**A Statistical Approach to Stratify Admissions for Drug Abuse in a Community Residential Setting**

IUPUI, Indianapolis IN 46202, USA

Abstract:The use of illicit drugs has increased over the last decade from 8.3% of the population using illicit drugs in the past month in 2002 to 10.2% (27 million people) in 2014. Of those, 7.1 million people met criteria for an illicit drug use disorder in the past year. The misuse of prescription drugs is second only to marijuana as the nation’s most common drug problem after alcohol and tobacco, leading to troubling increases in opioid overdoses in the past decade. The main purpose of the study is to understand the instances of drug abuse which can be valuable for regulatory agencies to implement preventive measures to reduce drug abuse. Data is collected from ‘SAMHSA’, The Treatment Episode Data Set -- Admissions (TEDS-A) which is a national data system of annual admissions to substance abuse treatment facilities. Analysis was performed using Recursive feature elimination, Logistic Regression and Support Vector Machine. Our logistic regression model gave 86 percent accuracy. The results of our model depicted that, there is an association between age, gender, occupation, number of substance abused, Psychiatric problem and primary reason for admission.

**1 Introduction**

Drug addiction, both illicit and prescription drugs, represents a troubling trend in the United States. According to the National Institute on Drug Abuse, drug addiction is characterized “as a chronic, relapsing brain disease that is characterized by compulsive drug seeking and use, despite harmful consequences”("The Science of Drug Abuse and Addiction: The Basics,").  Honing in on illicit drug use, the numbers are staggering. According to Substance Abuse and Mental Health Services Administration’s (SAMHSA) National Survey on Drug Use and Health (NSDUH) – 2014, the percentage of those 12 or older in the United States that have  used of illicit drugs, has risen from 8.3% in 2002 to 10.2% in 2014. This value of 10.2% represents approximately 27 million people(Hedden, 2015). When this issue expands to include prescription drug abuse, the data becomes even more staggering.   Alcohol, nicotine, illicit and prescription drug abuse represents a wastage of nearly 700 billion dollars annually, which is represented by increased health costs, crime, and lost productivity.  Perhaps the gravest and most tragic part of this crisis is the health and social ramifications of this situation. Sadly, every year over 90,000 Americans lose their lives due as result of the drug addiction and abuse of both illicit and prescription drugs as well as from alcohol addiction and abuse("The Science of Drug Abuse and Addiction: The Basics,"). While the aforementioned facts on the pervasiveness and economic impact of this situation flags our attention to investigate this issue, the social and health ramifications represent the factors that impel us most to further delve into this issue.

Our primary aim in our research project is to develop a model to predict admissions among substance abuse patients. A corollary aim is to analyze the predominant substance abused at state level and provide insights for regulatory agencies to reduce admissions. The goal of our project is to assess data at State level and assist the regulatory agencies to implement preventive measures for substance abuse.

**2 Methodology**

Our null hypothesis is that there is no relation between independent variables (such as age, race, ethnicity) and the patients being admitted for either marijuana, cocaine, or heroin. The alternative hypothesis is that there is in face a relation between these (and other) independent variables and the admission for those three substances.

**2.1 Data Collection**   
The data for this study is take from, Treatment Episode Data Set -- Admissions (TEDS-A) is a national data system of annual admissions to substance abuse treatment facilities. State laws require substance abuse treatment programs to report their publicly-funded admissions to the state. Some states collect only their publicly-funded admissions while other states are able to collect their privately-funded admissions from facilities that receive public funding. States then report these data from their state administrative systems to SAMHSA. The resulting data system is referred to as TEDS-A.

**2.2 Data Storage**Data was stored in SQl database. Original data set 1.6 mil record. Given the size of the Data, Data was then randomly selected using code in Python. As the result we stored and used 500k randomly selected data set in creating our model to predict readmissions.

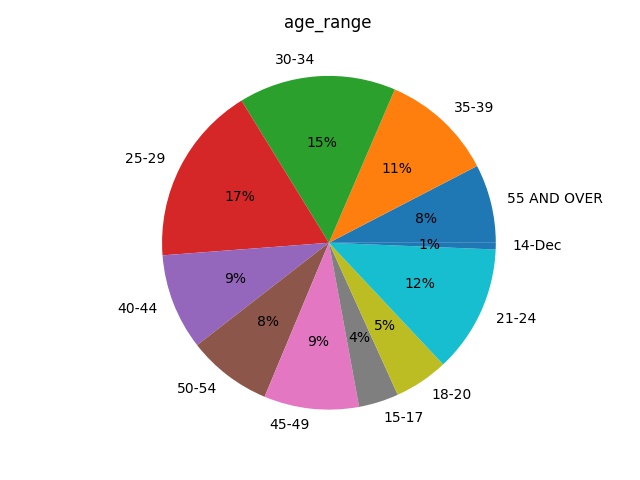
**2.3 Data Extraction**For data extraction we were able to use Python. The data set initially contained 60 variables like age, gender, race, ethnicity, education, type of substance abused etc. to develop a predictive model. 18 variables were selected for analysis. Categories in each variable and their composition are described here.

