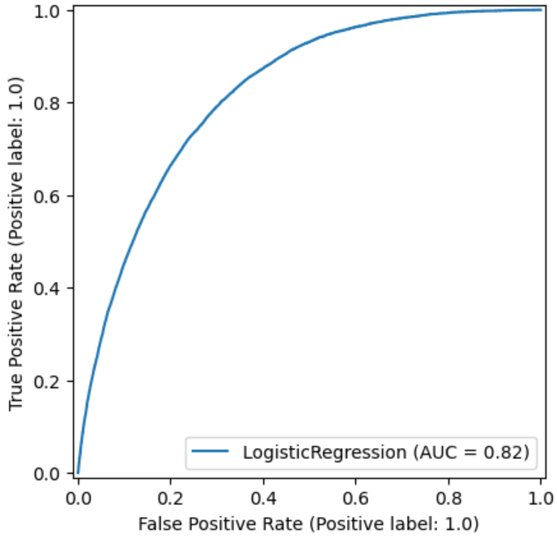
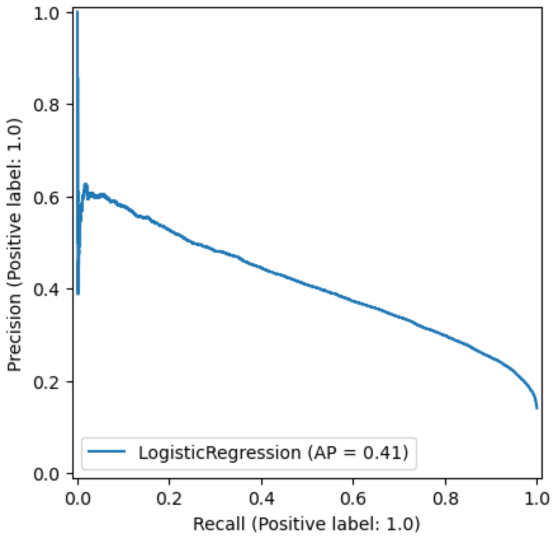
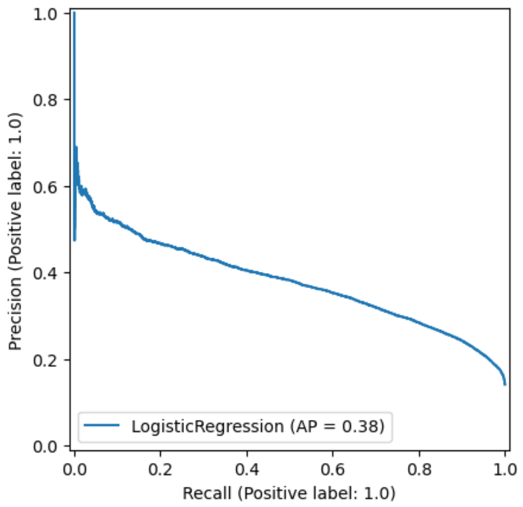
1. Build a logistic regression model. Doing so: What is the best predictor of diabetes and what is the AUC of this model?

First, I removed all rows containing missing data from the dataset. As stated in the spec sheet, there is not too much missing data, and the size of the dataset remains very large, so it is okay to remove rows. I then one-hot encoded the zodiac data, because the values make sense to interpret categorically and removed the original column. I then split the data 60/40 into training/test sets and fit a logistic regression model. This gave me the following ROC curve, PR curve, an AUROC of 0.82 and an AUPRC of 0.41. The AUPRC is more telling than the AUROC in this case because the classes are unbalanced, with only around 13.93% of the entries having diabetes. The large difference in AUROC and AUPRC confirms that the classes are unbalanced. This AUPRC is still pretty good, given that the baseline is 0.1393.



