Diabetes Prediction

Data Description:

Diabetes_binary : you have diabetes (0,1)

HighBP : Adults who have been told they have high blood pressure by a doctor, nurse, or other health professional (0,1)

HighChol : Have you EVER been told by a doctor, nurse or other health professional that your blood cholesterol is high? (0,1)

CholCheck : Cholesterol check within past five years (0,1)

BMI : Body Mass Index (BMI)

Smoker : Have you smoked at least 100 cigarettes in your entire life? [Note: 5 packs = 100 cigarettes] (0,1)

Stroke : (Ever told) you had a stroke. (0,1)

HeartDiseaseorAttack : Respondents that have ever reported having coronary heart disease (CHD) or myocardial infarction (MI) (0,1)

PhysActivity : Adults who reported doing physical activity or exercise during the past 30 days other than their regular job (0,1)

Fruits : Consume Fruit 1 or more times per day (0,1)

Veggies : Consume Vegetables 1 or more times per day (0,1)

HvyAlcoholConsump : Heavy drinkers (adult men having more than 14 drinks per week and adult women having more than 7 drinks per week)(0,1)

AnyHealthcare : Do you have any kind of health care coverage, including health insurance, prepaid plans such as HMOs, or government plans such as Medicare, or Indian Health Service? (0,1)

NoDocbcCost : Was there a time in the past 12 months when you needed to see a doctor but could not because of cost? (0,1)

GenHlth : Would you say that in general your health is: rate (1 ~ 5)

MentHlth : Now thinking about your mental health, which includes stress, depression, and problems with emotions, for how many days during the past 30 days was your mental health not good? (0 ~ 30)

PhysHlth : Now thinking about your physical health, which includes physical illness and injury, for how many days during the past 30 days was your physical health not good? (0 ~ 30)

DiffWalk : Do you have serious difficulty walking or climbing stairs? (0,1)

Sex : Indicate sex of respondent (0,1) (Female or Male)

Age : Fourteen-level age category (1 ~ 14)

Education : What is the highest grade or year of school you completed? (1 ~ 6)

Income : Is your annual household income from all sources: (If respondent refuses at any income level, code "Refused.") (1 ~ 8)

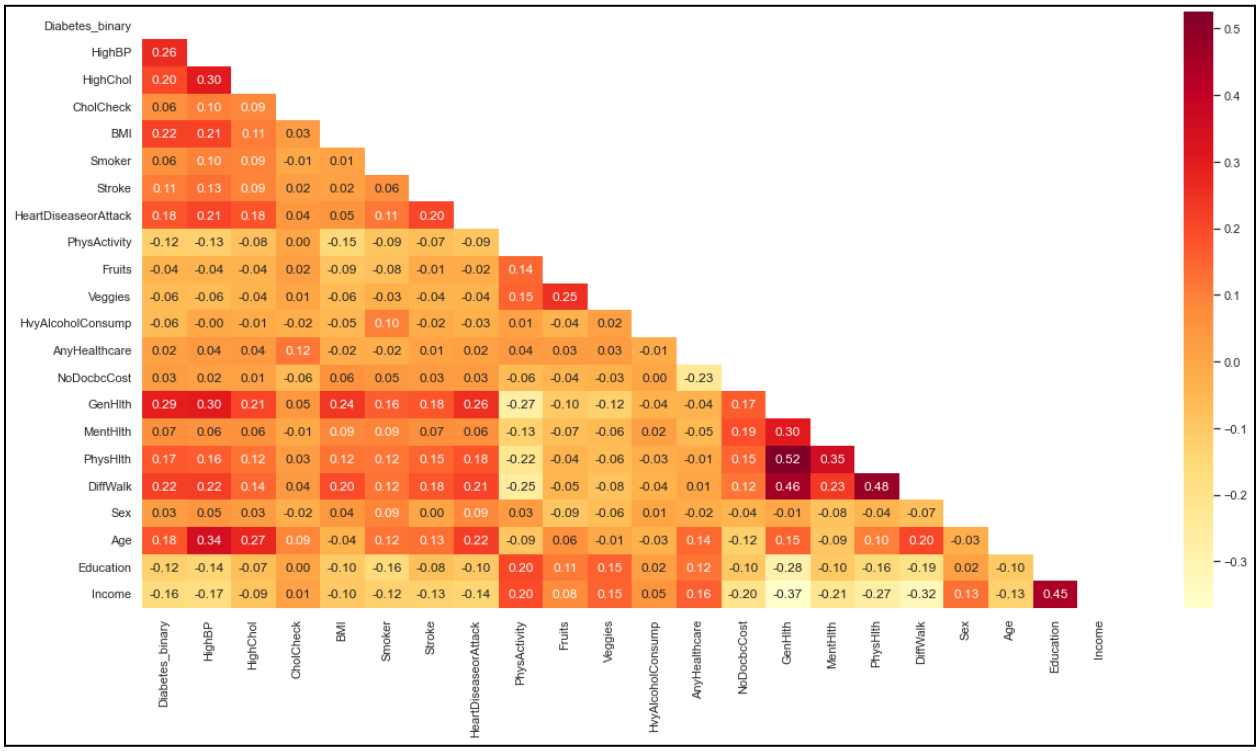
Data FIle: Diabetes_Health_Indicators_Dataset
Code File:

