Dear Candidate,

Thank you for your interest in the Senior ML Engineer position at Samay. We are excited to move forward with your application. As part of our evaluation process, we have a technical test, designed to assess your skills and compatibility with our team's needs.

Follow the instructions and submit your answers naming them as, *[Hiring] Technical test_Senior ML Engineer_Complete Name*

This test assesses candidates' ability to develop predictive models using acoustic data to diagnose respiratory health conditions. Specifically, candidates will be tasked with analyzing a dataset containing acoustic data related to certain health parameters and building a predictive model to identify and classify health conditions based on this data.

The primary goal of this technical test is to design, develop, and optimize a machine-learning solution for detecting and classifying respiratory diseases. This will be achieved using a dataset of lung sounds recorded with an electronic stethoscope. The comprehensive IPython Notebook delivered should demonstrate the candidate's proficiency in analyzing sensor data for health applications, specifically in pulmonary diseases.

Dataset Description:

The evolution of stethoscope technology has facilitated the high-quality recording of lung sounds from healthy individuals and those with various pulmonary conditions.

This dataset encompasses audio recordings from patients with seven different ailments, including asthma, heart failure, pneumonia, bronchitis, pleural effusion, lung fibrosis, and COPD, alongside normal breathing sounds. Recordings were taken from multiple positions on the chest, as determined by a specialist physician. Each sound was processed thrice using different frequency filters to highlight specific bodily sounds.

This valuable dataset supports the development of automated tools for diagnosing pulmonary diseases through lung sound analysis and can be extended to heart sound studies. The dataset is structured and may include features such as frequency, amplitude, duration, and spectral characteristics of the acoustic signals.

The dataset comprises audio recordings of lung sounds from 112 subjects, captured using an electronic stethoscope. It includes data from 35 healthy individuals and 77 subjects with various respiratory diseases [1, 2, 3].

- **Content:** The audio recordings have been filtered through Bell, Diaphragm, and Extended modes to ensure clarity and precision in sound quality.
- **Annotations:** Each audio file is annotated with comprehensive details including the type of lung sound, the disease diagnosis, recording location on the subject's chest, as well as the age and gender of the subjects. This information is crucial for the analysis and classification tasks.

Task:
Use the data provided on Kaggle Dataset: Lung Sound Dataset to solve the following assessment questions

1. In this part, the test will evaluate the candidate's ability to load a dataset into a Python environment (likely using libraries like Pandas) and perform preliminary analysis.

Questions:

a. How did you load the dataset into your Python environment? What libraries were used, and why? Provide a code snippet that verifies the correct loading of the data and displays the first few rows.
b. What is the structure of the dataset and the data types of each column after loading? Are there any columns that require type conversion for further analysis? How would you handle these conversions?
c. Are there missing or anomalous values in the dataset? Describe the method(s) you used to identify these issues and your approach to handling them.
d. Generate and interpret the descriptive statistics of the dataset. What insights can you derive from these statistics, and how might they influence your preprocessing steps?
e. Analyze the balance of the dataset across the different conditions (e.g., asthma, pneumonia). How many recordings are there per condition? Provide a visualization (such as a bar chart) showing the distribution of classes and discuss potential model training challenges due to these distributions.

2. This part focuses on signal processing and feature engineering, particularly in the context of audio or time-series data. Candidates may be asked to perform tasks like analyzing audio signals, visualizing data (e.g., waveforms), extracting relevant features (e.g., MFCC features), and explaining their significance.

- Explore and preprocess the acoustic dataset, including data cleaning, normalization, and feature engineering as needed.
- Conduct exploratory data analysis (EDA) to gain insights into the distribution and characteristics of the acoustic data.

Note: If you are not familiar with audio processing techniques, it is suggested to do an exploratory search although you can feel free to work with the signals as vectors or time series and extract characteristics that allow you to feed the proposed classification model.

Questions:

a. Describe the process you used to analyze the audio signals from the dataset. What specific characteristics did you examine, and what tools or libraries did you use for this analysis?
b. Provide a Python code snippet that visualizes the data. What types of visualizations did you choose (e.g., waveform, spectrogram, data transformations) and why? Discuss any patterns or anomalies observed from these visualizations.
c. Which features did you extract from the audio signals, and why did you choose them? Explain the significance of these features (e.g., MFCC) in the context of classifying respiratory conditions.
d. What preprocessing steps were necessary for the audio data before feature extraction? Discuss any techniques you applied for data cleaning, normalization, or transformation and justify their use.
e. Summarize the findings from your exploratory data analysis on the acoustic data. What insights were gained about the distribution and characteristics of the data? How do these insights impact the subsequent steps in your data analysis and model development process?
f. How is the distribution of the labels? Provide a detailed explanation of the label distribution within the dataset and discuss any imbalances.

Deliverables:

Candidates are required to submit the following deliverables:

1. Python code or Jupyter notebook containing the data preprocessing, analysis, and modeling process.
2. A written report summarizing the candidate's approach, findings, and recommendations.
3. Optionally, candidates may include visualizations, model performance metrics, and any additional insights derived from the analysis.

Additional Information:

- Candidates are encouraged to utilize relevant libraries and frameworks such as NumPy, pandas, sci-kit-learn, and TensorFlow/PyTorch for their analysis and modeling.
- The use of appropriate data visualization techniques is highly recommended to facilitate understanding and interpretation of the results.
- Candidates should adhere to best practices in data science, including proper documentation, code readability, and reproducibility.
- Candidates should consider best practices when sharing their Python code or Jupyter notebook: how would other people easily run and verify your code? consider how to best document/explain the steps to run your code.

References

1. Fraiwan, Mohammad; Fraiwan, Luay; Khassawneh, Basheer; Ibnian, Ali (2021), “A dataset of lung sounds recorded from the chest wall using an electronic stethoscope”, Mendeley Data, V3, doi: 10.17632/jwyy9np4gv.3.
2. Luay Fraiwan, Omnia Hassanin, Mohammad Fraiwan, Basheer Khassawneh, Ali M. Ibnian, Mohanad Alkhodari, "Automatic identification of respiratory diseases from stethoscopic lung sound signals using ensemble classifiers," Biocybernetics and Biomedical Engineering, Volume 41, Issue 1, 2021, Pages 1-14, ISSN 0208-5216, https://doi.org/10.1016/j.bbe.2020.11.003.
3. Kaggle Dataset: Lung Sound Dataset**.**