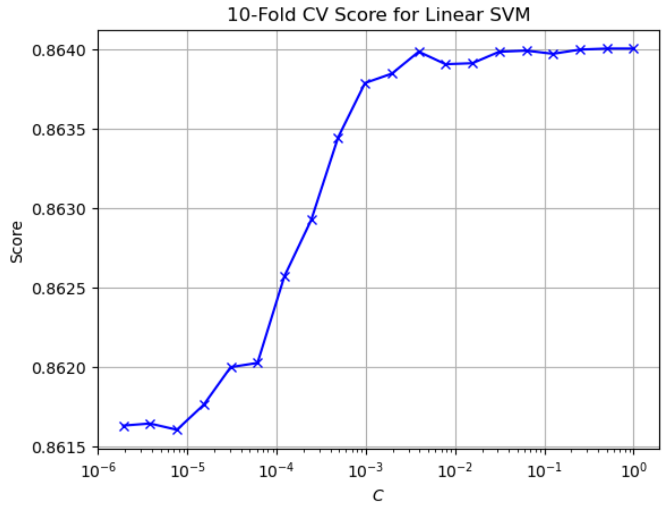


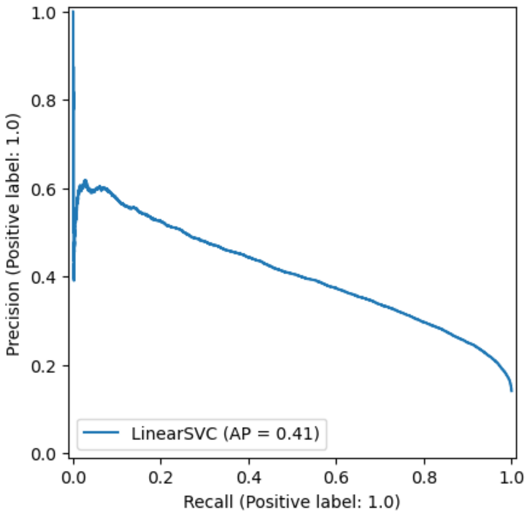
To determine the best predictor of diabetes, I fit a model removing 1 feature at a time in order to see which feature being removed causes the AUPRC to drop the most. I found that this was Body Mass Index, whose model with its removal gave an AUPRC of 0.38, meaning that it is the best predictor of diabetes from this dataset.



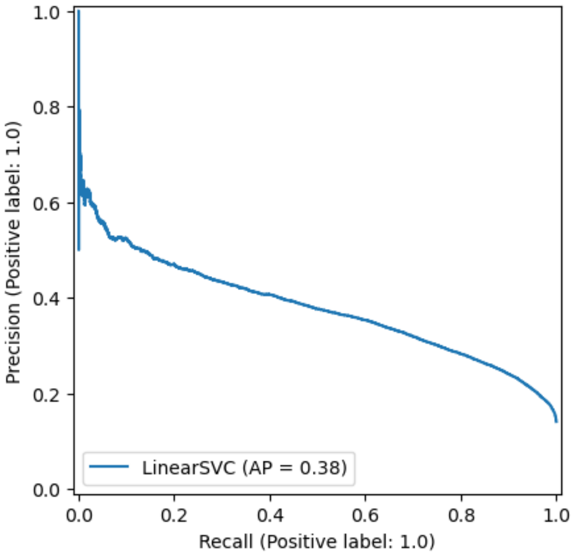
2. Build a SVM. Doing so: What is the best predictor of diabetes and what is the AUC of this model?

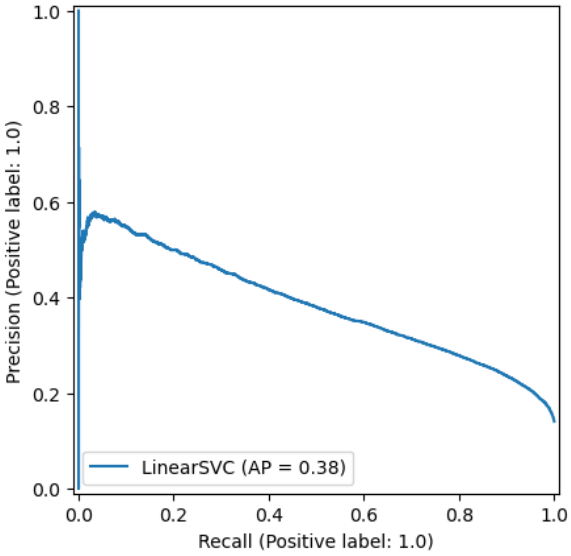
I split the data into a 60/40 training/test split. Then, I fit linear SVC models with different levels of regularization strength C in order to find the best value creating the model with the best cross-validation score, which created the graph below. From this graph, I concluded that the optimal C was 1, because it had the highest cross-validation score. I then used this value to fit a linear SVC model, giving the PR curve below, with an AUPRC of 0.41. This is the same value as before, using the logistic regression model.





