Investigating the relationship between age of preliminary exposure to substances and likelihood of trying methamphetamine

Trevor Fush

1 Introduction

According to the CDC, the United States is in a “Drug Overdose Epidemic”, where the amount of deaths due to drug overdoses is not only high, but is also increasing with time. There were a reported 70,630 deaths due to drug-involved overdoses in 2019 alone (National Institute on Drug Abuse). This said, it is important to understand why people are being drawn to these deadly substances, and find a way to counteract these tendencies. This project attempts to use the age at which individuals are first exposed to substances such as alcohol, marijuana, and cigarettes to predict whether or not they are likely to use methamphetamine in the future, using K-Nearest Neighbors (KNN) classification models in addition to Support Vector Machine (SVM) classification models to attempt to accomplish this task.

2 Data

In 2016, the National Survey on Drug Use and Health (NSDUH) surveyed a target population of “the civilian, noninstitutionalized population of the United States (including civilians living on military bases) who were 12 years of age or older at the time of the survey,” (National Survey on Drug Use and Health 2016). The survey maintains the goal of tracking trends in specific substance use and mental illness measures, and assesses the consequences of these conditions by examining mental and/or substance use disorders and treatment for these disorders. The dataset resulting from this survey that is used in this project contains 56,897 observations (or responses) for 2668 separate variables (corresponding to answers from survey questions).

2.1 Previous Work

Previously published literature summarized the same data from the year 2012, and analyzed statistics for specific age groups and their use of illicit substances. Barry-Jester and Flowers found that among baby boomers (individuals aged between 50 and 64), they accounted for less overall substance use than their younger counterparts, and they also found that alcohol and marijuana tended to be the most commonly used drugs of all those included in the survey. They also provide more specific statistics as to the amounts of drugs that baby boomers consume compared to younger members of society (Barry-Jester, A. M., & Flowers, A.).

Because the previously published work only provided a surface level analysis of the survey data, and it was based on data from 2012, this project is targeted to provide a more in depth investigation into a particular subset of the large survey data from 2016, which will be discussed in the next section. In addition, the data produced by fivethirtyeight and used by Barry-Jester and Flowers with the 2012 data was replicated for the 2016 survey that is being used in this study, and the results can be viewed in Appendix A.

2.2 Filtering Data

Due to the large number of features in this dataset, this project is designed to focus on specific features that are of higher importance to the goals of the investigation. The relevant features and their corresponding names in the survey dataset that were selected for this study were:

1. *cigtry* → Age when individual first smoked a cigarette
2. *alctry* → Age when individual first drank alcoholic beverage
3. *mjage* → Age when individual first used marijuana/hashish
4. *methamevr* → Whether or not an individual has ever used methamphetamine

Again, the goal for this investigation is to attempt to fit a model to predict whether or not an individual will use methamphetamine based on the age at which they first used cigarettes, alcohol, and marijuana.

For the initial processing of the data, the features of interest (described above) were extracted from the larger dataset, and this subset of the original data was filtered to only include responses from individuals that were complete and not skipped (individuals had the opportunity to skip questions when taking the survey, and these skips were denoted in the dataset and removed in this stage of the analysis). After cleaning the data, plots of the data were made to build intuition into the structure of the data and are shown in Figure 1 below.