Data Exploring

Descriptive Statistics:

	count	mean	std	min	25%	50%	75%	max
Diabetes_binary	253680.0	0.139333	0.346294	0.0	0.0	0.0	0.0	1.0
HighBP	253680.0	0.429001	0.494934	0.0	0.0	0.0	1.0	1.0
HighChol	253680.0	0.424121	0.494210	0.0	0.0	0.0	1.0	1.0
CholCheck	253680.0	0.962670	0.189571	0.0	1.0	1.0	1.0	1.0
BMI	253680.0	28.382364	6.608694	12.0	24.0	27.0	31.0	98.0
Smoker	253680.0	0.443169	0.496761	0.0	0.0	0.0	1.0	1.0
Stroke	253680.0	0.040571	0.197294	0.0	0.0	0.0	0.0	1.0
HeartDiseaseorAttack	253680.0	0.094186	0.292087	0.0	0.0	0.0	0.0	1.0
PhysActivity	253680.0	0.756544	0.429169	0.0	1.0	1.0	1.0	1.0
Fruits	253680.0	0.634256	0.481639	0.0	0.0	1.0	1.0	1.0
Veggies	253680.0	0.811420	0.391175	0.0	1.0	1.0	1.0	1.0
HvyAlcoholConsump	253680.0	0.056197	0.230302	0.0	0.0	0.0	0.0	1.0
AnyHealthcare	253680.0	0.951053	0.215759	0.0	1.0	1.0	1.0	1.0
NoDocbcCost	253680.0	0.084177	0.277654	0.0	0.0	0.0	0.0	1.0
GenHlth	253680.0	2.511392	1.068477	1.0	2.0	2.0	3.0	5.0
MentHlth	253680.0	3.184772	7.412847	0.0	0.0	0.0	2.0	30.0
PhysHlth	253680.0	4.242081	8.717951	0.0	0.0	0.0	3.0	30.0
DiffWalk	253680.0	0.168224	0.374066	0.0	0.0	0.0	0.0	1.0
Sex	253680.0	0.440342	0.496429	0.0	0.0	0.0	1.0	1.0
Age	253680.0	8.032119	3.054220	1.0	6.0	8.0	10.0	13.0
Education	253680.0	5.050434	0.985774	1.0	4.0	5.0	6.0	6.0
Income	253680.0	6.053875	2.071148	1.0	5.0	7.0	8.0	8.0

Correlation Matrix:

