**ANALYZING ACCIDENTAL DRUG OVERDOSE IN CONNECTICUT**

**Introduction**

A substantial increase in the cases of accidental death due to the drug overdose, in the state of Connecticut, over the last few years has caused a detrimental effect on society.

During last 15 years, accidental death due to drug overdose has increased dramatically throughout the United States. Moreover, the Center for Disease Control(CDC) has declared that death due to the drug overdose is one of the leading reason for death among the Americans under the age of 50[1]. In particular, the project is focused on the drug-related health crisis in Connecticut.

A survey report portrays that in Connecticut around 3,600 deaths have been recorded in each year between 2012 and 2017[2]. According to the ctpost, in Connecticut, drug-related deaths hit their highest point in the year 2013, a report prepared of last 10 years [3]. Other report says that accidental deaths due to drug abuse in 2016 have increased 25% compared to the death rate of previous year [4]. The vast array of reasons in different parts of the state that form a part of toxicity report generated for each case shows the nuisance of easily available drugs. After each incident of accidental death is notarized by the Chief Medical Officer, the reports are further documented, cataloged, and stored in the database for public references. Our team understood the seriousness of this problem which has caused in loss of such precious human lives in the state of Connecticut. There are some existing conventional analysis methods that deal with the basic interpretation of the data. However, these methods fail to explore the hidden pattern associated with the datasets. This raises the need for analysis using the data mining techniques. Our team believes that these hidden patterns such as relationships between one drug to other drugs and drug relationship with the victims’ demographic would help to mitigate the issue of accidental death related to drug abuse.

The report begins with a problem statement that explains the increased rate of accidental death due to an overdose of drugs in Connecticut over the last few years. The claim of the increased death rate because of drug death rate is backed up by various facts and figures. These facts and figures provide several pieces of evidence to treat this issue as a potential threat to the society, resulting in a data mining problem. Later, the report explains our team’s belief of mitigating the issue with a data mining patterns. Further, the report describes the data exploration part performed in checkpoint#2. Data exploration consists of the source of data, type of data file, number of records, number of attributes, initial level plan for preprocessing, and some basic level of explorations. Going ahead, the report also explains the preprocessing techniques and algorithms used in checkpoint#3. For our data, we have chosen a clustering algorithm. So, the report discusses why clustering being used and how the performance is measured once classification and clustering algorithm is being applied to the dataset. The report later discusses the pros and cons of each step of the logic of problem. Report explains about the model and discovered pattern in the dataset and compares different models and suggests which one would perform better. Finally, we conclude our project findings and discuss future work.

**Data Exploration**

The data motivating for the study is taken from www.data.gov’s website from database Accidental Drug Related Deaths 2012-June 2017 [2]. The data consists of 3,584 records and partitioned into 32 attributes. These attributes are: incident number, date of death, victim’s gender, victim’s race, victim’s age, city of residence, state of residence, county of residence, city of death, state of death, county of death, place of death, prognosis of incident, place of incident, cause of death, manner of death, diagnosis of incident, location of death, heroin, cocaine, fentanyl, oxycodone, oxymorphone, etoH, hydrocodone, benzodiazepine, methadone, amphet, tramad, morphine, other, and opioid.

After having a close look at the data, our team has found some insignificant attributes. For example, the incident number is an attribute which is a unique identifier and used to index each death. The incident number may be useful for the medical community to keep track of records, but it does not contribute to a model formation and hence is insignificant for data mining process. In addition, there are other attributes such as the state of residence and state of death, in which almost all of the records belong to Connecticut and so insignificant to data mining process.

There is an attribute so-called date of death, which is a composite attribute consists of day, month, and year of victim’s death. The date attribute needs to be fragmented to perform a better comparison between datasets. In addition, there are attributes like race, the city of residence, and place of injury, which has some missing values and replaced by ‘unknown’. Further, the team has found that the field manner of death has natural death value only for three instances out of 3,584 and hence needs to be removed. Finally, after data exploration phase we are left with 3,581 records and 30 attributes, which will be cleaned further in preprocessing phase.

