Disease Prediction Dashboard

Nick Bagley, Ava Duggan, Mia Huebscher, Katerine Osorio, Krina Patel

Northeastern University, Boston, MA, USA

Abstract

The main purpose of this project is to offer an interactive multi-purpose dashboard to inform our target audience about disease trends and symptom-disease correlation. Our dashboard is composed of several tabs that offer unique information based on what the user is looking for. To ensure the dashboard is user-friendly and gives users context, the dashboard first offers a brief explanation of its contents and how a user can benefit from the information presented. The first interactive component allows users to input their symptom/symptoms and choose a prediction algorithm to output a predicted disease and offer recommended precautions. Secondly, the dashboard provides a Sankey diagram that allows users to choose one-to-many diseases to visualize the relationship between diseases and symptoms. Using data that documents people's diseases and their symptoms, this visualization can allow users to see when diseases have identical symptoms. A third visualization allows a user to input a disease and create a heatmap to show the percentage of a population with a given disease in each state. Finally, a disease prevalence map allows users to choose from a list of diseases to display a bubble map of a disease's prevalence, where bubble size correlates to prevalence. As the users hover over the visualization, they can see the exact location, name of the area, and count of the disease. Whether a user would like to learn about a disease relevant to them or disease patterns in general, this dashboard can be a useful tool.

Introduction

The motivation behind our project has evolved as we have continued to refine our goals. Our overarching goal is to provide an interactive dashboard that can empower people of all ages to be proactive about their health by providing a diagnosis based on symptom(s) and other relevant disease information. By providing information on a disease, such as a relationship between diseases and symptoms, and visualizing disease prevalence, the idea is that a user can gain insight into a disease they have or think they might have.

For many people, looking for information on the internet regarding health/disease statistics may be confusing or misleading given that there are many different outlets one can get their information from. People often use the internet as a way to self-diagnose an illness given the symptoms they are experiencing. An example would be when searching, 'Why am I coughing?' The first result reads, 'Tobacco use, postnasal drip, asthma, and acid reflux.'. This can be misleading because one would have to make an educated guess on the presented options, taking into account that many people have poor health literacy. The intention is to offer the most accurate diagnosis possible given a list of symptoms so a person

can take the necessary steps for treatment. This project could serve as a prototype for a potentially more advanced system that could be used at home for anyone and serve to take some burden off the healthcare system. Not all, but many illnesses can be self-diagnosed and only require home remedies for treatment. This would assist to make healthcare workers less strained, so there is more effort and resources going toward people with severe diseases/illnesses.

Disease is a broad term used to encompass disorders in the structure or function of a human and is generally a negative term. However, it is important for users to understand that many conditions fall under the umbrella of 'disease', with many of these being temporary and treatable. We also focus on illness, whereby the model can diagnose a user with a condition that affects the body in several different capacities. Illness can be self-diagnosed, but often disease must be diagnosed by a medical expert. An example of an illness would be the Common Cold or a Migraine, whereas an example of a disease would be Tuberculosis. Information within this dashboard serves to advise people, but not be used as a replacement for professional medical guidance.

Methods

For this project, we pulled data from two different sources. The first dataset is pulled from Kaggle, created by user Pranay Patil. [1] This dataset contains different diseases associated with their respective symptoms. Each disease has multiple entries corresponding to different instances of reported symptoms for that specific disease. We used this dataset to do both our disease prediction and our Sankey diagram. Both of these portions import this dataset from a CSV file into a Pandas data frame to perform machine learning and visualizations on the data.

The second dataset that we used is pulled from the CDC website. [2] This dataset contains information showing the prevalence of different diseases in different states and counties across the USA. It contained data from both 2019 and 2020. We filtered it to only include only the data from 2020 so that it was more recent and we did not double-count any locations. The dataset contained over 1.7 million rows of data. In order to create the disease prevalence diagram we had to extract a longitude and latitude from each location. This allowed us to pinpoint locations on the map showing the counts of different diseases. For the heat map, we had to take a different approach. The data showed the total population of each county along with a percentage of the people there that had a certain disease. We applied a function to each row to find the actual count of people in each county using the population and the percentage. Then we grouped the data by disease and state to find a total count of each disease by state. We then had to import a dataset that showed each state's population in order to find the percentage of each state that had the disease. Once all of this preprocessing was done we were able to create the heat map that showed each state with the percentage of that respective disease.

