1 Problem 1

1.1 Part A

The stakeholders I am considering for this exercise are the members of the judicial system, the defendants, and people concerned with community safety.

Accuracy: The judicial system would benefit from a model which optimizes accuracy. If COMPAS can accurately predict recidivism, the judicial system can feel safe relying on their produced scores to predict defendants' likeliness to recidivate. I don't believe this would be reasonable to optimize because it does nothing to ensure fairness across protected groups.

Positive Predictive Value: Positive predictive value is the probability that defendants labeled as high risk are actually high risk. Individuals concerned with community safety would benefit from a model which optimizes positive predictive value. It would result in a high probability that defendants categorized as high risk are actually high risk individuals. They can then be given the needed care and treatment to help them rehabilitate, improving overall safety.

False Positive Rate: Defendants would benefit from a model that optimizes false positive rate. An optimized false positive rate would lead to a lower chance of defendants being incorrectly classified as high risk for recidivism. This would lower the chance that defendants are incorrectly given a harsher sentence than deserved.

False Negative Rate: The judicial system would also prioritize optimizing FNR because it would help ensure high risk criminals aren't mistakenly being classified as low-risk. A high FNR could result in high risk criminals receiving reduced sentences, reflecting poorly on judges and prosecutors.

Statistical Parity: Statistical parity is satisfied when the proportion of individuals classified as high-risk is the same for both groups. Defendants theoretically could benefit from a model which ensures statistical parity because it

indicates that the model is treating individuals from different protected groups equally. I wouldn't recommend optimizing statistical parity in this case because the different groups likely have unequal base rates.

1.2 Part B

Pre-existing bias may arise in this scenario if the data Prophecy is trained on is biased due to existing biases in society. For example, women historically have been underrepresented in the workforce, especially in tech. The data Prophecy is trained may reflect this bias and Prophecy would then habitually categorize women as less qualified candidates than other candidates with similar qualifications but traditionally male names. This may harm candidates of protected classes, such as women, black people, or people with disabilities. To mitigate this type of bias, you could improve the dataset Prophecy is being trained on and ensure there are equal instances of candidates of all races, genders, etc.

Technical bias may emerge if a flaw in the model causes it to habitually categorize female candidates as less promising than male candidates with similar attributes. This kind of bias could potentially harm any candidate as the model flaw could result in discrimination against any group of candidates. To mitigate this type of bias, you could improve your model by ensuring it meets certain measures of fairness, such as equalizing the FNR rate across protected groups.

Emergent bias may occur if there is a major shift in opinion on what constitutes a promising candidate for data scientist roles but Prophecy isn't able to incorporate these shifts into their model. Both candidates and companies using Prophecy could be harmed by this type of bias. Candidates who are considered competitive under the current expectations for data scientists could still be categorized as less competitive due to the outdated model. The companies using Prophecy wouldn't receive top candidates based on their current expectations

for data scientists. To mitigate emergent bias, Prophecy could be periodically updated and re-deployed to reflect the changing times.

2 Problem 2

2.1 Part A

To start, I will define false positive rate, false negative rate, and prevalence using definitions from the provided Wikipedia page. I will then rearrange each equation to achieve preferable forms.

False Positive Rate = $FPR = \frac{FP}{N}$

$FP = N - TN$

$N - TN = FPR * N$

$TN = N(1 - FPR)$

False Negative Rate = $FNR = \frac{FN}{P}$

$FN = P - TP$

$P - TP = FNR * P$

$TP = P(1 - FNR)$

p = Prevalence = $\frac{P}{P+N}$

$P = p(P + N)$

$N = (1 - p)(P + N)$

Now, I will define accuracy using the definition from the provided Wikipedia page. I will then substitute in some of the equations derived above to achieve a definition of accuracy which only consists of the terms prevalence, FNR, and FPR.

