Machine Learning Project Report

**1. Name of the Data**

Heart Disease Dataset

**2. Source of the Data**

The data is sourced from the UCI Machine Learning Repository.

**3. Link to the Original Data**

[**https://github.com/sharmaroshan/Heart-UCI-Dataset/blob/master/heart.csv**](https://github.com/sharmaroshan/Heart-UCI-Dataset/blob/master/heart.csv)

**4. Explain the Data in Words**

This dataset contains attributes related to heart health such as age, sex, chest pain type, blood pressure, cholesterol level, and others. The goal is to predict whether a patient is likely to have heart disease.

**5. Type of Problem**

Classification Problem

**6. Number of Attributes**

13 attributes (excluding the target variable).

**7. Number of Samples**

303 samples.

**8. Properties of the Data (Statistics)**

- Mean, Median, Min, Max, Standard Deviation calculated for numerical features such as age, cholesterol, blood pressure, etc.

**9. Missing Data**

No missing values were found in the dataset. If missing values were present, imputation with mean or median would be considered.

**10. Visualization of the Data**

Various visualizations such as histograms, boxplots, and correlation heatmaps were created using matplotlib and seaborn libraries to understand the distribution and relationships between features.

**11. Normalization or Standardization**

Data was standardized using StandardScaler to ensure all features contribute equally to the model performance.

**12. Preprocessing Applied**

- Label encoding for categorical variables.  
- Standardization for numerical features.  
- Splitting dataset into training and testing sets.  
- Saving preprocessed data into separate folders.

**13. Train and Test Split**

80% training data and 20% testing data.

**14. Machine Learning Models and Results**

- Logistic Regression  
- Support Vector Machine (SVM)  
- Random Forest Classifier

Best Performing Model: Random Forest Classifier  
Worst Performing Model: Logistic Regression

**15. Accuracy Table**

|  |  |
| --- | --- |
| Model | Accuracy |
| Logistic Regression | 82% |
| SVM | 85% |
| Random Forest | 90% |

**16. Bonus Visualization**

Advanced plots using seaborn library were created, including pairplots, heatmaps, and classification report visualizations.

**17. Reason for Choosing the Data, Importance of the Data, and Model Insights**

I selected the Heart Disease dataset because cardiovascular diseases are among the leading causes of death globally. Early detection can save lives by enabling preventive care. This dataset provides insight into critical factors influencing heart health. The Random Forest model, being the best-performing model, is important because it handles imbalanced and complex datasets effectively, offering high accuracy. Insights from the model showed that features like chest pain type, maximum heart rate, and exercise-induced angina are strong indicators of heart disease risk. Therefore, focusing on these factors in real medical screenings can improve patient outcomes significantly.

**18. Link to Code and Data**

GitHub Repository: <https://github.com/Shahd-B-2002/project-shahad/tree/main/project_heart>