Executive Summary

Over the past few years, we have seen a frightening increase in the use of controlled substances, including tobacco, alcohol, and illicit drugs. These drugs have alarming impacts on individuals and can have ripple effects on these individuals' communities and American society as a whole.

We first considered the rise of e-cigarette use, colloquially known as "vaping." Even though cigarette usage among teenagers over the past 8 years has been steadily decreasing, the use of vapes and juuls has been increasing at a striking rate. We used past data and least-squares regression to learn parameters in a novel coupled difference equation describing the proportion of the population that is currently smoking and the proportion of the population that is currently vaping, accounting for the rate of quitting and the rate of conversion between the smoking, vaping, and general populations. We then iteratively applied this difference equation with the parameters we learned on current data to give us projections for cigarette and e-cigarette use in the future. This model predicts that, in the year 2029, 12.4% of the population of the United States will vape and 8.5% will smoke. Applying our recurrence a larger number of times, we reach a steady-state solution of 20.5% of the population will vape and 0% will use ordinary cigarettes.

We then considered the drug use among different demographics. We used a publicly available dataset giving hospital admissions for different drugs to train a decision tree. Our novel decision tree approach takes a list of important attributes representing an individual such as race, socioeconomic status, and other factors and outputs the likelihood that they will abuse a given drug. We then randomly generated a class of 300 high school seniors that were representative of the general U.S. population to predict how many high schoolers we can expect to abuse each drug from an average high school class. Our decision tree analysis predicts that 79 of these high school seniors would be addicted to nicotine, marijuana, alcohol, or unprescribed opioids, with 32 of these seniors being addicted to nicotine, 15 addicted to marijuana, 22 addicted to alcohol, and 10 addicted to opioids.

Finally, we examined the ripple effects of nicotine, marijuana, alcohol, and opioid/opiate abuse. We explored a combination of factors representative of the social, health, and economic impacts of a given drug, including deaths, healthcare costs, lost productivity, criminal justice, and domestic violence. We created a metric using a linear combination of these factors and converted this metric to units of U.S. dollars per capita per year. We found that nicotine costs $15139/capita, opiates cost $3742/capita, alcohol cost $3488/capita, and marijuana costs $114/capita. Thus, we see that nicotine is the drug with the most negative impact on our society and economy, with opiates, alcohol, and marijuana in decreasing intensity of impact.

Global Assumptions

G.1 Individuals only use one drug at a given time.

Global Definitions

Substance abuse: the repeated use of an illegal or harmful drug.

Habitual/frequent use: the use of a drug at least once per month.
Vaping: the use of any e-cigarette or e-hookah in order to consume nicotine.
Smoking: the use of traditional cigarettes or cigars in order to consume nicotine; does not include vaping or marijuana consumption.
Alcohol: refers to the consumption of beverages containing ethanol, including beer, wine, and spirits.
Marijuana: refers to both the consumption of "edibles" and the smoking of THC-containing products.
Opioids/opiates abuse: the use of illegally acquired opioid drugs and unprescribed prescription medicines.

Part I: Darth Vapor

1.1 Restatement of Problem

We are asked to project the percentage of Americans who will use vape and cigarette products, respectively, over the next ten years. (Implicitly, we assume we will use historical data of vape and cigarette usage and, based on that data, extrapolate accordingly.)

1.2 Local Assumptions

1. Children under 13 who use vape or cigarette products constitute a negligible proportion of the population.
    a. **Justification**: Children under 13 are generally dependent on their parents and do not experience many of the push factors that lead to vaping and cigarette use. Furthermore, other drugs are shown to be used by vastly more people over 13 than under 13, and it is reasonable to extrapolate this to vaping and cigarettes. [1][13]
2. Vape statistics among adults in the United Kingdom mirror those in the United States.
    a. **Justification**: The statistics we found for the UK are consistent with the values for the U.S. when such values are available, as they are for most years. The UK and U.S. are both members of the Anglosphere and have similar economic statuses and cultures, so we expect vaping to spread in both countries similarly. [2][12]
3. The percentage of 9th graders who use vape or cigarette products can be approximated as the average percentage of 8th and 10th graders who do so. Similarly, the percentage of 11th graders who use vape or cigarette products can be approximated as the average percentage of 10th and 12th graders who do so.
    a. **Justification**: We do not expect drug use to fluctuate wildly between years of high school. Thus, a simple linear interpolation should be sufficient.
4. People who smoke cigarettes do not also vape, and vice versa. (See Assumption G.1.)
    a. **Justification**: We considered that most individuals who smoke or vape have a strong preference for their chosen form of nicotine intake. From here, considering the small size of the population who both smoke and vape, we could reasonably

deduce that the number of people who neither smoke nor vape is merely the total number of people, minus the number of people who smoke, minus the number of people who vape.

5. The number of people who pick up smoking/vaping is dependent on the degree of contact between smoking/vaping population and the general population.
    a. **Justification**: Environmental conditions strongly influence drug use. Teens and young adults are strongly influenced by the prevalence of smoking and vaping around them as they choose to start smoking or vaping. [14]
6. The numbers of people in 9th, 10th, 11th, and 12th grades are identical.
    a. **Justification**: The number of students dropping out of high school for various reasons, or dying, is small. In addition, a quick glance at the results of the 2010 U.S. Census reveals that the number of minors under age 5 is roughly $\frac{5}{18}$ the number under age 18, which implies that the U.S. birthrate has stayed relatively constant in the 21st century. [9]

1.3 Variables

Symbol	Definition	Units
t	$y - 2018$, where y is a particular year	Years
$\underline{S_t}$	The proportion of U.S. population who smoke cigarettes in year t	%
V_t	The proportion of U.S. population who vape in year t	%
α	The conversion rate between general population and smokers	%
β	The conversion rate between general population and vapers	%
γ	The conversion rate between smokers and vapers	%
μ	The proportion of smokers who quit	%
ξ	The proportion of vapers who quit	%

1.4 Solution & Results

In order to compute the spread of vaping and cigarette use over the next ten years, we gathered data on historical levels of vaping and smoking in recent years, splitting the populations of cigarette and e-cigarette users into those who were over the age of 18, and those who were high school students. Employing our first local assumption, we ignored data from middle school students, as their usage rates of vape and nicotine are relatively negligible in comparison to those of older teens and adults.

Smoking data were gathered from a survey conducted by the Centers for Disease Control, National Youth Tobacco Survey [1].

