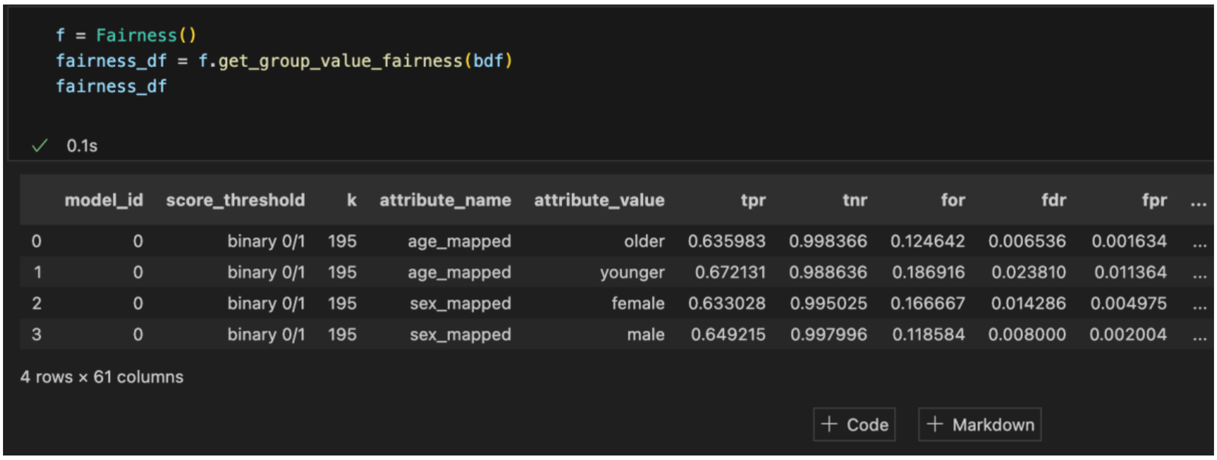
AIKI210 Assignment 2

Task 1



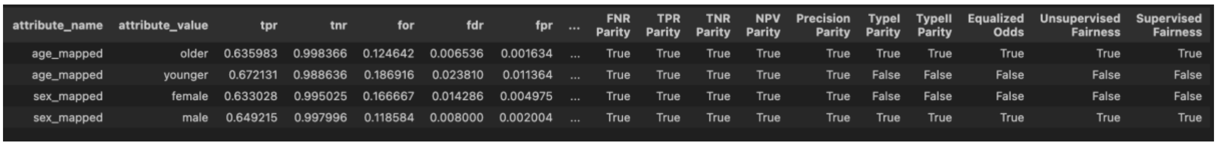
```
f = Fairness()
fairness_df = f.get_group_value_fairness(bdf)
fairness_df
```

✓ 0.1s

	model_id	score_threshold	k	attribute_name	attribute_value	tpr	tnr	for	fdr	fpr	...
0	0	binary 0/1	195	age_mapped	older	0.635983	0.998366	0.124642	0.006536	0.001634	...
1	0	binary 0/1	195	age_mapped	younger	0.672131	0.988636	0.186916	0.023810	0.011364	...
2	0	binary 0/1	195	sex_mapped	female	0.633028	0.995025	0.166667	0.014286	0.004975	...
3	0	binary 0/1	195	sex_mapped	male	0.649215	0.997996	0.118584	0.008000	0.002004	...

4 rows × 61 columns

+ Code
+ Markdown



attribute_name	attribute_value	tpr	tnr	for	fdr	fpr	...	FNR Parity	TPR Parity	TNR Parity	NPV Parity	Precision Parity	TypeI Parity	TypeII Parity	Equalized Odds	Unsupervised Fairness	Supervised Fairness
age_mapped	older	0.635983	0.998366	0.124642	0.006536	0.001634	...	True	True	True	True	True	True	True	True	True	True
age_mapped	younger	0.672131	0.988636	0.186916	0.023810	0.011364	...	True	True	True	True	True	False	False	False	False	False
sex_mapped	female	0.633028	0.995025	0.166667	0.014286	0.004975	...	True	True	True	True	True	False	False	False	False	False
sex_mapped	male	0.649215	0.997996	0.118584	0.008000	0.002004	...	True	True	True	True	True	True	True	True	True	True

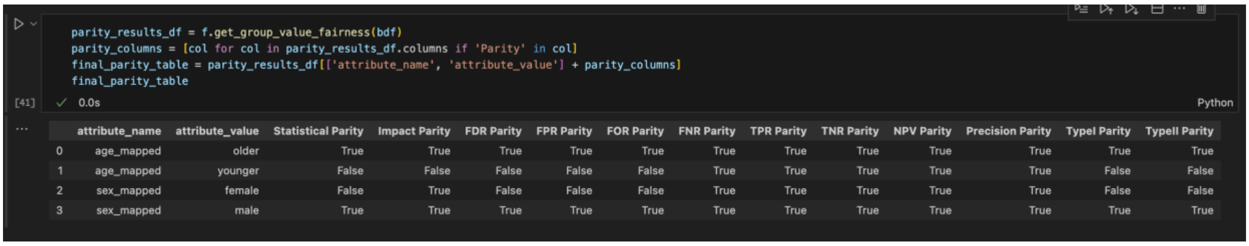
Disparities represent the differences in treatment and outcomes between the groups. In our case we have performance metrics such as false positive and negative, and we investigate whether different groups such as young/old and male/female get treated differently across these metrics. I.e., we find in our dataset that young people have a higher TPR (true positive rate) of 0.6721, compared to older people with a TPR of 0.6360, suggesting that the model is better at predicting "true" performance for younger people than for older individuals. Further, we also observe women (0.6330) having a slightly lower TPR than males (0.6492).

