# Using Machine Learning Methods to Discover: Does Personality Pick the Poison?

Kelsey Cherland

California State University, Fullerton

December 2021

# Origin and Basics of the Data

## Source

UCI Repository:
https://archive.ics.uci.edu/ml/machine-learning-databases/00373/

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 [https://arxiv.org/abs/1506.06297], 2015.

## Basic Data Details

After cleaning the data, 1,885 participants remained. Data collected includes education, nationality, age, personality, and their use of 18 psychoactive drugs ranging from chocolate to heroin.

# What Are the Researchers Asking?

- How do personality, gender, education, nationality, age, and other attributes affect an individual's risk of consuming and/or abusing drugs?
- What personality traits are the most important for evaluating a person's risk of consumption of a particular drug? Do they vary from drug to drug?
- Statisticians must be able to evaluate, given the data, IF these questions can be answered and also HOW.

# Background: Five-Factor Model for Personality

- Uses 5 independent traits on a continuum that are expected to remain stable throughout a person's life.
- Conscientiousness: impulsive, disorganized vs. disciplined, organized, careful
- Agreeableness: suspicious, uncooperative vs. trusting, cooperative, helpful
- Neuroticism: calm, confident vs. anxious, pessimistic
- Openness to Experience: prefers routine, practical vs. imaginative, appreciates art and unconventional ideas
- Extraversion: reserved, thoughtful vs. sociable, talkative, assertive

# Background: Impulsivity and Sensation-Seeking

- Impulsivity is defined as a tendency to act without adequate forethought. Measured via The Barratt Impulsiveness Scale (BIS-11; Patton et al., 1995)
- Sensation-Seeking is defined by the search for experiences and feelings that are varied, novel, complex, and intense, and by the readiness to take risks for the sake of such experiences. It is one of 5 other personality factors from Zuckerman's ZKPQ questionaire often used for psychometrics.

# Background: Drugs and their Pleiade

"Pleiades"are correlated modular groups (via biostatistics, introduced in 1931).

| Heroin | Ecstasy | Benzodiazepine | Other |
|---|---|---|---|
| Cocaine | Cocaine | Cocaine | Alcohol |
| Methadone | Amphetamines | Amphetamines | Amyl Nitrite |
| Crack | Cannabis | Methadone | Caffeine |
| Heroin | Ketamine | Benzodiazepine | Chocolate |
|  | Ecstasy |  | Nicotine |
|  | Legal Highs |  | VSA |
|  | Magic Mushrooms |  | Semeron |
|  | LSD |  |  |

Legal Highs: mephedrone, salvia, and various legal smoking mixtures; VSA: Volatile Substance Abuse, i.e. glues, gases, and aerosols; Semeron: A fictitious drug for survey design

# Background: Measuring Drug Use

- Drug use is nested and not discrete. Subjects in the category ‘Used in last day’ also belong to the categories ‘Used in last week’, ‘Used in last month’, ‘Used in last year’, and ‘Used in last decade’.
- For most studies, subjects who used a drug more than a decade ago are not considered a drug user.
- The original research considered a decade-, year-, month-, and week-based split of the data. I decided to use decade-based because it’s the simplest binary classification of use/non-user subjects.

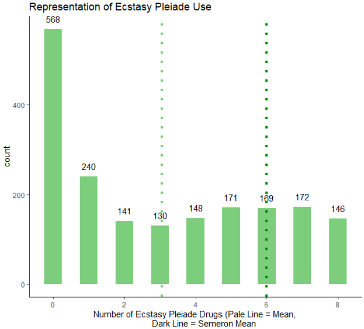
# Exploratory Data Analysis

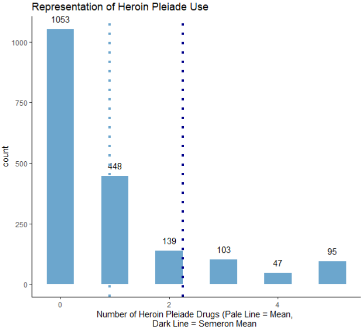
Fun facts about drug use in the dataset:

- Semeron is a fictitious drug used to identify "overclaimers". Eight (8) participants said they had used Semeron. Every person used drugs from the Ecstasy Pleiade with an average that was double the sample mean (6 vs 3). Seven (7) out of 8 over-claimers used drugs from the Heroin and Benzodiazepine Pleiades at a rate that was more than double the sample mean.
- Ecstasy is the most popular Pleiade and has two modes.
- The Heroin and Benzodiazepine Pleiades have a steeper decline and are overall less popular than drugs in the Ecstasy Pleiade.

