One Class Classification for Overdose Death Detection

Ram Nalluri

8/5/22

<Inspirit Ai 1:1 Mentorship Program>

Abstract

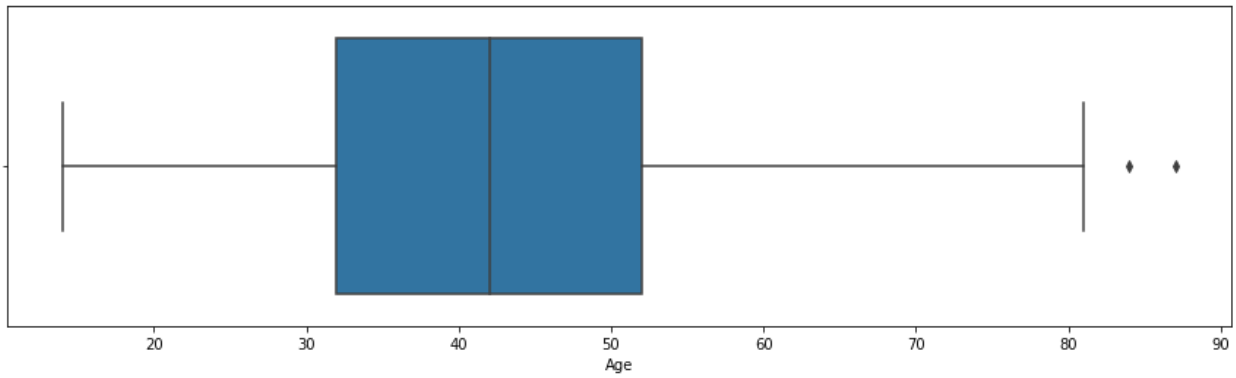
This paper describes a model intended to predict drug overdose deaths given a certain amount of information. This was done by analyzing and formatting data of drug overdose deaths in the US during 2012 to 2018. The data was then put into a One Class Classification model which would be able to predict the likelihood of a drug overdose death. Three different models were used returning an f1 score of 0.836, 0.849, and 0.85 respectively.

1. Introduction

In the past year alone there were an estimated 107,622 drug overdose deaths in just the United States [1]. With such an incredible amount of deaths from just this, I thought it would be beneficial to create a model to predict who is at risk of drug overdose deaths, help prevent as many potential deaths as possible. This can be achieved by placing extra checkups in order to ensure that those who are classified at risk stay safe. By doing so, this could prevent any addictions from developing and stop them early on, before they have any long-term effects. This research aims to determine the most important factors in determining drug overdose deaths and use that information to predict the likelihood of a person to fall victim to a drug overdose death. The first step was obtaining a useful dataset, removing its irrelevant data, and filling in its missing values. Next, the data was reformatted to be analyzed and given labels and used to train a model. Lastly, the model determined whether or not people were likely to fall victim to drug overdose deaths or not and have its accuracy determined.

