Retrospectively entering details of initial supervisor meeting:

Date: 5th March 2024

Time: 1215hrs

Platform: Zoom (Godfried's room)

Agenda: Review project summary

Notes of meeting:

We discussed how an ontology could be assessed without instances by using other published data. For example, what have other publications determined are good predictors of recidivism (even outside of machine learning) and are they included in the ontology? Following the discussion I drafted a revised project proposal:

The title will be “Addressing machine learning bias in criminal justice using ontology” (still subject to change as I develop the dissertation of course).

The research will focus on:

- Where is machine learning used in criminal justice? (high level, broad brush)
- Hone in on predicting recidivism (this will come from the research, but having started it I already know this is where I will focus)
- What are the key data points that predict recidivism?
- What data points have been used in machine learning algorithms to predict recidivism?
- Ethical issues/challenges with using machine learning to predict recidivism – many studies relate back to COMPAS in America
    - Are the predictions biased? (We already know that they are)
    - Why are they biased?
    - What has already been done to reduce/prevent bias?

From the research I will propose that using an ontology will reduce bias by exposing key attributes on which to train machine learning models. The ontology will also include protected attributes that will not be used to train the models, but will be used to validate the output i.e. ensure that the predicted profiles match actual profiles in terms of ethnicity, gender, age, and any other protected attributes that I discover/propose.

I will then use England and Wales as a case study to build an ontology, using the metadata from four different sources (magistrate’s court, crown court, probation service and prison service, available from the The Data First Project: An Introductory

User Guide (publishing.service.gov.uk) linked together into a single ontology. The ontology will separate the protected attributes for validation and will include a class for predicting recidivism that will contain the attributes that the research says are key for predicting recidivism. Some will come from the four sources already listed, and some might not be there so they will be added as “need to source this data elsewhere for an accurate model”.

The metadata from the recidivism prediction class can then be compared with the previous studies to demonstrate that the data would provide good predictions (since those studies will be used to generate the ontology in the first place they should have a good match rate). This will therefore show that if the ontology were to be loaded with real data, it would be able to predict recidivism to a similar standard as previous studies, but with the advantage of being able to instantly validate the profiles of the protected classes to ensure they match real-world profiles of the same classes. If they don’t within an agreed tolerance, the model should be re-trained with adjusted data (classes can be pulled in/out of the recidivism prediction class within the overall ontology, as long as the protected classes are always excluded) – sorry – I’m already starting to write a summary before I’ve even created the ontology!)

Finally – although the dissertation will focus on recidivism, there will be further work recommended to say that because the ontology will contain all of the data from the criminal justice system, not just what is needed to predict recidivism, it can be extended in many ways. For example, rather than just predicting recidivism, we could use the ontology to identify the attributes that impact recidivism that the criminal justice system can influence in order to reduce rather than predict recidivism. Looking at the interventions that probation make, or things that happen in the prison, that indicate high recidivism and try to change those attributes for individuals in the criminal justice system. The ontology could also be used to measure the impact of those interventions without even needing machine learning.