Prescription review: California Prescribed Opioids Model

Elias Fedai

Efedai9@yahoo.com

Abstract

In the recent years our society has seen a detrimental number of increased deaths from opioid related overdoses. The addiction of these type of drugs is a serious issue in our society and if not taking as a serious national crisis can have some very extreme negative ramification to our country's welfare. While in some instances these drugs may be introduced to people through an illicit manner. There are those instances where people become addicted to these drugs when prescribed from a medical provider and its these instances that this research project will be geared towards, particularly those in my home state of California. In this presentation there will be several key insight data displayed, displaying trends and patterns in areas where opioids are being prescribed. Application of a classification model could be a useful tool, in that taking the data provided we can determine and classify the severity level of total prescribed opioids by county.

Author Keywords

Data Science; Classification; Opioid crisis.

Introduction

Determining and establishing a functioning classification model uses several key analytical applications such as statistical methods, machine learning, exploratory analysis, and ability to derive the right questions in order to solve the problems needing to be addressed. The opioid problem in our country has been an apparent issue since the 90’s when the introduction of prescription opioids became available for doctors to prescribe [1]. Along with the thousands of people who have died due to opioid overdose, the financial burden on the country stemming from healthcare cost, addiction treatment, and lost productivity cost the United States somewhere in the ball park of 78.5 billion dollars a year [1]. In this analysis I will using a data set retrieved from the state of California Department of Justice Office of the Attorney General. In this project I would like to investigate several questions; In which regions are there increased number of opioids prescribed? Do these increased numbers subside in the same region in the state (north, east, west, south)? Between the years of 2015 to 2020 are the rates decreasing or increasing? While evaluating the data I will be able to not only answer these questions but also find new detailed insight into the data.

Why Classification model?

In this particular instance while there is data to indicate the total quantitation of prescribed medication for each county for a total of six years. I am going to want to establish a model in which we can accurately classify the level of risk associated with the number of prescribed opioids (low, moderate, high-risk group) for any of the counties in the state of California and to see if the model can accurately classify different levels. In this analysis I will be using a total of multiple different models (K-nearest neighbors, Linear Discriminant Analysis, Gaussian Naïve Bayes, Support Vector machine (linear, radial, polynomial), Random Forest, and a neural network; multi-layer perceptron. I will evaluate each of these models and see how they perform in respect to the data provided.

Variables

This data retrieved from the California Attorney Generals webpage consists of seven total columns (year, state, county, age group, run date/time, patient count, and population) and consists of 1740 total rows. In this particular analysis the date/time column will not be used. I did generate a couple of new columns; opioid percent which displays the total percent of prescribed opioids with respect to total population, a label column to determine the risk level for each row of the data-frame, and from this I generated a number label that I will use to create a classification model. During exploratory analysis it was observed in viewing the distribution of the categorical category 'label' we can see that there is a slight skew in the data favoring low risk (Fig 1.1). When determining the age spread of those prescribed opioids it became apparent that majority of these individuals were between the age of 45-64yrs (Fig 1.2)

Fig 1.1(Risk group by total count)

