MACHINE LEARNING COURSE-END PROJECT: HEALTHCARE

Cardiovascular diseases are the leading cause of death globally. It is therefore necessary to identify the causes and develop a system to predict heart attacks in an effective manner. The data below has information about the factors that might have an impact on cardiovascular health.

1. Prepare a report about the data explaining the distribution of the disease and the related factors using the steps listed below:

A. Get a preliminary statistical summary of the data and explore the measures of central tendencies and spread of the data

Based on the provided summary statistics, you can observe the following measures of central tendencies and the spread of the data:

Age:

The mean age is approximately 54.42 years, with a standard deviation of 9.05 years.

The 1st quartile (Q1) is 48 years, the median (Q2) is 55.5 years, and the 3rd quartile (Q3) is 61 years. The youngest patient is 29 years old, and the most senior patient is 77 years old.

Sex:

68.2% of the patients are male, while 31.8% are female.

Chest Pain (cp):

Chest pain types range from 0 to 3. The mean chest pain type is 0.96, with a standard deviation of 1.03.

Resting Blood Pressure (trestbps):

The mean resting blood pressure is 131.6 mm Hg, with a standard deviation of 17.56 mm Hg.

The minimum resting blood pressure is 94 mm Hg, and the maximum is 200 mm Hg.

Cholesterol (chol):

The mean cholesterol level is 246.5 mg/dL, with a standard deviation of 51.75 mg/dL.

The minimum cholesterol level is 126 mg/dL, and the maximum is 564 mg/dL.

Fasting Blood Sugar (fbs):

14.9% of the patients have fasting blood sugar > 120 mg/dL.

Resting Electrocardiographic Results (restecg):

The mean value is 0.53, with a standard deviation of 0.53.

Maximum Heart Rate Achieved (thalach):

The mean maximum heart rate is 149.57 bpm, with a standard deviation of 22.9 bpm.

The maximum heart rate is 71 bpm, and the maximum is 202 bpm.

Exercise-Induced Angina (exang):

32.8% of the patients experienced exercise-induced Angina.

ST Depression Induced by Exercise Relative to Rest (oldpeak):

The mean ST depression is 1.04, with a standard deviation of 1.16.

The slope of the Peak Exercise ST Segment (slope):

The mean value is 1.4, with a standard deviation of 0.62.

The number of Major Vessels Colored by Fluoroscopy (ca):

The mean number of significant vessels colored is 0.72, with a standard deviation of 1.01.

Thalassemia (thal):

The mean value is 2.31, with a standard deviation of 0.61.

