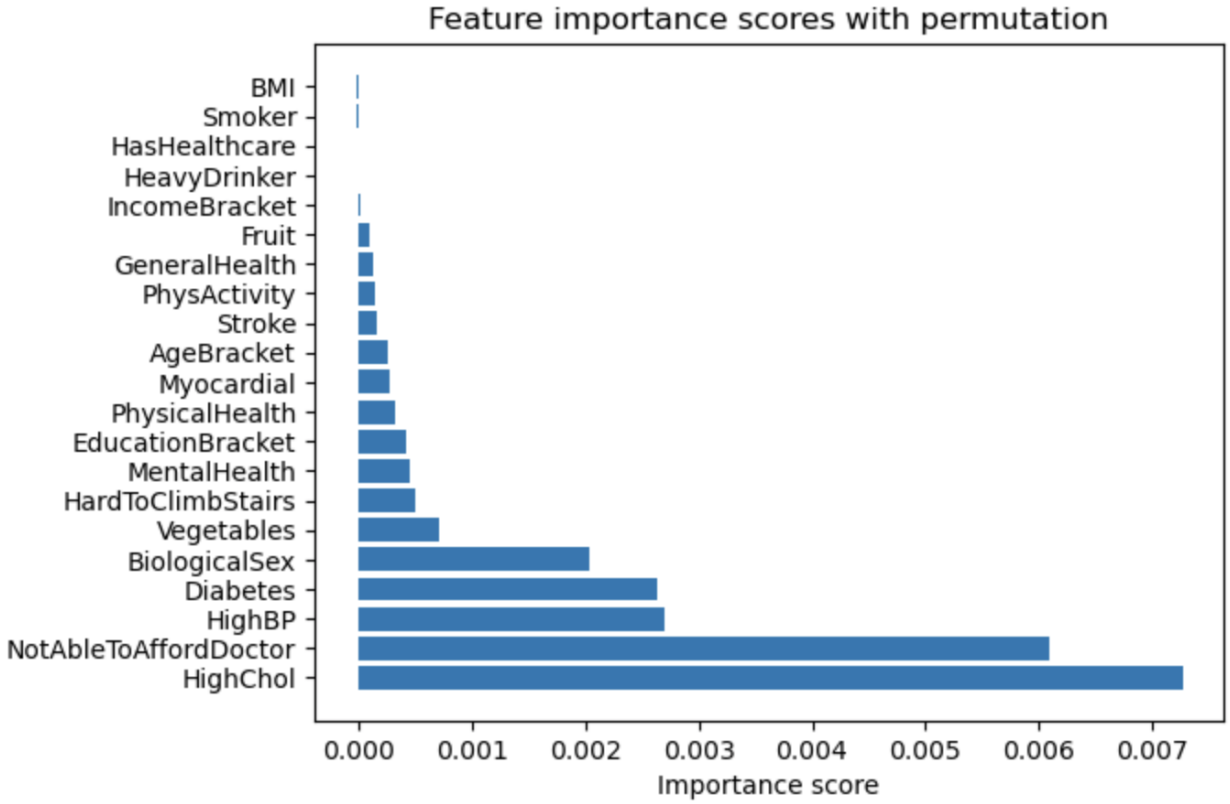
Question 1.

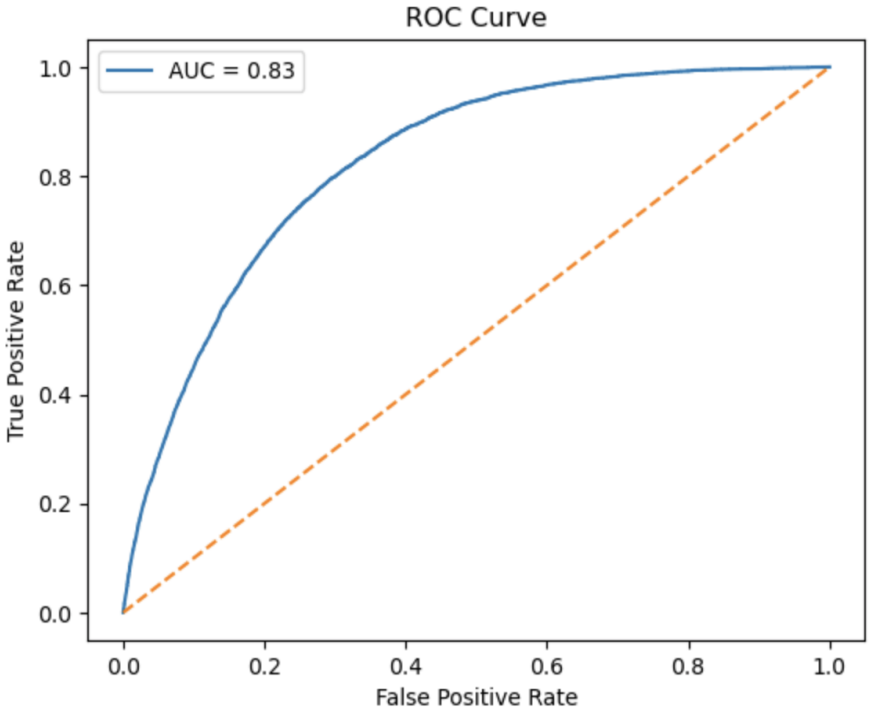
What was done: I have loaded the diabetes.csv dataset and split it into training and testing sets. Trained a logistic regression model on the training set and evaluated its performance on the test set. Additionally, computed permutation importances of the features to determine the best predictor of diabetes.