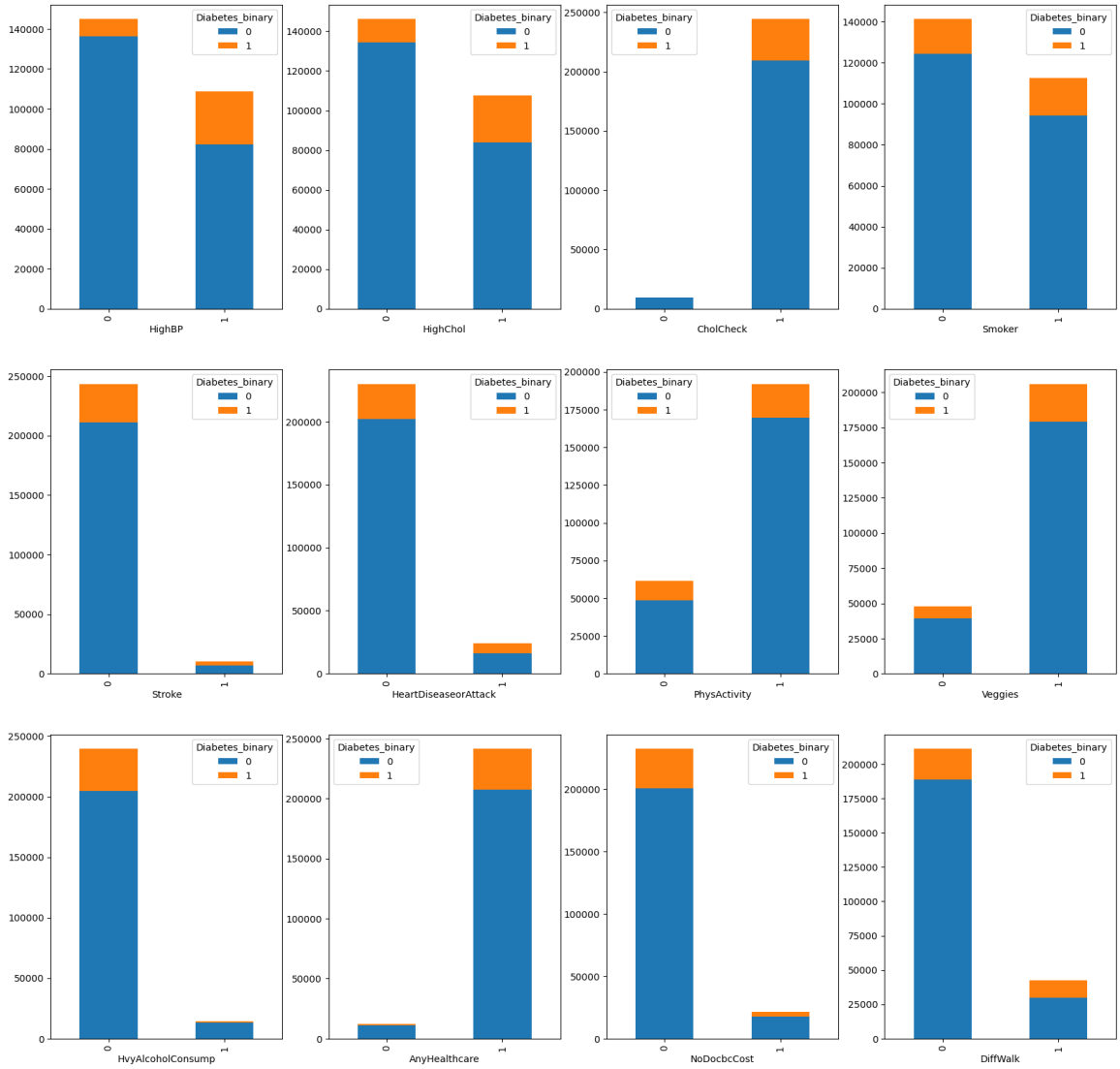


Correlation heatmap show relation between columns:

(GenHlth ,PhysHlth),(PhysHlth , DiffWalk),(GenHlth ,DiffWalk)are highly correlated with each other => **positive relation**

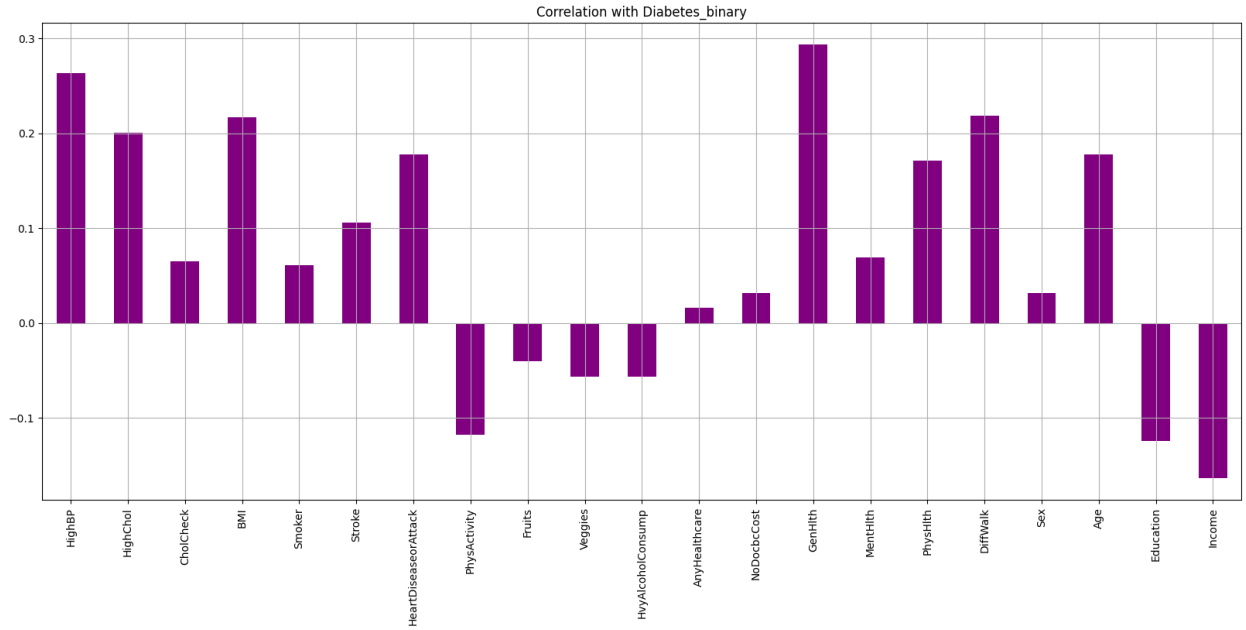
(GenHlth ,Income) , (DiffWalk , Income) are highly correlated with each other => **Negative relation**

