Symptom Based Disease Prediction using Machine Learning

1st Kyle Jerome Just
Computer Science Department
California Polytechnic University
Pomona, USA
kjjust7@gmail.com

2nd Joel Joshy
Computer Science Department
California Polytechnic University
Pomona, USA
joeljoshy3243@gmail.com

3rd Musa Waghu
Computer Science Department
California Polytechnic University
Pomona, USA
musawaghu@gmail.com

4th Joshua Furman
Computer Science Department
California Polytechnic University
Pomona, USA
jafbird52@gmail.com

5th Anita Mehrazarin
Computer Science Department
California Polytechnic University
Pomona, USA
mehrazarin.a@gmail.com

***Abstract*—Machine learning has shown great promise in disease prediction using clinical and demographic data. In this paper, we propose a novel approach for disease prediction using machine learning techniques and evaluate its effectiveness in predicting several diseases using real-world healthcare datasets. Each diagnosis will be based on collected data from a patient that is then applied to our classification model to give a single, final diagnosis as well as treatment options and more official tests for an identified illness.**

***Index Terms*—Disease, Symptoms, Automated Diagnosis, Decision Tree Classifier, Random Forest Classifier, Naive Bayes Classifier, Healthcare**

I. Introduction

The healthcare system, especially here in the USA is highly expensive and often avoided by people for this reason. Often people dismiss their symptoms to be a trivial illness and refuse to see a professional. Unfortunately, if these symptoms point to a more serious illness, it may fester until treatment may be much more difficult or even impossible. Depending on the accuracy of each model, we may decide to use a random forest model, Naive Bayes classifier, or decision tree to classify each illness. A dataset of symptoms with their subsequent diagnosis will be used with these models attempt to predict the illness of a patient. This patient will submit their symptoms, and severity of each, to the model and receive a preliminary diagnosis along with resources to follow up the preliminary with a more definitive test or doctor's visit.

This will be able to provide medical information and treatment to people without the doctors that can diagnose their specific illness. A simple internet connection is all that will be needed to provide help to the people most in need. It can also help divert trivial illnesses from doctors by providing safe, at home remedies and allow people with more pressing illnesses to see a doctor sooner.

II. Datasets

The datasets that we will be using include samples of a multitude of diagnosed conditions and the symptoms that were experienced before the diagnosis. This dataset includes about 9 features of symptoms per diagnosis with about 5000 data points. Each unique illness has 120 data points, resulting in a very balanced dataset. This dataset also contains data with a patients reported severity of symptoms as well. A second dataset that we are using includes very similar data, but also includes common medications, as well as common tests to detect such an illness. With these two datasets in tandem we want to create a model to diagnose and provide advice to patients at risk for the diseases we cover.

Most of our data will be sourced from kaggle.com and other data repositories as we are not able to collect our own data on the problem. These datasets will be named and linked within the references section at the end of this report.

III. Methodology

Initially, we will be training a decision tree to attempt to classify each disease based on symptoms and severity. However, due to the nature of the problem a Naive Bayesian model, neural network, non-linear SVG, random forest, or K-nearest neighbor may suffice as well.

During our testing phase, we will be using cross validation to active the most balanced and accurate models. We will split the data into 6 separate groups. Each group will contain 20 samples of each disease in it resulting in balanced test sets and training sets. Since this model will employ a batch learning training model, the expensive computation that will be required for the model will prove moot in the pursuit of the most accurate model.

In the goal to diagnose diseases, false positives are a safer option than false negatives. A false negative could lead to a sense of security and a negligence to illnesses that may need attention immediately. To reduce the amount of false negatives

given by the model, we have decided to maximize recall rather than accuracy or precision. We may even use a weighted metric with accuracy and recall selected for, but that is to be decided as we progress through the project.

We will be using a multi-class classification model, so each disease will be assumed to be mutually exclusive. This may not be the case in the real world, however to keep results simple and readable we will assume that the best match of the symptoms will be the proposed rather than all of the diseases they apply to.

REFERENCES

[1] Patil, Pranay. "Disease Symptom Prediction." Kaggle, 24 May 2020, https://www.kaggle.com/datasets/itachi9604/disease-symptom-description-dataset.

[2] HA.Gari. 2022 April. disease and their symptoms, Version 1. Retrived May 13th 2023 from https://www.kaggle.com/datasets/hagari/disease-and-their-symptoms