One is too many and a thousand not enough: Substance Use and Abuse

Team ID: 12774

March 2019

M3 Challenge

1 Executive Summary

Substance abuse has been a gradually increasing issue in our society, with terms such as "opioid epidemic" being coined to depict the severity and frequency with which substance abuse occurs. Abuse of tobacco, alcohol, and illicit drugs costs over 230 billion annually in health care expenses, and over 740 billion annually in total costs accounting for crime and lost work productivity. It is clear that substance abuse is taking a significant toll on our society, and the concern is only rising with the introduction of new potentially dangerous drug delivery systems, most notably vaping in teenagers and young adults.

Our team isolated the primary factors contributing to the spread of vaping and likelihood of substance abuse, then developed mathematical models which were optimized for the factors we isolated. Our model for the spread of nicotine use due to vaping uses cigarette data as a lower bound and utilizes a stochastic graph-based model to account for social dynamics and model an infectious spread through a population.

Our model for the likelihood of substance abuse was created by splitting the problem into two smaller models, one which examines how high school students transition into adulthood by looking at the evolution of SES (socioeconomic status) factors over time. This was done with a combination of stochastic modelling, normal distributions, and random forests. This revealed surprising results about the evolution of SES factors over time which we then utilized to create a model for drug usage for alcohol, tobacco, marijuana, and opiods. This was done with a random forest model which we then used to predict the proportions of the 300 students that will use these specific drugs over time.

We then developed a metric for the impact of the abuse of certain substances by considering the toxicity, dependence, harm, and financial strain of the substances. We recorded risk parameters for each of these factors and further created a weight vector to consider the significance of each risk parameter. We then took the dot product to compute final ranking values for each of the four substances.

Finally, we verified all of our models by comparing them to real trends, such as the growth of cigarette usage for our model for the growth of vaping.

Contents

2 Darth Vapor

2.1 Restatement of Problem

This problem asks us to do the following:

- Build a mathematical model that predicts the spread of nicotine use due to vaping over the next 10 years
- Analyze how the spread of vaping compares to the spread of cigarettes.

2.2 Assumptions and Justifications

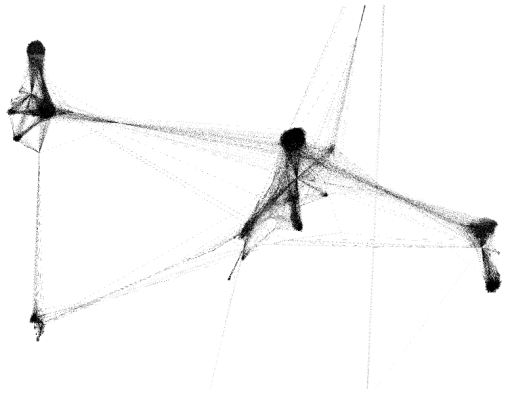
1. The probability that an individual will be influenced by a friend to vape is at least that of an individual accepting a cigarette from a friend.
   Justification: Vaping is more prevalent in the youth than cigarettes are, as is shown by the large percentages of youth who have vaped in their lifetime. This allows data for cigarettes to be used as a lower bound for a model of vaping.

2.3 Developing the Model

Originally, an approach focused around developing a compartmental model was deployed. The spread of vaping was compared to the spread of an infectious disease, as both involve susceptible and infectious groups, a disease "duration," and an infectious rate. An SIS model was determined to be the optimal compartmental model, as other compartmental models use factors such as immunity and carriers, which are not applicable to the spread of vaping. An SIS model is defined by the following differential equations:

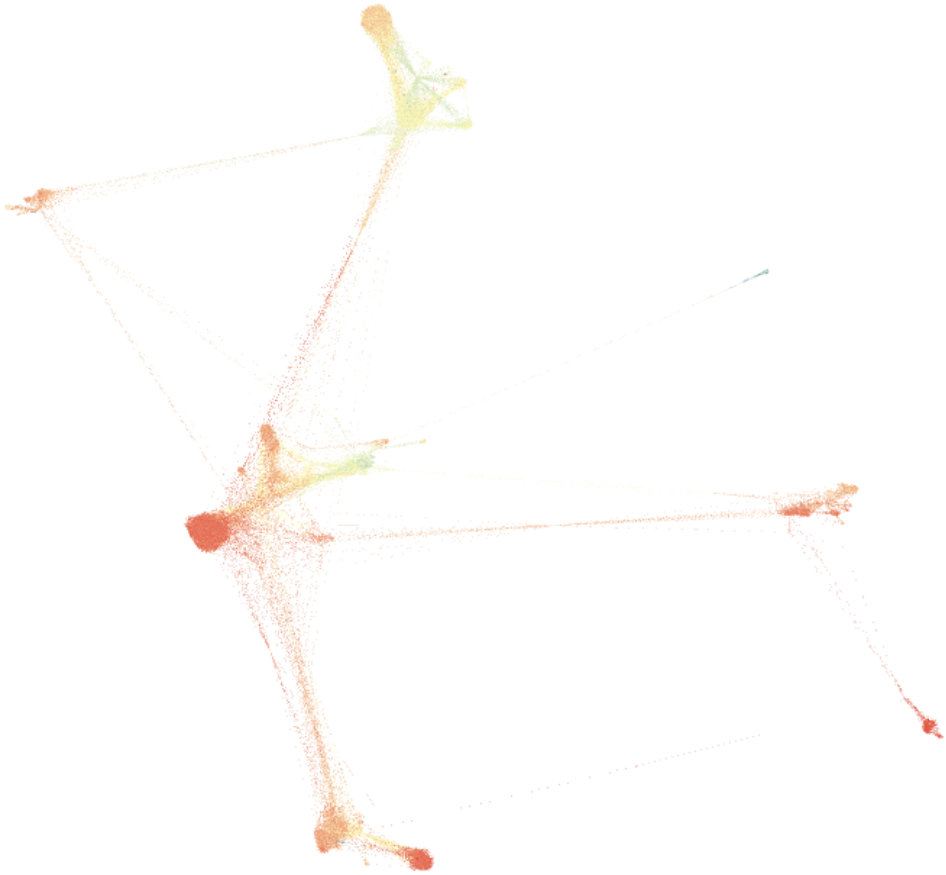
$$\frac{dS}{dt} = -\frac{\beta SI}{N} + \gamma I \qquad (1) \qquad\qquad \frac{dI}{dt} = \frac{\beta SI}{N} - \gamma I \qquad (2)$$

Solving this set of differential equations requires determining the total population N, which is trivial, and the infectious rate β. Factors influencing β were determined to be a combination of the current percentage of the population which vapes, and the probability that an individual would vape as a function of social factors such as age, friends who vape, and social dynamics. Based off of these requirements, it was determined that a superior way to model the spread of vaping through a population (and particularly through a young population) is with a graph-based model. In this model, each of $N = 4039$ people are modeled as a node with varying numbers of edges exiting the node, each representing a "relationship" to another person. A sample population was retrieved from the Stanford Network Analysis Project's[3] open dataset of Facebook circles:



A limitation of this data was that ages were not provided, but this limitation was circumvented by our custom labeling of each node with an age. To accomplish this, a

few ($N/20$) nodes were labeled at random according to the U.S. population distribution from census.gov[4]. From this seed, a breadth-first search was performed and each node was assigned an age according to the age of its neighbors (friends tend to be of similar ages). the result was the following (red = younger to green = older):



Each node is labeled with a boolean value representing whether or not they regularly vape (denoted as "addicted"). Using data from a survey from icpsr.umich.edu[9], a risk factor was determined for each age of the three age groups: 8th grader, 10th grader, and 12th grader. The answers to three questions are of particular interest:

1. Have you ever smoked a cigarette?
2. If your best friend offered you a cigarette, would you smoke the cigarette?
3. How old are you?