Analysis

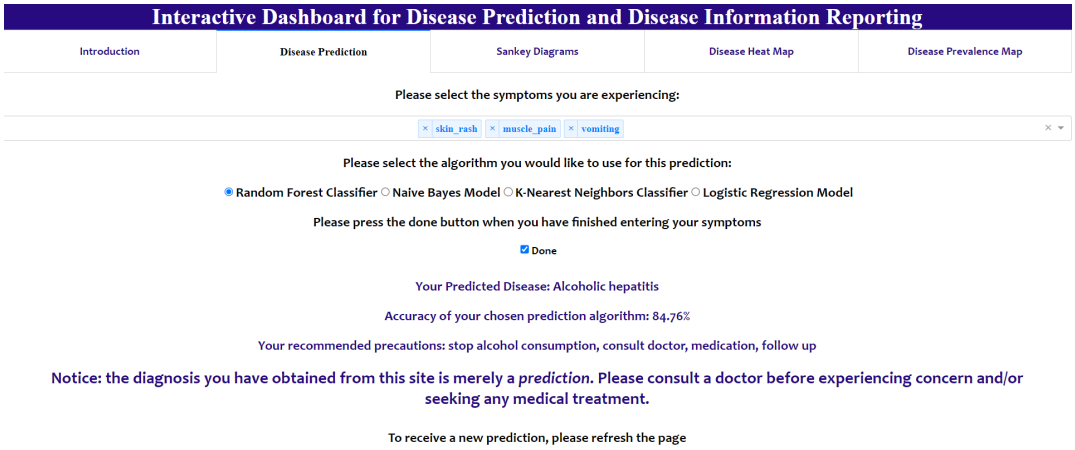

Image 1. Disease prediction section to analyze symptoms and output a possible disease (ALCOHOLIC HEPATITIS)

For our dashboard, we created a disease prediction section that helps with self-diagnosis and a dropdown menu to select a variety of symptoms that a user might be experiencing. As shown above, for image 1, I have selected skin rash, muscle pain, and vomiting as the symptoms, and the user also has the option to choose the desired algorithm to project and analyze the results based on the symptoms. In this case, I choose Random Forest Classifier. As a result, the predicted disease is Alcoholic Hepatitis, with an 84.76% accuracy. After the disease has been predicted, recommended precautions are given to the user depending on the seriousness of the illness. In this case, Alcoholic Hepatitis is a serious disease that requires an immediate end to alcohol consumption, a consultation with a doctor, medications, and a follow-up appointment to ensure improvement. With all predictions, users are given the warning to consult their doctor before making any serious medical decisions.