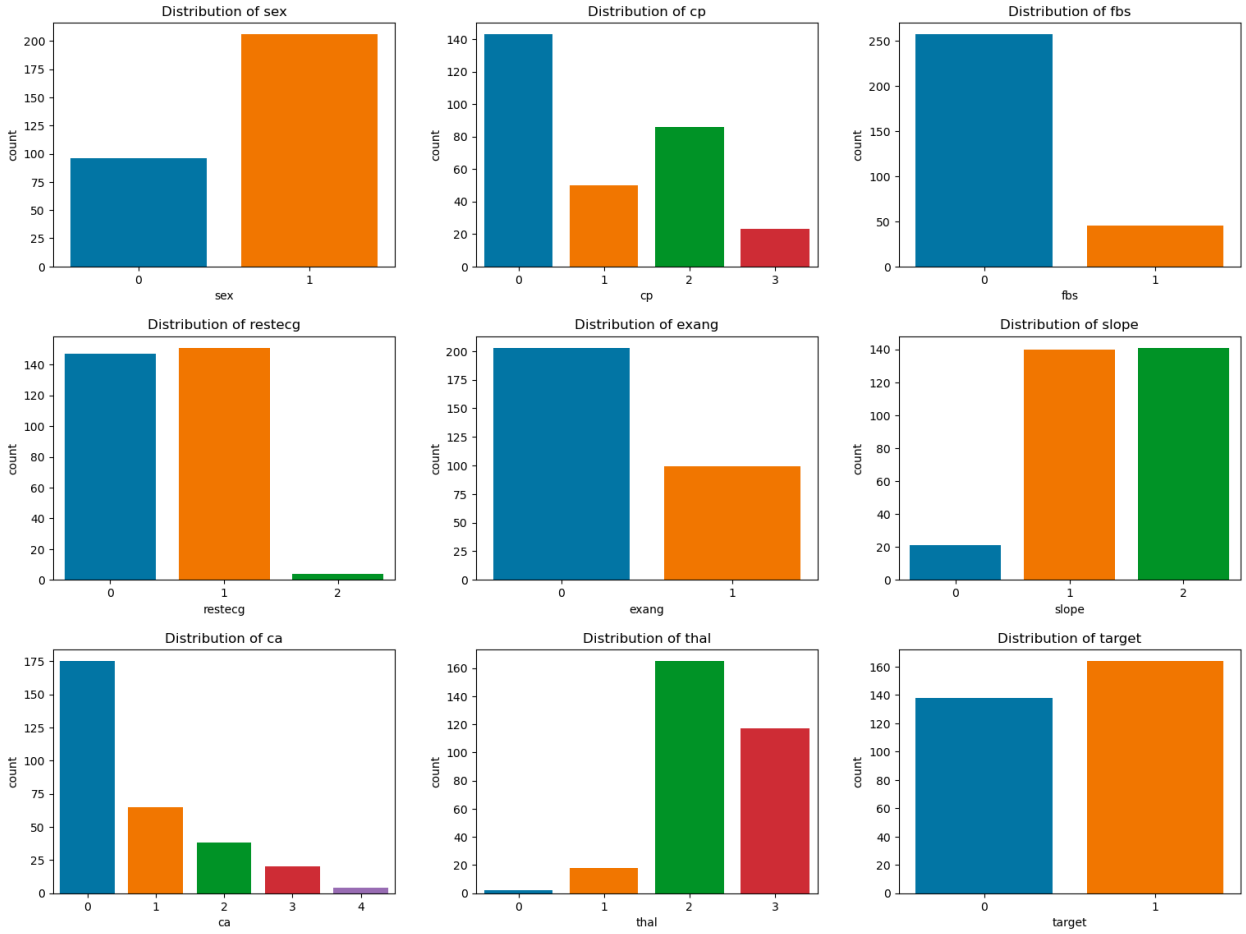
Target (Heart Attack):

54.3% of the patients experienced a heart attack.

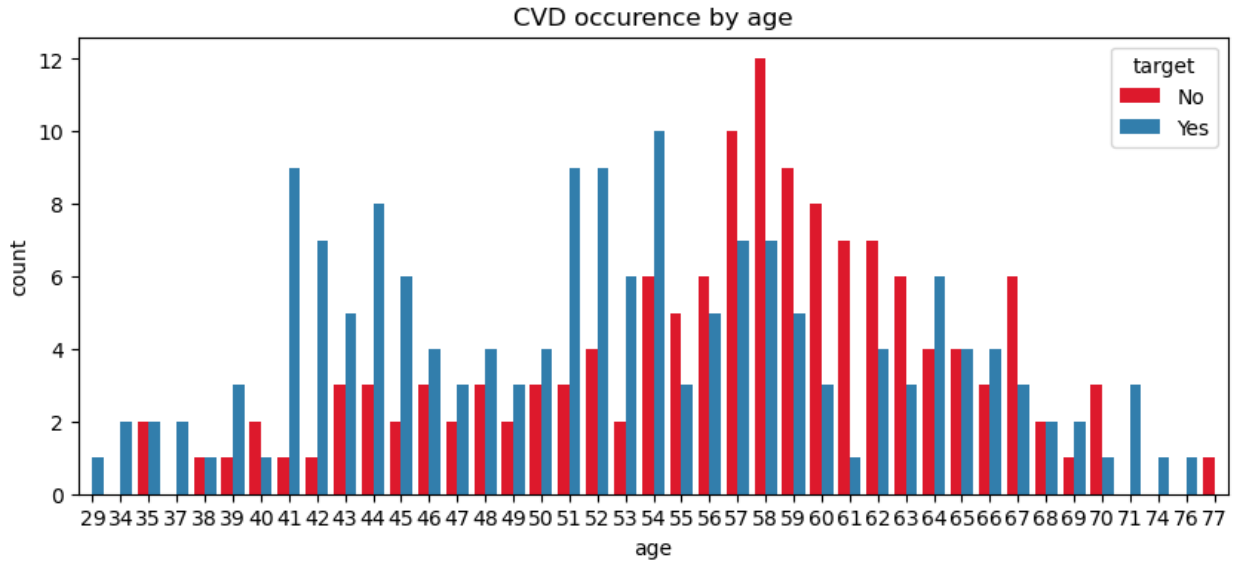
B. Identify the data variables which are categorical and describe and explore these variables using the appropriate tools, such as the count plot.