**Methodology**

In the preprocessing phase, our team found that one of the record (number 781) has an unknown age attribute value and with the assumption that age is a discrete attribute, the age attribute value is replaced by a mode value i.e. 50. Even though, if we have considered age attribute as a continuous attribute, the mean value would be approximately 50. Hence, 50 is a suitable value for the missing age value for the record.

Further, the team has converted numeric value of age and year attribute to a nominal value, so that clustering algorithms be performed on it. In addition, the team has found that attributes like residence county, death county, death location should be removed since they do not contribute to the model building but affecting the accuracy.

We have used two different learning algorithms for model building- clustering and classification. We thought to perform classification algorithm for our dataset but soon we realized that for our dataset there are not potential class values and hence decided to perform clustering first.

We initially started with Hierarchical cluster algorithm but it clustered 3579 records in one cluster and remaining 1 record in another cluster; so, then we moved to SimpleKMeans clustering algorithm. Here we tried different values for K like 2,3,4,5,6,7,8,9 for a learning curve and found that K=6 is an ideal K value because the curve has very small changes after it. This gave 6 clusters as listed in Appendix C. which we later transformed into 3 values which act as our class values for classification: Highly Vulnerable, Moderate Vulnerable, Less Vulnerable by grouping similar clusters.

Further, we performed classification and generated rules for the data using JRIP algorithm so that we can classify a test set easily. We also performed IBK for different values of K and Naive Bayes algorithm for classification purpose.

The main performance measure used was accuracy (correctly classified instances) in classification algorithms. We performed three main types of classification algorithms 1. IBK 2. Naive Bayes and 3. JRIP. We have attached the accuracies and confusion matrixes obtained for different measures like cross-validation, Percentage split in Appendix A. We believe Cross-validation is appropriate to measure as it gives all subset equal chance for contributing in model building and evaluation.

**Logic of Problem**

Creating and updating logic problem(LOP) using analytical thinking tool is a constant step while developing this project. The initial LOP in checkpoint 1 was the most crucial part as it involved predicting the approach to be undertaken while performing the data mining process for the problem. After checkpoint 1 through checkpoint 2 and 3, the procedure involved constantly updating the LOP based on the changes made during every checkpoint. The most important part was to update the concepts that were being used during each checkpoint while not changing the actual conclusion that the project will provide. For example, in checkpoint 1 our team predicted about using affinity analysis. However, after realizing the limitation of the data file in checkpoint 3 we first performed clustering and then performed classification in an ordered manner. LOP for checkpoint 1 was for project proposal stage. It involved developing LOP based on the preliminary findings and initially predicted approach. LOP updated for checkpoint 2 was based on data exploration of the data file. In Checkpoint 3 the LOP was re-updated based on the data pre-processing.

One of the advantages of LOP consists of a systematic approach with well-defined conditions for every step. This helps in creating an idea of the process to be undertaken while performing the LOP. Another advantage of LOP is that it is essential for analytical thinking as it helps to find solutions to a data mining problem based on the ability to analyze first and then implement the approach for a solution. LOP is a cyclic process, gets constantly updated during various checkpoints. Hence, it gives a better platform for the approach to be undertaken.

There are some grey areas in implementing LOP. One such is that sometimes the initial and final logic problem might be completely different due to the constant change involved in every step. Care should be taken that the goal doesn’t change much. Also, creating LOP at every step can be a time-consuming affair. The logic of problem helps in creating a path to solve the data mining problem, however, it doesn’t provide a way to evaluate the path undertaken.

Our team suggested one additional section called ‘Plan of Action’ where literature review and case studies of similar concepts can be analyzed. The case studies of similar types of problems (not exactly same problem) can help get team better perspective of solving the current problem. Professional teams can also conduct interviews with the people of the domain to understand their side of the story. The answer to the questions being asked in the interview can help in understanding the data in an effective way. This will provide a team a clear perspective for solving the problem using data mining based on the needs of people who require a solution. Based on the study of individuals responses, previously held case studies about the subject and remaining LOP can give efficient results for the project.

