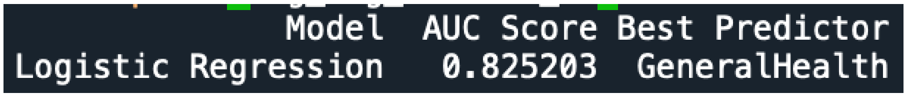
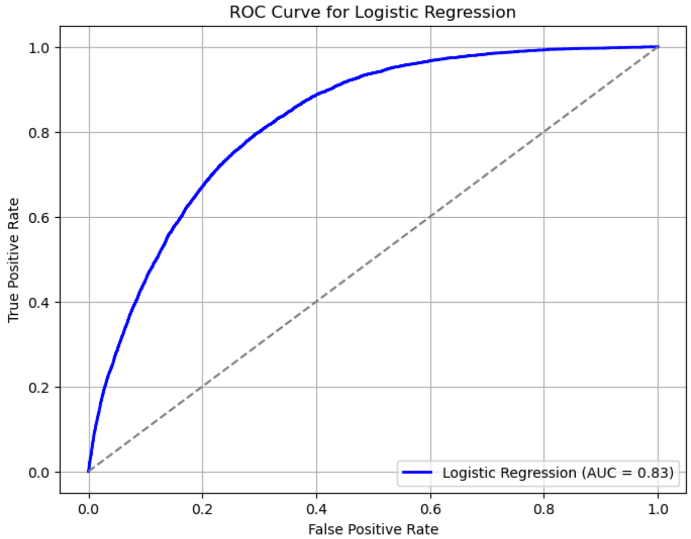
1. Build a logistic regression model. Doing so: What is the best predictor of diabetes and what is the AUC of this model?

I trained a Logistic Regression model to predict diabetes. I used an 80-20 train-test split for the data and the features were standardized. I calculated the AUC score along with the best predictor based on the absolute magnitude of the coefficients. An ROC curve was plotted to visualize the performance. Logistic regression was used because it’s a simple, interpretable model for binary classification. AUC was used as the evaluation metric because it measures the model’s ability to distinguish between classes. The AUC Score was 0.83, the Best Predictor was General Health, and the ROC curve showed that the model performs better than just randomly guessing the outcome. This means the model effectively discriminates between diabetic and non-diabetic individuals, with General Health being the strongest predictor of this outcome.



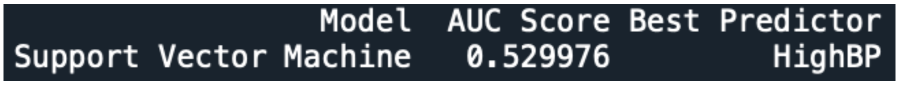
```
              Model  AUC Score Best Predictor
Logistic Regression   0.825203  GeneralHealth
```



