SMU Data Science - Professor Booth Project 4 - Group 3 Spring 2024

Group Members: Shanara Hawkins, Emmanuel Presley, Ivette Reese, & Jon Unger

From Data to Diagnosis: Enhancing Patient and Clinician Insights Using HEALTHhOUSE, a Disease Discovery Tool

Data is everywhere. The amount of data points at our disposal for consideration and analysis is expansive, immeasurable, and uniquely complex. For over two decades, the digital age has experienced a heightened focus on ensuring quality initiatives, valuable insights, proper allocation of resources, valid, reliable metrics, and the ability to perform a broad array of tasks and functions in a short amount of time. Without industry-specific tools and skills (programming languages, coding, data visualizations, cognitive academic language, evaluation, and analysis), efficiency, productivity, and creating sustainable outcomes would be a daunting and cumbersome task for data scientists. Data science is broad and applicable to the success of countless industries, such as finance, education, and transportation. It has revolutionized the healthcare industry and the various approaches to treatment, prognosis, and disease prevention. The number of successful healthcare initiatives in recent years can be attributed to advances in technology such as machine learning techniques, artificial intelligence tools, and predictive modeling.

Technology has proven to be a powerful tool aiding clinicians in deciding a patient's probability of acquiring specific health conditions and diseases. The likelihood that one will experience either an illness, a sickness, or other medical condition is inevitable. With this in mind, our group selected healthcare as our topic of interest to explore the ways predictive analytics could be used to accurately determine which disease a person would likely have given the symptoms they selected in the algorithm. The TV show "House" served as inspiration for our data science project, HEALTHhOUSE. Combining our knowledge of web application development, machine learning, coding, and algorithms we created a model that predicts diseases based on a pattern of symptoms. The goal of our predictive model is to help users in identifying potential illnesses more efficiently, mirroring Dr. House and his medical team's diagnostic prowess. In addition, we used Tableau to create our visualizations and dashboards to present a concise visual representation of the data in a manner that accurately displayed our targeted data points and was easy for our audience to understand.

Two datasets were used to train our machine-learning model and create data visualizations in Tableau. Both datasets were pulled from the Kaggle website. The first dataset is titled, Disease Prediction Using Machine Learning, and access to prior code is available for this dataset by clicking the link. This dataset includes 2 CSV files that contain 132 symptoms a person may experience, and the possible prognosis based on the selected symptoms. There are a total of 42 diseases to which the symptoms may be attributed. The primary use for this dataset will cover the Machine Learning components used to create HEALTHhOUSE's Discovery Tool Recommender. The second dataset is titled, Disease Prediction through Symptoms. Access to prior code is available for this dataset by clicking this link. This dataset includes 1 CSV file that contains 13 columns of information including the disease, symptoms, weight, height, intensity, severity, age, gender, BMI level, region, and season. The primary use for this dataset will be to create our HEALTHhOUSE Tableau dashboard/visualizations.

We preprocessed the data using Python and read the CSV files into the Pandas Library. We initiated the cleaning process by identifying and managing missing values in the dataset. We dropped rows and columns with missing values, removed duplicate entries, corrected spelling errors, updated and renamed columns, and created categories to organize the information contained in the dataset. More specifically, on the Machine Learning side, several symptoms were spelled incorrectly including *scurry* and *diarrhea*. We changed those to reflect the correct spelling and symptoms, which resulted in scurvy and diarrhea. Next, we compared both datasets to ensure there were exact matches of the named diseases. The first dataset had more diseases than the second dataset and both had unique named diseases that were not present in the other.
Therefore, we manually reviewed each dataset to locate diseases that were present in both. Although using a function in Pandas to merge the datasets and drop the diseases that were not present in both would have been a quicker process, our team chose to use all diseases present in the first dataset to yield a more robust application of diseases on the machine learning side. Therefore, to ensure the integrity and validity of the data visualization component of the project, the second dataset was filtered to include diseases that were present in both datasets, which resulted in a final total of 17 diseases.

Expanding on the Tableau-specific data cleaning process, first, we conducted data engineering before uploading the OHAS (Occupational Health and Safety) CSV dataset into Tableau. There were no duplicate or null values, but there were missing values present in the dataset.

After preprocessing the data, a total of 2,114 rows remained from the original count of rows, which was 2,129. The columns were unchanged and included a total of 13 columns. As mentioned previously we kept 17 out of 48 common diseases for further data analysis, after having a collaborative discussion and examining both datasets to identify the diseases they had in common. The final, cleaned dataset displayed no duplicate or null values and consisted of 218 rows and 13 columns which remained unchanged. We analyzed a total of 17 diseases and 130 symptoms with categorically unique values.