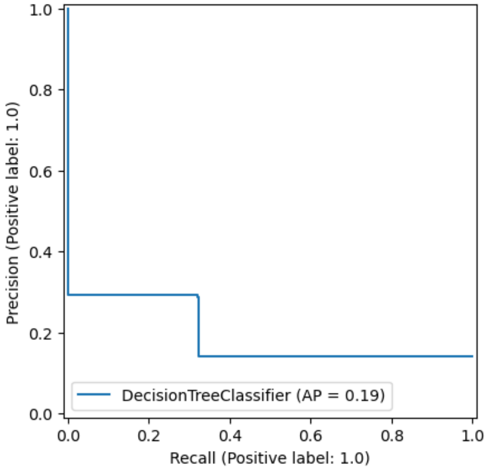
I then used the same process as the previous question to determine the best predictor of diabetes, and found BMI and self-assessment of health status to have the largest effect on the AUPRC, both dropping it to 0.38. This is consistent with what was found in the previous question, with the addition of self-assessment of health status. The graphs below represent the removal of BMI and self-assessment of health status respectively.



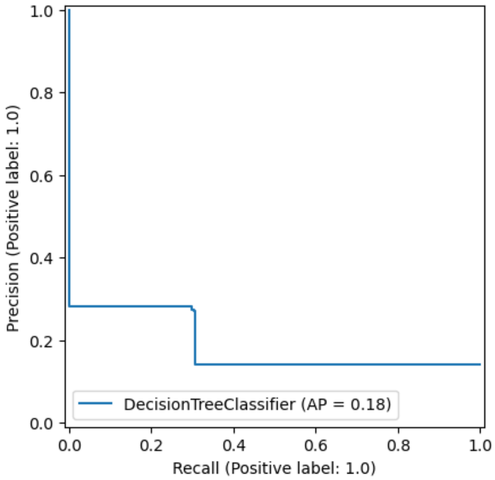


3. Use a single, individual decision tree. Doing so: What is the best predictor of diabetes and what is the AUC of this model?

I split the data into a 60/40 training/test split. I then fit a decision tree to this, splitting using Gini criterion, which gave the following PR curve, with an AUPRC of 0.19. This AUPRC is low, especially compared to the previous models but still higher than the proportion of the data with diabetes. This makes sense, as individual decision trees are weak learners, and the model did not perform well when faced with new data. I also fit a decision tree using entropy as the splitting criterion, which gave very similar results.

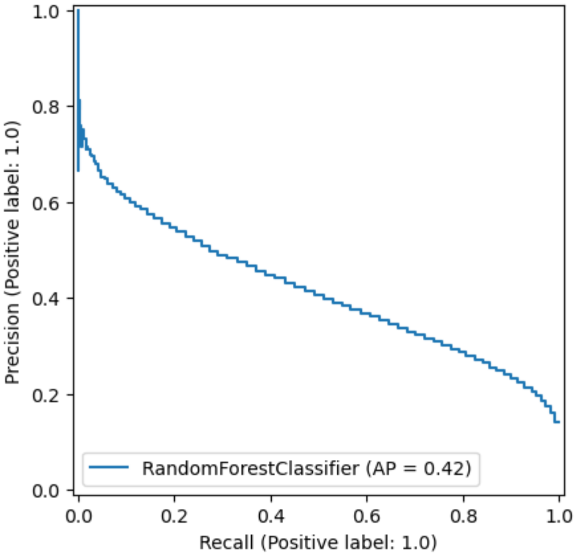


I then repeated the process from the previous problems, splitting using Gini. I found that self-assessment of health status was the best predictor of diabetes, dropping the AUPRC to 0.18. I also found that removing zodiac signs as a predictor increased the AUPRC to 0.20, which was strange. The graph below is the PR curve created from the model without self-assessment of health status.

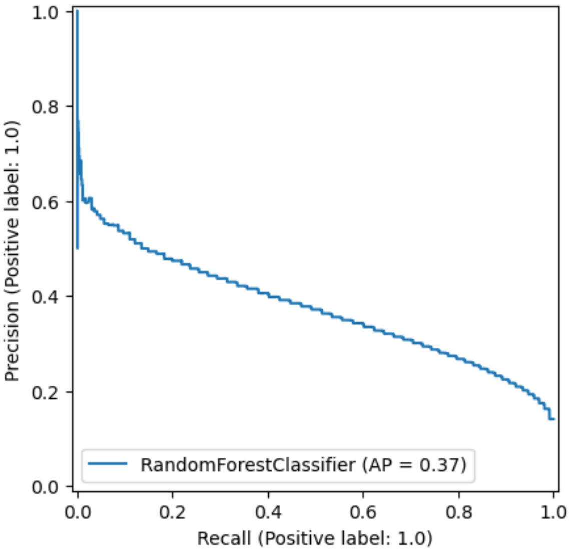


4. Build a random forest model. Doing so: What is the best predictor of diabetes and what is the AUC of this model?

I built a random forest model using a training/test split of 60/40 and Gini as splitting criterion. This gave the following PR curve, with an AUC of 0.42. This is notably slightly better than the logistic regression and SVM models, and much better than the single decision tree. This makes sense, as merging multiple decision trees allowed the model to handle new data much better.

