**AI/ML Bias & Fairness Mitigation**

The method that I researched regarding fairness and bias in Artificial Intelligence (AI) / Machine Learning (ML) models was the ‘pre-processing mitigation’ method to remove bias and promote fairness in the model’s results. This method is broad, including both simple anonymization techniques and complex removal of ethically questionable features using correlation of other variables, as well as many other use cases.

The simple anonymization uses could include removal of identifying information such as name, identification numbers, gender, weight, ethnicity, or any other demographic data point that could be used to create discriminatory decision boundaries in the model’s architecture. The tried-and-true method that I have used with this technique is whether I would like to be identified by the features in my model and running these ethical concerns by other team members. If members of my team / our legal team have issues with using certain features in the model, such as demographic data, we can simply remove these features before training the model.

However, sometimes there are more complex issues that need to be accounted for beyond just simple removal of identifying features. One instance is the correlation of other features to ethically questionable features that need to be analyzed for proper pre-processing mitigation. An example of this instance would be a hypothetical feature ‘male\_patterned\_baldness’ with results corresponding to a ‘gender’ feature (which we would have already removed). The corresponding results could look like ‘Yes’ / ‘No’ for all males, and an ‘N/A’ for all females. In this manner, the model could still learn gender-based decision boundaries even though we removed the ‘gender’ feature from our dataset, because the data still contains correlated features to the ‘gender’ feature [in this case, all males would have a result in the ‘male\_patterned\_baldness’ feature while all females would have nulls]. This scenario highlights the importance of nuanced use of the ‘pre-processing mitigation’ technique for specific use cases and the necessity of knowing all the features and their correlations to other features in the datasets.

For application of this method in my project, I plan on testing the model with and without demographic features to compare model results. My dataset contains features correlating to nurse attrition, and comparing results of model runs both with and without the demographic features (gender, marital\_status, age) could show the impact these features have on the model. If the model has much worse performance upon the removal of these features, I may need to consider keeping these features in the model. Furthermore, if I decided to keep these features in the model, I would need to analyze the results of the model to ensure that predictions are not creating groups based off these demographic features. The reason for not wanting these groups is because of the ethically murky implementation of the model’s results into the nursing profession. For example, if my model predicted that primarily women over the age of 55 were going to attrit, the business might rethink their staffing strategies to not hire these candidates going forward, leading to potential discriminatory hiring practices against this group. As a data scientist, it is important to think through and analyze all predictions because a lot of users of the predictions will not do their own analysis / not have the means to do their own analysis.

However, if the model did not have performance loss when removing the demographically identifying features, I would still need to reapply these ‘dropped’ features to the predicted results to check for discriminatory groups being created. If these groups are found, I would do a correlation of non-demographic features to these removed demographic features, to see if I am training my model to still learn demographically discriminatory decision boundaries. Lastly, if the data has ethically questionable patterns that cannot be accounted for without greatly reducing model performance, I would escalate these issues to subject matter experts on the data / problem and more experienced data scientists. Hopefully, as a group, we would be able to balance model performance with business needs to ensure the model performs in accordance to business goals while also mitigating bias and ensuring fairness of the model’s predictions.