True negative rates (TNR) are relatively similar for all groups, with the biggest disparity being between older (0.9984) and younger people (0.9886)

Parities on the other hand, indicate fairness or equality of outcomes across groups. In other words, it is the opposite of disparities. In the notebook, the parity cells are either True or False, showing whether the performance metrics for the groups are close or equal.

For instance, we see that the TPR and TNR are True across all groups (male, female, old, young), suggesting parity of the performance metrics in the model. Equalized odds on the other hand, which combines false positives and false negatives, is False for the groups of "younger" and "female" – indicating a lack of parity.

Task 2



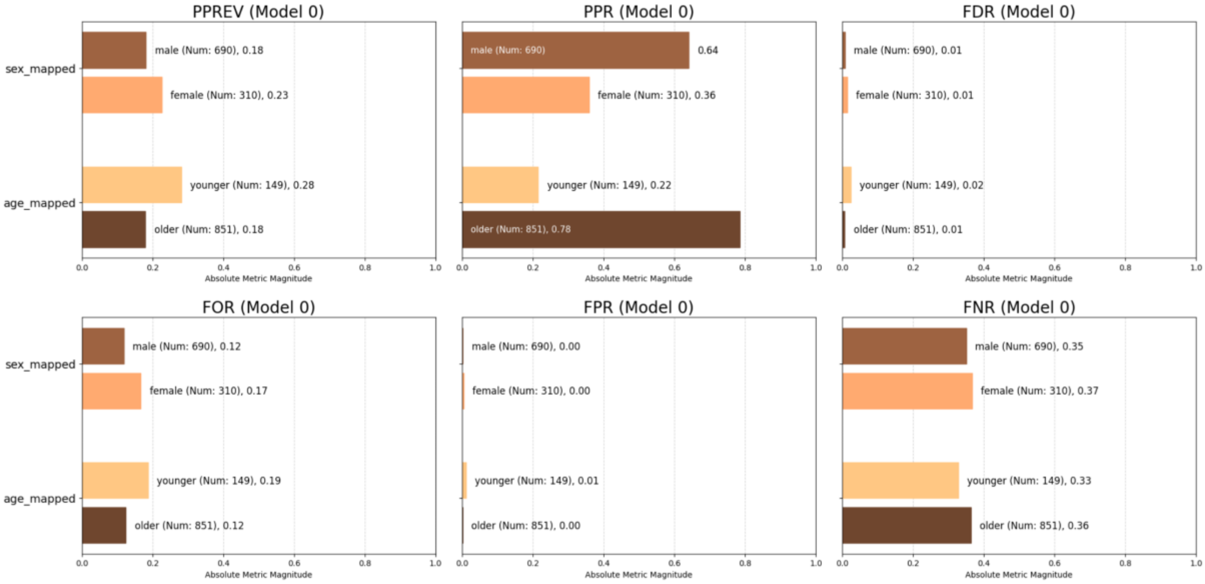
```
parity_results_df = f.get_group_value_fairness(bdf)
parity_columns = [col for col in parity_results_df.columns if 'Parity' in col]
final_parity_table = parity_results_df[['attribute_name', 'attribute_value'] + parity_columns]
final_parity_table
```

[41] ✓ 0.0s Python

	attribute_name	attribute_value	Statistical Parity	Impact Parity	FDR Parity	FPR Parity	FOR Parity	FNR Parity	TPR Parity	TNR Parity	NPV Parity	Precision Parity	TypeI Parity	TypeII Parity
0	age_mapped	older	True	True	True	True	True	True	True	True	True	True	True	True
1	age_mapped	younger	False	False	False	False	False	True	True	True	True	True	False	False
2	sex_mapped	female	False	True	False	False	False	True	True	True	True	True	False	False
3	sex_mapped	male	True	True	True	True	True	True	True	True	True	True	True	True

As requested, here is a table with the results from the fairness analysis which only includes the parities

Task 3



Starting with PPREV (Predicted Prevelance), we see how often credit loans are predicted for each category, where males and older individuals have a higher predicted prevelance.

Moving on to PPR (Positive Predictive Rate), we observe that that the model predicts credit approval for males much more often for males (0.64), compared to women (0.36). Further, older individuals are also predicted to get approved credit (0.78) significantly more often than younger people (0.22).

FDR, or the False Discovery Rate, measures the proportion of positive predictions that are actually false positives. Here we see that the FDR is very low across all groups.

