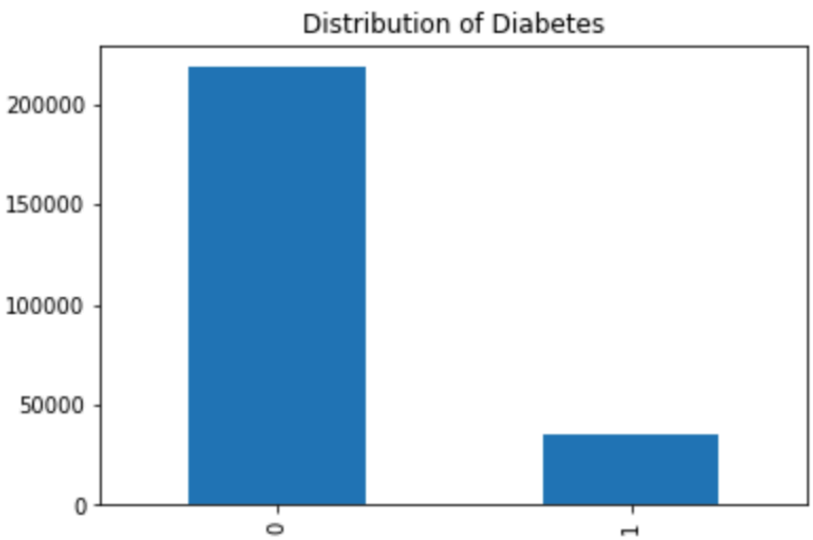
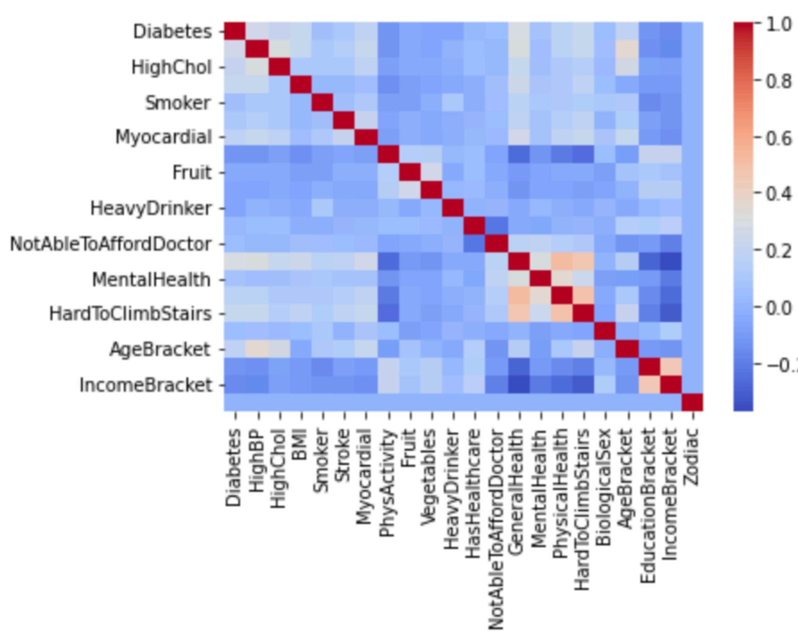
Data Preprocessing

It is stated that the data is carefully curated which indicates that there should not be too much missing data. Conducting exploratory analysis, it is suggested that the dataset is imbalanced as the number of individuals that have diabetes is very much lower than the number of individuals that do not have diabetes. This indicates that I should be including a class_weight parameter when running the models.