**DRUG ADMISSION DATA.** The states below are colored as per admission rate with the darkest color having highest rate.

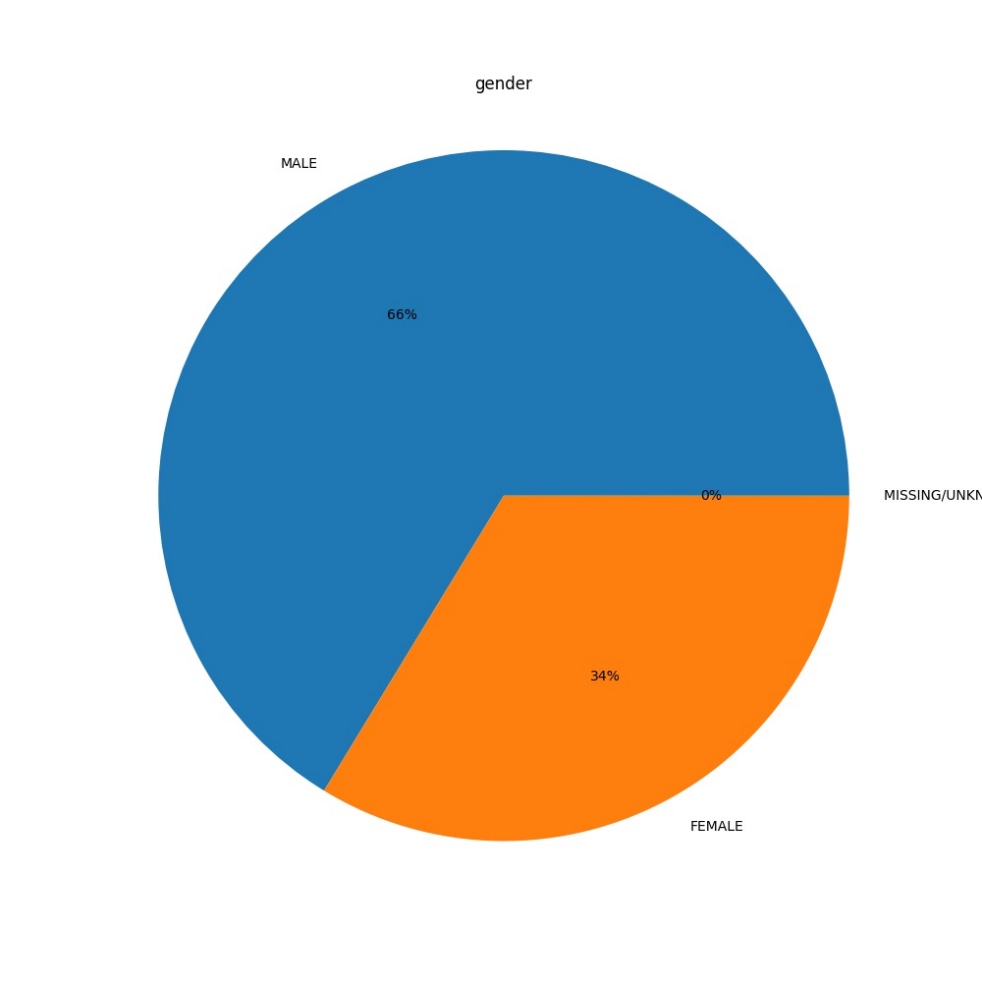
A close up of a map

Description generated with high confidence

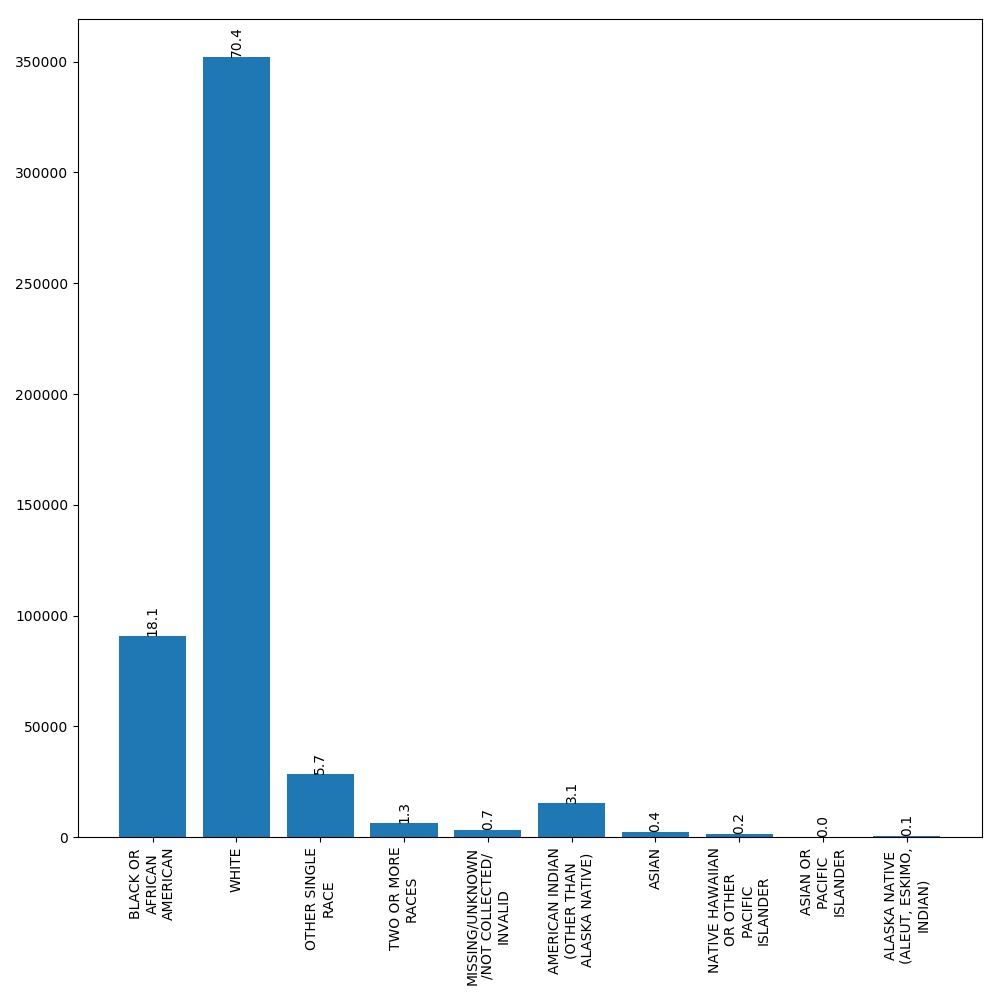
**AGE.** Out of the total population taken for the study, the age group which had maximum number of admissions were among 25-29yrs(17%) followed by 30-34(15%), 21-24(12%), 35-39(11%), 40-44(9%), 45-49(9%), 50-54(8%), 55 and over(8%) and 15-17yrs(4%) respectively.



