Electronics and Communication Engineering

Machine Learning

(UE18EC388)

Diabetes Prediction Using Pima Dataset

Made By :

Shreyam Kulshrestha	**6A**	**PES2201800237**
Suraj Naik	**6C**	**PES2201800692**

INDEX

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.
- 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.
- 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.

Problem Statement

- Early Prediction of diabetes so that diabetes can be controlled and save human lives.
- Prediction can be done using machine learning and dataset provided by Pima.
- Prediction can be done on the basis of glucose , insulin ,BMI , age.

Methodology

- **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.
- **K-Nearest Neighbor:** KNN is also a supervised machine learning algorithm. KNN helps to solve both the classification and regression problems. KNN is lazy prediction technique. KNN assumes that similar

things are near to each other. Many times, data points which are similar are very near to each other. KNN helps to group new work based on similarity measure.

- **Decision Tree**: Decision tree is a basic classification method. It is supervised learning method. Decision tree used when response variable is categorical. Decision tree has tree like structure-based model which describes classification process based on input feature. Input variables are any types like graph, text, discrete, continuous etc.
- **Support Vector Machine** : Support Vector Machine also known as svm is a supervised machine learning algorithm. Svm is most popular classification technique. Svm creates a hyperplane that separate two classes. It can create a hyperplane or set of hyperplanes in high dimensional space.

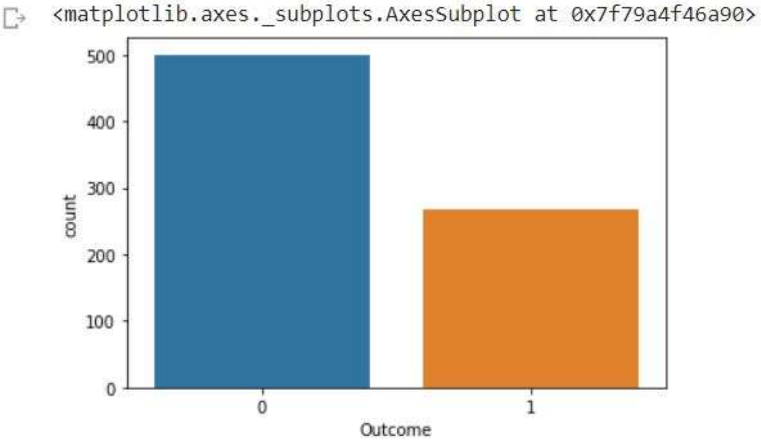
Software

- Colaboratory (Google Colab notebook)
  - Colaboratory, or "Colab" for short, allows you to write and execute Python in your browser.
  - https://colab.research.google.com/drive/1_uJAk-T9bSNxGyzT-NZefxXmf_TxvHwj#scrollTo=d-n7KrGzWpYs

Dataset details

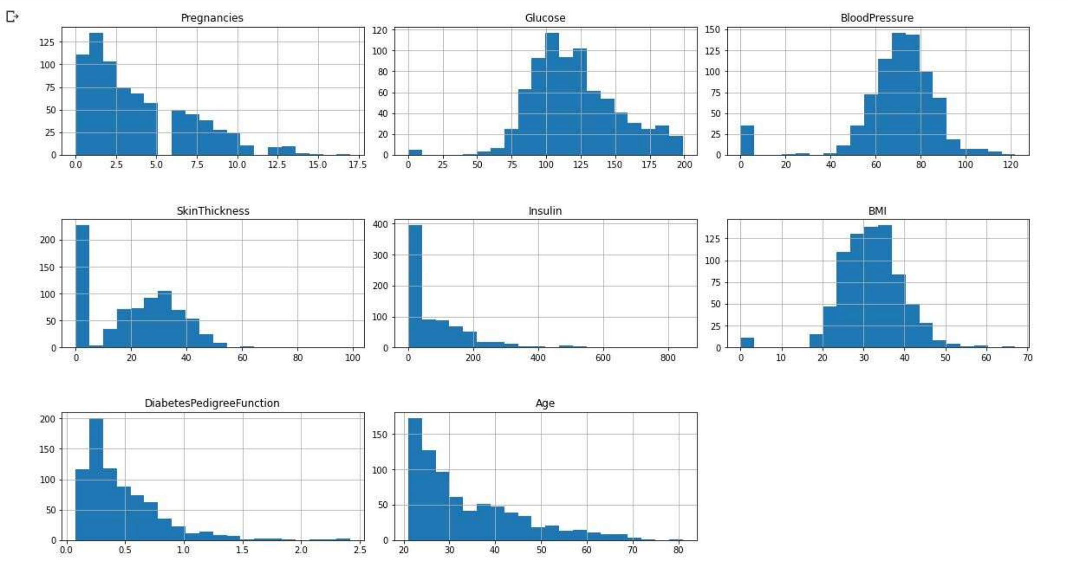
- Pima Indians Diabetes Database
  - The objective of the dataset is to predict whether or not a patient has diabetes, based on certain diagnostic measurements included in the dataset. The datasets consist of several medical predictor variables and one target variable, Outcome. Predictor variables includes the number of pregnancies the patient has had, their BMI, insulin level, age, and so on.
  - There are a total of 768 records and 9 features in the dataset.
  - Each feature can be either of integer or float datatype.
  - https://www.kaggle.com/uciml/pima-indians-diabetes-database