Using the responses to these three questions, a correlation was determined between age and likeliness of being influenced by a friend into smoking a cigarette for the first time, which was applied to the risk factor of each node. As online resources suggest that the start of the prevalence of vaping was in 2003, a time step of 1 day and a simulation start of mid-2004 was decided on. At each time step, a person u's susceptibility was computed as:

$$1 - (1 - R) \prod_{v} \frac{1 - S_u}{(u_{\text{age}} - v_{\text{age}})^4}$$

Where v iterates over each neighbor of u, R is a common "mutation" rate (meaning that u was not influenced but rather developed the habit on their own), S_u is the susceptibility

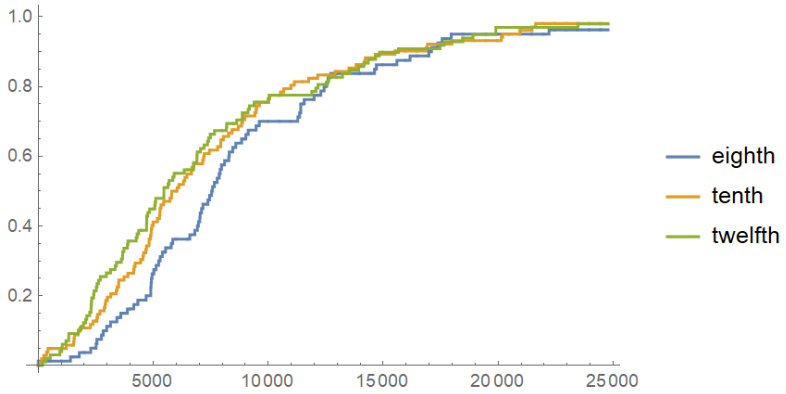
of u, and u_{age} is the age of the person represented by node u. This accounts for the fact that having multiple vaping friends can increasingly influence one to develop the habit themselves, with a friend being more influential if they are in a similar age group. The common mutation rate R was solved for by using the data provided by the National Institute on Drug Abuse[2], as the difference between the lifetime percentage of 8th graders who have vaped in 2015 and the 10th graders of 2017 can be explained by the latter living through more "susceptibility cycles" (once we remove the effect of influence of peers which is explicitly known).

It is worth noting that this susceptibility estimate is hugely dependent on the choice of time step, so once the model was established we tweaked the constant until the amount of time it took for cumulative vaping prevalence to reach 2018 levels was $2018 - 2004 = 14$ years (about 5000 days).

In order to equate vaping with cigarette smoking, we must divide the average amount of nicotine in an e-cigarette canister by that of a cigarette. Cigarettes deliver about 1 mg of nicotine per unit while a pod of JUUL contains 40 mg of nicotine. A four pack of JUUL pods costs 15.99 and the average user is estimated to spend $180 a month, putting one at $\frac{180}{15.99} * \frac{40}{1} = 450$ cigarette-equivalent units of nicotine per month and thus 5400 per year.

2.4 Validation of the Model

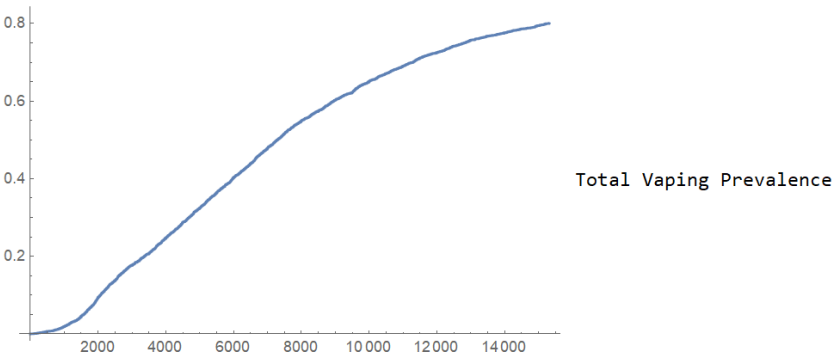
After running the model from the onset of 2004 (d = 0) until 2018 (d = 5000), we inspected the predicted prevalence of vaping in eighth, tenth, and twelfth graders (in percent of the population) and compared against the data provided by National Institute on Drug Abuse. The model proved to be accurate to these provided numbers, for example at d = 4900 we produced [22.5, 37.2549, 44.898] for these three age groups, as compared to the expected [21.50, 36.90, 42.50].



Despite the model being a lower bound based off of data for cigarette use, the growth rate for vaping still exceeded the growth rate of cigarettes pre-legislation, which demonstrates the infectious nature and social dynamics of vaping.

2.5 Strengths and Limitations

A token strength of this model is how robust it is for different types of people, and how accurately it models social dynamics. A graph-based model is a superb choice for modeling how an infectious disease spreads through a real population, as it accounts for critical social factors such as friends and social groups. Since it was demonstrated earlier that the spread of vaping can be represented as the spread of an infectious disease,



this model accounts for factors that other mathematical models, such as the SIS model, cannot.

A limitation of this model is the overarching fact that it is a lower bound, based on the assumption that vaping has been spreading (and will continue to spread) quicker than cigarette smoking. Additionally, there are a high number of constants and distributions for which we could do nothing but read literature online to estimate e.g. the age distribution of the friends of someone aged X.

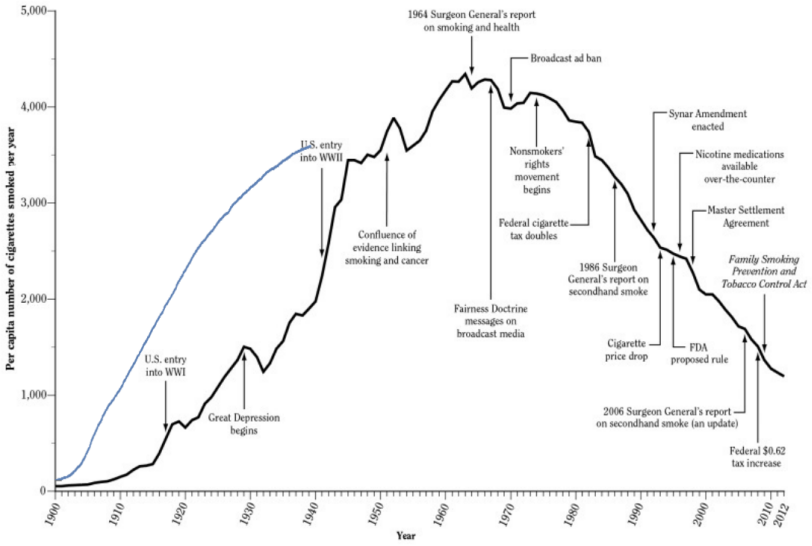
Another limitation of this model is the binary method in which individuals were labeled. Individuals were labeled as either vaping or non-vaping, but it is possible for an individual to have only vaped once and still be considered part of the susceptible population, which this model does not consider.

2.6 Summary

According to this simulation, in 10 years (d = 9000) we expect the total vaping prevalence to be at 60.26% as opposed to the current 31.74%. It is also important to note that the prevalence of vaping is highest in the ages 13-19, consistent with the idea that this is the age most targeted by advertisements of vaping.

