**Introduction:**

The aim of this AI project is to analyze a heart disease dataset and develop a predictive model to identify individuals at risk of heart disease. The dataset contains various features related to demographic information, medical history, and diagnostic test results. In this project, we will perform data preprocessing, feature selection, data analysis, and modeling using several algorithms, including logistic regression, SVM, decision tree, and random forest.

The preprocessing phase involves handling missing values by utilizing a random forest model and replacing categorical nulls with the mode. We will also perform label encoding and remove outliers to ensure the quality and integrity of the data. Feature selection techniques such as VarianceThreshold, SelectKBest with chi-squared test, Lasso regularization, SelectKBest with mutual information, and SelectKBest with ANOVA F-test will be employed to identify the most informative features for our predictive model.

During the data analysis phase, we will draw conclusions from the dataset. We will explore numeric columns, observe the differences between positive and negative heart disease cases, analyze the standard deviations of heart disease columns, investigate the relationship between age and heart disease, examine blood pressure values, compare heart disease occurrences between genders, and assess the association between chest pain type, age, and the likelihood of heart disease.

To build our predictive model, we will implement logistic regression, SVM, decision tree, and random forest algorithms. The performance of each model will be evaluated based on accuracy, with logistic regression achieving 85% accuracy, SVM achieving 84% accuracy, decision tree achieving 76% accuracy, and random forest achieving 81% accuracy. Finally, we will employ hyperparameter tuning techniques to optimize the models and select the best-performing model.

**1-Preprocessing:**

Checking Null Values: Perform an initial check for missing values in the dataset.

Replacing Null Values with Random Forest Model: Use a random forest model to predict and replace missing values in the dataset.

Replacing Categorical Null Values with Mode: For categorical features, replace missing values with the mode of the respective feature.

Label Encoding: Convert categorical features into numerical representations using label encoding.

Removing Outliers: Identify and remove outliers from the dataset using appropriate techniques such as Z-score or interquartile range (IQR).

**2-Feature Selection:**

VarianceThreshold: Remove features with low variance, assuming they contain little information for the classification task.

SelectKBest with Chi-squared Test: Select the top K features based on the chi-squared test, which measures the dependence between categorical variables.

Lasso Regularization: Use L1 regularization to select features based on their coefficients obtained from a logistic regression model.

SelectKBest with Mutual Information: Select the top K features based on their mutual information scores, which capture the statistical dependency between variables.

SelectKBest with ANOVA F-test: Select the top K features using the ANOVA F-test, which measures the difference in means between groups.

**3-Data Analysis:**

Numeric Columns: Note that the standard deviation of the "age" column is low, indicating limited variability. Additionally, consider removing or replacing values for children in the "work type" column.

Negative vs Positive Heart Disease Values: Positive heart disease values tend to have higher means or medians in most positive heart disease columns, except for the "Max Heart rate." Investigate this further during bivariate analysis.

Standard Deviation of Negative Heart Disease Columns: The standard deviation of negative heart disease columns is higher in some cases compared to positive columns. This suggests that positive heart disease values occupy lower ranges in most features and may be predictable.

Minimum Age of Negative and Positive Heart Disease: Observe that the minimum age for negative heart disease is 29, while for positive heart disease, it is 35. Determine if heart disease is more common for high ages or if these values are outliers.

BP Values: Note that the value ranges for blood pressure (BP) are similar in both positive and negative heart disease cases.

Male vs. Female: Male and female values are generally similar across many features, except for heart disease. Approximately 54% of females have heart disease compared to 23% of males.

Chest Pain Type with Age Ranges for Heart Disease: There is a clear association between chest pain type, age, and the likelihood of heart disease. As the chest pain type increases, particularly for individuals older than 55, the probability of having heart disease also increases significantly.

4-Modeling using Logistic Regression, SVM, Decision Tree, and Random Forest:

Perform classification using logistic regression, SVM, decision tree, and random forest algorithms.

Evaluate the models based on accuracy, with logistic regression achieving 85% accuracy, SVM achieving 84% accuracy, decision tree achieving 76% accuracy, and random forest achieving 81% accuracy.

**5-Hyperparameter Tuning:**

Fine-tune the models by optimizing their hyperparameters using techniques like grid search or random search.

Adjust hyperparameters such as regularization strength, kernel type, maximum tree depth, and the number of estimators.

Evaluate the tuned models and select the best-performing model based on accuracy or other relevant metrics.

**Conclusion:**

In conclusion, this AI project focused on analyzing a heart disease dataset and developing a predictive model to identify individuals at risk of heart disease. Through thorough preprocessing techniques, we ensured the quality and completeness of the dataset by handling missing values, performing label encoding, and removing outliers. Feature selection methods allowed us to identify the most relevant features for our predictive model, enhancing its accuracy and interpretability.

During the data analysis phase, we drew significant insights from the dataset. Notably, we observed that positive heart disease values tended to have higher mean or median values in most positive heart disease columns. The standard deviation of negative heart disease columns was higher in certain cases, indicating that positive heart disease cases occupied lower ranges and were potentially predictable. We also examined the impact of age, gender, and chest pain type on the likelihood of heart disease, providing valuable information for further medical investigation.

By employing logistic regression, SVM, decision tree, and random forest algorithms, we achieved promising accuracies for our predictive models. Logistic regression and SVM demonstrated the highest accuracies of 85% and 84%, respectively, while decision tree and random forest achieved accuracies of 76% and 81%. Hyperparameter tuning further optimized the models, ensuring their optimal performance.

This project's findings and predictive models can assist healthcare professionals in identifying individuals at risk of heart disease. By leveraging machine learning techniques, we can provide valuable insights to support early detection, prevention, and personalized treatment strategies for heart disease patients. Future work may involve exploring additional algorithms, incorporating more comprehensive datasets, and collaborating with domain experts to enhance the accuracy and clinical utility of our predictive models.