ASSESSING FACTORS INFLUENCING POISONINGS IN VIRGINIA

Emily Huo
School of Engineering and Applied Sciences
University of Virginia
Charlottesville, VA 22903
esh2nne@virginia.edu

Anupama Jayaraman
School of Engineering and Applied Sciences
University of Virginia
Charlottesville, VA 22903
aaj5dgv@virginia.edu

Mason Barnes
College of Arts and Sciences
University of Virginia
Charlottesville, VA 22903
mrb5ay@virginia.edu

December 8, 2022

1 Abstract

A CDC study shows that drug overdoses and poisonings were the third leading cause of death for children and teenagers ages one to 19 in 2020, an 83.6 percent increase from the year before. The sudden increase was in large part due to a 110.6 percent increase in unintentional poisonings. In Virginia, for the 1-19 age group, there were 26 fatal overdoses in 2019, which jumped to 40 in 2020, a 65 percent increase. Despite these studies, there appears to be a gap in the research done in poisonings in general since the focus is on specific poisonings. Due to this, we aim to assess factors influencing poisonings in Virginia by combining multiple datasets together to gain a better understanding of the situation. Then, by using classification, we aim to predict the accidental poisoning death-rate given a county in Virginia up to 7 different "severity levels" depending on the death rate per 100,000 people. Ultimately, we plan to predict future poisoning death-rates in order to identify areas and counties in Virginia that will need help in preventing poisoning-related deaths and treating poisoning-related issues. Our final model, a neural network, results in an accuracy of 79.7% and a weighted F1 score of 80.2%. This model can be used to identify counties that require more resources to deal with poisonings.

2 Introduction

2.1 Motivation

Our project looks at drug-poisoning deaths in Virginia, defined as drug-related deaths caused by unintentional means, like drug overdoses, suicide, and homicide. The CDC states that overdoses and poisonings were the third leading cause of death for children and teenagers ages 1-19 in 2020, with an 83.6% increase from 2019. In Virginia, for the 1-19 age group, there were 26 fatal overdoses in 2019, which jumped to 40 in 2020, a 65% increase, following the overall national trend in increasing drug related incidents. [1] In addition, 42%, of poisonings occur under 5 years old. Unfortunately, there appears to be a gap in the research. There is nothing about poisonings in general and current research, which primarily focuses on certain poisonings such as heavy metals or opioids. The primary goal of this project is to learn about the number of poisonings in Virginia, especially the important factors that impact the number of occurrences. We ultimately want to be able to predict rate of poisonings by predicting the poison-death rate in a given county of Virginia so the Virginia government can know much support to provide at future conditions. This will better help the Virginia government effectively allocate resources and save money.

2.2 Dataset

We will be using the Drug Poisoning Mortality by County dataset which outlines poisoning deaths/100,000 people in each county in the United States from 1999-2016:
`https://healthdata.gov/dataset/NCHS-Drug-Poisoning-Mortality-by-County-United-Sta/b7de-mf8t/data?no_mobile=true`

To append additional features and more further analyze the factors that may be affecting poisoning mortality, we appended school enrollment, drug use, demographics, labor force, medicare enrollment data, and geographical region.

