Weekly Report

During the spring break, I worked on the MIMIC - III Dataset and created a database from using 5 tables out of a total of 30 tables in the dataset acquired. A representation of the schema that I will be using this project is given below:

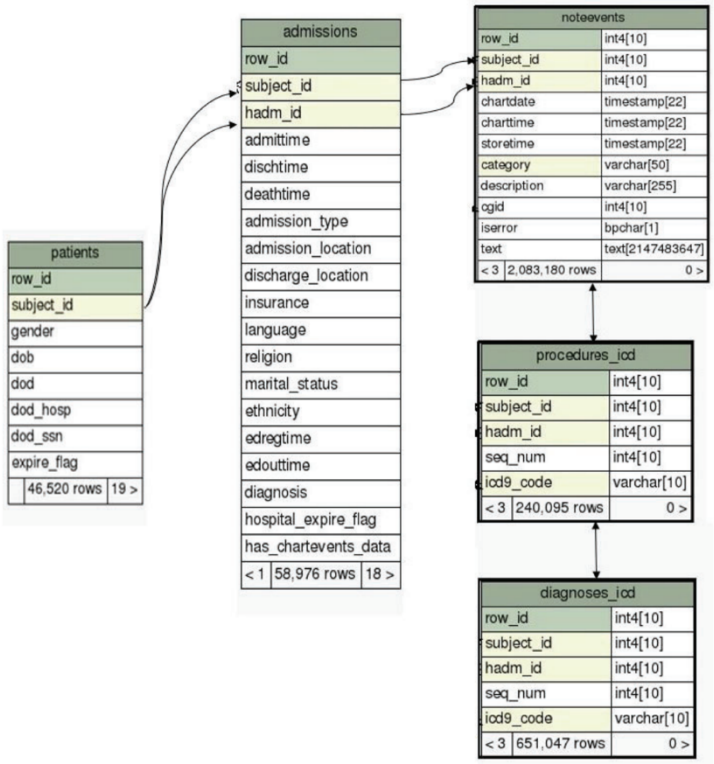

Fig.1: The schema of the tables from MIMIC

After all the table merges, the final dataset obtained contains about 3.9 Million records. But after I realized the potential problem with the size of the data and with the constraints of time and resources, I stopped using the entire note events, instead further filtering of the diagnosis and procedures tables was performed by me. I filtered out the textual data and corresponding diagnosis picking 4000 evenly distributed data for analysis.

I picked up 2 columns: Text and Diagnosis, ‘Text’ column consists of the clinical notes and ‘Diagnosis’ is the disease related to that clinical note:

TEXT	DIAGNOSIS
Admission Date: [**2151-7-16**] Dischar...	RT LOWER LOBE PNEUMONIA
Admission Date: [**2151-7-16**] Dischar...	RT LOWER LOBE PNEUMONIA
PATIENT/TEST INFORMATION:\nIndication: Aortic ...	RT LOWER LOBE PNEUMONIA
PATIENT/TEST INFORMATION:\nIndication: Endocar...	RT LOWER LOBE PNEUMONIA
Atrial fibrillation with a slow ventricular re...	RT LOWER LOBE PNEUMONIA

Fig. 2: Data For Analysis

Since I cannot work on the textual data directly, I pre-processed the data to extract symptoms from text, and remove stopwords. First, I removed punctuation and digits, and converted the text to lowercase using regular expressions. Then following a publicity available Disease-Symptom Knowledge Database (https://people.dbmi.columbia.edu/~friedma/Projects/DiseaseSymptomKB/) I made a list of about 400 commonly present symptoms in diseases. Then I built a function to check if a given word is a symptom by comparing it to the list of symptoms. The dataset after conversions contained the 'Diagnosis' and 'Text' in the below format:

	DIAGNOSIS	Keywords
0	RT LOWER LOBE PNEUMONIA	['cough', 'photophobia', 'lesion', 'asymptomat...
1	RT LOWER LOBE PNEUMONIA	['cough', 'photophobia', 'lesion', 'asymptomat...
2	RT LOWER LOBE PNEUMONIA	['cough', 'photophobia', 'lesion', 'asymptomat...
3	RT LOWER LOBE PNEUMONIA	['cough', 'photophobia', 'lesion', 'asymptomat...
4	RT LOWER LOBE PNEUMONIA	['cough', 'photophobia', 'lesion', 'asymptomat...

Fig.3: Diagnosis and Symptoms Keywords

While I did this manually, I also came across a library that could extract the key information from medical notes, called ScispaCy. It is a specialized version of spaCy that is trained specifically on scientific and biomedical text, which makes it ideal for processing medical text. I tried using it in my project to get the below output:

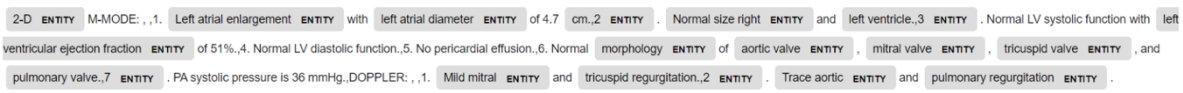

Fig.4: ScispaCy Usage

It helped me in tagging entities. However, along with medical terms, it also tagged generic entities. So, I went ahead with my data generated by text transformation and extraction.