Additionally, we can note that the growth of vaping prevalence in our simulated population appears to follow a skewed logistic curve, which is approximately what we would expect. A reason for the skewed nature could be our (somewhat) late start in simulation (as in not starting in 2003 when zero people vaped, but rather seeding the start with early statistics). If this is too much of a stretch, it is also important to notice that our graph looks very similar to that of the provided Adult per capita cigarette consumption[5] (before the major health reports of course), except that the inflection point occurs *much* earlier. Overlay of our model (blue) on that of cigarette smoking (black) adjusted for the average amount of nicotine that a vape smoker will intake in a year is below.

The nicotine spread due to vaping is simply the same graph as previously presented, except scaled to 38 years (14000 days) of vaping growth on the x-axis and the y-axis scaled by 5400 cigarette-equivalent doses of nicotine per capita per "addicted" node.



3 Above or Under the Influence?

3.1 Restatement of Problem

This problem asks us to do the following:

- Given a class of 300 high school seniors from a particular high school, create a robust model to predict how many of these seniors will use the following substances: nicotine, marijuana, alcohol, and un-prescribed opioids.

3.2 Assumptions and Justifications

1. When generating a predictive model for drug use, we can focus on specific factors involving socioeconomic status and personal situation
   Justification: The most important factors that affect predicted drug use are household income, employment, if any, family structure, and geographical location. (9,10)

2. Some other factors that can also affect drug usage are demographics
   Justification: When considering drug abuse statistics, some relevant factors include race, ethnicity and gender (11)

3.3 Mathematical Model

We begin the creation of the mathematical model by splitting the problem into two separate models. The first model determines how a student in a high school will progress into their their future given categorical variables of their upbringing. The second model takes any given person and predict the probabilities of them using nicotine, marijuana, alcohol, and/or opiods. We consider the following construction for the model.

Define

$$\mathbb{F} = \bigtimes_{i=1}^{n} F_i = F_1 \times F_2 \dots F_n$$

where $\times$ is the Cartesian product. Henceforth, we call F the *feature space*. We now define the functions from the first model. Let

$$h = \langle h_1, h_2, \dots, h_n \rangle \in \mathbb{F}$$

denote a high school student. We want to analyze the evolution of the features $h_1, h_2, \ldots h_n$ over time. Thus, we define functions

$$C_i : \mathbb{F} \longrightarrow F_i$$

that predict the evolution of feature i in the feature space until they are an adult, at which point we assume that they remain in a stable situation. Hence, the total transformation that occurs to student h is $C : \mathbb{F} \longrightarrow \mathbb{F}$ with

$$C(h) = \langle C_1(h), C_2(h), \ldots, C_n(h) \rangle.$$

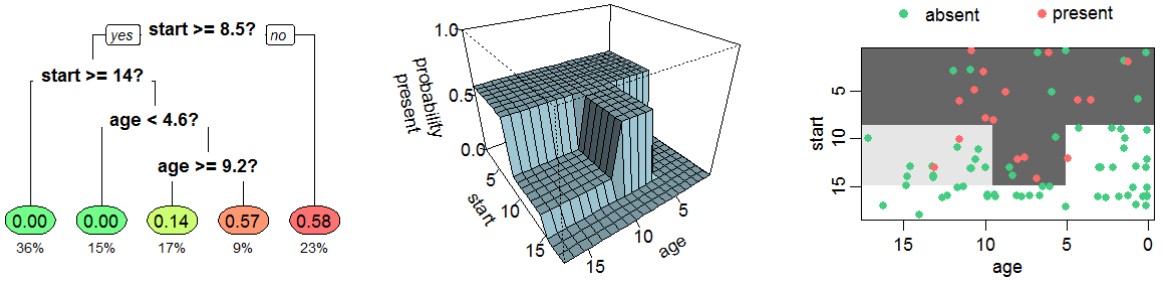
For the second model, we want to create four predictions functions that will give the probability that a given person will use any of the four drugs. Let these functions be P_1, P_2, P_3, P_4 such that they provide a prediction on whether a person will consume alcohol, tobacco, marijuana, and/or opiods, respectively. Any given function P_i is defined as $P_i : \mathbb{F} \longrightarrow \{0, 1\}$.

3.3.1 Random Forest Model

In this section we will describe the general random forest model which will be used for many of the predictive models that will be used in the next two sections that describe both model one and model two. The main idea of random forest is to consider many different CART models and aggregate them to create a more robust and stable prediction from a given feature space to a single output (e.g. $P : \{R_1, R_2, \ldots, R_n\} \longrightarrow g$).

3.3.1.1 CART Model

The CART model is the backbone of the random forest model in which classification tree is built as a predictive model. A classification tree is a tree in which each branch splits based on a given rule. For example, if F_1 is in the feature space, one of the junctions in the tree may be $f > \alpha$ where $f \in F_1$. All points that are run through the tree and arrive at this junction will be split based on whether or not their value for f_1 is greater than α. In essence, we are partitioning an n-dimensional space into various n-dimensional rectangular prisms. For each of these partitions, a constant function is fit to the data. A sample of the process can be seen below with only a two-dimensional feature space. Specifically, we state that the tree partitions the feature space into M regions. We define



the following

$$\hat{p}_{m,k} = \frac{1}{N_m} \sum_{x_i \in R_m} I(y_i = k)$$

where N_m is the number of observations in region R_m for class k in the prediction set. Essentially, this is the proportion of all points that are in R_m that are observed to be of

class k and node m of the tree itself. In addition $I(y_i = k)$ is 1 if y_i is k and 0 otherwise. We define our prediction for a point that falls in this region to be

$$k(m) = \arg\max_k \hat{p}_{m,k}$$

which is just the class that is most observed in the training set within a specific region. Finally, we want an error function which we will seek to minimize to perfect the tree. Specifically, we use the Gini Coefficient which is defined as

$$Q_m(T) = \sum_{k=1}^{k} \hat{p}_{m,k}(1 - \hat{p}_{m,k})$$

Where T is a tree defined by its nodes one of which is m.

3.3.1.2 Random Forest

The limitation with the CART model is that it is very rigid and is prone to error in the validation set. Hence, we create an aggregate method that is based on many CART models used in unison. The key idea is to create many different trees that do a good job of fitting the data and have them "vote" as to the true classification of a point.

To make sure that different CART models are made each time. We only allow the CART model to choose from a random sample of the features for each split that it considers. From this we create the final model $T = \{T_1, T_2, \ldots, T_l\}$. For a given point $p \in \mathbb{F}$, we define

$$T(p) = \text{mode}\{T_1(p), \ldots, T_n(p)\}$$

Empirically, this is very stable and is not prone to deviations in the validation set. Note that for the purposes of this competition, our implementation for the CART and random forest model will come from the randomForest R package which is available on CRAN for free usage. (13, 16)

3.3.2 Model One: Transition

For these two models we used two datasets from the ICPSR institute at the University of Michigan. The first dataset is a long-term longitudal study of students as they progress from high school to adulthood. Hence, this dataset is perfect for determining the evolution in a student from an earlier time to adulthood. The second is a large national survey of adults that includes their situational features and then presents their drug usage. For model one we will only be using the first study. (12,14)

