Group 24: Jiawen Zhou (jz3416), Aman Anand (aa4821), Yining Ma (ym2928), Ziwei Yang (zy2533)

Background and Context

According to statistical data, 50% of people aged 12 and over have illicitly used drugs in their lifetime in the US and over 70,000 drug overdose deaths occur annually. Therefore, it is useful to discover what factors affect drug consumption behavior and predict people's abusing degree. We use the Drug Consumption Dataset from UCI to make predictions about drug usage of participants given their personality measurements and past drug use. This is a multiclass classification problem involving 1885 samples, 12 features and 19 classes.

Dataset

Links to the dataset:

Kaggle: https://www.kaggle.com/code/obeykhadija/drug-consumption-prediction

Original: https://archive.ics.uci.edu/ml/datasets/Drug+consumption+%28quantified%29#

Features: NEO-FFI-R neuroticism, NEO-FFI-R extraversion, NEO-FFI-R openness to experience, NEO-FFI-R agreeableness, NEO-FFI-R conscientiousness, BIS-11 (impulsivity), ImpSS (sensation seeking), level of education, age, gender, country of residence, ethnicity. Features were originally categorical and integer, and are quantified to be treated as real numeric variables. 12 in total.

Targets: Use of 18 legal and illegal drugs: alcohol, amphetamines, amyl nitrite, benzodiazepine, cannabis, chocolate, cocaine, caffeine, crack, ecstasy, heroin, ketamine, legal highs, LSD, methadone, mushrooms, nicotine and volatile substance abuse

Plus one fictitious drug (Semeron) which was introduced to identify over-claimers.

Target variables values: CL0: never used the drug, CL1: used it over a decade ago, CL2: used it in the last decade, CL3: used it in the last year, CL4: used it in the last month, CL5: used it in the last week, CL6: used it in the last day.

Data Size: 1885 * 32 (1 feature is id, 12 features, 18 targets, 1 artificial target)

Questions to be answered: Multi-class classification on each separate drug. Transform to binary classification of drug use. Evaluation of risk to be a drug consumer for each drug.

ML Techniques

Baseline: Decision Tree, KNN

- KNN is a simple method to start with a classification problem. Since our dataset is not too big, we will get a first glance of a baseline model performance
- Alternatively, decision tree works well if total purity is achieved but such results are expensive. A pruned tree would give a decent baseline model performance glance.

Candidates: Logistic Regression, Random Forest, Boosting, Support Vector Machine, Neural Network

- Logistic Regression with regularization is simple for understanding and still has robust performance in most circumstances.
- Random Forest does not require explicit train-test split as it uses subsets of columns and bootstrapped samples to train, and it performs better than baseline decision trees.
- Boosting takes weak learners to train sequentially to eventually increase performance. Gradient Boosting and its variants are popular in industry and proved to perform well.
- Support Vector Machines can be applied with both Primal and Dual to analyze their performance and trade-offs. Since the number of features is smaller than the training examples, we would expect Primal to be effective.
- Neural Networks are expected to have better performance due to their ability in exploiting non-linearity in the dataset.