Investigating Emergency Department Statistics to Prevent Hospitalizations

Ava Downey

University of Hartford

December 10, 2021

Abstract:

Learning about the specifics of how, when, and why people end up in the emergency department can help to tackle many common problems faced in healthcare seen today. This includes the upgrade of medical quality, reduction of patient complaints, and less waste of medical resources. It can also unveil hidden medical knowledge through looking at the correlation and association of apparently independent variables. This new knowledge can be used to combat many common complaints and issues in the emergency department such as increased wait time, ambulance diversion, reduced staff morale, and adverse patient outcomes.

Though this correlation does not seem to be linear, there are commonalities between different demographics and the injuries they receive. Finding these correlations will help to make emergency department care more beneficial, as well as can be used to help prevent injuries that can be avoided.

Introduction:

People end up in the emergency department because they are getting hurt, so finding the root cause of why people are injuring themselves can help to greatly increase people's experience in the emergency department. By determining the source of different types of injuries, measures can be taken to help prevent them. This in turn will provide a more positive experience for people who do end up in the emergency department because decreased foot traffic will lead to lower wait times, happier and more attentive staff, and less death.

This type of data mining is often used to determine insurance rates for different people because people who are more likely to get hurt and end up in the emergency department will have to pay higher insurance premiums than someone who will most likely not get hurt. Insurance companies look only at demographics such as age, sex, and race, and not other factors such as fire involvement, body part injuries, and other more dependent variables. It is simpler to look at overall how likely a certain sex or age range is to end up getting hurt rather than how they are specifically getting hurt, which is what differentiates this project from what has previously been done.

The goal of this project is to find correlations between what is causing people to be hurt. This can include the demographics often studied such as age, sex, and race, but it also extends out to more specific and dependent variables such as body parts, location, fire involvement, and specific injuries. By learning the predispositions of how people get hurt, preventative measures can be taken to make sure that these unnecessary injuries do not occur, thus keeping people out of the emergency room.

Background:

Data mining data from emergency department data is not a new concept and has been done to establish insurance rates, reduce hospital overcrowding, and determine better treatment among others. One such paper, "The effects of demographic characteristics and insurance arrangement on the utilization of hospital emergency rooms" [1] talks about usage of emergency department services among different demographic groups. Through her research, she concluded that females, children and young adults, and ethnic minorities are the most likely to visit the emergency room for non-emergency services. The root of this issue is lack of proper resources such as outpatient clinic services, reliable insurance, and efficiency of resources for those privately insured.

This misuse of the emergency department can also lead to crowding which has significant negative impacts on the patients and staff. This can take the form of increased wait time, ambulance diversion, reduced staff morale, death of patients, and more, making it something that will make a strong impact if able to be minimized [2]. In order to combat this, analysis of patient age, previous history, month of the year, day of the week and time of the day can be used to predict future admissions which can be used to combat the above issues [3]. One paper, "The Necessity of Data Mining in Clinical Emergency Medicine; A Narrative Review of the Current Literature", sums up why data mining in emergency department datasets is so crucial stating, "data mining used in medical related research to explore the reduction of patient complaints which arise from insufficient and improper treatment… will upgrade the medical quality and also save the waste of medical resources" [4]. In learning about trends in the emergency room, not only money can be saved, but also peoples lives. Educating the public about what is most likely to get them hurt depending on their demographics can lead to a better experience in th emergency department for both patients and staff.

Methodology:

Both supervised and unsupervised algorithms were used to help formulate conclusions from the data. The unsupervised algorithms were used first to find general trends in the data to learn about what to look into further. The first algorithm used for this purpose was a correlation matrix. A correlation matrix takes the features in a dataset and ranks their correlation between each other. In the case of this project, all of the features in the preprocessed dataset were scaled from 0 to 1, then ran through the algorithm to produce the result which was used to help determine the next steps in the project.

Histograms were also used to help visualize the data before running supervised algorithms. The histograms help to show how the data is distributed to highlight potential bias in the data, as well as certain feature values that are more prevalent in the data for other reasons. For this project, histograms were created for the age, sex, race, location, diagnosis, disposition, and body part features.