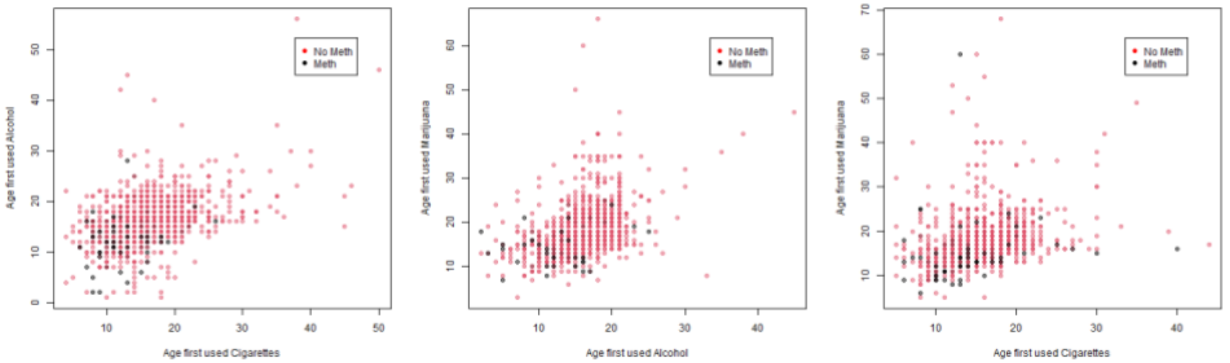


Figure 1: Distribution of respondents based on ages. *Left:* Age first used alcohol versus age first used cigarettes, colored by whether respondent used meth. *Center:* Age first used marijuana versus age first used alcohol, colored by whether respondent used meth. *Right:* Age first used marijuana versus age first used cigarettes, colored by whether respondent used meth.

When plotting the data, 10% of the actual data was randomly subsetted to plot due to the fact that there were over 18,000 observations for the processed data, which significantly cluttered the plotting space if 100% of the data was plotted. From this processed data, it can be seen that there is no clear general trend of which individuals used meth, but it can be seen that they tend to reside in the area of the parameter space that represents individuals using alcohol, cigarettes, and marijuana at younger ages.

After the data has been processed into a format that is conducive to machine learning, the predictive models will be implemented to assess the ability of the different models to predict meth use.

3 Methods

As briefly discussed in previous sections, in order to attempt to predict whether a respondent tried methamphetamine based on the age they tried other substances, this investigation will use KNN classification and SVM classification models. Prior to implementing the models, the data of 18,000 observations was split into training and testing datasets, where the training data consisted of 80% of the original data and the testing consisted of the remaining 20%. The observations selected for the training data were randomly selected. *All equations and methods described in this section were adapted from Introduction to Statistical Learning: with applications in R.*

3.1 K-Nearest Neighbor Classification

Given a positive integer, K, and a test observation, x_0, a KNN classifier identifies the K "nearest neighbors", or the K points with the smallest euclidean distance to the test point x_0 denoted by *N*, and estimates the conditional probability for the test observation being in class *j*. The conditional probability is represented by the following equation:

$$Pr(Y = j|X = x_0) = \frac{1}{K}\sum_{i \in N} I(y_i = j)$$

The KNN classifier then classifies the test observation to the class that has the highest conditional probability given by the equation above. For this investigation, a KNN classification model was fit to 18 values of K ranging from 1 to 18, and the testing misclassification error rate was recorded for each model. The testing misclassification error rate used in this investigation is described by the equation below:

$$Rate = \frac{1}{n}\sum_{i=0}^{n} I(y_i \neq \widehat{y_i})$$

Where n is the number of testing observations, y_i is the actual classification from the dataset, and $\widehat{y}_i$ is the predicted class from the KNN classifier. The results of this model will be discussed in the **Results** section of the investigation.

3.2 Support Vector Machine Classification

The SVM classification model creates a hyperplane with a soft margin that classifies the data points based on their location relative to the hyperplane in the parameter space. A p-dimensional hyperplane has the form:

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots\ + \beta_p X_p = 0$$

In order to build an SVM classification model, one can take the p-dimensional hyperplane and introducing a soft margin *M* that provides somewhat of a buffer for the classification hyperplane, while also introducing slack variables that allow individual observations to be on the opposite side of the hyperplane for overall robustness of the model. This creates an optimization problem to which the solution is the desired classification hyperplane. The optimization problem is outlined below.

$$\textit{Maximize M with respect to } \beta_0 \ldots \beta_p,\ \epsilon_1 \ldots \epsilon_n,\ M$$

$$\text{Subject to } \sum_{j=1}^{p} \beta_j^2 = 1,$$

$$y_i(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots\ + \beta_p x_{ip})\ \geq M(1 - \epsilon_i)$$

