Data Exploration using pandas and numpy library

1. How many missing values are there in each column of the dataset?
2. What is the distribution of ages in the dataset, and how does it vary with the presence of heart disease? 3. Are there any outliers in the resting blood pressure and cholesterol features, and how might they affect the analysis?
4. What is the correlation between the continuous features, and how might these relationships impact feature selection?

Data Visualization using matplotlib and seaborn library

5. Can you create a histogram of cholesterol levels and annotate the mean value on the plot?
6. How does the maximum heart rate achieved vary across different types of chest pain?
7. Can you visualize the distribution of resting blood pressure using a violin plot, separated by heart disease presence?
8. Create a pair plot for the continuous features, colored by the presence of heart disease. What patterns do you observe?

Feature Engineering

9. How would you handle categorical variables like chest pain type and resting ECG results for modeling? 10. Based on the distributions of continuous features, would any benefit from normalization or scaling?
11. Can new features be engineered from existing ones, such as a feature indicating high risk based on age and cholesterol levels combined?
12. How would you address any imbalances in the dataset regarding the presence of heart disease?

Classification: Logistic Regression, Decision Tree, Naïve Bayes, Support Vector Machine, Random Forest

13. Train a Logistic Regression model and evaluate its performance using accuracy and the ROC curve. How does it perform?
14. How does the performance of a Decision Tree classifier compare with a Random Forest for predicting heart disease presence?
15. Implement a Naïve Bayes classifier and discuss its suitability for this dataset. What assumptions does it make about the features?
16. Which kernel in Support Vector Machine works best for this dataset, and how do you determine the optimal parameters?

Regression: Linear Regression, Decision Tree Regression, Support Vector Machine Regression, Random Forest Regression

17. Although it's a classification problem, as an exercise, predict the 'Oldpeak' using Linear Regression. What is the R^2 score?
18. How would a Decision Tree Regressor perform in predicting resting blood pressure, and how can overfitting be prevented?
19. Use Support Vector Machine Regression to predict maximum heart rate. How do different kernels affect the prediction?
20. Apply Random Forest Regression to predict cholesterol levels. How does the model performance vary with the number of estimators?