**Conclusion**

Our team has run classification algorithms such as IBK, Naive Bayes, and JRIP to record and compare the results. In team’s observation, rule base algorithm(JRIP) has the highest accuracy rate both for cross-validation (87.62%) and percentage-split (88.24%) as mentioned in Appendix A. Furthermore, the accuracy rate of Naïve Bayes algorithm is least for both cross-validation (80.47%) and percentage split (79.12%). In addition, JRIP generates 28 rules that would be significant for analysis that is shown in Appendix B. One of the rule says that if the death happened at “Hartford” city and presence of heroin not found in the gender male then the severity measure categorizes it as Vulnerable. Another rule is generated based on presence/absence of some of the drugs in victim’s body and on the toxicity report. For example, if the drug EtOH is found, Oxycodone is not found, and toxicity report says multiple drug toxicity, for male, weka categorizes the severity of it to less Vulnerable.

One of the patterns found in a different model is that most of the males between the age group of 26-50 years, who are the resident of city “Waterbury”, are addicts of drugs- “Heroine” and “Oxycodone”.

For our dataset, there is very few numeric attribute (age and year) and most of the attributes are related to the presence/absence of various drugs in victim’s body. We feel JRIP would be the best model since it generates various rules related to the comparison of drugs presence/absence along with the relationship of them with different attributes such as gender, age, and death city.

Our project is focused on drug overdose-related death in the Connecticut. Overall, around 3600 deaths have happened in Connecticut between the 2012-May 2017. Death due to drug abuse is increasing every year and it already reached the alarming stage. The accidental death due to excessive drug dose has become so severe that Center of Disease Control (CDC) has declared it as a “leading cause of deaths among American under age 15” [1]. Now, the problem raises a question that if this issue is so severe then why are there no proper regulation in place to mitigate the issue. One of the probable answers could be that probably regulation is present but are not so effective. Another, possible answer could be that authority is not yet able to find a pattern of the accidental death related issue and hence they are not able to mitigate this crisis. Considering the severity of the problem, our team has decided to put some afford in finding the hidden set of patterns that could probably help the authority to put some strong regulation to reduce the number of deaths. In our observation and research, we found that it is hard to stop a drug-related death for a prescribed drug. Since these drugs are prescribed by a medical practitioner in most of the cases and it is difficult to stop a patient to take excessive medication. So, the team has decided to focus more on other opioid drugs such as Heroin, Cocaine, Morphine, and so on.

To find the relationships between other opioid drugs, we initially decided to go with JRIP and Apriori algorithm. We tried running apriori algorithms to find the association between drugs, but it generates few rules that were insignificant to the problem. So, we used simpleKMeans clustering technique to form the cluster and later JRIP classification algorithm to build the model and generate rules. The model that we build generated 28 rules. The conclusion drawn from these rules is that most the people who are involved in drug overdose-related deaths are “white male” from “Waterbury city”, and their age is between “26-50 years”. In addition, most of these deaths that happened are because of the presence of “Heroin” and “Oxycodone”. In our team’s view, if the government or higher authority will try to regulate these two drugs which are present in the city of Waterbury, the death rates can be mitigated.

As mentioned earlier, our data source does not have any numeric fields. So, for further action, our team will try to find and add more numeric attributes, that would make the analysis easier. Another alternative is that our team may have to generate some numeric field out of the existing nominal fields. For example, we can use some standard harm score value of different drugs to generate an attribute field, which will be numeric. This harm score would be assigned to each of the entry present in the dataset. Some standard harm score of opioid drugs are listed in Appendix D. The harm score will help to accurately categorize each entry in different severity level. We believe that the severity level in an amalgamation of other attributes such as age, sex, and death city will be able to show some pattern in existing dataset. Further, at the end of project analysis and model building, we will generate a report and sent it to the police department and government, so that they can apply some strong rules to mitigate the issue of drug-related deaths.