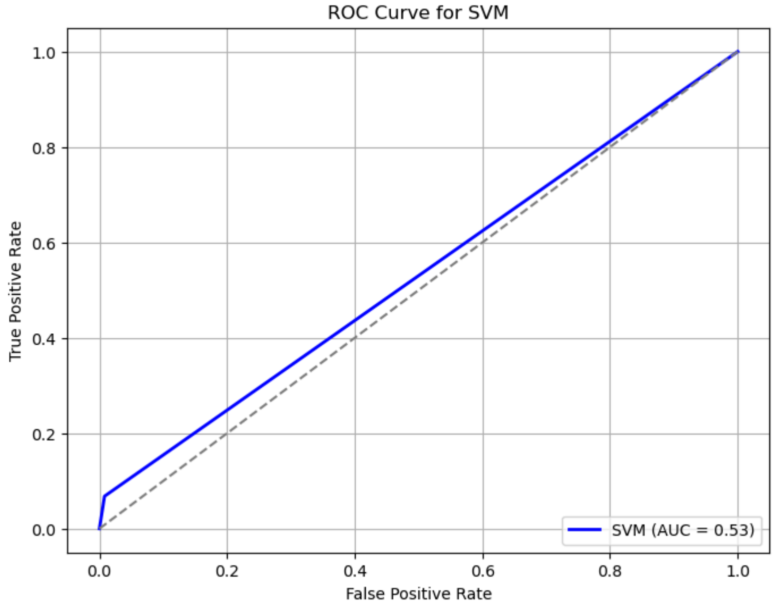
ROC Curve for Logistic Regression

2. Build a SVM. Doing so: What is the best predictor of diabetes and what is the AUC of this model?

I trained an SVM model to predict diabetes. I used an 80-20 train-test split for the data. Multiple values of the slack variable, C, were tested using cross-validation to find the best regularization parameter. Through this I ended up using C = 10 and calculated the AUC score along with the best predictor using permutation importance. An ROC curve was plotted to visualize the performance. SVM was appropriate to use when data is non-linearly separable, such as in this situation. Testing different values of C helps balance margin size vs. misclassification. AUC was used as the evaluation metric because it measures the model’s ability to distinguish between classes. The AUC Score was 0.53 and the Best Predictor is High Blood Pressure. This means the model does not effectively discriminate between diabetic and non-diabetic individuals, given that the AUC is close to 0.50. The model performs slightly better than randomly guessing the outcome.

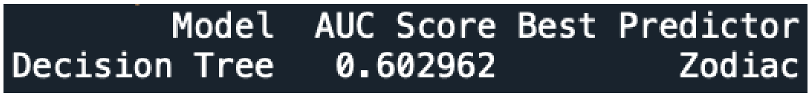


```
                  Model  AUC Score Best Predictor
Support Vector Machine   0.529976         HighBP
```

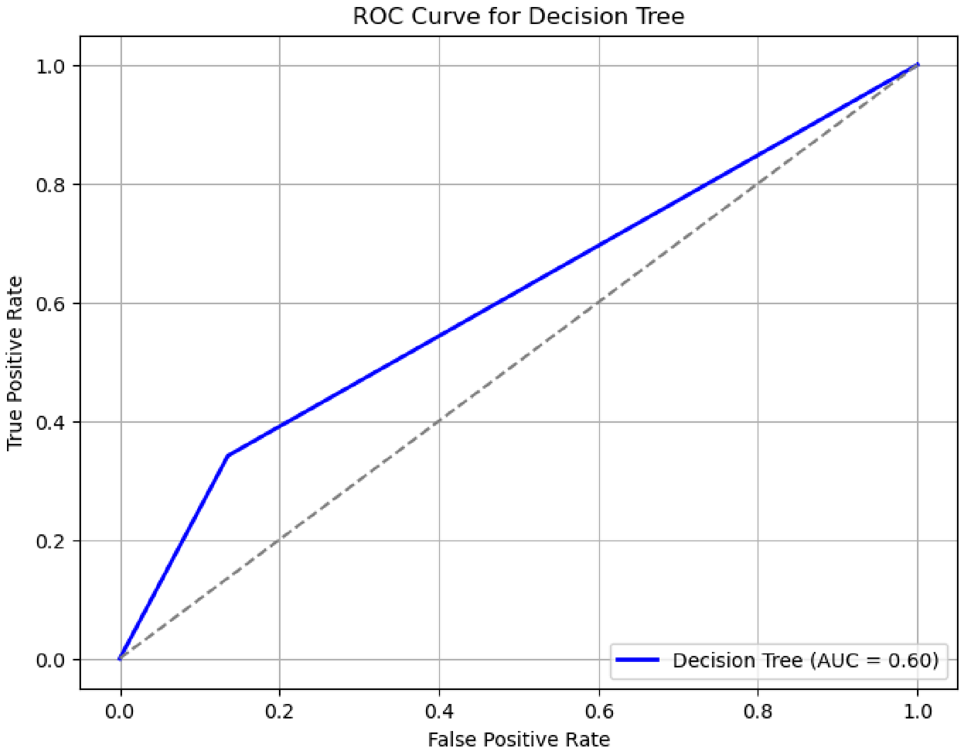


3. Use a single, individual decision tree. Doing so: What is the best predictor of diabetes and what is the AUC of this model?

I trained a Decision Tree model to predict diabetes. The model was built using Gini Impurity. I computed the AUC Score to assess performance and I found the most important feature by using feature importance scores. I also plotted the ROC curve to visualize the model's performance. We used Decision Trees because they are easily interpretable models, which, when implemented with Gini Impurity, helps determine the best splits for classification. AUC was used as the evaluation metric because it measures the model's ability to distinguish between classes. The AUC Score was 0.60 and the Best Predictor was Zodiac. While the AUC Score of the model performs better than randomly guessing outcomes, the decision tree in this case likely overfit our data as evidenced by the Best Predictor being Zodiac, which is unlikely to have a meaningful relationship with diabetes. The model did not pick up on truly predictive features.