For school enrollment, we added the ACS School Enrollment survey which contains information on school enrollment (including preschool enrollment) for counties in the United States year over year:
`https://data.world/uscensusbureau/acs-2016-5-e-schoolenrollment/workspace/project-summary?agentid=uscensusbureau&datasetid=acs-2016-5-e-schoolenrollment`
For drug use, we added the DEA's pain pill database data for Virginia, which includes information regarding drug purchases, buyer information, and investigations:
`https://www.washingtonpost.com/graphics/2019/investigations/dea-pain-pill-database/`
For demographics, we used government census data to determine the population and the number of people of different racial and ethnic backgrounds in each county:
`https://data.census.gov/cedsci/table?q=demographics%20virginia%202016&tid=ACSDP1Y2016.DP05`
For labor force, we added the Virginia government's Virginia Labor Force and Unemployment Estimates by Month by County which provides information about the total labor force, the number employed and unemployed, and the unemployment rate:
`https://data.virginia.gov/Economy/Virginia-Labor-Force-and-Unemployment-estimates-by/f6ym-7273/data`
Since this dataset listed counties in terms of codes, we needed to use an additional dataset `https://github.com/jalbertbowden/open-virginia-data-toolkit/blob/master/fips/fips-codes-virginia.csv`
For medicare enrollment, we used data from the Dartmouth Atlas Project, which described medicare spending per county in 2014:
`https://data.world/adamhelsinger/county-state-medicare-spend`
For geographical region, we used data from the UVA Weldon Cooper Center for Public Service's Demographics Research Group, which lists the counties and cities that are in each of the 8 demographic regions of VA:
`https://demographics.coopercenter.org/virginia-regions`

Further data could be used for this project including marital status and income for each county in the United States. Much of this data should be available through government census data and other national departments' surveys.

2.3 Related work

Lobo et al. predicts childhood lead exposure or lead in blood by using socioeconomic, housing, and water quality predictive features. Out of the 5 machine models that the scientists tested, Random Forest, Logistic Regression, k-Nearest Neighbor (kNN), Decision Trees, and Support Vector Machine SVM), Random Forest performed the best. The machine learning model predictions are at a fine aggregated geo-spatial scale so it can identify state-level hotspots where testing should be a priority. However, the models were implemented by establishing an arbitrary exposure thresholds. In the future the researchers could figure out another way to do it without establishing thresholds. Additionally, this research topic is limited by the lack of publicly available data. This is also why although it can accurately predict the risk of childhood lead exposure for a given area, it does not provide information regarding the variability within each area. In a large community or zip code, the aggregated data would fail to accurately describe the population [2].

In Potash et al., researchers studied whether a machine learning approach could be used to predict childhood lead poisoning. Both random forest and logistic regression methods were used to characterize risk of childhood lead poisoning with 95.5% and 95.1% specificity respectively. This research shows promising results with respect to the efficacy of using machine learning in predicting poisoning risks for children. However, this research only considered a dataset of about 6000 Chicago families and focused specifically on lead poisonings. While they used 2015 census demographic data for their models, they also considered housing characteristics which is more specifically related to lead poisonings rather than the fatal poisonings we propose to study in our project [3].

Dong et al. analyzes the risk of opioid overdosing in patients based on electronic health records and using machine learning techniques. The aim of their research was classification. The researchers gathered data from clinical databases regarding New York inpatient data and labeled patients as either opioid overdose or other patients. They determined that the most important feature for predicting overdose status was the Clinical Events feature, which are related symptoms like pain level or smoking history. For the model, the researches converted their categorical data using one-hot encoding and employed the random forest, decision tree, and logistic regression models. They also generated a deep neural network for classification. Analysis of the data demonstrated that the random forest method had the best recall and the neural network had the best precision. Although precision and recall were high, Doug et al. failed to use an Imputer and thus incorporate more features with missing values. Moreover, they could have used a more advanced architecture for neural networks. Ultimately, the paper demonstrates that given clinical data, machine learning can aid healthcare workers in combating opioid poisoning. This suggests that finding summary clinical data for VA counties will be beneficial in our proposed project [4].

3 Method

In order to perform our initial experiments, we first compiled the datasheets described above (section 3) into one larger, master dataset. Subsequently, we attempted to visualize and characterize the data to better understand the data we were working with and attempt to uncover any hidden relationships. Finally, we generated a Random Forest Classifier model for classifying poisoning death rates and assessed the model's performance. We specifically chose to use Random Forest as we believed that ensemble learning generate a decent initial model with intermediate complexity, compared to logistic regression and neural networks. Detailed methodology is described below (sections 7-9).

3.1 Data Compilation

Since our research did not turn up a good dataset that had many entries and features to use, we decided to combine multiple datasets so we could increase the number of features. In order to combine our dataset, each teammate took at least one dataset to read in and prepare.