The final unsupervised algorithm used was the K-means algorithm. K-means clustering shows how two different features in a dataset are related. Clusters will appear at specific variables that have correlations to each other. The amount of clusters depends on how many are specified in the algorithm and should be the number of clusters that create a meaningful visual. This algorithm provided a good basis as to what to look deeper into for the supervised algorithms because supervised algorithms will only have a high accuracy if there is some correlation between the variables. A flowchart describing the process can be seen below in figure one.

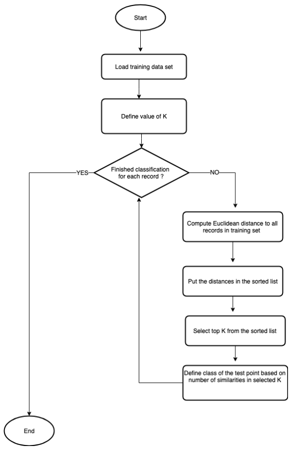

Start
Load training data set
Define value of K
Finished classification for each record ?
YES
NO
Compute Euclidean distance to all records in training set
Put the distances in the sorted list
Select top K from the sorted list
Define class of the test point based on number of similarities in selected K
End

Figure One [5]

The decision tree algorithm was the first supervised algorithm used for this project. A decision tree is able to find nonlinear correlation between several variables. It takes several independent variables, such as location and diagnosis to determine a dependent variable such as disposition. This is represented in Figure Two. A high accuracy will lead to a high correlation between the variables being looked at. This algorithm is a good choice for this project because the correlation between the data cannot be represented linearly.

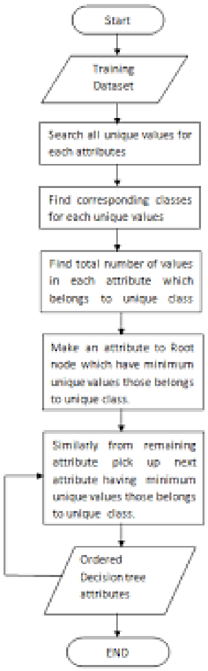

Start
Training Dataset
Search all unique values for each attributes
Find corresponding classes for each unique values
Find total number of values in each attribute which belongs to unique class
Make an attribute to Root node which have minimum unique values those belongs to unique class.
Similarly from remaining attribute pick up next attribute having minimum unique values those belongs to unique class.
Ordered Decision tree attributes
END

Figure Two [6]

Random forest was the next logical step, because it is very similar to the decision tree algorithm. The random forest algorithm also creates a decision tree, but instead of looking at the entire dataset at once, it looks at smaller subsections to create a more accurate decision. This oftentimes results in a higher accuracy than a regular decision tree. This is simply shown in Figure Three.

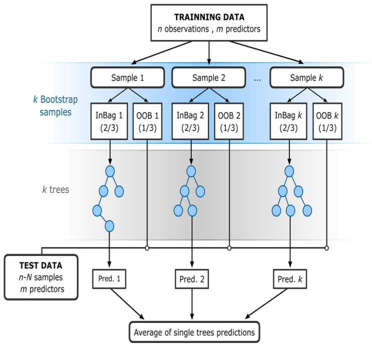


Figure Three [7]

The last algorithm looked at for this project uses natural language processing. The algorithm used for this was the spaCy named entity recognition (NER) algorithm. This algorithm extracts and labels different words from a string in order to better understand key elements in a string of text which helps to better understand large datasets such as the one used for this project. This is represented in Figure Four.