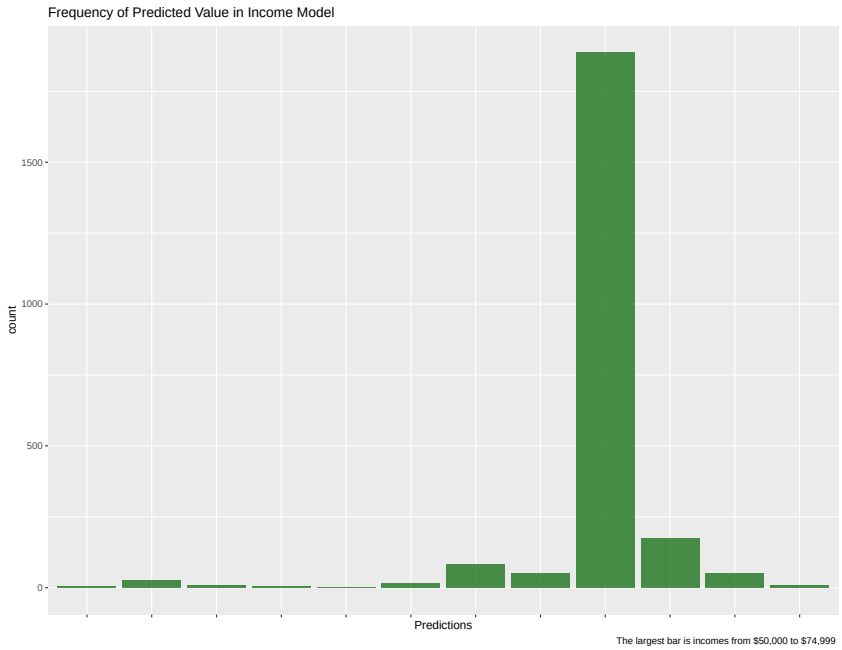
Within the longitudal study there were two tables that we used. The first was the measurement of the variables in the study from when the students were in high-school and the second is from when they were adults. From this we find our feature space which we define as

$$\mathbb{F} = \langle \text{Income}, \text{Race}, \text{Gender}, \text{Mother}, \text{Age}, \text{Work}, \text{Geography}, \text{Father} \rangle$$

Factors such as "Mother" and "Father" indicate whether they grew up with such a figure in their household. We noticed that certain features were immutable such as race, gender, mother, father, and age. Hence, what this leaves was that we had to model the change in income, work status, and geography of residence.

3.3.2.1 Income Model

When we run the income model using the random forest algorithm that was described above, we get a function $I : \mathbb{F} \longrightarrow$ Income. The model, however, had a 77.72% error in the validation set which indicated to us that there was almost no correlation between a student's previous status and their future income-earning ability. This analysis is further bolstered by the following histogram that shows predicted values by the model. Hence,



as there is no correlation between a students past and their current economic standing, we decided to model their future income using a normal distribution (15) with $\mu = 72641$ and $\sigma = 5000$. Hence,

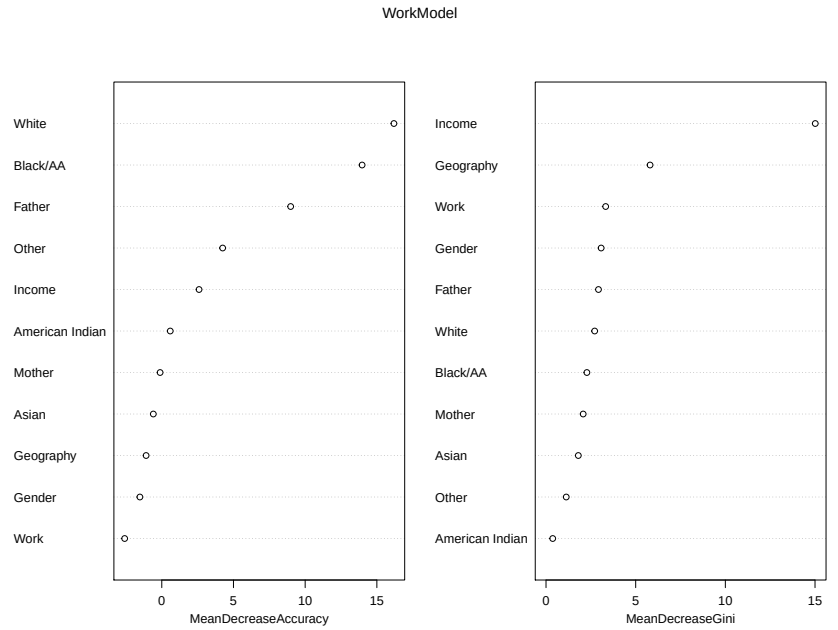
$$I(f) = \frac{1}{\sqrt{50000000\pi}} e^{\frac{(f_{income}-72641)^2}{50000000}}$$

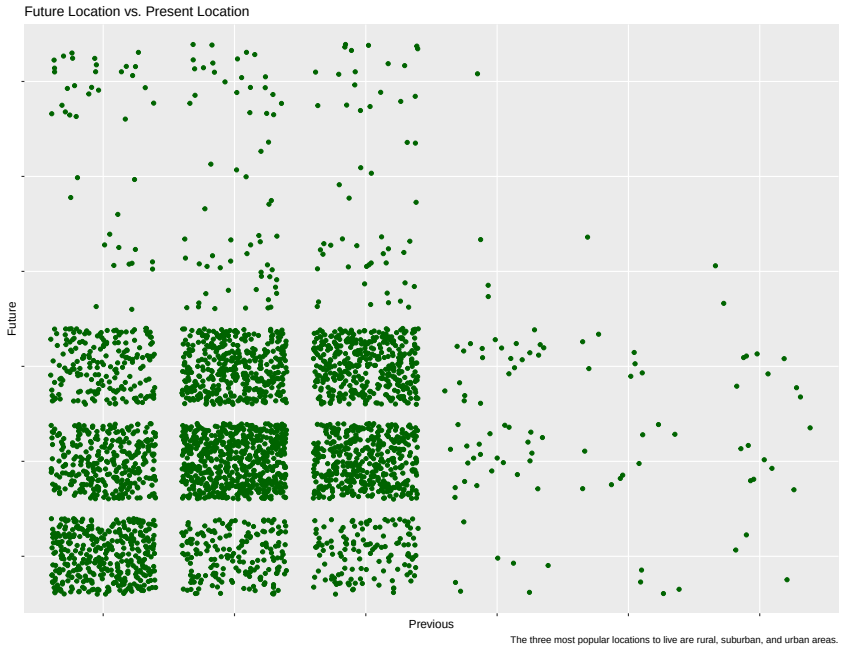
3.3.2.2 Work Situation Model

Again we describe a random forest model to predict whether or not a given student will be working in the future. The function is $W : \mathbb{F} \longrightarrow$ Work. This model performed extremely well with a test error of only 5.65%. The random forest model is very interesting as it is able to measure the importance of different variables by measuring the decrease in improperly classified variables and the Gini Coefficient. This is shown below in the following dotplot. Note that the numbers represent relative importance. As can be seen in the Gini Coefficient section, it makes sense that income is the largest factor as income usually means that someone who is inclined to working in high school will work later in life.

3.3.2.3 Geographic Location Model

We also utilized a random forest in this model and we noticed that with an error of 56.5% error in the model that previous factors are not quite correlated with future residence. Indeed, the following plot of previous residence to future residence shows that there are three locations that people tend to randomly end up in. We can define $G(f) = P$ where P is a probability function that returns "Rural" with probability $\frac{1}{5}$, "Suburban" with





probability $\frac{2}{5}$, and "Urban" with probability $\frac{2}{5}$ which are based on the final density of the students residence.

