The Problem: Enabling Automated Triage for Non-Medical Personnel

People without medical training often lack the prerequisite knowledge or experience to correctly assess when they should seek medical care. Some medical conditions require timely care for good outcomes. These can range from acute emergencies that require immediate care for survival to chronic conditions that should be treated within a few months to avoid severe future consequences. The process for sorting patients into classes based on the priority of care is called triage. Triage is predominantly used by emergency medical personnel to sort patients at mass casualty incidents to maximize the effective use of limited onsite resources and transportation options, but the same principles can be applied to all types of medical conditions. This could be used by medical offices to facilitate outreach from staff to patients or to allow patients to gain insight on how quickly they need to seek care. It is common for patients to hesitate seeking care based on the anticipated expense of seeking emergency room care vs urgent care vs a regular doctors' appointment. Any low cost or self-use option would help fill this gap and be a useful adjunct to the current system. To facilitate better patient triage, a model to quickly classify patients using a series of reported symptoms could be developed and deployed for use in medical apps, doctors' offices, or websites.

The goal is to create a model that sorts conditions into 4 categories:

1) Immediate care needed
    a. Seek medical care as soon as possible,
    b. Activate the local emergency medical system if needed.
2) Seek care soon (~ 2-5 days)
    a. While not immediate, timely care is required to ensure a good clinical outcome.
    b. The disease may progress significantly towards eventual death or disability without the appropriate care.
3) Address at next visit
    a. Condition is chronic and not likely to lead to immediate harm.
    b. The clinical outcome depends on a long-term care strategy with a regular physician, best addressed with medical providers that have or can establish long term relationships with the patient.
4) Need more information for triage
    a. Timeliness and type of care depend on the severity of symptoms that can vary widely.
    b. The next step would be to contact a nurse helpline or a similar service to receive personalized advice.

I will use a dataset collected from a third world location; it contains anonymous information for 4920 patients with 133 symptoms that are used to diagnose 41 diseases. First, I will assign a triage category based on the above criteria for each diagnosis in the dataset. Next, I will use a decision tree model to predict the triage category based on symptom features. I will then fine-tune the model to be as accurate and simple as possible; to eliminate features that would be impossible or difficult for non-medically trained people to assess accurately. I will also attempt to use various techniques to examine the explainability and significance of each feature. This information will then be used to further tune the model for optimal feature selection.

Initial Data Cleaning

I obtained a dataset that contained the medical information for 4920 patients, the dataset includes 133 binary symptom classifiers and 41 medical diagnoses. The dataset had no missing data, except for 1 column (fluid_overload), for which there was a duplicate column (fluid_overload.1). This was fixed by dropping 'fluid_overload' and renaming 'fluid_overload.1' to 'fluid_overload' Additionally, there was an extra unlabeled column where each observation was designated as null that had to be dropped. This was likely an index column that was incorrectly formatted and could not be read correctly into a pandas dataframe.

Oddly, the column for the disease diagnosis was mislabeled as prognosis, which is a medical term that indicates the likely outcome of a condition but is not always a classification per se. This was fixed by renaming the column from "prognosis" to "diagnosis".

Lastly, I generated the test variable (triage category), by assigning each diagnosis to a new triage category indicated by the rules above: 1 = Immediate care needed, 2 = Seek care soon, 3 = Address at next visit, 4 = Need more information for triage.

Exploratory Analysis

Curiously, the dataset had a uniform distribution for the count of each diagnosis (See appendix A). The count for the presence of each symptom was not uniform and ranged from 1932 for 'fatigue' to 102 for 'foul_smell_of_urine'. Some symptom counts were uniform between clusters of symptoms. This tended to occur for symptoms that had lower counts. Unfortunately, this phenomenon is suggestive of multicollinearity. The first decision tree model obtained had a testing accuracy score of 100%, indicated that there was data leakage in the model. Due to the uniformity of certain symptoms, it was highly likely that there was excessive multicollinearity in the data, and that certain symptoms or triage scores were too highly interrelated to be represented as different factors. To assess this, I performed a Cramer's V test across all variables to assess interrelatedness.

Statistical Analysis

The initial Cramer's V test indicated that the model had significant interrelationships between certain symptoms. With multiple pockets of high relationships occurring between certain symptom clusters (see figure 1).

