AI/ML Practical Interview Task: Predicting Patient Risk of Heart Disease

Objective:
The candidate will develop a machine learning model to predict the likelihood of a patient having heart disease based on various health-related features. The task will involve **data preprocessing**, **exploratory data analysis (EDA)**, **model selection**, and **evaluation** using the provided dataset.

Task Overview:

You are provided with a dataset (heart_disease_data.csv) containing information about patients' health conditions, such as age, cholesterol levels, blood pressure, exercise habits, etc. The goal is to build a model that predicts whether a patient is at high risk of heart disease based on these features.

Dataset Information:

The dataset contains the following columns:

- Age: Age of the patient (in years)
- Gender: Male or Female
- Cholesterol: Total cholesterol level (in mg/dL)
- BloodPressure: Systolic blood pressure (in mm Hg)
- MaxHeartRate: Maximum heart rate achieved during exercise (in bpm)
- Exercise: Whether the patient exercises regularly (Yes/No)
- Smoking: Whether the patient is a smoker (Yes/No)
- FamilyHistory: Whether the patient has a family history of heart disease (Yes/No)
- Risk: Target column (1 = High risk, 0 = Low risk of heart disease)

Steps to Perform:

1. Load and Explore the Dataset

- **Objective:** Familiarize yourself with the dataset and understand its structure.
  - Load the provided dataset (heart_disease_data.csv) into the working environment.

  - Display basic information about the dataset such as column names, data types, and the first few rows.
  - Check for any missing values and identify their distribution.
  - Investigate the class distribution of the target variable Risk (0 and 1).

2. Data Preprocessing

- **Objective:** Prepare the dataset for training by handling missing data and converting categorical variables into numeric format.
  - Handle missing values:
    - Impute or remove missing values depending on the feature type.
  - Convert categorical variables (Gender, Exercise, Smoking, Family History) into numerical format using **Label Encoding** or **One-Hot Encoding**.
  - Normalize numerical features (Cholesterol, Blood Pressure, MA Heartrate) to improve model performance.

3. Exploratory Data Analysis (EDA)

- **Objective:** Analyze the data to identify patterns or trends that can influence the prediction of heart disease risk.
  - Generate visualizations to explore relationships between features and the target variable:
    - **Bar charts**: Show the distribution of Risk across different categories such as Exercise, Family History, and Gender.
    - **Histograms**: Visualize the distribution of continuous features like Age, Cholesterol, Blood Pressure.
    - **Correlation heatmap**: Analyze the correlation between numeric features and the target variable.
  - Identify any potential patterns that might affect the prediction of heart disease (e.g., high cholesterol leading to higher risk).

4. Model Selection and Training

- **Objective:** Select and train a classification model to predict heart disease risk.
  - Split the data into **training (80%)** and **testing (20%)** sets.
  - Choose a classification algorithm to predict heart disease risk. You can choose one or more of the following:
    - **Logistic Regression**
    - **Random Forest**
    - **XGBoost**
    - **K-Nearest Neighbors (KNN)**

    - Train the model on the training dataset.
    - Use **cross-validation** to tune the hyperparameters and improve model performance.

5. Model Evaluation

- **Objective:** Evaluate the trained model using different metrics to assess its performance.
    - Test the model on the testing dataset and calculate the following:
        - **Accuracy**: The overall percentage of correct predictions.
        - **Precision, Recall, and F1-score**: Evaluate model performance for imbalanced classes (e.g., predicting minority class - high risk).
        - **Confusion Matrix**: Visualize the correct vs. incorrect predictions.
        - **ROC Curve and AUC**: Assess how well the model distinguishes between high and low-risk patients.

6. Model Deployment Using Prompt Engineering

- **Objective:** Generate a natural language interface for the trained model using prompt engineering.
    - Use **prompt engineering** to create queries that allow you to ask the model to predict heart disease risk based on patient details.
    - Example prompts:
        - “Given the patient details (Age: 60, Cholesterol: 240, Exercise: No, Smoking: Yes), predict the risk of heart disease.”
        - “Is a 45-year-old patient with high blood pressure and no family history at risk of heart disease?”
    - Test the system using a few queries and evaluate how well it provides risk predictions based on the dataset.

Deliverables:

1. **Dataset**: Use the provided heart_disease_data.csv for model training.
2. **Jupiter Notebook or Python Script**:
    a. Code for loading, cleaning, and preprocessing the data.
    b. Data visualizations and EDA results.
    c. Code for training and evaluating the machine learning model.
    d. Example prompt-based queries for predicting risk.

Evaluation Criteria:

- **Data Preprocessing:** Effective handling of missing data, encoding of categorical variables, and normalization of numerical features.
- **Feature Engineering and EDA:** Depth of data exploration and visualization of relationships between features and target variables.
- **Model Selection and Training:** The candidate's ability to choose the right model, train it, and optimize its performance.
- **Model Evaluation:** Evaluation of model performance using appropriate metrics (accuracy, precision, recall, F1-score).
- **Prompt Engineering:** Ability to generate meaningful queries to interact with the model using natural language.

I need the data and code for this task in a single ZIP file or in one organized folder. Along with that, I need a separate file that clearly explains the following:

1. What processes or steps you have completed in this task.
2. What steps I need to follow to run your code on my system (installation, libraries, commands, etc.).
3. If you have performed any visualizations, include a Word file explaining the visualizations and their purpose.

Please make sure everything is well-documented and easy to understand.