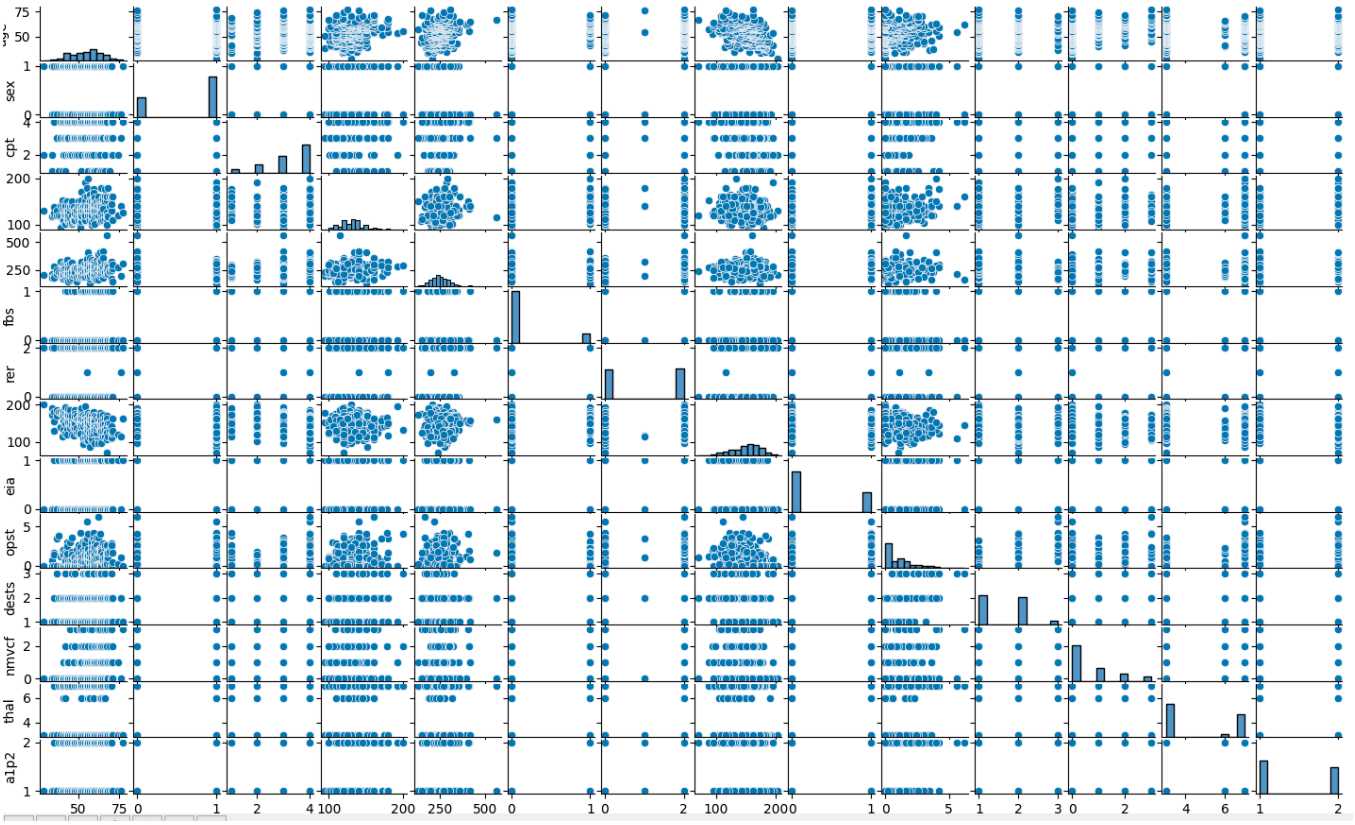
Problem 1: Read the database in from this heart1.csv file and analyze the data.