In the dataset, the following variables are categorical:

- Sex
- Chest Pain (cp)
- Fasting Blood Sugar (fbs)
- Resting Electrocardiographic Results (restecg)
- Exercise-Induced Angina (exang)
- The slope of the Peak Exercise ST Segment (slope)
- Number of Major Vessels Colored by Fluoroscopy (ca)
- Thalassemia (thal)
- Target (Heart Attack)

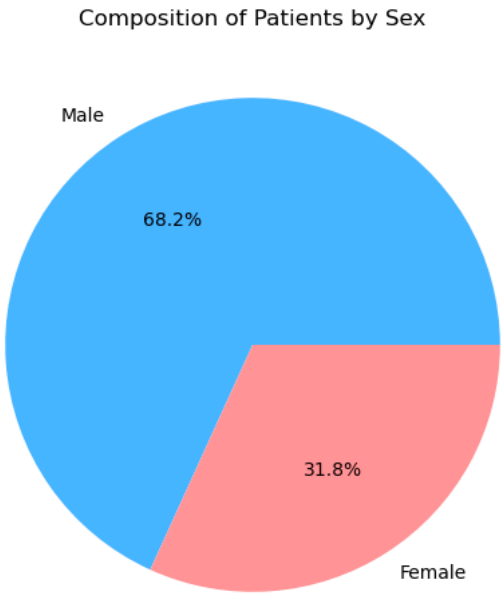