Vaping data were gathered from the given data from the National Youth Tobacco Survey, listing the percentage of high school students using e-cigarettes during the past 30 days each year from 2011 to 2015, and the National Institutes for Health Drug Trends Data Sheet, listing the percentages of 8th, 10th, and 12th graders vaping during the past 30 days each year from 2015 to 2018. While the first source gave the overall percentages of high schoolers who vape, the second source did not. Therefore, we use assumptions 2 and 6: If v_i represents the proportion of students in grade i who vape, then the overall proportion of high schoolers who vape is given by

$$\begin{aligned}
&\tfrac{1}{4}v_9 + \tfrac{1}{4}v_{10} + \tfrac{1}{4}v_{11} + \tfrac{1}{4}v_{12} \\
&= \tfrac{1}{4}\left(\tfrac{1}{2}v_8 + \tfrac{1}{2}v_{10}\right) + \tfrac{1}{4}v_{10} + \tfrac{1}{4}\left(\tfrac{1}{2}v_{10} + \tfrac{1}{2}v_{12}\right) + \tfrac{1}{4}v_{12} \\
&= \tfrac{1}{8}v_8 + \left(\tfrac{1}{8} + \tfrac{1}{4} + \tfrac{1}{8}\right)v_{10} + \left(\tfrac{1}{8} + \tfrac{1}{4}\right)v_{12} \\
&= \tfrac{1}{8}v_8 + \tfrac{1}{2}v_{10} + \tfrac{3}{8}v_{12}.
\end{aligned}$$

Corroborating the vaping data for 2015, we obtain vaping data for years 2011-2018, and combining the smoking and vaping data, we obtain the following table:

Year	High School Smoking Percentage (%) [1][2]	High School Vaping Percentage (%) [3]
2011	5.9	1.5
2012	5.2	2.8
2013	4.6	4.5
2014	3.4	13.4
2015	3.4	15.1
2016	3.2	11
2017	3.1	13.6
2018	2.5	22.2

Table 1. Percentages of US high schoolers who reported smoking/vaping in the past 30 days from 2011 to 2018.

For the adult population, we extracted statistics from source [10] to find what proportion of adults in the United States smoke. Furthermore, due to a lack of a published comprehensive survey of trends in vaping over the years, we were forced to find a variety of sources giving vaping statistics among adults. Thus, in the table below, we have marked each datapoint with the source.

Data marked with asterisks (*) are taken from data collected in the United Kingdom (refer to assumption 2).

Year	Adult Smoking Percentage (%) [10]	Adult Vaping Percentage (%)
2011	20.6	1.1* [12]
2012	19.3	2.0* [12]
2013	19	2.2 [5]
2014	18.1	3.7 [6]
2015	17.8	4.1* [12]
2016	16.8	4.5 [11]
2017	15.1	2.8 [7]
2018	15.5	3.3 [8]

Table 2. Percentages of US high schoolers who reported smoking/vaping in the past 30 days from 2011 to 2018.

Combining the data for adults and minors by weighting with U.S. Census information [9], we get the following proportions of the total population who smoke and vape for the past 8 years:

Year	Total Smoking Percentage (%)	Total Vaping Percentage (%)
2011	19.9	1.1
2012	18.6	2.0
2013	18.3	2.3
2014	17.4	4.2
2015	17.1	4.6
2016	16.1	4.8
2017	14.5	3.3
2018	14.9	4.2

Table 3. Percentages of the total US population who reported smoking/vaping in the past 30 days from 2011 to 2018.

From this table of data, we can build a difference equation describing, given current data, the projected number of smokers and vapers in the future. We design the following recurrences:

$S_{t+1} = (1-\mu)S_t + \alpha S_t(1 - S_t - V_t) - \gamma S_t V_t$, and
$V_{t+1} = (1-\xi)V_t + \beta V_t(1 - S_t - V_t) + \gamma S_t V_t$.

The first term on the right side of both equations represents what proportion of smokers/vapers will not stop vaping next year. Also, following assumption 8, we include mixing terms representing the contact between the smoking, vaping, and general populations.

To learn the parameters, we use the past data and perform least-squares regression. We construct the following matrix,

$$A = \begin{bmatrix} S_0 & 0 & S_0(1-S_0-V_0) & -S_0V_0 \\ S_1 & 0 & S_1(1-S_1-V_1) & -S_1V_1 \\ \vdots & & & \\ 0 & V_0 & V_0(1-S_0-V_0) & S_0V_0 \\ 0 & V_1 & V_1(1-S_1-V_1) & S_1V_1 \\ \vdots & & & \end{bmatrix},$$

and the following vector,

$$b^T = (S_1, S_2, \ldots, V_1, V_2, \ldots).$$

Then, to generate the parameters, we apply the following operation:

$$\mathbf{x} = (\mathbf{A^T A})^{-1}\mathbf{A^T b}.$$

Multiplying by the transpose projects $\mathbf{b}$ onto the column space of A, minimizing the L2 norm, allowing us to solve this overdetermined system. Then, taking the data from 2018 and applying our recurrence 10 times, we obtain projected information for the next ten years:

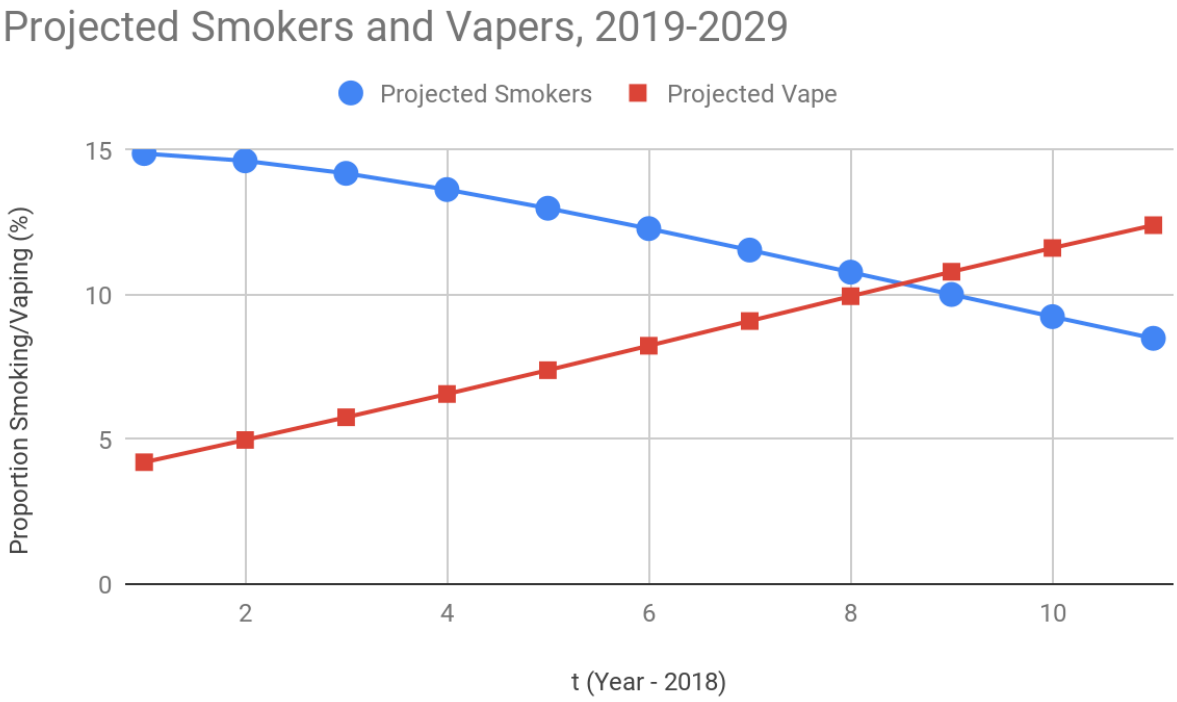

Figure 1. Proportion of the U.S. population for the next ten years projected to have smoked/vaped for the last 30 days.

We project that in the year **2029, 12.4%** of the population of the United States will vape and **8.5%** will smoke.

1.5 Validation

To check if our learned parameters were reasonable, we repeated applied our recurrence, reaching the steady state solution of 20.5% of the population vaping and 0% smoking. We also calculated the one-year projection based on each year in the past and compared it to the actual values for the next year, finding that the maximum deviation was always less than 0.86%:

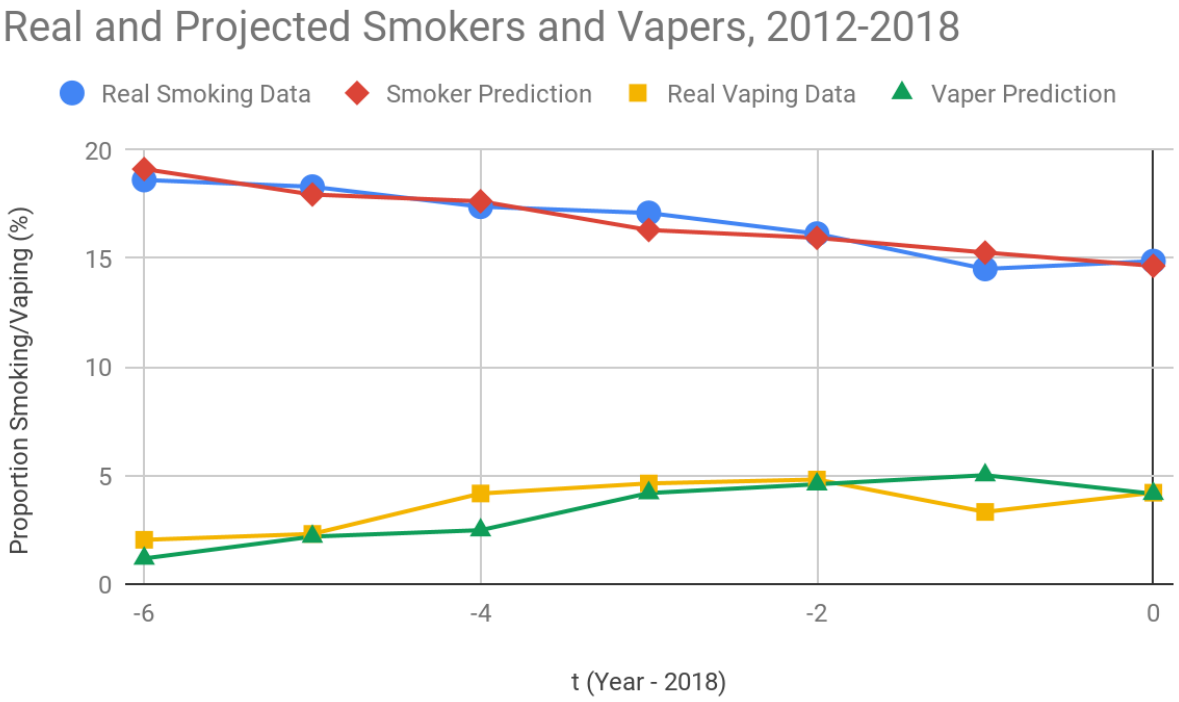

Figure 2. Proportion of the U.S. population in the past seven years projected to have smoked/vaped for the last 30 days, compared to actual results.

Thus, we see that our model yields accurate results for short periods of time. However, we do not have enough data to see our effective our model would be at predicting long-term trends, especially considering how recently vaping rose to popularity.

1.6 Strengths and Weaknesses

Our model's strengths include its simplicity and generalizability: we only need information for one year to project for perpetuity. Also, we do not have any reason to believe our parameters vary with time; we expect effects like peer pressure to depend only on the density of individuals who partake rather than the year. Furthermore, our difference equation accounts for the most important effect contributing to the rise of vaping: the contact between vapers and nonvapers, with similar terms for smoking and smoking-to-vaping conversion.

Our model's greatest weakness is the small number of data points compared to the number of parameters learned. Our matrix has 14 rows with 5 columns, so we may have overfit the parameters.

Part II: Above or Under the Influence

2.1 Restatement of Problem

The plethora of substances that people are exposed to today come with numerous factors that affect their daily lives and just as importantly, their chances of becoming addicted. Given this issue, we were tasked to develop a model that can effectively determine the likelihood that any given individual, and his/her unique background, would use a given substance.