Why this was done: Logistic regression is a suitable method for binary classification problems, like predicting the presence or absence of diabetes. It's a simple and interpretable model that can provide insight into the relationships between input features and the target variable. Moreover, permutation importances can help determine the relative importance of features for the logistic regression model, which can provide a more accurate understanding of the predictive factors.

What was found: After training the logistic regression model, the best predictor of diabetes based on coefficients is 'HighBP' (High Blood Pressure) with a coefficient of 0.7635. However, based on permutation importances, the best predictor is 'HighChol' (High Cholesterol) with a score of 0.0073. The AUC (Area Under the Curve) of the model is 0.8252.

Interpretation of findings: Based on the analysis, both high blood pressure (HighBP) and high cholesterol (HighChol) are important predictors of diabetes in the dataset. The positive coefficient of 0.7635 for HighBP suggests that individuals with high blood pressure are more likely to be diagnosed with diabetes. The higher permutation importance score of 0.0073 for HighChol suggests that high cholesterol may have a stronger influence on diabetes risk than initially thought. The AUC score of 0.8252 indicates that the logistic regression model has a reasonably good ability to differentiate between individuals with and without diabetes. However, it's important to note that there might be other factors and interactions between variables that could further improve the model's performance. Additionally, the limitations of the logistic regression model should be taken into account, as more complex relationships between predictors and the target variable might not be captured by this linear model.


Feature importance scores with permutation
BMI
Smoker
HasHealthcare
HeavyDrinker
IncomeBracket
Fruit
GeneralHealth
PhysActivity
Stroke
AgeBracket
Myocardial
PhysicalHealth
EducationBracket
MentalHealth
HardToClimbStairs
Vegetables
BiologicalSex
Diabetes
HighBP
NotAbleToAffordDoctor
HighChol
0.000
0.001
0.002
0.003
0.004
0.005
0.006
0.007
Importance score