**Appendix A:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Classification Algorithm used** | **Test Options** | **K-Value Used** | **Accuracy** | **Confusion Matrix** |
| IBK | Cross validation | K=1 | 82.09% | a b c <-- classified as  1748 129 39 | a = Highly Vulnerable  237 1012 27 | b = Vulnerable  153 56 179 | c = Less Vulnerable |
| K=3 | 83.82% | a b c <-- classified as  1804 96 16 | a = Highly Vulnerable  229 1039 8 | b = Vulnerable  169 61 158 | c = Less Vulnerable |
| K=5 | 83.99% | a b c <-- classified as  1813 95 8 | a = Highly Vulnerable  232 1039 5 | b = Vulnerable  176 57 155 | c = Less Vulnerable |
| Percentage Split (66%) | K=1 | 81.10% | a b c <-- classified as  582 42 5 | a = Highly Vulnerable  96 354 8 | b = Vulnerable  55 24 51 | c = Less Vulnerable |
| K=3 | 82.99% | a b c <-- classified as  604 23 2 | a = Highly Vulnerable  99 357 2 | b = Vulnerable  62 19 49 | c = Less Vulnerable |
| K=5 | 83.23% | a b c <-- classified as  604 23 2 | a = Highly Vulnerable  92 364 2 | b = Vulnerable  66 19 45 | c = Less Vulnerable |
| Naïve Bayes | Cross Validation | NA | 80.47% | a b c <-- classified as  1713 178 25 | a = Highly Vulnerable  276 975 25 | b = Vulnerable  142 53 193 | c = Less Vulnerable |
| Percentage Split (66%) | NA | 79.12% | a b c <-- classified as  570 55 4 | a = Highly Vulnerable  106 346 6 | b = Vulnerable  61 22 47 | c = Less Vulnerable |
| JRIP | Cross Validation | NA | 87.62% | a b c <-- classified as  1712 148 56 | a = Highly Vulnerable  111 1112 53 | b = Vulnerable  54 21 313 | c = Less Vulnerable |
| Percentage Split | NA | 88.24% | a b c <-- classified as  574 38 17 | a = Highly Vulnerable  33 402 23 | b = Vulnerable  24 8 98 | c = Less Vulnerable |

