1.

Observation:

1. The variables most correlated with the occurrence of heart disease (a1p2) include “thal”, “nmvcf”, “eia”, and “opst”. These four variables show significant influence on predicting the occurrence of heart disease, with correlation coefficients of 0.53, 0.46, 0.42, and 0.42, respectively. Among these variables, “thal” has the highest correlation at 0.53, indicating that it could be a strong indicator of heart disease.
2. The negative correlation of “mhr”, at -0.42, suggesting that the higher a person's maximum heart rate, the lower the risk of heart disease.
3. “fbs” shows a very small negative correlation (-0.02), which indicates a weak influence on heart disease.
4. The correlation between "opst" and "dests" reaches as high as 0.61, which may bring about some hidden issues, potentially leading to multicollinearity problems and affecting the prediction results. This can be addressed by dimensionality reduction through PCA or by directly removing one of the variables.

Conclusion:

Among these data, “thal”, “nmvcf”, “eia”, and “opst” are the most significant features for predicting heart disease, especially “thal” (0.53). On the other hand, mhr (-0.42) has a significant negative correlation, indicating an inverse relationship between heart health and heart disease risk. “fbs” and “dests” could be considered for removal: “fbs” (0.02) shows a correlation that is too small, while “dests” has a high correlation with “opst “(0.61), which could potentially affect the model's prediction.

2.

Table:

Number	Name	Test Accuracy	Training Accuracy
1	Perceptron	69.91%	74.07%
2	Logistic Regression	84.26%	90.74

3	SVM	84.26%	92.59%
4	Decision Tree	92.59%	77.78%
5	Random Forest	100.00%	90.74%
6	KNN	82.87%	88.89%

Observation:

1. SVM performed excellently in this analysis, especially after using a linear kernel function, achieving an accuracy of 92.59%. This indicates that the model captures the linear features of the data well.
2. Random Forest and KNN also demonstrated similarly high accuracy, suggesting that these two non-linear models handle the complexity of this dataset effectively.
3. Logistic Regression also performed well for this classification problem, with accuracy close to 91%, showing the strong applicability of linear classification models on this dataset.
4. Perceptron and Decision Tree had relatively lower accuracy, indicating that they may not be as effective as the other models, especially when the data features are more complex or non-linear.

Conclusion:

Based on the accuracy comparison, SVM, Random Forest, and KNN are the best models for heart disease prediction, as they can capture important patterns in the data and provide accurate predictions. If choosing between these three, SVM is the most recommended model, especially when the data is linearly separable. The high accuracy of these models indicates that they are suitable as a foundation for future heart disease prediction models.