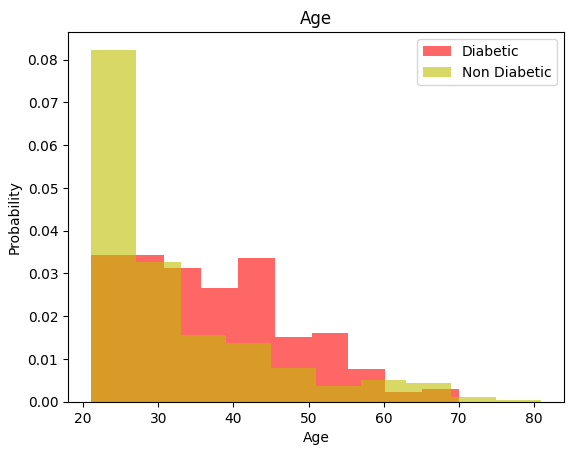
**5. Determine the accuracy.**



Histogram:

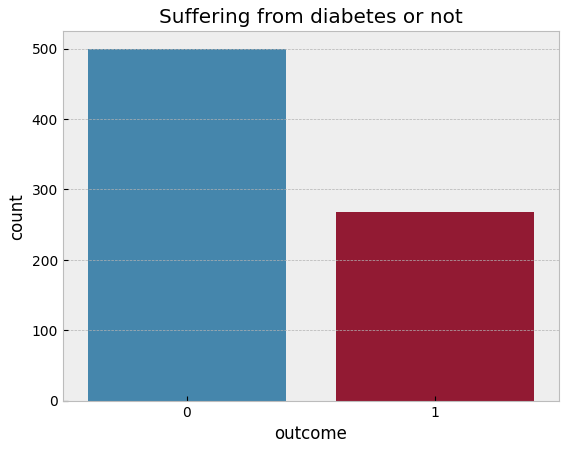


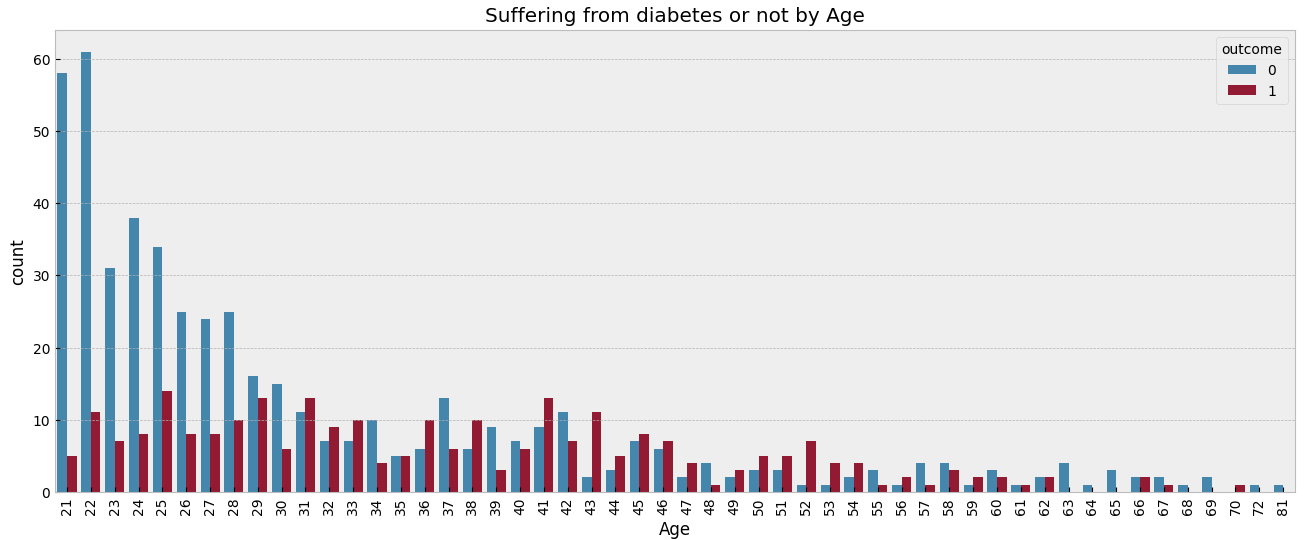
Let’s take a look at the plots. It shows how each feature and label is distributed along different ranges, which further confirms the need for scaling.

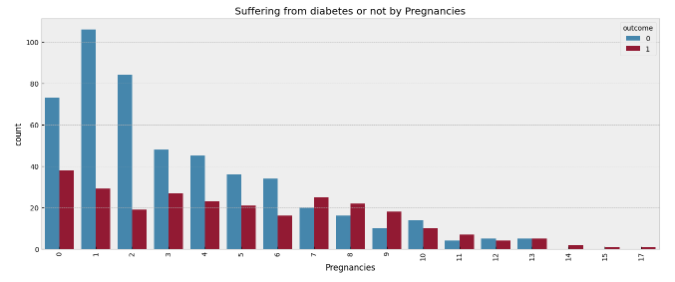
Countplot:

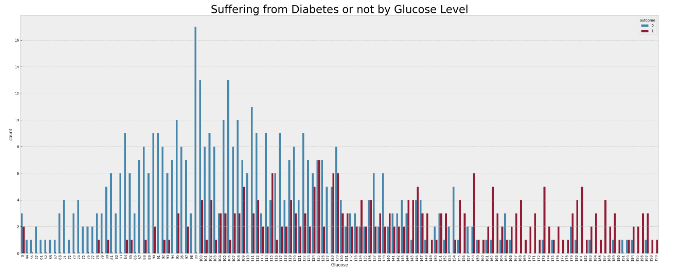
The following graphs depict the number of people suffering from diabetes with respect to different aspects such as age, insulin, BMI, glucose level, skin thickness etc.