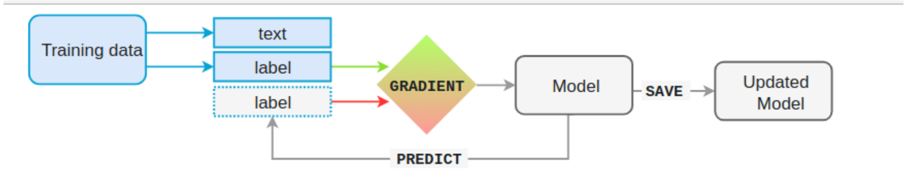


Figure Four [8]

Data:

This project uses a dataset from Kaggle called How People Get Hurt which details demographics about who ends up in the Emergency Department, as well as why they are there. This dataset is compiled from the National Electronic Injury Surveillance System (NEISS) and includes data from January 1, 2016, to December 31, 2020. It includes features such as Treatment Date, Age, Sex, Race, Body Part, Diagnosis, Disposition, Location, Product, Fire Involvement, Narrative, Stratum, PSU, and Weight.

Preprocessing this data included several steps, the first of which being deleting empty features such as Drug Involvement, and Alcohol Involvement. The next step was to delete empty values in columns that were otherwise populated. Since the dataset includes nearly two million different rows representing different patients, these values can just be deleted. The original dataset is also already label encoded, the specific contents of each cell can be left alone in the preprocessing stage. However, the data needs to be transformed into a format that is able to be read by the python programs, meaning that all numbers were transformed to be float64 data types, and all strings of text were transformed to be object data types. The dataset can then be split into a training set and a testing set to allow for supervised learning algorithms.

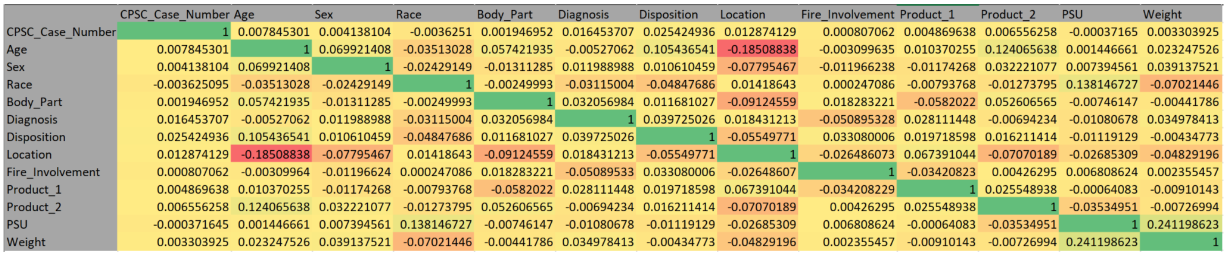
Experiments:

This project was completed using the language Python in Jupyter Notebook through the Visual Studio Code IDE. Python is an ideal language for this project because of the large number of packages and resources available for data mining projects, as well as the simple syntax. Jupyter Notebooks is very helpful for this project as well because of the flexibility to combine code along with visuals and plots, making it easy to test code while working.

In order to best use the dataset, it was split into a training and a testing dataset. The training set is ⅔ of the original dataset, while the testing set is ⅓ of the original dataset. This is used for supervised algorithms such as decision tree, random forest, and knn that need the testing set to validate the outcome of the training dataset on the algorithm. For the correlation matrix, and histograms, the full original dataset was used because there is no validation needed. It would have been ideal to use the full dataset for the named entity recognition algorithm, but unfortunately, due to lack of compute, only a smaller subsection of about 0.5% of the data was able to be used.

Many different python libraries were used to interpret the data for this project including pandas, sklearn, matplotlib, seaborn, graphviz, pydotplus, spacy, and word cloud. These libraries allow for the project to be completed efficiently, because, without the libraries, many of these algorithms would not have been able to have been implemented in the time frame given. By following the docs for the libraries, the algorithms used for the project were able to be simply implemented allowing more time to draw conclusions from the data.

