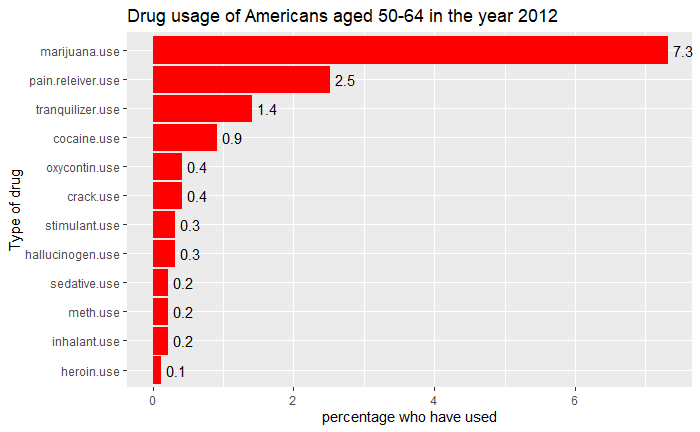
Exploring Drug Use Data Sets

1. **Introduction**

Drug abuse is an ever-growing problem in the United States. Thousands die every year and drugs continue to ruin lives and rip families apart. It is essential that we learn how to predict who will and will not decide to do drugs to aid our country as we fight this battle. MY goal with this project was to design a model (SVM) that can take in information about people and determine if they will take drugs or not. I will use my SVM model along with k-fold cross validation and tuning of hyperparameters to build the most accurate model I can for determining if someone will take drugs. With all the drugs out there in the world, I decided only to focus on cannabis. This will serve as a proof of concept to test weather you can predict weather someone will take drugs before they have taken them.

1. **Dataset**

I used two datasets for this project. The first of the two was the FiveThirtyEight dataset about drug use. It keeps track of summary statistics regarding age and drug use. Below is the recreation I did of the original report on this dataset.

This graph shows the percentage of people age 50-64 that used the various classes of drugs in 2012. This graph is largely what people would expect it to look like. The most popular drug among old people is marijuana and the second most is pain killers. The second dataset I used came from the machine learning repository. It has 1885 entries and characteristics describing impulsivity, level of education and things like age and gender. It also has data describing if the person has done many kinds of drugs in the past. This data set had a variable describing weather or not a person had claimed to have taken a made-up drug called semer. I removed all the people who claimed to have taken this drug because if they are lying about using an unknown drug, I do not want them to be part of my model. I then changed the data so that some of the columns would have binary output instead of many categorical outputs. I changed the outputs to be 1 if a person has consumed the drug recently, one month ago, and those that have not consumed recently. Once I did this in my R Markdown file, I turned the data into a csv and completed the rest of my analysis in python.

1. **Feature Selection**

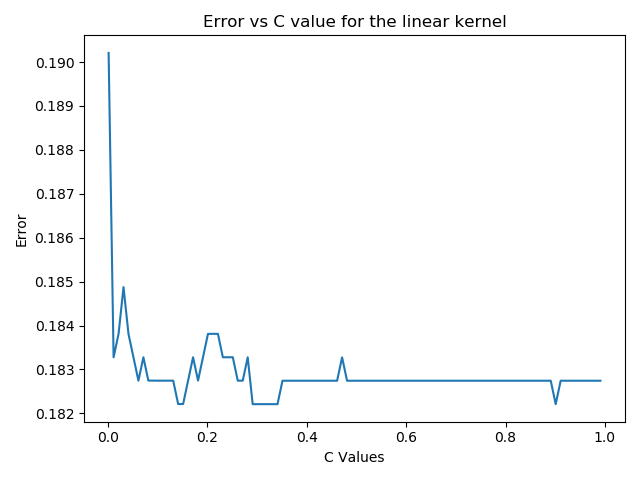
As cannabis was the most consumed drug of the many that were kept track of, I decided to use the new categorical variables I made on weather or not a person had consumed cannabis recently to perform my analysis. I wanted to give my model as much information to work with as possible so I decided to use al of the descriptive statistics that I could to train and test my model.

1. **Methods**

I was set on using soft margin SVM for my analysis but did not know which kernel or C value would work best with my data. I decided to run my training and testing with 4 different kernels and many different C values to finely tune my hyperparameters. I implemented the SVM, and K-fold cross validation found in the Sykit-learn library.