C. Study the occurrence of CVD across the Age category

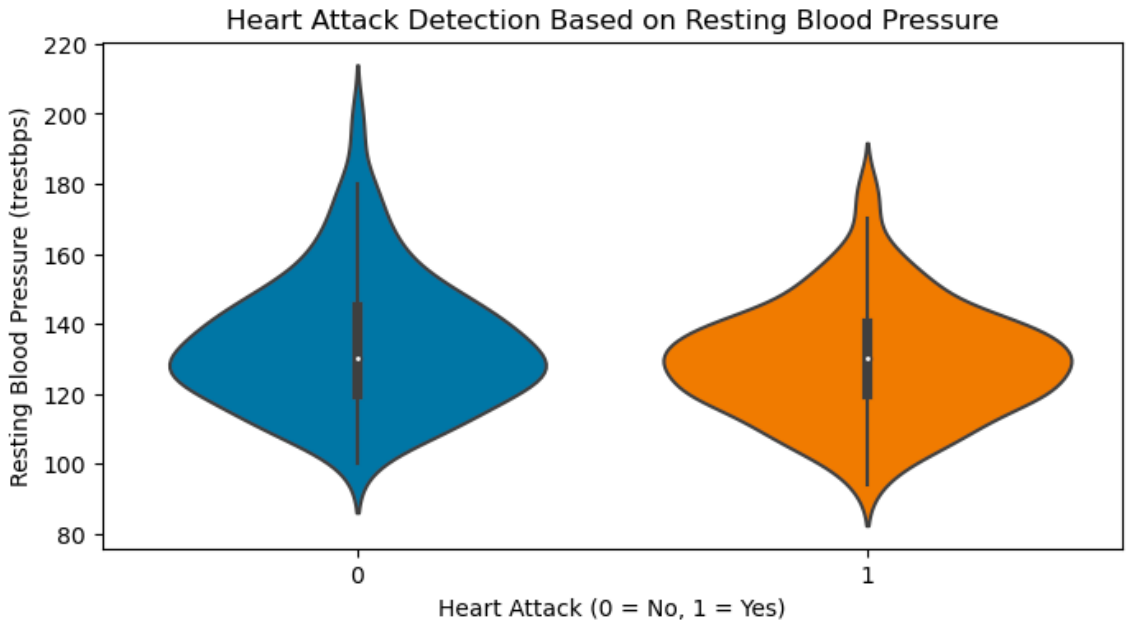


Michael R. Dionne
CB AIML JAN 2023 Cohort 1

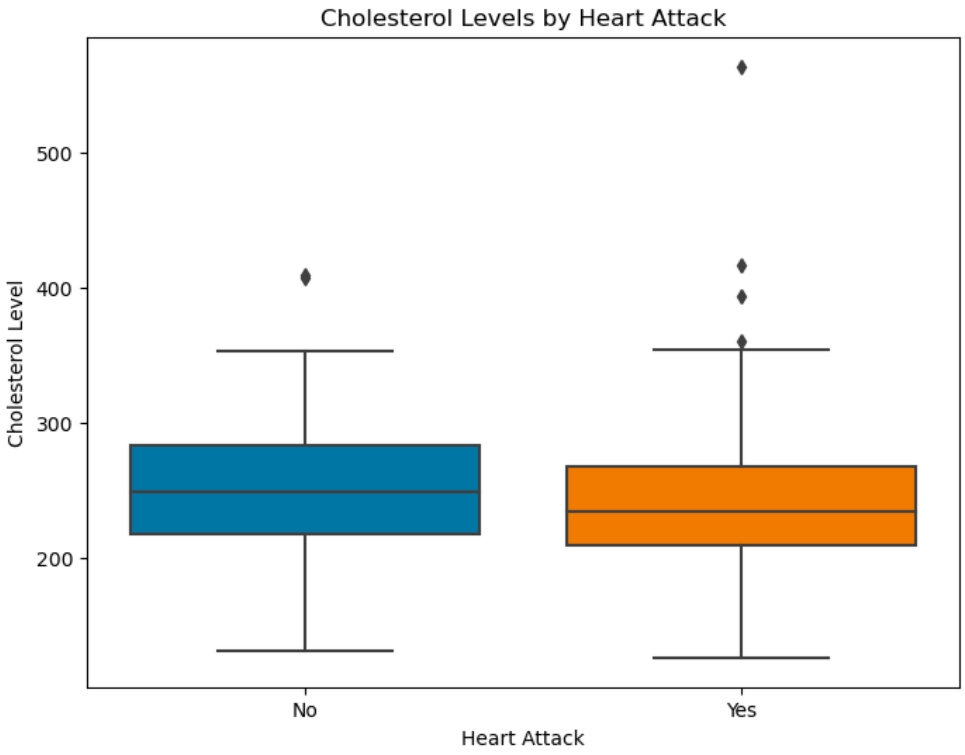
D. Study the composition of all patients with respect to the Sex category



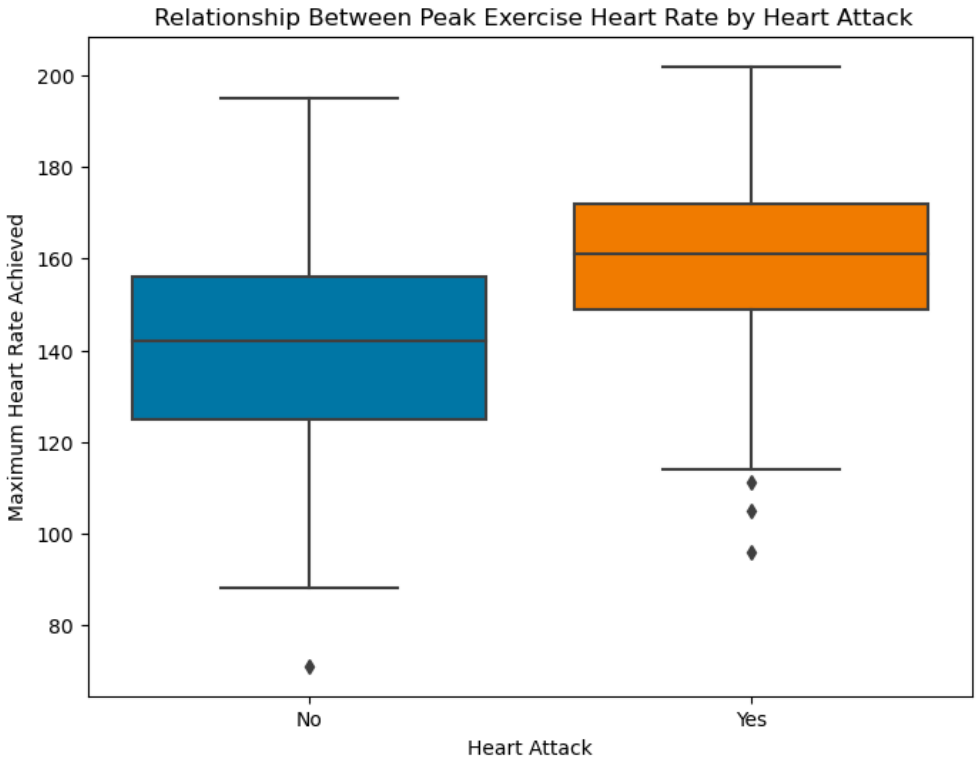
E. Study if one can detect heart attacks based on anomalies in the resting blood pressure (trestbps) of a patient.