3.3.2.4 Final Model

With all of this we can enumerate the final model in a table

Income	Race	Gender	Mother	Age	Work	Geography	Father
$I(f)$	f_{race}	f_{gender}	f_{mother}	$f_{age} + t$ (t is constant)	$W(f)$	$G(f)$	f_{father}

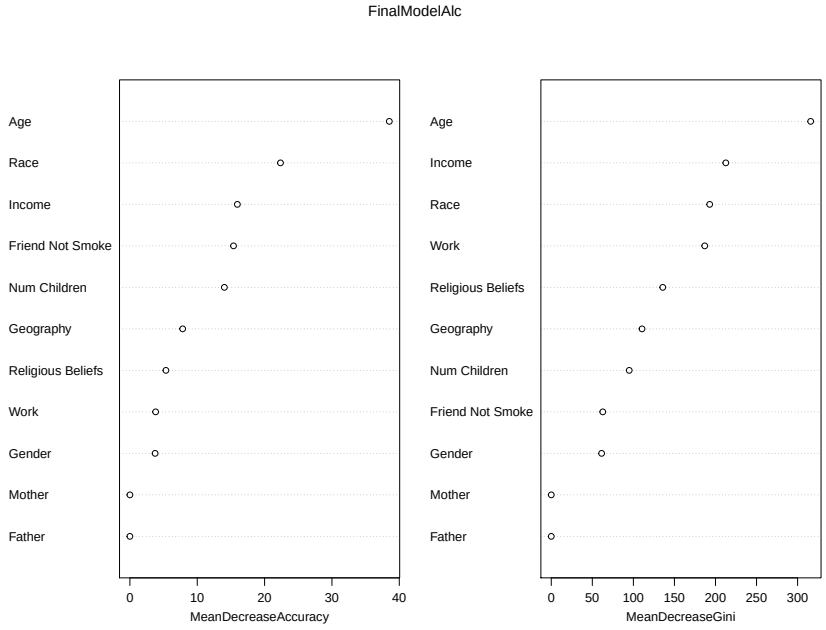
Table 1: **All Predictive Functions for the Transition Model**

3.3.3 Final Model

In the final model, we utilize a dataset with 30000 samples in which the categorical factors are quite similar to those that were presented in the transitional model. Again, here we have to create four random forest models for each substance: alcohol, cigarettes, marijuana, and opiods.

3.3.3.1 Alcohol Model

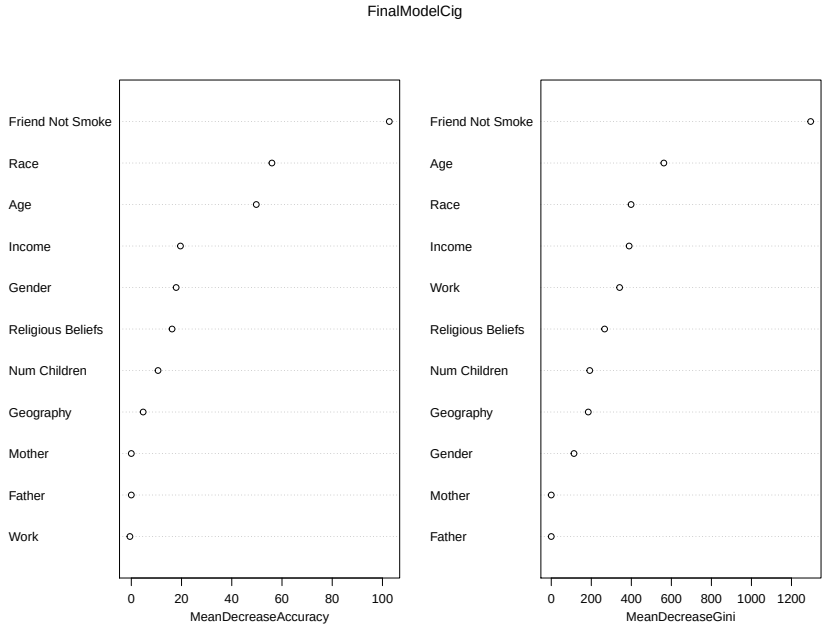
In this model we have a function $A : \mathbb{F} \longrightarrow \{0, 1\}$. This model on the validation set has an error of 9.19% which is quite good predictive ability as controlled substances tend to have a high amount of randomness in use. As was seen in the Work Function in section 3.3.2.2, we can create an importance score graph to examine what is important in the model. This model makes sense as the most important factor is age which should have



the largest impact according to intuition. Also note that the model gave a predicted 99.7% of people will use alcohol eventually.

3.3.3.2 Cigarette Model

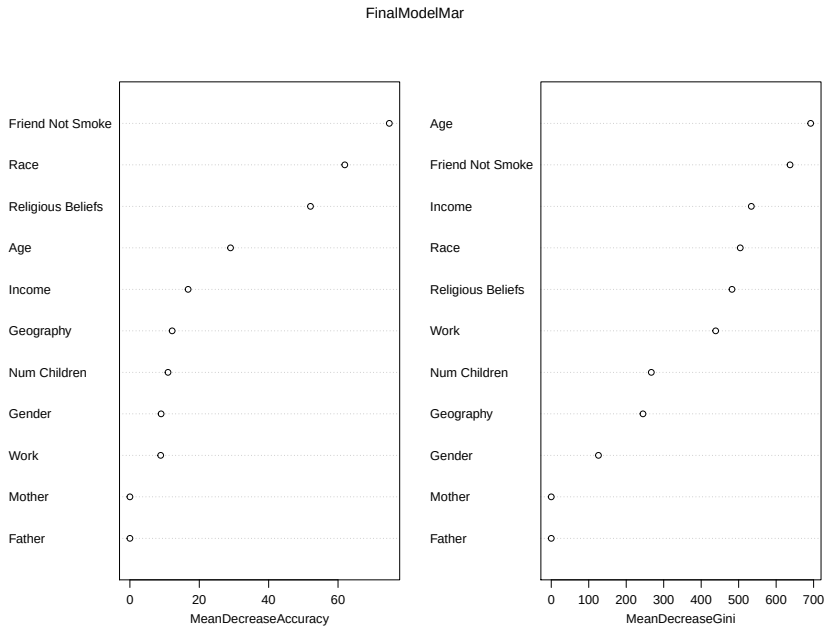
In this model we have a function $C : \mathbb{F} \longrightarrow \{0, 1\}$. The error rate for this one was higher at 28.9% but was expected due to the large variability in people that use drugs. The importance score is shown below



The percentage of people that were predicted to use cigarettes in this case is 75.3%.

