**PROJECT 1 - HEART DISEASE**

**MEMBER – Krishan Pandey, Hardeep Gumber, Anvita Iyer, Aleksanda Kutz**

**1. Exploratory Data Analysis (EDA)**

* **Objective**: Understand the dataset, identify patterns, and detect anomalies or missing data.
* **Methods**:
  + **Summary Statistics**: Measures like mean, median, mode, standard deviation, and quartiles can provide an overview of key variables (e.g., age, cholesterol levels, resting blood pressure).
  + **Visualization**: Use histograms, bar plots, box plots, scatter plots, and correlation matrices to visually explore relationships and distributions.
    - **Examples**:
      * Distribution of cholesterol levels across age groups.
      * Correlation between maximum heart rate and age.
  + **Correlation Analysis**: Identify relationships between continuous variables using Pearson or Spearman correlation coefficients.
    - **Example**: Exploring the correlation between cholesterol levels and the likelihood of exercise-induced angina.

**2. Statistical Hypothesis Testing**

* **Objective**: Test predefined hypotheses about the data to determine if observed relationships are statistically significant.
* **Methods**:
  + **T-tests**: Compare the means of two groups (e.g., compare the resting heart rates of males vs. females).
  + **ANOVA (Analysis of Variance)**: Compare the means across multiple groups (e.g., comparing cholesterol levels across different chest pain types).
  + **Chi-square Test**: Test for associations between categorical variables (e.g., testing the association between sex and the presence of exercise-induced angina).
  + **Logistic Regression**: Test whether a set of predictors (e.g., age, cholesterol, chest pain type) predict a binary outcome (e.g., presence of heart disease).

**3. Predictive Modeling**

* **Objective**: Build models to predict the likelihood of heart disease based on key features in the dataset.
* **Methods**:
  + **Logistic Regression**: A go-to method for binary classification tasks, such as predicting heart disease risk (yes/no).
  + **Random Forest**: An ensemble method that uses multiple decision trees to improve prediction accuracy. It is useful for both classification and regression tasks and can handle complex interactions between variables.
  + **Support Vector Machines (SVM)**: Effective for classification tasks, such as predicting if a patient will have heart disease based on multiple health indicators.
  + **K-Nearest Neighbors (KNN)**: A simple classification algorithm that assigns a class (heart disease or not) based on the most common outcome among similar data points.
  + **Gradient Boosting (e.g., XGBoost)**: Another ensemble method that often provides high accuracy by combining weak learners into a strong learner. Useful for classification and regression.

**4. Feature Selection and Dimensionality Reduction**

* **Objective**: Improve model performance by selecting the most important features and reducing noise in the data.
* **Methods**:
  + **Principal Component Analysis (PCA)**: Reduce the dimensionality of the dataset while retaining most of the variance. It can be used to visualize the data or improve model performance.
  + **Recursive Feature Elimination (RFE)**: A technique to select the most important features by recursively eliminating the least important ones.
  + **Feature Importance from Models**: Models like Random Forest or Gradient Boosting provide inherent feature importance metrics, helping you identify which features have the strongest impact on heart disease prediction.

**5. Time Series Analysis (if applicable)**

* **Objective**: Analyze trends over time (e.g., changes in heart rate, blood pressure, or cholesterol over repeated visits).
* **Methods**:
  + **Autoregressive Integrated Moving Average (ARIMA)**: Used for forecasting trends based on time-series data. It could be used to predict a patient’s health trajectory over time.
  + **Seasonal Decomposition of Time Series (STL)**: Can help identify patterns like seasonality in heart disease events (e.g., increased heart issues during certain months).

**6. Clustering Techniques**

* **Objective**: Identify natural groupings or patterns in the data that may not be immediately visible.
* **Methods**:
  + **K-Means Clustering**: Group similar patients together based on shared characteristics (e.g., cholesterol, age, and resting heart rate). This can help in identifying patient subgroups for targeted treatments.
  + **Hierarchical Clustering**: Create a hierarchy of clusters based on the data, which is useful for finding relationships between variables or patients.

**7. Survival Analysis (if longitudinal data is available)**

* **Objective**: Estimate the time until an event occurs (e.g., heart attack, hospitalization).
* **Methods**:
  + **Kaplan-Meier Estimate**: Estimate the survival function and plot survival curves to compare different patient groups (e.g., patients with and without angina).
  + **Cox Proportional Hazards Model**: Assess the effect of various risk factors (e.g., age, cholesterol) on survival time.

**8. Deep Learning (Advanced)**

* **Objective**: Use more complex models if you have large datasets with nonlinear relationships.
* **Methods**:
  + **Artificial Neural Networks (ANNs)**: Can be used to build more complex models for heart disease prediction, though they generally require more data and computational resources.
  + **Recurrent Neural Networks (RNNs)**: Useful for time-series analysis where sequential data (e.g., patient health records over time) is being used.

**Suggested Workflow for Analysis**

1. **Data Cleaning and Preprocessing**:
   * Remove or handle missing values.
   * Encode categorical variables (e.g., chest pain type, thal, sex).
   * Scale continuous variables (e.g., cholesterol, resting blood pressure) if necessary for modeling.
2. **Exploratory Data Analysis (EDA)**:
   * Generate summary statistics and visualizations.
   * Check for correlations between variables (e.g., age, cholesterol, and resting blood pressure).
   * Identify potential outliers or skewed data.
3. **Hypothesis Testing**:
   * Test the hypotheses developed earlier using appropriate statistical tests.
   * Example: Use a chi-square test to see if high fasting blood sugar (>120 mg/dl) is significantly associated with exercise-induced angina.
4. **Predictive Modeling**:
   * Build and evaluate predictive models (e.g., logistic regression, random forest) to classify heart disease risk.
   * Use cross-validation to tune hyperparameters and assess the model's generalization ability.
5. **Model Interpretation**:
   * Analyze feature importance or coefficients from your models to understand which factors contribute most to heart disease.
   * Example: Check which variables (age, chest pain type, cholesterol) most influence the model’s predictions.
6. **Visualization**:
   * Create 6–8 visualizations (as required) to clearly communicate findings (e.g., bar plots, correlation heatmaps, box plots, and decision tree diagrams).
   * Label each visualization clearly and provide context in the presentation.
7. **Conclusion and Reporting**:
   * Summarize key findings and support them with both numerical results and visualizations.
   * Provide recommendations based on the analysis (e.g., individuals with high cholesterol and blood pressure should be targeted for early interventions).
8. **Prepare Presentation**:
   * Ensure the final slide deck contains visualizations and explains the findings in a clear, professional manner.
   * Practice the presentation, making sure each team member is familiar with their part.

**Tools to Use**

* **Python Libraries**:
  + **Pandas**: For data manipulation and analysis.
  + **Matplotlib** and **Seaborn**: For visualizations.
  + **Scikit-learn**: For machine learning models.
  + **Statsmodels**: For statistical hypothesis testing.
  + **SciPy**: For statistical tests like t-tests and ANOVA.
* **GitHub**:
  + Track progress using **GitHub Projects**.
  + Store code, notebooks, and documentation in a well-organized repository.