* 1. **SVM modeling**

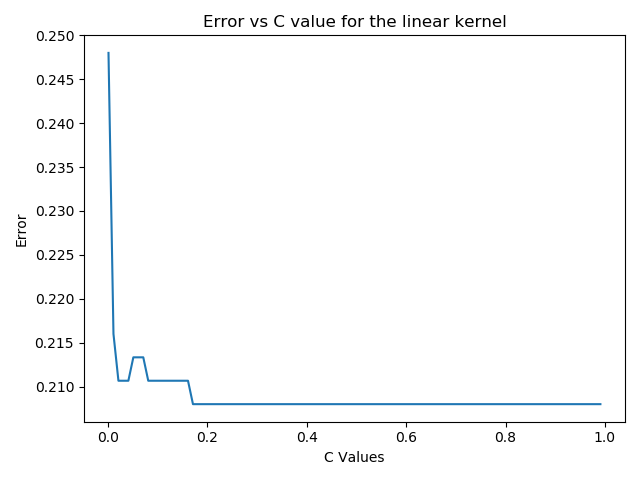
When I originally completed my analysis, I was finding that my results were largely dependent on the random seed and not the actual hyperparameters. When plotting the error vs the C values for each kernel, some of them showed much more variation that I was comfortable using as my final model. Specifically, the linear kernel.



This variance led to my decision to use K-fold cross validation to get averages and reduce the variance in the graph.

* 1. **SVM Modeling With K-fold Cross Validation**

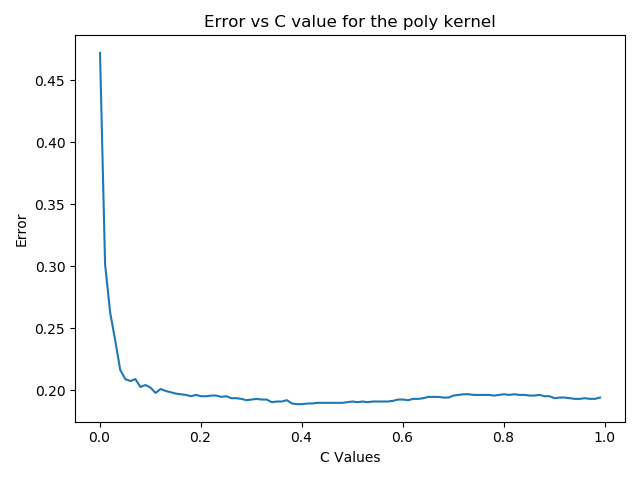
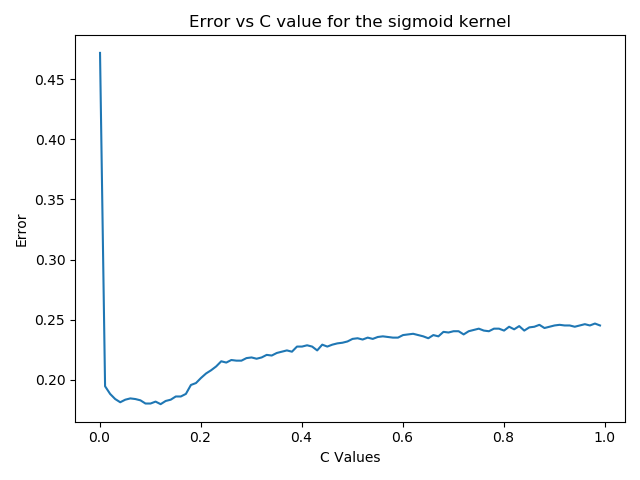
With computational complexity in mind, I decided to implement my K-fold cross validation with 5 folds for each combination of kernel and C value. This led to much more consistent graphs and a different optimal kernel and C value combination. This was the linear kernel graph after using K-fold cross validation.



As you can see, it is a much smoother graph and offers more confidence that the results can be trusted. When using K-fold cross validation, almost regardless of what the random seed was, my analysis found that the sigmoid kernel was the most accurate with a small C-value.

1. **Results**

My model worked better than expected. When I went into this project, I thought that it would be very difficult to predict weather someone will do drugs based on their characteristics, but my models turned out to be quite accurate. Even the worst of the models were still relatively accurate at making predictions.



The sigmoid kernel turned out to be the most accurate with an ideal error rate of just 17.6% but all the models had times when they were at most, only wrong 20% of the time.

1. **Conclusion and Future Work**

This project attempted to come up with the best SVM model for predicting weather a person will consume cannabis based on their personality and physical traits. My initial experimentation found interesting results for the ideal model but did not appear to be a very reliable result. When I added the K-fold cross validation with my models, the results became much more consistent. In the future, I would like to expand my model to make the same predictions for other drugs as well as cannabis. I would also like to implement lasso regression to control the feature selection.

**References**

E. Fehrman, A. K. Muhammad, E. M. Mirkes, V. Egan and A. N. Gorban, "The Five Factor Model of personality and evaluation of drug consumption risk.," arXiv [[Web Link]](https://archive.ics.uci.edu/ml/datasets/Drug+consumption+%28quantified%29), 2015

https://scikit-learn.org/stable/, scikit-learn Machine Learning in Python.

Flowers, Anna Maria Barry-Jester And Andrew. “How Baby Boomers Get High.”*FiveThirtyEight*, 21 Apr. 2017, fivethirtyeight.com/features/how-baby-boomers-get high.