3.3.3.3 Marijuana Model

In this model we have a function $M : \mathbb{F} \longrightarrow \{0, 1\}$. The error rate for this one was higher at 33.74% but was expected due to the large variability in people that use drugs. The importance score is shown below

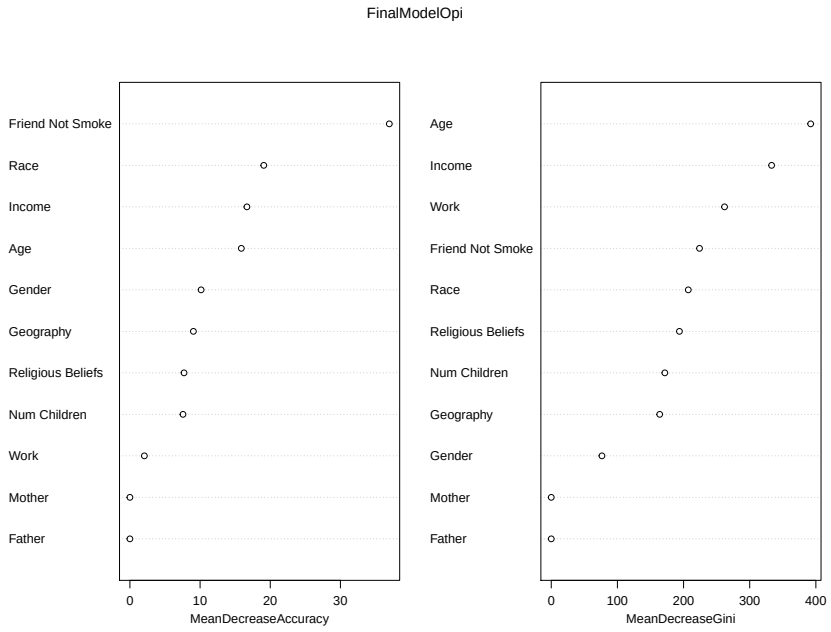


The model predicts that 56.6% of people will attempt to use marijuana in their life.

3.3.3.4 Opiod Model

In this model we have a function $O : \mathbb{F} \longrightarrow \{0, 1\}$. The error rate for this one was quite low at 15.6% and was expected due to the large variability in people that use drugs. The

importance score is shown below



The model predicts that 1.13% of people will attempt to use opiods in their life.

3.3.3.5 Results

From this we have our four predictive functions. We can thus find the number of the 300 studnets that will use each drug by multiplying 300 by the proportion preicted by each model which is shown in the table below.

Alcohol	Cigarettes	Marijuana	Opiods
299	226	170	34

Table 2: **Predicted Numbers of Students Using Each Drug**

3.4 Strengths and Limitations

There are a few strengths that we can see in our model development

- Highly robust due to the random forest algorithm
- Extremely general and will be able to handle any population sample given
- Can handle the evolution of a population over time
- Based on large US government surveys which provide a good random sample and also increase the confidence interval on our results
- Importance score plots remove "black box nature" of many machine learning algorithms as we can see exactly what impacts the model

We can also see a few limitations that can be mitigated

- High error on some predictions due to the random nature of drug use at times
- Lack of clear function at times means reliance on technology

3.5 Summary

The final summary of what can be seen in the data is very interesting in a few ways. The Specific transition based models provide a surprising result that most people's future income does not depend on their past. In addition, we can see that future propensity to work is heavily linked to current working status for students. Finally, geographic location is almost a Stochastic process with constant probabilities for transitions states. In terms of the final model, there were more interesting results. Probability of using alcohol is most highly linked to age. Propensity to use tobacco is linked extremely to friends that smoke which reinforces the idea that drugs travels in social circles. This is also reinforced by the first problem. Marijuana usage is most heavily linked to Race and Age and finally, opiod usage is very heavily linked to income as we should expect.

4 Ripples

4.1 Restatement of Problem

This problem asks us to do the following:

- Create a metric to analyze the impact of substance abuse
- Rank the given 4 substances (alcohol, marijuana, nicotine, opioids) with the metric based on the danger they pose to individuals and society

4.2 Assumptions and Justifications

1. We assume that the factors involved in assessing the danger of a substance are toxicity, harm, dependence, and financial strain
   Justification: The toxicity and dependence are both important measures of the danger of a substance to a given individual. In addition, the harm factor and financial strain further analyze the way that a person is affected in his/her interactions with society and the rest of the world as a result of their addiction (17)(18)

4.3 Metric Development

To develop a metric for the impact of substance use, consider the documented effects of the four drugs on toxicity, dependence, harm, and financial strain from (17) and (18). We use risk parameters to denote the extent of impact of each drug on each of the listed factors. The risk parameters are represented on a scale from 0.00 to 3.00, where 0.00 represents no effect and 3.00 represents the most potent effect. The data points from the first source were determined through the consultation of a national group of psychiatrists on the Royal College of Psychiatrists' register as specialists in addiction. (17) Those in the second source were computed through the assessment of nineteen Dutch experts with a variety in expertise who assessed the harmful effects of the given drugs. (18) The results from the two studies were then averaged to determine the effects of the four drugs on the non-financial factors of toxicity, dependence, and harm, as shown in the below three tables.

Drug	**Toxicity**		
	Acute Toxicity	Chronic Toxicity	Intravenous
Alcohol	1.90	2.44	N/A
Marijuana	1.04	1.47	N/A
Nicotine	0.72	2.90	N/A
Un-prescribed Opioids	2.59	2.27	3.00

Drug	**Dependence**		
	Physical Dependence	Psychological Dependence	Pleasure
Alcohol	1.90	1.87	2.30
Marijuana	1.70	0.97	1.90
Nicotine	2.60	2.31	2.30
Un-prescribed Opioids	3.0	2.91	3.00

Drug	**Harm**	
	Individual Harm	Social Harm
Alcohol	2.18	2.58
Marijuana	1.48	1.39
Nicotine	1.43	1.69
Un-prescribed Opioids	2.05	2.39

To determine the effects of the four drugs on financial factors, data on the cost of health care was recorded from the aforementioned studies. 3 of the 4 annual substance prices were taken from the Delphi Behavioral Health Group. (19) The 4^{th} price, which was for a nicotine addiction, was taken assuming a cigarette habit with the worst case scenario price. Smoking a pack of cigarettes every day for a year in New York City would cost $\$5.51 \cdot 365 = \$4,745$. (20) We then computed the Financial Cost Index of each of the drugs, which was determined by the following formula:

$$FCI = 1.00 + \frac{2.00}{49255} \cdot (C - 4745),$$

where C denotes the cost of the drug. This index was determined so that the least cost, $\$4,745$, is mapped to an index of 1.00 and the greatest cost, $\$54,000$, so mapped to an index of 3.00. The values were chosen since the least cost, although small compared to the greatest cost, is still a significant contribution to the individual's financial strain and the overall impact of the drug.

Drug	**Financial Strain**	
	Health Care Cost	Financial Cost Index
Alcohol	2.10	1.01
Marijuana	1.50	1.09
Nicotine	2.40	1.00
Un-prescribed Opioids	3.00	3.00

For each of the four tables, the values in each column were averaged (disregarding any unavailable values) to determine the average impact of the drugs on each of the four given factors, once again as a risk parameter computed out of 3.00. The resulting values were recorded in the below graph.

Drug	**Risk Parameters**			
	Toxicity	Dependence	Harm	Financial Strain
Alcohol	2.17	2.02	2.38	1.56
Marijuana	1.26	1.52	1.44	1.30
Nicotine	1.81	2.40	1.56	1.70
Un-prescribed Opioids	2.6	2.97	2.22	3.00

Finally, to determine the overall impact of substance abuse of any of the given four drugs, we express the risk of each drug as a risk vector of the four risk parameters as follows:

$$v_i = \langle t_i, d_i, h_i, f_i \rangle,$$

where t_i, d_i, h_i, f_i denote the toxicity, dependence, harm, and financial strain risk parameters, respectively. We now consider the weights of each risk parameter based on the significance of the effect of the parameter on the individual and on other parameters.

Dependence is given the highest weight of 0.35 since an individual's dependence on the drug will significantly affect the frequency of his or her use of the drug, which then affects the drug's toxicity and its harm. This frequency also directly correlates to the financial strain caused by the drug due to the cost and severity of the drug.