Results:



	CPSC_Case_Number	Age	Sex	Race	Body_Part	Diagnosis	Disposition	Location	Fire_Involvement	Product_1	Product_2	PSU	Weight
CPSC_Case_Number	1	0.007845301	0.004138104	-0.0036251	0.001946952	0.016453707	0.025424936	0.012874129	0.000807062	0.004869638	0.006556258	-0.00037165	0.003303925
Age	0.007845301	1	0.069921408	-0.03513028	0.057421935	-0.00527062	0.105436541	-0.18508838	-0.003099635	0.010370255	0.124065638	0.001446661	0.023247526
Sex	0.004138104	0.069921408	1	-0.02429149	-0.01311285	0.011988988	0.010610459	-0.07795467	-0.011966238	-0.01174268	0.032221077	0.007394561	0.039137521
Race	-0.003625095	-0.03513028	-0.02429149	1	-0.00249993	-0.03115004	-0.04847686	0.01418643	0.000247086	-0.00793768	-0.01273795	0.138146727	-0.07021446
Body_Part	0.001946952	0.057421935	-0.01311285	-0.00249993	1	0.032056984	0.011681027	-0.09124559	0.018283221	-0.0582022	0.052606565	-0.00746147	-0.00441786
Diagnosis	0.016453707	-0.00527062	0.011988988	-0.03115004	0.032056984	1	0.039725026	0.018431213	-0.050895328	0.028111448	-0.00694234	-0.01080678	0.034978413
Disposition	0.025424936	0.105436541	0.010610459	-0.04847686	0.011681027	0.039725026	1	-0.05549771	0.033080006	0.019718598	0.016211414	-0.01119129	-0.00434773
Location	0.012874129	-0.18508838	-0.07795467	0.01418643	-0.09124559	0.018431213	-0.05549771	1	-0.026486073	0.067391044	-0.07070189	-0.02685309	-0.04829196
Fire_Involvement	0.000807062	-0.00309964	-0.01196624	0.000247086	0.018283221	-0.05089533	0.033080006	-0.02648607	1	-0.03420823	0.00426295	0.006808624	0.002355457
Product_1	0.004869638	0.010370255	-0.01174268	-0.00793768	-0.0582022	0.028111448	0.019718598	0.067391044	-0.034208229	1	0.025548938	-0.00064083	-0.00910143
Product_2	0.006556258	0.124065638	0.032221077	-0.01273795	0.052606565	-0.00694234	0.016211414	-0.07070189	0.00426295	0.025548938	1	-0.03534951	-0.00726994
PSU	-0.000371645	0.001446661	0.007394561	0.138146727	-0.00746147	-0.01080678	-0.01119129	-0.02685309	0.006808624	-0.00064083	-0.03534951	1	0.241198623
Weight	0.003303925	0.023247526	0.039137521	-0.07021446	-0.00441786	0.034978413	-0.00434773	-0.04829196	0.002355457	-0.00910143	-0.00726994	0.241198623	1

Figure Five

Looking first at the correlation matrix of the dataset in Figure Five, this lack of linear correlation is very apparent. The green cells represent a strong positive linear correlation, the yellow cells represent no correlation, and the red cells represent a negative correlation. The strongest correlation was between PSU and Weight, but it still is not strong enough to amount to much statistical significance.

Histograms help to show how the data is laid out, to help determine possible bias as well as get a general idea of how the data is laid out. Several of the most helpful histograms are shown below in Figure Six. Looking at the race histogram, it can be determined that there could be bias in the dataset against non-white people. This might be due to the demographic of the location where the survey originates from, white people injuring themselves more, or non-white peoples statistics not being recorded in the dataset. The age histogram is also interesting because it shows how people are more likely to injure themselves earlier on in their life rather than later on, again possibly because older people might be passing away. The 200-250 age range represents children under the age of two years old, because in that stage of life, every month is statistically significant. The disposition histogram is also interesting because it shows that most people end up leaving the Emergency Department and not requiring additional serious treatment.

