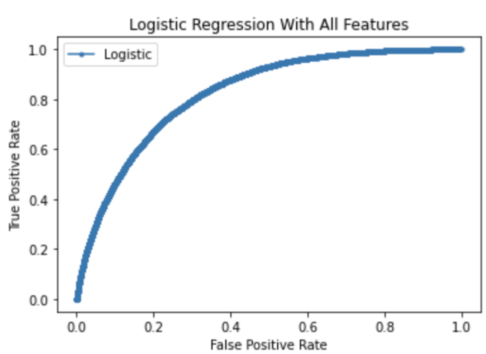
Question 1

1. To prepare the data (for all questions), I normalized all numerical data (BMI, General Health, Mental Health, Physical Health, Age Bracket, and Income Bracket) through by standardization (Z-score normalization), change Biological Sex to be represented by 0s and 1s instead of 1s and 2s (just to match the other categorical variables), and one-hot encoded the remaining categorical variables (Education Bracket and Zodiac). I then split the data into train and test sets with a 70/30 split. I then fit a logistic regression model with all the predictor variables as the inputs and "Diabetes" as the output. After running a logistic regression, I ran the predictions on the test set to calculate my accuracy. Then, I plotted the graph to show the ROC curve and the AUC. Finally, I calculated the AUC for when each individual predictor variable is dropped by one by one (leaving the remaining predictor variables as inputs).
2. I prepared the data the way I did (for all questions) because for some of the models used in this homework (ex. logistic regression), it's important to work with normalized data. Regarding specifically what I did for the logistic regression model, I did this to see how well all the predictor variables together did at accurately predicting diabetes using logistic regression and to find which predictor variable affects model performance the most (by finding which one when not included decreases AUC the most).
3. I found that when using all predictor variables, the logistic regression model has an accuracy of 72.857% and an AUC of 0.82424. The predictor variable that affected AUC the most was General Health because when that predictor variable is not included, the AUC is 0.80854 meaning the AUC decreased by 0.0157.
4. The AUC for the model using all the predictor variables is quite high at 0.82424 (so it is a well-performing model). General Health is the best predictor of diabetes because when this predictor variable is not included as an input, the AUC is affected the most in comparison to the dropping of other predictor variables.

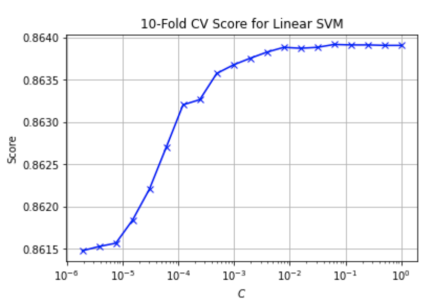


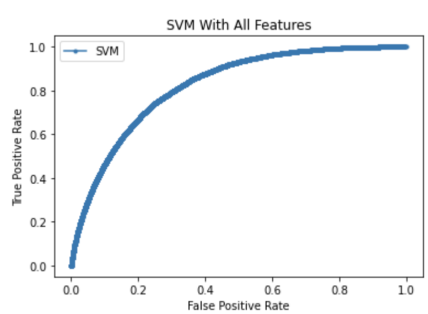

```
Logistic Regression Accuracy = 72.857%
Logistic Regression AUC = 0.82424
```

```
Logistic Regression AUC = 0.81700 HighBP
Logistic Regression AUC = 0.81904 HighChol
Logistic Regression AUC = 0.80873 BMI
Logistic Regression AUC = 0.82426 Smoker
Logistic Regression AUC = 0.82404 Stroke
Logistic Regression AUC = 0.82367 Myocardial
Logistic Regression AUC = 0.82425 PhysActivity
Logistic Regression AUC = 0.82422 Fruit
Logistic Regression AUC = 0.82423 Vegetables
Logistic Regression AUC = 0.82288 HeavyDrinker
Logistic Regression AUC = 0.82416 HasHealthcare
Logistic Regression AUC = 0.82425 NotAbleToAffordDoctor
Logistic Regression AUC = 0.80854 GeneralHealth
Logistic Regression AUC = 0.82417 MentalHealth
Logistic Regression AUC = 0.82404 PhysicalHealth
Logistic Regression AUC = 0.82412 HardToClimbStairs
Logistic Regression AUC = 0.82329 BiologicalSex
Logistic Regression AUC = 0.81567 AgeBracket
Logistic Regression AUC = 0.82347 IncomeBracket
Logistic Regression AUC = 0.82424 Kindergarten
Logistic Regression AUC = 0.82424 Elementary
Logistic Regression AUC = 0.82424 HighSchool
Logistic Regression AUC = 0.82424 GED
Logistic Regression AUC = 0.82424 College
Logistic Regression AUC = 0.82424 Graduate
Logistic Regression AUC = 0.82424 Aries
Logistic Regression AUC = 0.82424 Taurus
Logistic Regression AUC = 0.82424 Gemini
Logistic Regression AUC = 0.82424 Cancer
Logistic Regression AUC = 0.82424 Leo
Logistic Regression AUC = 0.82424 Virgo
Logistic Regression AUC = 0.82424 Libra
Logistic Regression AUC = 0.82424 Scorpio
Logistic Regression AUC = 0.82424 Sagittarius
Logistic Regression AUC = 0.82424 Capricorn
Logistic Regression AUC = 0.82424 Aquarius
Logistic Regression AUC = 0.82424 Pisces
```