Results and Analysis

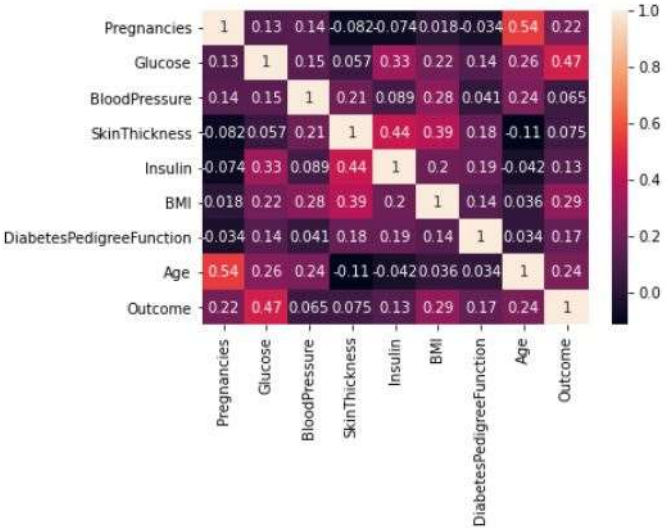


```
<matplotlib.axes._subplots.AxesSubplot at 0x7f79a4f46a90>
```

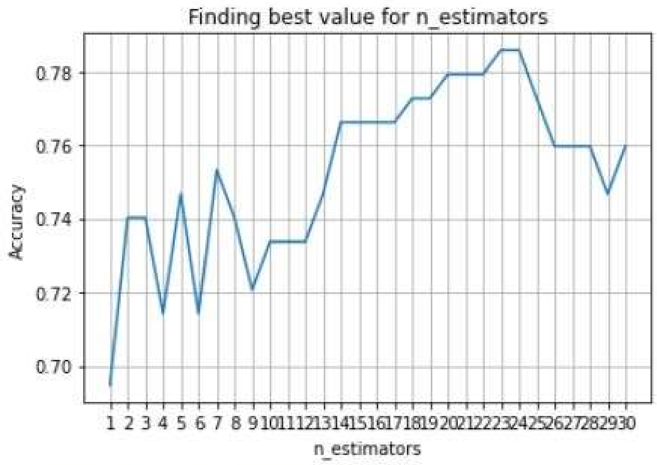
- ➢ This graph shows total number of diabetic people (1) and non-diabetic people (0).



- ➢ These graphs show data exploration for each feature variable . Y-axis represents the number of students.

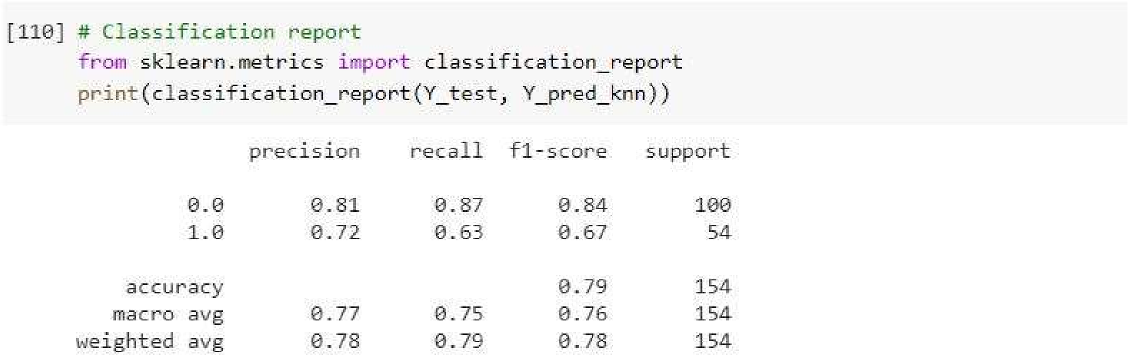


- This is a heat map shows correlation matrix i.e., darker the colour more the correlation.
- From the correlation heatmap, we can see that there is a high correlation between Outcome and [Glucose, BMI, Age, Insulin]. We can select these features to accept input from the user and predict the outcome.



```
Highest value:  0.7857142857142857
```

- This is the elbow method to get best value of n to be used for knn .