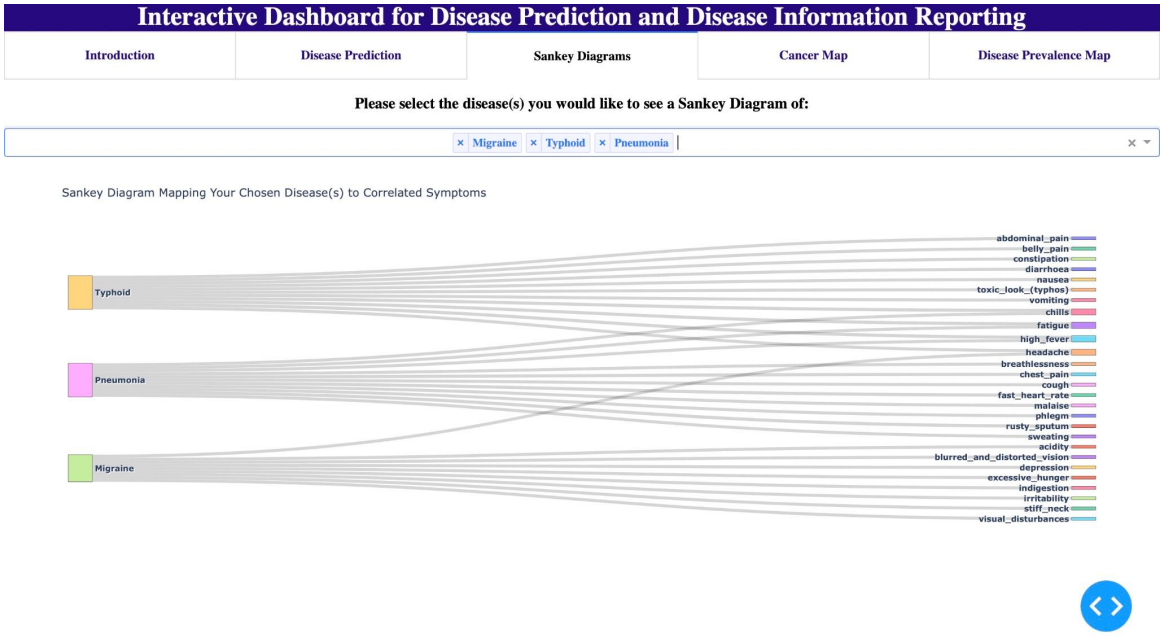

Image 2. Sankey diagram section maps and analyzes the correlation between Typhoid, Pneumonia, and Migraine and their symptoms

The Sankey diagram section is a great tool to visualize the correlation between different diseases based on their symptoms. In image 2, I have selected three different diseases: Migraine, Typhoid, and Pneumonia. The diagram maps the illness on the left side and it portrays the correlated symptoms on the right side. By comparing these diseases we can conclude that typhoid and pneumonia have more common symptoms and those migraine symptoms have a slight interrelationship with Pneumonia and Typhoid with just one common symptom which is headaches. On the other side, Typhoid and Pneumonia have a higher correlation since they have multiple common symptoms such as high fever, fatigue, chills, and headaches.

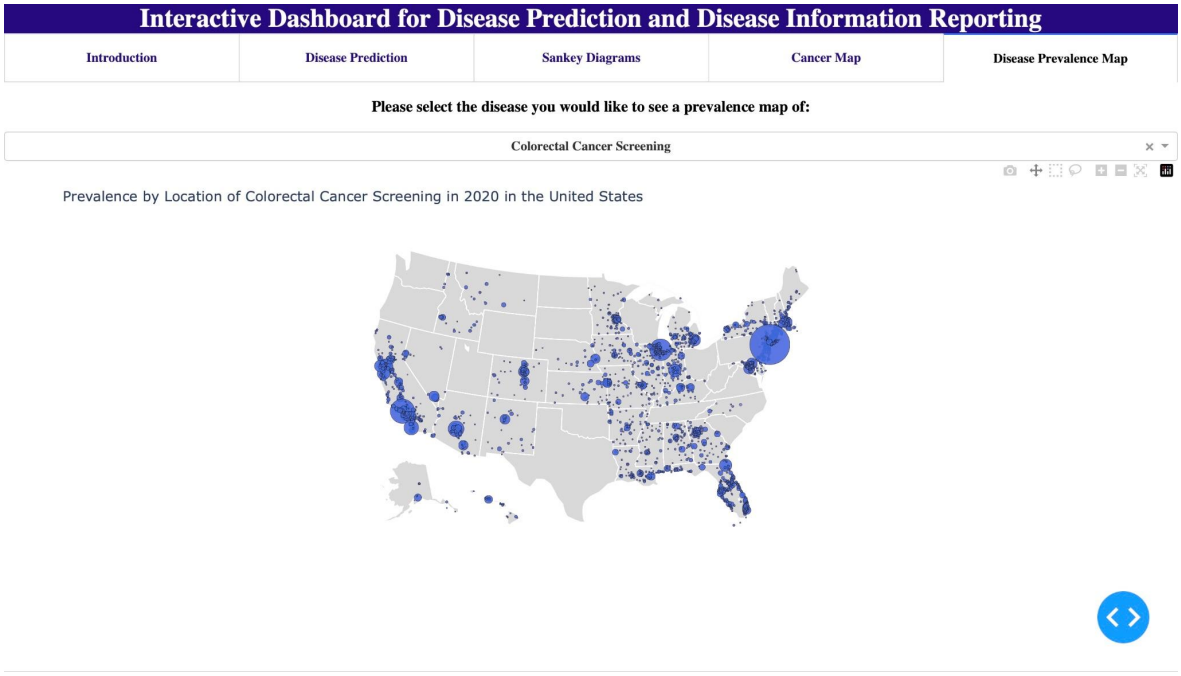

Image 3. Disease prevalence map that portrays areas with Colorectal Cancer Screening cases