Thirdly, we have FOR (False Omission Rate), which measures the proportion of negative predictions that are actually false negatives. Here we see that the model more likely predicts females and younger people to an incorrectly negative outcome (bank credit), which suggests a fairness issue for these groups.

False Positive Rates are extremely low for all groups.

False Negative Rates are pretty consistent for all groups, with a slightly higher rate for females and older people.

Task 4

Precision Parity – The proportion of positive credit predictions that are correct are measured by $\frac{True\ positives}{True\ Positives+False\ Positives}$.

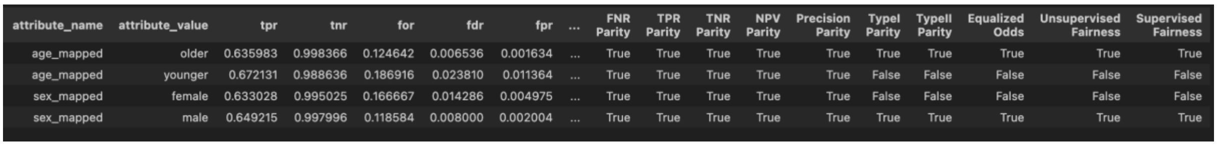
Aequitas measures this by comparing all groups such as male, female younger etc. and checking if the precision rates are similar. If parity is found, it means that the model is equally effective at making correct positive predictions across all groups.

Statistical Parity – In Aequitas, statistical parity is assessed by checking if the proportion of positive predictions is equal across the different demographic groups. For example, if a model predicts positive outcomes for younger and older peple at a similar rate, the model achieves statistical parity. In other words, one divides the amount of positive predictions by the total number of predictions for each group.

Equalized Odds – Aequitas checks Equalized Odds by comparing both the TPR and FPR across groups (e.g., by gender or age). If both rates are similar across groups, the model is considered to satisfy the criterion of equalized odds. This ensures that the model does not favor one group over another in terms of both making correct positive predictions and avoiding incorrect positive predictions. $TPR = \frac{TP}{TP+FN}$, $FPR = \frac{FP}{FP+TN}$.

Task 5

We can use (1) Precision Parity, (2) Statistical Parity and (3) Equalized Odds to assess the fairness of the model.



attribute_name	attribute_value	tpr	tnr	for	fdr	fpr	...	FNR Parity	TPR Parity	TNR Parity	NPV Parity	Precision Parity	TypeI Parity	TypeII Parity	Equalized Odds	Unsupervised Fairness	Supervised Fairness
age_mapped	older	0.635983	0.998366	0.124642	0.006536	0.001634	...	True	True	True	True	True	True	True	True	True	True
age_mapped	younger	0.672131	0.988636	0.186916	0.023810	0.011364	...	True	True	True	True	True	False	False	False	False	False
sex_mapped	female	0.633028	0.995025	0.166667	0.014286	0.004975	...	True	True	True	True	True	False	False	False	False	False
sex_mapped	male	0.649215	0.997996	0.118584	0.008000	0.002004	...	True	True	True	True	True	True	True	True	True	True

If we pull up our table findings from Task 1 we see that Precision Parity is satisfied for all groups, despite small variations in True Positive Rates.

Moving on to Statistical Parity, we investigate Type 1 Parity (False Positive) to see if there are variances. Indeed, we see that for younger individuals and females, parity is not maintained. Therefore, Statistical Parity is not fully satisfied.

Lastly, we look at Equalized Odds. Similarly to Statistical Parity, the Equalized Odds are not maintained for younger people and women. This can also be seen when comparing TPR and FPR for males and females as well as for young vs older people.

In conclusion, we can argue that the model is not fully fair as the Statistical Parity and Equalized Odds both are biased against females and younger people.

Task 6

Although there certainly are many limitations of observational criteria of fairness, I would like to highlight challenges related to (1) distinguishing groups, (2) exogenous explanations, and (3) threshold sensitivity.

Firstly, it is important to understand that making arbitrary groups such as "younger or older individuals" does not capture the complexity of human nature. Variables such as age, is a continuum, rather than a binary variable as it was in our case. The overarching problem with distinguishing groups is that it is static, while the real world is highly dynamic. However, it should also be mentioned that no effort would be able to fully capture human nature, as far as todays knowledge goes.

Secondly, observational fairness metrics don't account for broader societal inequalities. Metrics like equalized odds and statistical parity can only show whether the model treats groups the same, but they don't address underlying societal conditions that lead to disparities in the first place. I.e., young people have an economics disadvantage independent of whether Equalized Odds are achieved or not.

Lastly, metrics such as TPR and FPR can vary based on the decision thresholds, which could make fairness assessment inconsistent across different threshold values. Small changes in the decision threshold can lead to different fairness outcomes.

Comment: *Since I'm a first-year student with no coding experience, I have been using ChatGPT as a supplementary tool to make the notebook run, as well as create some parts of code where my knowledge was limited. I assume this should be okay as the focus of this course is not programming, but rather the interpretation and findings from the code. It should also be mentioned that the course description for AIKI210 states that no prior knowledge of code is needed, which further underscores the course's focus on interpretation rather than programming,*