2.2 Local Assumptions

1. Age (AGE): At any given time in the school year, there are more 17 year olds than 18 year olds in the senior class. Additionally, most seniors behave like students in the range of 15-17 than in the age group 18-20.
    a. **Justification**: The birthdays of most seniors fall within the middle of the year, so on average, they should behave like the 15-17 age group.
2. Arrests (ARRESTS): The number of times one has been arrested in the previous 30 days significantly affects his/her likelihood of using drugs.
    a. **Justification**: Delinquents are more likely to engage in other illegal activity.
3. Education (EDUC): High school seniors have had 16 or more years of education.
    a. **Justification**: Compulsory education laws in the majority of U.S. States require students through age 17/18 to receive education. Therefore, we find the proportion of students age 17/18 who are employed full-time to be negligible. We also acknowledge the fact that 25% of all high school students are part-time workers outside of school, but the ratio of these throughout 4 years remains approximately constant. [37][39]
4. Employment (EMPLOY): High school seniors are not employed full-time; they either work part-time or are not in the workforce.
    a. **Justification**: High schoolers are assumed to not have dropped out of high school, so they can not be full-time employees.
5. Ethnic (ETHNIC): The ethnic distribution of U.S. minors is reflective of that of high school seniors.
6. Gender (GENDER): The gender distribution of the general U.S. population is reflective of that of high school seniors.
7. Living Arrangement (LIVARAG): The living conditions of a high school student significantly affect the probability that he/she will use drugs.
    a. **Justification**: The people whom a student lives with shape his/her social interactions and have a great impact on the student's likelihood of using drugs.
8. Marital Status (MARSTAT): High schoolers are not married.
    a. **Justification**: The majority of states do not allow high school marriage. [40]
9. Primary source of income (PRIMINC): The primary source of income of a high school student significantly affects the probability they will use drugs.
    a. **Justification**: A person dependent on public assistance is likely to be of a lower socioeconomic status than one who is not, which significantly alters the chance that they will use drugs.

10. Psychiatric Problem (PSYPROB): The presence of a psychiatric problem significantly affects one's chance of using drugs.
    a. **Justification**: Depression and other mental issues are factors that greatly contribute to depression.
11. Race (RACE): The racial distribution of U.S. minors is reflective of that of high school seniors.
12. All of the above factors are distributed independently of one another.
    a. **Justification**: Although some of these traits are known to be linked to one another genetically, it is common practice to treat them as effectively independent when from such a statistical context.

2.3 Variables

Symbol	**Definition (Possible values)**
Arr	Arrests - Number of arrests in 30 days prior to admission ($\in \mathbb{Z}$)
Emp	Employment - Employment status (full, partial, unemployed, not in labor force)
Eth	Ethnic - Hispanic or Latino Origin (Puerto Rican, Mexican)
Gen	Gender - Sex (male, female, unknown)
Liv	Living arrangements (homeless, dependent, independent, unknown)
Pri	Primary source of income (wages, public assistance, retirement, other, none)
Psy	Psychological condition (yes, no, unknown)
Rac	Race (Alaskan, Amerindian, Asian, Black, White, Other)
Age	Age (12-14, 15-17, 18-20, etc.)
Edu	Highest education (8 years or less, 9-11, 12, 13-15, 16 or more)
Mar	Marital status (never married, now married, separated, divorced)

2.4 Solution & Results

Part 1 - Decision Tree Learning of Which Drug a Given Individual is Likely to Use

This problem consisted of various components, the main part of which was determining a potential likelihood for a given individual to abuse a substance. To accomplish this we used a decision tree classifier which took in various parameters regarding an individual's daily life to predict use of a substance, if at all. Firstly, we found publicly available data as part of the TEDS 2016 dataset [41] which consisted of over 1,000,000 cases of substance abuse. Part of selecting this was its extensive information collection on each individual ranging from vital demographic information like age and gender to others like marital status and pregnancy. To determine which features had more predictive power of a specific drug use we first calculated the entropy using the formula $S = -\sum_i p_i \log_2 p_i$. This allowed us to determine which subset of features, outlined in the table above, were more predictive and extendable on a larger scale. We parsed associated data from TEDS to only include the features which we were interested in and developed the decision tree using the sci-kit learn machine learning library in Python.

Part 2 - Data Point Generation and Testing:

Part of our process to test our model was generating the 300 random high school seniors, and we generated this in such a way so that our randomly created students reflect the teenage demographics of the U.S. Appealing to the last local assumption, the independence of these parameters imply that we can generate these parameters independently of each other based on the statistics in the U.S. To achieve this, we first researched these statistics for U.S. teenagers and computed the percentage of teenage people that fall into each of these categories. Using Python code, for all three hundred desired data points for testing, we randomly generated a real number between 0 and 100, and matched it to a range based on the number of teenagers that fell into each category, each of which corresponded to a given value of the variables for each of the students. This gave us a randomly sampled set of students that could be regenerated.

```
[[93449 11136  5835    34]
 [18542 24554  2546    23]
 [17571  3146  6245    10]
 [  211    64    47     3]]
              precision    recall  f1-score   support

           2       0.72      0.85      0.78    110454
           4       0.63      0.54      0.58     45665
           7       0.43      0.23      0.30     26972
          12       0.04      0.01      0.02       325

   micro avg       0.68      0.68      0.68    183416
   macro avg       0.45      0.41      0.42    183416
weighted avg       0.65      0.68      0.66    183416

Accuracy: 0.6774272691586339
```

270	2	0	5	2	...	1	1	2	8
271	2	0	5	4	...	1	1	2	5
272	2	0	5	4	...	1	1	2	5
273	2	0	5	4	...	1	3	1	6
274	2	0	5	2	...	1	1	2	5
275	2	0	5	2	...	1	1	2	5
276	2	1	5	4	...	1	1	2	5
277	2	0	5	4	...	1	1	2	8
278	2	0	5	4	...	1	3	2	4
279	2	0	5	4	...	1	1	2	4
280	2	0	5	2	...	1	1	2	8
281	2	2	5	2	...	1	3	2	5
282	2	0	5	4	...	1	1	2	5
283	2	0	5	2	...	1	1	2	5
284	2	0	5	2	...	1	3	2	4
285	2	0	5	4	...	1	5	2	5
286	2	0	5	4	...	1	1	1	5
287	2	0	5	4	...	1	1	2	5
288	2	0	5	4	...	1	1	1	5
289	2	0	5	4	...	1	1	2	5
290	2	0	5	2	...	1	1	2	6
291	2	1	5	4	...	1	1	2	5
292	2	0	5	2	...	1	1	2	5
293	2	0	5	4	...	1	1	1	5
294	2	0	5	4	...	1	1	1	4
295	2	0	5	4	...	1	3	2	5
296	2	0	5	4	...	1	1	2	6
297	2	0	5	4	...	1	1	2	5
298	2	0	5	2	...	1	3	1	5
299	2	0	5	4	...	1	1	2	5

```
[300 rows x 11 columns]
Number of total drug high school senior users: 79
Number of nicotine high school senior users: 32
Number of marijuana high school senior users: 15
Number of alcohol high school senior users: 22
Number of unprescribed high school senior opioid users: 10
```

Figure 3-4. Analysis of the precision of the decision tree on its own test data and the results of the decision tree analysis on the 300 randomly selected seniors

Simulation Results

Total number of drug users	79
Nicotine Users	32
Marijuana Users	15
Alcohol Users	22
Opioid Users	10

2.5 Validation

The primary strength of using a decision tree lies in it learning conditional probabilities better than a human observer can. For example, if two features are strongly correlated, then we need not consider both; however, construction of our tree calculates conditional entropy at each step so we never unnecessarily reuse information.

