CLASSIFICATION OF DIABETES USING NAIVE BAYES CLASSIFIER

A COURSE PROJECT REPORT

By

SRINAATH NARASIMHAN (RA2011003010309)

KANURI S V S SAI KUMAR(RA2011003010303)

APPALA SAI SURYA VARUN(RA2011003010296)

Under the guidance of

G ABIRAMI

In partial fulfillment for the Course

of

18CSE355T - DATA MINING AND ANALYSIS

in <School Of Computing>

FACULTY OF ENGINEERING AND TECHNOLOGY

SRM INSTITUTE OF SCIENCE AND TECHNOLOGY

Kattankulathur, Chenpalpattu District

NOVEMBER 2022

ABSTRACT

Approximately 422 million people across the world have diabetes, particularly in countries where the average income is in the middle and lower end of the economic spectrum. Statistics reveal that every year, about 1.6 million deaths are recorded which can be directly attributed to diabetes.Diabetes means blood sugar is above desired level on a sustained basis. The prime objective of this research work is to provide a better classification of diabetes. There are already several existing methods, which have been implemented for the classification of diabetes dataset. In the medical sector, the classifications systems have been widely used to exploit the patient's data and make predictive models or build sets of rules.

In this project, firstly the dataset obtained is preprocessed using the python libraries like numpy,pandas, matplotlib and sklearn. Then the processed data is taken and the Naive Bayesian algorithm is used for the classification on all the attributes and then Genetic Algorithm is used as an attribute selection and the Naive Bayesian used on that selected attribute for classification. The experimental results show the performance of this work on PIDD and provide better classification for diagnosis.

INTRODUCTION

Diabetes is a problem and a major public health challenge worldwide. This is one of the most wide- spread diseases, nowadays very common.

Diabetes mellitus or as it is simply called, diabetes is a disease which perpetuates in the metabolic method. It causes high blood sugar levels in an individual. The pancreas produces one of the most essential hormones of the human body, the insulin. The insulin extracts blood sugar and transports it for storage or to be used as cellular energy. For a diabetic patient, the body either produces insufficient insulin or is incapable of using the insulin that has been developed. Diabetes can be broadly categorized into four types: Type 1 , Type 2 , Pre-diabetes and Gestational Diabetes.

In this project, Genetic Algorithm (GA) has been used as an attribute (feature) selection method by which four attributes have been selected from eight attributes. The genetic algorithm works on the evolutionary generational cycle to generate high-quality solutions. These algorithms use different operations that either enhance or replace the population to give an improved fit solution. Naive Bayes (NBs) are statistical, super- vised learning methods for classification. Here, NBS has been used for the classification of the diabetes diagnosis.

TECHNIQUE USED:

Here, the proposed methodology is implemented by Genetic Algorithm as an Attribute Selection and Naive Bayes is used for Classification on the dataset which has been taken from UCI machine learning repository.

A Naive Bayes classifier is a probabilistic machine learning model that's used for classification tasks. The crux of the classifier is based on the Bayes theorem.

The implementation of the algorithm is as follows:

1. First, The dataset has been taken from UCI machine learning repository.
2. Then the data is preprocessed using the python libraries numpy, matplotlib and pandas.
3. Apply Genetic Algorithm as an Attribute Selection on the dataset.
4. Do the Classification by using Naive Bayes on selected attributes and all the attributes in the dataset.

PREPROCESSING:

Data Cleaning: Data Cleaning is a process which helps to detect and correct corrupt or inaccurate data from a dataset. It also involves identification of incomplete, inaccurate, or irrelevant parts of the data followed by their replacement, modification, or deletion of the coarse data.

Data Aggregation: In data aggregation, all the obtained information is gathered or aggregated as the name suggests and then expressed in the form of a summary. It serves several purposes such as statistical analysis. In this system, data aggregation is done using more relevant information about diabetic parameters based on specific attributes namely age, BMI, medical history (Diabetes Pedigree Function), glucose levels, skin thickness, Insulin, Blood Pressure of the patient.

IMPLEMENTATION OF GENETIC ALGORITHM:

Genetic Algorithm for attribute selection works in such a way that either the algorithm stops forming new iterations when a maximum number of iterations have been formed or a satisfactory fitness value is achieved for the problem. Using this algorithm, no. of attributes required for the analysis are selected.

USAGE OF NAIVE BAYES THEOREM:

Then,the naive bayes algorithm is used for classification in the described manner below:

Pseudocode Calculate diagnosis = "yes", diagnosis = "no" probabilities.