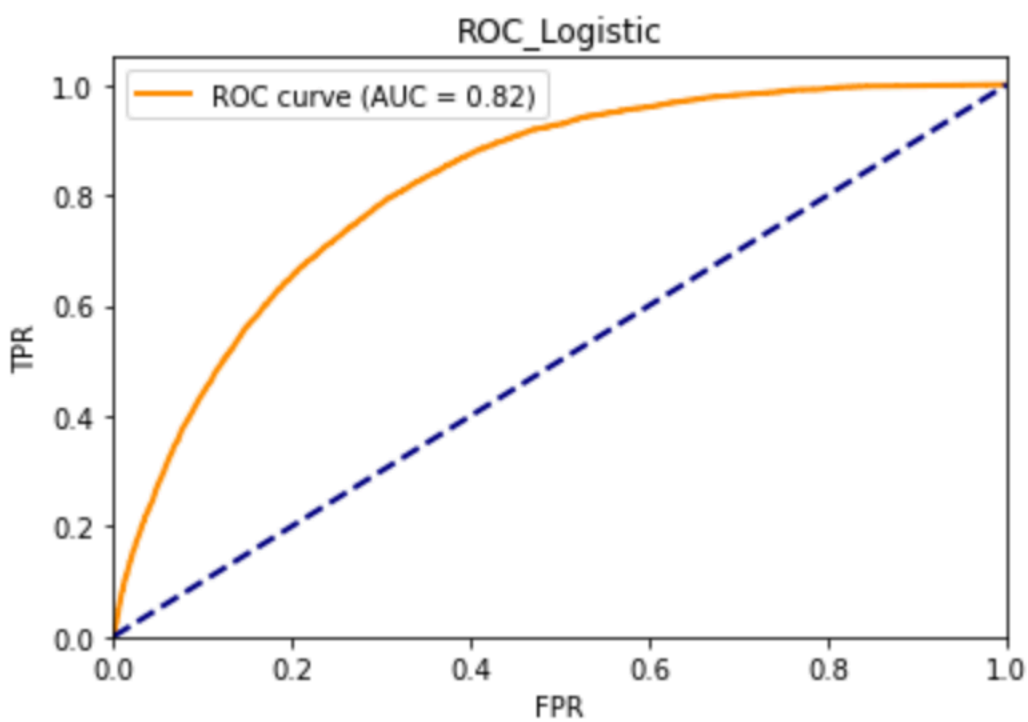




I drop the Diabetes column as the outcome variable and use the rest of the dataframe as the predictor variables. Then I do a train_test_split with a 20% ratio on test data. Looking at the correlation matrix above, there are no strong correlation between features.

Question 1

I build a logistic regression model with the parameters: solver = "saga", max_iter = 2500 and class_weight = "balanced", fit it with X_train and y_train. First, I obtained array of the predicted probabilities of the positive class for each sample in X_test with predict_proba function. I compute the AUC score with the roc_auc_score and display the ROC curve. To determine which among the predictor variables is the best predictor for diabetes, I calculate the feature importance using permutation_importace function. Selecting "saga" as the solver instead of the other is due to the consideration of the large number of sample. "Saga" utilize an adaptive learning rate and enable faster convergence than other solvers on large-scale dataset, which is suitable for this case as the number of sample for this dataset even after train_test_split is 2 million for the X_train and 50 thousands for the X_test. Moreover, as the heatmap above that indicate the correlation between the predictor variables is not very strong, hence I do not consider choosing the "newton-cg" as the solver. Considering that the outcome variable is imbalanced, class_weight = "balanced" is included to adjust the weights of the training samples so that each class is given equal weight during training. Utilizing the permutation_importance function enable me to assess the importance of each feature in the trained logistic regression model. Considering between feature_importance and permutation_importance, the latter is more robust, unbiased and model agnostic method for identifying the most important features in a model.



