Applied Data Science (COMSM0017) – Coursework

Sam James – Group 10 (Screen Time)

Reflection

Within the group my role was mainly that of data preparation. After the features that related to screen time had been selected, I performed analysis comparing the focused dataset to the whole dataset. For this analysis I calculated the proportions the different categories for each feature to give a comparative variable between the datasets of different sizes and indicating whether the focused dataset was representative of the whole dataset. I then researched into methods of feature selection as an alternative to dimensionality reduction, finding and performing recursive feature elimination. This was performed on the focused dataset and I experimented a little with changing the representation of each category to see this impact on the outcome of the feature selection.

I also worked on the visualisation of the results trying to have a consistent colour representation of depression and no depression diagnosis throughout the report. Also have the text in graphs matching that of the report to improve the look of the work. I spent time looking into the best file format to input graphics into the report to maintain image quality, finding that '.pdf' scaled images well. However, this was improved further by creating graphs that were the exact size for the report, 3.5 inches wide in the case of a single column in IEEE conference format.

I think that one of the biggest learning point over the project was online collaborative programming, which I hadn't performed up to this point in my degree with 6 people working on the same project. We utilised Google Collaboratory, although found recently they changed how it functions when multiple people are using it, which limited its effectiveness. It highlighted the importance of clearly laying out what had been done and making it clear the minimum that needed to be run to produce the output desired. Also keeping consistent terminology throughout the file and commenting the code. One problem I did find was trying to adjust some of the graphs into a consistent format was made difficult by the quantity of code in the file, a lot of the outputs of which were not put in the report. I have learned in future created sperate file or potentially a streamlined version would improve the usability of it, while keeping old version to ensure no work was lost.