SYNOPSIS

TITLE – HEARTH DISEASE

SATISH M – AF0378136

ABOUT THE TOPIC (DATASET):

Data Structure

The dataset should be organized in a tabular format, where each row represents a patient record on a specific date. The columns might include:

1. Date: The date of the record.
2. Patient ID: Unique identifier for the patient.
3. Age: Age of the patient.
4. Gender: Gender of the patient.
5. Chest Pain Type: Type of chest pain (e.g., typical angina, atypical angina, non-anginal pain, asymptomatic).
6. Resting Blood Pressure: Resting blood pressure (in mm Hg).
7. Cholesterol Level: Serum cholesterol level (in mg/dl).
8. Fasting Blood Sugar: Fasting blood sugar > 120 mg/dl (1 = true; 0 = false).
9. Resting ECG Results: Resting electrocardiographic results (e.g., normal, ST-T wave abnormality).
10. Max Heart Rate Achieved: Maximum heart rate achieved.
11. Exercise-Induced Angina: Exercise-induced angina (1 = yes; 0 = no).
12. Oldpeak: ST depression induced by exercise relative to rest.
13. Slope: The slope of the peak exercise ST segment.
14. Number of Major Vessels: Number of major vessels (0-3) colored by fluoroscopy.
15. Thalassemia: Thalassemia (e.g., normal, fixed defect, reversible defect).
16. Diagnosis of Heart Disease: Presence of heart disease (1 = yes; 0 = no).

Data Preprocessing

1. Data Cleaning: Handle missing values and ensure data consistency.

2. Normalization: Normalize numeric features if required (e.g., age, cholesterol levels).
3. Encoding Categorical Variables: Encode categorical variables (e.g., chest pain type, thalassemia) using one-hot encoding or label encoding.

Analysis Techniques

1. Time Series Analysis
- Trend Analysis: Identify trends in heart disease diagnosis rates over time.
- Seasonality Detection: Detect seasonal patterns in heart disease occurrences.
- Anomaly Detection: Identify anomalies in the data, which could indicate unusual spikes or drops in heart disease cases.

2. Predictive Modelling
- Classification Models: Use logistic regression, decision trees, or more advanced models like Random Forest, Gradient Boosting Machines, or neural networks to predict the likelihood of heart disease based on patient attributes.
- Survival Analysis: Analyze the time until a heart disease event occurs, using techniques like Kaplan-Meier estimation or Cox proportional hazards models.

3. Correlation Analysis
- Risk Factors: Explore correlations between various risk factors (e.g., cholesterol level, blood pressure) and the presence of heart disease.
- Demographic Analysis: Analyze correlations between demographic variables (e.g., age, gender) and heart disease incidence.

4. Cluster Analysis
- Patient Segmentation: Use clustering algorithms (e.g., K-means, hierarchical clustering) to segment patients into groups based on similarities in their attributes and risk factors.
- Risk Profiling: Identify common characteristics of high-risk groups.

Implementation Steps

1. Data Ingestion: Load the dataset into a data analysis environment (e.g., Python, R).
2. Preprocessing: Clean and prepare the data for analysis.
3. Exploratory Data Analysis (EDA): Conduct EDA to understand data distribution and initial patterns.
4. Modeling: Develop and validate models for classification, time series analysis, and clustering.
5. Visualization: Create visualizations to communicate insights effectively (e.g., correlation heatmaps, survival curves, cluster visualizations).
6. Reporting: Summarize findings in reports or dashboards for stakeholders.

Data set: Heart Disease Dataset (kaggle.com)

Technologies: pandas, Microsoft Excel, Microsoft PowerBi, seaborn, matplotlib.

Software Requirements:

Operating System – Windows, Linux and mac

IDLE – Jupyter Notebook

Hardware Requirements:

RAM – Minimum 4GB

Processor – Minimum intel i3