Assignment for Data Scientist/Data Engineer

You have been tasked with creating a new assistive technology that can identify the class of health condition described in a given patient description. The technology should assist doctors as they review many such descriptions each day during their hospital rounds, and help them quickly identify key information about a patient's condition. The dataset provided to you contains descriptions of 5 different health conditions: those related to the digestive system, cardiovascular system, neoplasms, nervous system, and general pathological conditions. There are a total of 14438 records in the dataset, which is presented as text.

To complete this assignment, you will need to:

- Split the dataset into training and test sets as appropriate.
- Perform exploratory data analysis on the data.
- Build a text classification model based on the dataset
- Monitor relevant evaluation metrics while training the model
- Present any interesting or important findings from the exploratory data analysis.
- Save the trained model, and make it available through an API (using either Flask or FastAPI).
- Dockerize the API.
- Host a local RabbitMQ message queue server, and send data for inference through the message queue.
- Create a RabbitMQ Consumer that consumes data from the queue and makes an API call for model inference.
- Save the inference results, along with the text on which model inference was made, in a Postgres/SQLite/MySQL Database (which can be locally hosted on your machine).

Please submit your assignment within 7 days to the following email addresses: aborgohain@inveniai.com and sanatan@inveniai.com.

If you have any questions or concerns, please feel free to contact aborgohain@inveniai.com.