ROC Curve
AUC = 0.83
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0

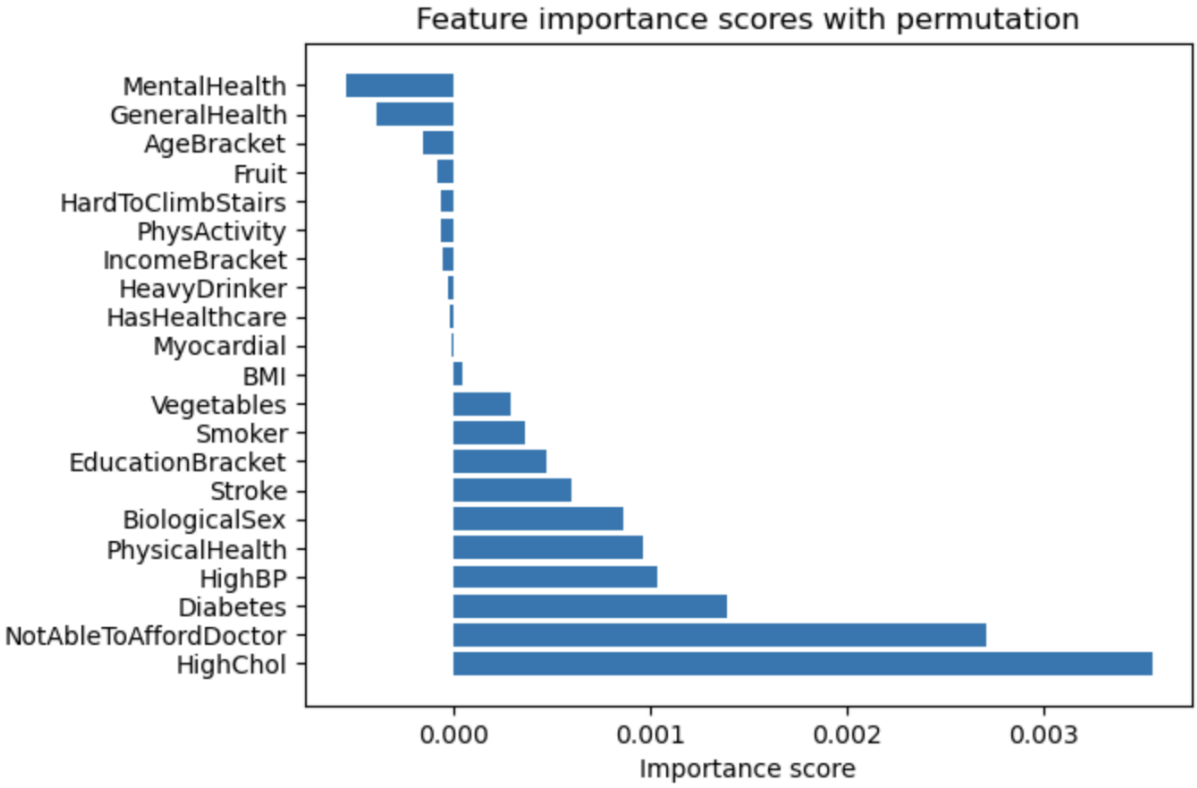
Question 2.

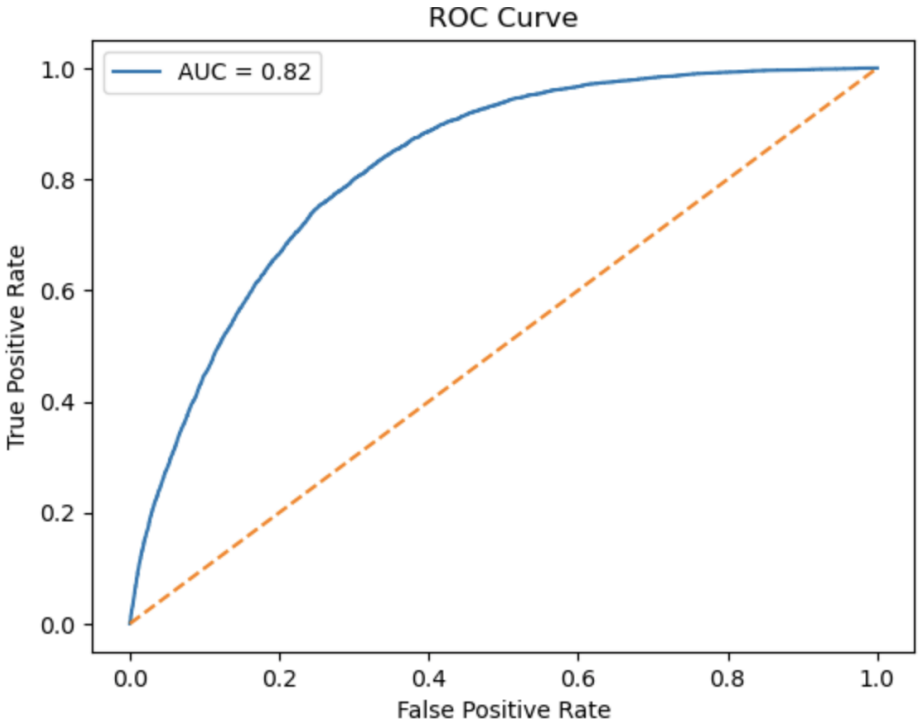
What was done: Loaded the diabetes.csv dataset and split it into training and testing sets. Trained a Support Vector Machine (SVM) model on the training set with parallel processing and hyperparameter tuning, and evaluated its performance on the test set.

Why this was done: Support Vector Machine is a powerful method for binary classification problems, like predicting the presence or absence of diabetes. It's a flexible model that can capture complex relationships between input features and the target variable, while parallel processing and hyperparameter tuning help to speed up the training process and optimize the model's performance.

