Gender Inequality - Project Plan

Tim Cross

July 16, 2025

1 Research Question

For this project I have decided to look at data relating to gender equality in different countries around the world, in data taken from the World Banks databank. Through analysis of this data I will aim to demonstrate the potential of AI and Machine Learning methods.

The dataset, Gender Statistics, is taken from the World Banks databank and includes 1162 features on key gender topics. Themes include demographics, education, health, labour force, and political participation. The data also includes more general data such as GDP, GNI, birth rate and population.

The target variable I decided to look at is 'Female share of employment in senior and middle management (%)', as I felt this was a good marker of general gender equality within a country. I will look at how other factors influence this target variable and what that can teach you about gender equality in a variety of countries, and the factors that contribute towards more or less equality.

This dataset is classified as public under the World Banks Access to Information Classification Policy, and therefore suitable to be used and shared in this project.

https://databank.worldbank.org/source/gender-statistics

2 Methodology

The data will initially be run through a cleaning pipeline, which will deal with missing data and adjust data as required, to leave it ready for further analysis and processing with ML models.

I will then carry out exploratory data analysis to get an understanding of the data, in particular looking at the target variable and how other key features interact with it. Through a variety of plots I will explore these relationships, which will help guide the next stage of ML modelling.

Through the exploratory data analysis I will have gained an understanding of the data, which will allow me to select relevant machine learning models to train on the data. Through the use of the models this will also allow an understand of which features show strong correlations with the target variable, and in turn may have a strong correlation with gender equality within a country. Following this will be an evaluation of these models results using various metrics, and coming to a conclusion on the best model for predicting the target variable.

Finally the project will be concluded with a Final Report, which will detail all the processes carried out, and what conclusions can be drawn from the results.

3 Timeline

With this project having a short duration, planning and time management will be of vital importance. Below is a timeline of the project. Major milestones will be: data prepared and cleaned by Monday 7th, EDA completed on Tuesday 8th, with ML modelling and evaluation completed by Friday 11th. This leaves Monday 14th and Tuesday 15th for completing the final report. This plan does have contingency built in: work can still be carried out on Wednesday 16th (before submission that day) if the project overruns, work can also be carried out on Sunday 13th if I feel the project is behind schedule at that point.

3/7	4/7	7/7	8/7	9/7	10/7	11/7	14/7	15/7
Issued	Project Plan	Clean Data	EDA	ML Model	ML Model	ML Model	Final Report	Final Report