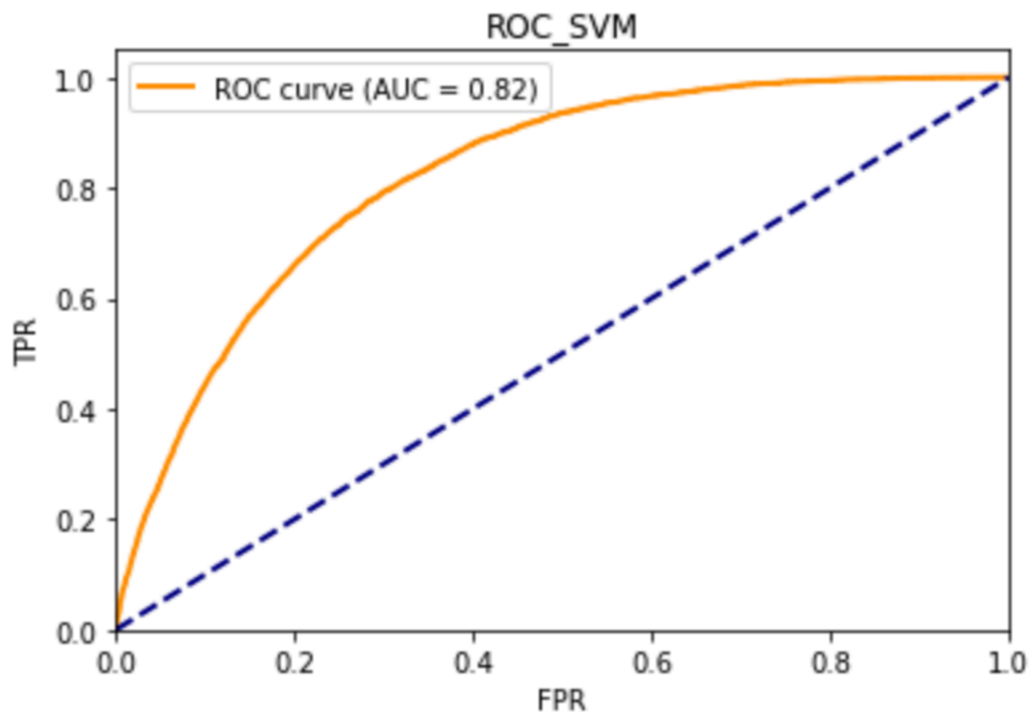
Above is the ROC curve for the model with an AUC score of 0.82156. The logistic regression model has a F1 score of 0.43939, Accuracy of 0.72966, Precision of 0.30802 and Recall of 0.76611. Although the question only require me to indicate the AUC score for the model but to determine the overall performance of the model it will not be sufficient to only include AUC score.

```
Feature Importances:
GeneralHealth: 0.009094134342478754
BMI: 0.005292100283822177
HeavyDrinker: 0.0010446231472721723
Myocardial: 0.0007154683065279221
BiologicalSex: 0.0001596499526963613
Stroke: 0.00015767896562601802
NotAbleToAffordDoctor: -5.912961210929879e-06
PhysActivity: -6.504257332070606e-05
Smoker: -8.672343109424884e-05
Zodiac: -0.0001281141595710910 7
HardToClimbStairs: -0.0001419110690633496
MentalHealth: -0.0002798801639861015
IncomeBracket: -0.00031141595711129 4
PhysicalHealth: -0.0004119362976978791
HasHealthcare: -0.0004217912330495066
Vegetables: -0.0004277041942604475 7
Fruit: -0.0004888047934405226
EducationBracket: -0.0005203405865657374
HighChol: -0.004921554714601051
AgeBracket: -0.007964758751182577
HighBP: -0.010097366761274018
```

This chart obtained from running the permutation_importance function indicates that the “GeneralHealth” is the best predictor for Diabetes. However, right below the “GeneralHealth” variable, the “BMI” also has a high importance very close to “GeneralHealth”. Hence, it is to conclude that both “GeneralHealth” and “BMI” are the best predictors for diabetes in the logistic regression model.

Question 2

Utilizing the same process as running the logistic regression model, I build a svm using LinearSVC function with parameter: dual = False and class_weight = “balanced” wrapped in a CalibratedClassiferCV function. Since I have to determine the AUC score for the svm, utilizing CalibratedClassiferCV function enable me to calibrate the probability estimates of a classifier since LinearSVC function does not output probabilities directly and allow the svm to run the predict_proba function. Setting dual = False will enable the model to run faster and more memory efficient whereas the class_weight parameter is included as there exist imbalance in outcome variable.



