Diabetes Prediction Using Machine Learning in Python

Somesh P. Panchal

Dept. of Computer Engineering

Sardar Vallabhbhai Patel Institute of technology

Vasad, Gujarat

Abstract: Diabetes is an illness caused because of high glucose level in a human body. Diabetes should not be ignored if it is untreated then Diabetes may cause some major issues in a person like: heart related problems, kidney problem, blood pressure, eye damage and it can also affects other organs of human body. Diabetes can be controlled if it is predicted earlier. To achieve this goal this project work we will do early prediction of Diabetes in a human body or a patient for a higher accuracy through applying, Various Machine Learning Techniques. Machine learning techniques Provide better result for prediction by constructing models from datasets collected from patients. In this work i will use Machine Learning Classification and ensemble techniques on a dataset to predict diabetes. Which are K-Nearest Neighbor (KNN), Logistic Regression (LR), Decision Tree (DT), Support Vector Machine (SVM), XGBoost and Random Forest (RF). The accuracy is different for every model when compared to other models. The Project work gives the accurate or higher accuracy model shows that the model is capable of predicting diabetes effectively. Our Result shows that XGBoost achieved higher accuracy compared to other machine learning techniques.

Introduction: Diabetes is noxious diseases in the world. 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. According to (WHO) World Health Organization about 422million 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 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.

About Dataset: 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. Such collected data can be useful to predict diabetes. Various techniques of Machine Learning can capable to do prediction, however it’s tough to choose best technique.

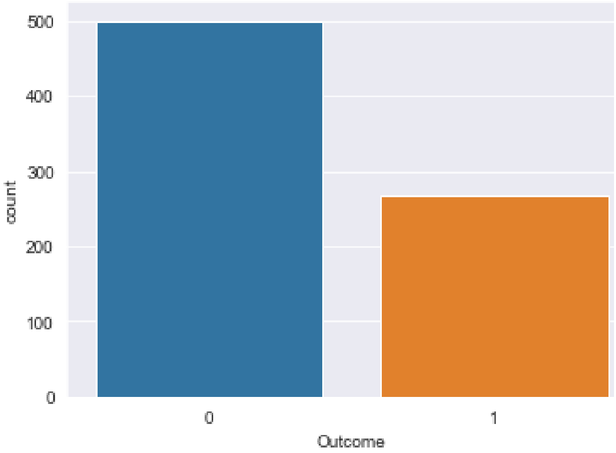
Pregnancies	Number of times pregnant
Glucose	Plasma glucose concentration a 2 hours in an oral glucose tolerance test
Blood Pressure	Diastolic blood pressure (mm Hg)
Skin Thickness	Triceps skin fold thickness (mm)
Insulin	2-Hour serum insulin (mu U/ml)
BMI	Body mass index
Diabetes pedigree function	Diabetes pedigree function
Age	Age(Years)
Outcome	0 – Absence of diabetes 1 – Presence of diabetes

In this project, i will use python3 and Jupyter notebook. We will go through the project by importing the dataset, conducting exploratory data analysis to get insights and understanding on how the dataset looks like and then build the model. i will further use Decision Trees, Random Forests, Support Vector Machines and XGBoost.

The 7 Steps of Machine Learning

- Data Collection
- Data Preparation
- Choosing a model
- Training the model
- Evaluating the model
- Parameter tuning
- Making prediction

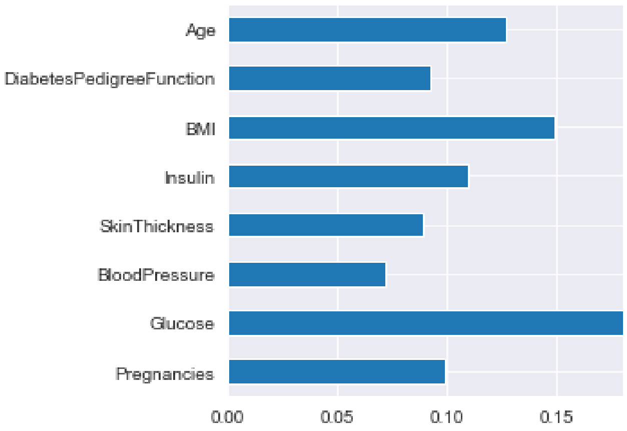
Distribution of Diabetic patient- We made a model to predict diabetes however the dataset was slightly imbalanced having around 500 classes labeled as 0 means negative means no diabeteS Data Science Project s and 268 labeled as 1 means positive means diabetic.