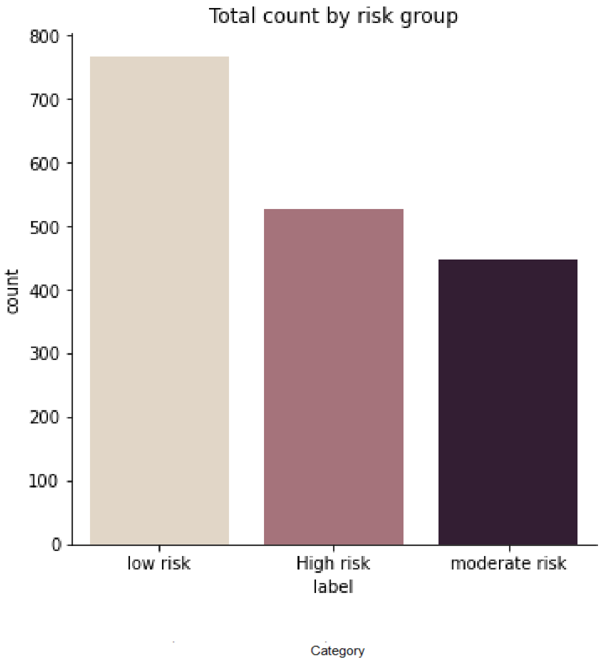
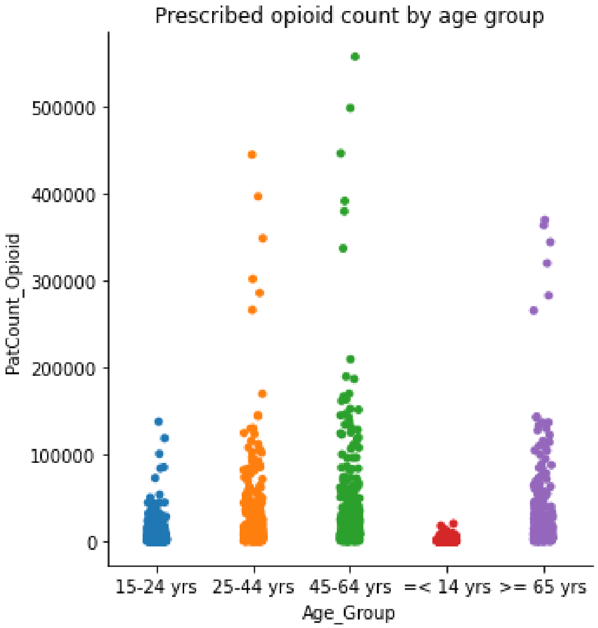


Fig 1.2 (Distribution of Categorical Variable)



Methods

During this study, several methods will be used in assessing the best possible outcome in which model will be the best fit for this study. It would appear as of now that my best approach on tackling this project is to use a multiclass classification model (low risk, moderate risk, high risk). Platforms that will assisting with the methods to be used are; Python, Jupyter Notebook, and excel. In this assignment I ended up electing five different models to compare; k-nearest neighbors, Decision Tree, Gaussian Naïve Bayes, Linear Discriminant Analysis. Each analysis will be split into a 75/25, and put through a ten-fold cross validation to ensure optimal performance of the model. Additionally, because I am dealing with a skewed data set in both the categorical variables and numerical variables. I will perform a normalization technique (min/max scaler). All analysis will be evaluated, models will be compared in both their overall metric evaluation and compared against each other.

Exploratory Analysis

In this data analysis, it was evident that I would need to explore the columns of value in order to determine if there was any added value there to be presented but also to determine the best route in preparing to create a classification model. During this exploratory analysis, it was discovered that Los Angeles led all other counties in California in most prescribed opioids from the years of 2015 to 2020 (Fig 2), but this chart can be misleading because rather than viewing the total prescribed amount it would be more beneficial to see the total prescribed amount in respect to total population (Fig 3). The reason for this is that cities with high population are at a disadvantage in that it would be obvious that they would have higher values because of the sheer number of people. When we look at the total percent it becomes evident that the more risker counties are different than those initially listed in Fig 2. In Fig 3 we can see that Lake County actually has the highest prescribed opioids in respect to total population followed by Calveras, Butte, Tuolumne, and Yuba County which is rather interesting. In addition, I decided to break down this total percent by each year to see if the county names would change for each year. Upon evaluating this it was interesting to see that Lake County remained number one for each respective year. Additionally, when taking all the counties and dividing the state into three general regions (Northern cali, SoCal, and Central cali) the region that averaged the highest percentage displaying close to 4% was Northern Cali and Southern and Central Cali where both averaging close to 3.3% and 3.1/% respectively.