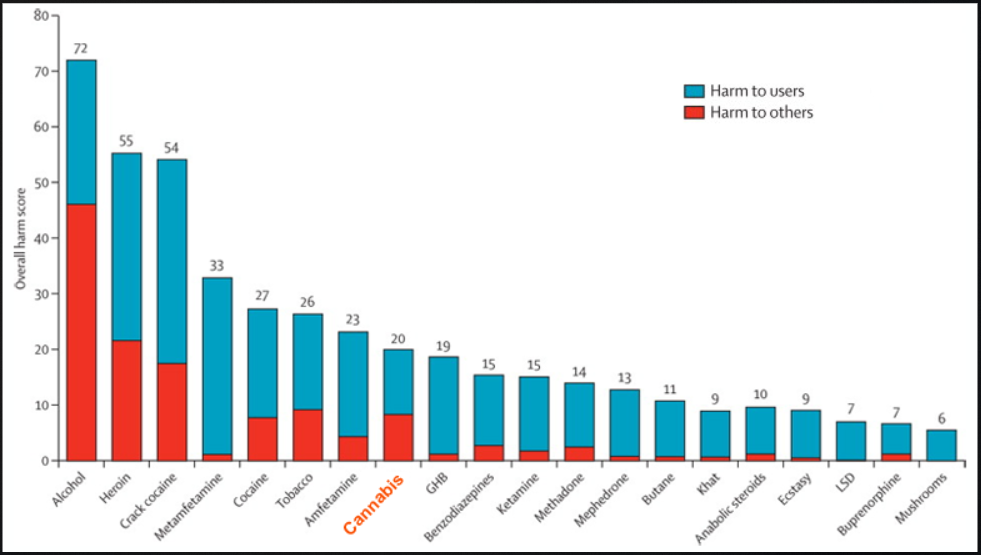
**Appendix B:**

|  |  |
| --- | --- |
| **Class Name** | **Rules** |
| Vulnerable | (Death City = HARTFORD) and (Heroin = Absent) and (Gender = Male) => Severity=Vulnerable (27.0/1.0) |
| (Death City = HARTFORD) and (Fentanyl = Present) => Severity=Vulnerable (17.0/3.0) |
| Less Vulnerable | (EtOH = Present) and (Death City = BRIDGEPORT) => Severity=Less Vulnerable (37.0/8.0) |
| (EtOH = Present) and (Toxicity Report = Multiple Drug Toxicity) and (Oxycodone = Absent) and (Gender = Male) => Severity=Less Vulnerable (9.0/1.0) |

**Appendix C:**

|  |  |
| --- | --- |
| Cluster0 | Apr, 2016, Male, White, 26, WALLINGFORD, MERIDEN, 'Other Recorded', Unknown, 'Hotel or Motel', 'Acute Intoxication due to the Combined Effects of Fentanyl Cocaine and Heroin', Present, Present, Present, Absent, Absent, Absent, Absent, Absent, Absent, Absent, Absent, Absent, Absent, Absent, Accident, Unknown |
| Cluster1 | Apr, 2017, Female, White, 60, 'NEW LONDON','NEW LONDON', Hospital, 'Abuse of Medication', Residence, 'Abuse of Medication', Absent, Absent, Absent, Present, Absent, Absent, Absent, Absent, Absent, Absent, Absent, Absent, Absent, Absent, Accident, 'Acute Intoxication due to the Combined Effects of Oxycodone and Dextro/Levo Methorphan' |
| Cluster2 | May, 2017, Female, White, 58, SEYMOUR, DERBY, Hospital, 'Substance Abuse', Residence, 'Substance Abuse', Present, Absent, Absent, Absent, Absent, Absent, Absent, Absent, Absent, Absent, Absent, Absent, Absent, Absent, Accident, 'Ethanol and Heroin' |
| Cluster3 | Jun,2016, Female, 'Asian Other', 35, BRIDGEPORT, BRIDGEPORT, Hospital, Unknown, Unknown, 'Acute Intoxication from the Combined Effects of Heroin Acetyl Fentanyl and Alcohol', Present, Absent, Present, Absent, Absent, Present, Absent, Absent, Absent, Absent, Absent, Absent, Absent, Absent, Accident, Unknown |
| Cluster4 | Feb,2012, Male, Black,30,'NEW BRITAIN','NEW BRITAIN', Hospital, Unknown, Other, 'Death Associated with Cocaine Use', Absent, Present, Absent, Absent, Absent, Absent, Absent, Absent, Absent, Absent, Absent, Absent, Absent, Absent, Accident, Unknown |
| Cluster5 | Sep, 2016, Male, 'Hispanic White', 58, HARTFORD, HARTFORD, Hospital, 'Substance Abuse', Unknown, 'Acute Fentanyl Intoxication', Absent, Absent, Present, Absent, Absent, Absent, Absent, Absent, Absent, Absent, Absent, Absent, Absent, Absent, Accident, Unknown |

**Appendix D:**



**References:**

[1] TheUpshot. 2017. Drug Deaths in America Are Rising Faster Than Ever. Retrieved September 27,2017 from <https://wonder.cdc.gov/controller/datarequest/D76;jsessionid=61D42FCE91996BE4F8E1C638A493A283>

[2] Data.gov. 2017. Accidental Drug Related Deaths 2012-June 2017. Retrieved September 27 2017 from

<https://catalog.data.gov/dataset/accidental-drug-related-deaths-january-2012-sept-2015>

[3] Cuda, A. 2014. Drug deaths hit 10-year high. Retrieved October 1, 2017, from <http://www.ctpost.com/local/article/Drug-deaths-hit-10-year-high-5259381.php>

[4] Kramer, J. 2017. Accidental Drug Overdoses Increase 25 Percent In Connecticut. Retrieved October 1, 2017, from

<http://www.ctnewsjunkie.com/archives/entry/accidental_drug_overdoses_increase_25_percent_in_connecticut/>