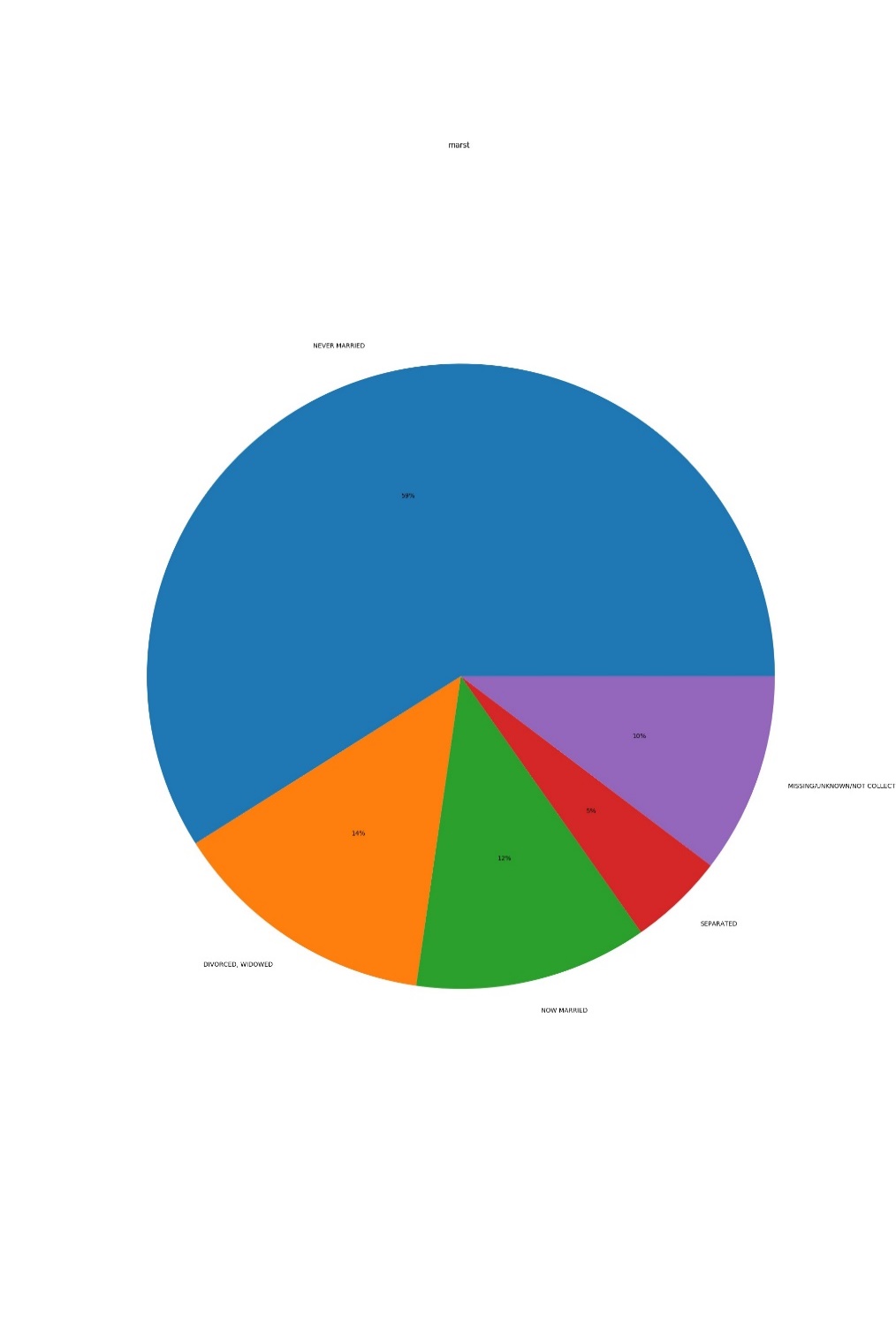
**GENDER.** Out of the total population taken for the study, the gender which had maximum number of admissions were male (66%) compared to female (34%)



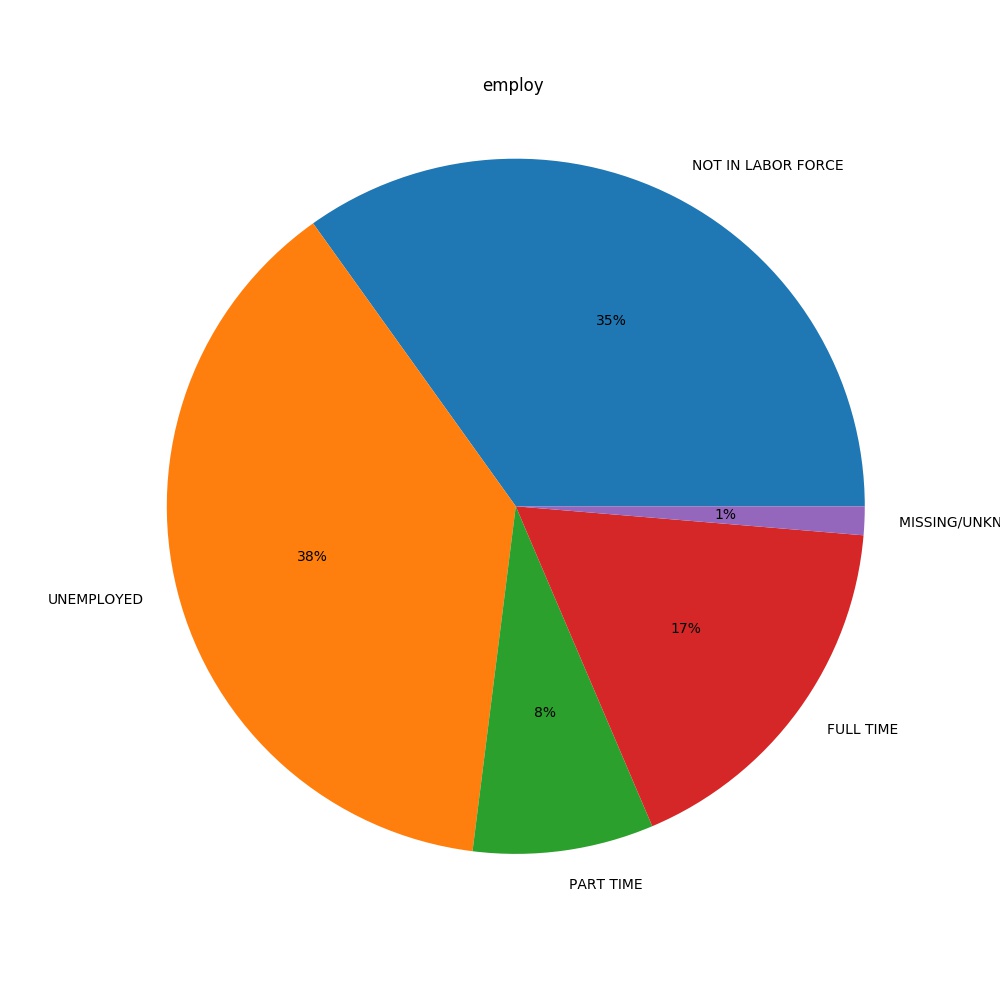
**RACE.** Out of the total population taken for the study, the group of race which had maximum number of admissions were Whites (70.382%) followed by black African Americans (18.130%), other single races (5.7%), American Indians (3.06%), two or more races (1.28%), Asians (0.42%), native Hawaiian (0.24%) and other Asian and Alaskan population (0.1%)