Fig 2 (Total Prescribed Opioids by County)

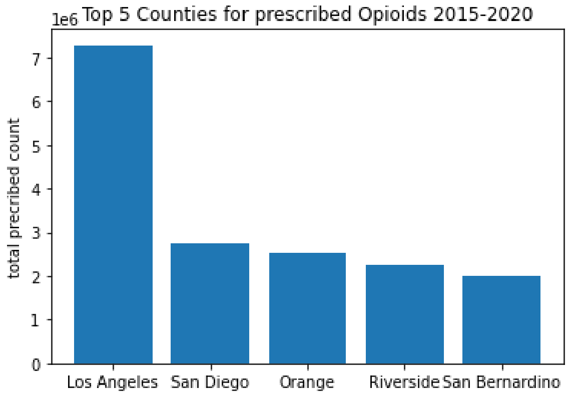
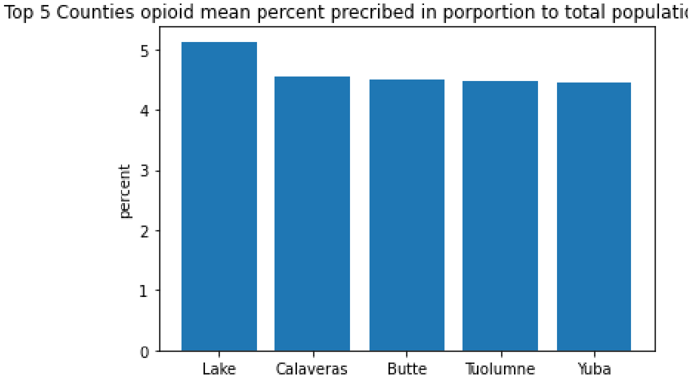


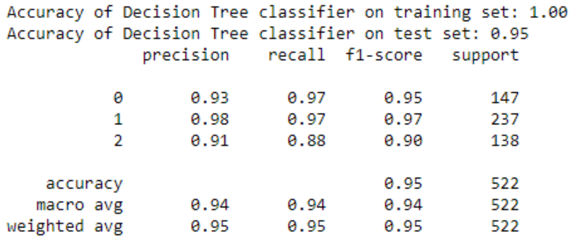
Fig 3 (Percent precribed opioids in respect to population)



Results

Once analysis was completed on the multiple models; k-nearest neighbors (KNN), Decision Tree, Gaussian Naïve Bayes, Linear Discriminant Analysis, Support Vector machine, Random Forest, and MLP. Metrics that were used to measure the performance of the model was accuracy. Going through each model it was surprising to see those vast differences in the performances of each respective model. Decision tree (Fig 3.0) displayed a 95% classification accuracy while maintaining desirable F1 score and precision values. KNN (Fig 3.1) displayed 98%, also displaying both desirable values for F1 and precision. Linear Discriminant Analysis displayed an accuracy of 58% which was clearly a less desirable result and also displayed subpar values for both F1 and precision. Gaussian Naïve Bayes which may actually be more suited for text data classification also displayed sub performance outcomes. Displaying an accuracy of 52% and an undesirable F1 and precision values. Additionally, I also committed to Support Vector machine. In this model I elected to try three different variations of this classifier. Polynomial, radial, and linear (Fig 3.2). Of the three different types the only one that yielded desirable outcome was the linear kernel, which gave a perfect outcome. The multi-layer perceptron model (Fig 3.3) also yielded very undesirable outcome displaying an accuracy less than 30%.the last model Random Forest (Fig 3.4) did display a more desirable accuracy value of 96% and worthy F1 and precision values. In my overall analysis of the different models. Those models who scored an overall accuracy of 95% are noteworthy for possible future instances using additional data. If we look at the overall classification report for each model the performance of each becomes very evident. (Fig 4.0).

Fig 3.0 (Classification report Decision Tree)

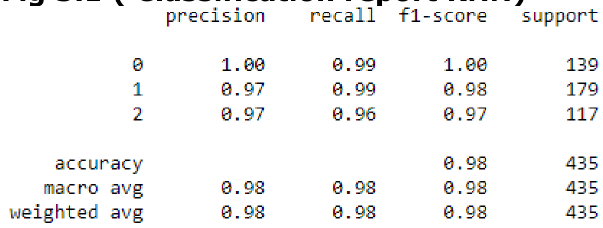


```
Accuracy of Decision Tree classifier on training set: 1.00
Accuracy of Decision Tree classifier on test set: 0.95
              precision    recall  f1-score   support

           0       0.93      0.97      0.95       147
           1       0.98      0.97      0.97       237
           2       0.91      0.88      0.90       138

    accuracy                           0.95       522
   macro avg       0.94      0.94      0.94       522
weighted avg       0.95      0.95      0.95       522
```