Stack Plot (Diabetes):



Feature Selection

Correlation between Independent Features and Binary Target Diabetes:



Diabetes_binary's relation with other columns Through bar Graph Result:

1. Fruits , AnyHealthcare , NoDocbccost and sex are least correlated with Diabetes_binary.
2. HighBP , HighChol , BMI , smoker , stroke , HeartDiseaseorAttack , PhysActivity , Veggies , MentHlth , HvyAlcoholconsump , GenHlth , PhysHlth , Age , Education , Income and DiffWalk have a significant correlation with Diabetes_binary.

VIF Test:
Variance inflation factor (VIF) is a measure of multicollinearity in a regression model. It is calculated as the ratio of the variance of the estimated coefficient of a variable to the variance of the coefficient, if the variable were the only independent variable in the model.
The constant is included in the VIF calculation because the constant term is typically represented by a column of 1's in the matrix. It is added to the matrix to account for the intercept term in the regression equation. The intercept term is the value of the dependent variable when all of the explanatory variables are equal to 0.
Without the constant term, the VIF will be calculated incorrectly. This could lead to incorrect conclusions about the multicollinearity of the explanatory variables.

Here are some general guidelines for interpreting VIF values:

- VIF < 1: No multicollinearity
- VIF 1-5: Moderate multicollinearity
- VIF > 5: High multicollinearity

const	116.856706
Diabetes_binary	1.193120
HighBP	1.344502
HighChol	1.180932
CholCheck	1.033501
BMI	1.160280
Smoker	1.091872
Stroke	1.081612
HeartDiseaseorAttack	1.175776
PhysActivity	1.157396
Fruits	1.112540
Veggies	1.112397
HvyAlcoholConsump	1.025418
AnyHealthcare	1.113209
NoDocbcCost	1.144200
GenHlth	1.821914
MentHlth	1.239497
PhysHlth	1.623288
DiffWalk	1.536636
Sex	1.075748
Age	1.354954
Education	1.326495
Income	1.505649
dtype: float64	

VIF for all the features is less than 5 this shows that there is no multicollinearity in the data

ANOVA Test:

ANOVA can be used for feature selection by comparing the means of the target variable for each feature. The F-value is a measure of the significance of the ANOVA test. **A high F-value** indicates that the means of the groups are **significantly differen**t, while a low F-value indicates that the means of the groups are not significantly different.
A higher F-value indicates a stronger association.

10 best features selected by ANOVA Test:

	Scores	Feature
13	23924.564885	GenHlth
0	18870.365816	HighBP
16	12699.341579	DiffWalk
3	12516.718642	BMI
1	10600.350806	HighChol
18	8246.866284	Age
6	8231.555129	HeartDiseaseorAttack
15	7672.267690	PhysHlth
20	7004.370724	Income
19	3991.111142	Education

Chi-Square Test:

The chi-square test is a statistical test that can be used for feature selection in machine learning. It is used to determine whether there is a significant association between two categorical variables
The chi-square statistic is then compared to a critical value from a chi-square distribution. If the chi-square statistic is greater than the critical value, then the null hypothesis of independence is rejected and it is concluded that there is a significant association between the two variables.