Pyes, Pno from training input for each test input samples.

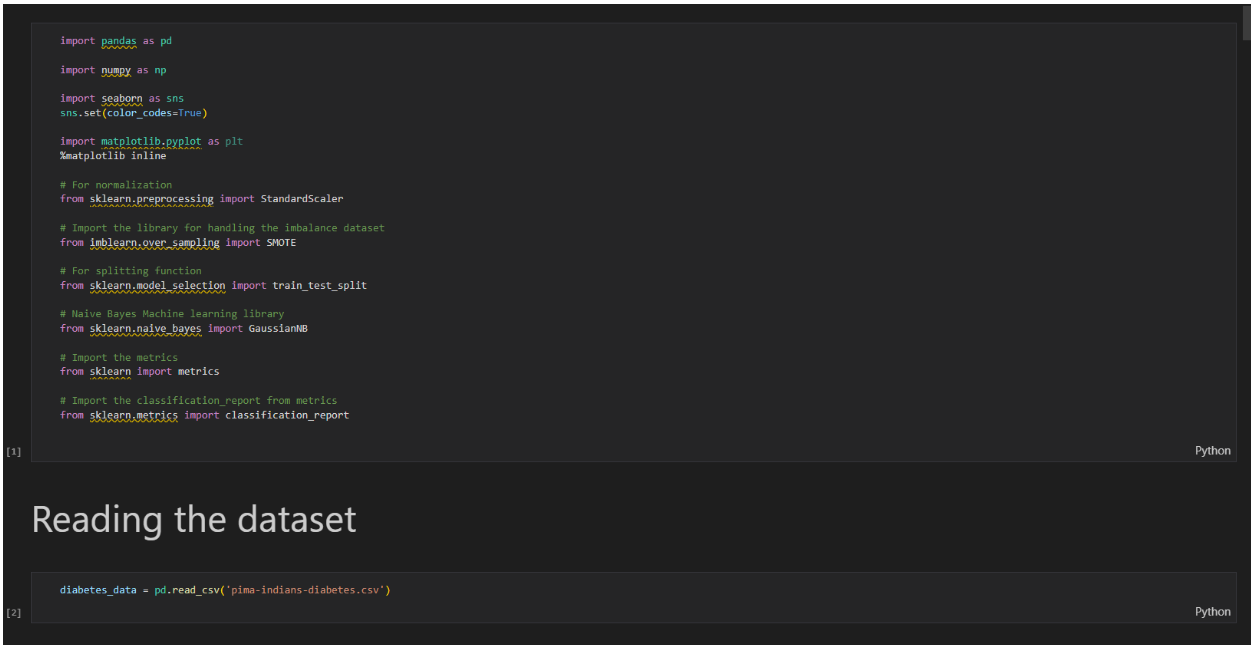
For each feature,

Calculation of category of feature based on categorical division.

Then, probabilities of diagnosis = "yes", diagnosis = "no" corresponds to that category P(feat, yes), P(feat, no) from training input are calculated.

For each feature, the result values are calculated. The algorithms are performed and verified and the results are shown in the below sections.

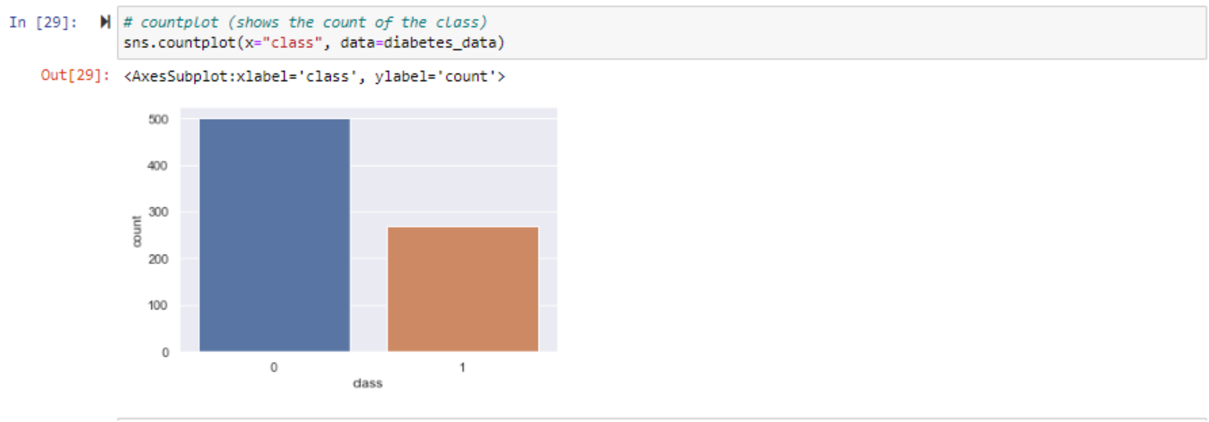
IMPLEMENTATION AND RESULTS:



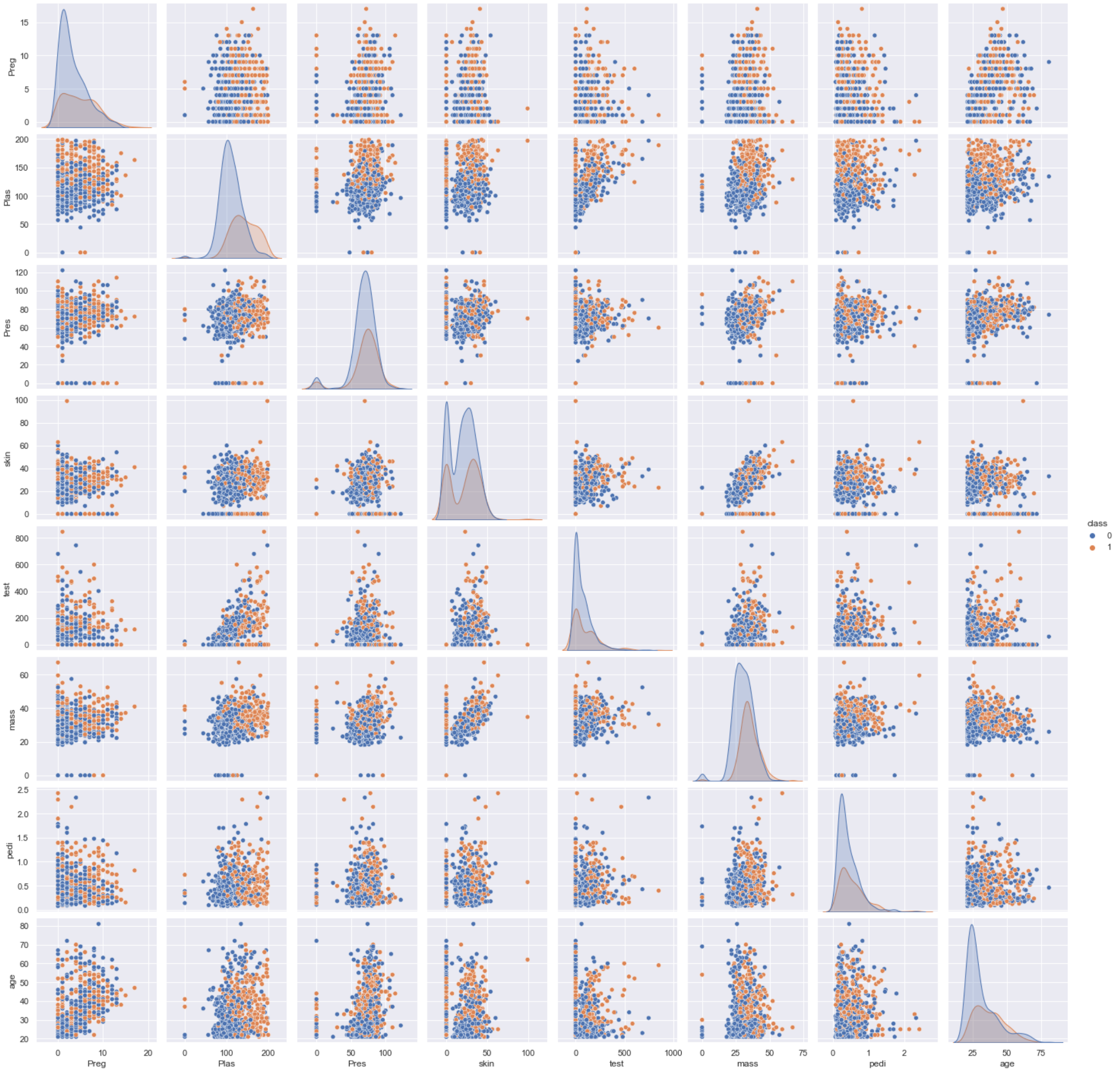
```
import pandas as pd

import numpy as np

import seaborn as sns
sns.set(color_codes=True)

import matplotlib.pyplot as plt
%matplotlib inline

# For normalization
from sklearn.preprocessing import StandardScaler

# Import the library for handling the imbalance dataset
from imblearn.over_sampling import SMOTE

# For splitting function
from sklearn.model_selection import train_test_split

# Naive Bayes Machine learning library
from sklearn.naive_bayes import GaussianNB

# Import the metrics
from sklearn import metrics

# Import the classification_report from metrics
from sklearn.metrics import classification_report
```