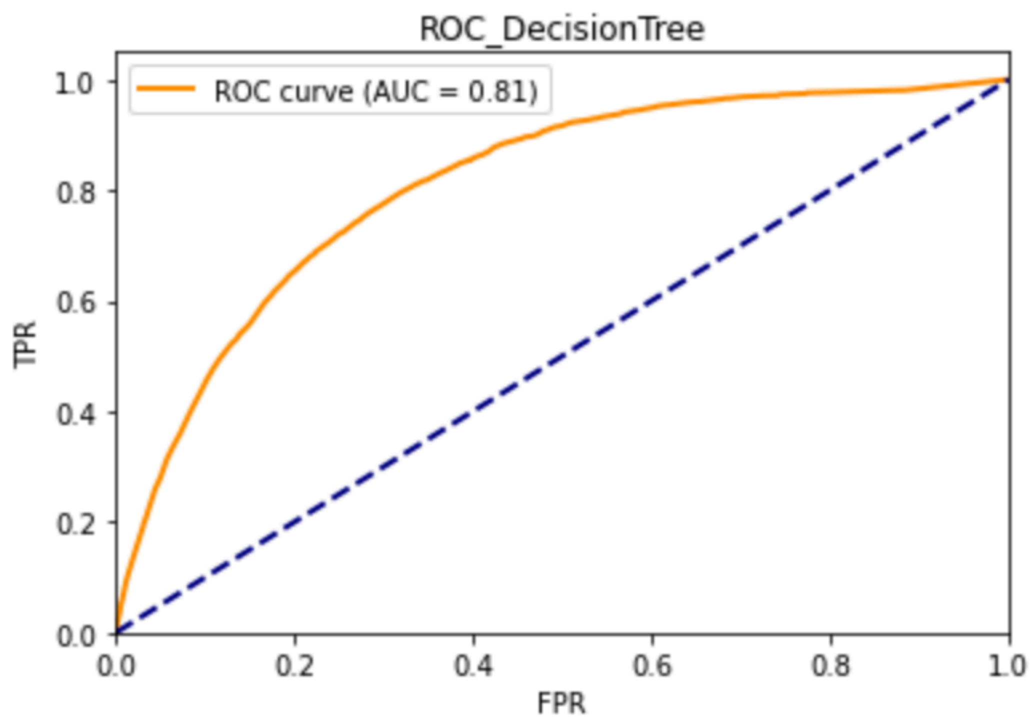
Above is the ROC curve for the model with an AUC score of 0.82121. The svm model has a F1 score of 0.22841, Accuracy of 0.86337, Precision of 0.52134 and Recall of 0.14624. Although the question only require me to indicate the AUC score for the model but to determine the overall performance of the model it will not be sufficient to only include AUC score.

```
Feature Importances:
BMI: 0.005743456322926466
GeneralHealth: 0.004007016713970302
HighBP: 0.0012298959318826409
HighChol: 0.0010564490696940766
AgeBracket: 0.0007943077893408645
PhysicalHealth: 0.0006878744875433052
BiologicalSex: 0.00037251655629135796
HeavyDrinker: 0.00029367707347834895
MentalHealth: 0.0001162882371491425
HasHealthcare: 0.00011234626300847817
Fruit: 7.883948281250941e-06
Zodiac: 0.0
Stroke: -5.551115123125783e-17
NotAbleToAffordDoctor: -1.7738883632933966e-05
Smoker: -1.9709870703288336e-05
Vegetables: -6.307158625044052e-05
PhysActivity: -0.0001222011983601612
EducationBracket: -0.00025031535793127444
HardToClimbStairs: -0.00033112582781458233
IncomeBracket: -0.0004158782718385767
Myocardial: -0.000683932513402763
```

This chart obtained from running the permutation_importance function indicates that the “BMI” and “GeneralHealth” are both the best predictor for Diabetes for single vector machine. This is concluded as the feature importance the both predictor variables are overly close.

Question 3

Utilizing the same process as running the logistic regression model, I build a single decision tree using DecisionTreeClassifier function with parameter: criterion = entropy, max_depth = 10 and class_weight = “balanced”. Setting max_depth to 10 which is not too low and not too high to prevent overfitting and underfitting. Setting class_weight to “balanced” help in balancing the contribution of each class to the training of the model, especially as indicated that the dataset is imbalanced. Both “gini” and “entropy” are the common criterion for classification. However, “entropy” is more sensitive to information gain which will be a better choice when the distribution of classes is imbalanced.