10 best features selected by Chi-Square Test:

	Scores	Feature
15	133424.406534	PhysHlth
14	21029.632228	MentHlth
3	18355.166400	BMI
16	10059.506391	DiffWalk
0	10029.013935	HighBP
13	9938.507776	GenHlth
18	9276.141199	Age
6	7221.975378	HeartDiseaseorAttack
1	5859.710582	HighChol
20	4829.816361	Income

Features Selected By ANOVA Test and Chi-Square Test:

ANOVA Test	Chi-square Test
GenHlth HighBP DiffWalk BMI HighChol Age HeartDiseaseorAttack PhysHlth Income Education	PhysHlth MentHlth BMI DiffWalk HighBP GenHlth Age HeartDiseaseorAttack HighChol Income

- If you have continuous data and are building a linear model, then ANOVA is the best choice for feature selection.
- If you have categorical data and are building a classification model, then chi-square is the best choice for feature selection.
- In some cases, you may want to use both ANOVA and chi-square for feature selection. This can be done by first using ANOVA to select a subset of features and then using chi-square to select the best features from the subset.
- Ultimately, the best way to choose between ANOVA and chi-square for feature selection is to experiment with both methods and see which one works best for your data and your model.

Logistic regression:

Logistic regression is a statistical model that is used to predict the probability of a categorical outcome. The outcome can be binary, such as whether a patient has a disease or not, or it can be multi-class, such as whether a customer will buy a product or not.
Logistic regression is a supervised learning algorithm, which means that it requires a training dataset that includes both the independent variables and the outcome variable.
The accuracy of the model will depend on the quality of the training data.
Disadvantage:

- It is not a perfect model.
- It can be sensitive to outliers.
- It can be difficult to interpret the results.

most important **limitations**

- Assumption of Linearity => **Check the assumptions of linearity**
- Sensitivity to outliers => **Remove outliers**
- No Multicollinearity =>**Check for multicollinearity**
- Overfitting => **Use cross-validation:**

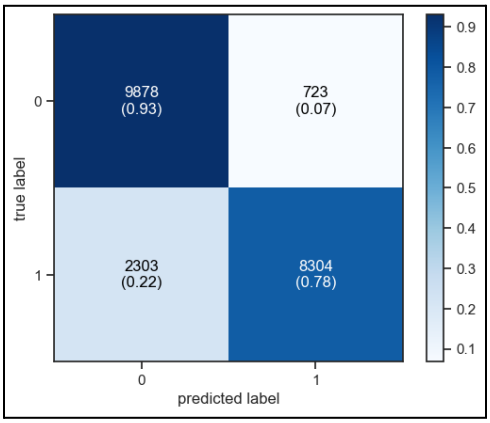
Logistic regression Result:

```
Training set score: 0.8596111874545308
Testing set score: 0.8573179932101094
------------------------------------------------------------
Mean Squared Error: 0.1426820067898906
Root MEan Square Error 0.3777327187177338
------------------------------------------------------------
              precision    recall  f1-score   support

           0       0.81      0.93      0.87     10601
           1       0.92      0.78      0.85     10607

    accuracy                           0.86     21208
   macro avg       0.87      0.86      0.86     21208
weighted avg       0.87      0.86      0.86     21208

------------------------------------------------------------
```



- There is 7% probability that the patients who actually does not have the Diabetes but the model predicted that the have diabetes (Type-I error)
- There is 22% probability that the patients who actually have the Diabetes but the model predicted that they does not have diabetes (Type-II error)
- In this case Type-II error is more dangerous.

- The training set score and testing score indicates that the model was able to correctly predict 85.96% of training cases and 85.73% testing cases. So,the model is a good fit.
- The mean squared error (MSE) of 0.14268and the root mean squared error (RMSE) of 0.3777 indicate that the model is not perfectly accurate, but it is still able to make reasonably good predictions
- The precision, recall, and f1-score for both the training and testing sets are all above 0.85, which indicates that the model is able to correctly identify both positive and negative cases with a high degree of accuracy.
- The accuracy of 0.86 for both the training and testing sets indicates that the model is able to correctly predict the outcome of the majority of cases.

Decision Tree

A decision tree is a supervised learning algorithm that can be used for both classification and regression problems. It is a tree-like model that consists of a root node, branches, internal nodes, and leaf nodes.

