Alcohol Abuse In New York State

Ting Chien Meng
The Department of Computer Science
New York University
tcm390@nyu.edu

Haoran Zhang
The Department of Computer Science
New York University
hz2613@nyu.edu

Henan Liu
The Department of Computer Science
New York University
hl4001@nyu.edu

***Abstract*—Alcohol has become a crucial factor that influences many aspects of society. When people enjoy the benefits it brings, some users also suffer from it. According to the analysis from the CDC, 95000 citizens die from alcohol-related accidents annually. Such an astounding number raise our interest to find out more about alcohol products. In this project, the descriptive analysis and correlation analysis about the relationship between alcohol and some social and economic factors are given, such as unemployment rate and average education level. Furthermore, a prediction model is formed based on multinomial logistic regression, forecasting the medical assistance that the substance users probably need. The study is expected to enrich the analysis of the impact of alcohol on society, particularly on substance users.**

Keywords—analytics, alcohol, prediction, unemployment, education, economics

I. Introduction

Alcohol plays a vital role in many aspects of society including but not limited to the chemical, food, and electronic industry. For regular people, alcohol beverage has become the most popular drink on some occasions, such as banquets, parties, or even family dinners. When people enjoy it, some of them also suffer from it. According to the research from the Centers for Disease Control and Prevention (CDC)[1], around 95000 people die due to alcohol-related accidents annually, which makes alcohol the third leading preventable cause of death in the United. States.

Some social problems are also the consequences of alcohol disorder. Substance users normally have worse performance in work and education, compared with regular employees and students. It is hard for them to control the frequency of use and focus on their issues. Many users experienced job terminations or work accidents due to their symptoms of alcohol abuse. Without financial support, these users must choose other ways to get money for buying alcohol products. Some of them will live dependently with their parents or relatives, some will apply for government assistance, others choose the worst case, committing crimes, as the source of funds. These behaviors all import huge burden on governments and society and the only way to solve it is to help the patients out of addiction.

TEDS is a national census data system that collects the information about publically-funded admissions in each state of the United. States. It includes all types of substance abuse treatment that may constitute the public burden. The dataset contains detailed records on admissions aged 12 or older, which describe many personal data, such as age, gender, and employment status, and some medical characteristics (substance types, age at first use, frequency of use, etc.). Sufficient data give us many possible combinations of factors and research directions on alcohol abuse

In this paper, a series of descriptive analysis and correlation analysis is first conducted to study the relationship between alcohol abuse and some social and economic factors in New York state. With the help of correlation analysis, we select several factors from TEDS to form a prediction model for New York state that takes patients' employment status, frequency of use, and age as input and predicts the type of medical service the patient may need. After the predictive modeling at the personal level, we attempt to transfer the modeling to the state level. The next model does a similar operation but in a different state.

This paper makes the following contributions:

- The results of descriptive analysis assist us to identify the trend of alcohol abuse in the New York state from 2000 to 2018 and to understand the relation between alcohol disorder and other substance use.
- The correlation analysis provides proof of the relation between alcohol disorder and other substance use. Also, it helps us to select the important factors that form the prediction model
- The prediction model of medical services can effectively forecast the service level for the incoming patients, reducing the cost of testing and evaluating the medical condition.

The paper is organized as follows: Section II state the motivation of this project. Section III summarizes the related work, while section IV and V provide details about the datasets and the design and implementation of the programs. Subsequently, sections VI and VII illustrate the analysis of the modeling results and future work. Finally, the summary is presented in section VIII.

II. Motivation

Alcohol beverage is an important part of American culture, taking an unignorable role in civilians' life. However, the damage it causes can be vast. There are 139.7 million citizens aged 12 or older having alcohol products every month, accounting for 50.8% of the population. 47.1% of them have

binge alcohol use and 11.5% have heavy alcohol use. Also, underage alcohol use is a common issue among the age group of 12 to 20, in which 18.5% are using alcohol per month[2]. These astounding numbers emphasize its impact on the young age population, and it deserves more attention from related administration.

