Addressing machine learning bias in criminal justice using ontology: a study of recidivism

Abstract (with assumptions that will be confirmed during the research):

Recidivism is when someone who has previously been convicted of a crime then reoffends. Reducing recidivism has obvious benefits to society in reducing crime, and the cost of crime. To that end machine learning has been used to predict recidivism, but for different purposes in different places. In part of the United States of America (USA), Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) is used to generate risk scores for offenders that are used in pretrial and sentencing to guide judges on bail and sentencing, and at parole to guide on who is safe to release. The impact of getting it wrong therefore has significant implications for the individual, who might be unfairly treated, and COMPAS has indeed been found to have racial bias (Angwin et al., 2022), although it doesn't use ethnicity as a feature. A further issue with COMPAS is the algorithms are proprietary and therefore not published, which increases mistrust in the outputs.

Ontology is proposed as a solution to biases caused both by direct use of inappropriate features such as ethnicity, and by indirect inference of those features from other features. The ontology allows protected features to be identified and excluded from any machine learning algorithms. Furthermore, it allows the remaining features to be classified according to their "safe use", with some features containing no inferential data about the protected features, while others might have some inference and should only be used with caution. Finally, the ontology allows the prediction results to be compared with the actual profiles for ethnicity, gender and age to ensure the results are uniformly accurate across the different groups.

An ontology was designed and built for England and Wales using metadata from Data First, a programme to combine data from Crown Courts, Magistrate Courts, Prison Service and Probation Service (Ministry of Justice, 2020a). Additional features were added from other studies to improve the recidivism predictions. The concept of the ontology was tested using an anonymised dataset from Georgia, USA (National Institute of Justice, N.D.). Exploratory Data Analysis (EDA) and statistical analysis were performed to identify the actual recidivism profiles for ethnicity, gender and age, and to identify features with high correlation to those protected features. The results were used to classify the features in the ontology to give guidance to future machine learning algorithms to minimise the risk of bias. [Method to be determined] was used to test the ontology, which was found to be [result to be summarised here].

The study has proven that an ontology is a practical way of mitigating the risk of bias in machine learning models by identifying dangerous features to be avoided, and then by validating that results conform to expected distributions. The method was tested using recidivism with a blended model using metadata from England and Wales and anonymised historic data from Georgia, USA, but would be applicable to any domain where potential bias can be predicted.

Future studies are recommended to populate the ontology with real data from England and Wales, available under strict conditions from Ministry of Justice (2020b), to train and test machine learning solutions and feed the results back into the ontology.

Research question:

- Can ontology be used to reduce bias when using machine learning to predict recidivism?

Aim:

- Machine learning is in use to predict recidivism but previous studies have indicated ethical issues such as racial bias. This study will demonstrate if biases can be identified and reduced with the use of ontologies by creating an ontology of criminal justice in England and Wales with features identified for safely predicting recidivism, and features only to be used with caution.

Objectives:

- Identify features that predict recidivism
- Identify features that potentially introduce unethical biases when predicting recidivism
- Assess if actual recidivism varies between characteristics such as ethnicity, gender and age
- Create an ontology of criminal justice using available metadata and illustrate how the ontology can be used to manage the features to reduce biases.

The literature review will examine:

- What are the key indicators to predict recidivism?
- How has machine learning been used to predict recidivism to date (real-world and academically)?
- What are the ethical challenges of using machine learning to predict recidivism?
    - Including performance measures, bias, interpretability of results.
- What measures have been taken to address the challenges?
    - Including feature engineering and if ontologies have been used to date.

