**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Frequency of Cannabis User Predictions**

**­­­­**

**Amanda Bowerman**

MSUNet ID: bowerm37

Project Category: General Machine Learning

Final Report

1. **Introduction**

Cannabis legalization has been a hot topic in the United States since California legalized it for medical purposes in 1996. By 2012 many states followed in the footsteps of California with Colorado and Washington being some of the first to fully legalize the sale of Cannabis for recreational purposes. One of the fears of legalization of Cannabis is that we are unaware of the effects of the drug on adolescents/young adults. This project aims to develop a prediction model that will predict whether someone is a cannabis user based on a multitude of different features. These features include age, gender, education, ethnicity, country and impulsiveness to name a few. The method used for this prediction model is K-Nearest Neighbors (KNN).

1. **Related Work**

Baby boomers grew up with the “Woodstock mentality”. How Baby Boomers Get High, a report on FiveThirtyEight, discusses exactly what this mentality is, and how baby boomers in a 2012 survey get high. With alcohol use being extremely high (most likely due to the age of the baby boomer group, which is between the ages of 50 to 64), it has been removed to determine the percentage of other substances. About 7.3% from that age group said they had used Marijuana (another name for Cannabis) in the past year. Heroin, being the lowest recorded percentage was at about 0.05%. Figure 1 in section 3 shows the dispersion of drug use in the baby boomers.

1. **Dataset**

The first dataset used for this project was the Drug Use by Age dataset from FiveThirtyEight. This data set included percentage of users for a variety of different substances along with the frequency of use. Figure 1 below, as mentioned in section 2, displays the percentage of baby boomers using each substance.

**Chart, waterfall chart

Description automatically generated**

Figure 1: Dispersion of Drug Use in Baby Boomers

The primary dataset used in the model is Drug Consumption dataset from UCI. This dataset contains 1,885 entries, with 32 features per entry. The first 12 attributes are known for each respondent: age, level of education, gender, country of residence, ethnicity, as well as their personality measurements. These measurements include NEO-FFI-R (neuroticism (Nscore), extraversion (Escore), openness to experience (Oscore), agreeableness (Ascore), and conscientiousness (Cscore)), as well as BIS-11 (impulsivity) and ImpSS (sensation seeking). Each of the substances the respondent chooses between the 7 options: “never used”, “used over a decade ago”, “used in the last decade”, “used in the last year”, “used in the last month”, “used in the last week”, and “used in the last day”. They are also given an ID number to allow for anonymity.

**4 Methods**

K-Nearest Neighbors was for this prediction model. Multiple models were tested, each with a different set of features to determine the best predictors. K-Nearest Neighbors estimates the conditional probability for class j as the fraction of points in N\_0 whose response values equal j:

KNN then applies the Bayes rule and classifies the test observation x0 to the class with the largest probability.

