**Heart Disease Prediction Analysis Report**

**Introduction**

This report summarizes the analysis performed on the heart disease dataset obtained from the UCI Machine Learning Repository. The primary objective of this analysis was to predict the presence and severity of heart disease in patients based on a set of 13 clinical features. The dataset includes information such as age, sex, chest pain type, resting blood pressure, cholesterol levels, and other diagnostic measurements. The target variable, **num**, represents the diagnosis of heart disease, ranging from 0 (no presence) to 4 (severe presence).

**Data Preprocessing**

The initial dataset contained 303 instances and 13 features. The following preprocessing steps were undertaken to prepare the data for modeling:

1. **Handling Missing Values**: The analysis identified missing values in the **ca** (number of major vessels colored by fluoroscopy) and **thal** (thalassemia) features. Although the specific imputation method wasn't detailed in the provided notebook conversion, handling these missing values is crucial for model training. *(Note: The notebook output shows missing values but doesn't explicitly show how they were handled before scaling/modeling. This is a potential gap.)*
2. **Feature Scaling**: Numerical features were scaled using both **StandardScaler** and **MinMaxScaler**. The **MinMaxScaler** scaled data was ultimately used for splitting and model training. Scaling ensures that features with larger value ranges do not disproportionately influence the model.
3. **Categorical Feature Encoding**: Categorical features such as **sex**, **cp**, **fbs**, **restecg**, **exang**, **slope**, and **thal** were converted into numerical format using one-hot encoding (specifically **pd.get\_dummies** with **drop\_first=True**). This allows machine learning algorithms to process these variables.

**Exploratory Data Analysis (EDA)**

EDA was performed to gain insights into the data distribution and relationships between features:

1. **Summary Statistics**: Descriptive statistics revealed the range, mean, and standard deviation for each feature, providing a basic understanding of the data's characteristics.
2. **Distribution Analysis**: Histograms were generated for numerical features to visualize their distributions. This helps in understanding the spread and skewness of the data.
3. **Correlation Analysis**: A correlation matrix heatmap was plotted to examine the linear relationships between features. This can help identify potential multicollinearity or features strongly correlated with the target.
4. **Pairwise Relationships**: A pairplot was generated to visualize pairwise relationships between features, although the specific plots were not visible in the converted markdown.

**Modelling**

After preprocessing and splitting the data into 80% training and 20% testing sets, three different classification models were trained to predict heart disease:

1. **Logistic Regression**: A linear model commonly used for binary and multi-class classification.
2. **Random Forest Classifier**: An ensemble method based on decision trees, known for its robustness and ability to handle complex interactions.
3. **Support Vector Classifier (SVC)**: A powerful model that finds an optimal hyperplane to separate different classes.

These models were fitted using the scaled training data (**X\_train**) and the corresponding target labels (**y\_train**).

**Results**

The performance of the trained models was evaluated on the unseen test set (**X\_test**, **y\_test**) using accuracy and classification reports:

* **Accuracy**:
  + Logistic Regression: 0.54
  + Random Forest: 0.49
  + Support Vector Classifier (SVC): 0.54
* **Classification Reports**:
  + **Logistic Regression**: Showed high precision and recall for class 0 (no heart disease), but significantly lower performance for classes 1-4, particularly for classes 3 and 4 where precision and recall were 0.
  + **Random Forest**: Also performed well on class 0 but had very poor precision and recall for classes 1-4.
  + **SVC**: Similar to Logistic Regression, performed well on class 0 but struggled with classes 1-4, showing slightly better recall for class 3 compared to Logistic Regression but still poor overall.

**Conclusion**

Based on the evaluation metrics, both **Logistic Regression** and **Support Vector Classifier (SVC)** achieved the highest overall accuracy (54%) on the test set. However, the classification reports highlight a significant challenge: all models performed poorly in distinguishing between the different levels of heart disease presence (classes 1-4). They were reasonably effective at identifying the absence of heart disease (class 0) but struggled with the minority classes representing disease presence.

This suggests potential issues such as class imbalance or the need for more sophisticated modeling techniques or feature engineering. While Logistic Regression or SVM might be considered the 'best' based purely on accuracy, their practical utility is limited due to the poor performance on positive classes. Further investigation, including addressing class imbalance (e.g., using SMOTE) and potentially exploring different algorithms or hyperparameter tuning, would be necessary to develop a more reliable model for predicting heart disease presence and severity.