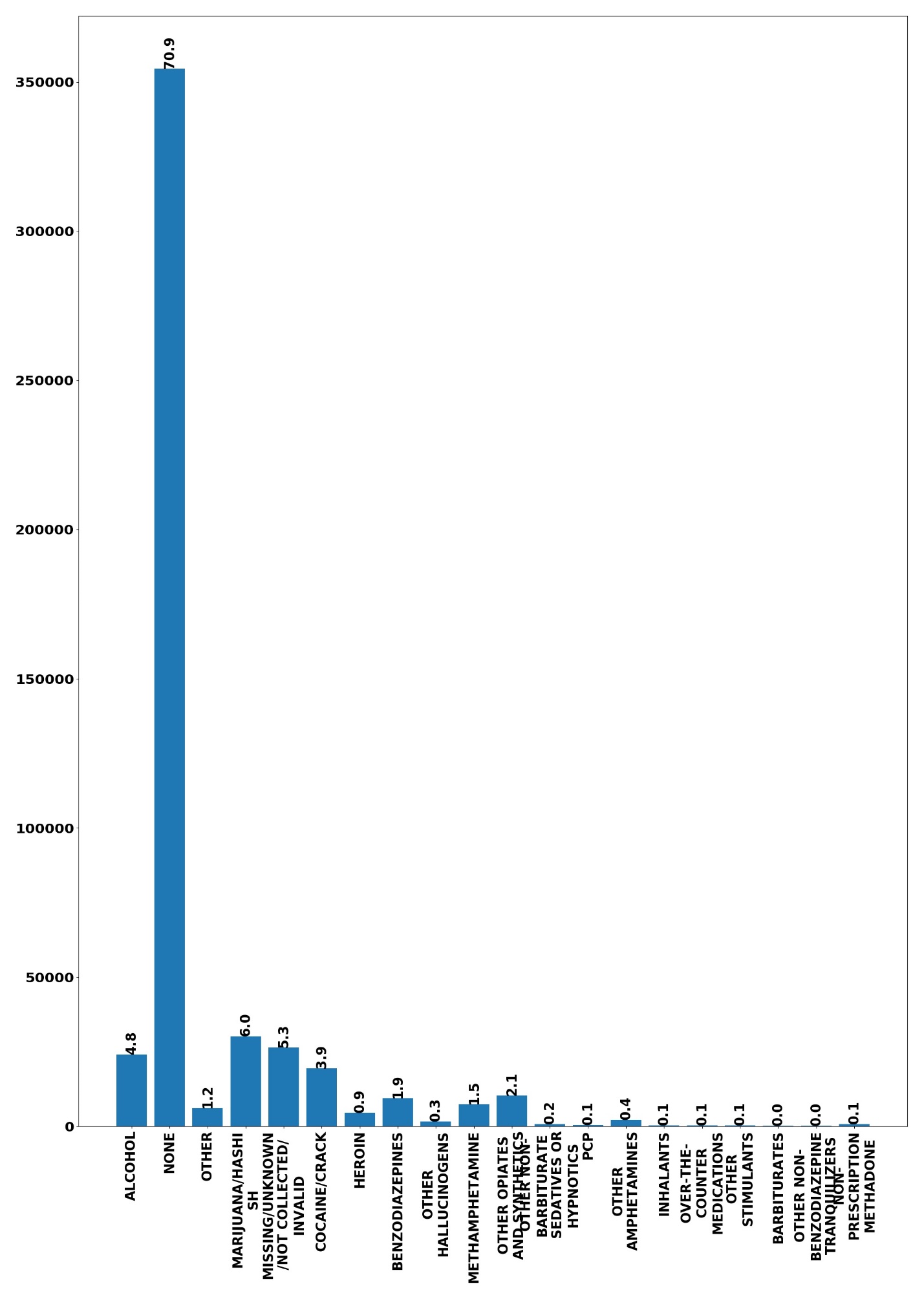
**MARITAL STATUS.** Out of the total population taken for the study, the group which had maximum number of admissions were never married (59%), followed by divorced (14%), now married (12%), missing (10%), separated (5%).