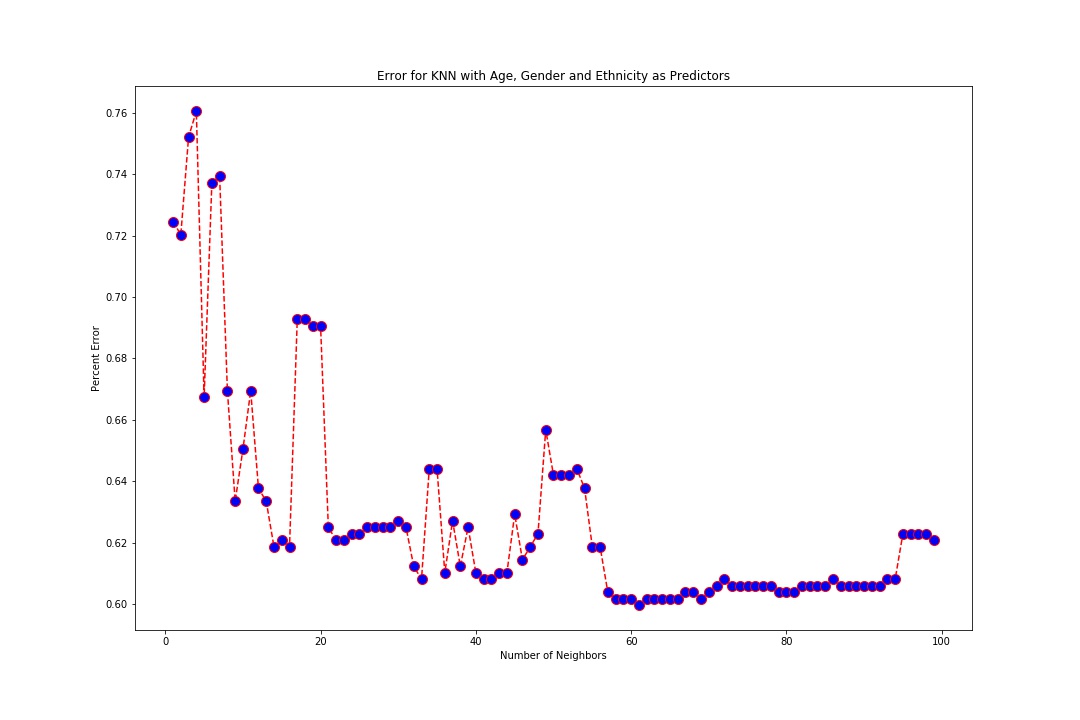
**5 Experiments and Discussion**

Graphical user interface, chart

Description automatically generatedError rate and F-1 score were used to evaluate the accuracy of the model. After splitting the data into a training (1,413 observations) and validation (472 observations), 4 different models were created. These four models contained different predictors to determine which were the most valuable in predicting the frequency of Cannabis users.

Figure 2: Error rate for predictors Age and Ethnicity

The results in Figure 2 show that the K with the lowest error rate is 34 with an error rate of 68%. This is results in an F-1 accuracy score of 29%. This model does a very poor job of predicting the class of frequency of Cannabis users.

Figure 3: Error rate for predictors Age, Gender and Ethnicity

Chart, scatter chart

Description automatically generatedThis model performed slightly better than the model above, with the best K being 61. This resulted in a F-1 accuracy score of 40%. This model does a better job of predicting the frequency of Cannabis users but is still not as accurate as it could be.

Figure 4: Error rate for predictors Education, Age, and Impulsiveness

Using education, age and impulsiveness resulted with the best K being 33 with an error rate of 68%. This results in a F-1 accuracy score of 32%. This performs better than the model with predictors age, gender, and ethnicity, but worse than the first model with predictors age and ethnicity.

Chart, scatter chart

Description automatically generated

Figure 5: Error rate for predictors Age, Gender, Ethnicity, Country, Impulsiveness, Nscore, Ascore, and SS

This model with multiple predictors had the lowest error rate, 62%, when K is 61. This results in a F-1 accuracy score of 38%. This performs worse than the first model, but better than the other two.

K-Nearest Neighbors does an okay job of predicting the frequency of Cannabis users, but the accuracy of the models could be increased. This could be done through more data collection or through the use of different methods.

**6 Conclusions and Future Work**

The goal of this project was to attempt to predict/classify the frequency of Cannabis users based on different features including age, gender, ethnicity, and education, etc. The machine learning technique used was K-Nearest Neighbors (KNN). The best model resulted from predictors age, gender, and ethnicity and K of 61. While the accuracy was only 40%, this could be improved upon with more observations.

For future works on this project could include: (i) collecting more data, and (ii) implementing different methods such as SVM or Logistic Regression.

**References**

[1] “Timeline of Cannabis Laws in the United States.” *Wikipedia*, Wikimedia Foundation, 14 Apr. 2021, en.wikipedia.org/wiki/Timeline\_of\_cannabis\_laws\_in\_the\_United\_States#:~:text=In%201996%2C%20California%20became%20the,legalize%20cannabis%20for%20recreational%20use.

[2] 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://arxiv.org/abs/1506.06297), 2015

[3] Barry-Jester, Anna Maria, and Andrew Flowers. “How Baby Boomers Get High.” *FiveThirtyEight*, FiveThirtyEight, 23 Apr. 2015, fivethirtyeight.com/features/how-baby-boomers-get-high/.