Financial strain is then given the next highest weight of 0.30 since this strain will impact the individual's stress and quality of life, which can affect the frequency of drug use.

Harm is given the weight 0.25 since it is the all-encompassing value reflecting the effect of the drug on the individual's lifestyle, which in turn, also affects the frequency of drug use. Finally, toxicity is given the weight of 0.10.

We represent these weights as the weight vector

$$w = \langle 0.10, 0.35, 0.25, 0.30 \rangle.$$

Thus, the overall impact of substance abuse of a given drug can be computed as the dot product of the drug's risk vector with the weight vector:

$$Impact = v_i \cdot w.$$

These computations yield that the four drugs are have increasing impacts of substance abuse in the following order: marijuana, alcohol, nicotine, and un-prescribed opioids.

5 Conclusion

5.1 Strengths

- Our first model avoided using many assumptions and was very robust
- Our second model was extremely general and was not susceptible to changes in population
- The third model is simple and is able to be updated when new factors arise

5.2 Weaknesses

- We had some error rates that were large
- Some models may be hard to update with new factors on the fly

6 Citations

1. https://www.cdc.gov/tobacco/data_statistics/sgr/e-cigarettes/pdfs/2016_sgr_entire_report_508.pdf
2. https://www.drugabuse.gov/trends-statistics/monitoring-future/monitoring-future-s
3. http://snap.stanford.edu/
4. https://factfinder.census.gov/faces/nav/jsf/pages/index.xhtml
5. https://www.ncbi.nlm.nih.gov/core/lw/2.0/html/tileshop_pmc/tileshop_pmc_inline.html?title=Figure%202.1.%20Adult*%20per%20capita%20cigarette%20consumption20and%20major%20smoking%20and%20health%20events%2C%20United%20States%2C%201900%020132012.&p=BOOKS&id=294310_ch2f1.jpg
6. https://www.cdc.gov/tobacco/basic_information/e-cigarettes/pdfs/Electronic-Cigarepdf
7. https://www.metro.us/body-and-mind/health/juul-cost-per-month
8. https://www.nih.gov/
9. https://doi.org/10.1300/J251v05n03_05
10. https://doi.org/10.3109/00952998509016846
11. https://doi.org/10.1300/J233v06n02_06
12. https://www.icpsr.umich.edu/icpsrweb/DSDR/studies/21600
13. https://web.stanford.edu/~hastie/Papers/ESLII.pd
14. https://www.icpsr.umich.edu/icpsrweb/DSDR/studies/36361
15. http://www.wikiwand.com/en/Normal_distribution
16. https://cran.r-project.org/web/packages/randomForest/randomForest.pdf
17. https://doi.org/10.1016/S0140-6736(07)60464-4
18. https://doi.org/10.1159/000317249
19. https://www.rehabspot.com/treatment/paying-for-rehab/cost-of-addiction
20. https://www.ny1.com/nyc/all-boroughs/health-and-medicine/2018/06/01/new-york-city

7 Appedix

```cpp
#include <bits/stdc++.h>
#define all(a) a.begin(), a.end()
#define forn(i,n) for(int i = 0; i < (int) n; i++)
#define ios ios::sync_with_stdio(false); cin.tie(0); cout.tie(0)

using namespace std;

vector<vector<int>> adj;
vector<bool> seen;
vector<double> age;

unsigned seed =
chrono::system_clock::now().time_since_epoch().count();
default_random_engine rnd(seed);
uniform_real_distribution<double> uniform(0.0, 1.0);

// agedist[i] = proportion of population that is less than
5*(i+1) years old
vector<double> agedist{0.0, 0.0, 0.068, 0.138, 0.211, 0.285,
0.359, 0.432, 0.508, 0.59, 0.676, 0.758, 0.827, 0.882, 0.923,
0.952, 0.974, 1.0};
double random_age() {
  double r = uniform(rnd);

  for (int i = 0; i < agedist.size(); i++)
    if (r <= agedist[i])
      return 5 * i + 5 * uniform(rnd);
}

int main() {
  freopen("age.txt", "w", stdout);
  freopen("graph.txt", "r", stdin);

  int n, m;
  cin >> n >> m;
//  cout << n << " " << m << "\n";

  seen.resize(n);
  adj.resize(n);
  age.resize(n);
  forn(i, m) {
    int a, b;
    cin >> a >> b;

    assert(a < n && b < n);
    adj[a].push_back(b);
    adj[b].push_back(a);
  }

  queue<int> q;
  forn(i, n / 20) {
    int r = uniform(rnd) * n;
    age[r] = random age();
    q.push(r);
  }

  while(!q.empty()) {
    int u = q.front();
    q.pop();
```

```
56        if (seen[u]) continue;
57        seen[u] = 1;
58
59        int nbrs = 0, sm = 0;
60        for(int v : adj[u])
61          if (age[v]) {
62            nbrs++;
63            sm += age[v];
64          } else
65            q.push(v);
66
67        if (nbrs && uniform(rnd) < 0.1 + nbrs * 0.009) {
68          // friends have similar ages
69          normal_distribution<double> distribution(1.0 * sm / nbrs,
    5);
70          age[u] = distribution(rnd);
71        } else
72          age[u] = random_age();
73      }
74
75      forn(i, n) cout << age[i] - 5 << "\n";
76    }
77
```

