Motivation

As I had no idea what to do for a project, I decided to orient on the dataset I find. For learning purposes, I searched dataset basing on the following criteria: pretty big, quite messy so that I could apply data processing techniques I learned, and train the most simple models (linear\logistic regression) in a short time and with good accuracy.

Raw NSDUH(the largest survey in the USA on drug usage and mental health of the population) data seemed like a good option for that. This survey-generated dataset has 2668 variables, each corresponding to a particular question in a survey with 2 to 20 values that mean different answers or the absence of an answer due to various reasons. For better understanding - questions in the survey are I must indicate though, that official dataset does not contain any personal data, because of respondent confidentiality.

Idea

My idea was to be able to train classifiers in a short time to identify a person as a certain drug user based on the answers on unrelated questions. For that I had to label the data using certain columns in the dataset, process the data and train classifiers on it.

Papers on the same dataset

Fairly speaking I could not find any works where people are trying to predict whether a person uses a particular type of drugs. Most of the works I found are trying either to describe data from the medical or social point of view or to work with specific columns from the dataset (20-30). I have not found any works where all dataset was used.
Anyway, here is the list of partially useful(in terms of this project) works, where I found some methods of working with the dataset:

https://www.researchgate.net/profile/Wilson_Compton/publication/7493381_Projecting_Drug_Use_Among_Aging_Baby_Boomers_in_2020/links/5b2014260f7e9b0e373ee4b7/Projecting-Drug-Use-Among-Aging-Baby-Boomers-in-2020.pdf

https://pdfs.semanticscholar.org/6cce/706bc99ed0344869c42b08b3f65c3b72a162.pdf

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3334449/

https://s3.amazonaws.com/academia.edu.documents/32185762/Sung_et_al_2005.pdf?AWSAccessKeyId=AKIAIWOWYYGZ2Y53UL3A&Expires=1547926587&Signature=zMcy%2BNt4xj11JWVOTACCCR2pzx0%3D&response-content-disposition=inline%3B%20filename%3DNonmedical_use_of_prescription_opioids_a.pdf

Work&code

Everything connected to processing and preparation is written in the notebooks, therefore I will not take these comments out of the context and dub them here, they belong with code.

Results

- NEW 13 binary columns indicating whether an observed person is using a certain type of drugs or not.
- 2 reduced datasets for training and testing of classifiers, each containing 13(for each drug-type) x 4(train_x, train_y, test_x, test_y) files.

  I can not load it to GitHub due to size limitation. It looks like that:

Why 2: in one dataset from train data all info about all drug usage was deleted, in the other only info about drug usage of the particular type(the one usage of which I am trying to predict) is deleted.

- Classifiers for each of the selected types of drugs with the following accuracies that were trained in few seconds:

for data with info about drug usage other than the predicted type:

1.0,	'hallucinogens',
1.0,	'inhalants',
0.9943057996485062,	'methamphetamine',
0.9597188049209139,	'pain relievers',
0.9796133567662566,	'tranqualizers',
0.9997188049209139,	'stimulants',
0.9951493848857645,	'sedatives',
0.9484710017574692,	'tobacco',
0.9958523725834798,	'alcohol',
0.9970474516695957,	'marijuana',
0.9824253075571178,	'cocaine',
0.9977504393673111,	'crack',
0.9945869947275923	'heroin

for data without (supposedly)any info on drug usage

1.0,	'hallucinogens',
1.0,	'inhalants',
0.9944463971880492,	'methamphetamine',
0.9617574692442882,	'pain relievers',
0.9833391915641476,	'tranqualizers',
0.9784182776801406,	'stimulants',
0.9953602811950791,	'sedatives',
0.9528998242530756,	'tobacco',
0.9960632688927944,	'alcohol',
0.9978910369068541,	'marijuana',
0.9995782073813708,	'cocaine',
0.9980316344463972,	'crack',
0.9958523725834798	'heroin

Conclusions

- Consequences of bad column namings
  Unfortunately, due to the initial messy naming of columns and different codes for different values, that were supposed to be the same, separation of data into different segments was not proper. I could not separate all data about every drug, because not all columns connected to a certain type of drug were used with an associated prefix. Also, I could not separate them manually, because there are 2668 variables and 13 types of drugs. Therefore many columns with info about usage of target drug were left behind lifting accuracy too high, as in case of inhalants and hallucinogens.