Specifically, the detriment of alcohol not only impairs the growth of brains but also affects other organs, such as livers, hearts, and stomachs. According to the report from WHO, over 200 health conditions are highly related to alcohol disorder, including cancers, cardiovascular diseases, and road injuries. Apart from its damage to the human body, its impacts on society are also unignorable. Car accidents and violent crimes dramatically increase government expenditures on public safety and medical services. Consequently, it imports a heavy burden on economics, which is proved by the 249 million dollars cost of excessive alcohol using, 40.4% of which is paid by the U.S. government[3]

We attempt to analyze the influence of alcohol on different age groups and the relation between it and some social factors. With these results, some suggestions about substance control will be provided. Also, we want to form a prediction model for the health center to forecast the medical service that substance users may need, according to their personal information. In the meanwhile, the patients can estimate the service type before they going to the health department so that they can receive medical assistance as soon as possible without wasting time on consulting and waiting.

III. Related Work

In [4], the authors argue that real-time big data analytics using Hadoop provides meaningful insight and helps predict emergencies by understanding data patterns in the huge amount of data generated by healthcare and government. Such an argument proofs the utility of Hadoop-related frames will be a powerful tool in our projects. In our project, it is Spark MLlib that takes the responsibility for calculation and analysis. MLlib's tight integration with Spark results in several benefits: it is well suited for the large-scale machine learning applications because they are iterative; Spark's vibrant open-source community has led to rapid growth and adoption of MLlib; MLlib is a high-level library on top of Spark. These advantages make analytics efficient and fast.

Another vital component is the selection of mathematical models. In [5], the authors proposed a social media data warehouse of DrugAbuse and AdverseDrugReaction events from tweets on Storm and Hadoop and designed a multidimensional model for conceptual interpretation. While in our project, a multinomial logistic regression model is formed to accomplish the prediction.

Our project aims at analyzing the relationship between alcohol disorder and other personal characteristics, such as employment status and age. Therefore, similar research can be helpful. In Homish et al.'s research[6], authors use multilevel regression models to discover the relationship between the use of illicit drugs and marital satisfaction in the first four years of the marriage. The analysis not only states the couples who use drugs regularly have a relatively huge decline in satisfaction compared with nonuse couples but also shows that male and female users will have different attitudes to marriage depending on if their partners are also addicts. In the tested groups, the discrepant illicit drug use group, in which one of the partners does not use drugs while the other one does, has the most despondent marriage. The marriages in the group are evaluated as the worst by the wives in it, even worse than the marriage of the congruent user group. This result could mean the discrepancies in drug use not only leads to family conflicts but also represents the dissimilarities among the couples.

The research from Lee, J. O. et al. [7] proves the fact that unemployment can be a reason for substance abuse. The authors select the life history calendar (LHC) to obtain the data from participants and use generalized linear mixed modeling to analyze the data. The analysis shows a significant difference in substance use between the adults who grew up in a low childhood SES family and those who did not. A highly correlated relation between drinking and unemployment can be found in adults who experienced low childhood SES. The increase of abuse possibility, 20% to 60%, shows the long term of unemployment is associated with the high possibility of alcohol addiction among these people. However, for those who spent their childhood in a normal SES family, the data shows no relationship between losing jobs and drinking. In our project, unemployment is also an important factor in correlation analysis and predictive modeling

IV. Design and Implementation

The main workflow of our project is divided into 6 parts:

1. Clean and profile the datasets and store them to HDFS
2. Design schema for the dataset in Hive
3. Launch the data analysis using Scala in Spark
4. Transfer the output to Hive
5. Connect tableau with Hive to produce the data visualization