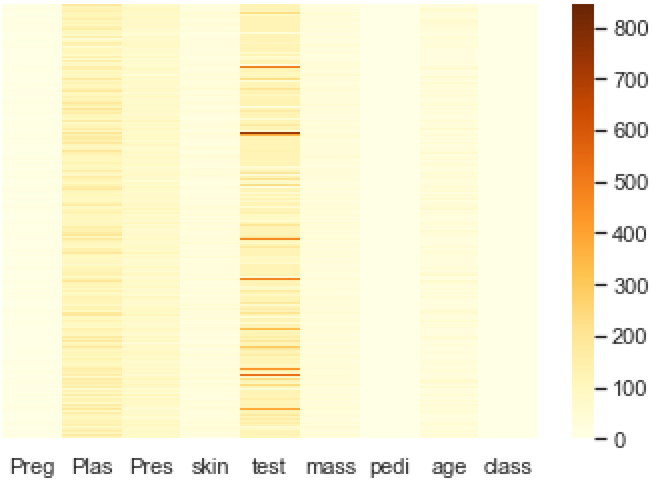
Reading the dataset

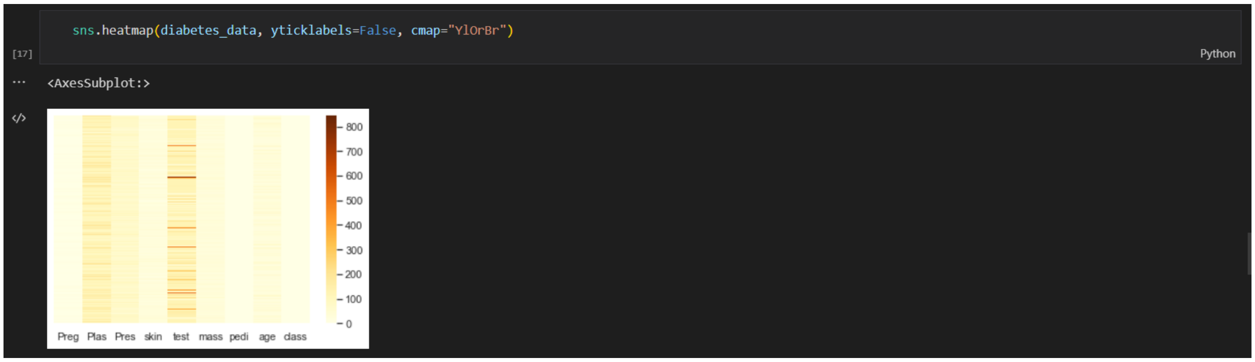
```
diabetes_data = pd.read_csv('pima-indians-diabetes.csv')
```



```
In [29]:  # countplot (shows the count of the class)
          sns.countplot(x="class", data=diabetes_data)

Out[29]: <AxesSubplot:xlabel='class', ylabel='count'>
```

