Machine Learning Engineer Nanodegree

Capstone Proposal

Hatem Kamal
Jule 28th, 2021

Proposal

Domain Background

In the last year we had a project to make the diagnosis system for a hospital this project was about the database and the system of the hospital only. But as an improvement I thought about adding what I've learnt here to the system where the user adds the symptoms and get an output of what disease he might have.

Problem Statement

Before the patient goes to the doctor we need to provide an initial point of what disease the patient might have. Also if the patient wants to see what disease he might have from the website. I got some diseases with their symptoms from a kaggle dataset that needed to be trained by a model to get the disease.

Datasets and Inputs

I found a dataset in kaggle about this topic, this dataset has the disease in a column with its symptoms in the rest of the row. This dataset has 41 diseases with 130 different symptoms also the diseases is repeated with each possible symptom attached to each row in 304 row after removing dublicates.

The data isn't quite balanced from 5:10 sample per disease.

The data set is provided in this link:

https://www.kaggle.com/itachi9604/disease-symptom-description-dataset?select=dataset.csv

Solution Statement

As the user inputs his symptoms, we have to pass these symptoms to the endpoint, which have a trained model based on the collected dataset, to get what disease he might have most. The symptoms the user will choose from the template he sees. Then after predicting the disease the results will be shown to him.

I'll make the input to the model to be all the possible symptoms, these inputs are fed to the model to get the diseases with more than 70% probability and show them in the template to the user.

Benchmark Model

For this problem support vector machine model might be a good starting we will be fed the data into the model and get the disease.

Evaluation Metrics

Based on the dataset, the evaluation metrics would be 80% F1, as the data isn't quite balanced, on the validation data.

Project Design

Many part of the project will be in the preprocessing the data to have the desired input. As the csv data has to be the symptoms in the columns and each cell indicates is this symptom and disease related or not (1 or 0). This cleaned data is to be fed to the model to be trained. Same process will be applied to the user input before it's fed to the trained model.