University of Ottawa

School of Electrical Engineering and Computer Science

CSI5155 - Fall 2022

Assignment 1: Drug Consumption Dataset

TOTAL MARKS 70

The problem of evaluating an individual's risk of drug consumption and misuse is an important first step to address this serious problem globally. According to [1], a number of factors are correlated with initial drug use including psychological, social, individual, environmental, and economic factors, and these factors are likewise associated with a number of personality traits.

In [1], an online survey methodology was employed to collect data including Big Five personality traits (NEO-FFI-R), impulsivity (BIS-11), sensation seeking (ImpSS), and demographic information, resulting in the Drug Consumption dataset that contained information on the consumption of eighteen (18) central nervous system psychoactive drugs.

The following paper contains a detailed description of the data and the process of data quantification: https://link.springer.com/book/10.1007/978-3-030-10442-9 or https://arxiv.org/abs/1506.06297.

You are tasked to analyse this data through the construction of machine learning models. For this assignment, please use the Drug Consumption dataset from the UCI Machine Learning Repository. (The original link is located at https://archive.ics.uci.edu/ml/datasets/Drug+consumption+%28quantified%29 and the beta version may be found at this link: https://archive-beta.ics.uci.edu/ml/datasets/drug+consumption+quantified).

You are invited to consult the original paper [1] to obtain a deeper understanding of the domain and the descriptions of the features. You will note that the dataset was used by other researchers in multiple other studies, including:
https://ieeexplore.ieee.org/document/8234516
https://ieeexplore.ieee.org/document/8986131

Instruction:

1. This is an individual assignment. Submit your assignment using BrightSpace, before the due date.
2. For the implementation, you should either upload your code on BrightSpace or provide a link to a GitHub repository. Note that, if you choose to use GitHub, the date and time of last change to your repository should be **before** the assignment deadline.
3. Use Scikit-Learn to complete the assignment.

Topic: Supervised learning – Binary classification

The aim of this learning task is to identify the profiles of persons prone to consume drugs, when contrasted with those that do not, i.e., the class label is **Consume (Non-user, User).**

In terms of the dataset, this implies that we are converting the original multi-class learning problem into a binary learning problem. Specifically, for feature number 32 in the original data set, we combine classes C1 (Never Used) and C2 (Used over a Decade Ago) into "Non-User" while the data for the remaining four (4) classes are grouped together and labeled as "User".

Note this dataset is imbalanced, where most individuals surveyed never consumed drugs. For now, we are not considering any form of data rebalancing, prior to learning.

You are asked to follow the following steps.

Import the data into your machine learning environment and conduct feature engineering. Next, construct models using the following four (4) types of algorithms: a single decision tree (DT), a random forest (RF) learner, a support vector machine (SVM), and a k-nearest neighbor (k-NN) classifier. You should use the holdout method of evaluation, namely use 67% of the data for training, and 33% for testing.

Please complete the following steps.

A Programming [50]

a. Feature engineering: feature transformation and feature selection. [**10 marks**]
Feature transformation and feature selection are pre-processing steps followed before conducting machine learning. Refer to the reference paper for more details [1].
Feature transformations are useful to prepare the data for learning and include converting categorical data to numerical data, or normalising numerical data prior to training.
Feature selection techniques remove unnecessary features prior to training. To this end, you may use (any) one (1) feature selection algorithm as available in Scikit-Learn.

b. Model construction: Use the four (4) algorithms - DT, RF, SVM, k-NN - to construct four (4) models against the data. **[24 marks]**
c. Evaluation: Show the four (4) confusion matrices corresponding to the four (4) models and calculate the recalls and precisions. **[8 marks]**
d. Evaluation: Draw a figure to show the ROC Curves for the four (4) models. **[8 marks]**

B Reporting [20]

Submit a **250 to 300** words written summary, detailing the following.

e. Discuss the results you obtained and the lessons you learned when analysing this data. **[10 marks]**
f. Contrast the results you obtained during this assignment with those of the reference paper [1]. Be sure to discuss any differences in methodologies, and results, and to highlight similarities. **[10 marks]**

Main Reference

[1] 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. ISBN 978-3-030-10441-2