- The **root node** represents the entire population or sample.
- The **leaf nodes** represent the individual classes or predictions.
- The **branches and internal nodes** represent the decisions that are made to split the population or sample into smaller and smaller groups.

Decision trees are a powerful tool that can be used to solve a variety of problems.
One of the biggest limitations of decision trees is **overfitting.**
Decision trees are not suitable for all problems. They are best suited for problems where the **target variable can be broken down into a series of decisions.**

Logistic	Decision Tree
It is a **parametric** model, which means that it makes assumptions about the underlying distribution of the data.	It is a **non-parametric** model, which means that they do not make any assumptions about the underlying distribution of the data.
It is used when the relationship between the independent variables and the dependent variable is **linear**.	It can be used when the relationship between the independent variables and the dependent variable is **nonlinear**
Used when when **interpretability** is important	Used when **accuracy** is more important than interpretability.

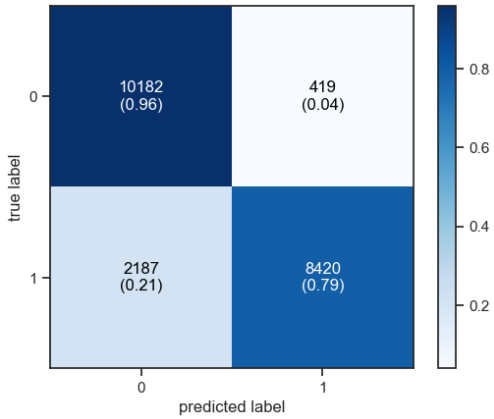
Decision Tree Result:

```
Training set score: 0.9132244765984965
Testing set score: 0.8771218408147868
------------------------------------------------------------
Mean Squared Error: 0.12287815918521312
Root MEan Square Error 0.35053981112737126
------------------------------------------------------------
              precision    recall  f1-score   support

           0       0.82      0.96      0.89     10601
           1       0.95      0.79      0.87     10607

    accuracy                           0.88     21208
   macro avg       0.89      0.88      0.88     21208
weighted avg       0.89      0.88      0.88     21208


------------------------------------------------------------
```



- There is 4% probability that the patients who actually does not have the Diabetes but the model predicted that the have diabetes (Type-I error)
- There is 21% probability that the patients who actually have the Diabetes but the model predicted that they does not have diabetes (Type-II error)
- In this case Type-II error is more dangerous.

- The training set score and testing score indicates that the model was able to correctly predict 91.32% of training cases and 87.71% testing cases. So,the model is a good fit.
- The mean squared error (MSE) of 0.1228 and the root mean squared error (RMSE) of 0.3505 indicate that the model is not perfectly accurate, but it is still able to make reasonably good predictions
- The precision, recall, and f1-score for both the training and testing sets are all above 0.88, which indicates that the model is able to correctly identify both positive and negative cases with a high degree of accuracy.
- The accuracy of 0.88 for both the training and testing sets indicates that the model is able to correctly predict the outcome of the majority of cases.

RandomForest

Random forest is an ensemble learning method for classification.

Ensemble learning: - Ensemble learning is a machine learning technique that combines the predictions of multiple models to improve the overall accuracy of the prediction. It is often used when a single model is not able to achieve the desired level of accuracy.

In a random forest, each decision tree is trained on a random subset of the features. This helps to reduce the variance of the model and improve its accuracy.

- Random forests are a way of averaging multiple decision trees, trained on different parts of the same training set, with the goal of reducing the variance.
- Forests are like the pulling together of decision tree algorithm efforts. Taking the teamwork of many trees thus improving the performance of a single random tree.

forests give the effects of a k-fold cross validation.

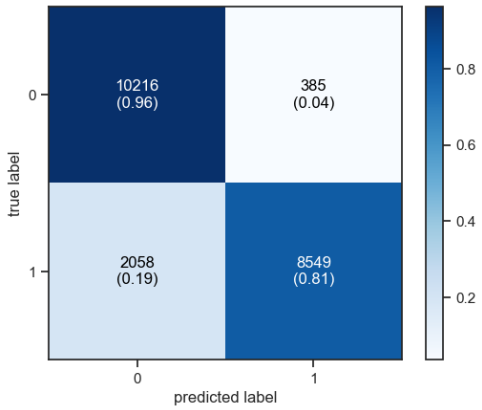
Random Forest Result:

```
Training set score: 0.889863390186727
Testing set score: 0.884807619766126
------------------------------------------------------------
Mean Squared Error: 0.11519238023387401
Root MEan Square Error 0.33940002980829864
------------------------------------------------------------
              precision    recall  f1-score   support

           0       0.83      0.96      0.89     10601
           1       0.96      0.81      0.87     10607

    accuracy                           0.88     21208
   macro avg       0.89      0.88      0.88     21208
weighted avg       0.89      0.88      0.88     21208

------------------------------------------------------------
```