Once I had this for my analysis, I split the 'Keywords' column and performed one hot encoding. The final dataset for analysis looked like this:

	DIAGNOSIS	shortness of breath	dizziness	asthenia	fall	syncope	vertigo	sweat	sweating increased	palpitation	...	feces in rectum	prodrome	hypoproteinemia	alcohol binge episode	abdomen acute	air fluid level	catching breath	large-for-dates fetus	immol
0	hypertensive disease	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
1	hypertensive disease	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
2	hypertensive disease	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
3	hypertensive disease	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
4	hypertensive disease	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	

5 rows × 405 columns

Fig. 5: Diagnosis and Symptoms after One hot Encoding

After all these transformations, the data was now ready to be trained on model. So, in my next step, I built a Decision Tree Classifier.

45]:

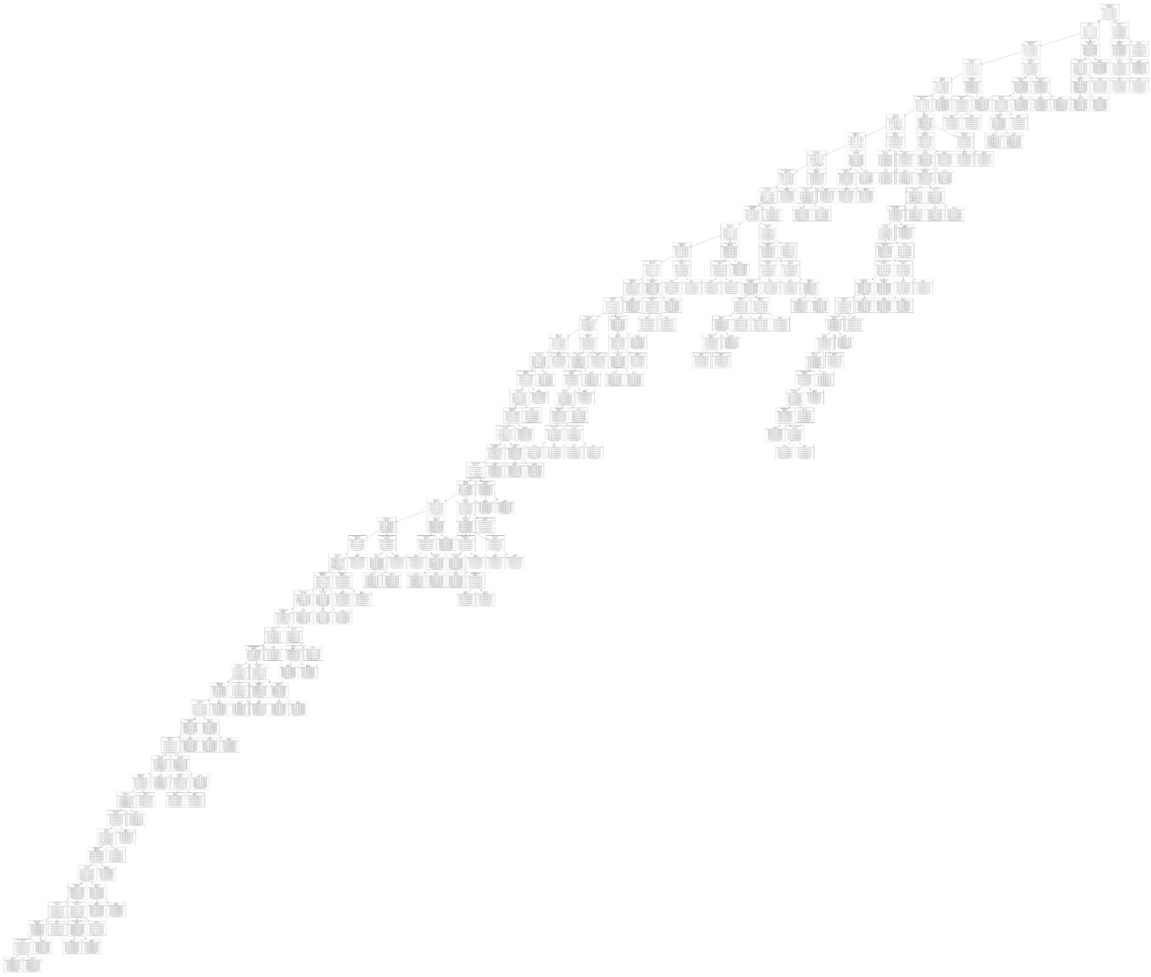

Fig. 6: Decision Tree

Once the model ran successfully, I compared actual vs predicted values as below:

```
Pred: Alzheimer's disease
Actual: Alzheimer's disease

Pred: HIV
Actual: HIV

Pred: Pneumocystis carinii pneumonia
Actual: Pneumocystis carinii pneumonia

Pred: accident cerebrovascular
Actual: accident cerebrovascular

Pred: acquired immuno-deficiency syndrome
Actual: acquired immuno-deficiency syndrome

Pred: adenocarcinoma
Actual: adenocarcinoma

Pred: adhesion
Actual: adhesion

Pred: affect labile
Actual: affect labile

Pred: anemia
...

Pred: upper respiratory infection
Actual: upper respiratory infection
```

Fig. 7: Prediction vs Actual