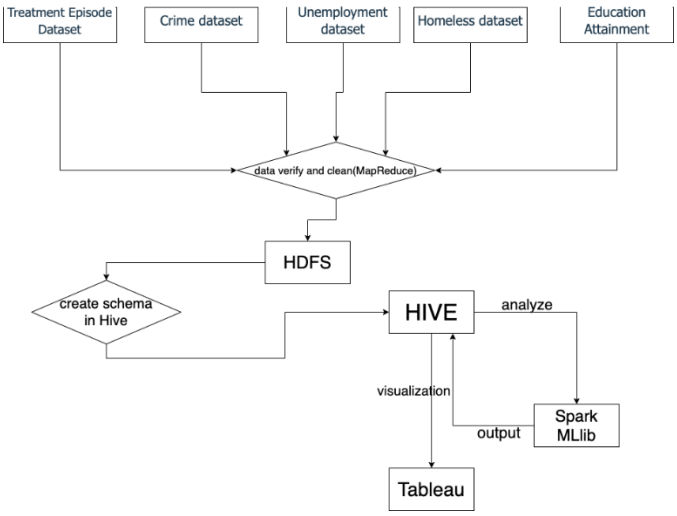


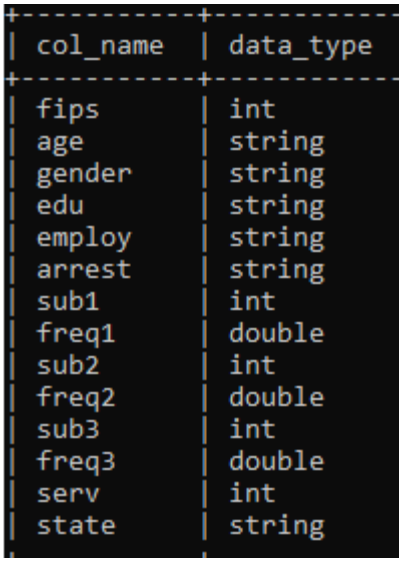
Figure A The design diagram

Data clean:

The datasets we collect contain a large amount of incomplete, incorrect, or irrelevant parts. Therefore, we use MapReduce programs in Java to clean them and get the well-formed datasets. Also, profiling in MapReduce gives us a general understanding of our datasets.

Hive:

Hive is a data warehouse built on top of Apache Hadoop. It could transfer HiveQL scripts to do MapReduce operations. We load our well-formed datasets into and design the data schema for future work.



col_name	data_type
fips	int
age	string
gender	string
edu	string
employ	string
arrest	string
sub1	int
freq1	double
sub2	int
freq2	double
sub3	int
freq3	double
serv	int
state	string

Figure B Example schema in Hive

Spark MLlib:

MLlib is a machine learning library of Spark and it could access all data sources in Hadoop (Hive, HBase, HDFS, etc.), which makes machine learning and data analysis scalable and easy.

V. DATASETS

To accomplish the predictive modeling not only at the personal level (only datasets A, B, and C are needed) but also at the state level, we introduce the datasets D, E, F, and G as complements. Since only a portion of data from them are utilized and the length limit of the paper, detailed introductions are not given in this section.

A. The Treatment Episode Data Set – Admissions

The Treatment Episode Data Set - Admissions (TEDS-A) [8] is a national data system of annual admissions to substance abuse treatment facilities. TEDS-A provides specific data on the personal level of patients admitted to official substance abuse treatment programs. The data in TEDS are preprocessed into a definition table and are represented as numbers in a particular range according to the table. For instance, the education level (EDUC) of each patient is transferred into numbers as Figure C. This reduces the difficulty of screening some data but makes some other data classification unclear. In Figure D, the arrest records in TEDS-A are proof of unclearness. The information is translated to four levels and the final level combines all records over two times such that it is hard to classify those patients by their criminal records.

Value	Label	Unweighted Frequency	%
1	8 YEARS OR LESS	103,837	6.8%
2	9-11	339,613	22.1%
3	12	659,052	42.9%
4	13-15	268,686	17.5%
5	16 OR MORE	95,465	6.2%
-9	MISSING/UNKNOWN/NOT COLLECTED/INVALID	70,372	4.6%

Figure C Classification of education level

Value	Label	Unweighted Frequency	%
0	NONE	1,302,836	84.8%
1	ONCE	97,836	6.4%
2	2 OR MORE TIMES	16,089	1.0%
-9	MISSING/UNKNOWN/NOT COLLECTED/INVALID	120,264	7.8%

Figure D Classification of arrest records

In total, there are 62 types of information are recorded in TEDS-A, which gives researchers many possible directions of data analysis. In this project, we mainly focus on TEDS-A in 2015 that contains 1,537,025 observations. The size of the dataset is 225 MB, and the format is comma-separated values (CSV).

B. Unemployment

A seasonally adjusted unemployment dataset from the U.S. Bureau of Labor Statistics is also used in the project, recording the labor force participation rate, the employment-population ratio, the number of the labor force, and the unemployment rate[9].

