Weekly Report: Predictive Healthcare

Summary of Activities:
This week, I focused on acquiring and analysing a dataset for the disease prediction capstone project. The primary objective was to locate a suitable dataset and perform initial data exploration and pre-processing.

1. Dataset Acquisition:
I spent a considerable amount of time searching for appropriate datasets for disease prediction. After thorough research, I selected the Columbia's Disease - Symptom Knowledge Database (link) as it appeared to be comprehensive and well-suited for the project's objectives. This table serves as a repository of disease-symptom associations, derived through an automated process utilizing textual discharge summaries of patients admitted to New York Presbyterian Hospital in 2004. The first column displays the diseases, followed by the frequency of their mention in discharge summaries, along with associated symptoms. The associations for the 150 most prevalent diseases were determined based on these notes, with symptoms ranked according to their strength of association. The methodology involved utilizing the MedLEE natural language processing system to extract UMLS codes for diseases and symptoms from the notes, followed by statistical analysis leveraging frequencies and co-occurrences to establish associations.

Disease	Count of Disease Occurrence	Symptom
UMLS:C0020538_hypertensive disease	3363	UMLS:C0008031_pain chest
		UMLS:C0392680_shortness of breath
		UMLS:C0012833_dizziness
		UMLS:C0004093_asthenia
		UMLS:C0085639_fall
		UMLS:C0039070_syncope
		UMLS:C0042571_vertigo
		UMLS:C0038990_sweat^UMLS:C0700590_sweating increased
		UMLS:C0030252_palpitation
		UMLS:C0027497_nausea
		UMLS:C0002962_angina pectoris
		UMLS:C0438716_pressure chest
UMLS:C0011847_diabetes	1421	UMLS:C0032617_polyuria
		UMLS:C0085602_polydypsia
		UMLS:C0392680_shortness of breath
		UMLS:C0008031_pain chest
		UMLS:C0004093_asthenia
		UMLS:C0027497_nausea
		UMLS:C0085619_orthopnea
		UMLS:C0034642_rale
		UMLS:C0038990_sweat^UMLS:C0700590_sweating increased
		UMLS:C0241526_unresponsiveness
		UMLS:C0856054_mental status changes
		UMLS:C0042571_vertigo
		UMLS:C0042963_vomiting
		UMLS:C0553668_labored breathing
UMLS:C0011570_depression mental^UMLS:C0011581_depressive disorder	1337	UMLS:C0424000_feeling suicidal
		UMLS:C0438696_suicidal
		UMLS:C0233762_hallucinations auditory
		UMLS:C0150041_feeling hopeless
		UMLS:C0424109_weepiness
		UMLS:C0917801_sleeplessness
		UMLS:C0424230_motor retardation
		UMLS:C0022107_irritable mood
		UMLS:C0312422_blackout
		UMLS:C0344315_mood depressed
		UMLS:C0233763_hallucinations visual
		UMLS:C0233481_worry
		UMLS:C0085631_agitation
		UMLS:C0040822_tremor
		UMLS:C0728899_intoxication
		UMLS:C0424068_verbal auditory hallucinations
		UMLS:C0455769_energy increased
		UMLS:C1299586_difficulty
		UMLS:C0028084_nightmare
		UMLS:C0235198_unable to concentrate
		UMLS:C0237154_homelessness
UMLS:C0010054_coronary arteriosclerosis^UMLS:C0010068_coronary heart disease	1284	UMLS:C0008031_pain chest
		UMLS:C0002962_angina pectoris

Fig. 1: Raw Dataset

2. Dataset Analysis and Pre-processing:
Upon acquiring the dataset, I conducted a basic analysis to understand its structure, variables, and potential relevance to our project. The dataset was initially in raw format, containing diverse medical information.

The data extraction process involved copying website data in .html format and saving it into an Excel file for further analysis. Basic data cleaning, column segmentation, and string formatting were carried out within Excel. Subsequently, the Excel sheet was imported into a Jupyter Notebook for additional processing.

Data pre-processing steps included:

1. Correcting spelling mistakes in disease or symptom names and their codes.
2. Removing duplicate symptoms associated with the same or similar disease names.
3. Separating multiple symptoms listed within the same row.
4. Eliminating irrelevant codes assigned to diseases and symptoms.
5. Compiling a comprehensive list of all symptoms.
6. Assigning Boolean values (0 or 1) to each symptom for every disease, indicating its presence or absence.
7. Adding the corresponding disease in the final column.

The cleaned data for analysis looked like this:

	Disease	Heberden's node	Murphy's sign	Stahli's line	abdomen acute	abdominal bloating	abdominal tenderness	abnormal sensation	abnormally hard consistency	abortion	...	vision blurred
0	Alzheimer's disease	0	0	0	0	0	0	0	0	0	...	0
1	HIV	0	0	0	0	0	0	0	0	0	...	0
2	PneumocystisÃ-Â cariniiÃ-Â pneumonia	0	0	0	0	0	0	0	0	0	...	0
3	accidentÃ-Â cerebrovascular	0	0	0	0	0	0	0	0	0	...	0
4	acquiredÃ-Â immuno-deficiency syndrome	0	0	0	0	0	0	0	0	0	...	0

5 rows × 405 columns

Fig. 2: Cleaned Final Dataset

There are a total of 149 unique diseases and 405 unique symptoms in this dataset.
Below graph shows a spread of disease with their frequencies in the dataset:

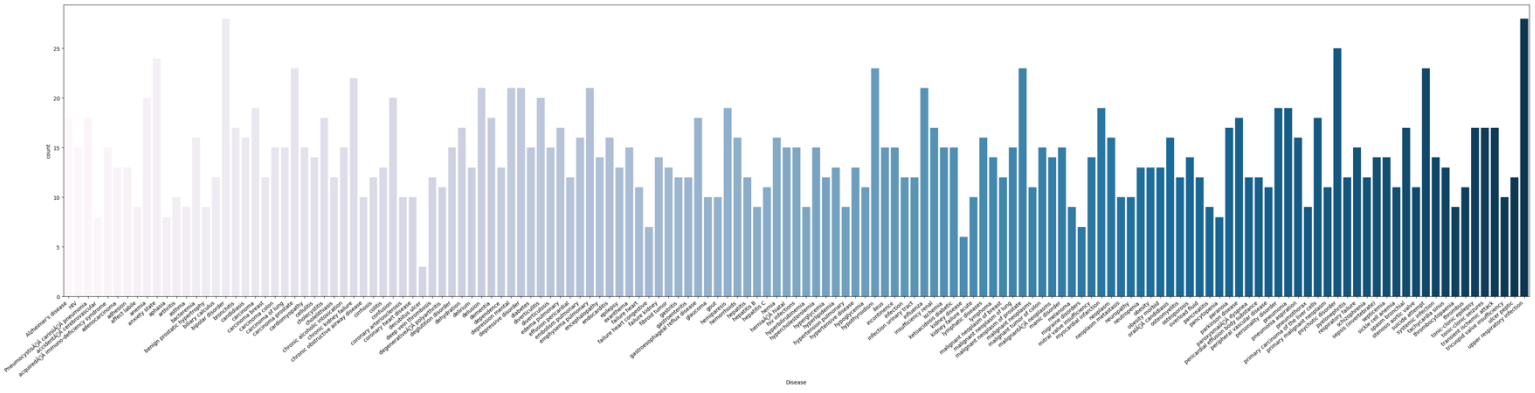

Fig. 3: Spread of Diseases in the dataset

From the graph we can see that the top 5 diseases are: “Bipolar Disorder”, “Upper Respiratory Infection”, “psychotic disorder”, “anxiety state”, “malignant neoplasm of prostate”.
Below are there counts:

```
Disease
bipolar disorder                 28
upper respiratory infection      28
psychotic disorder               25
anxiety state                    24
malignant neoplasm of prostate   23
                                 ..
aphasia                           8
migraine disorders                7
failure heart congestive          7
kidney disease                    6
decubitus ulcer                   3
Name: count, Length: 149, dtype: int64
```

Fig. 4: Count of Symptoms

3. Common Symptoms Identification:
Following data pre-processing, I analysed the dataset to identify common symptoms present across various diseases. This analysis aimed to provide insights into the prevalence and distribution of symptoms within the dataset. By understanding the most frequent symptoms, we can better comprehend the dataset's characteristics and potential patterns.

The top 10 common symptoms identified are:

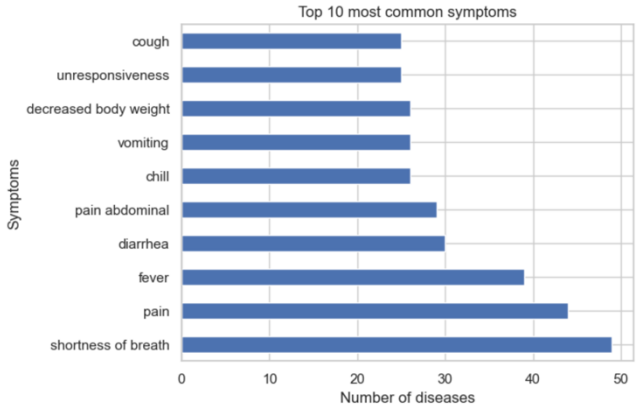

Fig. 5: Top 10 identified symptoms

Below are some plots indicating the presence or absence of some symptoms in diseases:

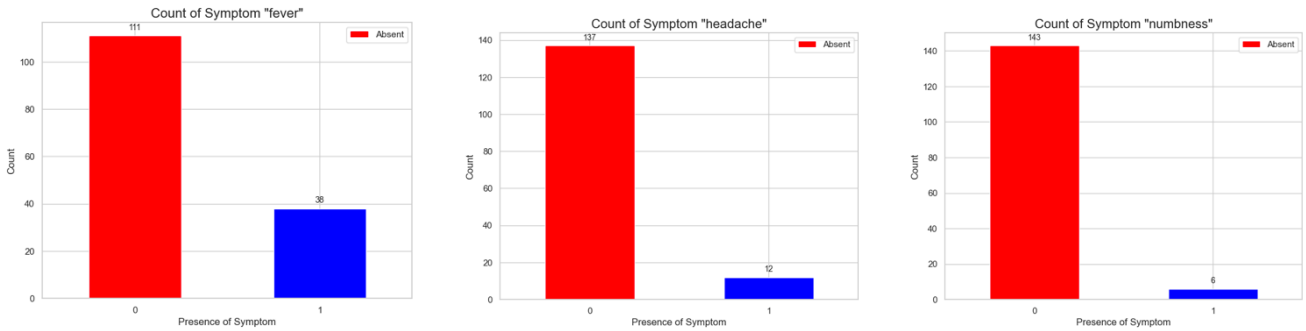

Fig. 6: Symptoms presence in disease

4. Challenges Faced:
While the dataset selection process was challenging, the Columbia dataset emerged as a promising choice due to its richness and relevance to the project. However, data pre-processing presented its own set of challenges, particularly in structuring the data into a suitable format for analysis.

5. Conclusion:
In conclusion, this week's progress laid a solid foundation for the disease prediction capstone project. With the dataset acquired, cleaned, and basic analysis conducted, I plan to advance to the next stages of model development and evaluation. Furthermore, I will be following professor’s suggestion to read the clinical notes and try to extract relevant information from the same for my analysis. I will explore more dataset for the same and try working on that pre-processing in the coming week.x