Fig 3.1 (Classification report KNN)

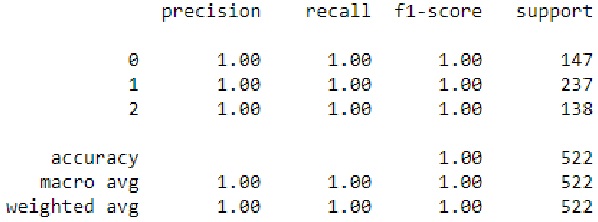


```
              precision    recall  f1-score   support

           0       1.00      0.99      1.00       139
           1       0.97      0.99      0.98       179
           2       0.97      0.96      0.97       117

    accuracy                           0.98       435
   macro avg       0.98      0.98      0.98       435
weighted avg       0.98      0.98      0.98       435
```

Fig 3.2 (Classification report SVM Linear)

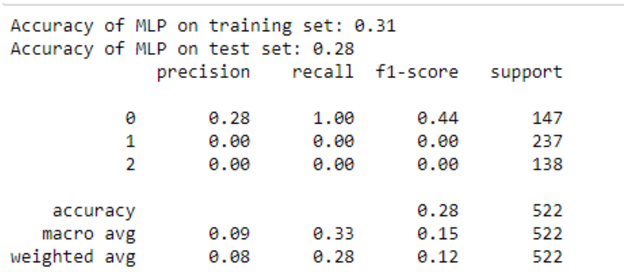


```
              precision    recall  f1-score   support

           0       1.00      1.00      1.00       147
           1       1.00      1.00      1.00       237
           2       1.00      1.00      1.00       138

    accuracy                           1.00       522
   macro avg       1.00      1.00      1.00       522
weighted avg       1.00      1.00      1.00       522
```

Fig 3.3 (Classification report MLP)

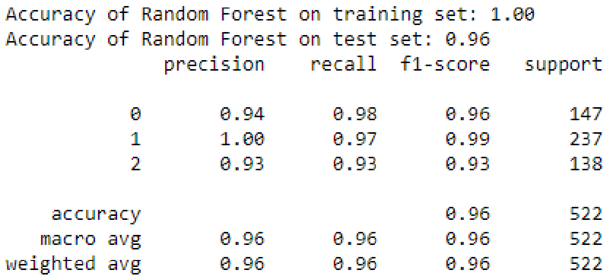


```
Accuracy of MLP on training set: 0.31
Accuracy of MLP on test set: 0.28
              precision    recall  f1-score   support

           0       0.28      1.00      0.44       147
           1       0.00      0.00      0.00       237
           2       0.00      0.00      0.00       138

    accuracy                           0.28       522
   macro avg       0.09      0.33      0.15       522
weighted avg       0.08      0.28      0.12       522
```

Fig 3.4 (Classification report Random Forest)

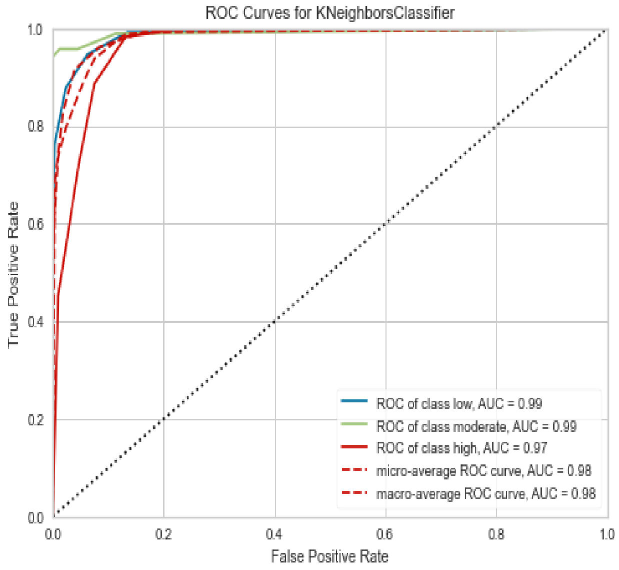


```
Accuracy of Random Forest on training set: 1.00
Accuracy of Random Forest on test set: 0.96
              precision    recall  f1-score   support

           0       0.94      0.98      0.96       147
           1       1.00      0.97      0.99       237
           2       0.93      0.93      0.93       138

    accuracy                           0.96       522
   macro avg       0.96      0.96      0.96       522
weighted avg       0.96      0.96      0.96       522
```