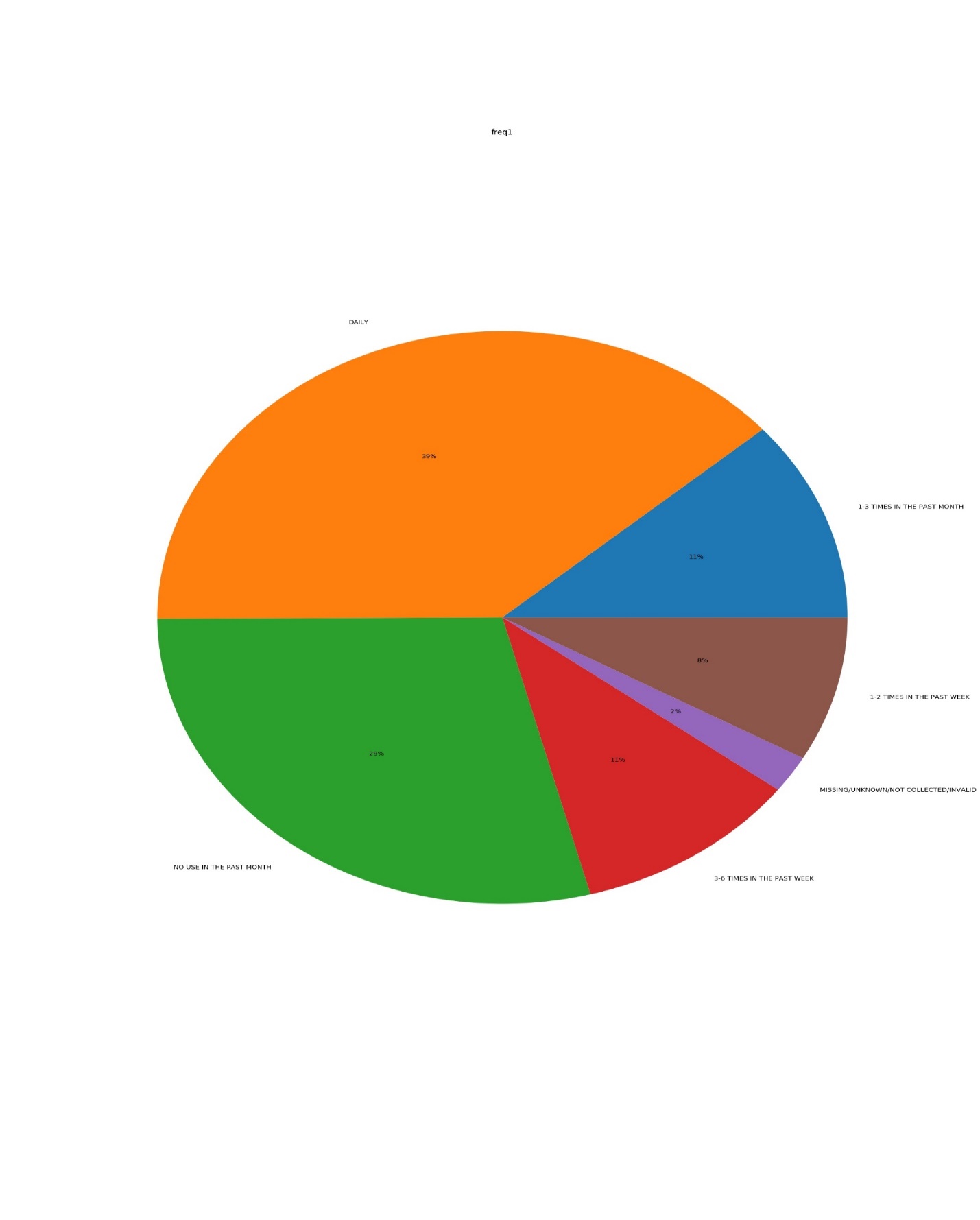
**EMPLOYMENT.** Out of the total population taken for the study, the group which had maximum number of admissions were unemployed(38%) followed by not in labor force(35%), full time(17%), parttime(8%) and missing(1%) respectively



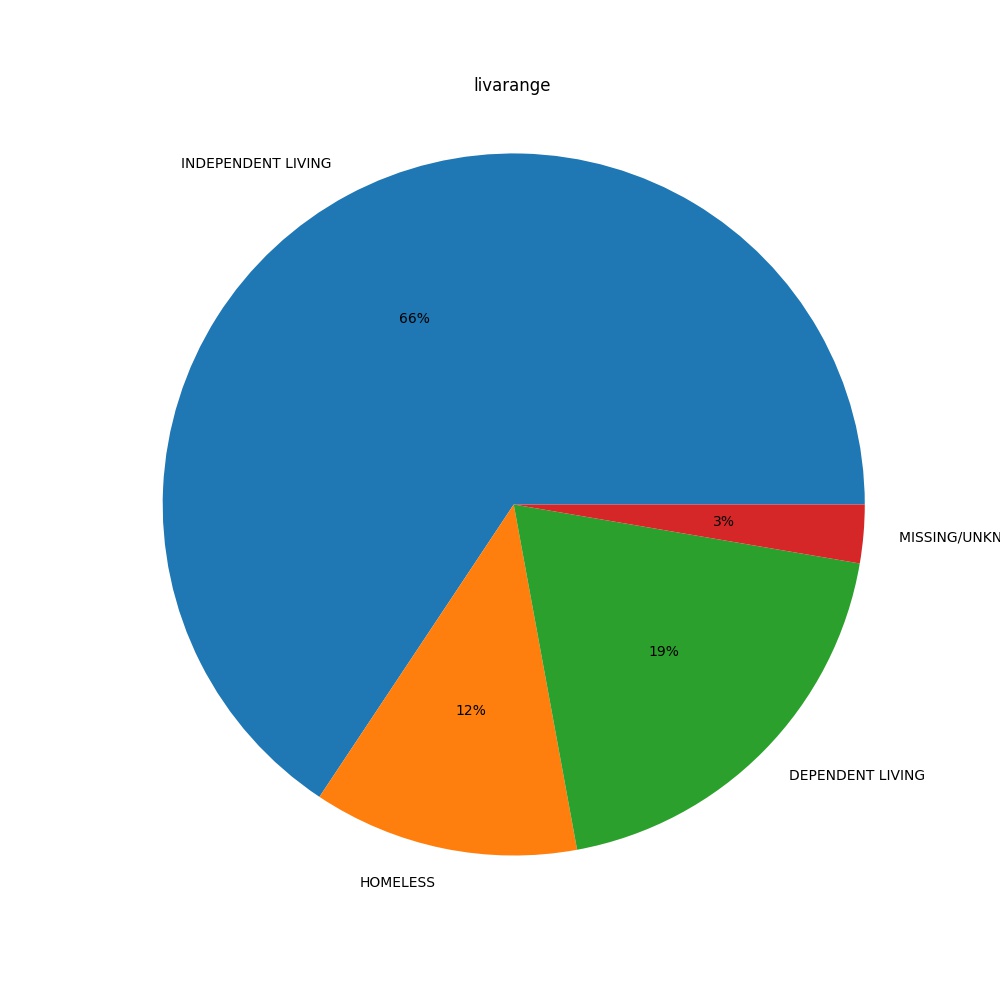
**SUBSTANCE ABUSED.** Out of the total population taken for the study, the group which had maximum use of substance as none(70.89%) followed by marijuana(6.04%), unknown drugs(5.92%), alcohol(4.83%), cocaine(3.93%), other opiates(2.06%), benzodiazepines(1.97%), methamphetamine(1.48%), heroine(0.89%) and other drugs like non barbiturates, PCP, stimulants (0.829%) respectively