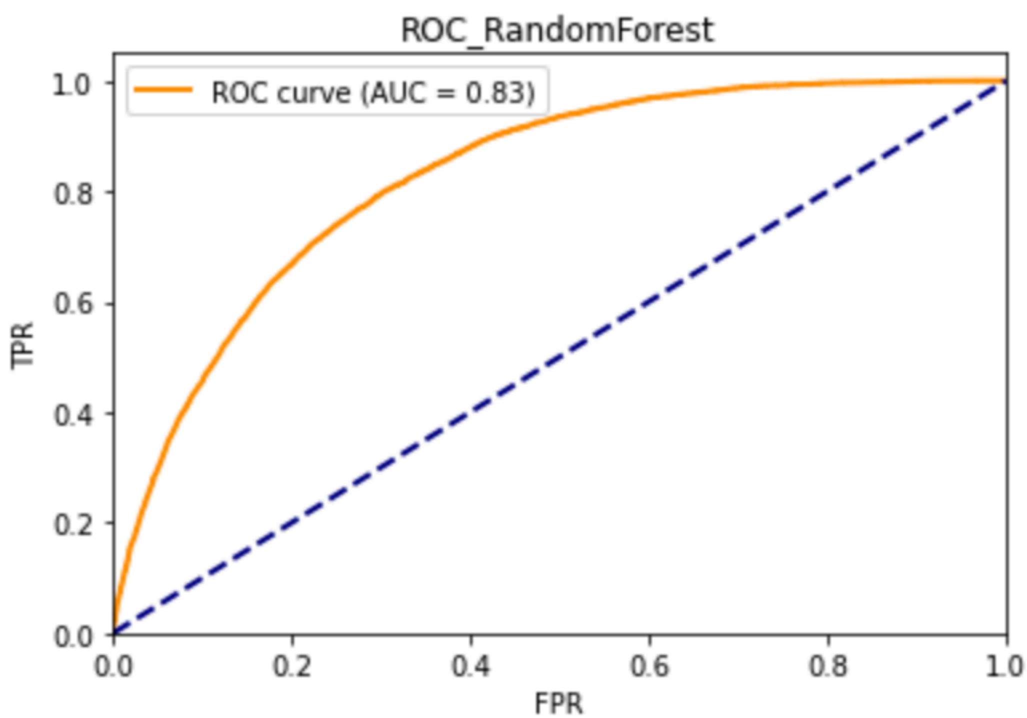
Above is the ROC curve for the model with an AUC score of 0.80896. The single decision tree model has a F1 score of 0.43030, Accuracy of 0.72146, Precision of 0.3 and Recall of 0.76069. Although the question only require me to indicate the AUC score for the model but to determine the overall performance of the model it will not be sufficient to only include AUC score.

```
Feature Importances:
GeneralHealth: 0.023470514033427946
BMI: 0.012409334594765054
AgeBracket: 0.007373462630085137
IncomeBracket: 0.0029150898770104127
HardToClimbStairs: 0.0028520182907599946
PhysicalHealth: 0.0019236833806370091
Myocardial: 0.0018744087038788937
Stroke: 0.0011766792809839166
HeavyDrinker: 0.0005183695994954273
BiologicalSex: 0.0005065436770734677
HighChol: 0.0004217912330495177
Fruit: 0.00032915484074422794
Smoker: 0.00026608325449385407
HasHealthcare: 0.00015373699148532038
PhysActivity: 0.00014979501734466716
Vegetables: 0.00010643330179754829
MentalHealth: 4.336171554712998e-05
NotAbleToAffordDoctor: 2.3651844843908253e-05
EducationBracket: -0.0001635919268369479
Zodiac: -0.0002483443708609312
HighBP: -0.003742904446546835
```

This chart obtained from running the permutation_importance function indicates that the "GeneralHealth" and "BMI" are both the best predictor for Diabetes for single decision tree. This is concluded as the feature importance the both predictor variables are overly close. However, the "AgeBracket" is closely following both the predictor stated above. Hence, for single decision tree, it can be suggested that there are three predictor variables that make best prediction for Diabetes, which is "GeneralHealth", "BMI" and "AgeBracket".

