* **What processes or steps you have completed in this task.**

Here is a summary of the processes and steps we have completed in this task:

1. **Data Loading and Initial Exploration:**
   * Loaded the heart disease dataset.
   * Examined the basic structure and information about the data (e.g., using df.info()).
   * Checked for missing values in each column.
2. **Data Cleaning:**
   * Handled missing values in numerical columns ('Age', 'Cholesterol', 'BloodPressure', 'MaxHeartRate') using imputation (mean for Age and median for Cholesterol, and median within Gender groups for BloodPressure and MaxHeartRate).
   * Handled missing values in categorical columns ('Gender', 'Exercise', 'Smoking', 'FamilyHistory') using mode imputation.
3. **Data Preprocessing:**
   * Encoded categorical features ('Gender', 'Exercise', 'Smoking', 'FamilyHistory') into numerical format using Label Encoding.
   * Scaled numerical features ('Cholesterol', 'BloodPressure', 'MaxHeartRate') using StandardScaler.
4. **Exploratory Data Analysis (EDA):**
   * Visualized the distribution of categorical features using bar charts.
   * Visualized the distribution of numerical features using histograms and box plots.
   * Investigated the class distribution of the target variable 'Risk' and identified class imbalance.
   * Generated bar charts to explore the relationship between categorical features and 'Risk'.
   * Generated histograms to explore the relationship between continuous features and 'Risk'.
   * Calculated and visualized the correlation matrix for numerical features and 'Risk' using a heatmap.
   * Analyzed patterns and trends influencing heart disease risk based on the visualizations and correlations.
5. **Model Selection and Training:**
   * Split the data into training (80%) and testing (20%) sets.
   * Initialized and trained four classification models: Logistic Regression, Random Forest, XGBoost, and K-Nearest Neighbors.
   * Performed hyperparameter tuning with cross-validation for Logistic Regression, Random Forest, XGBoost, and K-Nearest Neighbors to improve model performance.
6. **Model Evaluation:**
   * Evaluated the trained models on the testing dataset using various metrics:
     + Accuracy
     + Precision, Recall, and F1-score
     + Confusion Matrix
     + ROC Curve and AUC
7. **Natural Language Interface (Simple Implementation):**
   * Created a basic natural language interface function (process\_simple\_query) to take a natural language prompt with patient details.
   * Implemented logic to extract patient details from the prompt using regular expressions.
   * Processed and formatted the extracted details for the trained model (including scaling and encoding).
   * Used the best-performing model (Random Forest) to make a prediction.
   * Generated a simple natural language explanation of the prediction.
   * Tested the interface with a few example queries.

We have covered the essential steps from data preparation and exploration to model building, evaluation, and creating a basic interface for interaction.