**Take-Home Exam: Data Scientist Position**

**Instructions**

1. You may use either **R** or **Python** to complete the assignment.
2. Your code should adhere to best practices, including clear documentation, consistent formatting, and meaningful variable/function names.
3. Please create a GitHub repository, upload your solution there, and share the link with us. Ensure the repository contains:
   * The code
   * A README file explaining your approach, how to run your code, and the key results.
   * A requirements.txt or equivalent file listing dependencies.

**Scenario: Drug Response Analysis**

You are provided with a dataset containing gene expression data and patient metadata from a clinical trial evaluating a new drug for an autoimmune disease.

**Dataset Details**

1. **gene\_expression.csv**:
   * Rows: Samples
   * Columns: Gene symbols (e.g., TP53, IL6) and their expression levels.
2. **meta\_data.csv**:
   * SampleID: Unique identifier for each sample.
   * Response: Binary variable indicating response to treatment (Responder or Non-Responder).
   * Das28: Clinical score given by the physician corresponds to disease severity.
   * Gender: Patient gender (Male, Female).

**Tasks**

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

* Provide an overview of the dataset.
* Visualize the distribution of key variables (e.g., gene expression levels, response rates).
* Highlight any missing values or outliers and describe your approach to handle them.

**2. Feature Selection**

* Identify genes significantly associated with the treatment response. Use appropriate statistical methods (e.g., t-tests, ANOVA, or feature selection techniques).
* Summarize your findings with a list of the top 10 genes and a brief justification for their selection.

**3. Predictive Modeling**

* Build a classification model to predict treatment response using gene expression and patient metadata.
* Use appropriate feature engineering and model evaluation techniques (e.g., train/test split, cross-validation, performance metrics such as accuracy, precision, recall, and AUC).
* Justify your choice of the model.
* Evaluate the performance of your classification model using relevant metrics (e.g., accuracy, precision, recall, F1 score, AUC-ROC).
* Provide a confusion matrix and interpret the results in the context of treatment response.
* Discuss the trade-offs between different performance metrics and explain which metric(s) are most appropriate for this clinical scenario.

**4. Explainability and Feature Importance**

* Identify the most important features (genes or metadata) influencing model predictions.
* Use explainability methods (e.g., SHAP values, permutation importance) to demonstrate which features contribute most to treatment response predictions.
* Provide an interpretation of these important features in the clinical context.

**5. Code Standards**

* Ensure your code is modular, reusable, and well-documented.
* Include unit tests for key functions to demonstrate testing practices.

**6. GitHub Workflow**

* Use a GitHub repository to document your work.

**Deliverables**

1. GitHub repository link.
2. Short report in the README file, summarizing:
   * Key insights from EDA.
   * Top genes identified in feature selection.
   * Model performance and interpretation.