Question 4

Utilizing the same process as running the logistic regression model, I build a random forest using RandomForestClassifier function with parameter: criterion = entropy, max_depth = 10 and class_weight = “balanced”. Setting max_depth to 10 which is not too low and not too high to prevent overfitting and underfitting. Setting class_weight to “balanced” help in balancing the contribution of each class to the training of the model, especially as indicated that the dataset is imbalanced. Both “gini” and “entropy” are the common criterion for classification. However, “entropy” is more sensitive to information gain which will be a better choice when the distribution of classes is imbalanced.



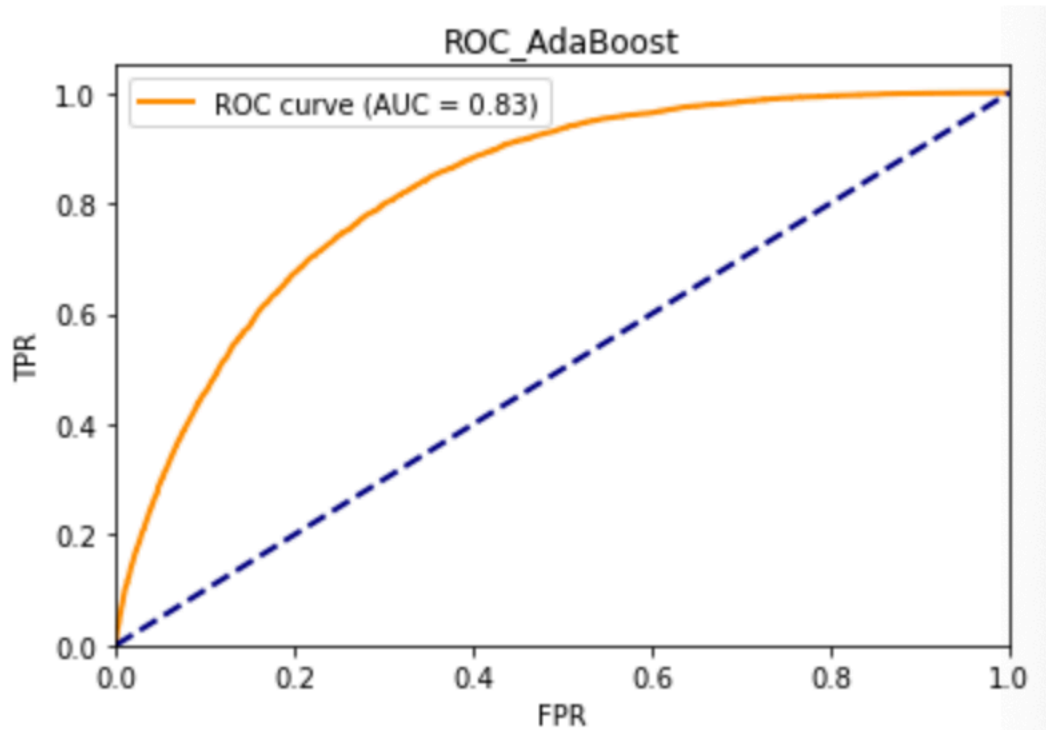
Above is the ROC curve for the model with an AUC score of 0.82559. The random forest model has a F1 score of 0.44212, Accuracy of 0.73354, Precision of 0.31115 and Recall of 0.76354. Although the question only require me to indicate the AUC score for the model but to determine the overall performance of the model it will not be sufficient to only include AUC score.

```
Feature Importances:
GeneralHealth: 0.00968740145064646
BMI: 0.005739514348785823
HardToClimbStairs: 0.0015176600441500709
HeavyDrinker: 0.0007509460737937012
Myocardial: 0.0005696152633238305
MentalHealth: 0.0002089246294543878 4
Fruit: 0.00016950488804788887
HasHealthcare: 0.00014585304320400284
NotAbleToAffordDoctor: 5.9129612109709574e-05
Stroke: 5.5187637969056345e-05
PhysActivity: 5.5187637969056345e-05
Smoker: -2.7593818984583684e-05
Vegetables: -0.00015570797855570805
Zodiac: -0.0003863134657837164
BiologicalSex: -0.0003902554399243696
PhysicalHealth: -0.0005341374960580846
EducationBracket: -0.0008238725953958248
IncomeBracket: -0.0009421318196153105
HighChol: -0.0009815515610218095
AgeBracket: -0.003161463260801067
HighBP: -0.009165089877010446
```