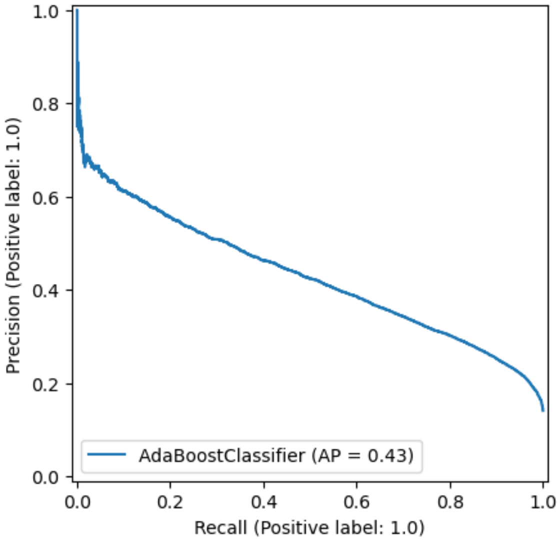


I then repeated the process from the previous problems, splitting using Gini. I found that BMI was the best predictor of diabetes, dropping the AUPRC to 0.37 The graph below is the PR curve created from the model without BMI.

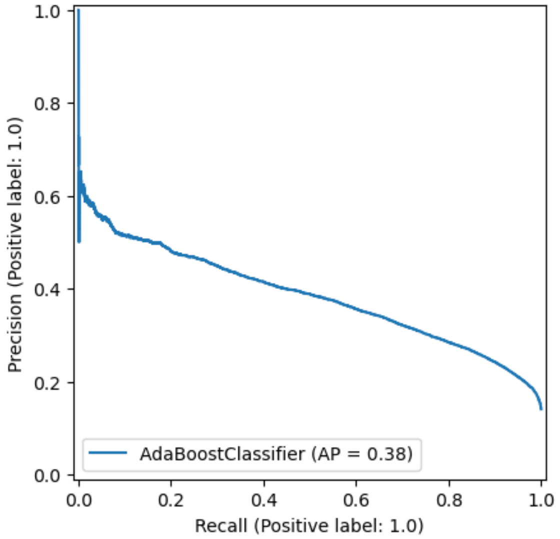


5. Build a model using adaBoost. Doing so: What is the best predictor of diabetes and what is the AUC of this model

I built an adaBoost model using a training/test split of 60/40. This gave the following PR curve, with an AUC of 0.43. This means that this model performed better than the others, notably missing the sharp downward spike at low recall. It makes sense that this model did much better than the individual decision tree.

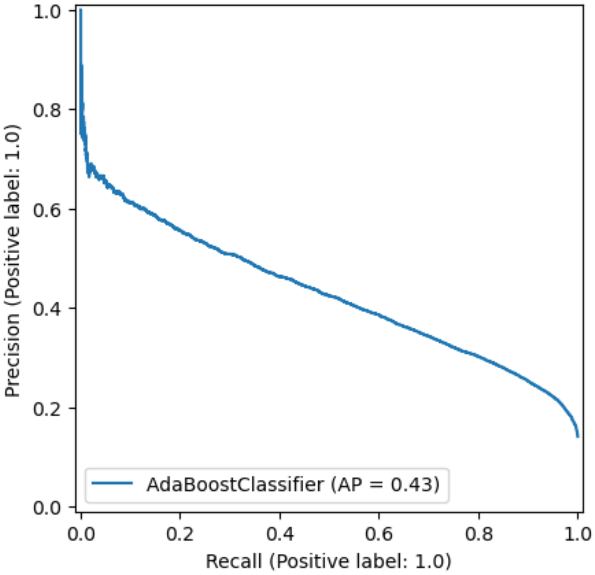


I then repeated the process from the previous problems. I found that BMI was the best predictor of diabetes, dropping the AUPRC to 0.38. The graph below is the PR curve created from the model without BMI as a predictor.



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

The best model for predicting diabetes from the models used above is the one made using adaBoost. I think this because it gave the highest AUPRC, 0.43. The graph below is the PR curve for the model.


Precision (Positive label: 1.0)
AdaBoostClassifier (AP = 0.43)
Recall (Positive label: 1.0)