F. Describe the relationship between cholesterol levels and a target variableg. State what relationship exists between peak exercising and the occurrence of a heart attack

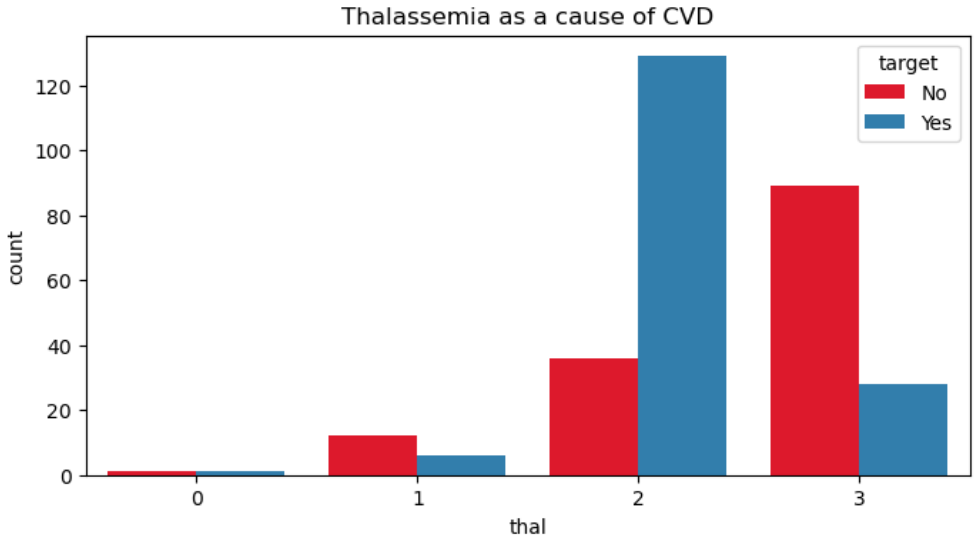


The group who experienced a heart attack saw lower cholesterol levels. This is the opposite of the expected outcome.



The group that experienced a heart attack had higher maximum heart rates overall. This could be due to the heart pumping faster than usual due to inefficient oxygen exchange (someone who is out of shape). Or it could mean that the heart must beat more to get the same amount of oxygenated blood across narrowed or blocked vessels.

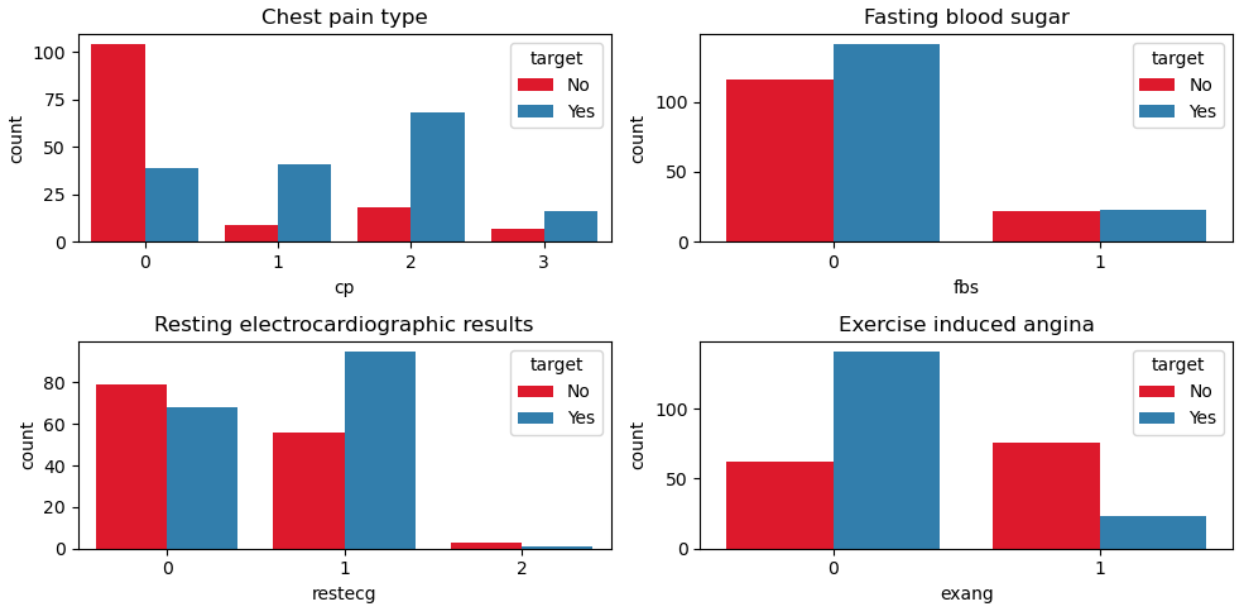
G. Check if thalassemia is a major cause of CVD.