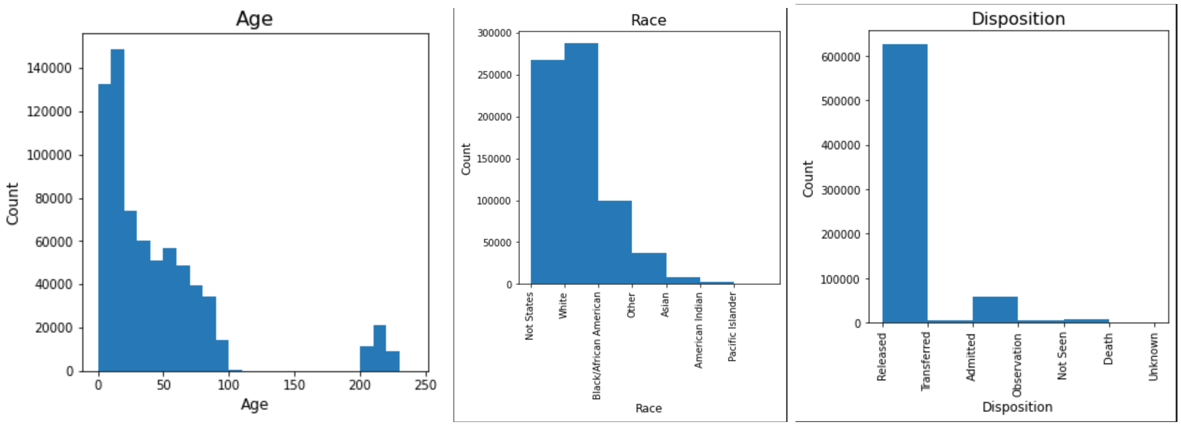


Figure Six

K-means algorithm was also used to get a feel for the dataset. The most helpful graph produced from this algorithm is shown in Figure Seven. It shows the relationship between the location of injuries and the disposition of the patient. The algorithm shows clusters that injuries occurring in a home, farm, ranch, street, or other public property are more likely to lead to serious injury such as death than the other locations. Injuries occurring at a sports center, or school also seem to have no strong correlation towards any disposition.

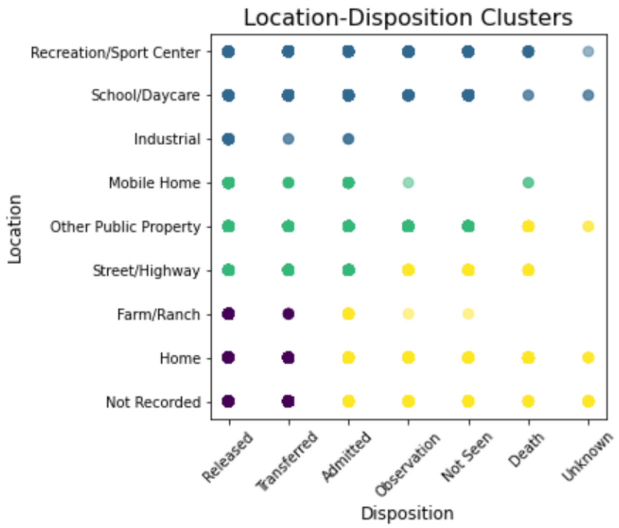


Figure Seven

The decision tree predicting disposition from diagnosis and location can be seen in Figure Eight. The orange nodes represent patients that are released from the Emergency Department after being treated, while the green nodes represent the other outcomes including being transferred, admitted, put under observation, or dying. This decision tree helps to reflect the disposition histogram in Figure Six because most cases lead to an outcome of being released. This decision tree places burns, aspiration, and ingestion as some of the most common diagnoses to lead to death. Location does not seem to play a statistically significant role in disposition.

Figure Eight

This decision tree for was one of the most accurate models in this project, with an accuracy of 0.8904, a precision of 0.83, a recall of 0.89, and an f1-score of 0.84. The random forest of the same variables produced the same classification report.

Figure Nine represents a decision tree looking at how age, sex, and race affect the burn type of a patient. Since burns appear to be the leading cause of death, it is important to know who is the most at risk for this diagnosis. The blue nodes represent a chemical burn, and the orange nodes represent a thermal burn. Chemical burns appear to be more common than thermal burns which are also reflected in the histograms of the dataset. People over the age of 12.5 seem to be more likely to burn themselves, though the type of burn does not seem dependent on age. It also appears that white people tend to receive thermal burns more often than any other race, though this might be due to bias in the data.

