**Project Title - Medical Chatbot**

**Prepared for UMBC Data Science Master Degree Capstone by Dr Chaojie (Jay) Wang**

**Author Name- Sai Thanmai Peddader Pally**

**Link to the author's GitHub profile**

**Link to the author's LinkedIn profile**

**Background Information**

The chosen topic revolves around the development of a healthcare chatbot using a dataset that includes information about various diseases and their associated symptoms. This dataset comprises two primary files: "testing.csv" and "training.csv." The former contains data on 41 diseases and their symptoms.

**Why It Matters**

This topic is of significance due to several reasons:

* Healthcare chatbots and symptom checkers play a vital role in providing accessible and preliminary healthcare guidance.
* With the increasing popularity of telemedicine, such tools can assist individuals in assessing their symptoms and identifying potential health issues.
* Efficient symptom checkers can support healthcare professionals during patient consultations by providing additional context.

**Research Questions**

* How effective is the machine learning model trained on this dataset in accurately identifying diseases based on reported symptoms?
* Can the chatbot provide reliable recommendations for medical actions (e.g., suggesting self-care, consulting a doctor) based on the identified diseases?
* How can the chatbot be optimized for user-friendliness and ease of understanding for individuals with varying levels of medical knowledge?

**Data Information**

* **Data sources -** Kaggle
* **Data size -** 30KB
* **Data shape -** 133 Columns and 4921 Rows
* **What does each row represent? -** The values in the dataset represent whether a specific symptom is associated with a disease
* **Data dictionary**
  + The values in this dataset reflect the presence or absence of symptoms in each training example, with the ultimate goal of associating symptoms with disease names.
  + Potential Values: Categorical variables (symptoms) are binary, with '1' indicating symptom presence and '0' indicating symptom absence.

**Which variable/column will be your target/label in your ML model?**

The target/label variable in my ML model is the Disease(prognosis) column, which contains the names of the diseases that I want to predict based on the symptoms provided. The features/predictors variables are the 132 symptom columns, which contain binary values indicating whether each symptom is present or absent for each disease. These variables can be used as the input (X) of my ML model, while the Disease column can be used as the output (y) of my ML model.