This chart obtained from running the permutation_importance function indicates that the "GeneralHealth" and "BMI" are both the best predictor for Diabetes for random forest. This is concluded as the feature importance the both predictor variables are overly close.

Question 5

Utilizing the same process as running the logistic regression model, I build a adaBoost using AdaBoostClassifier function with parameter: max_depth = 1 and class_weight = “balanced”. Setting max_depth to 1 indicating there is a weak classifier in this case single decision tree with a single split. This will prevent overfitting when there are many features in the dataset and allowing a more generalizable decision rule for the model as the model will not memorize the training data. Setting class_weight to “balanced” help in balancing the contribution of each class to the training of the model, especially as indicated that the dataset is imbalanced. Both “gini” and “entropy” are the common criterion for classification. Including criterion or not including criterion will result in the same performance result indicating no necessity in including criterion for this model.



Above is the ROC curve for the model with an AUC score of 0.82613. The adaBoost Classifier model has a F1 score of 0.44274, Accuracy of 0.73098, Precision of 0.31024 and Recall of 0.77281. Although the question only require me to indicate the AUC score for the model but to determine the overall performance of the model it will not be sufficient to only include AUC score.

```
Feature Importances:
BMI: 0.006484547461368595
GeneralHealth: 0.004269157994323503
HeavyDrinker: 0.0011412015137180153
Myocardial: 0.0005676442762534761
EducationBracket: 0.00020892462945438784
HardToClimbStairs: 0.00017541784925886316
BiologicalSex: 0.00015373699148528707
HasHealthcare: 8.869441816458101e-05
Stroke: 1.379690949222523e-05
MentalHealth: 7.883948281250941e-06
Fruit: 0.0
Smoker: 0.0
PhysActivity: 0.0
Zodiac: 0.0
Vegetables: 0.0
NotAbleToAffordDoctor: 0.0
PhysicalHealth: 0.0
IncomeBracket: -0.0003823714916430632
HighChol: -0.003196941028066913
AgeBracket: -0.0036758908861558416
HighBP: -0.007529170608640867
```

This chart obtained from running the permutation_importance function indicates that the “BMI” and “GeneralHealth” are both the best predictor for Diabetes for adaBoost. This is concluded as the feature importance the both predictor variables are overly close.

