BIRMINGHAM CITY
University

CMP6230 Draft Pipeline

DRAFT VERSION

Lewis Higgins - Student ID 22133848

Word count: XXXX

Contents

Candidate Data Sources

For the first stage of the pipeline, data ingestion, three data sources will be identified in order to find the one that would be most optimal for the production and deployment of a machine learning model to complete a supervised learning task.

1.1 Notes - DELETE BEFORE SUBMISSION

Lucidchart can generate ERDs from CSVs. For each, show the pandas head, column data types and ERD. Finding one with multiple CSVs can be good to make the ERD look more complex. **So far, all of your data is from Kaggle. Consider another source like data.gov.uk, especially considering that it can give you raw data for preprocessing.** On Kaggle, the "Provenance" section will have the source if the description doesn't. If neither have a source, it's probably fake data.

- Smoke detection
  - Real data
  - Lots to explain (how the alarms work etc)
  - Preprocessed, but you could still do more (remove timestamp etc)
  - Classification - Should the smoke/fire alarm sound?
- Employee data
  - Allegedly real data though it seems hard to believe.
  - Classification - Is the employee likely to find another job instead?
- Australian weather
  - Promising. Real data, but very large. Can do lots of preprocessing.
  - Classification - Will it rain tomorrow?
- Cardiovascular disease
  - Good data, enormous amount of it (laptop might not handle it), good source.
  - However, it's already been processed. Check if that's fine or not.
  - Classification - Do they have heart disease?
- Diabetes
  - In consideration for CMP6200, cannot use a dataset in both.
  - It's actually real data according to the Kaggle page, from "multiple healthcare providers" and the electronic health records (EHRs) they keep.
  - Preprocessed already, but duplicates exist in it.
  - Only 9 features, is that a bad thing?

  - Classification - Do they have diabetes?

All tasks of sheet 1, sheet 2 "should be analysed and you should write the plan for it", because sheet 2 refers moreso to the final report itself due both Dec 13 (draft) and Jan 10 (final). Because the amount of rows is not relevant in this particular module, you can use smaller ones like the 305 row Heart set. Week 8's lab is likely to be of vital importance to the final assessment of this module, as it goes through the use of MLFlow.

It's infeasible to do this on your laptop. Tuesday and Wednesday you'll need to smash this out *quickly.*

1.2 Candidate 1 - Smoke Detection Dataset

t