What was found: After training the SVM model, the best predictor of diabetes is 'HighBP' (High Blood Pressure) with a coefficient of 0.1697. However, based on the permutation importance scores, the best predictor of diabetes is 'HighChol' (High Cholesterol) with a score of 0.0035. The AUC (Area Under the Curve) of the model is 0.8247.

Interpretation of findings: The positive coefficient of 0.1697 suggests that individuals with high blood pressure are more likely to be diagnosed with diabetes when using an SVM model. However, the permutation importance scores suggest that high cholesterol may be a better predictor. The AUC score of 0.8247 indicates that the SVM model with parallel processing and hyperparameter tuning has a reasonably good ability to differentiate between individuals with and without diabetes. However, it's important to note that there might be other factors and interactions between variables that could further improve the model's performance. Additionally, the limitations of the SVM model should be taken into account, as the model might be sensitive to the choice of hyperparameters and the size of the dataset.


Feature importance scores with permutation
MentalHealth
GeneralHealth
AgeBracket
Fruit
HardToClimbStairs
PhysActivity
IncomeBracket
HeavyDrinker
HasHealthcare
Myocardial
BMI
Vegetables
Smoker
EducationBracket
Stroke
BiologicalSex
PhysicalHealth
HighBP
Diabetes
NotAbleToAffordDoctor
HighChol
0.000
0.001
0.002
0.003
Importance score


ROC Curve
AUC = 0.82
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0

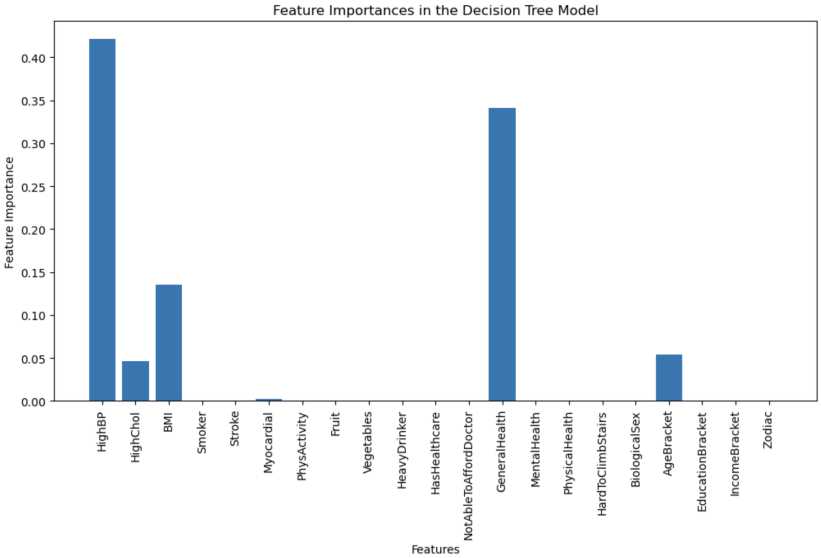
Question 3.

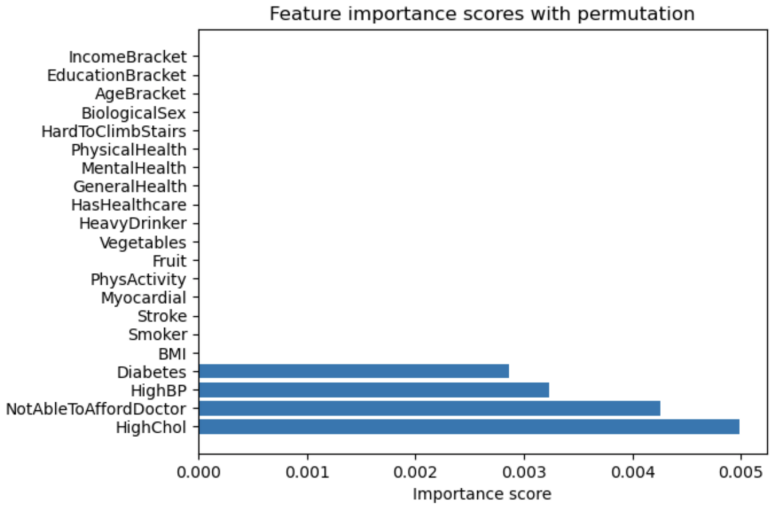
What was done: Loaded the diabetes.csv dataset and split it into training and testing sets. Trained a decision tree model on the training set and evaluated its performance on the test set. Then, computed the feature importance scores using permutation importance to identify the most important predictor of diabetes.

