|  |
| --- |
|  |
| Prediction of Diabetes using various machine learning models |
|  |

|  |
| --- |
| **S.Y.V.Manidhar;V.Pujitha;K.Veera Vidya;P.Bharathi;P.Siva Nanda Reddy**  21/06/2018 |
|  |

**ABSTRACT:**

**Diabetes has become a serious health problem in present world.** About one in seven U.S adults has diabetes now, according to the [Centers for Disease Control and Prevention](https://www.cdc.gov/). But by 2050, that rate could skyrocket to as many as one in three. In 2013 it was estimated that over 382 million people throughout the world had diabetes. Diabetes could cause hypoglycemia(low BP) and hyperglycemia(high BP) which could cause bad effect on patient. There are many factors that cause diabetes for example Glucose, Blood Pressure, Skin Thickness, etc .Data sciences could help us solving these severe problem of diabetes Here we use various Machine Learning models to predict the probability of patient being effected by Diabetes.

**INTRODUCTION:**

**The evolution in data sciences has lead to huge benefits in almost every field of real-world working environment such as weather report prediction, super-market sales analysis, hospital data analysis. Packages in python like NumPy,Pandas,Sklearn,matplotlib help us in data analytics.Using those packages we can moodily the data,plot the dependencies of features,predict the outcome of certain tuple of features.**

**In this project, we implement a python code for predicting the probability of getting affected by diabetes using different machine learning models.**

**PROPOSED METHOD:**

**There are many machine learning models but in the project we concentrate on Logistic Regression,Decision Tree,Random Forest,SVM,K nearest neighbors.**

**Pseudocode:**

1. **Load the data set.**
2. **Find the co-relation, plot the partial regression plot on dependent variables vs independent variables and conclude whick feature affect outcome most and least. .**
3. **Split the dataset into train set and test set.**
4. **Implent different machine learning models.**
5. **Train the different models using train set.**
6. **Now predict the outcome using the model and test data.**
7. **Obtain the confusion matrix and classification report of different classifiers.**
8. **Conclude the analysis by specifying the classifier used to obtain the maximum accuracy for the taken data set.**

**EXPERIMENTAL SETUP:**

**The Dataset used in this project is Diabetes Dataset. It contains 768 entries and 8 attributes. The data collected is based on No.of Pregnancies ,Glucose level in Blood, Blood Pressure, Skin thickness, Insulin,Body Mass Index,Diabetes Pedigree Function,Age.**

**Attribures Description:**

|  |
| --- |
| **Attribute Name No.of attributes Type of attribute** |

**Pregnancies: 768 non-null int64**

**Glucose: 768 non-null int64**

**BloodPressure: 768 non-null int64**

**SkinThicknes: 768 non-null int64**

**Insulin: 768 non-null int64**

**BMI: 768 non-null float64**

**DiabetesPedigreeFunction : 768 non-null float64**

**Age: 768 non-null int64**

**No.of Pregnancies: Number of births given by the patient .**

**Glucose level in Blood: Glucose content in blood noted through blood test.**

**Blood Pressure: Noted using sphygmomanometer,blood pressure monitor,etc.**

**Skin thickness: Noted using some ultrasonic scans.**

**Insulin: Insulin contain present in pancreas.**

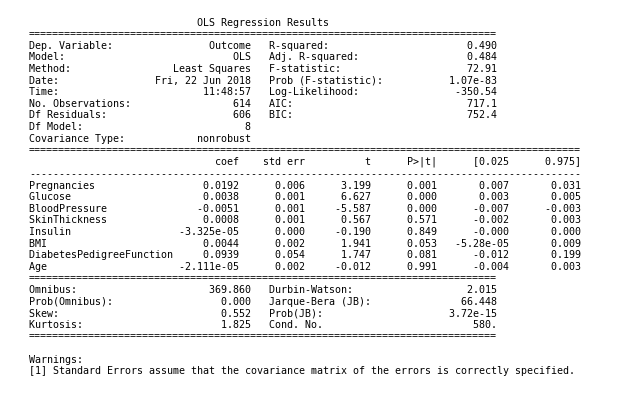
**Body Mass Index: BMI is a measure of body fat based on height and weight that applies to adult men and women.**

**Diabetes Pedigree Function: It provides some data about diabetes mellitus history in relatives and genetic relationship of those relatives to the patient.**

