📑 Heart Disease Analysis –Project Report

🔹 Introduction

The main objective of this project is to **predict heart disease** and analyze the dataset features to identify the most significant risk factors. Various techniques such as Exploratory Data Analysis (EDA), feature selection, and machine learning models have been applied to understand the data and build predictive models.

🔹 Dataset Information

The dataset contains several patient-related medical attributes that help in predicting the presence of heart disease. The features are:

1. **Age** – Patient's age (higher age increases risk).
2. **Sex** – Gender (male/female).
3. **Chest Pain Type (cp)** – Different categories of chest pain; strongly correlated with heart disease.
4. **Resting Blood Pressure (trestbps)** – High BP is a critical risk factor.
5. **Serum Cholesterol (chol)** – High cholesterol indicates possible blockages.
6. **Fasting Blood Sugar (fbs)** – Shows diabetes correlation with heart disease.
7. **Resting ECG (restecg)** – Electrocardiographic results (normal or abnormal).
8. **Maximum Heart Rate Achieved (thalach)** – Peak heart rate during exercise.

9. **Exercise Induced Angina (exang)** – Whether chest pain occurs during exercise.
10. **Oldpeak** – ST depression (indicator of abnormal ECG).
11. **Slope** – Slope of the ST segment.
12. **Number of Major Vessels (ca)** – Count of major blood vessels blocked.
13. **Thal** – Thalassemia test result.
14. **Target** – 1 = Heart disease present, 0 = No heart disease.

🔹 Exploratory Data Analysis (EDA)

- Checked for missing values, duplicates, and outliers.
- Visualized data distributions using **histograms and boxplots**.
- Created a **correlation heatmap** which showed that chest pain, cholesterol, maximum heart rate, and blood pressure are strongly related to heart disease.
- Used **countplots and pie charts** to analyze categorical variables such as sex, chest pain type, and target.

🔹 Preprocessing & Feature Engineering

- **Data Cleaning** – Handled missing values and encoded categorical features.
- **Feature Scaling** – Normalized continuous features (age, trestbps, chol, thalach, oldpeak).
- **Feature Selection** – Selected the most influential features based on correlation and importance scores.

🔹 Machine Learning Models Used

1. **Logistic Regression** – A baseline classification model for predicting presence/absence of disease.
2. **Random Forest Classifier** – An ensemble method using multiple decision trees for higher accuracy.
3. **Support Vector Machine (SVM)** – Classifies patients by finding an optimal decision boundary.
4. **K-Nearest Neighbors (KNN)** – Predicts based on similarity with neighboring data points.
5. **Naive Bayes** – A probability-based classifier.

🔹 Model Evaluation

- Models were evaluated using **accuracy, precision, recall, F1-score, and confusion matrix**.
- **Random Forest and Logistic Regression** showed the best performance.
- Achieved accuracy in the range of **80–85%**, which is quite good for this dataset.

🔹 Results & Conclusion

- **Key Findings**:
    - Age, chest pain type, cholesterol, blood pressure, and thalach are the most important predictors.
    - Male patients were found to have a higher risk of heart disease.
    - Exercise induced angina and oldpeak are also strong indicators.

#..Conclusion:

This project demonstrates that **machine learning models can effectively predict the risk of heart disease** when provided with the correct medical attributes. Such models can be highly useful in hospitals and healthcare systems for **early diagnosis and prevention** of heart-related issues.