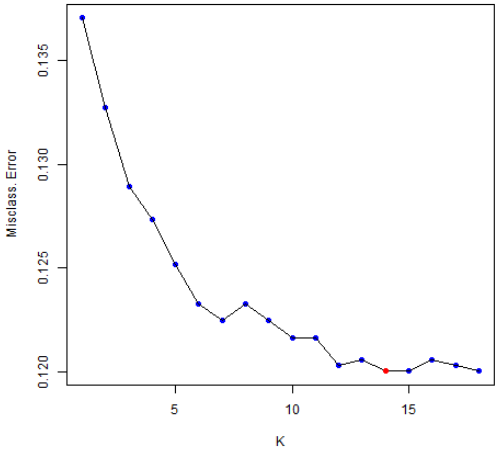
$$\epsilon_i \geq 0,\ \sum_{i=1}^{n} \epsilon_i \leq C$$

Where ϵ_i are the slack variables, and C is a nonnegative tuning parameter. For the purposes of this investigation, an SVM model was fitted with a linear kernel and varying values of the tuning parameter C, in addition to an SVM model with a radial kernel (non-linear decision boundary) and varying model tuning parameters.

The model tuning parameters for the SVM model for each kernel are selected by cross-validation on the training dataset, and the parameters that result in the lowest misclassification error, or the lowest number of support vectors, are selected. The results are discussed in the following **Results** section.

4 Results

4.1 KNN Classifier Results



As previously mentioned, for the KNN classifier, 18 values of K were used to fit a KNN model to the training data, and the model was applied to the testing dataset to observe the testing misclassification error for each value of K. A plot of the testing misclassification error as a function of K for this data is shown in Figure 2 to the right. The model found that the optimal value of K within this range was 14, and the corresponding test misclassification rate was 12.0%. This result means that 88% of the time, this model was able to correctly predict whether or not an individual used meth. This seems interesting at first, however after closer inspection of the data, if the model were to predict that every single person didn’t try meth, the test misclassification error rate would be 11.951%. This means that this model did worse than if it had predicted that each of the observations in the testing dataset were an individual who didn’t use meth. After consideration, this result isn’t satisfactory, so a more robust model that is more effective in higher dimensions is needed in order to have the opportunity to classify the data with a higher accuracy.

Figure 2: Plot of misclassification error rate and K for varying values of K.

4.2 SVM Classifier Results

4.2.1 Linear Kernel

For the SVM classifier with a linear kernel, the classifier was tuned with differing values of the cost variable (which is correlated with the nonnegative tuning parameter discussed in the **Methods** section for the SVM model). However, since the training data was very large and the cross validation process was time consuming for each kernel, only a few values of the tuning parameters were able to be explored in this investigation. Of these few values of tuning parameters, the misclassification errors were all relatively the same, so the model that resulted in the least number of support vectors was selected to be used for prediction on the testing dataset, and this model was with the cost parameter equal to 1. After predicting the classification of the testing dataset, the misclassification error was 11.95%, which sounds slightly better than the KNN result, however this is just because the model predicted that each individual would not use meth. From Figure 3 on the right, one can see that as an example, the entire parameter space of age first tried alcohol (x-axis) and age first tried cigarettes (y-axis) is classified as 2 by the SVM model (2 is the classification for no meth). Thus, it is clear why the error of the testing set is the same as predicting each individual as not doing meth, because that is exactly what the SVM classifier with a linear kernel is doing. This is the motivation behind trying a non-linear kernel which will provide a non-linear decision boundary, and may possibly be able to provide a more precise prediction.

