Data Mining & Data Warehousing

BSCS - 633

Classification Model Report

Submitted By:

Mohammad Taha B18101055

Course Incharge:

Miss Maryam Feroze

Diabetes Health Indicators Dataset

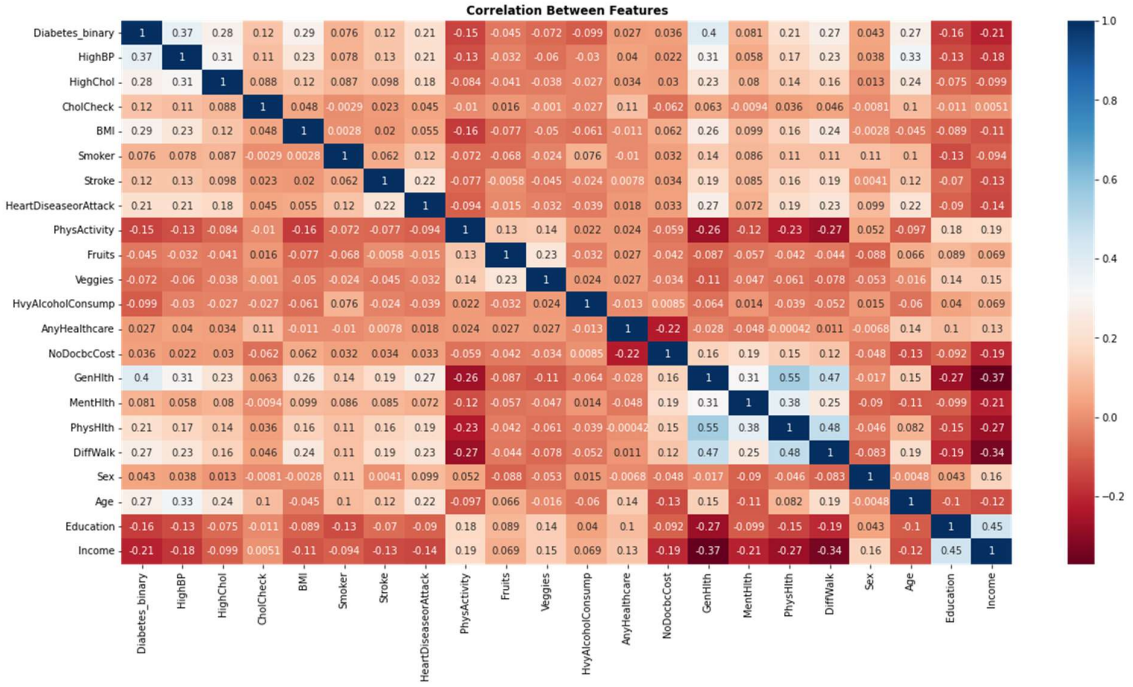
Data Description:

Transactions Count	:	70,692
Classification Class Positive %	:	50 %
Classification Class Negative %	:	50 %
No# of Features	:	21
Feature Description	:	

- **Diabetes_binary**: This is the target column which declares that an individual has no diabetes, prediabetes, or diabetes
- **HighBP and HighChol**: These two separate columns shows that if a person has blood pressure or cholesterol so it is of high category which is represented by 1 or normal denoted by 0.
- **CholCheck**: It checks that an individual has checked its cholesterol in 5 years by representing it by 1 and 0 for the opposite criteria.
- **BMI**: It is Body Mass Index of the individual. The maximum value is 98 and minimum value is 12.
- **Smoker**: It denotes the count of 100 cigarettes smoked by a person by 1 and 0 for person who smoked less than hundred cigarettes.
- **Stroke or HeartDiseaseorAttack**: These two columns show that people having stroked or heart disease such as coronial heart disease or myocardial infarction are shown by 1 and 0 for vice versa.
- **PhysActivity**: The value is 1 for those individuals involved in Physical Activity in last 30 days and 0 for no physical activity.
- **Fruits and Veggies**: These two columns give value as 1 for those who have consumed fruit or vegetable more than once per day.
- **HvyAlcoholConsump**: This column identifies for adult men having 14 or greater drinks per week and women having 7 or greater drinks per week as 1 for 'yes' and 0 for 'no'.
- **AnyHealthcare**: It gives the value 1 for those having health care coverage and 0 for vice versa.
- **NoDocbcCost**: This column identifies those individuals who didn't see a doctor for 12 months due to cost, 1 indicates yes while 0 shows no.
- **GenHlth**: It is the rating of one's general health from scale 1-5, 1 being the highest and 5 the lowest. This column is in ordinal manner.
- **MentHlth and PhysHlth**: These two columns are similar as they keep a check of having poor mental health in past 30 days for MentHlth and for PhysHlth it keeps check of having any physical injury in past 30 days.
- **DiffWalk**: This column tells that if an individual is having difficulty in walking or climbing stairs, so the value is 1 otherwise 0.
- **Sex**: The value 0 is for female and 1 for male.
- **Age**: This column gives value 1 for age between 18 and 24, 9 for age between 60 and 64 and 13 for 80 years and older.
- **Education**: This column indicates the level of education from a scale 1 – 6.
- **Income**: This column shows the level of income on a scale 1 – 8.