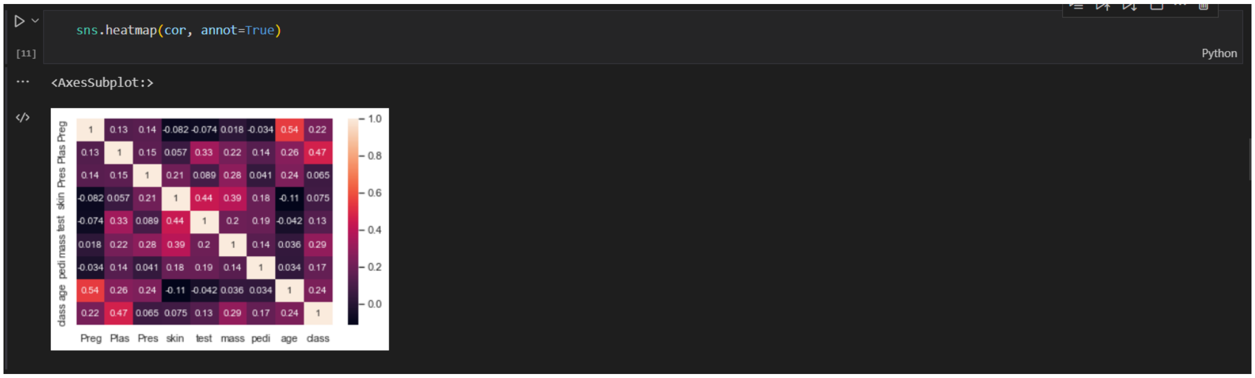

Preg
Plas
Pres
skin
test
mass
pedi
age
class
0
1
0
5
10
15
20
0
50
100
150
200
0
20
40
60
80
100
120
0
200
400
600
800
1000
0.0
0.5
1.0
1.5
2.0
2.5
25
30
40
50
60
70
75
80





```
sns.heatmap(diabetes_data, yticklabels=False, cmap="YlOrBr")
```

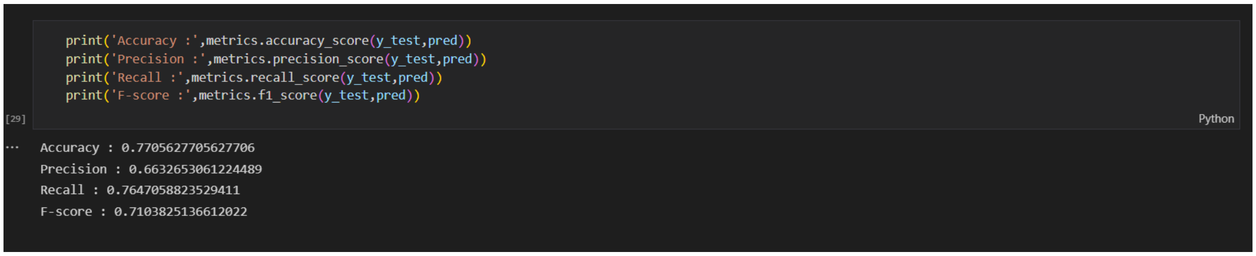
```
<AxesSubplot:>
```



```
sns.heatmap(cor, annot=True)
```