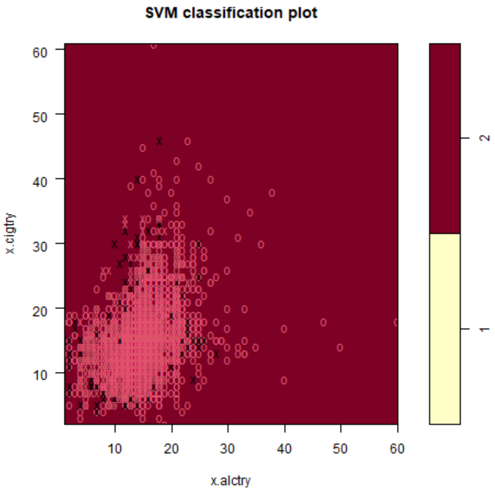


Figure 3: Visualization of SVM classification region in a projection of the parameter space, with a linear kernel. The value 2 corresponds to no meth and the value 1 corresponds to meth. The X’s correspond to meth in the training data.

4.2.2 Radial Kernel

The next step in this investigation was to fit a SVM classification model with a radial kernel, following the same steps as for the linear kernel with cross-validation on the training set to fit the model parameters *cost* and *gamma*. After running the model tuning, *cost* was set to 0.001 and *gamma* was set to 0.01. The number of support vectors for this model was less than that for the linear kernel, where the radial kernel needed 3644 support vectors and linear kernel needed 4089 support vectors. After using the model with the radial kernel to predict the testing data, the misclassification error rate was found to be 11.95%, which is the exact same as the linear kernel in the previous section. By looking at the plot of the parameter space for the linear kernel in Figure 4, it is clear that even with the radial kernel, the SVM classifier is simply just classifying each of the observations as no meth, which agrees with the values of the misclassification error rate as well. So again, this model was unable to provide a more accurate prediction of whether an individual is or is not going to use meth based on which age they first use alcohol, marijuana, and cigarettes.

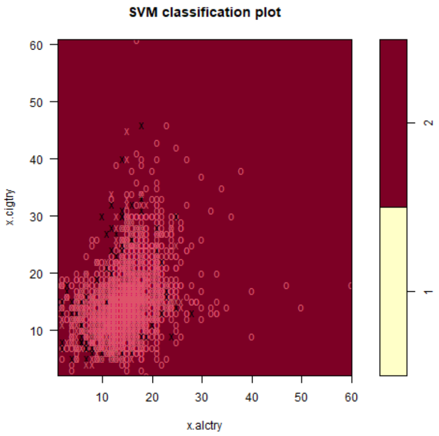


Figure 4: Visualization of SVM classification region in a projection of the parameter space, with a radial kernel. The value 2 corresponds to no meth and the value 1 corresponds to meth. The X's correspond to meth in the training data.

5 Discussion and Conclusion

Overall, as it was seen in the **Results** section, neither the KNN model, SVM model with a linear kernel, or the SVM model with a radial kernel were able to accurately predict whether or not an individual was going to use meth based on the age they first used alcohol, marijuana, and cigarettes. There are many reasons why this may be the case:

1. *There was little to no distinct separation between individuals who used and didn't use methamphetamine*. This means it was very difficult for the SVM models to separate the data into unique sections, and this is also the reason why there were so many support vectors for both of the models with linear and radial kernels. No matter what decision boundary was drawn, there would be a large number of classifications on either side of the decision boundary, resulting in the large number of support vectors.
2. *There were many more data points of individuals who didn't use meth than there were for individuals who used meth.* From the selected and filtered data that was used in the models, only 2268 out of 18537 individuals marked that they used meth, which means the majority of respondents would fall into the category of not using meth. This means that for models like the KNN classification model, for any given x_0, it's neighborhood is

almost always going to contain more points of individuals who never used meth, so the conditional probability will always favor individuals who do not use meth.