**FREQUENCY OF SUBSTANCE ABUSE.** Out of the total population taken for the study, the group which had maximum number of admissions were having the frequency of abuse as daily(39%) followed by no use in past one year(29%), 3-6 times used in past week(11%), 1-3 times used in past week(11%), missing(2%)respectively.



**LIVING RANGE.** Out of the total population taken for the study, the group which had maximum number of admissions were having independent living (66%) followed by dependent living (19%), homeless (12%), missing (3%) respectively



**2.4 Data Analysis**For analysis we used Python to perform Logistic regression and support vector machines to develop a predictive model to analyze data and predict readmissions was either due to Heroin/Marijuana/Cocain or remaining others.

First, we connected the database with MySQLdb, then read the data in the database into a data frame DF. We then performed the cleaning of the data, and dropped the un-related columns. And we found out of 60 variables, 18 variables are related to the research were selected. As each variable was having too many categories, categories were merged together based on the literature. Furthermore, we created dummy variables on dataset, in order to create columns for logistic regression with binary values. We then dropped missing values from dataset, to allow predictive model to perform prediction accurately. In addition, we implemented Recursive feature elimination to consider smaller and smaller set of features. With the features from RFE, we performed spearman’s correlation and assessed P-values. we built logistic regression model and support vector machine model. Accuracy of both the models were assessed. After which 10-fold cross validation was performed to further assure the accuracy of the model. Confusion matrix, sensitivity, specificity and F-scores were calculated. ROC curves were drawn for both the models to assess AUC. Finally, the weights of all variables used in the model were calculated and found thesignificance of a variable in the model built.

Figure 1 - Creating Dummy Variables  
A screenshot of text

Description generated with very high confidence

Figure 2 - Drop missing values  
A screenshot of a social media post

Description generated with very high confidence

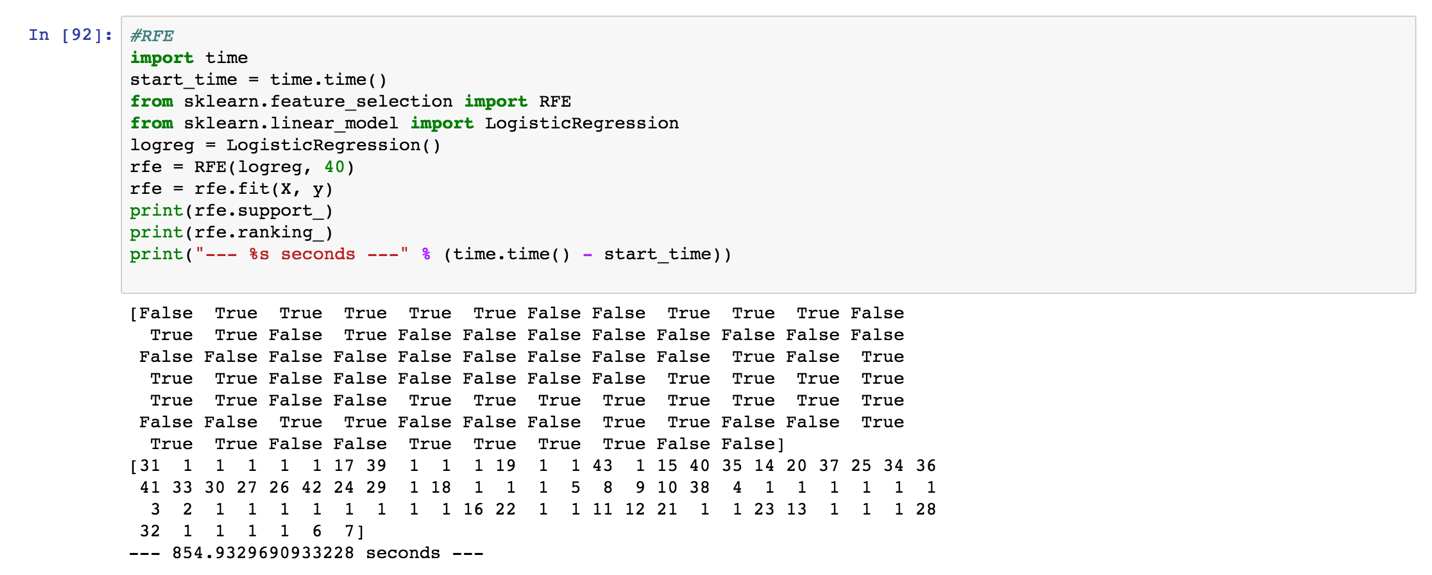
Figure 3 - Recursive Feature Elimination (RFE)  


Figure 4 - Logistic Regression Model  
A screenshot of a cell phone

Description generated with high confidence

Figure 5- Confusion Matrix  
A screenshot of a cell phone

Description generated with very high confidence

Figure 6 - ROC Curve  
A screenshot of a cell phone

Description generated with high confidence

Figure 7 - Variable Weights  
A screenshot of a cell phone

Description generated with very high confidence

The logistic regression model was tested on 2011, 2012 and 2013. The code used for analysis follows and the results are shown

**2013**

*Figure 8 – Logistic Regression for 2013*

A screenshot of text

Description generated with very high confidence

A screenshot of a social media post

Description generated with very high confidence

A screenshot of a map

Description generated with high confidence

*Figure 9 – Logistic Regression results for 2012*

A close up of a map

Description generated with high confidence

*Figure 10 – Logistic Regression results for 2011*

**A close up of a map

Description generated with high confidence**

**Support Vector Machine (SVM)**

Results of the logistic model developed on 2014 was compared against the Support vector machine model.