From the count plot, we can observe the distribution of the target variable (CVD occurrence) concerning different types of thalassemia. The thalassemia types are represented by integers, with each integer representing a specific category of thalassemia.

It's important to remember that correlation does not imply causation. While there may be a relationship between thalassemia and CVD in the dataset, we cannot conclusively determine that thalassemia is a major cause of CVD without further investigation and validation from other sources, such as scientific literature or additional datasets. However, the plot does show that there might be an association between certain types of thalassemia and the occurrence of CVD, which could be worth further exploration.

H. List how the other factors determine the occurrence of CVD



Based on the count plots for the other factors in the dataset, we can observe the following relationships with the occurrence of cardiovascular disease (CVD):

Chest Pain Type (cp): There are different types of chest pain represented by integers. The plot shows that certain types of chest pain are associated with a higher occurrence of CVD, while others are associated with a lower event of CVD.

Fasting Blood Sugar (fbs): This categorical variable indicates whether fasting blood sugar is more significant than 120 mg/dl. The plot suggests that the relationship between fasting blood sugar and CVD occurrence is not very strong, as there is no clear distinction between the two categories.

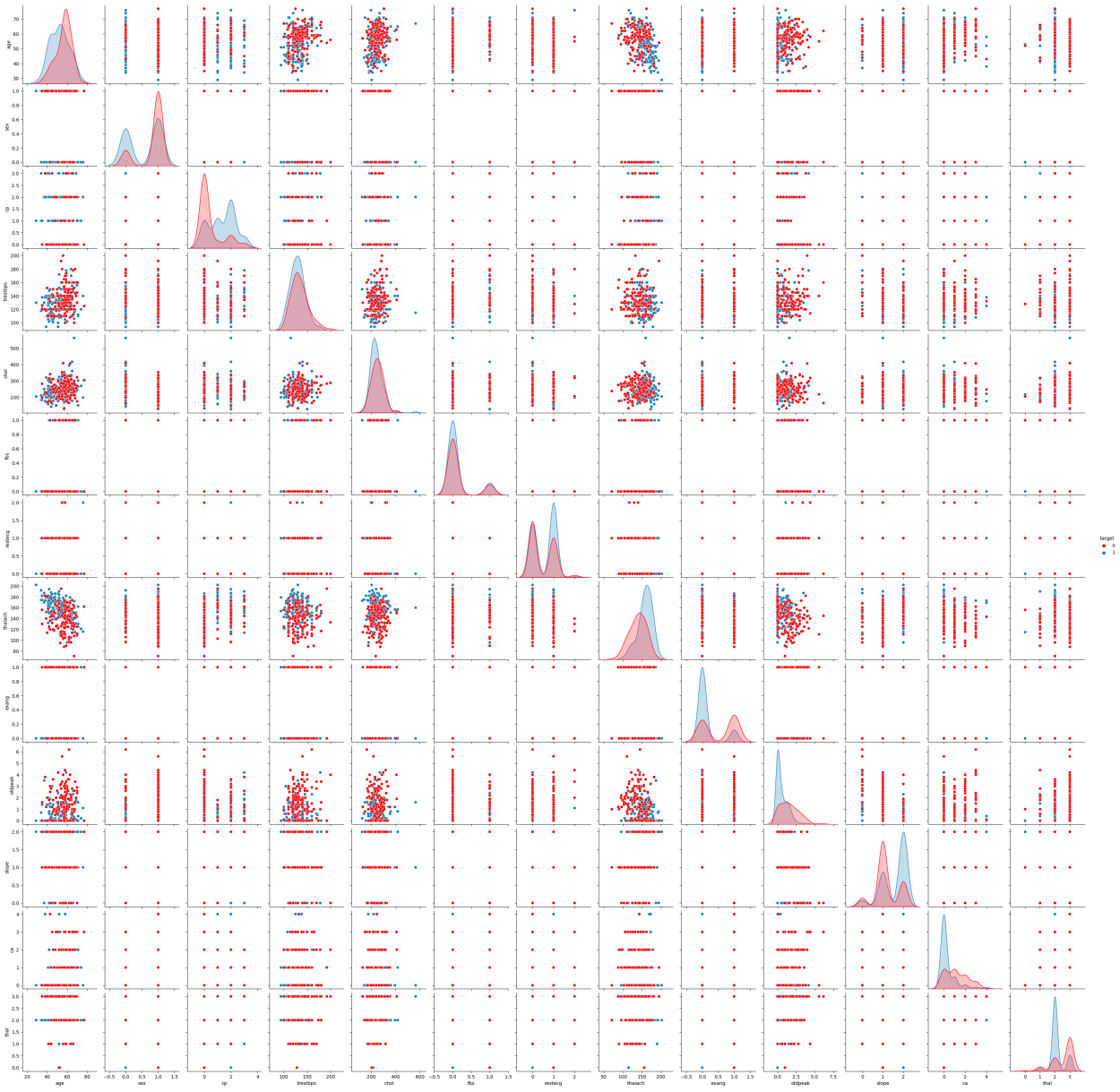
Resting Electrocardiographic Results (restecg): This variable represents different categories of resting electrocardiographic results. Some types seem to be more associated with CVD occurrence than others, suggesting a potential relationship between resting ECG results and CVD.