- There is 4% probability that the patients who actually does not have the Diabetes but the model predicted that the have diabetes (Type-I error)
- There is 19% probability that the patients who actually have the Diabetes but the model predicted that they does not have diabetes (Type-II error)
- In this case Type-II error is more dangerous.

- The training set score and testing score indicates that the model was able to correctly predict 88.98% of training cases and 88.48% testing cases. So,the model is a good fit.
- The mean squared error (MSE) of 0.1151 and the root mean squared error (RMSE) of 0.3393 indicate that the model is not perfectly accurate, but it is still able to make reasonably good predictions
- The precision, recall, and f1-score for both the training and testing sets are all above 0.88, which indicates that the model is able to correctly identify both positive and negative cases with a high degree of accuracy.
- The accuracy of 0.88 for both the training and testing sets indicates that the model is able to correctly predict the outcome of the majority of cases.

Support vector machines (SVM)

Support vector machines (SVMs) are a type of supervised learning algorithm that can be used for both **classification and regression** tasks. SVMs are based on the idea of finding a hyperplane that separates the data points into two classes. The hyperplane is chosen such that the distance between the hyperplane and the closest data points on either side is maximized. This distance is called the margin.

SVMs are known for their **high accuracy** and their ability to **handle large datasets**. However, they can be computationally expensive to train and can be sensitive to the choice of hyperparameters.

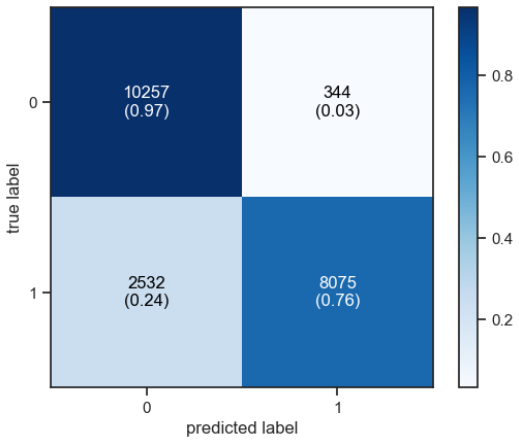
Support vector machines (SVM) Result:

```
Training set score: 0.8652089564303613
Testing set score: 0.8643907959260656
-----------------------------------------------------------
Mean Squared Error: 0.13560920407393437
Root MEan Square Error 0.3682515499952911
-----------------------------------------------------------
              precision    recall  f1-score   support

           0       0.80      0.97      0.88     10601
           1       0.96      0.76      0.85     10607

    accuracy                           0.86     21208
   macro avg       0.88      0.86      0.86     21208
weighted avg       0.88      0.86      0.86     21208


-----------------------------------------------------------
```



- There is 3% probability that the patients who actually does not have the Diabetes but the model predicted that the have diabetes (Type-I error)
- There is 24% probability that the patients who actually have the Diabetes but the model predicted that they does not have diabetes (Type-II error)
- In this case Type-II error is more dangerous.

- The training set score and testing score indicates that the model was able to correctly predict 86.52% of training cases and 86.43% testing cases. So,the model is a good fit.
- The mean squared error (MSE) of 0.1356 and the root mean squared error (RMSE) of 0.3682 indicate that the model is not perfectly accurate, but it is still able to make reasonably good predictions
- The precision, recall, and f1-score for both the training and testing sets are all above 0.86, which indicates that the model is able to correctly identify both positive and negative cases with a high degree of accuracy.
- The accuracy of 0.86 for both the training and testing sets indicates that the model is able to correctly predict the outcome of the majority of cases.

XGBoost

XGBoost is an open-source software library for machine learning that uses the gradient boosting algorithm. It is one of the most popular machine learning libraries in the world, and it has been used to win many machine learning competitions.
XGBoost is a powerful algorithm that can be used for a variety of tasks, including classification, regression, and ranking. It is also very efficient, and it can be used to train models on large datasets.

