IN3062: Written Report

INTRODUCTION TO ARTIFICIAL INTELIGENCE

Sasha Sokolan, Munish Peenith , Gera Jahja

City university of london  Class of 2022

Table of Contents

[Introduction 1](#_Toc86703287)

[Dataset 2](#_Toc86703288)

[Methodology 2](#_Toc86703289)

[Plan 2](#_Toc86703290)

[Analysis 2](#_Toc86703291)

[Reflection 3](#_Toc86703292)

Checklist for Report: **(delete before submission)**

* Report should be 8 pages max ( not including table of contents and title page)
* Report in pdf format, single column, standard margins, font Arial 11, maximum 8 pages (see below for details), including all figures

Following the above analytical process, make sure that in your report you answer the following questions (where appropriate):

* What is your dataset, problem domain?

Music recommendation based on personality type

Data sets: <https://www.kaggle.com/akdagmelih/five-personality-clusters-k-means/notebook> and

Problem domain: Psychology of music preference:

Hypothesis: How does someone’s personality type affect their musical preference?

Assumptions:

Using the Big Five personality traits to mesure personality:

* Openness to experience (inventive/curious vs. consistent/cautious)
* Conscientiousness (efficient/organized vs. easy-going/careless)
* Extroversion (outgoing/energetic vs. solitary/reserved)
* Agreeableness (friendly/compassionate vs. challenging/detached)
* Neuroticism (sensitive/nervous vs. secure/confident)
* Is your model classification or regression?

-Classification since we are predicting a discrete class label (the correlation between musical preference and personality type)

Possible: A problem with two classes is often called a two-class or binary classification problem.

(refference: <https://machinelearningmastery.com/classification-versus-regression-in-machine-learning/>)

• Did you have any missing, corrupt or misleading data? If so, how did you cope it?

• Have you omitted some data? If so, why?

• Did you apply techniques to understand your dataset?

• What models did you use?

• How did you encode the input variables?

• What are the criteria for selecting model performance evaluation tools?

• What were your outputs?

• Did you have any problems or difficulties working with the dataset?

# Introduction

Description and motivation of the problem, description of the dataset including data types (e.g. discrete, continuous) (15% importance )

Define questions and analysis tasks (a brief overview of the domain, analytical questions that are being asked, a list of your objectives and the expected output(s) of your analysis)

# Dataset

* Define the domain and dataset(s) (you are free to choose the domain and the dataset that you want to investigate).
* Perform an initial investigation of the dataset and the characteristics of the data. Develop a plan as to how you might transform the data to make it useable.

# Methodology

Include: summary of the models used, with their pros and cons, a hypothesis statement, description of choice of training and evaluation methodology (20% Importance)

## Plan

* Develop a plan as to which artificial intelligence techniques you might use and what sorts of potential observations these can lead to, and how you will evaluate these.
* Split your dataset (train/validate/test, some datasets come pre-split). If you have a holdout test set then you most likely don’t want to use this until the near the end of your work.

## Analysis

Include:

* Results: description and presentation of the output. The code acts as an appendix to this section, and code quality (e.g. commenting) contributes. (35%)
* Evaluation: analysis and critical evaluation of results. (10%)
* Perform the analysis. Get the data ready for analysis, carry out your analysis/modelling as needed, validate your results and communicate observations, iterating through this process. Analytical operations can include data processing to an extent that is needed (not all datasets are messy) to prepare a useful and robust dataset to work within, and data derivation (such as feature engineering).
* You might establish a baseline result first, computing metrics on training and validation sets, analyse errors, work on succeeding iterations, and alternative models. (If initial metrics are amazing and there are no errors is the problem too easy?)
* Generally, be close to your data (visualise the dataset, collect summary statistics, look at

errors, analyse how different parameters affect performance, try out different model

variants).

Conclusions and referencing: lessons learned, references (using Harvard format) and future work. (10%)

# Reflection

• Reflection (10%)

The most important point to be addressed in the reflection is who did what. However, this is

also a chance to report further conclusions and discussion of your work from your individual

perspective.