Title: Heart Disease Risk Prediction

1. Introduction

This project implements a Heart Disease Prediction System using machine learning. It integrates a trained predictive model with a web-based interface for clinical decision support. The system automatically predicts the risk level of heart disease based on patient features, allowing healthcare providers to make informed decisions.

2. Tools and Libraries

The following Python libraries are used throughout the notebook:

- Pandas : data loading and manipulation
- NumPy : numerical computations
- Matplotlib & Seaborn : data visualization
- Scikit-learn :machine learning models, preprocessing, and evaluation metrics

These tools collectively support efficient data analysis and predictive modeling.

3. Dataset Description

The dataset contains clinical and demographic attributes commonly associated with heart disease diagnosis. Typical features include:

- Age
- Sex
- Chest pain type
- Resting blood pressure
- Cholesterol level
- Fasting blood sugar
- Maximum heart rate achieved

- Exercise-induced angina
- Other ECG-related and clinical indicators
- The target variable represents the presence or absence of heart disease.

---

4. Data Loading and Initial Exploration

The dataset is loaded into a Pandas DataFrame. Initial exploration includes:

- Displaying the first few records
- Checking the dataset shape
- Inspecting data types
- Generating descriptive statistics using describe()

This step helps to understand the structure, scale, and basic distribution of the data.

---

5. Exploratory Data Analysis (EDA)

EDA is performed to uncover patterns, trends, and potential issues in the dataset.

5.1 Missing Values and Class Balance

- The notebook checks for missing values to ensure data quality.
- The target variable distribution is analyzed to verify whether the classes are balanced or imbalanced.

5.2 Visualizations

Several plots are used to better understand the data:

- Histograms to analyze feature distributions
- Count plots to observe categorical variable frequencies
- Correlation heatmap to examine relationships between numerical features

The correlation heatmap is particularly useful in identifying strongly related variables that may influence heart disease prediction.

6. Data Preprocessing

Before model training, preprocessing steps are applied:

- Feature selection
- Encoding categorical variables (if applicable)
- Splitting the dataset into training and testing sets
- Feature scaling using standardization techniques

These steps ensure that the machine learning models perform optimally.

7. Machine Learning Models

The notebook trains and evaluates one or more classification models using Scikit-learn. Commonly applied models in this type of analysis include:

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

Each model is trained on the training dataset and evaluated using the test dataset.

---

8. Model Evaluation

Model performance was evaluated using cross-validation accuracy, training accuracy, and test accuracy. A comparison of multiple machine learning algorithms was conducted.

8.1 Model Comparison Results

The table below summarizes the performance of all trained models:

8.2 Best Model Performance

	Model	CV Accuracy	Training Accuracy	Test Accuracy	Overfitting Gap	Fit Category
0	Random Forest	0.9995	1.0000	1.000	0.0000	Best Fit
1	SVM	0.9985	1.0000	0.998	0.0020	Best Fit
2	KNN	0.9960	0.9975	0.997	0.0005	Best Fit
3	Gradient Boosting	0.9992	1.0000	0.999	0.0010	Best Fit
4	MLP (ANN)	0.9978	1.0000	0.998	0.0020	Best Fit

The Random Forest Classifier was selected as the best-performing model based on the highest test accuracy.

- Best Model: Random Forest
- Test Accuracy: 100.000%

The extremely high accuracy indicates that the model is highly effective at distinguishing between patients with and without heart disease. The low overfitting score suggests good generalization performance on unseen data.

9. Results and Discussion

The results indicate that machine learning can effectively predict heart disease risk based on clinical attributes. Models with higher accuracy and balanced precision–recall values are considered more reliable for practical use.

Correlation analysis and EDA results also highlight key health indicators that strongly influence heart disease prediction.

10. Conclusion

The Heart Disease Prediction System successfully integrates ML models with a web interface.

Provides fast, accurate risk assessment based on patient features.

Can support clinical decision-making and improve early detection of heart disease.

11. Future Work

Possible improvements include:

- Hyperparameter tuning for better model performance
- Using cross-validation
- Testing advanced models such as XGBoost or Neural Networks
- Deploying the model as a web or mobile application