Question 2

1. I first split the data into train and test sets with a 70/30 split. I then found the optimal C value that will be used as a parameter in our SVM model through using a for-loop that looped over different C values ranging from 1 to 1e-6 and identified the C-value that resulted in the highest validation score. I then fit a SVM model with all the predictor variables as the inputs and “Diabetes” as the output. After running a logistic regression, I ran the predictions on the test set to calculate my accuracy. Then, I plotted the graph to show the ROC curve and the AUC. Finally, I calculated the AUC for when each individual predictor variable is dropped by one by one (leaving the remaining predictor variables as inputs).
2. I first found the optimal C value so that our SVM model performs at its best with tuned hyperparameters. Regarding specifically what I did for the SVM model, I did this to see how well all the predictor variables together did at accurately predicting diabetes using a SVM and to find which predictor variable affects model performance the most (by finding which one when not included decreases AUC the most).
3. First of all, I found the optimal C value to be 0.0625 (so I used it as a parameter for the model). I found that when using all predictor variables, the SVM has an accuracy of 86.323% and an AUC of 0.82348. The predictor variable that affected AUC the most was General Health because when that predictor variable is not included, the AUC is 0.80764 meaning the AUC decreased by 0.01584.
4. The AUC for the model using all the predictor variables is quite high at 0.82348 (so it is a well-performing model). General Health is the best predictor of diabetes because when this predictor variable is not included as an input, the AUC is affected the most in comparison to the dropping of other predictor variables. This model is better than the logistic regression model because it has a comparable AUC but a much higher accuracy.



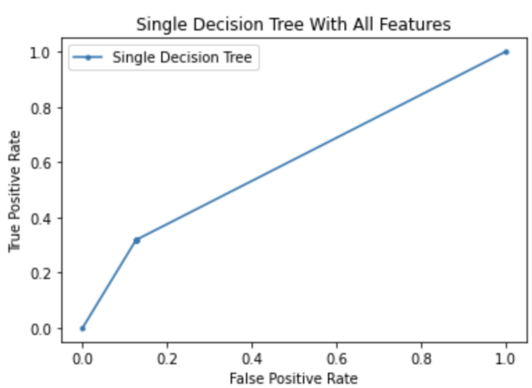



```
SVM Accuracy = 86.323%
SVM AUC = 0.82348
```

```
SVM AUC = 0.81622 HighBP
SVM AUC = 0.81831 HighChol
SVM AUC = 0.80800 BMI
SVM AUC = 0.82357 Smoker
SVM AUC = 0.82321 Stroke
SVM AUC = 0.82306 Myocardial
SVM AUC = 0.82352 PhysActivity
SVM AUC = 0.82345 Fruit
SVM AUC = 0.82347 Vegetables
SVM AUC = 0.82214 HeavyDrinker
SVM AUC = 0.82339 HasHealthcare
SVM AUC = 0.82347 NotAbleToAffordDoctor
SVM AUC = 0.80764 GeneralHealth
SVM AUC = 0.82342 MentalHealth
SVM AUC = 0.82336 PhysicalHealth
SVM AUC = 0.82352 HardToClimbStairs
SVM AUC = 0.82246 BiologicalSex
SVM AUC = 0.81531 AgeBracket
SVM AUC = 0.82273 IncomeBracket
SVM AUC = 0.82347 Kindergarten
SVM AUC = 0.82348 Elementary
SVM AUC = 0.82348 HighSchool
SVM AUC = 0.82348 GED
SVM AUC = 0.82348 College
SVM AUC = 0.82348 Graduate
SVM AUC = 0.82348 Aries
SVM AUC = 0.82348 Taurus
SVM AUC = 0.82348 Gemini
SVM AUC = 0.82348 Cancer
SVM AUC = 0.82348 Leo
SVM AUC = 0.82348 Virgo
SVM AUC = 0.82348 Libra
SVM AUC = 0.82348 Scorpio
SVM AUC = 0.82348 Sagittarius
SVM AUC = 0.82348 Capricorn
SVM AUC = 0.82347 Aquarius
SVM AUC = 0.82348 Pisces
```