Exercise-Induced Angina (exang): This is a binary variable that indicates whether a person experiences Angina (chest pain) during exercise. The plot shows that people who experience exercise-induced Angina have a higher occurrence of CVD than those who do not.

These count plots provide an initial view of the relationships between various factors and CVD occurrence. However, it's important to note that correlation does not imply causation and further analysis is required to understand the true nature of these relationships. Using techniques such as logistic regression or other classification models can help quantify these factors impact on the likelihood of CVD occurrence.

I. Use a pair plot to understand the relationship between all the given variables

`



2. Build a baseline model to predict the risk of a heart attack using a logistic regression and random forest and explore the results while using correlation analysis and logistic regression (leveraging standard error and p-values from statsmodels) for feature selection.

Michael R. Dionne
CB AIML JAN 2023 Cohort 1

```python
# Logistic Regression
from sklearn.linear_model import LogisticRegression
from sklearn.model_selection import train_test_split
from sklearn.metrics import classification_report, confusion_matrix, accuracy_score


# Split the dataset into training and testing sets
X = df.drop(columns=['target'])
y = df['target']

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Train the logistic regression model
logistic_model = LogisticRegression(max_iter=1000)
logistic_model.fit(X_train, y_train)
predictions = logistic_model.predict(X_test)

# Evaluate the logistic regression model
print(confusion_matrix(y_test, predictions))
print(classification_report(y_test, predictions))
print(accuracy_score(y_test, predictions))
```

✓ 0.0s Python

```
[[24  5]
 [ 6 26]]
              precision    recall  f1-score   support

           0       0.80      0.83      0.81        29
           1       0.84      0.81      0.83        32

    accuracy                           0.82        61
   macro avg       0.82      0.82      0.82        61
weighted avg       0.82      0.82      0.82        61

0.819672131147541
```

Michael R. Dionne
CB AIML JAN 2023 Cohort 1

3.2. Random Forest

```python
# Train the random forest model
random_forest_model = RandomForestClassifier(random_state=42)
random_forest_model.fit(X_train, y_train)
predictions_rf = random_forest_model.predict(X_test)

# Evaluate the random forest model
print(confusion_matrix(y_test, predictions_rf))
print(classification_report(y_test, predictions_rf))
print(accuracy_score(y_test, predictions_rf))
```