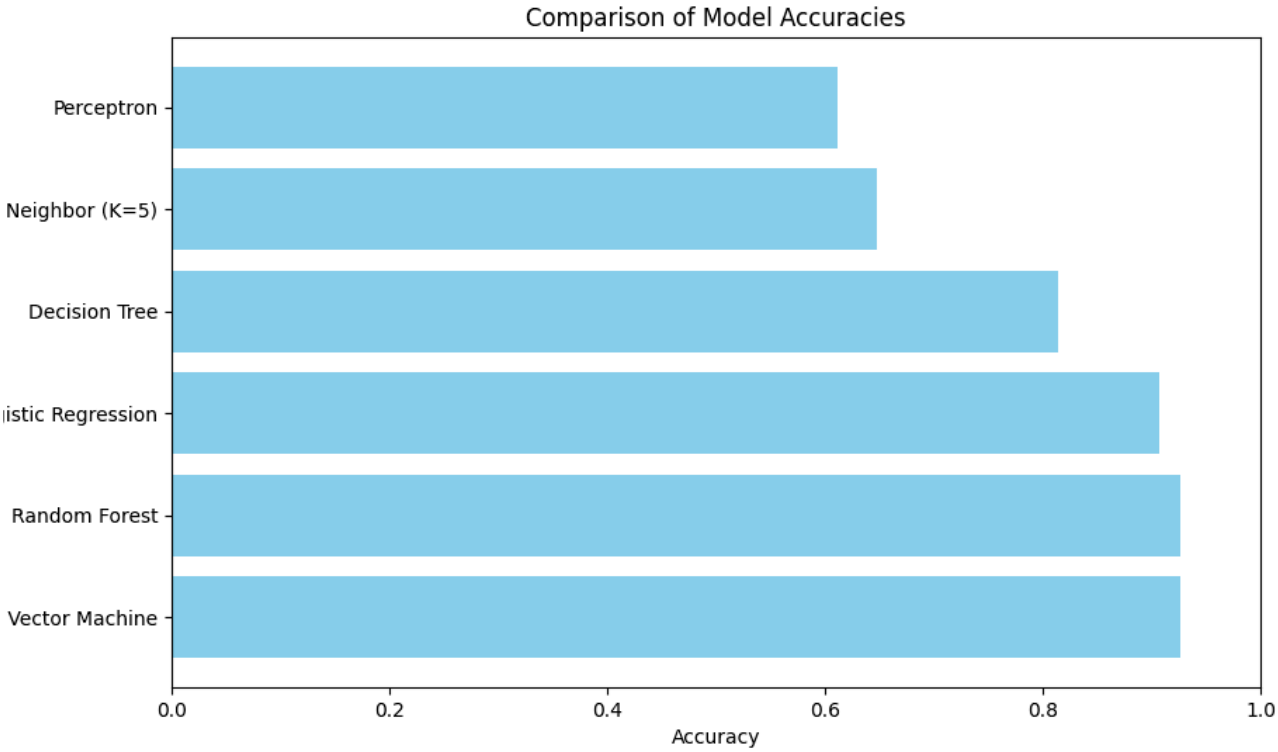


Results: The pairs you listed have covariance values beyond the threshold range of -10 to +10. Here's a breakdown of the covariance for each pair:

1. **('rbp', 'age'):** Covariance = 44.43
   This indicates a relatively strong positive relationship between resting blood pressure (rbp) and age.
2. **('sc', 'age'):** Covariance = 103.61
   There's a strong positive covariance between serum cholesterol (sc) and age, suggesting that older individuals tend to have higher cholesterol levels.
3. **('sc', 'rbp'):** Covariance = 159.73
   A strong positive covariance suggests that higher serum cholesterol is correlated with higher resting blood pressure.
4. **('mhr', 'age'):** Covariance = -84.87
   This negative covariance indicates that as age increases, maximum heart rate (mhr) tends to decrease.
5. **('mhr', 'rbp'):** Covariance = -16.19
   A slight negative relationship between maximum heart rate and resting blood pressure.
6. **('mhr', 'sc'):** Covariance = -22.44
   A slight negative covariance between maximum heart rate and serum cholesterol.
7. **('thal', 'mhr'):** Covariance = -11.39
   This indicates a weak negative relationship between thalassemia (thal) and maximum heart rate.

Observation: The correlation and covariance analysis presents relationships among various variables such as age, sex, chest pain type (cpt), resting blood pressure (rbp), serum cholesterol (sc), fasting blood sugar (fbs), and other medical indicators. Age has a positive correlation with rbp, sc, and a1p2, but a negative one with maximum heart rate (mhr). Sex shows notable correlations with thal and a1p2. Highly correlated pairs include age with mhr, opst with dests, and a1p2 with several factors like cpt, mhr, eia, nmvcf, and thal. These insights indicate potential associations useful for further medical or statistical analysis.

Problem 2: Split your heart1.cvs data into training and test datasets.



Results:

Model Accuracies:
Perceptron: 61.11%
Logistic Regression: 90.74%
Support Vector Machine: 92.59%
Decision Tree: 81.48%
Random Forest: 92.59%
K-Nearest Neighbor (K=5): 64.81%

Observation Based on the Model Accuracies:

From the model accuracies obtained, it is clear that **Support Vector Machine (SVM)** and **Random Forest** are the top-performing models, both achieving an accuracy of **92.59%**, indicating that they handle the complexity of the heart disease dataset well. **Logistic Regression** also performs robustly with an accuracy of **90.74%**, suggesting that the dataset has a relatively linear structure, as Logistic Regression is a linear model. On the other hand, **Decision Tree** performs decently with **81.48%**, which is lower than the ensemble Random Forest, reinforcing the idea that individual decision trees can overfit the data, while Random Forest mitigates this issue by averaging over many trees. The **K-Nearest Neighbors (K=5)** and **Perceptron** models, with accuracies of **64.81%** and **61.11%**, respectively, are the least effective in this context. KNN's performance may be affected by the choice of K and feature scaling, while Perceptron, being a simple linear classifier, likely struggles with the complexity of the dataset. Overall, SVM and Random Forest emerge as the most reliable models for predicting heart disease in this dataset.