Question 3

1. I first split the data into train and test sets with a 70/30 split. I then fit a single decision tree model with gini as a criterion with all the predictor variables as the inputs and "Diabetes" as the output. After running a single decision tree, I ran the predictions on the test set to calculate my accuracy. Then, I plotted the graph to show the ROC curve and the AUC. Finally, I calculated the AUC for when each individual predictor variable is dropped by one by one (leaving the remaining predictor variables as inputs).
2. I did this to see how well all the predictor variables together did at accurately predicting diabetes using a single decision tree and to find which predictor variable affects model performance the most (by finding which one when not included decreases AUC the most). I specifically used gini as a criterion since this would allow me to use gini index to decide what the best split is from a root node and subsequent splits later on.
3. I found that when using all predictor variables, the single decision tree has an accuracy of 79.491% and an AUC of 0.59595. The predictor variable that affected AUC the most was BMI because when that predictor variable is not included, the AUC is 0.58176 meaning the AUC decreased by 0.01419.
4. The AUC for the model using all the predictor variables is not high at 0.59595 (so the model is not that good at discriminating). BMI is the best predictor of diabetes because when this predictor variable is not included as an input, the AUC is affected the most in comparison to the dropping of other predictor variables. Since the AUC for this model is lower than the AUC for the SVM and logistic regression models, we can conclude that a single decision tree is a poor model for predicting diabetes in our dataset.

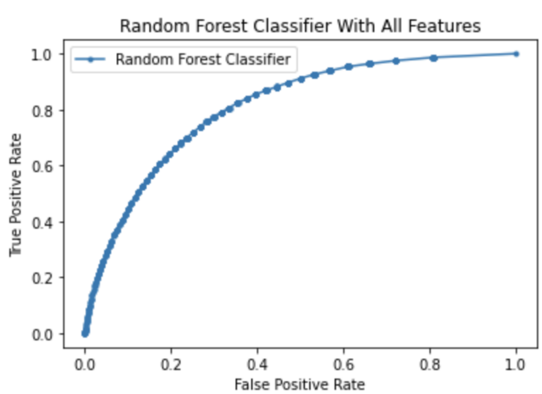



```
Single Decision Tree Accuracy = 79.491%
Single Decision Tree AUC = 0.59595
```

```
Single Decision Tree AUC = 0.59857 HighBP
Single Decision Tree AUC = 0.59144 HighChol
Single Decision Tree AUC = 0.58176 BMI
Single Decision Tree AUC = 0.59986 Smoker
Single Decision Tree AUC = 0.59784 Stroke
Single Decision Tree AUC = 0.59506 Myocardial
Single Decision Tree AUC = 0.59732 PhysActivity
Single Decision Tree AUC = 0.59754 Fruit
Single Decision Tree AUC = 0.59742 Vegetables
Single Decision Tree AUC = 0.59910 HeavyDrinker
Single Decision Tree AUC = 0.59816 HasHealthcare
Single Decision Tree AUC = 0.59724 NotAbleToAffordDoctor
Single Decision Tree AUC = 0.59036 GeneralHealth
Single Decision Tree AUC = 0.59749 MentalHealth
Single Decision Tree AUC = 0.59722 PhysicalHealth
Single Decision Tree AUC = 0.59405 HardToClimbStairs
Single Decision Tree AUC = 0.59442 BiologicalSex
Single Decision Tree AUC = 0.59263 AgeBracket
Single Decision Tree AUC = 0.59394 IncomeBracket
Single Decision Tree AUC = 0.59591 Kindergarten
Single Decision Tree AUC = 0.59816 Elementary
Single Decision Tree AUC = 0.59685 HighSchool
Single Decision Tree AUC = 0.59762 GED
Single Decision Tree AUC = 0.59755 College
Single Decision Tree AUC = 0.59859 Graduate
Single Decision Tree AUC = 0.59573 Aries
Single Decision Tree AUC = 0.59545 Taurus
Single Decision Tree AUC = 0.59675 Gemini
Single Decision Tree AUC = 0.59641 Cancer
Single Decision Tree AUC = 0.59685 Leo
Single Decision Tree AUC = 0.59772 Virgo
Single Decision Tree AUC = 0.59629 Libra
Single Decision Tree AUC = 0.59541 Scorpio
Single Decision Tree AUC = 0.59813 Sagittarius
Single Decision Tree AUC = 0.59667 Capricorn
Single Decision Tree AUC = 0.59760 Aquarius
Single Decision Tree AUC = 0.59782 Pisces
```

