### **Olive Diagnostics Home Exercise**

### Dear Candidate,

### Thank you for taking the time to complete our exercise. I believe this task will best demonstrate your work style at Olive Diagnostics.

### The first question involves a full-cycle exercise. I’m interested in understanding your thought process and evaluating your programming skills. The objective is to perform exploratory data analysis (EDA) on a given dataset and showcase your machine learning abilities. Please generate meaningful visualizations to gain insights into the data. Be creative and thorough, explaining each step clearly and suggesting relevant solutions.

The second question focuses on signal processing. I’m interested in seeing how you handle the signal, what manipulations you can perform, and how you extract relevant features. You can read more about the Beer-Lambert law in spectroscopy for additional context, but it’s not a requirement.

### Please remember that I can only assess your thought process if you articulate it clearly in your submission.

For this exercise we will have 3 days, in case of any question please feel free to contact me at [yuval.cohen@olive.earth](about:blank)

### 

### **Exercise 1: Full Cycle of Data Handling, Prediction, and Analysis**

**The data:**

**Patient\_ID:** Unique identifier for each patient.

**Protein:** Protein levels in the urine.

pH: pH levels of the urine.

**Specific\_Gravity:** Specific gravity of the urine.

**Blood:** Binary indicator of the presence of blood in the urine (0 or 1).

**Age:** Age of the patient.

**Gender:** Gender of the patient (Male or Female).

**Glucose:** Glucose levels in the urine.

**Calcium:** Calcium levels in the urine.

**Condition:** Binary indicator of whether the patient has a specific condition (0: No, 1: Yes)

1. **Data Handling:**
   * You are given a dataset containing patient urine analysis results along with labels indicating the presence or absence of a specific medical condition (e.g., kidney stones).
   * The dataset has missing values, outliers, and various types of features (e.g., numeric, categorical).
2. **Data Cleaning & Exploration:**
   * Clean the dataset by handling missing values, identifying and treating outliers, and preparing features for modeling.
   * Perform exploratory data analysis (EDA) to understand the distributions, correlations, and key features that might influence the target variable.
3. **Model Building:**
   * Build a predictive model to classify whether a patient has the medical condition based on the urine analysis features.
   * Use any machine learning algorithm of your choice, and justify your selection.
4. **Model Evaluation:**
   * Evaluate the model using appropriate metrics (e.g., accuracy, precision, recall, F1-score).
   * Perform cross-validation to assess the model’s robustness.
5. **Result Analysis:**
   * Provide a critical analysis of the model’s performance. Discuss any limitations, potential improvements, and the implications of your findings.
   * Do errors analysis, use visualization and summarize the results in a few sentences.
   * If you had an option to reject some of your predictions, what was your criterion? What was the effect on your metrics? Suggest a way to show this trade-off.
6. **A/B testing:** 
   * try another model and test if one model is better than the other using appropriate statistical tests. Explain your test, assumptions, etc.
7. **Software skills:**
   * Write a python class, divide your code into appropriate functions (according to best practice roles.
   * Add unit-test in the places you find it necessary (please read first about unit test before you are doing it)

**Deliverable:**

* A Jupyter notebook containing all code, analysis, and explanations.
* A brief report summarizing your findings and conclusions.
* Please be creative and thorough.

### **Exercise 2: Signal Preprocessing and Analysis**

**Task:**

1. **Signal Data:**
   * You are provided with a raw signal dataset representing time-series data from urine spectrometry measurements (e.g., light intensity readings over time).
2. **Preprocessing:**
   * Apply necessary preprocessing steps to the signal, including filtering noise, detrending, and normalizing the data.
   * Visualize the original and preprocessed signals to demonstrate the effectiveness of your techniques.
3. **Frequency Analysis:**
   * Perform a Fast Fourier Transform (FFT) on the signal to analyze its frequency components.
   * Identify any significant frequency components and explain their potential relevance to the underlying chemical processes in the urine.
4. **Feature Extraction and Engineering:**
   * Extract key features from the signal (e.g., peak frequencies, amplitude, etc.) that could be used in a predictive model. Try to generate new features based on your raw data (any transformation?)

**Deliverable:**

* A Jupyter notebook containing all preprocessing steps, FFT analysis, visualizations, and explanations.
* A short summary describing your approach and findings.