Project Report: Heart Disease Prediction

1. Project Objective

The primary objective of this project was to build a predictive system that can accurately determine if a patient has heart disease based on a set of 11 clinical features. The goal involved:

1. Exploring the provided patient dataset to understand its features.
2. Developing an approach to preprocess the data for machine learning.
3. Building and evaluating several predictive models to find the most accurate system.

2. The Dataset

The project used `dataset.csv`, which contains 1,190 patient records and 12 columns.

- **Target Variable:** `target` (0 = No Heart Disease, 1 = Heart Disease)
- **Key Features:** The models were trained on 11 features, including:
  - `age`: Age of the patient
  - `sex`: (0 = female, 1 = male)
  - `chest pain type`: (1-4)
  - `resting bp s`: Resting blood pressure
  - `cholesterol`: Serum cholesterol
  - `max heart rate`: Maximum heart rate achieved
  - `exercise angina`: Exercise-induced angina
  - `oldpeak`: ST depression induced by exercise
  - `ST slope`: Slope of the peak exercise ST segment

3. Methodology

The project was executed in three main phases:

Phase 1: Exploratory Data Analysis (EDA)

We first explored the data to uncover patterns. This involved:

- Checking the balance of the `target` variable to see if our dataset was skewed (it was fairly balanced).

- Visualizing the distributions of key features like `age`, `cholesterol`, and `max heart rate` for patients with and without heart disease.
- Plotting how categorical features like `sex` and `chest pain type` correlate with a heart disease diagnosis.
- Creating a correlation heatmap to understand how all numerical features relate to each other and to the `target`.

Phase 2: Data Preprocessing

To prepare the data for machine learning, we:

1. **Converted Categorical Data:** Used one-hot encoding (via `pd.get_dummies`) to convert categorical columns (like `chest pain type`) into a numerical format that models can understand.
2. **Split the Data:** Divided the dataset into a training set (80%) to teach the models and a testing set (20%) to evaluate their performance.
3. **Feature Scaling:** Applied `StandardScaler` to all numerical features. This normalizes the data, ensuring that features with larger ranges (like `cholesterol`) do not unfairly dominate features with smaller ranges (like `oldpeak`).

Phase 3: Model Building and Evaluation

We trained and evaluated three different classification models to find the most effective "system" for prediction.

1. **Logistic Regression:** A solid, reliable baseline model for binary classification.
2. **Random Forest Classifier:** A powerful ensemble model that builds multiple decision trees and combines their votes.
3. **XGBoost Classifier:** A highly advanced and often top-performing gradient-boosting model.

Each model was trained on the `X_train` data and then tested on the unseen `X_test` data. We measured their performance using accuracy, precision, recall, and a confusion matrix.

4. Results and Conclusion

The models' performance on the unseen test data was as follows:

Model	Accuracy

Random Forest	**94.12%**
XGBoost	92.44%
Logistic Regression	86.13%

Conclusion

The **Random Forest** model provided the highest accuracy at **94.12%**, making it the most effective "system" for predicting heart disease from this dataset.

A feature importance analysis from the Random Forest model also revealed that `ST slope`, `chest pain type`, `max heart rate`, and `oldpeak` were among the most influential factors in its predictions.

This project successfully met its objective by exploring the data, establishing a preprocessing pipeline, and building and evaluating multiple models to find a high-accuracy solution for detecting heart disease.