University of Colorado, Boulder

Sami Lababidi
CSCI 4022 Fall 2021
Dr. Zachary Mullen

Disease Identification System

Dec 15, 2021

OVERVIEW

The healthcare system has been under a lot of pressure recently due to the COVID-19 pandemic. Urgent care clinics and emergency rooms struggle to keep up with the large number of patients seeking medical attention. Some hospitals are turning patients away when they run out of beds. Healthcare workers are doing everything they can to help those in need of medical attention. However, we need to find a way to make it easier and faster for them to diagnose patients.

I care about this topic because my brother is a doctor and I saw how much he struggled during the hardest days of the pandemic. I am also a close friend of an ER nurse who now suffers from depression as a result of working long hours and seeing many people die in intensive care from COVID-19. These two reasons and in an attempt to help healthcare workers who worked tirelessly to help their community, I started thinking about how I can help them and came up with an alternative solution.

DATASET

The dataset available is divided into four csv tables. The first table contains a list of diseases and their descriptions, 41 rows. The second contains a set of recommendations for each disease, 41 rows. The third table is a list of all unique symptoms reported by patients, 134 rows.

The final table has a list of cases with a combination of symptoms that previous patients reported and their final diagnoses, 4920 rows.

The dataset was published on kaggle.com by Pranay Patil under the name “Disease Symptom Prediction” for education pupusas. The link for the data along with more details can be found by clicking here.

Real World Impact

With a large sample of data from patients across the world, we can implement a system that can accurately diagnose patients without the interference of doctors in many cases which can help them work more efficiently and help patients reach medical care faster. This project is still in its early stages, the long term goal is to develop a system that won’t only allow patients to record their symptoms, but also record the severity of these symptoms. The system should be able to order lab tests, recommend basic patient care, and maybe even describe basic medication under the supervision of nurses. With data science developing everyday, I believe that we can in the future push for regulations that will allow such a system to exist. The real world impact for that is not only to help reduce pressure on doctors but also to provide a cheaper basic medical care for patients without health insurance and can’t afford one.

Exploratory Results

Testing by adding a new patient that is experiencing the following symptoms

```
# the patient is experiencing these symptoms from the list above
curr_sym = [reported_symptoms.index('vomiting')]
curr_sym.append(reported_symptoms.index('fatigue'))
curr_sym.append(reported_symptoms.index('high_fever'))
```

Then we run A-Priori algorithm to find the most frequent symptoms with the current symptoms:

```
# create new combination
new_combo = add_sym(all_sym, curr_sym, ignored_sym)
supp = A_priori(new_combo, case_buckets)
sym_supp = return_supp(new_combo, supp, 0.04)
print_sym_supp(sym_supp, reported_symptoms)
```

```
chills 0.062195121951219512
yellowing_of_eyes 0.041463414634146344
malaise 0.04024390243902439
```

After that, we present the patient with the returned symptoms to see if any apply:

```
# add symptoms
curr_sym.append(reported_symptoms.index('yellowing_of_eyes'))

# patient not experiencing the symptom
ignored_sym.append(reported_symptoms.index('chills'))
ignored_sym.append(reported_symptoms.index('malaise'))

# create a row for the new patient to be added to recommendation array
new_patient = add_patient(all_sym, curr_sym, ignored_sym)
case_buckets[len(case_buckets)] = curr_sym
```

Run collaborative filtering to find the k-nearest patients:

```
k_Nearest = nearest_k_neighbors(case_symptom_matrix, 50, new_patient)

print(suggestion_set(patient_diagnosis, k_Nearest, diseases))
```

```
{'Chronic cholestasis', 'Hepatitis D', 'Hepatitis C'}
```

Results

Collaborative filtering suggesting checking for the above three medical conditions. The results returned only three diagnoses from 50 similar patients. This means that these patients were experiencing similar symptoms when diagnosed with the same disease. This test is successful because the system was able to recommend similar cases that have a common diagnosis. Only 3 recommended medical conditions out of 50 nearest patients.

Conclusion

To get better accuracy from this system, we need to collect data from patients with ratings for their symptom severity from 1 to 10. The rating will allow us to distinguish between diseases that have a similar set of reported symptoms. In addition, we will need a much larger sample than we currently have. The long term goal from this project is to create a system that is able to order lab testing, provide medical care instructions, and recommend treatments under the supervision of nurses. This can be a great tool for healthcare providers to help them narrow down the list of possible diagnoses to test for. It can also provide a virtual tool for patients to use at home who can’t reach a clinic and to limit crowding at hospitals.