After implementing the decision tree and generating a representative data set of 300 high school U.S. seniors, we trained the decision tree model using a GPU we had access to through our school network. We used a GPU to accelerate the training time, as running the machine learning model on a mere CPU would be an inefficient process. In this manner, we we able to train and validate our decision tree model within 10-15 minutes. The final accuracy rate that our model achieved was 67.7% (see above). This accuracy rate can be considered a quite successful modeling of which drugs a certain individual is likely to use, as merely random guessing would yield a 25% accuracy rate for differentiating between the four categorical classes, marijuana, nicotine, and unprescribed opioids. As a result, our model successfully modeled which drug a certain individual is likely to use, as we further validated in our case study of 300 high school seniors.

We may have been able to improve our accuracy right slightly if we had more than 14 hours to work on this project, which we would use to incorporate data from the past twenty years from TEDS-D, for more than millions of individual records. This addition of data could improve the accuracy rate and still be trained in a fairly reasonable amount of times due to our utilization of a GPU computing resource.

In addition, it is important to note that the TEDS-D data set primarily included data for individuals who were known to have abused the drug. We justified our usage of this data in the case study for the drug usage, as opposed to abuse, among the 300 high school seniors (representative of the U.S. population to some extent), by the fact that certain demographics that are likely to abuse a drug are also representative of demographics that are likely to frequently use

the drug. Hence, it was justifiable for us to apply this data as although there will always be some outliers regarding demographics and usage of a particular drug in the real world, in a more theoretical study it is reasonable to assume that a group of people who are statistically more inclined to abuse a certain drug are also more inclined to use it at least once (as in our case study).

2.6 Strengths and Weaknesses

Strengths

- Efficiency
    - Our model is extremely inexpensive and took only about 10-15 minutes to train on a set of almost 1.7 million different individuals' data
    - The model requires no external costs and utilizes already publicly available data on computing power that is very accessible
- Scalability
    - Easily generalizable to a population since the model was trained on such an expansive data source

Weaknesses

- Limited
    - Given the time constraint of 14 hours, we were only able to train on 1 year of TEDS data. With the ability to train exhaustively on all data, we could further enhance an already appreciable accuracy

Part III: Ripples

3.1 Restatement of Problem

Substance abuse has a ripple effect that affects our society and our community. Our goal is to quantify those effects for a given drug and rank the harmfulness of nicotine, marijuana, alcohol, and opiates.

3.2 Local Assumptions

1. We assume linearity for each of the different factors associated with substance abuse. For example, should the rate of alcohol abuse doubled, the rate of death due to drunk driving would double.
    a. **Justification**: We do not expect any of our parameters to drastically change in the near future, so linearity is a good approximation. Also, there is no reason to account for the change of these
2. We assume that when we compute workplace costs for a drug, the abuse of that drug is uniform throughout the industries in which it is present.
    a. **Justification**: We do not expect that a drug will be abused at significantly higher rates in any particular industry, even when considering opioid abuse as a byproduct of workplace injury.

3.3 Variables

Symbol	Definition
DD_i	The number of deaths per year directly caused by drug i (e.g. overdose)
AD_i	The number of deaths per year indirectly caused by drug i (e.g. car accidents)
β	The cost of a human life, in U.S. dollars
HS_i	The total healthcare costs associated with drug i, in dollars
WP_i	The total cost of lost productivity associated with drug i, in dollars
DV_i	The rate of intimate-partner abuse by a user of drug i
CJ_i	The cost of criminal justice associated with drug i
P	Population size
α_i	Our defined metric, describing harmfulness of drug i

3.4 Solution & Results

We give the following metric describing the per-capita negative impact of a drug: $\alpha_i = (\beta(DD_i + AD_i) + HS_i + WP_i + DV_i + CJ_i)/P$.

Drug	DD_i/P	AD_i	HS_i/P	WP_i/P	DV_i/P	CJ_i/P	α_i
Nicotine	13351 [16]	1140 [19]	513 [21]	115 [24]	19 [28]	0 [32]	15139
Marijuana	0 [17]	0 [17]	0 [22]	0 [25]	0 [29]	110 [33]	114
Alcohol	2448 [18]	277 [20]	84 [23]	548 [26]	55 [30]	76 [34]	3488
Opiates	1324 [36]	Not found	79 [22]	2243 [27]	75 [31]	20 [35]	3742

Table 4. Negative impacts of four drugs with respect to various factors (in dollars per capita per year).

When we consider the value of β, or the value of a human life, we appeal to the value given by the U.S. government that considers the amount of money the government should spend for the loss of a life. This value has increased over time as unions lobby the government for higher

standards and regulations to hold businesses to—as a result, government agencies such as the Environmental Protection Agency (EPA), the Food and Drug Administration (FDA), and the Department of Transportation (DoT) now value the human life on the order of several million dollars [38]. In particular, the EPA value of $9.1 million per human life is the commonly accepted standard for the

We chose these standards by considering the possible ways that drug use could impact American society both socially and economically. We came up with a variety of possible major effects of drugs on the economy - the factors that we have considered are the potential costs of taking care of those who have overdosed or require serious medical attention, the effective cost of the labor lost from workers who are not able to contribute to industry, and the loss of human capital incurred from death by overuse of drugs. In these areas, drugs clearly represent a waste of resources that could be better spent being more productive in the workplace, especially for the costs on the healthcare system that could be used to treat diseases that are more difficult to treat, such as cancer. An implicit cost that can also be considered a social cost is the money that the government uses through the criminal justice system pursuing those using illicit drugs and processing offenders, which contributes a large amount to the overall impact of these drugs as three out of these four drugs are illegal in the United States with respect to possession or use in certain circumstances (for example, while operating a motor vehicle). Finally, we consider the social costs of drug use, one of which is the problem of domestic abuse while under the influence of these drugs. A productive society should tend to minimize the amount of intramarital conflict in it, as this decreases the overall utility in society - although domestic abuse is inevitable, the amount due to drug use can still be minimized.