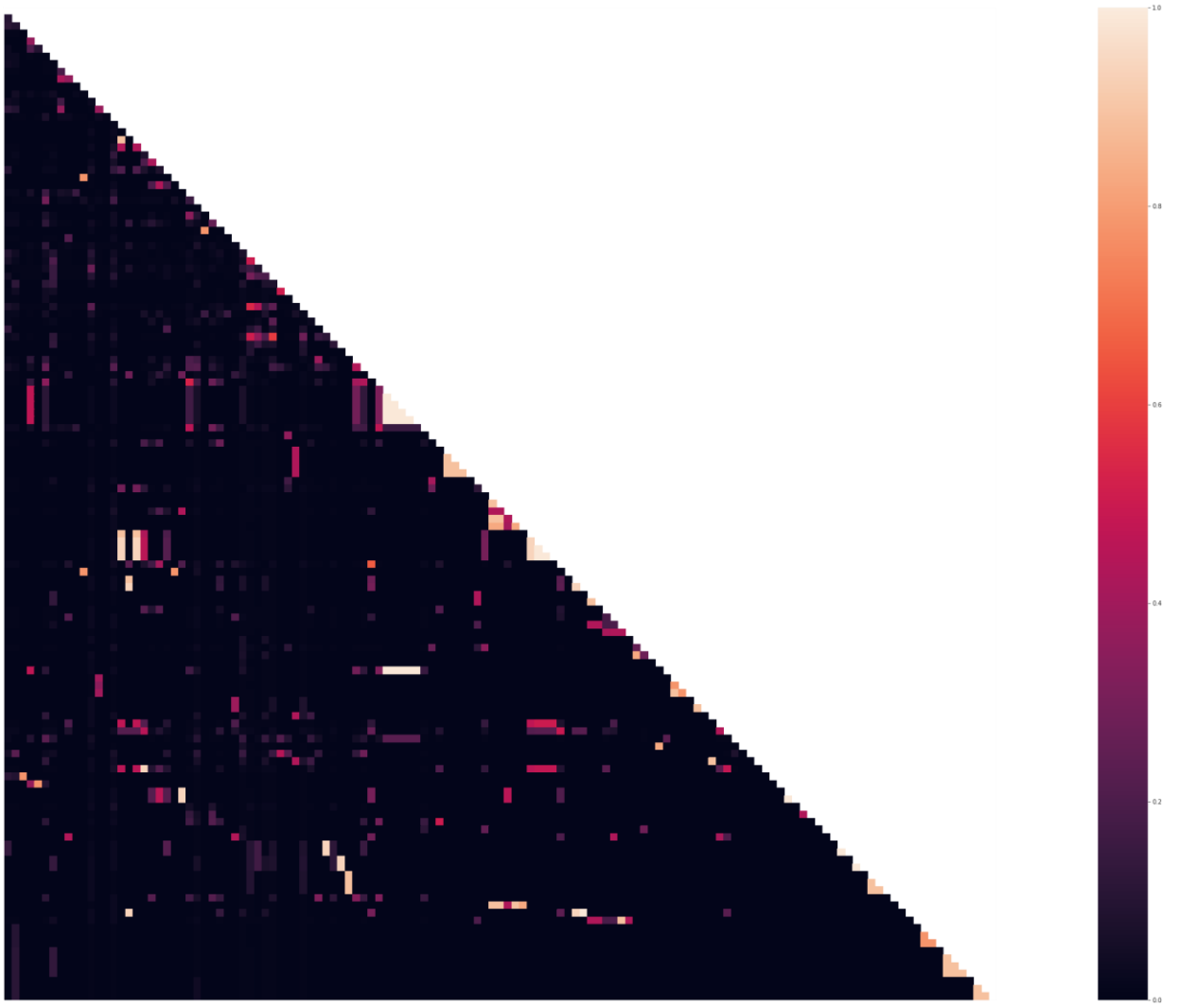

Figure 1: Initial Heat Map Generated Using Cramer's V.

I used the output of the first Cramer's test to remove symptoms that had a strong relationship (> 0.4) with other symptoms using a simple set of logic. I first dropped any symptom that a nonmedically trained person would be unable or highly unlikely to identify correctly (for example 'coma', or 'acidity'). Second I dropped any symptom that was highly related to a set of other symptoms, for example, 'loss_of_smell" would be associated with 'phlegm', 'nasal_discharge', or any other symptom related to a cold or upper respiratory condition. Lastly, I cleaned any binary interrelationship by dropping the symptom that would be more difficult to detect or diagnosis accurately. After applying this set of logical rules, I ran another Cramer's test over the remaining symptoms to confirm there were no longer any symptoms with strong relationships.