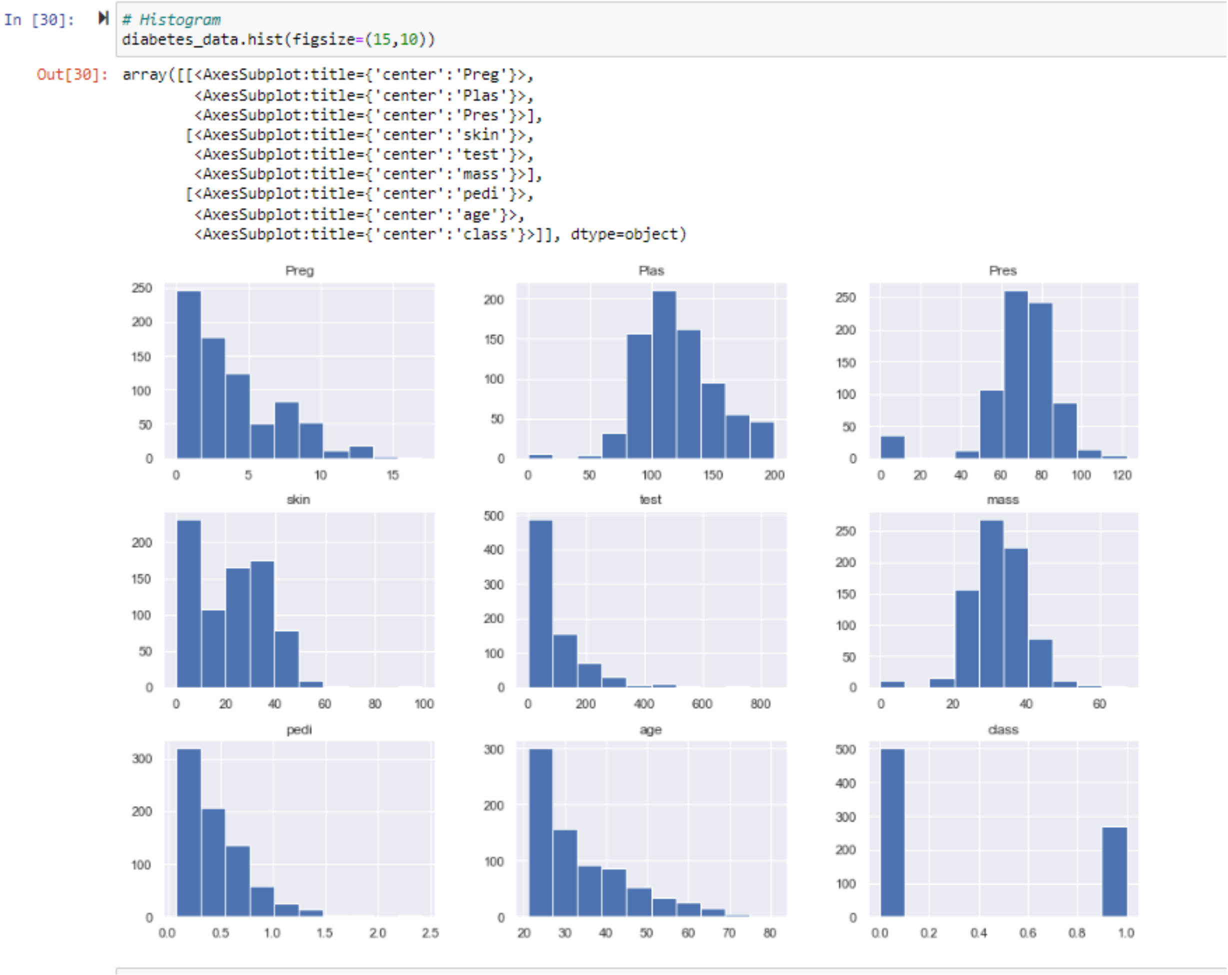
```
<AxesSubplot:>
```

	Preg	Plas	Pres	skin	test	mass	pedi	age	class
Preg	1	0.13	0.14	-0.082	-0.074	0.018	-0.034	0.54	0.22
Plas	0.13	1	0.15	0.057	0.33	0.22	0.14	0.26	0.47
Pres	0.14	0.15	1	0.21	0.089	0.28	0.041	0.24	0.065
skin	-0.082	0.057	0.21	1	0.44	0.39	0.18	-0.11	0.075
test	-0.074	0.33	0.089	0.44	1	0.2	0.19	-0.042	0.13
mass	0.018	0.22	0.28	0.39	0.2	1	0.14	0.036	0.29
pedi	-0.034	0.14	0.041	0.18	0.19	0.14	1	0.034	0.17
age	0.54	0.26	0.24	-0.11	-0.042	0.036	0.034	1	0.24
class	0.22	0.47	0.065	0.075	0.13	0.29	0.17	0.24	1



```
[29]
print('Accuracy :',metrics.accuracy_score(y_test,pred))
print('Precision :',metrics.precision_score(y_test,pred))
print('Recall :',metrics.recall_score(y_test,pred))
print('F-score :',metrics.f1_score(y_test,pred))
```

```
Accuracy : 0.7705627705627706
Precision : 0.6632653061224489
Recall : 0.7647058823529411
F-score : 0.7103825136612022
```



```
In [30]:  # Histogram
          diabetes_data.hist(figsize=(15,10))
```

```
Out[30]: array([[<AxesSubplot:title={'center':'Preg'}>,
                 <AxesSubplot:title={'center':'Plas'}>,
                 <AxesSubplot:title={'center':'Pres'}>],
                [<AxesSubplot:title={'center':'skin'}>,
                 <AxesSubplot:title={'center':'test'}>,
                 <AxesSubplot:title={'center':'mass'}>],
                [<AxesSubplot:title={'center':'pedi'}>,
                 <AxesSubplot:title={'center':'age'}>,
                 <AxesSubplot:title={'center':'class'}>]], dtype=object)
```

CONCLUSION:

In Experimental studies the dataset has been par- titioned between 70-30% (538-230) for training & test of NBS, GA_NBs. It has been performed on a dataset and the results compared with several existing methods which are noted in Table 5.

By applying the GA method, four attributes have been selected from eight attributes. This means the cost has been reduced to s(x) = 4/8 = 0.5 from 1 and an improvement on the training and classification by a factor of 2.

This project focuses widely on the methodological approach of Naive Bayes Classifier to detect diabetes accurately and precisely. The proposed approach can handle both large and small amounts of data. This process being a classification algorithm is fast, secure, easy to adapt and provides accurate results. It can be useful in medical research and a mobile application/website can also be built based on these findings to make it more accessible to common users and users who have a risk of developing diabetes.