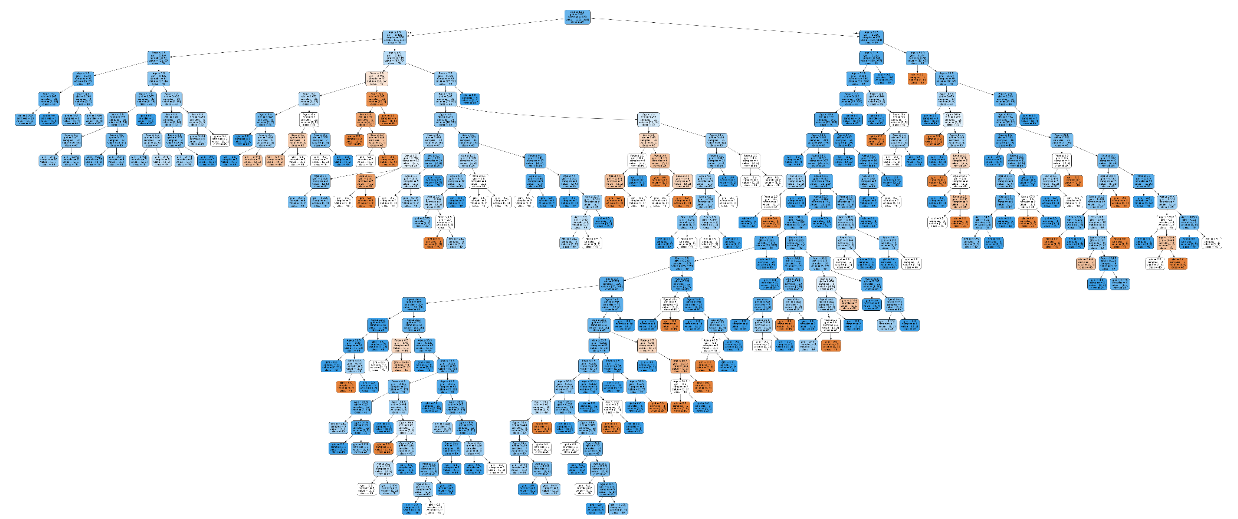


Figure Nine

The decision tree for the burn type to disposition decision tree had an accuracy of 0.7137, a precision of 0.72, a recall of 0.71, and an f1-score of 0.72. The random forest model for the same variables had an accuracy of 0.8676, a precision of 0.87, a recall of 0.87, and an f1-score of 0.87.

The last algorithm looked at for this project was the spaCy NER algorithm. From looking at the word clouds able to be extracted from this algorithm, many conclusions about time and location can be made. Figure Ten shows the common words relating to date and time, while Figure Eleven shows common words relating to location. It can be concluded that many people injure themselves at night, as well as that many people end up in the Emergency Department because of falling. People tend to injure themselves at school or at home, and often injure body parts such as their head, knee, and ankle. This is all supported by the histograms looked at earlier. As seen in Figure Eleven, medical jargon gets misclassified by the algorithm.