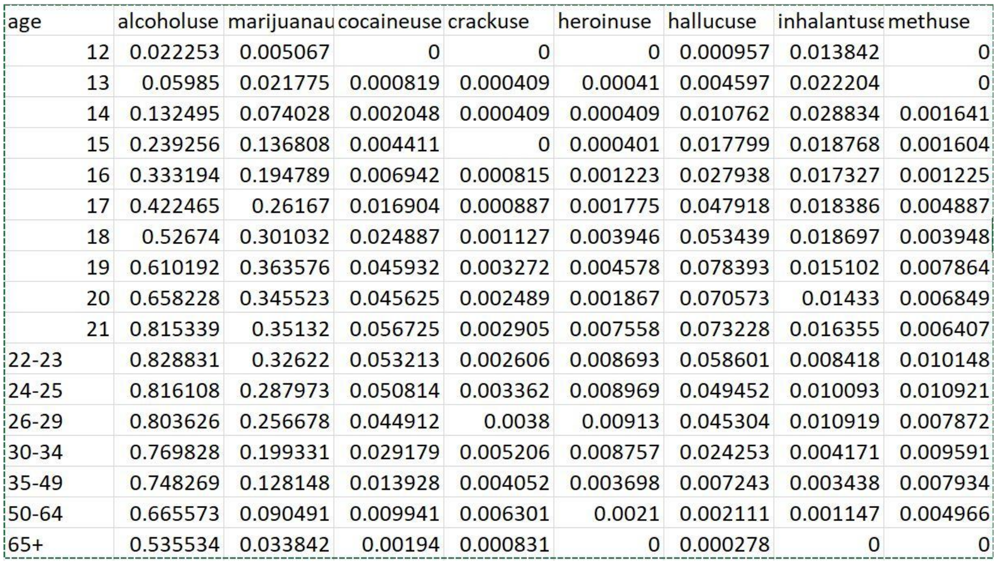
For these reasons, with the data given by the NSDUH survey, this investigation was unable to determine if there was a correlation between age that individuals first use alcohol, marijuana, and cigarettes and whether or not they will use methamphetamine.

5.1 Future Work

For future research into this dataset, it would be interesting to see what effects demographics have on the likelihood of individuals to use different substances, and at what frequencies. With a dataset this large, there are a lot of possibilities for future research into this dataset that would be interesting.

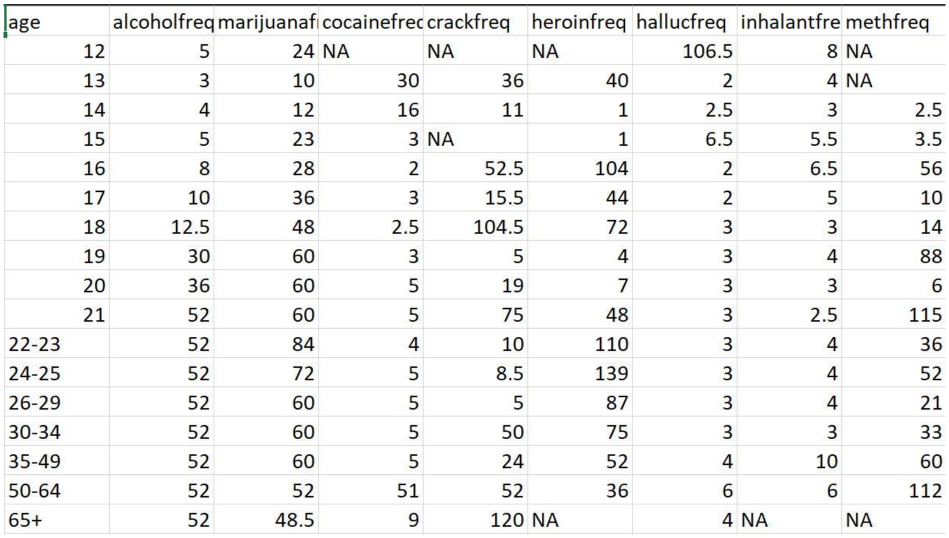
Appendix A

This appendix contains statistics mimicking the data found on fivethirtyeight's github for the drug-use-by-age dataset (Fivethirtyeight).