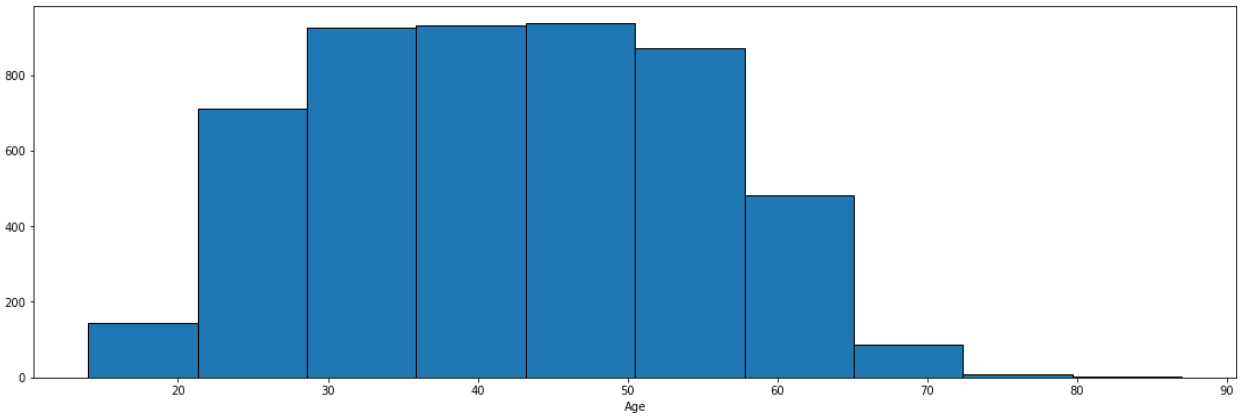
2. Background

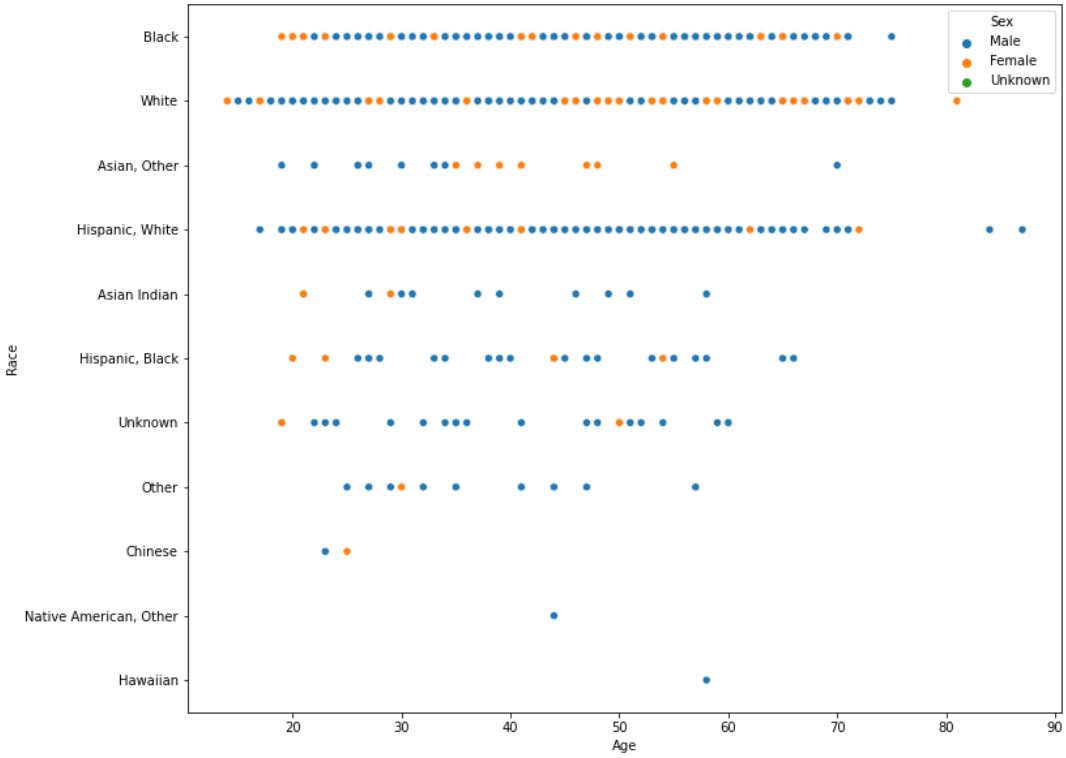
Prior to developing the model, some research regarding drug overdose deaths was gathered. One specific resource used was from the cdc, which described the frequent drug overdose deaths from 2011-2016[2]. These helped in providing a starting point as well as some background knowledge in understanding the subject. The current strategies to combat drug overdose deaths typically include going to therapy or being prescribed certain substances to prevent further damage. However, these methods are largely reliant on identifying or noticing a problem. The issue with that is sometimes the problem is not noticed until it's too late. By using this, one can determine the risk of someone consuming drugs and minimize their chances of a drug overdose death. Normally one would have to notice something wrong with a person to suggest they seek help, however, via this model there are additional resources to identify people at risk. The normal methods combined with this auxiliary method will allow for a decrease in drug overdose deaths and further aid for those who could potentially suffer from a drug overdose death.



3. Dataset

The dataset used in the research gave information about the victims of drug overdose deaths from 2012-2018. It provides information about the type of substance, demographics, and the details about the death. It had 5105 different victims with 41 different details on each one. However, after the preprocessing it was decreased to 5099 victims with 19 different details on each one. The preprocessing removed unnecessary data and assigned numerical values to all of the columns. In addition, some columns that were null and could not be assigned an average placeholder value were removed, resulting in a slightly smaller number of victims. The first graph below displays the distribution of drug overdose deaths by age while the second adds on to that displaying the frequencies of different ages, races, and sexes.





