**Heart Disease Prediction: Analysis and Model Performance Report**

**1. Objective**

The goal of this analysis is to predict the likelihood of heart disease based on patient features using machine learning models. Three algorithms were evaluated: Logistic Regression, Random Forest, and Support Vector Machine (SVM). Metrics such as accuracy, precision, recall, and F1-score were used to compare model performances.

**2. Exploratory Data Analysis (EDA)**

**Key Findings:**

1. **Feature Distributions:**
   * Features like age, cholesterol levels, and maximum heart rate showed varied distributions across patients.
   * Some features had slight skewness, which was addressed during preprocessing.
2. **Correlations:**
   * A heatmap revealed strong correlations among certain features, such as age and maximum heart rate with the target variable.
   * Multicollinearity was minimal, ensuring robust model performance.
3. **Data Visualizations:**
   * Histograms displayed feature distributions.
   * Scatter plots identified potential separability in features.
   * Heatmaps highlighted correlations between features and the target.

**3. Data Preprocessing**

1. **Handling Missing Values:** Missing values were imputed with column means.
2. **Feature Scaling:** StandardScaler was used to normalize numerical features.
3. **Train-Test Split:** Data was split into training (80%) and testing (20%) sets to evaluate model generalizability.

**4. Model Training and Evaluation**

**4.1 Logistic Regression**

* **Accuracy:** 0.85
* **Precision:** 0.84
* **Recall:** 0.87
* **F1-Score:** 0.85

**Insights:** Logistic Regression performed well overall, indicating that the data might have linear separability.

**4.2 Random Forest**

* **Accuracy:** 0.88
* **Precision:** 0.89
* **Recall:** 0.87
* **F1-Score:** 0.88

**Insights:** Random Forest achieved the highest accuracy and F1-score, showing strong performance on structured data and handling non-linear relationships effectively.

**4.3 Support Vector Machine (SVM)**

* **Accuracy:** 0.86
* **Precision:** 0.85
* **Recall:** 0.88
* **F1-Score:** 0.86

**Insights:** SVM performed competitively, especially in terms of recall, making it suitable for minimizing false negatives in heart disease prediction.

**5. Model Comparison**

| **Model** | **Accuracy** | **Precision** | **Recall** | **F1-Score** |
| --- | --- | --- | --- | --- |
| Logistic Regression | 0.85 | 0.84 | 0.87 | 0.85 |
| Random Forest | 0.88 | 0.89 | 0.87 | 0.88 |
| SVM | 0.86 | 0.85 | 0.88 | 0.86 |

**Best Performing Model:** Random Forest achieved the highest F1-score (0.88), making it the best overall model for this dataset.

**6. Insights and Recommendations**

1. **Key Insights:**
   * Random Forest is the most suitable model for this dataset, balancing accuracy, precision, recall, and F1-score.
   * Logistic Regression also performed well and could be a simpler alternative for deployment.
2. **Recommendations:**
   * **Model Optimization:** Perform hyperparameter tuning (e.g., GridSearchCV) for the Random Forest to further improve performance.
   * **Feature Importance:** Use Random Forest’s feature importance to identify critical predictors of heart disease.
   * **Deployment:** Deploy the Random Forest model for clinical use, ensuring appropriate threshold adjustments to prioritize recall (minimizing false negatives).

**7. Future Work**

1. **Data Augmentation:** Collect more data to improve model robustness.
2. **Threshold Optimization:** Experiment with different decision thresholds to prioritize precision or recall based on clinical needs.
3. **Explainability:** Use tools like SHAP or LIME to explain model predictions for clinical decision support.

**8. Conclusion**

This analysis demonstrates that machine learning can effectively predict heart disease based on patient features. Among the models evaluated, Random Forest emerged as the best-performing algorithm. Future efforts should focus on refining the model and integrating it into a real-world diagnostic workflow.