Year	Period	labor force participation rate	employment-population ratio	labor force	employment	unemployment	unemployment rate
2010	Jan	59.0	52.1	2144182	1893169	251013	11.7
2010	Feb	58.9	52.1	2143881	1895920	247961	11.6
2010	Mar	58.9	52.2	2142133	1898914	243219	11.4

Figure E Data example in unemployment dataset

C. Average Wage

The average wage comes from the Quarterly Census of Employment and Wages (QCEW), presented by the U.S. Department of Labor[10].

D. Crimes

The Summary Reporting System (SRS) [11]designed by the Federal Bureau of Investigation is used to get crime data at the state level, which is also used in predictive modeling.

population	violent_crime	homicide	rape_legacy
262803276	1798792	21606	97470
265228572	1688540	19645	96252
267783607	1636099	18211	96153
270248003	1533887	16974	93144
272690813	1426044	15522	89411
281421906	1425486	15586	90178
285317559	1439480	16037	90863

Figure F Data example in crimes dataset

E. Educational Attainment

We also use educational attainment [12]from the U.S. census bureau as a complementary dataset when designing the

prediction model. It refers to the highest level of education that a responder has completed. The table records the educational attainment statistics in each state

F. Homelessness

The Annual Homeless Assessment Report (AHAR)[13] is used in this project to provide an estimate of the number of citizens experiencing homelessness in each state, which is an important factor in predictive modeling

G. Mental Health

The dataset of mental health named MH-CLD and MH-TEDS[14] describes the personal data that are highly related to mental health diagnoses and the mental health treatment outcomes and demographic

VI. Results

A. Descriptive Analysis On The Secondary Substance Of Alcohol User

In this section, we analyze the secondary substance of the patients whose primary substance is alcohol using TEDS in 2015. We assume that alcohol abuse may induce some other substance abuse problems.

We study the composition of alcohol abusers' secondary substance with Hadoop MapReduce. As Figure G, 53 percent of alcohol abusers had other substance problems in 2015. Among them, cocaine/crack and marijuana/hashish are the most common problems, both of which account for 20%.

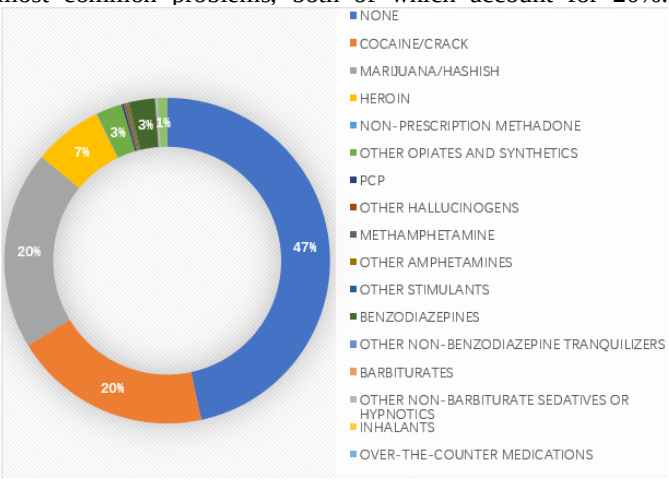


Figure G Result diagram

Also, we evaluate the relationship between the frequency of substance use and alcohol disorder by calculating the correlation coefficients. As shown in Table 1, there are 14 types of coefficients over 0.5. It indicates that alcohol abusers whose frequency of drinking is higher may also consume other substances more frequently.

Table 1

Substance Type	Correlation Coefficient
COCAINE/CRACK	0.72
MARIJUANA/HASHISH	0.65
HEROIN	0.69
NON-PRESCRIPTION METHADONE	0.65
OTHER OPIATES AND SYNTHETICS	0.66
PCP	0.51
OTHER HALLUCINOGENS	0.72
METHAMPHETAMINE	0.61
OTHER AMPHETAMINES	0.7
OTHER STIMULANTS	0.22
BENZODIAZEPINES	0.65
OTHER NON-BENZODIAZEPINE TRANQUILIZERS	0.45
BARBITURATES	0.21
OTHER NON-BARBITURATE SEDATIVES OR HYPNOTICS	0.59
INHALANTS	0.78
OVER-THE-COUNTER MEDICATIONS	0.52
OTHER	0.33

Consequently, the analytic result agrees with our hypothesis that alcohol abuse negatively affects preventing and controlling other substance problems

B. Correlation Analysis On Alcohol Disorder And Social And Economic Factors

The correlation coefficients presented above raise our interest in the influence of alcohol on society. Based on previous research, we make the following assumptions:

1) A positive economic situation contributes to fewer alcohol disorder problems. To verify this, we introduce the average wage in New York state in 2015 as the second variable.