The following are some statistical findings uncovered by extracting insights and patterns from the data. First, the average height ranged from 62 inches to 76, the average weight from 92 to 600 pounds, and the average BMI from 17.1 to 53. The age ranged from 16 to 57 years old. Noted within the dataset was an imbalance in gender classification with a total of 97 males and 121 females. For clarification of the verbiage and acronyms present in the dataset, the columns named Disease and Symptoms CUIs relate to the disease CUIs, which is an acronym for Concept of Unique Identifier (CUIs) in the Unified Medical Language System (UMLS).

We created three dashboards in Tableau. The first dashboard, Symptoms and Diseases Overview has two bar charts. The bar charts display the types and count of diseases, and the most prevalent symptoms by disease type. The second dashboard, Regional Disease Trends, consists of a stacked bar chart and a circle chart. The stacked bar chart includes the count of diseases by fourth regions: northwest, northeast, Southwest, and Southeast. The circle charts display the number of diseases by region and gender. The third dashboard, Severity Factors Comparison, displays a scatter plot, and a box plot. The box and scatter plot displays high, low, and medium disease severity classifications. The scatter plot compares the severity of the diseases by BMI and weight. The box plot establishes the gender differences by age in each disease severity.

Following are additional results combining multiple data points where insights and patterns are used to provide insights that could be valuable to improving healthcare outcomes for the affected groups of individuals. Hypothyroidism appeared as the most predominant disease, with the highest count of 23, while migraine was the lowest at 7. Shortness of breath, asthenia, and fever were the most prevalent symptoms experienced.

In taking a closer look at geographical location and how it affects healthcare, it was noted that the Southeast region showed the highest prevalence of diseases, with a total of 75 counts of diseases. In contrast to the Southeast region, the Southwest region had a lower prevalence, with a total count of 44 diseases. Notably, chronic diseases were the most predominant, affecting 95 out of 218 individuals and underscoring the need for region-specific healthcare initiatives and strategies. In addition, males suffered more chronic diseases in nearly every region with a total count of 52 in comparison to females with a total count of 43. Except in the Southwest region, females suffered more chronic diseases with a total count of 11 while their male counterparts had a total of 9.

The impact of health metrics and demographic factors on disease severity was also noted. BMI and weight were found to affect disease severity, with higher values correlating to higher severity. The higher the BMI, the individual is either overweight (BMI 25-29.9) or obese (BMI 30 or higher), the higher the disease severity. Males were found to suffer from a higher severity of diseases than females, and this trend was observed to affect younger males more. These findings highlight the importance of health metrics and demographic factors in understanding disease severity.

To conclude the Tableau component of the analysis, the prevalence of diseases varied across regions and chronic diseases were the most predominant. This emphasizes the importance of region-specific healthcare strategies. Finally, gender differences in disease prevalence and severity were observed and important for analysis. Further analysis indicated that males generally experience higher severity in diseases in comparison to females, which was particularly pronounced in younger males. Healthcare and data science play an important role together in terms of enhancing outcomes for patients, preventing future illnesses, and detecting diseases based on various factors like age, gender, and geographic location. Data-driven approaches that are implemented and assessed for reliability and accuracy can improve the health of patients, increase overall longevity, and reduce early-onset diseases by identifying the most common characteristic risk factors and associated symptoms.

Bias, limitations, and implications are present concerning healthcare, and the impacts were highlighted by a variety of influential factors. The first Limitations and biases present included the quality and availability of inclusive datasets. As with many other forms of data sources, information may be missing or groups of people may be left out of the conversation, leading to inaccurate, incomplete, and biased. In addition, privacy regulation and interoperability issues such as HIPPA compliance laws present challenges in data gathering, particularly if individuals decide to opt out of allowing their private healthcare information to be shared without their consent. Such restrictions can create additional barriers to appropriate treatment, diagnosis, and individualized care plans if there is not enough data to form accurate conclusions. Also, there are future implications that must be considered when dealing with topics as delicate, personal, and specialized as patient and clinician healthcare. For instance, healthcare professionals may get so accustomed to using algorithms and other diagnostic tools, that they may come to over-rely on them to assist with making decisions for their patients, especially if they contend with a heavy workload of people who require immediate, emergency, or specialized care.

Data is everywhere. From the moment you wake up to the moment you go to sleep, data is continuously being created and collected. Even the most common daily tasks such as monitoring the temperature in your home, shopping for household products online, and the rate of attendance at various places in your local community rely on data. Data science influences advancements, and innovations, and creates meaningful changes based on reliable outcomes leading to a better quality of life. As progress in machine learning algorithms continues and new technologies emerge, concurrently, we can also ensure improvements in efficacy, validity, and accuracy. Predictive modeling working in tandem with personalized medicine has its advantages such as empowering healthcare practitioners to deliver targeted treatment and interventions that will enhance patient healthcare outcomes and satisfaction.