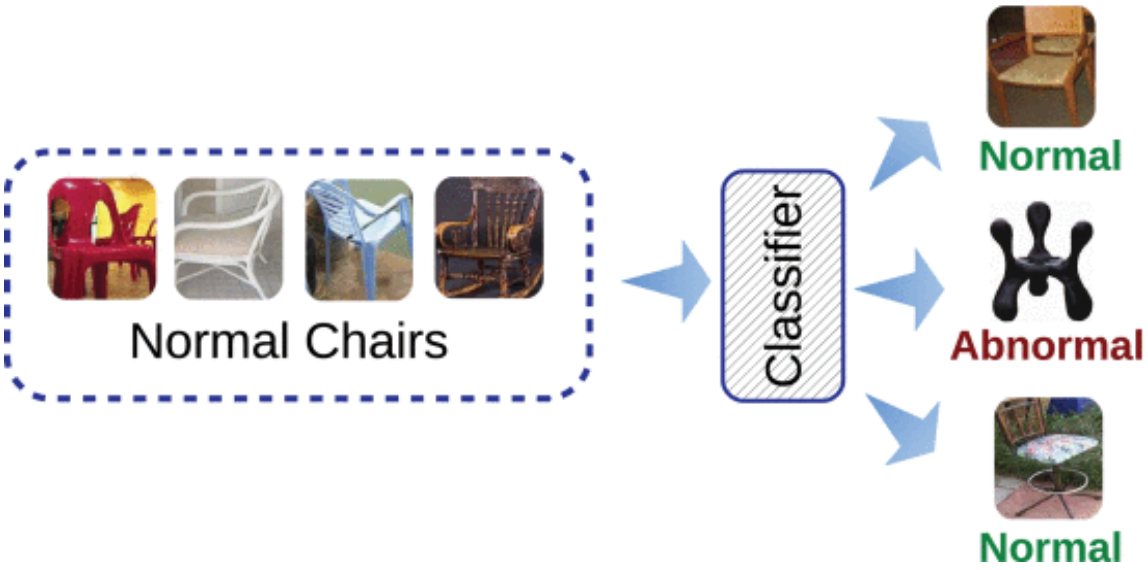
4. Methodology / Models

The data used in this project had information about drug overdose deaths in the US from 2012-2018, but it did not include every drug consumption in the US from 2012-2018. As a result, normal classification would not work, resulting in the usage of One Class Classification.

What is One Class Classification?

One Class Classification, rather than labeling the data with different classes, divides the data into either inliers or outliers. It operates as anomaly detection.

For example, in this image the chairs in the data are registered as normal, and thus, any that are different from the chairs in the data are classified as abnormal.[3]



One Class Classification using Local Outlier Factor

One way to identify outliers is to locate instances that are separated from other instances in the feature space. Although, if there are too many features, this method lessens in accuracy. Local Outlier Factor, or LOF for short, achieves outlier detection by utilizing the idea of nearest neighbors, by assigning each example a score of how likely it is to be and outlier or inlier based on the size of its local neighborhood. The examples with a higher score are the ones that have the highest chance to be an outlier.

One Class Classification Using Support Vector Machines

Another method of identifying outliers is capturing the density of the majority class and classifying the other examples as outliers. Support Vector Machines or SVMs for short utilize this.

One Class Classification Using Minimum Covariance Determinant

The final method used is Minimum Covariance Determinant or MCD for short. It works by defining an ellipsoid that covers the normal data. The data that is located outside of the ellipsoid is considered an outlier while the data within the ellipsoid is an inlier.[4]

5. Results and Discussion

The three models used scored relatively similarly with Minimum Covariance Determinant scoring the highest and Local Outlier Factor scoring the lowest. Local Outlier Factor had an f1 score of 0.836, and an accuracy and recall of 0.8362745098039216. Support Vector Machines had an f1 score of 0.849, and an accuracy and recall of 0.8490196078431372. Minimum Covariance Determinant had an f1 score of 0.85, and an accuracy and recall of 0.85. However, the dataset used for the models could lead to potential error. Since the training and testing data is completely one sided, giving no example of living drug users, all of the data should be classified as an overdose death. However, since everything is a drug overdose death, there is no way to truly test the models accuracy.

Model Name	Accuracy	Recall	F1 Score
Local Outlier Factor	0.8362745098039216	0.8362745098039216	0.836
Support Vector Machines	0.8490196078431372	0.8490196078431372	0.849
Minimum Covariance Determinant	0.85	0.85	0.85

6. Conclusions

This model can assist in preventing further drug overdose deaths. While this does not have 100% accuracy, it can help in identifying people at risk and allow preemptive actions. The issue with the current model, however, is that it i solely using one-sided data, resulting in outcomes with questionable credibility. If I were to extend the project, I would fit it with up to data, from both sides and attempt to increase the scores. Additionally, by refitting it with new information, the model can be continually utilized for years to come.

Acknowledgments

This paper would not have been possible without my mentor, Eric Bradford. His continued support, assistance, and knowledge aided me throughout the process. I am also grateful to my parents who aided me throughout the process.

References

[1]CDC. 2022. *U.S. Overdose Deaths In 2021 Increased Half as Much as in 2020 – But Are Still Up 15%*. [online] Available at: <https://www.cdc.gov/nchs/pressroom/nchs_press_releases/2022/202205.htm> [Accessed 9 August 2022].

[2]Hedegaard, H., Bastian, B., Trinidad, J., Spencer, M. and Warner, M., 2022. *Drugs most frequently involved in drug overdose deaths : United States, 2011–2016*. [online] CDC Stacks. Available at: <https://stacks.cdc.gov/view/cdc/61308> [Accessed 13 August 2022].

[3]Takamatsu, M., 2020. *Deep One-Class Classification*. [online] Medium. Available at: <https://medium.com/analytics-vidhya/paper-summary-deep-one-class-classification-doc-adc4368af75c> [Accessed 9 August 2022].

[4]Brownlee, J., 2020. *One-Class Classification Algorithms for Imbalanced Datasets*. [online] Machine Learning Mastery. Available at: <https://machinelearningmastery.com/one-class-classification-algorithms/> [Accessed 9 August 2022].