2) Crimes have a strong correlation with the illicit use of alcohol products. We use data from the Summary Reporting System (SRS) presented by the FBI to represent the crimes.

3) Unemployment could lead to serious alcohol abuse.

First, MapReduce programs screen the data representing the personal case of alcohol abuse from 2000 to 2018, compute the population, and transfer them to Hive. A similar operation is done on the dataset of average wage from 2000 to 2018. After these, a correlation analysis is conducted.

As shown in Figure H, the average wage has a noticeable negative correlation with alcohol disorder (correlation coefficients=-0.9109), while the correlation between alcohol abuse and crime rate reaches 0. 9117 in Figure I. These results

agree with the first two assumptions. However, the third result fails to explain our assumption, which shows nearly no correlation between illicit use of alcohol and unemployment by coefficient 0.0059 in Figure J.

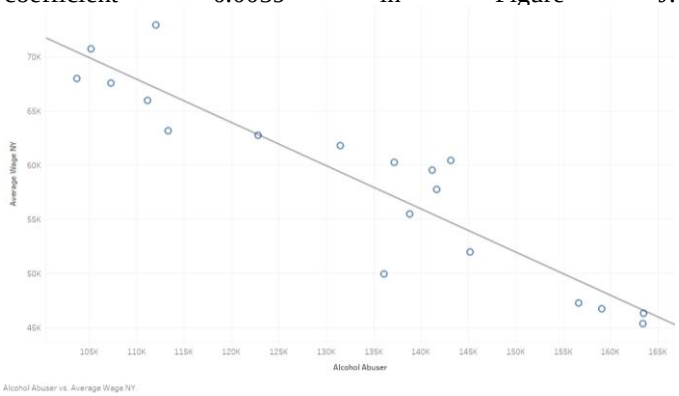


Figure H The correlation between alcohol abuse and average wage

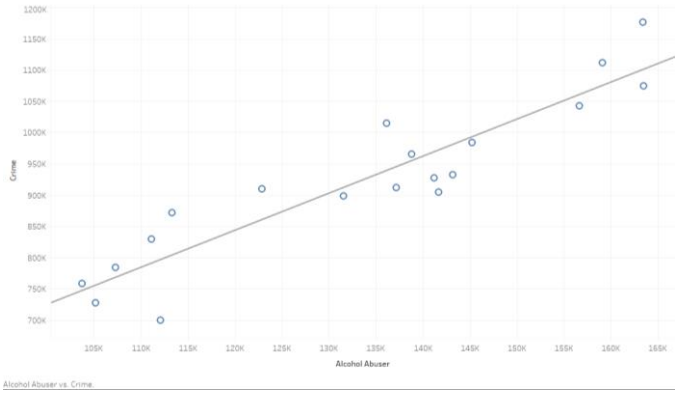


Figure I The correlation between alcohol abuse and crimes

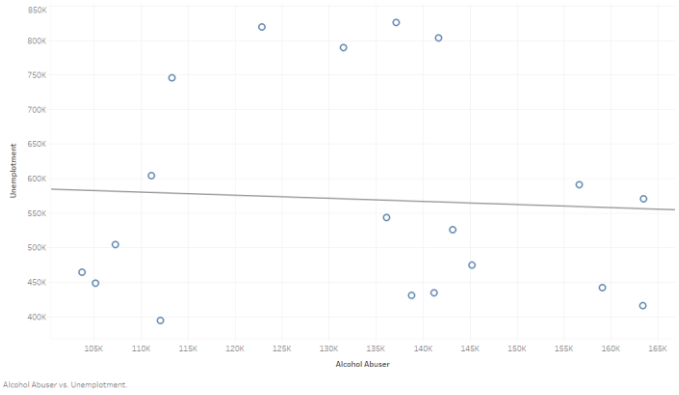


Figure J The correlation between alcohol abuse and unemployment

C. Service Prediction Model for Alcohol Abusers in New York State

In this section, we design a service prediction model for alcohol abusers given personal information and characteristics using Logistic Regression in SparkML. We limit the TEDS records in New York state in 2015 because we assume that different states own different features which may result in differences in the prediction model.