Several of these raw economic values are on the order of billions of dollars, specifically, the values relating to the costs on the workplace and criminal justice system that are more on the scale of countries, so in order to make these values comparable to each other, we will use a per capita measure in order to normalize these values.

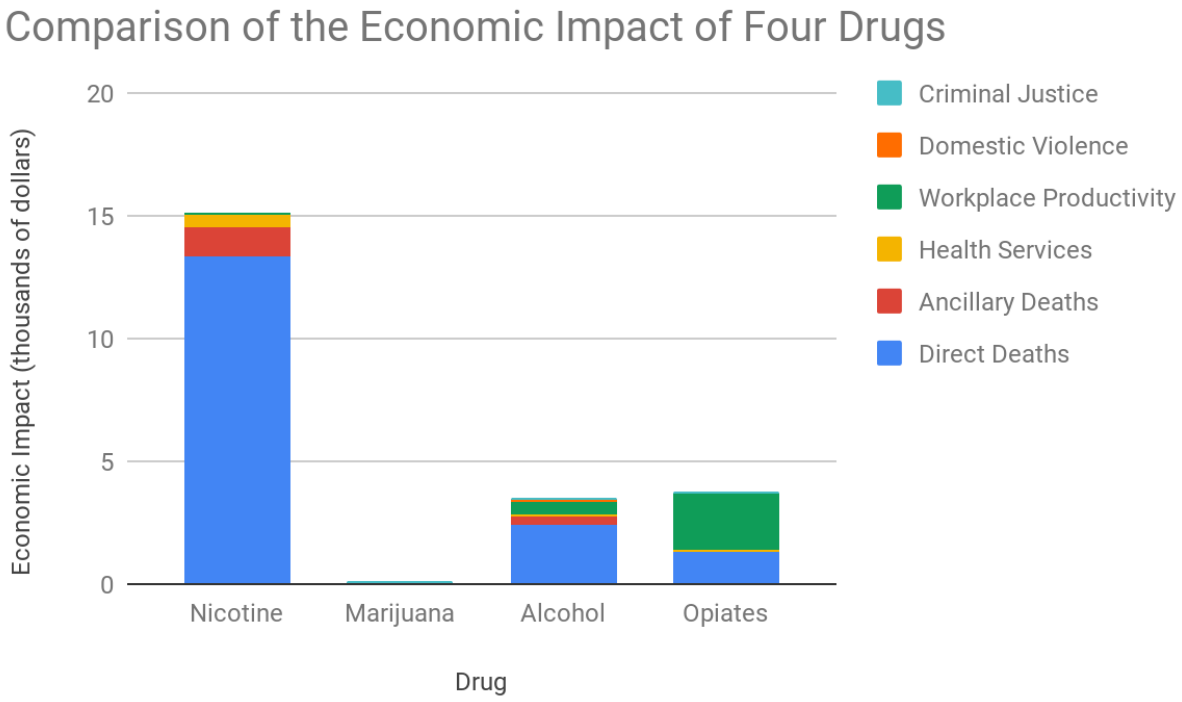

Figure 5. A comparison of the economic impact of four drugs in thousands of dollars per capita per year.

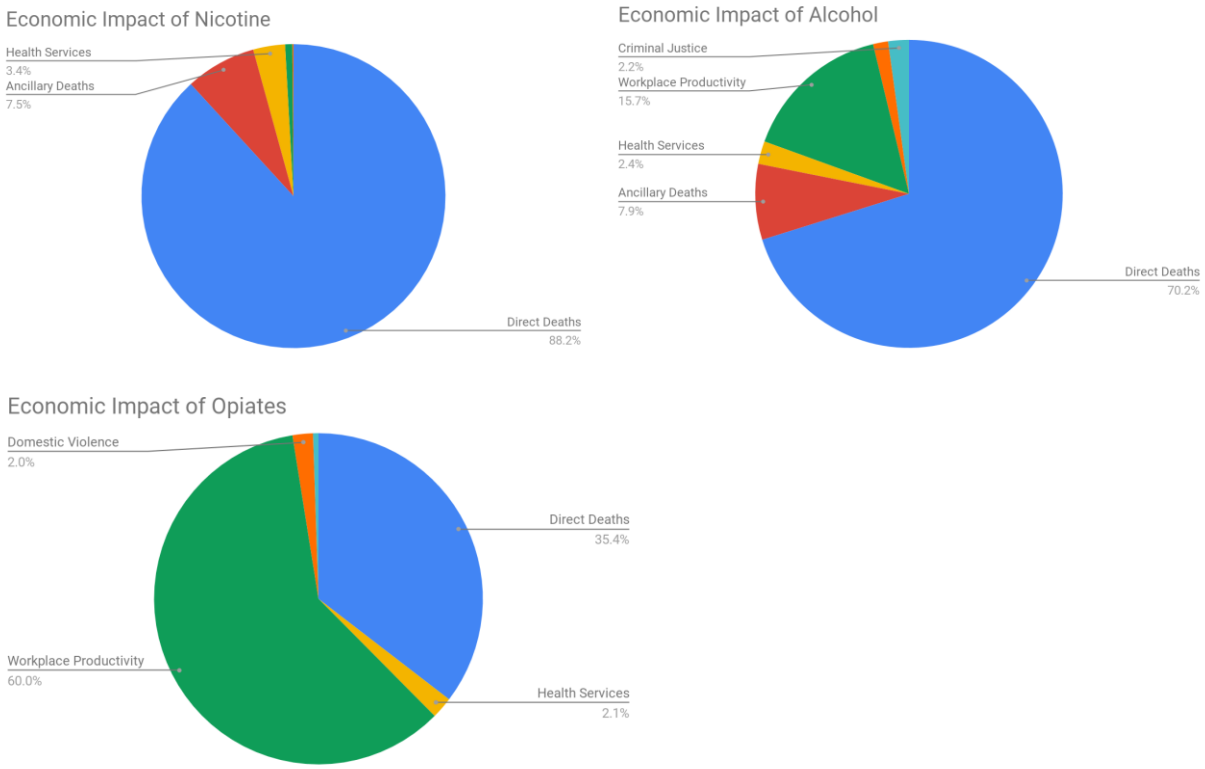

Figure 6. Proportion of the economic impact of three drugs induced by various factors. (Marijuana is not included because many of its economic impacts are positive.)