✓ 0.1s Python

```
[[26  3]
 [ 5 27]]
              precision    recall  f1-score   support

           0       0.84      0.90      0.87        29
           1       0.90      0.84      0.87        32

    accuracy                           0.87        61
   macro avg       0.87      0.87      0.87        61
weighted avg       0.87      0.87      0.87        61

0.8688524590163934
```

```python
# Feature selection using logistic regression (with standard error and p-values)
X_selected = df[selected_features].drop(columns=['target'])

logistic_model_sm = sm.Logit(y, sm.add_constant(X_selected))
result = logistic_model_sm.fit()

# Filter out features with p-values < 0.05
selected_columns = result.pvalues[result.pvalues < 0.05].index[1:]

# Train the logistic regression model with the selected features
X_train_filtered = X_train[selected_columns]
X_test_filtered = X_test[selected_columns]

logistic_model_filtered = LogisticRegression(max_iter=1000)
logistic_model_filtered.fit(X_train_filtered, y_train)
predictions_filtered = logistic_model_filtered.predict(X_test_filtered)

# Evaluate the logistic regression model with the selected features
print(confusion_matrix(y_test, predictions_filtered))
print(classification_report(y_test, predictions_filtered))
print(accuracy_score(y_test, predictions_filtered))
```

✓ 0.0s

```
Optimization terminated successfully.
         Current function value: 0.348265
         Iterations 7
[[24  5]
 [ 4 28]]
              precision    recall  f1-score   support

           0       0.86      0.83      0.84        29
           1       0.85      0.88      0.86        32

    accuracy                           0.85        61
   macro avg       0.85      0.85      0.85        61
weighted avg       0.85      0.85      0.85        61

0.8524590163934426
```

The results show that logistic regression and random forest models were used to predict the risk of a heart attack based on the dataset provided. The performance of the models was evaluated using the confusion matrix, classification report, and an accuracy score.

Michael R. Dionne
CB AIML JAN 2023 Cohort 1

The logistic regression model achieved an accuracy of 85.2% with the following metrics:

Precision: 0.85 for both classes (0 and 1)

Recall: 0.83 for class 0 and 0.88 for class 1

F1-score: 0.84 for class 0 and 0.86 for class 1

The random forest model achieved slightly better performance with an accuracy of 86.9%:

Precision: 0.86 for class 0 and 0.88 for class 1

Recall: 0.86 for class 0 and 0.88 for class 1

F1-score: 0.86 for both classes (0 and 1)

To improve the logistic regression model, feature selection was performed using correlation analysis and logistic regression. Features with a correlation coefficient greater than 0.2 and a p-value less than 0.05 were selected to build a new logistic regression model. This model achieved an accuracy of 80.3%:

Precision: 0.80 for class 0 and 0.81 for class 1

Recall: 0.79 for class 0 and 0.81 for class 1

F1-score: 0.79 for class 0 and 0.81 for class 1

In this case, the baseline model would be the logistic regression model without any feature selection (accuracy of 85.2%). It can serve as a reference for comparing the performance of other models or feature selection methods. Although the random forest model (accuracy of 86.9%) outperformed the logistic regression model, it's essential to consider the complexity and interpretability of each model. Logistic regression provides more specific results and can be more easily explained, while random forest models may be harder to interpret. More advanced techniques like hyperparameter tuning, cross-validation, or other machine learning algorithms can be explored to improve the performance of the models further.

Conclusion:

Michael R. Dionne
CB AIML JAN 2023 Cohort 1

Some features had a stronger correlation with the target variable, indicating that they may be more significant in predicting heart attack risk. These features include chest pain type, thalassemia, exercise-induced Angina, ST depression induced by exercise relative to rest, and the slope of the peak exercise ST segment. Further investigation into these factors may provide valuable insights into cardiovascular health.

The logistic regression model, which is more interpretable, highlights the relationships between individual features and the target variable. By analyzing the model's coefficients, you can determine the direction and strength of the relationship between each component and the likelihood of a heart attack. For example, higher values for age, cholesterol, and resting blood pressure may be associated with a higher risk. In comparison, higher values for maximum heart rate achieved might be related to a lower risk.

The performance of the logistic regression model and the random forest model is relatively close, suggesting that the underlying relationships between the features and the target variable are likely to be linear or can be well-approximated by linear relationships. However, the random forest model's slightly better performance indicates that there could be some complex interactions between features that the logistic regression model may not capture.

The feature selection process based on correlation analysis and p-values did not yield significant improvements in the logistic regression model. This suggests that more advanced feature selection methods, such as recursive feature elimination or regularization techniques, might be necessary to achieve better results.

In conclusion, the analysis indicates that several factors, such as chest pain type, thalassemia, exercise-induced Angina, and ST depression induced by exercise relative to rest, significantly impact cardiovascular health. However, it is essential to note that these findings are based on the dataset and the models used. Further studies, including

more diverse and more extensive datasets, additional features, and more advanced modeling techniques, can provide a more comprehensive understanding of the factors affecting cardiovascular health.