age	alcoholuse	marijuanau	cocaineuse	crackuse	heroinuse	hallucuse	inhalantuse	methuse
12	0.022253	0.005067	0	0	0	0.000957	0.013842	0
13	0.05985	0.021775	0.000819	0.000409	0.00041	0.004597	0.022204	0
14	0.132495	0.074028	0.002048	0.000409	0.000409	0.010762	0.028834	0.001641
15	0.239256	0.136808	0.004411	0	0.000401	0.017799	0.018768	0.001604
16	0.333194	0.194789	0.006942	0.000815	0.001223	0.027938	0.017327	0.001225
17	0.422465	0.26167	0.016904	0.000887	0.001775	0.047918	0.018386	0.004887
18	0.52674	0.301032	0.024887	0.001127	0.003946	0.053439	0.018697	0.003948
19	0.610192	0.363576	0.045932	0.003272	0.004578	0.078393	0.015102	0.007864
20	0.658228	0.345523	0.045625	0.002489	0.001867	0.070573	0.01433	0.006849
21	0.815339	0.35132	0.056725	0.002905	0.007558	0.073228	0.016355	0.006407
22-23	0.828831	0.32622	0.053213	0.002606	0.008693	0.058601	0.008418	0.010148
24-25	0.816108	0.287973	0.050814	0.003362	0.008969	0.049452	0.010093	0.010921
26-29	0.803626	0.256678	0.044912	0.0038	0.00913	0.045304	0.010919	0.007872
30-34	0.769828	0.199331	0.029179	0.005206	0.008757	0.024253	0.004171	0.009591
35-49	0.748269	0.128148	0.013928	0.004052	0.003698	0.007243	0.003438	0.007934
50-64	0.665573	0.090491	0.009941	0.006301	0.0021	0.002111	0.001147	0.004966
65+	0.535534	0.033842	0.00194	0.000831	0	0.000278	0	0

Figure 5: This table contains the fraction of respondents who used a given substance for each age group. The substances included in this table are alcohol, marijuana, cocaine, crack, heroin, hallucinogens, inhalants, and methamphetamine respectively. The age groups included are shown in the "age" column.



age	alcoholfreq	marijuanaf	cocainefrec	crackfreq	heroinfreq	hallucfreq	inhalantfre	methfreq
12	5	24	NA	NA	NA	106.5	8	NA
13	3	10	30	36	40	2	4	NA
14	4	12	16	11	1	2.5	3	2.5
15	5	23	3	NA	1	6.5	5.5	3.5
16	8	28	2	52.5	104	2	6.5	56
17	10	36	3	15.5	44	2	5	10
18	12.5	48	2.5	104.5	72	3	3	14
19	30	60	3	5	4	3	4	88
20	36	60	5	19	7	3	3	6
21	52	60	5	75	48	3	2.5	115
22-23	52	84	4	10	110	3	4	36
24-25	52	72	5	8.5	139	3	4	52
26-29	52	60	5	5	87	3	4	21
30-34	52	60	5	50	75	3	3	33
35-49	52	60	5	24	52	4	10	60
50-64	52	52	51	52	36	6	6	112
65+	52	48.5	9	120	NA	4	NA	NA

Figure 6: This table contains the median frequency at which the respondents use a given substance (in days) in the last 12 months. The substances included in this table are the same as for Figure 5.

Bibliography

Barry-Jester, A. M., & Flowers, A. (2015, April 23). How baby boomers get high. Retrieved April 18, 2021, from https://fivethirtyeight.com/features/how-baby-boomers-get-high/

Fivethirtyeight. (n.d.). Fivethirtyeight/data. Retrieved April 18, 2021, from https://github.com/fivethirtyeight/data/tree/master/drug-use-by-age

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An introduction to statistical learning: With applications in R*. Boston: Springer.

National Institute on Drug Abuse. (2021, February 25). Overdose death rates. Retrieved April 18, 2021, from https://www.drugabuse.gov/drug-topics/trends-statistics/overdose-death-rates

National Survey on Drug Use and Health 2016 (NSDUH-2016-DS0001). (n.d.). Retrieved April 18, 2021, from https://www.datafiles.samhsa.gov/study-dataset/national-survey-drug-use-and-health-2016-nsduh-2016-ds0001-nid17185