*Figure 11– SVM*

**A screenshot of a social media post

Description generated with very high confidence**

**3 Results**

From the results, it was identified as the accuracy of our model in 86%. When 10-fold cross validation was performed, the accuracy of the model is identified close to our logistic regression model i.e., 85.8%. The model is also supported by high precision and recall scores.

Also, interpreting the confusion matrix, True positives and True negatives were high relative to True negative and False positive scores. When the ROC curve is interpreted, high AUC score supports the model.

After which, we identified the weights of the available data searching for the elements that have the strongest positive and negative correlation to admittance for substance abuse. Some of the most indicative elements in terms of positive correlation to admission include:

• Daily use

• Alcohol as a secondary or tertiary abused substance

• Ethnicity

Some of the driving indicators with a negative correlation to admission include:

• Just use of alcohol (no other substances)

If a random selection of a record from the dataset was pulled, there is a 43% chance that the record would indicate an admission. Attempting to identify the factors that could predict an admission for substance abuse, our model shows an 86% accuracy in identifying the population based on the other available variables. This is confirmed by an 85.8% accuracy when executing a 10-fold cross validation. Thereby, we were able to increase the predictive ability of our model to nearly double the accuracy of a random selection baseline.

When we tested our model on 2011, 2012 and 2013 datasets, the accuracy was observed to be 58%, 59% and 59%. The precision, recall and AUC overall all the years was found to be between 54% to 60%. This implies that the model built could more accurately predict the admissions than their baseline accuracy.

In the SVM model, the accuracy was found to be like that of logistic regression i.e., 85%. It was supported by 10-fold cross validation score of 86%. The precision and recall scores were similar to that of logistic regression. So, on our data, both the models were identified to perform similarly.

Working as a team

Our team has people with various backgrounds and work experiences. Most of them are not relevant with the skills necessary for the analysis despite of which we were able to finish all the tasks proposed in the project proposal. Every team members contributed to finish the project. With each person having different strengths, every did well throughout the project. The following were the tasks fulfilled by each team member:

Individual contributions:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Name | Background study | Problem design | Data collection | Data extraction | Data analysis | Document writing | Presentation |
| Anita | X | X |  |  |  | X | X |
| Asha | X | X | X |  | X | X |  |
| Hassan |  | X |  | X | X |  | X |
| Eric | X |  | X | X |  | X | X |
| Parvathi |  | X | X | X | X |  | X |
| Siddhartha |  |  | X | X | X | X |  |

4 Discussion and Conclusion

At the superficial level, our group determined that we reject the null hypothesis. There was a statistically significant association between hospital admission due to heroin/marijuana/cocain and variables like age, gender, and so on in the data set. At a deeper level, our paper raises the deep and disturbing trends of the rising incidence of substance use and abuse in the United States. The goal of this paper was to analyze the data and provide the strongest predictors of hospital admission for drug use and abuse. This data should be relayed to regulatory agencies to implement preventive measures for substance abuse. However, ultimately solving this issue requires the coordinated and passionate involvement from various segments within society, including primary care offices, emergency departments, educational settings, and the judicial court system. We believe if the data is relayed in a timely and effective manner to all of the aforementioned groups, it may help reduce the extraneous economic cost, social anguish, and most importantly, loss of life.

Overall, we believe that our data collection, analysis, and conclusions were fairly well founded and discussed. However, it is entirely possible that there may be limitations to our study. One limitation may be evidenced in the data storage component of our study. While we truly believe that our data is strongly representative of all demographic segments, it may be possible that because of the limited storage, we may not have been able to collect every data point from every sub-segment of the population. We believe that with more data storage, we may have collected data and reduce the narrow possibility of missing any sub-segment of the population.

Limitations:

The main limitation of the study observed was the size of the data, which is 2 GB with 1.6 million rows and 60 columns. Other constraints were data cleaning that includes removing 40 unnecessary columns which were not related to admissions. After data cleaning, performing RFE took a bit more time than for regular size data. Also, building SVM was the hardest part which took approximately 5 hours to build. Because of which, it was not able to perform tests on 2011, 2012 and 2013 data.

Appendix

**Codes:**

**A.1 SQL connection:**

import MySQLdb

conn = MySQLdb.connect(host="localhost", user="snuthakk", passwd="6115096", db="group2\_db")

cursor = conn.cursor() #cursor allows you to execute statements

cursor.execute('select \* from admission\_records');

rows = cursor.fetchall()

print(len(rows))

**A.2 Reading into Dataframes:**

x = list(rows)

import pandas as pd

df = pd.DataFrame(x, columns=['id', 'caseid', 'year', 'age\_range', 'gender', 'race', 'ethnicity', 'marst', 'educ', 'employ', 'detnlf', 'preg', 'vet', 'livarange', 'primincome', 'arrests', 'stfips', 'cbsa', 'region', 'division', 'servseta', 'methuse', 'daywait', 'psource', 'detcrim', 'noprior', 'sub1', 'route1', 'freq1', 'frstuse1', 'sub2', 'route2', 'freq2', 'frstuse2', 'sub3', 'route3', 'freq3', 'frstuse3', 'numsubs', 'idu', 'alcflg', 'cokeflg', 'marflg', 'herflg', 'methflg', 'opsynflg', 'pcpflg', 'hallflg', 'mthamflg', 'amphflg', 'stimflg', 'benzflg', 'trnqflg', 'barbflg', 'sedhpflg', 'inhflg', 'otcflg', 'otherflg', 'alcdrug', 'dsmcrit', 'psyprob', 'hlthins', 'primpay'])