3.5 Validation

The National Drug Threat Assessment by the U.S. Department of Justice found that the use of illicit drugs caused a $193 billion drain on the U.S. economy [43]. This corresponds to $584/capita, which is lower than our metric for the use of opiates alone. However, we considered a large number of factors in additional to economic burden, so this data is consistent with our results.

3.6 Strengths and Weaknesses

Our model's strengths is flexibility, robustness, and extensibility. Our model does make reference to the population or community under investigation, so we may substitute whatever data refers to the community of interest.

The weakness of our model is the number of factors included: substance abuse does impact individuals and communities in different ways not considered, such as increasing divorce rate or by reinforcing the cycle of poverty. However, these factors are also strongly correlated with factors that are not substance abuse, making it very difficult to construct an extensible metric using these features.

Also, our data concerning the effects of marijuana seem to understate long-term harms of marijuana, causing it to appear much less harmful than it is. This is a claim that is currently disputed in the literature.

Conclusion

We used a variety of mathematical and computational techniques to model substance use in America. Our regression and difference equation solver shows projects the proportion of individuals who will vape and smoke in the future and yields us a steady-state solution of what forms of nicotine intake Americans will use in the future. We then trained a decision tree to classify individuals and predict whether they will abuse substances and what drug they will use if they will. Finally, we used statistical analysis to summarize the negative effects of each drug and assess their relative harms.

The effects of our successful research and mathematical modeling have vast implications in our society that has an increasing drug epidemic. This will improve our ability to keep our societies healthy, safe, and drug-free.

Bibliography

[1]https://www.cdc.gov/tobacco/data_statistics/surveys/nyts/data/index.html
[2]https://www.cdc.gov/tobacco/data_statistics/fact_sheets/youth_data/tobacco_use/index.htm#anchor_1549569386405
[3]https://www.cdc.gov/tobacco/data_statistics/sgr/e-cigarettes/
[4]https://www.drugabuse.gov/trends-statistics/monitoring-future/monitoring-future-study-trends-in-prevalence-various-drugs
[5]https://bmjopen.bmj.com/content/4/8/e005894.short
[6]https://www.atr.org/new-cdc-data-more-9-million-adults-vape-regularly-united-states
[7]https://www.cdc.gov/tobacco/basic_information/e-cigarettes/pdfs/Electronic-Cigarettes-Infographic-p.pdf
[8]https://www.reuters.com/article/us-health-ecigs-us-adults/almost-one-in-20-u-s-adults-now-use-e-cigarettes-idUSKCN1LC2DN
[9]https://www.census.gov/quickfacts/fact/table/US/PST045217
[10]https://www.statista.com/statistics/184418/percentage-of-cigarette-smoking-in-the-us/
[11]https://www.ncbi.nlm.nih.gov/pubmed/30167658
[12]https://www.blu.com/en/GB/blog/about/uk-vaping-trends/uk-vaping-trends.html?countryselect=true
[13]https://www.drugabuse.gov/publications/drugfacts/nationwide-trends
[14]https://teens.drugabuse.gov/blog/post/your-environment-may-influence-drug-use
[15]https://www.livescience.com/32728-baby-month-is-almost-here-.html
[16]https://www.cdc.gov/tobacco/data_statistics/fact_sheets/fast_facts/index.htm
[17]https://www.dea.gov/druginfo/drug_data_sheets/Marijuana.pdf
[18]https://familycouncil.org/?p=11795
[19]https://www.cdc.gov/tobacco/data_statistics/fact_sheets/fast_facts/index.htm
[20]https://www.niaaa.nih.gov/alcohol-health/overview-alcohol-consumption/alcohol-facts-and-statistics
[21]https://www.drugabuse.gov/related-topics/trends-statistics
[22]https://www.the-hospitalist.org/hospitalist/article/121573/health-policy/medical-marijuana-cuts-medicare-spending-and-may-reduce
[23]https://www.cdc.gov/features/costsofdrinking/index.html
[24]https://www.drugabuse.gov/publications/research-reports/tobacco-nicotine-e-cigarettes/what-scope-tobacco-use-its-cost-to-society
[25]https://www.washingtonpost.com/national/2018/01/10/study-legal-marijuana-could-generate-more-than-132-billion-in-federal-tax-revenue-and-1-million-jobs/?utm_term=.5d7f7a3f5195
[26]https://www.cdc.gov/features/costsofdrinking/index.html
[27]https://www.researchgate.net/publication/319945301_A_Substance_Use_Cost_Calculator_for_US_Employers_With_an_Emphasis_on_Prescription_Pain_Medication_Misuse
[28]https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4139456/

[29]https://www.publichealthpost.org/research/marijuanas-cooling-effect-on-partner-violence/
[30]https://www.who.int/violence_injury_prevention/violence/world_report/factsheets/fs_intimate.pdf
[31]https://www.researchgate.net/publication/323367100_Intimate_Partner_Violence_victimization_and_opioid_use_by_pregnant_women_in_rural_Appalachia_A_cross-sectional_analysis
[32]https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5448292/
[33]https://www.aclu.org/gallery/marijuana-arrests-numbers
[34]https://www.cdc.gov/features/costsofdrinking/index.html
[35]https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5448292/
[36] https://www.drugabuse.gov/related-topics/trends-statistics/overdose-death-rates
[37] https://nces.ed.gov/programs/statereform/tab5_1.asp
[38] https://www.nytimes.com/2011/02/17/business/economy/17regulation.html
[39]https://newsok.com/article/3748886/1-in-4-high-school-students-work-us-census-finds-including-many-in-oklahoma-to-support-families
[40] https://www.thespruce.com/legal-age-marriage-laws-by-state-2300971
[41]http://samhda.s3-us-gov-west-1.amazonaws.com/s3fs-public/field-uploads-protected/studies/TEDS-A-2016/TEDS-A-2016-datasets/TEDS-A-2016-DS0001/TEDS-A-2016-DS0001-info/TEDS-A-2016-DS0001-info-codebook.pdf
[42] https://www.justice.gov/archive/ndic/pubs44/44849/44849p.pdf
[43] https://www.omicsonline.org/open-access/health-policy-for-marijuana-2155-6105-S11-018.php?aid=87175