Data Pre-Processing:

- Understanding Data using Visualizations / EDA:



Insights:

1) Based on this visualization, we can conclude that HighBP, BMI, GenHlth and Age are some of the features positively correlated to Diabetes_binary while Income, Education, PhysActivity are some of the negatively correlated feature

2) A person with Diabetes is also likely to have a problem of High BP and High Cholesterol

3) A person with Diabetes is not likely to have a heavy Alcohol consumption

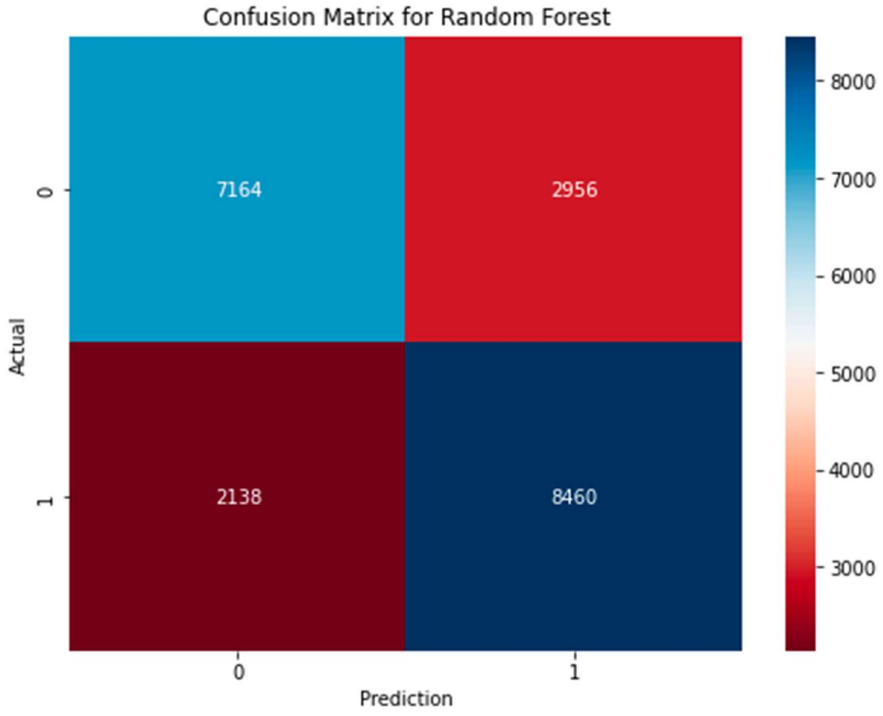
- Data Transformations (If needed):

Duplicate Records	:	Dropped
Null or NA Values	:	Data do not contain such values
Handling Typos or Strange Names	:	All columns have binary values
Training Data : Testing Data	:	70 : 30

Modeling:

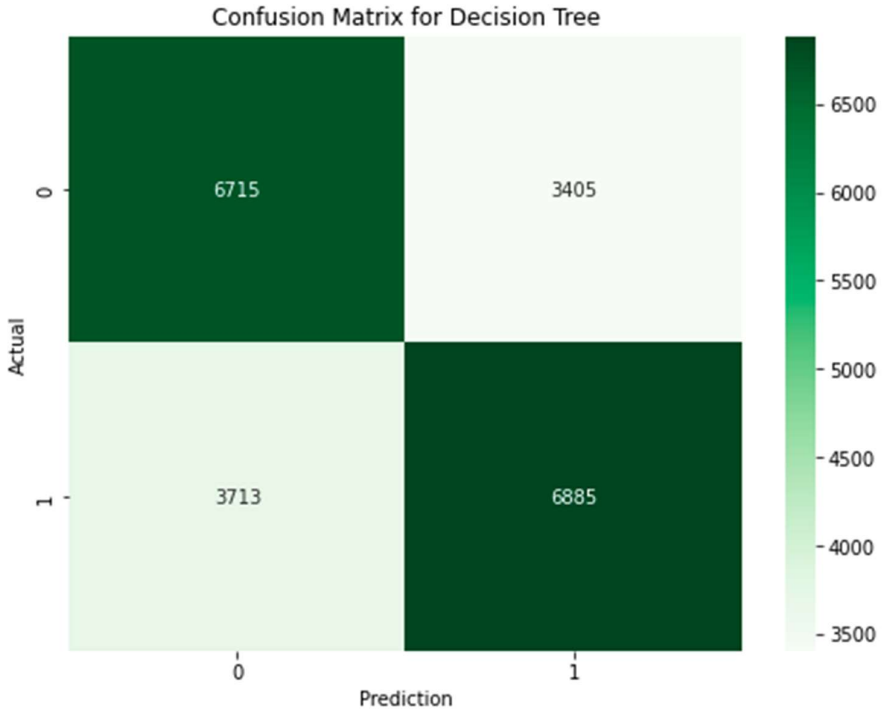
Under the process of modelling, I tried several known models such as Decision Trees, Random Forest, Logistic Regression and Naïve Bayes, while working on all of these models we got to experience different output either in terms of raising our accuracy for target variable or in decreasing as well.

- Random Forest:



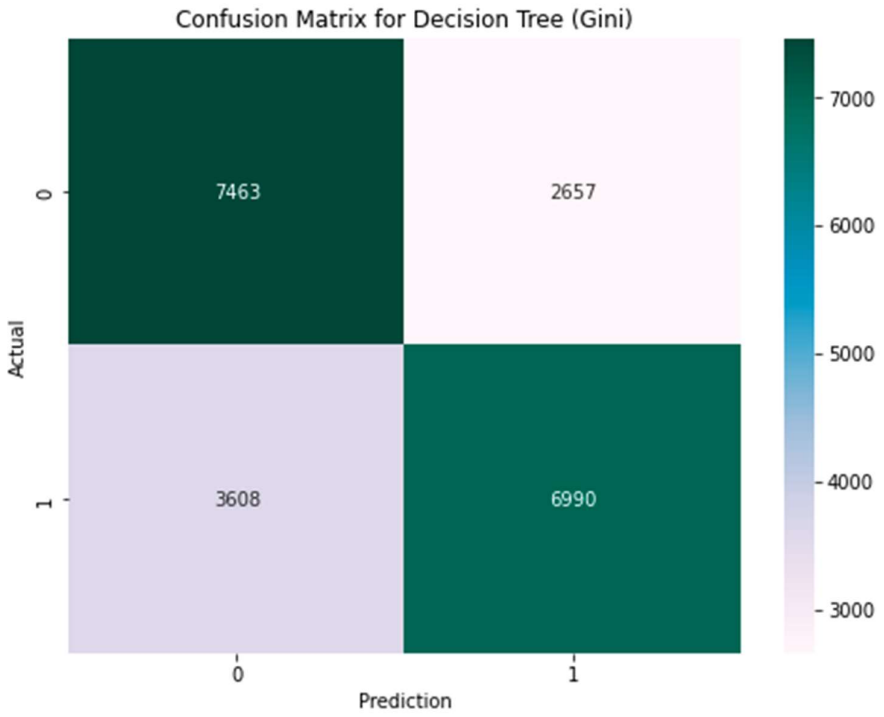
	precision	recall	f1-score	support
0.0	0.77	0.71	0.74	10120
1.0	0.74	0.80	0.77	10598
accuracy			0.75	20718
macro avg	0.76	0.75	0.75	20718
weighted avg	0.76	0.75	0.75	20718