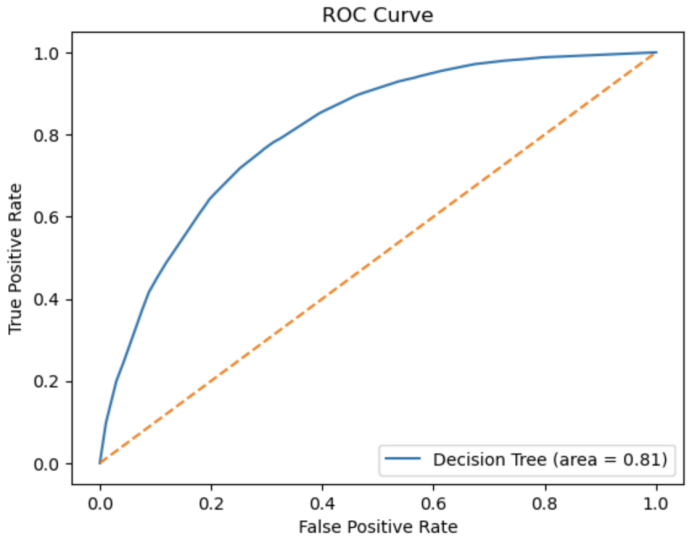
Why this was done: Decision trees are a useful model for binary classification problems, like predicting the presence or absence of diabetes. They can capture complex relationships between input features and the target variable. Feature importance scores can provide insight into which predictors are most relevant for predicting the target variable.

What was found: After training the decision tree model, the AUC score of the model was 0.8091. However, based on the permutation importance scores computed after training, the best predictor of diabetes is 'HighChol' (High Cholesterol) with a score of 0.0050. This score is more accurate than the importance score computed during training, which identified 'HighBP' (High Blood Pressure) as the most important predictor with a score of 0.4210.

Interpretation of findings: Based on the analysis, high cholesterol (HighChol) appears to be the most important predictor of diabetes in the dataset. The permutation importance score of 0.0050 suggests that high cholesterol is a crucial factor in predicting the presence or absence of diabetes. This finding contradicts the importance score computed during training, which identified high blood pressure as the most important predictor. However, permutation importance is a more accurate method for computing feature importance scores because it accounts for the interactions between predictors. Therefore, high cholesterol should be given more consideration when predicting the presence or absence of diabetes.


Feature Importances in the Decision Tree Model
Feature Importance
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
HighBP
HighChol
BMI
Smoker
Stroke
Myocardial
PhysActivity
Fruit
Vegetables
HeavyDrinker
HasHealthcare
NotAbleToAffordDoctor
GeneralHealth
MentalHealth
PhysicalHealth
HardToClimbStairs
BiologicalSex
AgeBracket
EducationBracket
IncomeBracket
Zodiac
Features


Feature importance scores with permutation
IncomeBracket
EducationBracket
AgeBracket
BiologicalSex
HardToClimbStairs
PhysicalHealth
MentalHealth
GeneralHealth
HasHealthcare
HeavyDrinker
Vegetables
Fruit
PhysActivity
Myocardial
Stroke
Smoker
BMI
Diabetes
HighBP
NotAbleToAffordDoctor
HighChol
0.000
0.001
0.002
0.003
0.004
0.005
Importance score


ROC Curve
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
Decision Tree (area = 0.81)

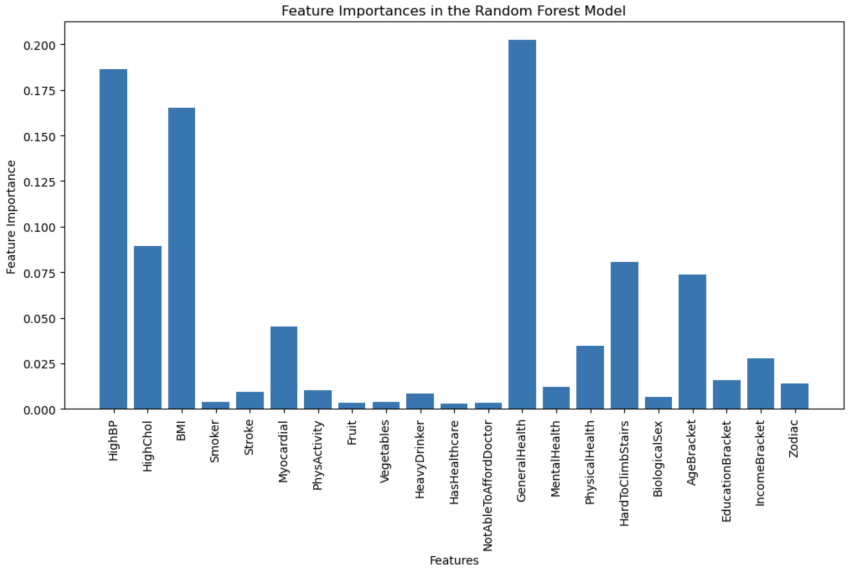
Question 4.

