Mass General Brigham
Mass Eye and Ear
HARVARD MEDICAL SCHOOL

Shalhout Lab Research Tech I Exam

This exam will test a candidate's ability to perform tasks related to machine learning (ML) modeling, natural language processing (NLP), data cleaning, data merging, and data analysis/visualization.

Instructions:

Exam Duration: 1 week from this email (please reach out directly to Sophia_Shalhout@meei.harvard.edu if you need more time)

Submission Format: Submit an R Markdown or Jupyter Notebook as an **html** via email with answers and figures, along with well annotated code.

Tools Allowed: Open book, internet resources, and external packages/libraries allowed.

Expected Output: Clean, well-commented, annotated R/python code with explanations and a final report summarizing findings. **Please deliver 1 markdown/notebook that contains answers to all 6 problems**.

Problem 1: Data Cleaning and Merging Incompatible Datasets

You are given two csv files (**Dataset_A and Dataset_B**) that contain genetic mutation information from a targeted mutation panel. However, the two datasets come from different sources, and column/row names differ. Your task is to merge the two datasets based on the appropriate columns. Assumptions may need to be made to merge the files. If you make an assumption to complete the task, mention it in the code and defend your assumption (ie, justify the assumption). After merging, identify and remove any duplicate entries. Identify any missing or anomalous data in the merged dataset. Describe your strategy for handling these missing values or anomalies.

Deliverables: Submit the merged data, code used for merging the datasets and code for cleaning the data. Provide a brief explanation of the steps you took to handle column name incompatibility, missing values, and duplicates.

Problem 2: Natural Language Processing for Biological Text

You are given a collection of research articles (**in plain text format, see article_1 through article_5**) related to genetic mutations and their implications. Your task is to perform basic NLP tasks to analyze these articles. Calculate the frequency of the most common entities and plot the results in a bar chart.

Deliverables: Submit the code used for the preprocessing. Submit a visualization showing the most frequent biological entities in the dataset. Provide a brief summary of your findings.

Problem 3: Mutation Prediction Using Machine Learning

You are provided with a dataset (**ML_dataset**) of genetic mutations and their clinical outcomes. The dataset contains the following columns:

`Gene`: The gene where the mutation occurs.

`Mutation`: The specific mutation.

`Location`: The location of the mutation on the gene.

`Effect`: The functional effect of the mutation.

`ClinicalOutcome`: The clinical outcome associated with the mutation (binary: 1 for positive outcome, 0 for negative outcome).

Your task is to build a machine learning model to predict the clinical outcome based on the available features. Evaluate your model's performance on the test set using appropriate metrics (e.g., accuracy, precision, recall, F1-score). Provide a brief explanation of the model selection process and your rationale behind it.

Deliverables: Submit the code used for training and evaluating the model. Submit the performance metrics of the model along with a short interpretation. Provide a ROC curve if applicable.

Problem 4: Visualizing Mutation Data

You are given a dataset containing information about the frequency of specific genetic mutations observed in a targeted panel test (**mutation_data**). The dataset has columns like `Gene`, `Mutation`, `MutationType`, and `Frequency`.

Your task is to: Create a heatmap to visualize mutation frequency across different genes. Create a scatter plot to visualize the relationship between mutation type and frequency. Identify any patterns or insights from the visualization and explain them.

Deliverables: Submit the code used for creating the visualizations. Provide the heatmap and scatter plot as part of the submission. Write a short interpretation of the visual patterns observed.

Problem 5: Aligning Multiple Panels

You are provided with data from two different mutation panels. One is a **commercial panel**, and the other is an **in-house panel**. Both panels list genes and mutation types, but the commercial panel includes additional annotations. Your task is to align the data from the two panels and identify which genes and mutations are shared between them and which are unique to each panel. Highlight any discrepancies in mutation annotations and suggest how to resolve these. Present the aligned data in a clean, understandable format.

Deliverables: Submit the R/python code used for aligning the panels. Provide a table or visualization showing the shared and unique mutations across the panels. Write a brief discussion on any discrepancies found.

Problem 6: Identifying and Resolving Medication Data Duplications

You have been provided with a dataset containing **medication records for patients**. The dataset has three columns: **patient_id, date_of_med, and medication_name**. The dataset includes a total of 5000 instances of a patient receiving a medication for 500 patients, where each patient can have multiple entries across various dates. Your task is to clean the data by removing duplications. Please make sure with such data, all duplications are removed, even those that may not be immediately obvious.

Deliverables: Present a summary of how many duplications were found and removed. Provide a cleaned version of the dataset and annotate code to inform how this was carried out.

Good Luck!