- Decision Tree (Entropy):



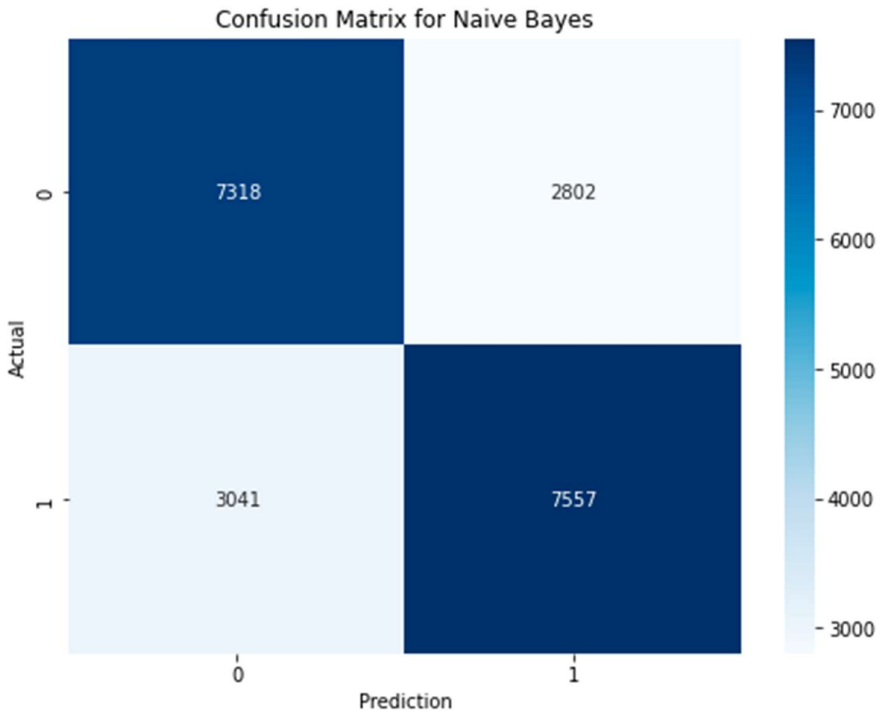
	precision	recall	f1-score	support
0.0	0.64	0.66	0.65	10120
1.0	0.67	0.65	0.66	10598
accuracy			0.66	20718
macro avg	0.66	0.66	0.66	20718
weighted avg	0.66	0.66	0.66	20718

- Decision Tree (Gini):



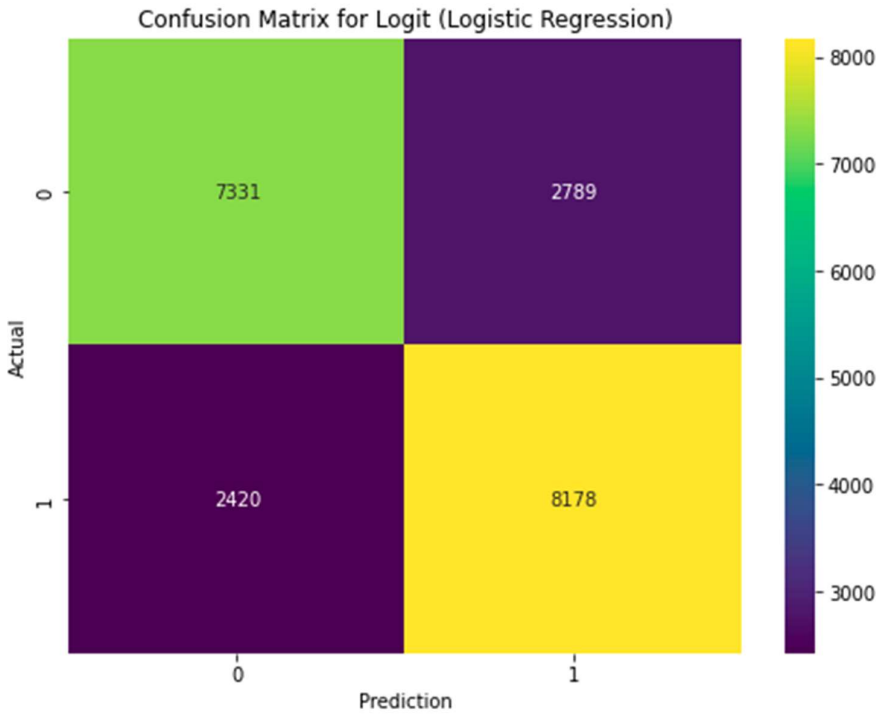
	precision	recall	f1-score	support
0.0	0.67	0.74	0.70	10120
1.0	0.72	0.66	0.69	10598
accuracy			0.70	20718
macro avg	0.70	0.70	0.70	20718
weighted avg	0.70	0.70	0.70	20718