3.1.1 VA Counties and Codes

Since the labor force dataset did not have a column with county name, a way to convert the county codes to county name had to be found. We found the dataset fips-codes-virginia.csv, which had a column for county code, and a column for county name. The other irrelevant columns were deleted so those were the only ones left. Then, to make it easier to lookup the county name given the county code, the dataframe was converted into a dictionary. This dictionary then gets used by the labor force dataset.

3.1.2 Labor Force

The labor force dataset was prepared by adding in a new column called Names which lists the county a data entry corresponded with. This was added using the County Code that the dataset already had and referencing the above mentioned dictionary of [county code : county name]. This way, it could be combined with the other datasets which only have county name. Then, the irrelevant columns were dropped so the dataset only included Name, Year, LaborForce, Employment, Unemployment, and UnemploymentRate. Finally, the dataset was restricted to only contain entries from the years 2010-2016.

3.1.3 Demographics

Demographic data for the years 2010-2016 was collected from the census website and matched to the poisoning data by comparing across county names and years. Features included population breakdown by race and ethnicity for each county in Virginia.

3.1.4 School Enrollment

School Enrollment data was collected from 2016 census data and was incorporated into our poisoning data set by comparing across county names. As we only have reliable school enrollment data for one year at this point, we generalized the school enrollment across all years 2010-2016 in the poisoning data set. Later, we may want to find more accurate school enrollment data across various years in order to better capture the effect of school enrollment on poisoning rates.

3.1.5 Poisoning Mortality

Poisoning data was restricted to Virginia counties during the years 2010-2016. This was because these were the years for which we had both reliable demographic and labor force data.

3.1.6 Drug Use

Drug use in Virginia data, from the DEA's pain pill database, included a variety of features, including buyer and transaction details. In order to get data that we believed would better train a model to predict poisoning mortality, we processed the data to get the number of transaction per Virginia county per year. The final dataset ranged from 2006 to 2014, but we restricted the data to 2010-2014 to conform with the same lower year limit as the other datasets.

3.1.7 Medicare

For the medicare enrollment data, we extracted only the number of medicare enrollees in each county. This data was only present 2014, so we generalized the 2014 number for a specific county to all of the years for that county.

3.1.8 Geographic Region

For geographic region, we used the county and city information already within the dataset to tie the specific geographic region to each data point.

3.1.9 Putting the Data all Together

Before combining all of the data, we first saved our processed data from each dataset into its own .csv file so that we would not have to repeat those steps in the future. In order to combine all of our data into a master dataset, we iterated through each of the data points in the initial poisoning mortality dataset. For a given datapoint, if another dataset had a datapoint with a matching county and year, we appended those feature values to the datapoint in the poisoning dataset. We ultimately saved this dataset as poisoning_full_data.csv.

4 Experiments

Initially in order to process the data, we will want to exclusively parse our data for only Virginia counties in order to make sure our solution is specifically applicable to Virginia. Next, we will want to append time and information data to the current dataset to incorporate more features that we can evaluate and subsequently train our models with. Ultimately, we have three aims with this project:

Our first aim is to learn about our problem and view any existing correlations between our features. We will look at processing our data in different ways to accomplish this task, including using a correlation matrix for our continuous features. This type of analysis will help us determine which features we should be incorporating into our model.

Our second aim to develop a model that will predict the poison death-rate in a given county of Virginia. We will examine this using random forest, support vector machines, and neural networks to classify the death rate as those models worked well for similar work in lead and opioid poisoning as noted in previous literature (see Section 3). Moreover, we will consider simplifying the classification task by grouping together and reducing the total number of categories, which may help improve our classification performance. We will test the model performance and fine-tune parameters by using a train-test split and cross-validation.

Our third and final aim is to use our model to identify Virginia counties that we should be concerned about and that we should redirect resources towards in order to deal with the drug poisoning issue.