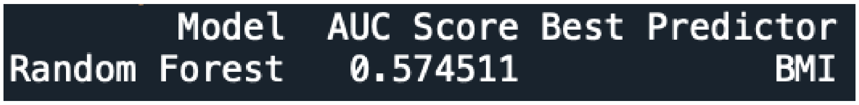


```
        Model  AUC Score Best Predictor
Decision Tree   0.602962         Zodiac
```

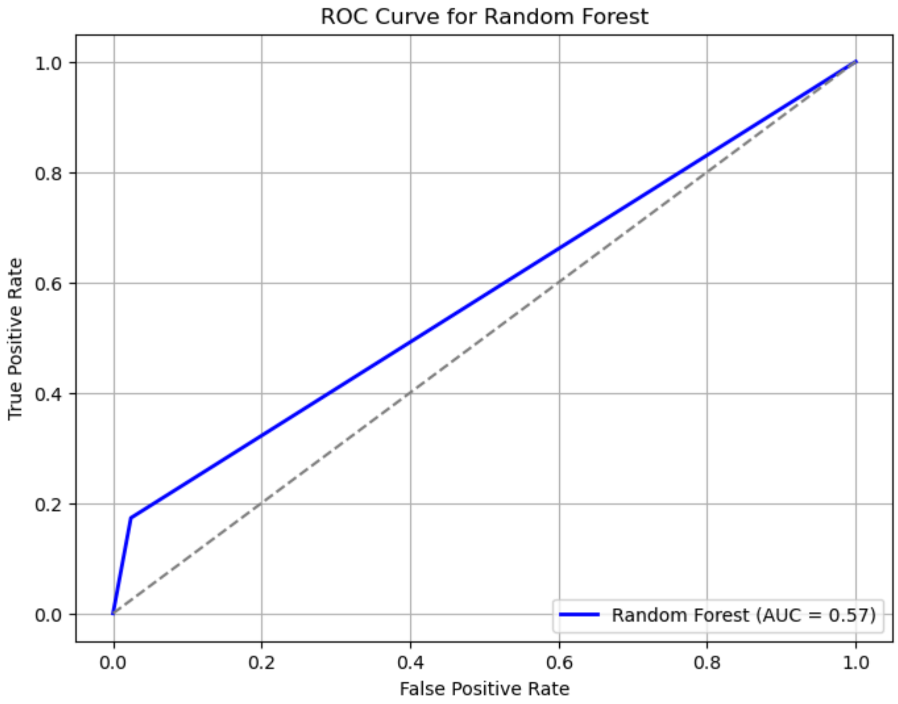


4. Build a random forest model. Doing so: What is the best predictor of diabetes and what is the AUC of this model?

I trained a Random Forest Classifier using 100 trees, with each tree using a random subset (50%) of both samples and features to reduce overfitting. I evaluated the model using the AUC Score and extracted the best feature by using feature importance scores. I plotted the AUC Curve to visualize the model's performance. I used this because Random Forests improve upon Decision Trees by reducing overfitting through ensemble learning, averaging multiple tree predictions. Using random features and sample selection enhances generalization. AUC was used as the evaluation metric because it measures the model's ability to distinguish between classes. The AUC Score for this model is 0.58 and the Best Predictor is BMI. This means the AUC Score of the model performs better than randomly guessing outcomes.

