CSC 499 Summary

Brianna MacDonald

At the end of this research, I have learned and accomplished the following:

- Learned about different definitions of fairness, as well as how to define “fair” from “unfair”.
- Learned about the different types of bias and how to mitigate them in terms of fairness or accuracy (and the trade-off that goes on between the two).
- Learned about the different algorithms from IBM’s AIF360 library for in-processing, post-processing, and pre-processing.
- Learned how to take into account different protected attributes in different datasets (notably the COMPAS and Adult datasets from the AIF360 library).
- Learned how to rank the “privileged value” within a dataset depending on the protected attribute of that dataset
- Learned about the differences between discriminatioin and unwanted bias when it comes to sexual orientation, gender, race, and age.
- Learned how the different classifiers/algorithms in AIF360 can...
    - Mitigate bias in training data (pre-processing)
    - Mitigate bias in classifiers (in-processing)
    - Mitigate bias in predictions (post-processing)
- Learned and read from several different papers focusing on prejudice removal in classifiers, as well as the different processes used.
- Determined the best type of algorithm to use based on accuracy and F1 score for the COMPAS dataset.
- Computed and plotted confusion matrices and ROC-AUC curves depending on the type of model as input.
- Observed differences in protected attributes for the COMPAS dataset for the following metrics:
    - Statistical parity difference
    - Equal opportunity difference
    - Average absolute odds difference
    - Disparate impact
    - Theil index
- Observed the change in eta when it came to evaluating the above metrics.
- **Note: there is not a lot of change in metrics when dealing with ETA in small changes (for example: the metrics for ETA = 1.0 and ETA = 25.0 are almost identical, however there is significant change when ETA= 250.0).**