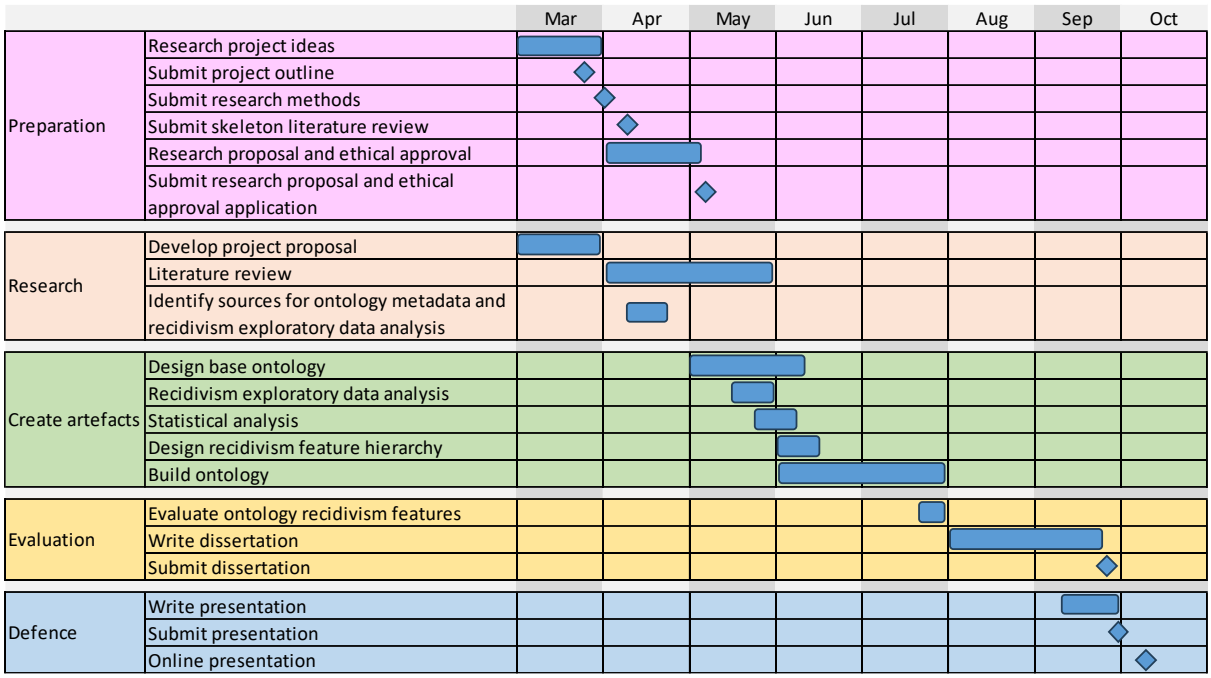
Research design:

- Methods to evaluate ontologies will be researched. A method that allows evaluation of an ontology using the metadata without instances will be selected to evaluate the ontology created.
- Statistical analysis will be used to determine if statistically significant differences exist between recidivism and ethnicity, gender and age.
- Primary research will not be required.

Artefacts to be created:

- Exploratory data analysis (EDA) of a non-UK data set (UK not available) to identify features that correlate with protected features such as ethnicity and gender.
- Statistical analysis of relevant feature relationships. For example:
    - Is there a statistically significant difference between actual recidivism rates for different ethnicities, genders or ages?
    - Are the correlations identified in the EDA statistically significant?
- Ontology of criminal justice using metadata from magistrate court, crown court, prison service and probation in England and Wales as a baseline, supplemented with additional features identified in the literature review, if applicable. The features will be organised to provide a method for segregating data using the indicative (non-UK) findings from the EDA and statistical analysis to identify and reduce bias when predicting recidivism.

Project plan:



Key risks:

- Time to get it all done alongside a full time job! Sticking to the project plan will be crucial.
- The research proposal and ethical approval isn't finalised until May, by which time the literature review should be practically complete. If the proposal is not approved it will be very difficult to complete revisions in time.
- There is a risk that the exploratory data analysis and statistical analysis don't uncover features that can be used to address the ethical challenges. In that case the resultant ontology and conclusions will be modified to visualising results in future to uncover biases rather than addressing previously identified biases.

References

Angwin, J., Larson, J., Mattu, S. & Kirchner, L. (2022) ‘Machine Bias’, in: Martin, K. (ed) *Ethics of Data and Analytics*. New York: Auerbach Publications. 254-264.

Ministry of Justice (2020a) Ministry of Justice: Data First. Available from: https://www.gov.uk/guidance/ministry-of-justice-data-first [Accessed 18 March 2024].

Ministry of Justice (2020a) MOJ: Data First, application form for secure access to data. Available from: https://www.gov.uk/government/publications/moj-data-first-application-form-for-secure-access-to-data [Accessed 18 March 2024].

National Institute of Justice (N.D.) Recidivism Forecasting Challenge. Available from: https://nij.ojp.gov/funding/recidivism-forecasting-challenge [Accessed 18 March 2024].