4.1 Python Notebooks for Analysis

Our machine learning code is written the following 2 notebooks: `https://colab.research.google.com/drive/1ImSEsETGTWLMtKy4Amky6jLUvW8OnOHX?usp=sharing` and `https://colab.research.google.com/drive/1-icXN5ES5FQEBOCiYuYpa_RbpMIwH--U?usp=sharing`. Hard copies of these documents are submitted on UVA Collab along with our final dataset.

5 Results

5.1 Dataset Characterization

Our final dataset use 152.9 kB of memory to store 931 data points with 21 features. These features include 4 categorical features (County name, year, Poison Death Rate, and geographic region) and 17 numerical features: population, labor force, unemployment, unemployment rate, total Hispanic/Latino population, total white population, total black or African American population, total native American or Alaskan native population, total Asian population, total Hawaiian or Pacific Islander population, total other race, preschool enrollment, kindergarten enrollment, 1st-4th grade enrollment, number of drug buys, and medicare enrollment. In this dataset, we had 323 missing values among the number of drug buys data, which can be attributed to the lack of complete overlap between the years of the rest of the dataset and the drug use dataset. In order to begin viewing any correlations, we plotted the death rate categories against each of the numerical features. The plots are visible in our ML4VA_Project_ScikitLearners.ipynb code.

In our data, we see that larger populations tend to have lower death rates, possibly due to more infrastructure that can deal with poisonings. We see similar trends with labor force, employment, and unemployment. However, there is no apparent trend with the unemployment rate, likely because labor force, employment, and unemployment are not normalized by the population and thus are better representations of population instead of the work force. With all races, except African Americans, we see a strong right skew. African Americans appear to have a bimodal distributions of death rates. Grade level enrollment and Medicare enrollment seem to have the same impact as labor force, but this may again be due to the lack of normalization. The number of drug buys appears to have an inconsistent effect on the death rate.

In looking at Virginia geography, we see a roughly normal distribution of death rates in most regions. Interestingly, Central, Eastern, Hampton Roads, and Northern regions all have fairly tight distributions, while the Valley and West Central regions have a bit more spread. In addition, the Southwest Region is significantly skewed right. This may be due to the fact that southwest Virginia has one of the highest rates of poverty, with more than 50% living under the poverty line in 2018.

5.1.1 Feature Engineering

In order to improve the accuracy of our model, we took a look at engineering our features by dividing the current features by population. The results demonstrate that labor force, employment, and school enrollment, normalized by population, appear to have little impact on the death rate. Death rates appear to decrease with increasing Hispanic, Black, Asian, and Pacific Islander population percentages, while it appears to increase with increasing White population percentages. In addition, normalized drug buys appear to have a positive correlation with death rates, and medicare enrollment appears to correspond to a medium death rate.

In order to further improve model performance, we considered simplifying the classification problem by compiling the 15 death ranges in the dataset into 7 ranges: 2-5.9, 6-9.9, 10-13.9, 14-17.9, 18-21.9, 22-25.9, and 26+ deaths per 100,000.

5.2 Random Forest Model

We initially used the Random Forest Classifier in order to predict the category of poisoning death rates as given in the NCHS drug poisoning mortality data set. The data set matches all counties in the United States to 15 different ranges of deaths: 2-3.9, 4-5.9, 6-7.9, 8-9.9, 10-11.9, 12-13.9, 14-15.9, 16-17.9, 18-19.9, 20-21.9, 22-23.9, 24-25.9, 26-27.9, 28-29.9, and 30+ deaths per 100,000 people.

First, we cleaned the data using an Imputer with a 'median' strategy to take care of any missing data, scaled, and used a OneHotEncoder to take care of our one categorical variable, the year (dropping "County Name" from our feature list). We then used randomized search cross validation in order to find hyperparamaters for the Random Forest Classifier which resulted in the greatest weighted F1 score. Our best classifier had a maximum depth of 19, a minimum sample leaf node size of 1, and a minimum sample split size of 2.