df

df.columns = ['ID', 'CASEID', 'YEAR', 'AGE', 'GENDER', 'RACE', 'ETHNIC', 'MARSTAT', 'EDUC',

'EMPLOY', 'DETNLF', 'PREG', 'VET', 'LIVARAG', 'PRIMINC', 'ARRESTS',

'STFIPS', 'CBSA', 'REGION', 'DIVISION', 'SERVSETA', 'METHUSE',

'DAYWAIT', 'PSOURCE', 'DETCRIM', 'NOPRIOR', 'SUB1', 'ROUTE1', 'FREQ1',

'FRSTUSE1', 'SUB2', 'ROUTE2', 'FREQ2', 'FRSTUSE2', 'SUB3', 'ROUTE3',

'FREQ3', 'FRSTUSE3', 'NUMSUBS', 'IDU', 'ALCFLG', 'COKEFLG', 'MARFLG',

'HERFLG', 'METHFLG', 'OPSYNFLG', 'PCPFLG', 'HALLFLG', 'MTHAMFLG', 'AMPHFLG', 'STIMFLG', 'BENZFLG', 'TRNQFLG', 'BARBFLG', 'SEDHPFLG','INHFLG', 'OTCFLG', 'OTHERFLG', 'ALCDRUG', 'DSMCRIT', 'PSYPROB','HLTHINS', 'PRIMPAY']

df

**A.3 Data Cleaning:**

#Drop unnecessary columns

df1 = dfx.drop(dfx.columns[[0,1,2,6,10,11,12,15,16,17,19,20,21,22,23,24,27,31,32,33,35,36,37,39,40,41,42,43,44,45,46,47,48,49,50,51, 52, 53, 54, 55, 56,57,59,61,62]], axis=1)

df1.head(3)

**A.4 Creating Dummy Variables:**

df4 = pd.get\_dummies(df3, columns = ['AGE', 'GENDER', 'RACE', 'MARSTAT', 'EDUC', 'EMPLOY', 'LIVARAG',

'PRIMINC', 'REGION', 'NOPRIOR', 'SUB1', 'FREQ1', 'FRSTUSE1', 'SUB2',

'SUB3', 'NUMSUBS', 'ALCDRUG', 'PSYPROB'])

df4

**A.5 Drop missing values:**

#Drop columns with missing values

df5 = df4.drop(df4.columns[[0,8,13,17,21,26,28,35,36,52,55,60,63,68,76,86,97]], axis=1)

df5.shape

for i in enumerate(df5.columns):

print (i)

**A.6 Recursive Feature Elimination:**

#RFE

import time

start\_time = time.time()

from sklearn.feature\_selection import RFE

from sklearn.linear\_model import LogisticRegression

logreg = LogisticRegression()

rfe = RFE(logreg, 40)

rfe = rfe.fit(X, y)

print(rfe.support\_)

print(rfe.ranking\_)

print("--- %s seconds ---" % (time.time() - start\_time))

**A.7 Spearman Correlation:**

corr = df7.corr(method = 'spearman')

corr

**A.8 Logistic Regression:**

#Logistic regression

from sklearn.cross\_validation import train\_test\_split

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.3, random\_state=0)

from sklearn.linear\_model import LogisticRegression

from sklearn import metrics

logreg = LogisticRegression()

logreg.fit(X\_train, y\_train)

y\_pred = logreg.predict(X\_test)

print('Accuracy of logistic regression classifier on test set: {:.2f} '.format(logreg.score(X\_test, y\_test)))

from sklearn import model\_selection

from sklearn.model\_selection import cross\_val\_score

kfold = model\_selection.KFold(n\_splits=10, random\_state=7)

modelCV = LogisticRegression()

scoring = 'accuracy'

results = model\_selection.cross\_val\_score(modelCV, X\_train, y\_train, cv=kfold, scoring=scoring)

print("10-fold cross validation average accuracy: %.3f" % (results.mean()))

**A.9 Confusion Matrix:**

from sklearn.metrics import confusion\_matrix

confusion\_matrix = confusion\_matrix(y\_test, y\_pred)

print(confusion\_matrix)

**A.10 ROC**

#Logistic ROC

from sklearn.metrics import roc\_auc\_score

import matplotlib.pyplot as plt

from sklearn.metrics import roc\_curve

logit\_roc\_auc = roc\_auc\_score(y\_test, logreg.predict(X\_test))

fpr, tpr, thresholds = roc\_curve(y\_test, logreg.predict\_proba(X\_test)[:,1])

plt.figure()

plt.plot(fpr, tpr, label='Logistic Regression (area = %0.2f)' % logit\_roc\_auc)

plt.plot([0, 1], [0, 1],'r--')

plt.xlim([0.0, 1.0])

plt.ylim([0.0, 1.05])

plt.xlabel('False Positive Rate')

plt.ylabel('True Positive Rate')

plt.title('Receiver operating characteristic')

plt.legend(loc="lower right")

plt.savefig('Log\_ROC')

plt.show()

A.11 Variable weights:

#Calculate weights

b = logreg.coef\_

print(b)

#Calculate weights

b1 = b[0].tolist()

print(b1)

#Calculate weights

b2 = sorted(zip(b1, cols), reverse=True)

b3 = pd.DataFrame(b2, columns=['Coef', 'Variables'])

b3

**REFERENCES**

Hedden, S. L., Kennet, J., Lipari, R., Medley, G., Tice, P., Copello, E.A.P., Kroutil, L.A., Hunter, D. (2015). Behavioral Health Trends in the United States: Results from the 2014 National Survey on Drug Use and Health. In (September 2015 ed.). SAMHSA.gov: Substance Abuse and Mental Health Services Administration.

The Science of Drug Abuse and Addiction: The Basics. Retrieved from <https://www.drugabuse.gov/publications/media-guide/science-drug-abuse-addiction-basics>

Swensen, I. D. (2015). Substance-abuse treatment and mortality. *Journal of Public Economics, 122*(Supplement C), 13-30. doi:<https://doi.org/10.1016/j.jpubeco.2014.12.008>

Smith, M. W., Stocks, C., & Santora, P. B. (2015, February). Hospital readmission rates and emergency department visits for mental health and substance abuse conditions. Retrieved October 02, 2017, from <https://www.ncbi.nlm.nih.gov/pubmed/25563483>