Question 4

1. I first split the data into train and test sets with a 70/30 split. I then fit a Random Forest Classifier model with gini as a criterion with all the predictor variables as the inputs and "Diabetes" as the output. After running a Random Forest Classifier, I ran the predictions on the test set to calculate my accuracy. Then, I plotted the graph to show the ROC curve and the AUC. Finally, I calculated the AUC for when each individual predictor variable is dropped by one by one (leaving the remaining predictor variables as inputs).
2. I did this to see how well all the predictor variables together did at accurately predicting diabetes using a Random Forest Classifier and to find which predictor variable affects model performance the most (by finding which one when not included decreases AUC the most). I specifically used gini as a criterion since this would allow me to use gini index to decide what the best split is from a root node and subsequent splits later on.
3. I found that when using all predictor variables, the Random Forest Classifier has an accuracy of 86.341% and an AUC of 0.81015. The predictor variable that affected AUC the most was BMI because when that predictor variable is not included, the AUC is 0.77869 meaning the AUC decreased by 0.03146.
4. The AUC for the model using all the predictor variables is quite high at 0.81015 (so it is a well-performing model). BMI is the best predictor of diabetes because when this predictor variable is not included as an input, the AUC is affected the most in comparison to the dropping of other predictor variables. The AUC for this model is comparable to that of the logistic regression model except the Random Forest Classifier model also has a notably higher accuracy (86.341% vs. 72.857%).

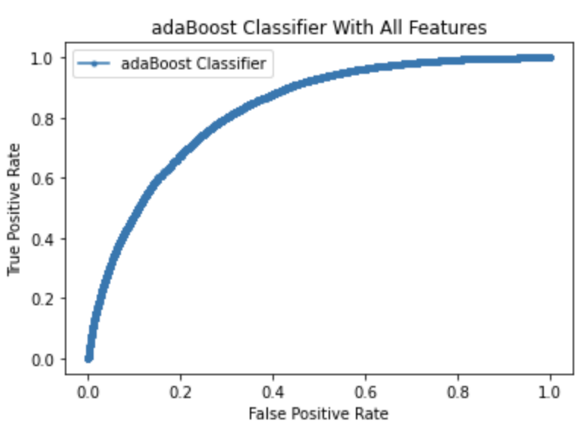



```
Random Forest Classifier Accuracy = 86.341%
Random Forest Classifier AUC = 0.81015
```

```
Random Forest Classifier AUC = 0.80023 HighBP
Random Forest Classifier AUC = 0.80192 HighChol
Random Forest Classifier AUC = 0.77869 BMI
Random Forest Classifier AUC = 0.80818 Smoker
Random Forest Classifier AUC = 0.80839 Stroke
Random Forest Classifier AUC = 0.80814 Myocardial
Random Forest Classifier AUC = 0.80860 PhysActivity
Random Forest Classifier AUC = 0.80761 Fruit
Random Forest Classifier AUC = 0.80835 Vegetables
Random Forest Classifier AUC = 0.80807 HeavyDrinker
Random Forest Classifier AUC = 0.80979 HasHealthcare
Random Forest Classifier AUC = 0.80978 NotAbleToAffordDoctor
Random Forest Classifier AUC = 0.78964 GeneralHealth
Random Forest Classifier AUC = 0.80802 MentalHealth
Random Forest Classifier AUC = 0.80688 PhysicalHealth
Random Forest Classifier AUC = 0.80905 HardToClimbStairs
Random Forest Classifier AUC = 0.80743 BiologicalSex
Random Forest Classifier AUC = 0.78941 AgeBracket
Random Forest Classifier AUC = 0.80365 IncomeBracket
Random Forest Classifier AUC = 0.80974 Kindergarten
Random Forest Classifier AUC = 0.80963 Elementary
Random Forest Classifier AUC = 0.81070 HighSchool
Random Forest Classifier AUC = 0.80978 GED
Random Forest Classifier AUC = 0.81006 College
Random Forest Classifier AUC = 0.81020 Graduate
Random Forest Classifier AUC = 0.80929 Aries
Random Forest Classifier AUC = 0.80882 Taurus
Random Forest Classifier AUC = 0.80755 Gemini
Random Forest Classifier AUC = 0.81029 Cancer
Random Forest Classifier AUC = 0.80942 Leo
Random Forest Classifier AUC = 0.80845 Virgo
Random Forest Classifier AUC = 0.80871 Libra
Random Forest Classifier AUC = 0.80862 Scorpio
Random Forest Classifier AUC = 0.80920 Sagittarius
Random Forest Classifier AUC = 0.80935 Capricorn
Random Forest Classifier AUC = 0.80923 Aquarius
Random Forest Classifier AUC = 0.80847 Pisces
```