Using raw data, unaltered with engineered or dropped features, the model yielded a weighted precision score of 49.6%, a weighted recall score of 49.7%, and a weighted F1 score of 48.9%. Once we incorporated the engineered features, dropped the features that did not appear to influence death rates (labor force, employment, unemployment, nursery enrollment, kindergarten enrollment, and grades 1-4 enrollment per year, and unemployment), and used the simplified death rate groupings, we found the same model hyperparameters optimized performance, which increased to a weighted precision of 73.9%, weighted recall of 74.3%, and a weighted F1 of 74.0%.

5.3 Support Vector Classification Models

To see if we could improve upon our Random Forest Model, we attempted to use Support Vector Machines on our engineered dataset. However, we found that a tuned linear SVM yields an F1 score of 53.8%, polynomial yields an F1 score of 64.4%, and gaussian yields an F1 score of 68.9%,

5.4 Neural Network Model

Finally, we created an ANN with 6 dense layers and one dropout regularization layer. The model was ran on the engineered features constructed for the Random Forest model as described above and predicted the "threat level" of each county in Virginia. Our final trained model resulted in a validation accuracy of 79.68% and yielded a weighted f1 score of about .8022 on an unseen test set, outperforming our Random Forest model by a few points. The model was made of 6 dense layers and one dropout layer. Each dense layer used the 'Relu' activation function besides the last layer which used the 'softmax' function. The model had a total of 3,291,399 parameters.

6 Conclusion

In the end, we had a model that could predict the poisoning death-rates in Virginia counties and rank the different counties according to their poisoning severity level given information like demographic and medicare enrollment. By applying our neural network to data points for each year of available data, we predicted that 5 of the counties with some of the consistently highest predicted poisoning death rates are Buchanan, Dickenson, Wise, Russell, and Pulaski County. This is very useful because we have learned that we should redirect Virginia's health resources to serve these areas. In the future, our work can be built upon by incorporating more recent data and assessing new features. For example, we notice that most of these counties are in southwest Virginia, which has a disproportionately high poverty rate. Including income or poverty data might improve the performance of our model.

7 Contribution

In order to complete this project, we split up the work. The task breakdown is as follows:

Mason Barnes: Processed the school enrollment, demographics, and poisoning death-rate data. Used random forest to get a preliminary model for the data and designed the neural network. Wrote relevant sections in the paper.

Emily Huo: Processed the labor force and county names/codes data. Contributed to preliminary data visualization and characterization. Made the video for the presentation. Wrote relevant sections in the paper.

Anupama Jayaraman: Processed the medicare enrollment and drug use data. Contributed to preliminary data visualization and characterization. Did feature engineering for an improved Random Forest and SVM. Wrote relevant sections in the paper.

References

[1] Lindsey Kennett. 'Drugs are just taking over': Virginia sees 65% increase in fatal overdoses in 2020" https://www.wsls.com/news/local/2022/04/29/drugs-are-just-taking-over-virginia-sees-65-increase-in-fatal-overdoses-in-2020/. (accessed: 11.02.2022)

[2] G.P. Lobo, B. Kalyan, & A.J. Gadgil. Predicting childhoos lead exposure at an aggregated level using machine learning. In *Intenation Journal of Hygiene and Environmental Health*, 238. 2021.

[3] Eric Potash, Rayid Ghani, Joe Walsh, Emile Jorgensen, Cortland Lohff, Nik Prachand, & Raed Mansour. Validation of a Machine Learning Model to Predict Childhood Lead Poisoning. In *JAMA Network Open*, 3(9). 2020.

[4] Xinyu Dong, Sina Rashidian, Yu Wang, Janos Hajagos, Xia Zhao, Richard Rosenthal, Jun Kong, Mary Saltz, Joel Saltz, & Fushend Wang. Machine Learning Based Opioid Overdose Prediction Using Electronic Health Records. In *SSRN Electronic Journal*, pages 389–398. 2019.