- Naiive Bayes:



	precision	recall	f1-score	support
0.0	0.71	0.72	0.71	10120
1.0	0.73	0.71	0.72	10598
accuracy			0.72	20718
macro avg	0.72	0.72	0.72	20718
weighted avg	0.72	0.72	0.72	20718

- Logistic Regression (Log-It):



	precision	recall	f1-score	support
0.0	0.75	0.72	0.74	10120
1.0	0.75	0.77	0.76	10598
accuracy			0.75	20718
macro avg	0.75	0.75	0.75	20718
weighted avg	0.75	0.75	0.75	20718

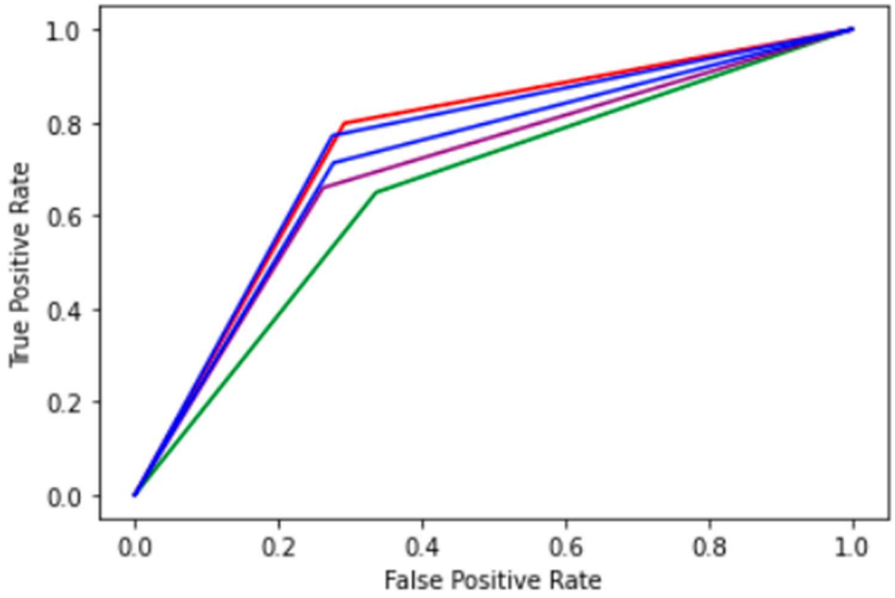
Model Comparison (Using ROC Curve):

Plotting of TPR (True Positive Rate) to the FPR (False Positive Rate) gives the ROC Curve. Near to the diagonal more inefficient the model is because the increase in TPR also cost FPR. Which should be minimized.

- Red: Random Forest
- Green: Decision Tree (Entropy)
- Purple: Decision Tree (Gini)
- Blue: Naiive Bayes
- Blue: Logistic Regression

```
plt.plot(fpr_rf,tpr_rf,color="red")
plt.plot(fpr_dt,tpr_dt,color="green")
plt.plot(fpr_Gdt,tpr_Gdt,color="purple")
plt.plot(fpr_nb,tpr_nb,color="blue")
plt.plot(fpr_lr,tpr_lr,color="blue")

plt.ylabel('True Positive Rate')
plt.xlabel('False Positive Rate')
```



According to the research the model will work more efficiently if we remove some of the factors which is making our predictions with more accuracy and precision.