Question 5

1. I first split the data into train and test sets with a 70/30 split. I then fit an adaBoost Classifier model with all the predictor variables as the inputs and “Diabetes” as the output. After running a adaBoost Classifier, I ran the predictions on the test set to calculate my accuracy. Then, I plotted the graph to show the ROC curve and the AUC. Finally, I calculated the AUC for when each individual predictor variable is dropped by one by one (leaving the remaining predictor variables as inputs).
2. I did this to see how well all the predictor variables together did at accurately predicting diabetes using an adaBoost Classifier and to find which predictor variable affects model performance the most (by finding which one when not included decreases AUC the most).
3. I found that when using all predictor variables, the adaBoost Classifier has an accuracy of 86.529% and an AUC of 0.82791. The predictor variable that affected AUC the most was BMI because when that predictor variable is not included, the AUC is 0.81113 meaning the AUC decreased by 0.01673.
4. The AUC for the model using all the predictor variables is quite high at 0.82803 (so it is a well-performing model). BMI is the best predictor of diabetes because when this predictor variable is not included as an input, the AUC is affected the most in comparison to the dropping of other predictor variables.




```
adaBoost Classifier Accuracy = 86.529%
adaBoost Classifier AUC = 0.82791
```

```
adaBoost Classifier AUC = 0.81991 HighBP
adaBoost Classifier AUC = 0.82355 HighChol
adaBoost Classifier AUC = 0.81118 BMI
adaBoost Classifier AUC = 0.82791 Smoker
adaBoost Classifier AUC = 0.82794 Stroke
adaBoost Classifier AUC = 0.82713 Myocardial
adaBoost Classifier AUC = 0.82791 PhysActivity
adaBoost Classifier AUC = 0.82791 Fruit
adaBoost Classifier AUC = 0.82791 Vegetables
adaBoost Classifier AUC = 0.82689 HeavyDrinker
adaBoost Classifier AUC = 0.82791 HasHealthcare
adaBoost Classifier AUC = 0.82791 NotAbleToAffordDoctor
adaBoost Classifier AUC = 0.81213 GeneralHealth
adaBoost Classifier AUC = 0.82803 MentalHealth
adaBoost Classifier AUC = 0.82791 PhysicalHealth
adaBoost Classifier AUC = 0.82805 HardToClimbStairs
adaBoost Classifier AUC = 0.82719 BiologicalSex
adaBoost Classifier AUC = 0.81819 AgeBracket
adaBoost Classifier AUC = 0.82698 IncomeBracket
adaBoost Classifier AUC = 0.82791 Kindergarten
adaBoost Classifier AUC = 0.82791 Elementary
adaBoost Classifier AUC = 0.82791 HighSchool
adaBoost Classifier AUC = 0.82791 GED
adaBoost Classifier AUC = 0.82791 College
adaBoost Classifier AUC = 0.82771 Graduate
adaBoost Classifier AUC = 0.82791 Aries
adaBoost Classifier AUC = 0.82791 Taurus
adaBoost Classifier AUC = 0.82791 Gemini
adaBoost Classifier AUC = 0.82791 Cancer
adaBoost Classifier AUC = 0.82791 Leo
adaBoost Classifier AUC = 0.82791 Virgo
adaBoost Classifier AUC = 0.82791 Libra
adaBoost Classifier AUC = 0.82791 Scorpio
adaBoost Classifier AUC = 0.82791 Sagittarius
adaBoost Classifier AUC = 0.82791 Capricorn
adaBoost Classifier AUC = 0.82791 Aquarius
adaBoost Classifier AUC = 0.82791 Pisces
```

Extra Credit 1

The adaBoost Classifier model is the best at predicting diabetes in this dataset because it has the highest accuracy and the the highest AUC among the five models.

Extra Credit 2

Something interesting about this dataset is that mental health has a medium positive correlation with physical health (they have a correlation coefficient of 0.353618867841803). This is likely because maintaining a healthy and fit body often promotes good mental health and well-being because if you physically feel good, you are more likely to have an elevated mood and be less stressed.

```
df['PhysicalHealth'].corr(df['MentalHealth'])
```

```
0.353618867841803
```