1. Data Exploration and Preprocessing

- **Data Understanding:**
  - Explored the dataset to gain a comprehensive understanding of its structure, columns, and potential issues.
- **Issues Identified in the Data:**
  - **Nationality Issues:** Noted that the **Nationality** column had biased data, particularly with a majority of the entries being **Saudi**. This could lead to model bias and needed careful treatment.
  - **Avg_HBA1C Results Problem:** Found that some records in the **Avg_HBA1C Results** column had a value of 0, which doesn't make sense. This was treated as `null` since a value of 0 for HBA1C is outside the expected range.
  - **Religion Data:** The **Religion** column contained a lot of non-meaningful and inconsistent data. Decided to exclude this column as it wouldn't add value and could lead to incorrect model behavior.
  - **Missing Age Information:** Although the dataset contained a **Date of Birth** column, it lacked an **Age** column. Therefore, you calculated **Age** using the date of birth.
  - **Unnecessary Columns:** Removed columns like **Diagnosis Type** (all entries were for Type 2) and **Patient Identifier** (not relevant to the analysis).
  - **Bias in Data:** The dataset was found to be biased, especially in terms of nationality (with a large percentage of Saudi patients). This bias needed to be addressed carefully when training the model.
  - **Highly Correlated Columns:** Identified that there were no highly correlated columns with the output column, requiring alternative feature engineering strategies to improve predictive performance.

2. Data Cleaning and Transformation

- **Handling Text Data:**
  - Binary encoded categorical text data (such as **Nationality** and **Gender**) to make it usable for machine learning models.
- **Handling Missing Values:**
  - For the **Avg_HBA1C Results** column, you used an **Iterative Imputer** to predict and replace the missing values. This was done because dropping rows with missing values would have resulted in a significant loss of data (~25%).
- **Outlier Removal:**
  - Applied **Interquartile Range (IQR)** method to detect and remove outliers in the dataset.
  - This step was essential for ensuring that the model's predictions were not skewed due to extreme values.
- **Data Scaling:**
  - Scaled the features to ensure they all fell within the same range, making it easier for machine learning models to learn effectively.
- **Handling Imbalanced Data:**

    - Since the target variable **Chronic_flag** was imbalanced (majority class being non-chronic), you used **SMOTE** (Synthetic Minority Over-sampling Technique) to increase the number of chronic patients (class 1), making the data more balanced for model training.

3. Model Building and Training

- **Model Selection and Evaluation:**
    - Created an **evaluate_model** function using **Repeated Stratified K-Fold Cross Validation** to assess the performance of different models.
    - Trained multiple models to evaluate their performance:
        - **MLPClassifier**
        - **RandomForestClassifier**
        - **ExtraTreesClassifier**
- **Grid Search for Hyperparameter Tuning:**
    - For the **RandomForestClassifier**, you applied **GridSearchCV** to find the best hyperparameters, ensuring that the model was optimized for the best performance.
- **Model Training and Evaluation:**
    - Trained the models on the cleaned and preprocessed dataset.
    - Evaluated the performance of each model using metrics like accuracy, precision, recall, and F1-score.

4. Model Predictions and Saving Results

- After training, you used the models to predict the **Chronic_flag** on the dataset.
- Saved the predictions into the original dataset by adding the **Chronic_flag** predictions as a new column.
- The updated dataset (with the predicted values) was saved into an Excel file for further analysis and reporting.

5. Model Deployment:

- Saved all the trained models for future use, making it easy to apply them to new data when needed.

6. Data Visualization (Power BI Dashboard)

- To further understand the relationships in the data, you built a **Power BI Dashboard**:
    - **Charts:** You used a combination of bar charts, line charts, and scatter plots to visualize key relationships like **HBA1C Test Compliance**, **Age Range**, **Gender**, and **Chronic_flag**.
    - **Segmentation and Trends:** The dashboard allowed you to segment the data based on **Age Range**, **Gender**, and other variables, helping identify trends in the **Chronic_flag** variable (such as which age range has the highest rate of chronic conditions).

  - **Data Analysis:** The dashboard provided visual insights into the impact of different features (like **Age**, **HBA1C Results**, and **Comorbidities**) on the chronic condition prediction.

7. Result Analysis:

- **Insights from Power BI:** The dashboard gave you valuable insights, including demographic trends, testing compliance, and how different factors (such as age and gender) correlate with the likelihood of being diagnosed with chronic conditions.
- **Business Decision Support:** The dashboard served as an effective tool for stakeholders to make data-driven decisions based on the analysis of trends in the **Chronic_flag** variable.

Conclusion:

By following these steps, you effectively cleaned and preprocessed your data, handled biases, dealt with missing values, and built machine learning models to predict chronic conditions. You then visualized the results using a Power BI dashboard, providing a comprehensive understanding of the dataset and the relationships between various factors and the target variable **Chronic_flag**.