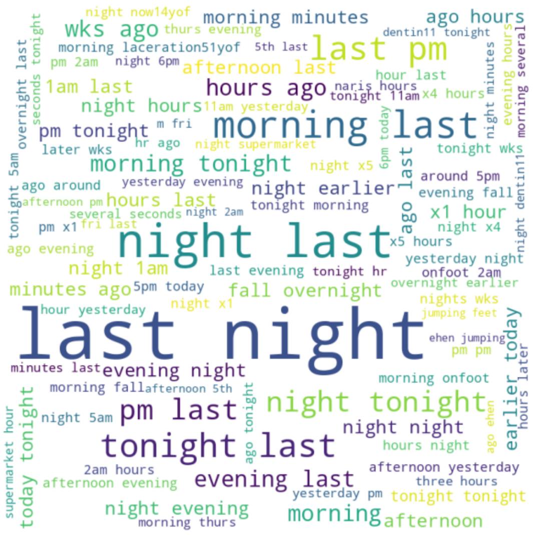


Figure Ten

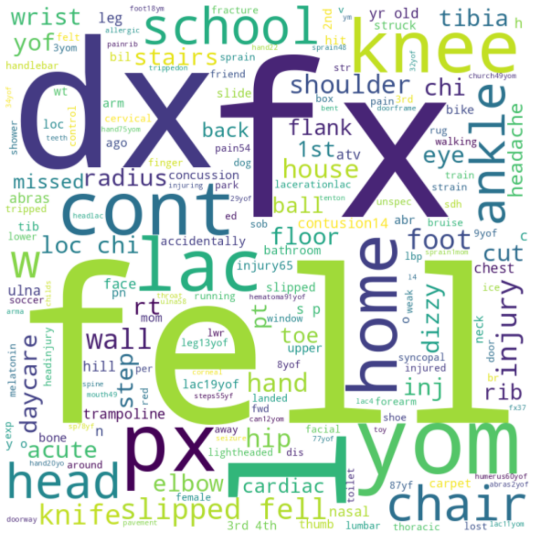


Figure Eleven

Conclusion:

After running several models on this dataset, the main conclusion found is that the relationship between many of the variables including age, sex, race, location, and etc. are not linearly correlated to the final diagnosis or disposition of a patient. However, through looking at non-linear models such as decision tree, and random forest, it becomes apparent that there is some non-linear correlation. Though the accuracy is not perfect, it is above the threshold for statistical significance. From these decision trees, it can be concluded that there are correlations between demographics, location, and injury to disposition and diagnosis. The most likely reason for someone to die in the Emergency Department would be if they have suffered from burns, aspiration, or ingestion of something inedible. The demographic most likely to experience this diagnosis are white people over the age of 12.5.

Looking into the future, adding a second dataset to supplement what is currently available in the original dataset might help to draw more meaningful conclusions. In this project, it was difficult to find the conclusions that were found, and even then, they are only barely statistically significant. It will also be helpful to spend more time running algorithms such as decision tree or random forest on more factors into disposition as it seemed to have the strongest correlation in the data. It would also be interesting to be able to look further into how drug and alcohol use affect emergency department rates as well. This dataset also has a lot of room to study natural language processing through the scope of the narrative portion of the dataset-- something that was only briefly explored in this project. It will also be more meaningful if the whole dataset is able to be included in this model, rather than the very small subset done in this project. In general, the conclusions drawn from this project can be greatly improved through more research and trial on the dataset in the future.

References:

[1] Sharp, T. D. (2002). *The effects of demographic characteristics and insurance arrangement on the utilization of hospital emergency rooms* (Order No. 1411441). Available from Healthcare Administration Database; ProQuest Dissertations & Theses Global: The Humanities and Social Sciences Collection. (238074194). Retrieved from http://libill.hartford.edu:2048/login?url=https://www.proquest.com/dissertations-theses/effects-demographic-characteristics-insurance/docview/238074194/se-2?accountid=11308

[2] Graham, Byron & Bond, Raymond & Quinn, Michael & Mulvenna, Maurice. (2018). *Using Data Mining to Predict Hospital Admissions From the Emergency Department*. IEEE Access. 6. 1-1. 10.1109/ACCESS.2018.2808843.

[3] G.Abhiram , T.P.Anithaashri. (2020). *Hospital Admissions Using Data Mining*. European Journal of Molecular & Clinical Medicine. From https://ejmcm.com/pdf_4533_44508d4ed900a9de385d60bafad4e67b.html

[4] Parva, E., Boostani, R., Ghahramani, Z., & Paydar, S. (2017). The Necessity of Data Mining in Clinical Emergency Medicine; A Narrative Review of the Current Literatrue. *Bulletin of emergency and trauma*, *5*(2), 90–95.

[5]https://towardsdatascience.com/introduction-to-knn-machine-learning-algorithm-by-experiment-on-khmer-handwriting-classification-66a64652a02c

[6] https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.695.2653&rep=rep1&type=pdf

[7]https://www.researchgate.net/figure/The-flowchart-of-random-forest-RF-for-regression-adapted-from-Rodriguez-Galiano-et_fig3_303835073

[8] https://spacy.io/usage/spacy-101

[9] https://www.kaggle.com/jpmiller/how-people-get-hurt