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

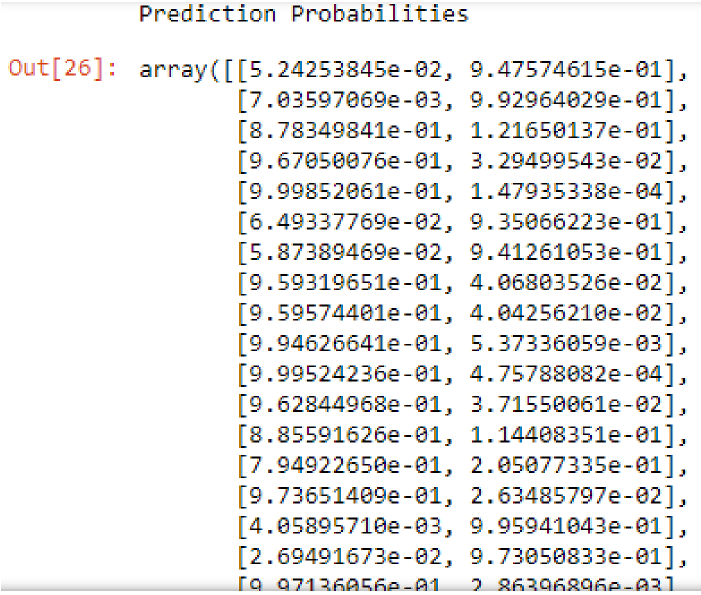
Missing Values removal- Remove all the instances that have zero (0) as worth. Having zero as worth is not possible. Therefore this instance is eliminated. Through eliminating irrelevant features/instances we make feature subset and this process is called features subset selection, which reduces dimensionality of data and help to work faster. Splitting of data- After cleaning the data, data is normalized in training and testing the model. When data is spitted then we train algorithm on the training data set and keep 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 same scale.

Feature Importance: It is important to understand how the features (variables) in our model contribute to prediction. We will then look at the features that are most important when we use the XGBoost model for diabetes prediction.



The graph shows that the most important feature in this diabetes prediction is **Glucose** followed by **BMI.**

Prediction*:* Finally, let us use XGBoost Model to predict the possibility of a patient having Diabetes or not (1 or 0). The following are the prediction probabilities of absence or presence of Diabetes respectively.



```
          Prediction Probabilities

Out[26]: array([[5.24253845e-02, 9.47574615e-01],
                [7.03597069e-03, 9.92964029e-01],
                [8.78349841e-01, 1.21650137e-01],
                [9.67050076e-01, 3.29499543e-02],
                [9.99852061e-01, 1.47935338e-04],
                [6.49337769e-02, 9.35066223e-01],
                [5.87389469e-02, 9.41261053e-01],
                [9.59319651e-01, 4.06803526e-02],
                [9.59574401e-01, 4.04256210e-02],
                [9.94626641e-01, 5.37336059e-03],
                [9.99524236e-01, 4.75788082e-04],
                [9.62844968e-01, 3.71550061e-02],
                [8.85591626e-01, 1.14408351e-01],
                [7.94922650e-01, 2.05077335e-01],
                [9.73651409e-01, 2.63485797e-02],
                [4.05895710e-03, 9.95941043e-01],
                [2.69491673e-02, 9.73050833e-01],
                [9.97136056e-01, 2.86396896e-03]
```

the patient at index 0 has a 98.1% chance of absence of diabetes, while the patient at index 1 has a 91.8% predicted chance of having diabetes.

Conclusion: The main aim of this project was to design and implement Diabetes Prediction Using Machine Learning Methods and Performance Analysis of that methods and it has been achieved successfully. The proposed approach uses various classification and ensemble learning method in which Gradient Boosting classifiers are used. And 78% classification accuracy has been achieved. The Experimental results can be asst health care to take early prediction and make early decision to cure diabetes and save humans life.