Data Mining 1 Assignment 3

Classification and Clustering using a drug consumption dataset

Dr Bernard Butler, SETU

24 November 2023

Issue Date: 24 November 2023

Submission Date: Sunday 17th December 2023 at 21:00

1.1 Aim

The purpose of this assessment is to give students the opportunity to show their knowledge and skills in Exploratory Data Analysis (EDA), Classification and Clustering.

1.2 Format of submissions

Students are asked to submit a Jupyter notebook named according to the following format: `[studentId]-[FirstnameLastname]-CA3.ipynb`. Students should add any supporting resources (other than the data files) needed to run the analysis, so that the notebook can be "run" by the lecturer. If a student uses any additional python packages, please indicate where they can be found too. There is no need to add the files to a zip archive before uploading.

As with CA1 and CA2, students are reminded to use the markdown cells to explain their reasoning, discuss outputs and report on the progress of their data investigation. Students are also advised to use both tabular and plotted output to present their work. Each notebook should read like a technical report, with all findings supported by evidence. Other recommendations from the CA2 brief regarding

- use of python comments,
- flow of analysis,
- making full use of markdown features
- collecting frequently used code in functions, etc.

also apply to CA3.

1.3 The drug consumption data set.

The data and notebook spec are provided in addition to this briefing.

The data is described further here.

The following additional comments may prove helpful to students:

1. The features might appear to be numerical, but are mainly derived from ordinal-valued psycho-social scores that have been passed through a PCA-based process.
2. Therefore, there of this preliminary treatment, should be no need to derive dummy variables from the features - they can be treated as either numerical or categorical, as needed, for modelling purposes.
3. The meaning of the features is presented in the link above.
4. The targets are ordered categories, represeenting the level of drug consumption, as explained in the link above.

1.4 Submission mode and Deadline

Assignment attempts should be uploaded via the relevant moodle submission page. The deadline for submission is **Sunday 17 December 2023 at 21:00**.

id	Criterion	Marks
1	Read the data, split into train and test datasets and perform EDA.	20
2	For a decision tree and another classifier, build the best model to predict the alcohol target.	20
3	Repeat Task 2 but with the caffeine, cocaine and heroine, explaining which is the easiest to predict.	25
4	Use hierarchical and positional clustering to find hidden groups in the study participants.	35

This assessment is worth 30% of your overall marks.

The key to success in this assessment is good data preparation and a structured approach to building models, keeping track of various options and paying close attention to multiple performance measures, most of which are subject to trade-offs, i.e., you cannot improve performance measure *A* indefinitely without harming at least one of the other performance measures.

1.5 Specification

For more details, please refer to the attached `CA3spec.ipynb` notebook.

Students are invited to be creative when answering the questions, i.e., marks will be assigned for "flair" and going beyond the treatment given in the notes and labs.

1.6 Other

If you use any resources (such as code, data, analysis) from others, *you must cite the source*. This is easy to do in practice: markdown has support for citing/referencing both online resources and books. Otherwise you are passing off someone else's work as your own - this is known as *plagiarism* and is not acceptable in SETU.

As with CA1 and CA2, you are advised to upload your work frequently, but please remember to *submit* the version you wish to be graded.