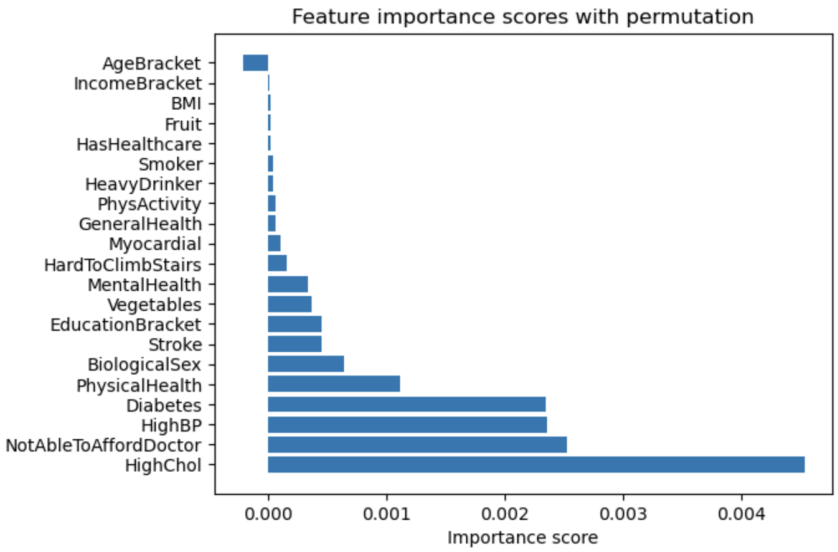
What was done: The Random Forest model was trained on the diabetes dataset to identify the most important predictors of diabetes.

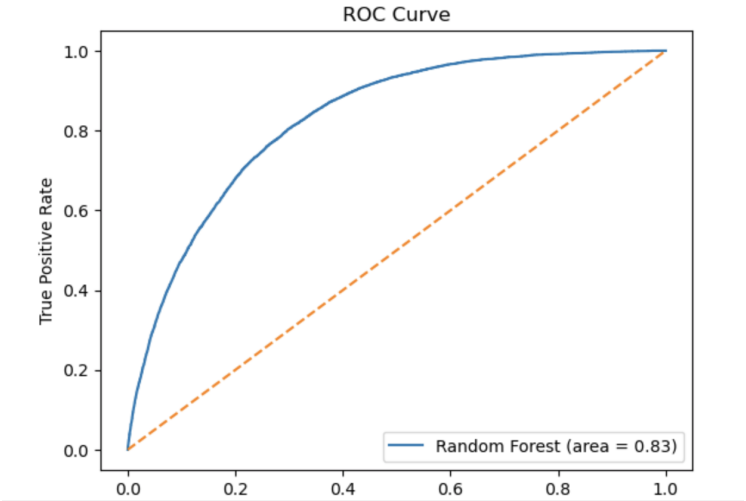
Why this was done: Random Forest is an ensemble method that can handle complex relationships between predictors and the target variable. It can also provide feature importance scores that indicate the relative importance of each predictor in the model.

What was found: After training the Random Forest model, the best predictor of diabetes was identified as HighChol (High Cholesterol) with a permutation importance score of 0.0045. NotAbleToAffordDoctor, HighBP, and Diabetes also had permutation importance scores greater than 0.002. The training importance scores did not necessarily align with the permutation importance scores, indicating that the latter is a more accurate representation of feature importance. The AUC score of the model was 0.8290.