The output of the prediction model is the type of services that abusers would be placed in. The information about services and treatment settings is provided in the field 'SERVSETA'. As Table 2, we divide the detailed services into 3 groups—DETOXIFICATION, REHABILITATION or RESIDENTIAL, AMBULATORY according to the prefix description of SERVSETA value. The value of model output generally coordinates with the intensity of the abuse. To be specific, a smaller value indicates that the client may have more serious alcohol abuse and need more tender health care service

Table 2

Value	Service Type	SERVSETA Value
1	DETOXIFICATION	DETOX, 24 HR, HOSPITAL INPATIENT
		DETOX, 24 HR, FREE-STANDING RESIDENTIAL
2	REHABILITATION/RESIDENTIAL	REHAB/RES, HOSPITAL (NON-DETOX)
		REHAB/RES, SHORT TERM (30 DAYS OR FEWER)
		REHAB/RES, LONG TERM (MORE THAN 30 DAYS)
3	AMBULATORY	AMBULATORY, INTENSIVE OUTPATIENT
		AMBULATORY, NON-INTENSIVE OUTPATIENT
		AMBULATORY, DETOXIFICATION

We pick up several fields from TEDS which are likely to have an impact on the intensity of personal alcohol use, including age, frequency of alcohol use, education level, employment status, and the number of arrests in 30 days. Their correlation coefficients with the service type we define are calculated on Spark, which helps to decide which to use to predict service type.

Table 3

Age	**Frequency**	Education	**Employment**	Arrests
-0.242	-0.636	0.126	-0.373	0.017

As in Table 3, the coefficients of Age, Frequency, and Employment is relatively high compared with others. Consequently, we choose these three factors to predict which service type is suitable for an alcohol abuse client. The definition of these three factors is listed in the following tables.

Value	Age Range
2	12-14
3	15-17
4	18-20
5	21-24
6	25-29
7	30-34
8	35-39
9	40-44
10	Over 45

Value	Frequency Description
1	NO USE IN THE PAST MONTH
2	1-3 TIMES IN THE PAST MONTH
3	1-2 TIMES IN THE PAST WEEK
4	1-2 TIMES IN THE PAST WEEK
5	DAILY

Value	Employment Status
1	FULL TIME OR PART TIME
2	UNEMPLOYED
3	NOT IN LABOR FORCE

Logistic Regression in Spark MLlib is utilized for the classification of services in our project. The accuracy of prediction in the TEDS dataset in New York state in 2015 is 68.5%. There are 107123 records in the dataset.

Users can input the age, frequency of alcohol use, and employment status of a client. The model will predict which type of service the client may need, which also indicates the potential severity of the alcohol abuse problem.

D. Service prediction model for alcohol abusers in the Northeastern United States

In this section, we want to apply this model across states. New York state belongs to the Northeastern United States geographically. The Northeastern region is the nation's most economically developed, densely populated, and culturally diverse region. We assume that the states in the Northeastern region may share the similarity in the prediction model. Northeastern United States contains Connecticut, Maine, Massachusetts, New Hampshire, New Jersey, New York, Pennsylvania, Rhode Island, Vermont, Delaware, Maryland, and Federal districts Washington, D.C.

We introduce some state-level feature to balance the differences between areas: the number of records receiving mental health treatment services in different states, crime rate, unemployment rate, population with a college degree, and sheltered population in our project. These features are likely to associate with alcohol use conventions in different areas and their correlation coefficients are also relatively high (i.e. 0.238, -0.136, -0.135, -0.147, -0.172 respectively), compared with the computation results of other features.

The accuracy of prediction in the TEDS dataset in the Northeastern United States in 2015 is 70.8%. There are 195906 records in the dataset. And it works better than the model without state-level features whose accuracy is 69.21%.

VII. Future Work

Our future work of this paper will involve the research into expanding the prediction model to the whole United States. More advanced statistical methods will be used to define whether factors are appropriate for the prediction and the results will be evaluated more accurately. Additionally, we will try some high-level machine learning algorithms to achieve better performance.