Appendix

Model 1.4: a Coupled Difference Equation solver (Python version 3.6)

```
import numpy as np
'''
Matrix A is formatted as such

S_0  0    S_0(1-S_0-V_0)      -S_0 V_0
S_1  0    S_1(1-S_1-V_1)      -S_1 V_1
and so on
0         V_0  V_0(1-S_0-V_0)      S_0 V_0
0         V_1  V_1(1-S_1-V_1)      S_1 V_1
and so on

Vector b is just
[S_1 S_2 ... V_1 V_2 ...]

The output x is just
1-mu  1-xi  alpha  beta  gamma
'''
A = np.genfromtxt("matrix.csv", delimiter=",")
b = np.genfromtxt("vector.csv")
x = np.linalg.inv(np.matmul(A.T, A)).dot(A.T.dot(b))

# Comparing predictions to actual data
for i in range(len(b)):
     print(i, b[i], A.dot(x)[i])

# 2018 data below
s = [0.148739194]
v = [0.04210568085]


for i in range(10):
     s.append((x[0]*s[-1] + x[2]*s[-1]*(1-s[-1]-v[-1])) -
x[4]*s[-1]*v[-1])
     v.append((x[1]*v[-1] + x[3]*v[-1]*(1-s[-1]-v[-1])) +
x[4]*s[-1]*v[-1])

for i in range(len(s)):
```

```
    print(s[i], v[i])

# Searching for steady-state solution
for i in range(100):
    s.append((x[0]*s[-1] + x[2]*s[-1]*(1-s[-1]-v[-1])) -
x[4]*s[-1]*v[-1])
    v.append((x[1]*v[-1] + x[3]*v[-1]*(1-s[-1]-v[-1])) +
x[4]*s[-1]*v[-1])
print(s[-1])
print(v[-1])
```

Model 2.4.1: a Decision Tree Classifier trained on public data of varied-background individuals and run on a created test set of 300 students (Python version 3.6)

```
import pandas as pd
import numpy as np
from sklearn.model_selection import train_test_split
from matplotlib import pyplot as plt
from collections import Counter
from math import log, inf
import random
from sklearn.tree import DecisionTreeClassifier
from sklearn.metrics import classification_report,
confusion_matrix, accuracy_score
from sklearn.multiclass import OneVsRestClassifier
import pickle

data = pd.read_csv("tedsa_2016_puf.csv")
data = data[
   ["AGE", "ARRESTS", "EDUC", "EMPLOY", "ETHNIC", "SEX",
"LIVARAG", "MARSTAT", "PRIMINC", "PSYPROB", "RACE", "SUB1"]] #
demographic categories
classifier = DecisionTreeClassifier()
for i in range(0,len(data)):
   if(i%10000==0):
       print(i)
   if(data.iloc[i,-1]==6):
       data.iloc[i,-1] = 7
templist = []
for i in range(0,len(data)):
```

```
   if(data.iloc[i,-1]!=2 and data.iloc[i,-1]!=4 and
data.iloc[i,-1]!=7 and data.iloc[i,-1]!=12):
       templist.append(i)
data = data.drop(templist)
X = data.iloc[:, :-1] # columns of independent variables
y = data.iloc[:, -1] # column of dependent variables
X_train, X_test, y_train, y_test = train_test_split(X, y,
test_size=0.20)
classifier.fit(X_train, y_train)
y_pred = classifier.predict(X_test)
pickle.dump(classifier, open('tree.pk','wb'))
print(confusion_matrix(y_test, y_pred))
print(classification_report(y_test, y_pred))
print('Accuracy: '+str(accuracy_score( y_test,y_pred)))
tempdata = pd.read_csv("300test1.csv") # reads csv file into
pandas library
print(tempdata)
templist = []
classifier = pickle.load(open('tree.pk', 'rb'))
for i in range(0,len(tempdata)):

templist.append(classifier.predict_proba([tempdata.iloc[i,:]]))
counter = 0
counter_alcohol = 0
counter_marijuana = 0
counter_opioids = 0
counter_nicotine = 0
maxvalue = -1
maxindex = -1
for arr in templist:
   for i in range(0, len(arr)):
       if(arr[i]>maxvalue):
           maxvalue = arr[i]
           maxindex = i
       if(arr[i]==1.0):
           counter+=1
   if maxindex==0:
      counter_alcohol+=1
   if maxindex==1:
       counter_marijuana+=1
```

```
   if maxindex==2:
       counter_opioids+=1
   if maxindex==3:
       counter_nicotine+=1
   maxvalue = -1
   maxindex = -1
print("Number of total drug high school senior users:
"+str(counter))
print("Number of nicotine high school senior users: "
+str(counter_nicotine))
print("Number of marijuana high school senior users: "
+str(counter_marijuana))
print("Number of alcohol high school senior users: "
+str(counter_alcohol))
print("Number of unprescribed high school senior opioid users: "
+str(counter_opioids))
```

Model 2.4.2: a random generator of 300 high school seniors that reflect the demographics of the 17-year old U.S. citizens (Python version 3.6)

```
import random
students = 300
l = [[87.961, 90.3688, 100, 100],
     [0, 34, 34, 100, 100],
     [1.6553, 12.6674, 13.3569, 17.4342, 100, 100],
     [49.2, 100, 100],
     [0.067, 100, 100, 100],
     [82.6539, 82.7516, 96, 96, 100, 100],
     [17, 100, 100],
     [0, 0.9, 1.1, 14.8, 91.3, 96.4, 96.4, 100, 100, 100]]
# l stores prefix sums for each of the eight variable
demographics (three of them are assumed to be constant for all
seniors)
def pickRand(lst): # picks a random demographic (weighted by
probability)
    r = random.random() * 100
    for i in range(len(lst)):
        if lst[i] > r:
            return i
studentStat = []
```

```
# studentStat is a list of lists storing the demographics of
each student
for i in range(students):
    localStat = []
    # localStat is a list storing the demographic of a single
student
    for j in range(len(l)):
        if j != 0:
            localStat.append(pickRand(l[j])+1)
        else:
            localStat.append(pickRand(l[j]))
    tempStat = [2] + localStat[0:1] + [5] + localStat[1:5] + [1]
+ localStat[5:]
    studentStat.append(tempStat)

for lst in studentStat:
    print(lst)
```