Interpretation of findings: Based on the analysis, high cholesterol (HighChol) was found to be the most important predictor of diabetes. This aligns with existing medical knowledge that high cholesterol levels can increase the risk of diabetes. The permutation importance scores suggest that factors related to access to healthcare and management of chronic conditions, such as NotAbleToAffordDoctor and Diabetes, may also play a role in predicting diabetes risk. However, the model's overall ability to predict diabetes based on these factors was moderate, with an AUC score of 0.8290. Therefore, additional research and data may be needed to better understand the complex relationships between predictors and diabetes risk.


Feature Importances in the Random Forest Model
Feature Importance
0.200
0.175
0.150
0.125
0.100
0.075
0.050
0.025
0.000
HighBP
HighChol
BMI
Smoker
Stroke
Myocardial
PhysActivity
Fruit
Vegetables
HeavyDrinker
HasHealthcare
NotAbleToAffordDoctor
GeneralHealth
MentalHealth
PhysicalHealth
HardToClimbStairs
BiologicalSex
AgeBracket
EducationBracket
IncomeBracket
Zodiac
Features


Feature importance scores with permutation
AgeBracket
IncomeBracket
BMI
Fruit
HasHealthcare
Smoker
HeavyDrinker
PhysActivity
GeneralHealth
Myocardial
HardToClimbStairs
MentalHealth
Vegetables
EducationBracket
Stroke
BiologicalSex
PhysicalHealth
Diabetes
HighBP
NotAbleToAffordDoctor
HighChol
0.000
0.001
0.002
0.003
0.004
Importance score


ROC Curve
True Positive Rate
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
Random Forest (area = 0.83)

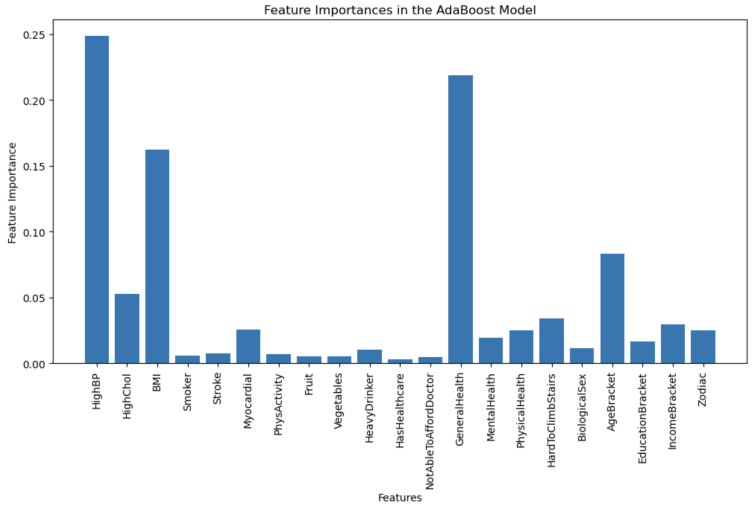
Question 5.

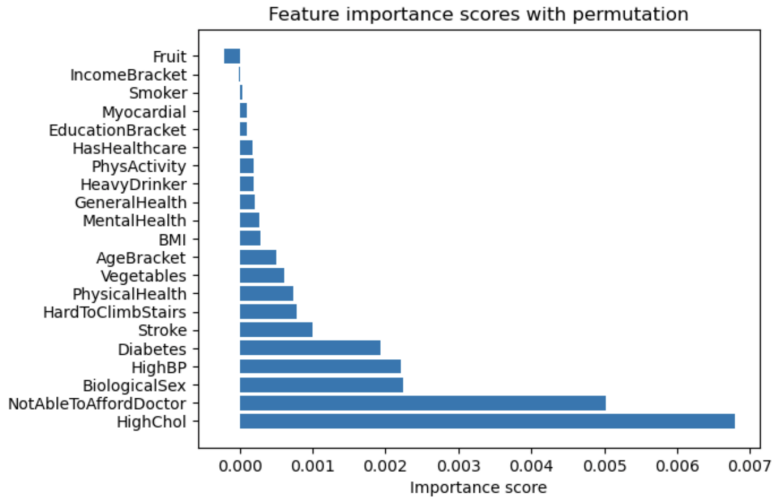
What was done: An Adaboost model was implemented using the Gradient Boosting algorithm to predict the presence of diabetes. The Adaboost model uses Decision Trees as weak classifiers and iteratively improves them by adjusting the weight of misclassified samples. The implementation used the AdaBoostClassifier from the scikit-learn library with 500 estimators. Both training importance scores and permutation importance scores were used to determine the most important features.