# Exploratory Data Analysis

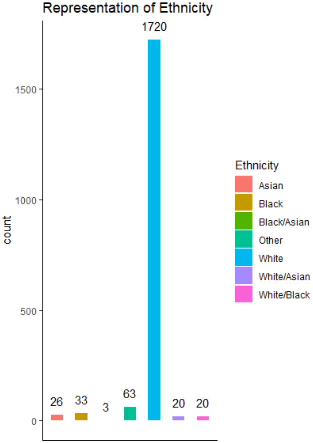
Comparing sample mean with Semeron/"overclaimer"mean for fun:



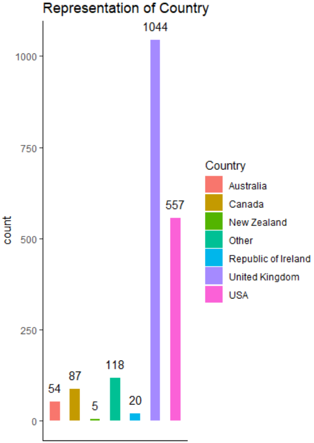


# Exploratory Data Analysis

This data-set skews White and UK / USA, so it would be difficult to draw meaningful inference based on the Nominal variables Ethnicity and Country.



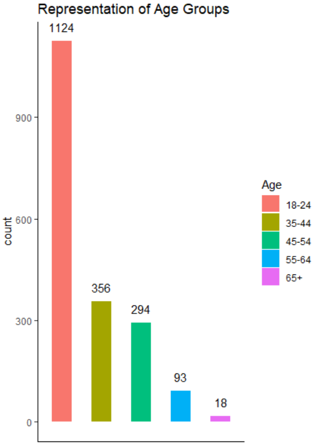
Representation of Ethnicity

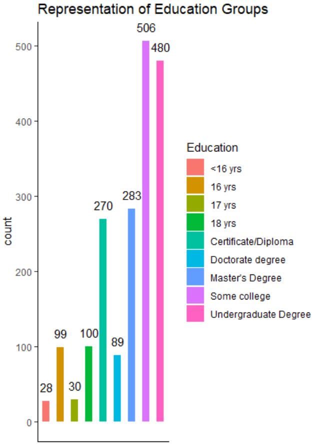


Representation of Country

# Exploratory Data Analysis

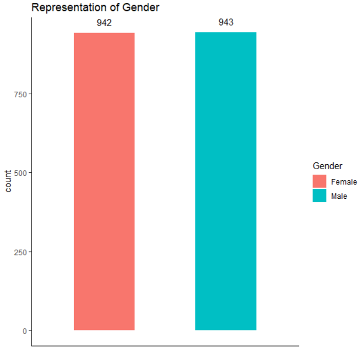
This data-set also skews toward higher education attainment and ages 18-24. Yet, these are categorical variables with the helpful characteristic of being Ordinal.

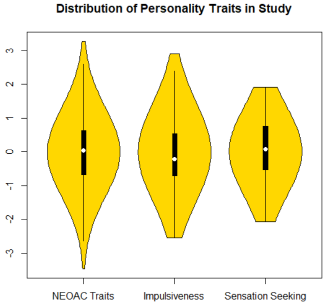




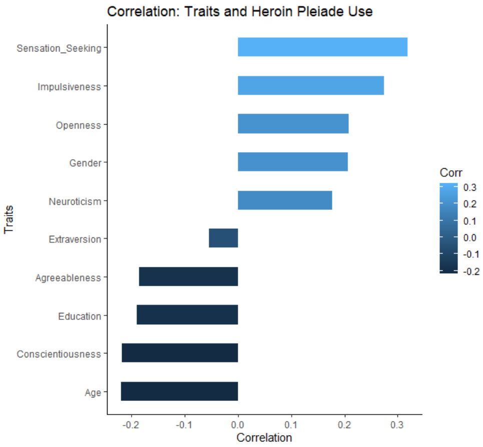
# Exploratory Data Analysis

This data-set has a fairly even representation of Gender (Male or Female) and the Personality Traits Measured.