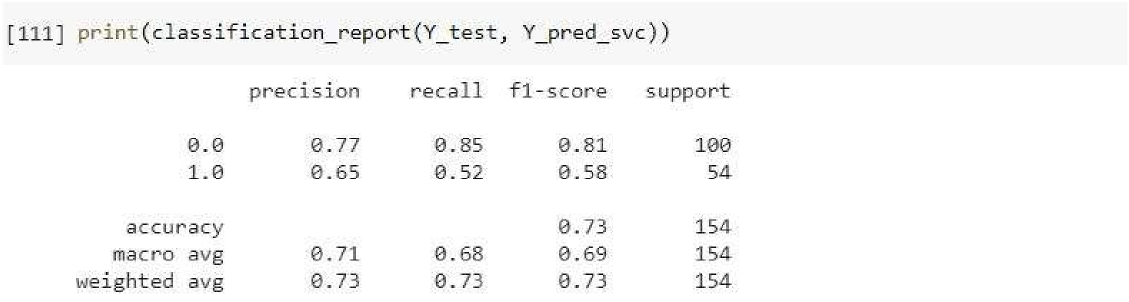


```
[110] # Classification report
      from sklearn.metrics import classification_report
      print(classification_report(Y_test, Y_pred_knn))
```

```
              precision    recall  f1-score   support

         0.0       0.81      0.87      0.84       100
         1.0       0.72      0.63      0.67        54

    accuracy                           0.79       154
   macro avg       0.77      0.75      0.76       154
weighted avg       0.78      0.79      0.78       154
```



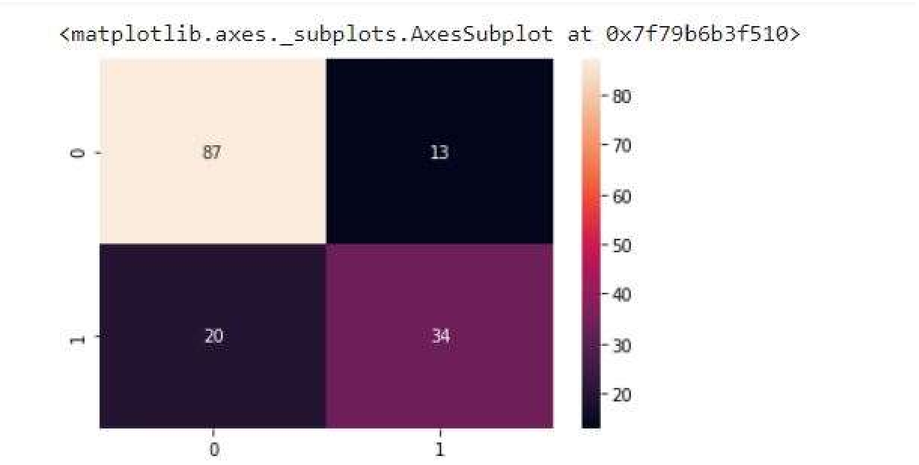
```
[111] print(classification_report(Y_test, Y_pred_svc))
```

```
              precision    recall  f1-score   support

         0.0       0.77      0.85      0.81       100
         1.0       0.65      0.52      0.58        54

    accuracy                           0.73       154
   macro avg       0.71      0.68      0.69       154
weighted avg       0.73      0.73      0.73       154
```

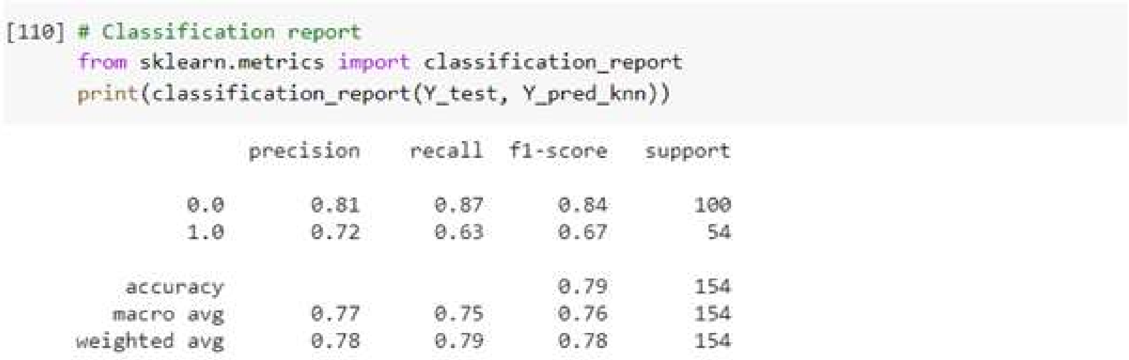
- ➢ This is the accuracy matrix that shows how much precision recall is there.



```
<matplotlib.axes._subplots.AxesSubplot at 0x7f79b6b3f510>
```

- ➢ This is the graphical representation of accuracy matrix that show how accurate model is.

Conclusion



```
[110] # Classification report
      from sklearn.metrics import classification_report
      print(classification_report(Y_test, Y_pred_knn))
```

```
              precision    recall  f1-score   support

         0.0       0.81      0.87      0.84       100
         1.0       0.72      0.63      0.67        54

    accuracy                           0.79       154
   macro avg       0.77      0.75      0.76       154
weighted avg       0.78      0.79      0.78       154
```

Future Scope

Combined with other parameters such as weight, height , Activeness level , daily routine . The model could be modified to predict the fitness Index which could help on spreading health awareness .

Reference

- Kaggle.com
    - https://www.kaggle.com/uciml/pima-indians-diabetes-database
- https://www.ijert.org/diabetes-prediction-using-machine-learning-techniques
- https://www.sciencedirect.com/science/article/pii/S1877050920300557