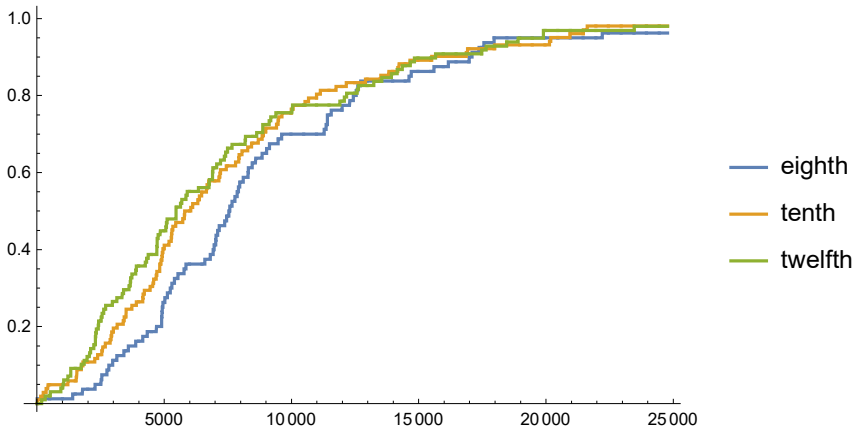
```cpp
#include <bits/stdc++.h>
#define all(a) a.begin(), a.end()
#define forn(i,n) for(int i = 0; i < (int) n; i++)

using namespace std;

vector<vector<int>> adj;
vector<double> age;
int n;
vector<bool> addicted;

unsigned seed =
chrono::system_clock::now().time_since_epoch().count();
default_random_engine rnd(seed);
uniform_real_distribution<double> uniform(0.0, 1.0);

double measure(double lo, double hi) {
  int num = 0, den = 0;
  forn(i, n) if (lo <= age[i] && age[i] < hi) {
    den++;
    num += addicted[i];
  }

  return 1.0 * num / den;
}

vector<pair<double, double>> influence{
  {0, 0}, // 0 year-olds know to abstain
  {12, 2.227966},
  {13, 3.6134655},
  {14, 4.097354},
  {15, 4.300107},
  {16, 5.20536375},
  {17, 5.4757184},
  {19, 5.9609712},
  {23.5, 4.3742811},
  {37, 2.723509},
  {48.5, 1.910963},
  {60, 1.853268},
  {65, 1.750422},
  {1000, 1.750422} // lets say anyone over 65 has the same
influenceability as a 65 year old
};

double susceptibility(double age, double friend age) {
  auto lo = lower_bound(all(influence),
pair<double,double>{age, 0});
  auto prev = lo-1;

  double x1 = prev->first, y1 = prev->second;
  double x2 = lo->first, y2 = lo->second;

  double factor = (y2-y1) / (x2-x1) * (age - x1) + y1;

  return 1e-5 * factor / pow(abs(age - friend_age), 0.25);
}

void time_step() {
  vector<bool> nw_addicted(addicted);
```

```
57      vector<bool> seen(n);
58
59      queue<int> q;
60      q.push(0);
61
62      while(!q.empty()) {
63        int u = q.front();
64        q.pop();
65        if (seen[u]) continue;
66        seen[u] = 1;
67
68        double ni = (1 - 0.000013); // probability of not influenced
69        for (int v : adj[u]) {
70          if (addicted[v])
71            ni *= (1 - susceptibility(age[u], age[v]));
72          q.push(v);
73        }
74
75        if (uniform(rnd) > ni)
76          nw_addicted[u] = 1;
77      }
78
79      swap(nw_addicted, addicted);
80    }
81
82    int main() {
83      cout << seed << "\n";
84      freopen("graph.txt", "r", stdin);
85      ifstream age_in("age.txt");
86
87      int m;
88      cin >> n >> m;
89
90      adj.resize(n);
91      age.resize(n);
92      addicted.resize(n);
93      forn(i, m) {
94        int a, b;
95        cin >> a >> b;
96
97        assert(a < n && b < n);
98        adj[a].push_back(b);
99        adj[b].push_back(a);
100     }
101
102     forn(i, n) age_in >> age[i];
103
104     vector<double> grade8, grade10, grade12, total;
105     for(int t = 0;; t++) {
106       int a = 0;
107       forn(i, n) a += addicted[i];
108
109       grade8.push_back(measure(13, 15));
110       grade10.push_back(measure(15, 17));
111       grade12.push_back(measure(17, 19));
112       total.push_back(1.0 * a / n);
113
114       if (t % 1000 == 0){
115         cout << 100.0 * a / n << "\n";
```

```
116         cout << "\t08: " << grade8.back() << "\n";
117         cout << "\t10: " << grade10.back() << "\n";
118         cout << "\t12: " << grade12.back() << "\n";
119       }
120
121       if (a > 0.80 * n) break;
122       time_step();
123     }
124
125     ofstream o8("grade8.txt");
126     o8 << "{";
127     forn(i, grade8.size()) o8 << grade8[i] << ",}"[i ==
    grade8.size() - 1];
128
129     ofstream o10("grade10.txt");
130     o10 << "{";
131     forn(i, grade10.size()) o10 << grade10[i] << ",}"[i ==
    grade10.size() - 1];
132
133     ofstream o12("grade12.txt");
134     o12 << "{";
135     forn(i, grade12.size()) o12 << grade12[i] << ",}"[i ==
    grade12.size() - 1];
136
137     ofstream o("total.txt");
138     o << "{";
139     forn(i, total.size()) o << total[i] << ",}"[i == total.size()
    - 1];
140
141     o8.close();
142     o10.close();
143     o12.close();
144     o.close();
145   }
146
```

In[100]:=
```
g8 = ReadList[
     "C:\\Users\\Ben Mirtchouk\\Desktop\\Computer Stuff\\C++\\moody\\facebook\\grade8.txt"
     ][[1]];
g10 = ReadList[
     "C:\\Users\\Ben Mirtchouk\\Desktop\\Computer Stuff\\C++\\moody\\facebook\\grade10.txt
       "][[1]];
g12 = ReadList[
     "C:\\Users\\Ben Mirtchouk\\Desktop\\Computer Stuff\\C++\\moody\\facebook\\grade12.txt
       "][[1]];
ListLinePlot[{g8, g10, g12}, PlotLegends → {"eighth", "tenth", "twelfth"}]
```

Out[103]=



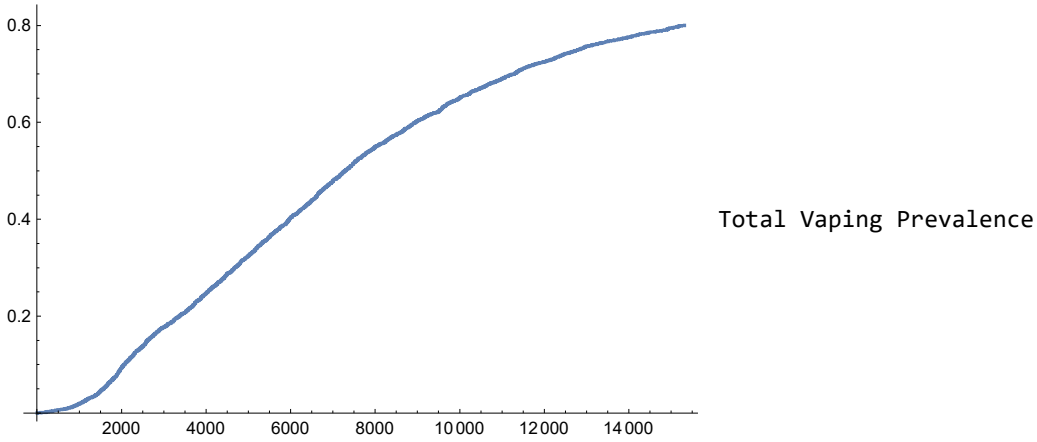
In[104]:=
```
f[n_] := {g8[[n]], g10[[n]], g12[[n]]};
```

In[112]:=
```
f[4900]
```

Out[112]=
```
{0.225, 0.372549, 0.44898}
```

In[131]:=
```
tot = ReadList[
     "C:\\Users\\Ben Mirtchouk\\Desktop\\Computer Stuff\\C++\\moody\\facebook\\total.txt"
     ][[1]];
ListLinePlot[tot, PlotLegends → "Total Vaping Prevalence"]
```

Out[132]=

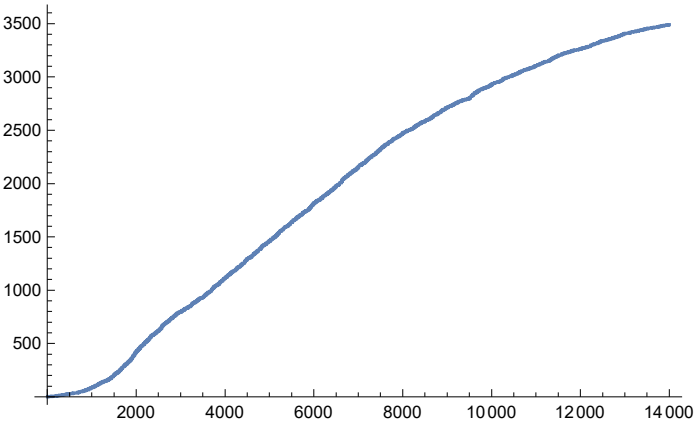


In[134]:=
```
tot[[4900]]
```

Out[134]=
```
0.317405
```

In[147]:=
```
ListLinePlot[Table[tot[[i]] * 4500 , {i, 1, 14000}]]
```

Out[147]=



In[146]:=
```
14000 / 365.
```

Out[146]=
```
38.3562
```