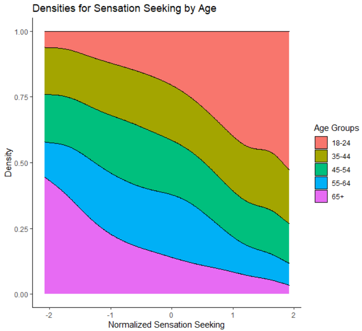


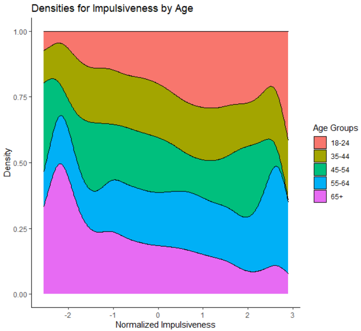
# Exploratory Data Analysis



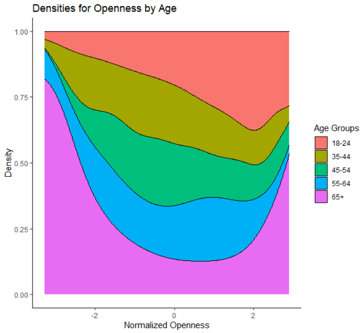
Sensation Seeking, Youth, Impulsiveness, and lower Conscientiousness correlate with Heroin Pleiade Use