The disease prevalence section helps the user to select a variety of illnesses from the dropdown menu. For image 3, I have chosen Colorectal Cancer Screening. The maps give a geographical visualization across different states in the United States that shows Colorectal Cancer Screening cases. The areas with bigger dots represent a higher number of cases in that specific location and where the dots are smaller it represents that number of cases is particularly low. The map also helps to visualize locations where there is a higher percentage of the population tends to have a higher tendency for diseases. In this case, there's a huge dot located on the Northeast side which is where New York is located, and on the west coast where California is located, we can see that we have a lot of cases of colorectal cancer screening cases.

Conclusions

Our project provides users with information on their symptoms based on 2019 and 2020 datasets from the CDC and Kaggle. We are able to provide them with their predicted disease based on their entered symptoms, based on four different algorithms. These include a random forest classifier, naive Bayes model, k-nearest neighbors classifier, and logistic regression model, which each give scores of about 80% accuracy. After providing the user with their predicted disease, we also include suggested precautions they should take afterward. For example, if a user inputted "vomiting" and "stomach_pain" as their symptoms, their predicted disease would be GERD, and they would be recommended to "avoid fatty food, avoid lying down after eating, maintain a healthy weight, and exercise". On the second tab, we include more information on Sankey diagrams. The user can select their desired disease(s) and then see a Sankey displaying diseases(Left) vs. symptoms(Right). That way the user can see all the symptoms associated with as many of their selected diseases. On the following tabs, we are displaying a heat map on the percentage prevalence of any selected disease in the United States. The user can select their desired

disease from a dropdown menu, and then see a different heat map depending on their selection. The user can also hover over different states to see the number of cases as a percentage of each of the states' populations. On the last tab, we have our disease prevalence map displaying the prevalence of any selected disease, in cities across the United States. Cities with a higher count of the selected disease, are shown with a blue bubble which is larger depending on the size of the percentage. The user can hover over each circle to see the city and the case counts of the disease. Our interactive dashboard is able to provide medical information through written and visual displays, along with reference materials about diseases relevant to our users.

Author Contributions

Krina created a Sankey comparing symptoms to disease. She allowed the user to input their choice of symptoms to see all the diseases associated with their inputted disease. She also created another Sankey comparing diseases to U.S. states by prevalence. The user was able to input a disease to see which U.S. states the disease was most prevalent in. These Sankeys were not implemented in our final dashboard. For the report, she completed the conclusion section. Ava initialized the dropdown option for the disease prevalence map so that a user may choose the disease to be displayed. She also made a function to turn the symptom/disease data into a data frame that gives values for diseases and their symptoms. This data was used for the final Sankey, where the user can choose one-to-many diseases from a dropdown to see overlapping relationships with diseases and their symptoms, as well as their value. She added an introduction tab to the disease dashboard to provide a quick overview of its components. For the report, she wrote the abstract and introduction. Katerine created the presentation for the Data Science fair and was one of the presenters. For the report, she worked on the analysis section and added the major insights and main outcomes from the different visualizations on the dashboard. Nick imported and cleaned the datasets to prepare them for analysis and visualization. This was done through Jupyter notebooks mainly using Pandas. He also created the disease prevalence and disease heat map visualizations using this data. For the report, he worked on the methods section. Mia created the dashboard and styled its components. She also produced the prediction capability of the dashboard, added her group members' work to the respective tabs of the dashboard, and implemented the dropdown menu in the Sankey diagram tab. In the introduction tab, she added an image and links to our data sources.

References

[1] Patil, P. (2020, May 24). *Disease symptom prediction*. Kaggle. From https://www.kaggle.com/datasets/itachi9604/disease-symptom-description-dataset

[2] Centers for Disease Control and Prevention. (n.d.). *Places: Local data for Better Health, Place Data 2022 release*. Centers for Disease Control and Prevention. From https://chronicdata.cdc.gov/500-Cities-Places/PLACES-Local-Data-for-Better-Health-Place-Data-202/eav7-hnsx