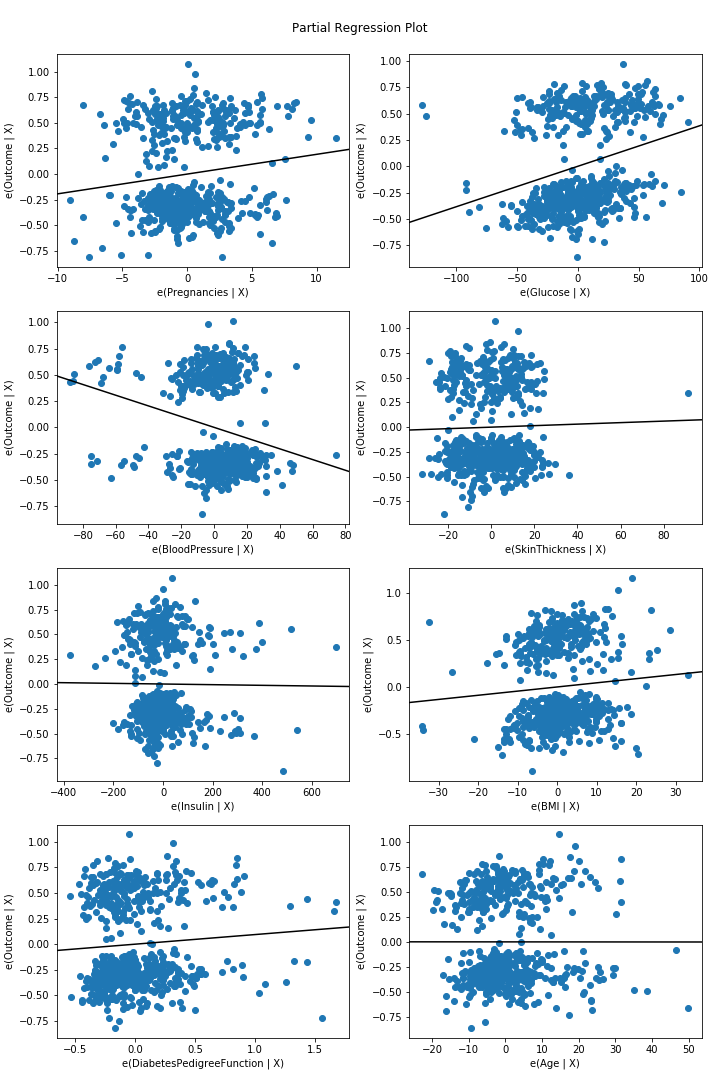
**Age:Age of the patient at the time of data entry.**

**ANALYSIS:**

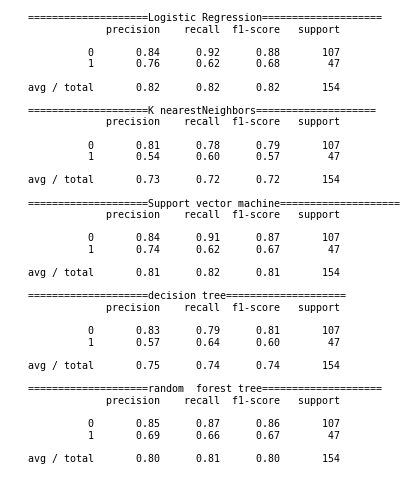
**We use python with necessary libraries to explore the dataset.First, we import those libraries.Next, we import the dataset and analyze the dataset.We then find the partial regression plot and statistical summary of the data the results are shown below**



**We then plot the partial regression plot of dependent variables Vs outcome the plot is as shown below:**



**Next we split the data into train set and test set .Using that train set we train different models and predict the outcomes using those models and test set. The Classification reports of the classifiers as show below**



**From the results we see that the accuracies of different classifier models are**

1. **Logistic Regression : 82%**
2. **KNearest Neighbors: 73%**
3. **SVM : 81%**
4. **Decision tree : 75%**
5. **Random forest tree : 80%**

**Conclusion:**

**A brief study of five classifiers is made on the taken dataset and the results are tabulates. It is found that Logistic regression has the maximum accuracy of 82%.Through the regression plot and correlation summary we say that the Glucose feature is most affecting and age is the least affecting feature for getting diabetes.This shows the importance of machine learning models in predicting whether the patient getting affected by diabetes**