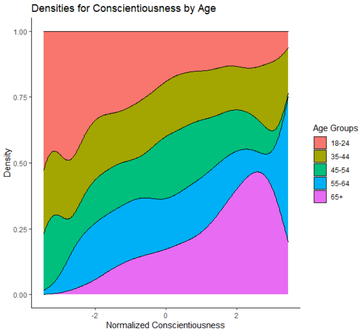
# Exploratory Data Analysis



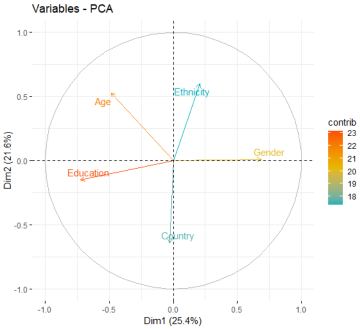


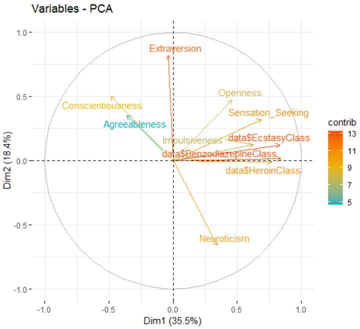
# Exploratory Data Analysis



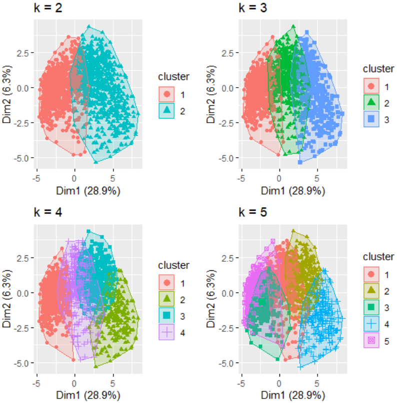


# Exploratory Data Analysis: PCA

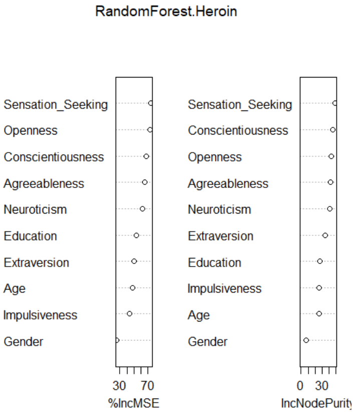


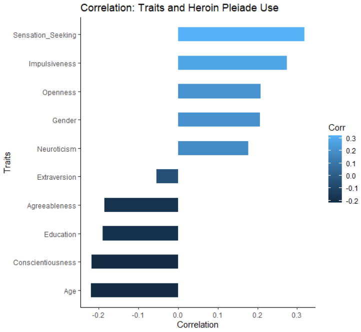


# Exploratory Data Analysis: K-Means

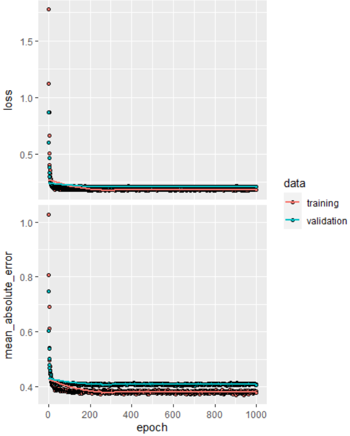


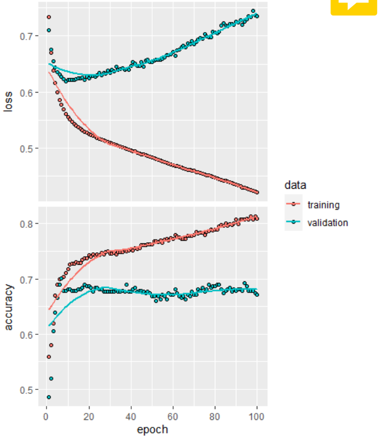
# Unsupervised Learning: Random Forest



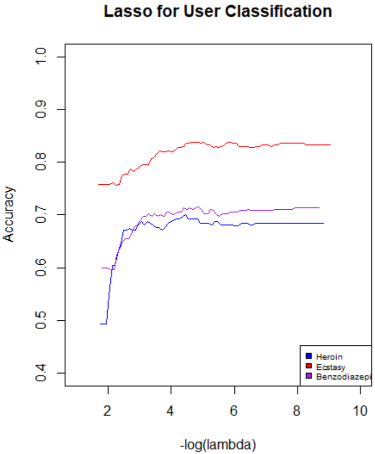


# Neural Networks: 1 layer vs 2 layers





# Supervised Learning: GLM



# Comparing MSE and MAE

| Measure: | MAE | MSE |
|---|---|---|
| Random Forest | 0.3876 | 0.2067 |
| NN 1 | 0.4070 | 0.2090 |
| NN 2 | 0.3804 | 0.2285 |
| GLM | 0.3896 | 0.1994 |

# Conclusion / What's Left

- Work through the Ecstasy and Benzodiazepine Pleides
- The literature suggests that Random Forest works especially well for this dataset. My work agrees.
- For its simplicity, the glm worked well.
- Overall, Sensation Seeking seemed to be the best indicator of whether or not someone is a drug user.

# Discussion

What would have made this research more interesting?

- Other variables: average income over the past year, five years; Partnership status (single, married, divorced)
- Does the individual use alone or in groups? Where (i.e. home, club, etc.)?
- Data that uses the same psychometric measures but is not dominated by a presence of White individuals from the UK and the United States; Different country and/or culture. Different rules and history.

# References

This is a list of major references used for the presentation

- 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], 2015
- E. Fehrman, V. Egan, A. N. Gorban, J. Levesley, E. M. Mirkes, A. K. Muhammad, "Personality Traits and Drug Consumption. A Story Told by Data."Springer, Cham, 2019. [Web Link]. ISBN 978-3-030-10441-2 [a book of Original Owners of Database]
- J. Gareth, D. Witten, T. Hastie, R. Tibshirani. "An Introduction to Statistical Learning : with Applications in R."New York :Springer, 2013.

# End

## Special Thanks to...

- Professor Behseta
- fellow classmates
- the statistical research community