VIII. Conclusion

In this paper, we present an analysis of alcohol abuser records in the TEDS. First, we do some descriptive research on the characteristics of secondary abuse problems of alcohol abusers. The result shows that it is necessary to attach importance to illicit drug use of alcohol abusers. Second, we analyze the correlation between the population of alcohol abusers in New York state and other social factors. Finally, we implement a service type prediction model using Logistic Regression for patients in New York. The model uses fields in TEDS to predict which service is appropriate for a patient. Also, we expand the prediction model to the Northeastern United States by adding some state-level features. The modified model shows a positive result.

Acknowledgment

This work was supported by NYU HPC, Prof. Lakshminarayanan Subramanian, and Prof. Suzanne Mcintosh, in part by Substance Abuse and Mental Health Services Administration (SAMSHA), in part by United States Census Bureau, and in part by Tableau.

Reference

[1] CDC. "Deaths from Excessive Alcohol Use in the U.S." https://www.cdc.gov/alcohol/features/excessive-alcohol-deaths.html (accessed.

[2] U. S. D. o. H. a. H. Services. "Key Substance Use and Mental Health Indicators in the United States: Results from the 2019 National Survey on Drug Use and Health." https://www.samhsa.gov/data/sites/default/files/reports/rpt29393/2019NSDUHFFRPDFWHTML/2019NSDUHFFR090120.htm (accessed 12/11, 2020).

[3] WHO. "Global status report on alcohol and health 2018." https://www.who.int/publications/i/item/9789241565639 (accessed 11/17, 2020).

[4] J. Archenaa and E. A. M. Anita, "A Survey of Big Data Analytics in Healthcare and Government," *Procedia Computer Science*, vol. 50, pp. 408-413, 2015/01/01/ 2015, doi: https://doi.org/10.1016/j.procs.2015.04.021.

[5] F. Jenhani, M. S. Gouider, and L. B. Said, "Streaming Social Media Data Analysis for Events Extraction and Warehousing using Hadoop and Storm: Drug Abuse Case Study," *Procedia Computer Science*, vol. 159, pp. 1459-1467, 2019/01/01/ 2019, doi: https://doi.org/10.1016/j.procs.2019.09.316.

[6] G. G. Homish, K. E. Leonard, and J. R. Cornelius, "Illicit drug use and marital satisfaction," (in eng), *Addict Behav*, vol. 33, no. 2, pp. 279-291, 2008, doi: 10.1016/j.addbeh.2007.09.015.

[7] J. O. Lee *et al.*, "Unemployment and substance use problems among young adults: Does childhood low socioeconomic status exacerbate the effect?," (in eng), *Soc Sci Med*, vol. 143, pp. 36-44, 2015, doi: 10.1016/j.socscimed.2015.08.016.

[8] SAMDHA. "Treatment Episode Data Set: Admissions 2015 (TEDS-A-2015-DS0001)." https://www.datafiles.samhsa.gov/study-dataset/treatment-episode-data-set-admissions-2015-teds-2015-ds0001-nid17208 (accessed 11/25, 2020).

[9] "Local Area Unemployment Statistics." https://data.bls.gov/PDQWeb/la (accessed 12/07, 2020).

[10] "Quarterly Census of Employment and Wages (QCEW)." https://labor.ny.gov/stats/lsqcew.shtm (accessed 12/11, 2020).

[11] "Summary (SRS) Data with Estimates." https://crime-data-explorer.fr.cloud.gov/downloads-and-docs (accessed 11/23, 2020).

[12] "EDUCATIONAL ATTAINMENT." https://data.census.gov/cedsci/table?q=s1501&tid=ACSST1Y2018.S1501&hidePreview=false (accessed 12/14, 2020).

[13] "2019 AHAR: Part 1 - PIT Estimates of Homelessness in the U.S." https://www.hudexchange.info/resource/5948/2019-ahar-part-1-pit-estimates-of-homelessness-in-the-us/ (accessed 12/15, 2020).

[14] "Mental Health Client-Level Data (MH-CLD)." https://wwwdasis.samhsa.gov/dasis2/mhcld.htm (accessed 11/28, 2020).