Fig 4.0 (ROC/AUC results for KNN)



Conclusion

Since the implementation of prescribing opioids for medicinal purpose, we have seen our country shift to a somewhat dependency on these types of medications. While opioids do carry significant medicinal purposes, the fact that these types of drugs are extremely addictive does carry many risks some of which are irreversible. Some examples of these risks are injuries due to medication intoxications, medication abuse, overdose, and many more. Due to these serious risks the outcome that may lie from these actions are but not limited to; death, long extensive and expensive medical cost, job loss, and risk of injuring others. Which in turn is extremely detrimental to not only the families that will suffer horrible from this, but also our state that will be susceptible to the social and economic ramifications that this type of disease carries. In analyzing the data for the years of 2015 to 2020 the State of California has shown a positive trend in reducing prescription opioids, but the need to continue on-going monitoring is essential as the benefits of having such a system outweighs the risk significantly. I do believe rather than creating a system to penalize or render shame, fault, accusations to our great medical providers. I believe that we must keep them informed on the numbers and allow them to make the correct judgement decisions as they are a trusted intricate part of our community. In comparing the multiple classification models in this report. A couple of the models did yield some promising outcomes and may be of interest. These models consisted of; Decision Tree, KNN, SVM, and random forest. While these models performed really well on the data set provided. I would recommend that we reanalyze these models are much larger sample pool as the amount they were tested consisted of 1740 rows. In conclusion, the analysis did display the great progression our state is showing in the recent years, but has also shown that there still some opportunities that we can take advantage of. Assessing and addressing this problem is not only an issue of the individuals suffering from this disease but is the responsibility of all of us as a community.

Acknowledgements

I would like to thank all the researchers and authors in the review of the opioid crisis in our country and data science thank you for the great insight in the subject matter.

References

1. Opioid overdose crisis (2021). Retrieved September 1, 2021 from Opioid Overdose Crisis | National Institute on Drug Abuse (NIDA)

2. Assessment of Racial/Ethnic and Income Disparities in the Prescription of Opioids and other Controlled medications in California (2021). Retrieved on September 1, 2021 from Assessment of Racial/Ethnic and Income Disparities in the Prescription of Opioids and Other Controlled Medications in California | Health Disparities | JAMA Internal Medicine | JAMA Network

3. CURES Statistics, California Schedule II-V Drug Acquistion, Prescription and dispensation public statistic (2021). Retrieved on September 1, 2021 from CURES Statistics | State of California - Department of Justice - Office of the Attorney General

4. Evidence on Strategies for addressing the opioid epidemic- Pain management and the opioid epidemic (2021). Retrieved on September 2, 2021 from Evidence on Strategies for Addressing the Opioid Epidemic - Pain Management and the Opioid Epidemic - NCBI Bookshelf (nih.gov)

5. CMS Opioid Prescribing (2021). Retrieved on September 2, 2021 from CMS Opioid Prescribing | CMS

6. Opioid Crisis Statistics (2021). Retrieved on September 2, 2021 from Opioid Crisis Statistics | HHS.gov

7. Medicare Part D Opioid Prescribing Mapping Tool (2021). Retrieved on September 2, 2021 from Medicare Part D Opioid Prescribing Mapping Tool | CMS

8. Overdose Deaths Rates (2021). Retrieved on September 2, 2021 from Overdose Death Rates | National Institute on Drug Abuse (NIDA)

9. Prescriber Information (2021). Retrieved on September 2, 2021 from Prescriber Information (ct.gov)

10. Opioid Overdose (2021). Retrieved on September 3, 2021 from Opioid overdose (who.int)

11. Systematic Evaluation of State Policy Interventions Targeting the US Opioid Epidemic (2021). Retrieved on September 2, 2021 from Systematic Evaluation of State Policy Interventions Targeting the US Opioid Epidemic, 2007-2018 | Psychiatry and Behavioral Health | JAMA Network Open | JAMA Network