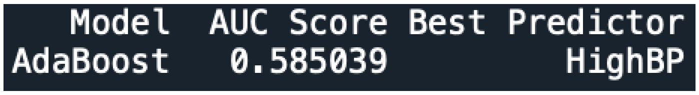


```
         Model  AUC Score Best Predictor
Random Forest   0.574511            BMI
```

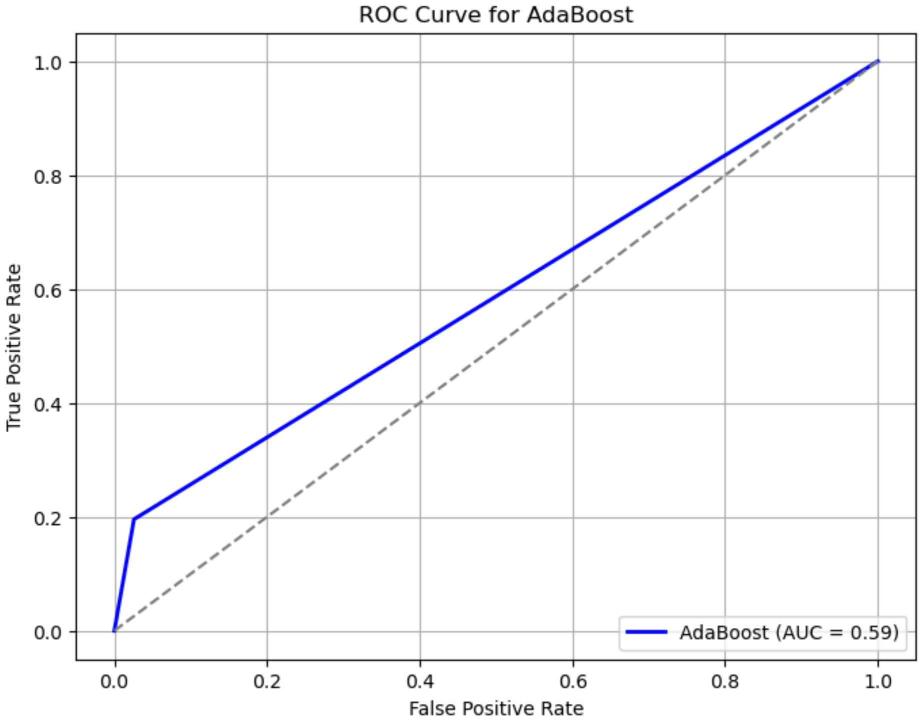


5. Build a model using adaBoost. Doing so: What is the best predictor of diabetes and what is the AUC of this model?

I used an adaBoost classifier to train 2000 weak decision tree learners with max depth = 1. The model sequentially adjusted weights to focus on misclassified instances. I computed the AUC Score to evaluate the model and used feature importance scores to extract the best predictor. I plotted the AUC Curve to visualize the model’s performance. I used adaBoost because it effectively improves weak learners by boosting performance iteratively. Using shallow tree depths (depth = 1) prevents overfitting while still capturing important relationships. AUC was used as the evaluation metric because it measures the model’s ability to distinguish between classes. The AUC Score is 0.59 and the Best Predictor is High Blood Pressure. AdaBoost actually had a weaker performance than a singular decision tree, and the adaBoost model only performed slightly better than randomly guessing outcomes. Since the outcome also depended on hyperparameters, further tuning may yield even better results.

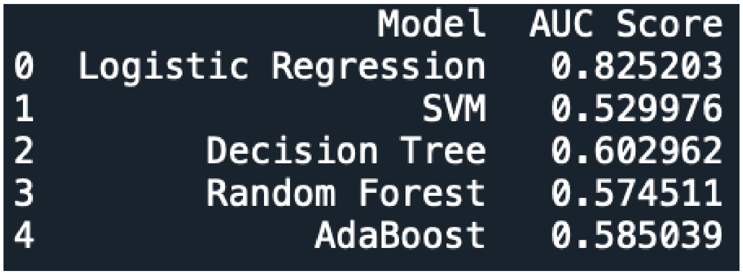


```
   Model  AUC Score Best Predictor
AdaBoost   0.585039         HighBP
```



Extra credit:
a) Which of these 5 models is the best to predict diabetes in this dataset?

Logistic Regression is the best model to predict diabetes in this dataset, as evidenced by the highest AUC Score of 0.83. This means that out of all the models, Logistic Regression was able to best discriminate between diabetic and non-diabetic individuals, with the best ratio of true positives (people with diabetes predicted to have diabetes) to false positives (people without diabetes predicted to have diabetes).



```
                 Model  AUC Score
0  Logistic Regression   0.825203
1                  SVM   0.529976
2        Decision Tree   0.602962
3        Random Forest   0.574511
4             AdaBoost   0.585039
```