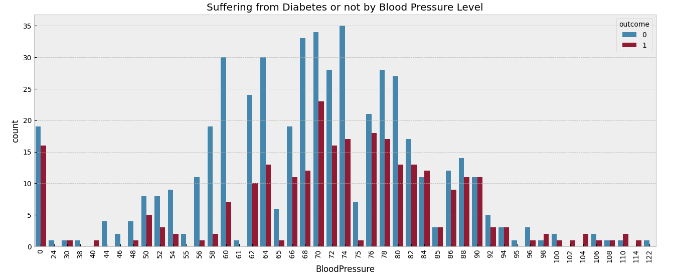


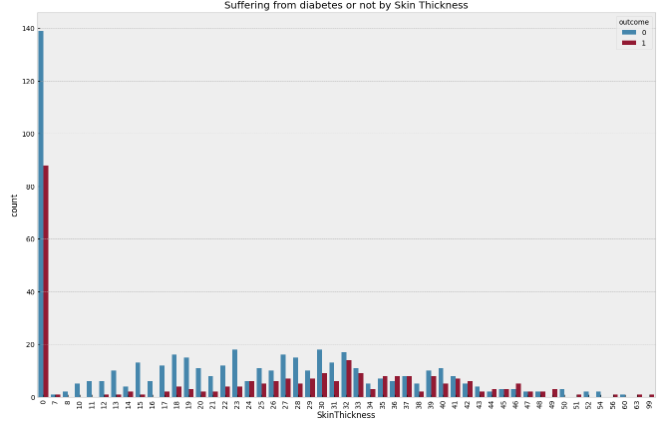
The above graph shows that the data is biased towards data points having outcome value as 0 where it means that diabetes was not present actually. The number of non-diabetics is almost twice the number of diabetic patients.

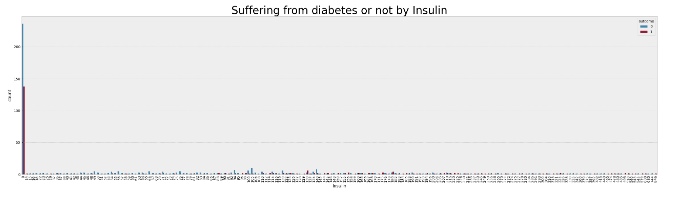


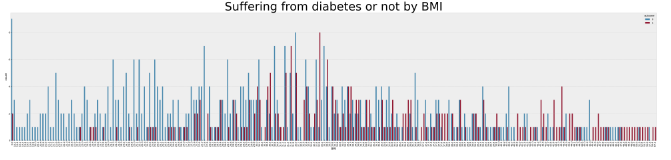












k-Nearest Neighbors:

The k-NN algorithm is arguably the simplest machine learning algorithm. Building the model consists only of storing the training data set. To make a prediction for a new data point, the algorithm finds the closest data points in the training data set, its “nearest neighbors.”

Considering if we choose one single nearest neighbor, the prediction on the training set is perfect. But when more neighbors are considered, the training accuracy drops, indicating that using the single nearest neighbor leads to a model that is too complex.

The accuracy is around 0.81.

Logistic regression:

Logistic regression is one of the most popular Machine Learning algorithms. It is used for predicting the categorical dependent variable using a given set of independent variables.

Logistic regression predicts the output of a categorical dependent variable.

The accuracy is around 0.79

Naïve Bayes classification:

Naive Bayes is a statistical classification technique based on Bayes Theorem. It is one of the simplest supervised learning algorithms. Naive Bayes classifier is the fast, accurate and reliable algorithm. Naive Bayes classifiers have high accuracy and speed on large datasets.

The accuracy is 0.78

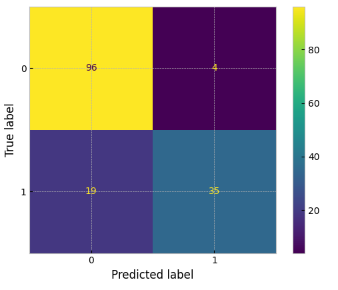
Support Vector Machine (SVM):

Support Vector Machine(SVM) is a supervised machine learning algorithm used for both classification and regression. Though we say regression problems as well its best suited for classification. The objective of SVM algorithm is to find a hyperplane in an N-dimensional space that distinctly classifies the data points.

The accuracy is 0.85

Confusion matrix:

A confusion matrix is a table that is used to define the performance of a classification algorithm. A confusion matrix visualizes and summarizes the performance of a classification algorithm.



Conclusion:

One of the important real-world medical problems is the detection of diabetes at its early stage. In this study, systematic efforts are made in designing a system which results in the prediction of diabetes. During this work, five machine learning classification algorithms are studied and evaluated on various measures. Experiments are performed on john Diabetes Database. In future, the designed system with the used machine learning classification algorithms can be used to predict or diagnose other diseases. The work can be extended and improved for the automation of diabetes analysis including some other machine learning algorithms.