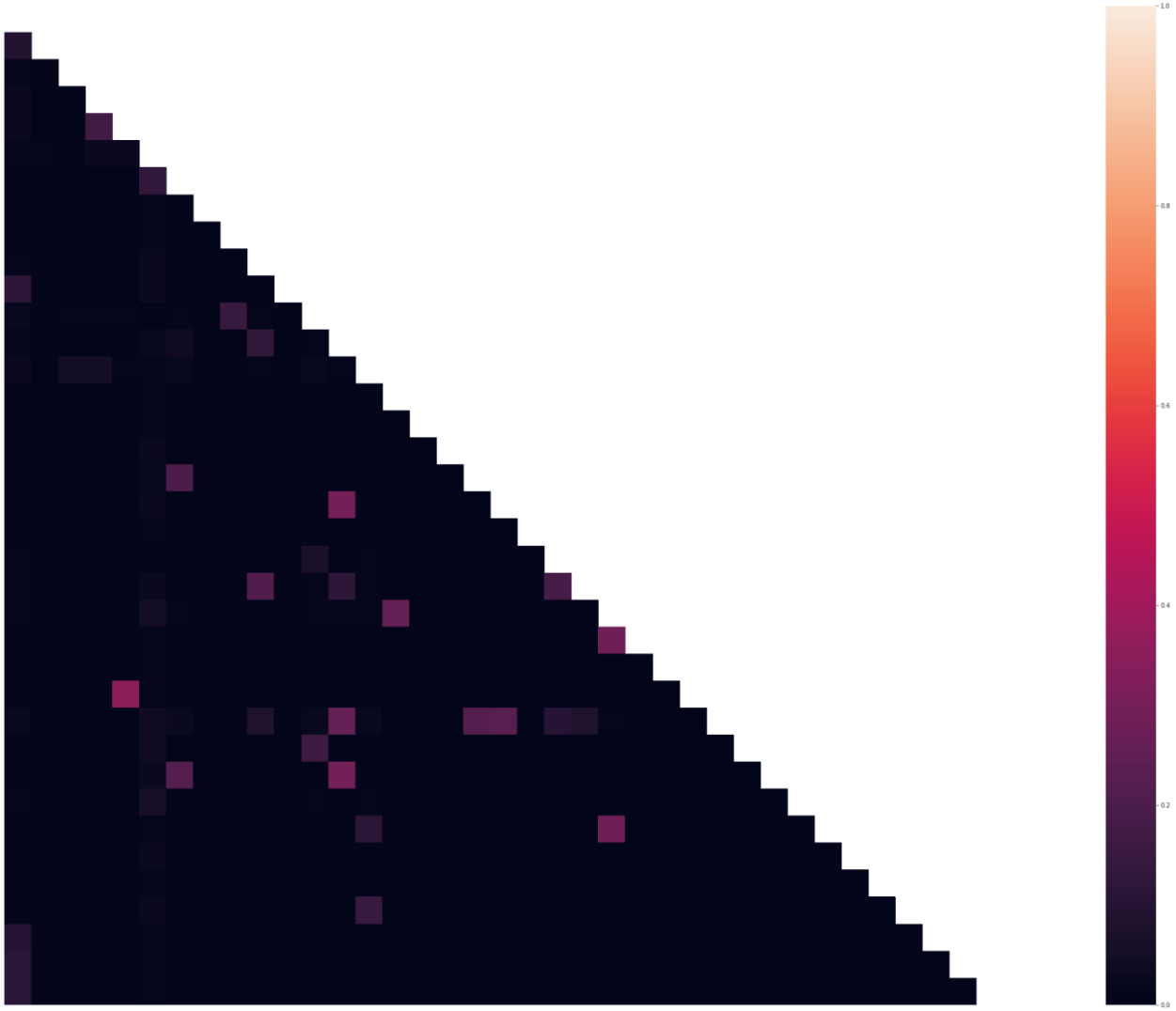

Figure 2: Second Heatmap Generated Using Cramer's Test After Cleaning.

As you can see from the above figure, The second decision trees model using the remaining features no longer had accuracy scores of 100%. Indicating that the data leakage problem had been solved.

Diagnosis	Count
(vertigo) Paroymsal Positional Vertigo	120
AIDS	120
Acne	120
Alcoholic hepatitis	120
Allergy	120
Arthritis	120
Bronchial Asthma	120
Cervical spondylosis	120
Chicken pox	120
Chronic cholestasis	120
Common Cold	120
Dengue	120
Diabetes	120
Dimorphic hemmorhoids(piles)	120
Drug Reaction	120
Fungal infection	120
GERD	120
Gastroenteritis	120
Heart attack	120
Hepatitis B	120
Hepatitis C	120
Hepatitis D	120
Hepatitis E	120
Hypertension	120
Hyperthyroidism	120
Hypoglycemia	120
Hypothyroidism	120
Impetigo	120
Jaundice	120
Malaria	120
Migraine	120
Osteoarthristis	120
Paralysis (brain hemorrhage)	120
Peptic ulcer diseae	120
Pneumonia	120
Psoriasis	120
Tuberculosis	120
Typhoid	120
Urinary tract infection	120
Varicose veins	120
hepatitis A	120

Appendix A: Dataset Diagnoses

Appendix B: Remaining Symptoms by Frequency

fatigue	**1932**
skin_rash	**786**
chest_pain	**696**
diarrhoea	**564**
irritability	**474**
lethargy	**456**
blurred_and_distorted_vision	**342**
loss_of_balance	**342**
muscle_weakness	**234**
stiff_neck	**228**
family_history	**228**
continuous_sneezing	**222**
stomach_pain	**222**
indigestion	**222**
burning_micturition	**216**
pain_behind_the_eyes	**120**
slurred_speech	**120**
polyuria	**120**
stomach_bleeding	**120**
blood_in_sputum	**120**
pain_during_bowel_movements	**114**
swollen_legs	**114**
puffy_face_and_eyes	**114**
knee_pain	**114**
unsteadiness	**114**
belly_pain	**114**
lack_of_concentration	**114**
history_of_alcohol_consumption	**114**
silver_like_dusting	**114**
red_sore_around_nose	**114**
nodal_skin_eruptions	**108**
patches_in_throat	**108**
dehydration	**108**
weakness_in_limbs	**108**
weakness_of_one_body_side	**108**
pus_filled_pimples	**108**
foul_smell_of urine	**102**