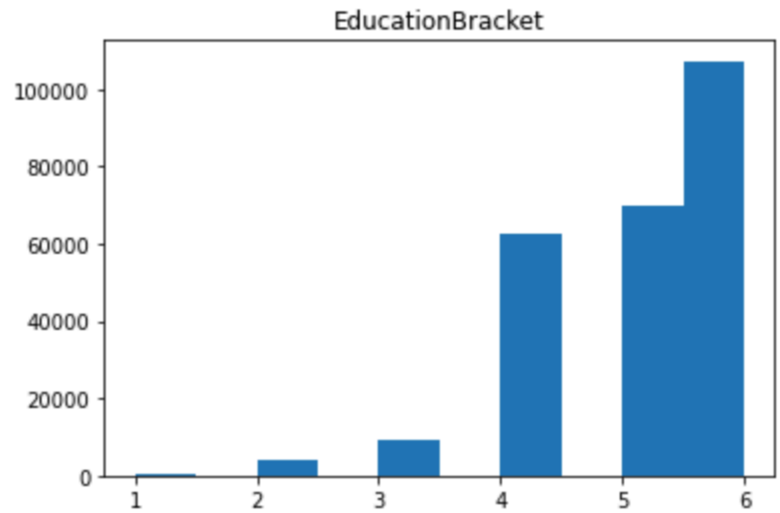
Extra Credit 1

	AUC Score	F1 Score	Accuracy	Precision	Recall
Logistic Regression	0.82156	0.43939	0.72966	0.30802	0.76611
SVM	0.82121	0.22841	0.86337	0.52134	0.14624
Single Decision Tree	0.80896	0.43030	0.72146	0.3	0.76069
Random Forest	0.82559	0.44212	0.73354	0.31115	0.76354
adaBoost Classifier	0.82613	0.44274	0.73098	0.31024	0.77281

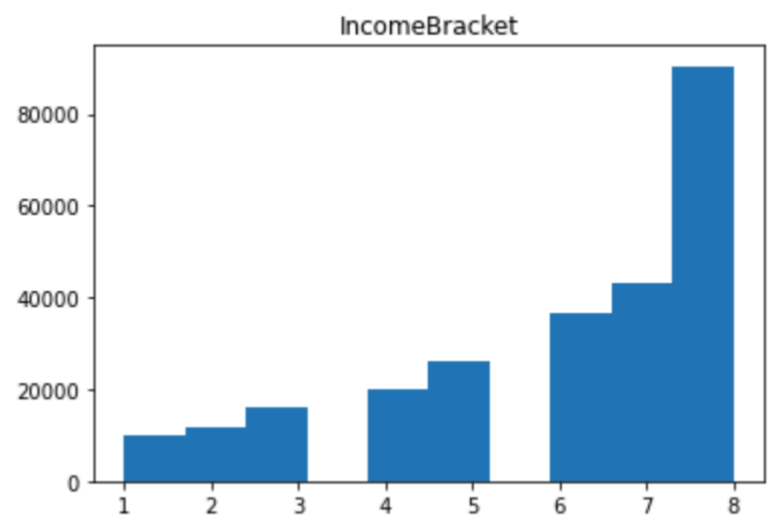
According to the conclusion table above, the evaluation metrics indicated that the best model to predict diabetes in this dataset is the AdaBoost Classifier model. This is due to the reason that the AUC score for the model is the highest among the models which indicates that AdaBoost has a better overall performance in distinguishing between positive and negative cases. The weighted average of precision and recall, which is the F1 score, is also the highest for the AdaBoost Classifier model. This implies that the model has the best balance between precision and recall, resulting better at correctly identifying positive cases and avoiding false positives. Although, the accuracy score for the AdaBoost Classifier is not the highest among the models, but it is relatively high. This indicates that the model are able to correctly classify a large proportion of cases. Moreover, the recall among the model is also the highest. This implies that the model is able to correctly identify large proportion of positive cases out of all the actual positive cases in the dataset. On the other hand, the precision is not the highest but reasonable for AdaBoost. This scenario may be due to the reason of the trade off that high recall score may come at a cost of a lower precision score.

Extra Credit 2

Observing the correlation matrix heatmap displayed in the data preprocessing section, it is interesting to conclude that there is a rather strong positive correlation between EducationBracket and IncomeBracket. This could be suggesting that the higher the education level would indicate a higher income. There is also a strong negative correlation between IncomeBracket and GeneralHealth. This could be suggesting that the higher the individual's income the better is their health. Performing explanatory analysis, I obtained some interesting result.



The plot indicates that in this dataset there are a relatively large number of samples that are educated individuals that college graduates. This may be the sole cause that there are large number of sample that do not have diabetes, as college graduates that have a higher level of health literacy and awareness may be more likely to engage in healthy behaviors and better access to healthcare as indicated with the relatively negative correlation in the heatmap.



The plot indicates that in this dataset there are a relatively large number of samples that are individuals that have a relatively high income. This may be the sole cause that there are large number of sample that do not have diabetes, as individuals with higher income tend to be more educated implying that they may have a higher level of health literacy and awareness may be more likely to engage in healthy behaviors and better access to healthcare as indicated with the relatively negative correlation in the heatmap.