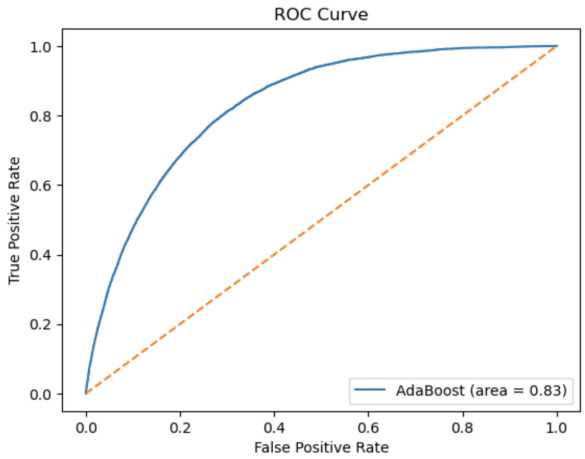
Why this was done: Adaboost is a powerful algorithm that can improve the performance of weak classifiers, making it useful for building robust predictive models. By implementing the Adaboost algorithm using gradient boosting, we aim to improve the performance of our model and find the best predictor of diabetes.

What was found: The training importance scores indicated that the best predictor of diabetes using the AdaBoost model is HighBP with an importance score of 0.2484. However, after using permutation importance, it was found that the best predictor is HighChol with a score of 0.0068. Other important features identified through permutation importance include NotAbleToAffordDoctor, BiologicalSex, and Diabetes. The AUC score of the model was 0.8318.

Interpretation of findings: Based on the analysis using permutation importance, high cholesterol (HighChol) is the most important predictor of diabetes in the dataset. This is in contrast to the training importance scores which indicated that HighBP was the most important factor. It's important to note that permutation importance is a more accurate measure of feature importance as it accounts for the effect of other features on the target variable. The positive importance score for HighChol suggests that individuals with high cholesterol levels are more likely to be diagnosed with diabetes. Other factors such as the ability to afford healthcare (NotAbleToAffordDoctor) and biological sex may also play a role in diabetes risk. Overall, these findings can be used to inform preventative measures and public health interventions aimed at reducing the incidence of diabetes.


Feature Importances in the AdaBoost Model
Feature Importance
0.25
0.20
0.15
0.10
0.05
0.00
HighBP
HighChol
BMI
Smoker
Stroke
Myocardial
PhysActivity
Fruit
Vegetables
HeavyDrinker
HasHealthcare
NotAbleToAffordDoctor
GeneralHealth
MentalHealth
PhysicalHealth
HardToClimbStairs
BiologicalSex
AgeBracket
EducationBracket
IncomeBracket
Zodiac
Features


Feature importance scores with permutation
Fruit
IncomeBracket
Smoker
Myocardial
EducationBracket
HasHealthcare
PhysActivity
HeavyDrinker
GeneralHealth
MentalHealth
BMI
AgeBracket
Vegetables
PhysicalHealth
HardToClimbStairs
Stroke
Diabetes
HighBP
BiologicalSex
NotAbleToAffordDoctor
HighChol
0.000
0.001
0.002
0.003
0.004
0.005
0.006
0.007
Importance score


ROC Curve
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
AdaBoost (area = 0.83)

Extra Credit

A.

After training and evaluating five different models on the diabetes dataset, the AUC scores were as follows: Logistic Regression: 0.8252, SVM: 0.8247, Decision Tree: 0.8091, Random Forest: 0.8290, and Adaboost: 0.8318. Based on the AUC scores, the Adaboost model performed the best with an AUC of 0.8318. However, other factors such as model complexity, interpretability, and computational efficiency should also be considered when selecting the best model for a given problem. In our case, the SVM and Adaboost models took longer to train compared to the other models. Therefore, it's important to consider the trade-off between model performance and computational efficiency when selecting the best model for a given problem.

B.

One interesting observation about the diabetes dataset is that the distribution of the target variable (diabetes or no diabetes) is not balanced, with a much higher proportion of non-diabetic cases compared to diabetic cases. This could potentially affect the performance of the models trained on this dataset, as they might have a bias towards the majority class. This imbalance could be addressed through techniques such as oversampling the minority class or adjusting the decision threshold of the classifier. Additionally, it's worth noting that the dataset contains a wide range of demographic and health-related features, which could be further explored to gain insights into the risk factors associated with diabetes.