- other failures:
  - Such methods as KNN imputation of missing values were hard to apply due to the time constraint

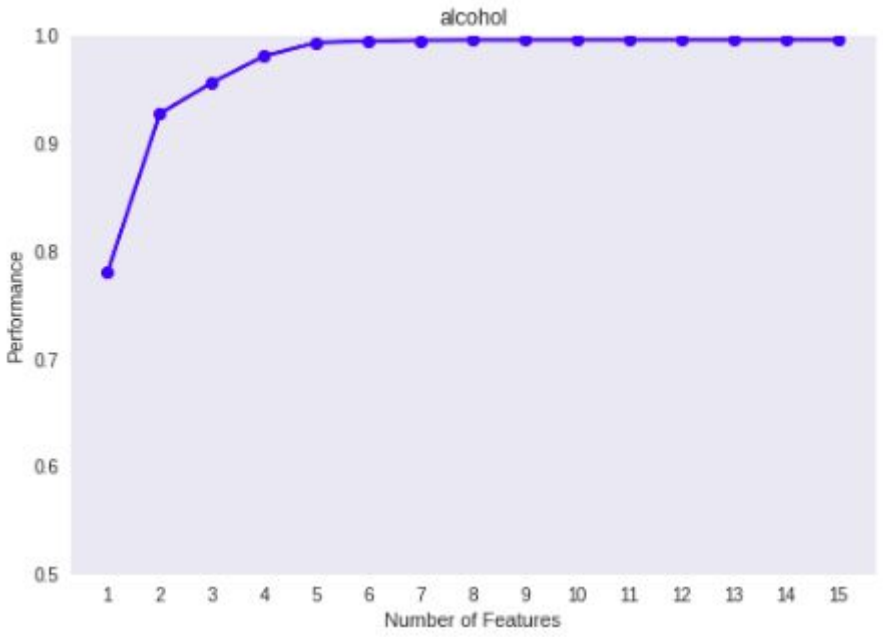
  - Factor Analysis can be harmful as well. There were some cases when a generated feature was so relevant that provided 99% accuracy. That feature was a zero column… Of course, most people do not use meth!

- possible improvements
  - to work properly with these columns and badly chosen indicator values I would have to use special software, such as SAS or Stata. Settings for such software are available in official page for downloading the dataset. Using such software with already predefined settings I probably would be able to separate data according to drug-type
  - To measure accuracy I should have used accuracy evaluation methods for underrepresented data. (most people do not use cocaine, so accuracy 95 may be actually really bad)
  - possible further analysis:
    it would be interesting to see which variables are influencing a particular type of drug usage. Due to chosen preprocessing techniques, I could not figure it out, as feature selection was performed after dataset was transformed by Factor analysis. But I could run decision tree feature selection as a more fast method on not yet transformed by any data reduction method data, and see which variable influence on prediction. I must note though, that it would only make sense if I separate data properly.

- Successes:

  - for such popular drugs as alcohol, tobacco, and marijuana the results

were really good. As that data was popular there was nearly no mess in column naming, I could separate data properly and classifiers gave good results on reduced data. It was interesting to see how feature selection was performed for the creation of reduced dataset for example for alcohol:



- Gained new experience of working with quite big, badly structured surveys
- Knowledge, that people work with this type of data using special software, such as SAS or Stata
- Such time-consuming methods as FactorAnalysis and Stepwise Forward Feature Selection worked in an adequate amount of time on this quite big dataset