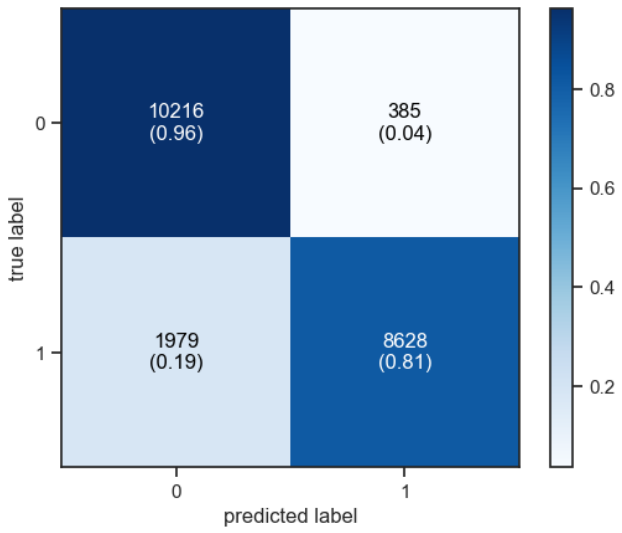
XGBoost Result:

```
Training set score: 0.8972193032091181
Testing set score: 0.8885326291965296
------------------------------------------------------------
Mean Squared Error: 0.11146737080347038
Root MEan Square Error 0.3338672951989613
------------------------------------------------------------
              precision    recall  f1-score   support

           0       0.84      0.96      0.90     10601
           1       0.96      0.81      0.88     10607

    accuracy                           0.89     21208
   macro avg       0.90      0.89      0.89     21208
weighted avg       0.90      0.89      0.89     21208

------------------------------------------------------------
```



- There is 4% probability that the patients who actually does not have the Diabetes but the model predicted that the have diabetes (Type-I error)

- There is 19% probability that the patients who actually have the Diabetes but the model predicted that they does not have diabetes (Type-II error)
- In this case Type-II error is more dangerous

- The training set score and testing score indicates that the model was able to correctly predict 89.72% of training cases and 88.88% testing cases. So,the model is a good fit.
- The mean squared error (MSE) of 0.1114 and the root mean squared error (RMSE) of 0.3338 indicate that the model is not perfectly accurate, but it is still able to make reasonably good predictions
- The precision, recall, and f1-score for both the training and testing sets are all above 0.89, which indicates that the model is able to correctly identify both positive and negative cases with a high degree of accuracy.
- The accuracy of 0.89 for both the training and testing sets indicates that the model is able to correctly predict the outcome of the majority of cases.

SVM V/s XGBoost

- **Consider the size of the dataset:** If you have a small dataset, then you may want to use **SVM**, as it is less computationally expensive to train. However, if you have a large dataset, then you may want to use **XGBoost,** as it can be more accurate.
- **Consider the accuracy requirements:** If you need a very accurate model, then you may want to use **XGBoost**. However, if you are willing to sacrifice some accuracy for simplicity, then you may want to use **SVM**.
- **Consider the time constraints:** If you are under time constraints, then you may want to use **SVM**, as it is less computationally expensive to train. However, if you have more time, then you may want to use **XGBoost**, as it can be more accurate.

Advantages and Disadvantages:

Algorithm	Advantage	Disadvantage
Logistic Regression	Simple to understand & interpret	Sensitive to outliers.
Decision Tree	Easy to train & measure accuracy	Overfit to the training data
Random Forest	Can reduce overfitting and improve accuracy	Computationally expensive to train
SVM	High accuracy, robust to noise & outliers, scalable.	Computationally expensive to train, sensitive to hyperparameters, difficult to interpret
XGBoost	High accuracy, robust to noise & outliers, fast scalable	Computationally expensive to train, sensitive to hyperparameters, difficult to interpret

Summary:

Algorithm	Training score	Testing score	RMSE	Type-I error	Type-II error	Classification Metrics (recall)	Model accuracy score
Logistic Regression	0.8596	0.8573	0.3777	0.07	0.22	0.86	0.86
Decision Tree	0.9132	0.8771	0.3505	0.04	0.21	0.88	0.88
Random Forest	0.8898	0.8848	0.3394	0.04	0.19	0.88	0.88
SVM	0.8652	0.8643	0.3682	0.03	0.24	0.86	0.86
XGBoost	0.8972	08885	0.3338	0.04	0.19	0.89	0.89

Conclusions:

By comparing all the above models we can say that XGBoost provides best results as compared to others.