Accuracy = $\frac{TP+TN}{P+N}$

Accuracy = $\frac{P(1-FNR)+N(1-FPR)}{P+N}$

Accuracy = $\frac{p(P+N)(1-FNR)+(1-p)(P+N)(1-FPR)}{P+N}$

Accuracy $= p(1 - FNR) + (1 - p)(1 - FPR)$

We can now compare the accuracy of Group A to the accuracy of Group B:

$p_A(1-FNR_A)+(1-p_A)(1-FPR_A) = p_B(1-FNR_B)+(1-p_B)(1-FPR_B)$

Clearly, Group A and Group B's accuracies cannot be equal if $FNR_A = FNR_B$, $FPR_A = FPR_B$, but $p_A \neq p_B$. Therefore, we can conclude that a classifier cannot simultaneously achieve equal accuracy, equal false positive rates, and equal false negative rates if Group A and Group B have different base rates.

2.2 Part B

We are given the following definitions:

$ACC_A = ACC_B + \delta$

$FPR_A = FPR_B + \delta$

$FNR_A = FNR_B + \delta$

$p_A = p_B + \epsilon_p$

We will now expand the definition of accuracy derived in Part A.

Accuracy $= p(1 - FNR) + (1 - p)(1 - FPR)$

Accuracy $= 1 - pFNR + pFPR - FPR$

Using this expanded definition of accuracy, we will now compare accuracies for Group A and Group B. We will perform substitution using the provided definitions to simplify the equation to the desired form:

$ACC_A = ACC_B + \delta$

$1-p_AFNR_A+p_AFPR_A-FPR_A = 1-p_BFNR_B+p_BFPR_B-FPR_B+\delta$

$1 - (p_B + \epsilon_p) * (FNR_B + \delta) + (p_B + \epsilon_p) * (FPR_B + \delta) - (FPR_B + \delta) = 1 - p_BFNR_B + p_BFPR_B - FPR_B + \delta$

$1-p_BFNR_B-p_B\delta-\epsilon_pFNR_B-\epsilon_p\delta+p_BFPR_B+p_B\delta+\epsilon_pFPR_B+\epsilon_p\delta-FPR_B-\delta = 1-p_BFNR_B+p_BFPR_B-FPR_B+\delta$

$-\epsilon_p FNR_B + \epsilon_p FPR_B - \delta = \delta$

$\epsilon_p FPR_B = \epsilon_p FNR_B + 2\delta$

$FNR_B = FPR_B - \frac{2\delta}{\epsilon_p}$

This function is derived from the equation $ACC_A = ACC_B + \delta$. It illustrates the relationship between the False Negative Rate and False Positive Rate of Group B when the base rates between Group A and Group B are not exactly equal and their accuracies are approximately balanced. It also shows that in order to maintain fairness between Group A and Group B when they have different base rates, Group B's FPR is constrained by the value of Group B's FNR. Specifically, Group B's FPR can me no more than $\frac{2\delta}{\epsilon_p}$ less than Group B's FNR. Note that δ is some constant margin of error and ϵ_p represents the difference between Group A and Group B's base rates.

3 Problem 3

The data science application discussed in this lecture is the use of data science in hiring and recruitment. Specifically, it talks about AI's role in screening job applicants and judging their fit. The purpose of this data science application is to speed up the recruitment process by eliminating the need for a human to look at and evaluate every application. Additionally, the creators and proponents of this technology argue that it will eliminate human bias by replacing it with a model.

One stakeholder who could benefit from this application are the companies who would save time and money by eliminating the need for a human to evaluate every application. Specifically, HR representatives will save a lot of time. Any candidate could be adversely affected by this technology. The model could incorrectly categorize an applicant as undesirable, costing them the chance to move forward in the interview process. However, the lecture specifies that members of

protected groups (such as disabled individuals or gender and racial minorities) are more likely to be affected.

This lecture describes several harms that may occur as a result of the use of AI technology in recruitment and hiring. For one, it illustrates how they can discriminate against gender and racial minorities. This is because models are representative of the data fed to them. Since our society has long seen under-representation of gender and racial minorities in desirable job fields, the model may be fed biased data and discriminate against them. The lecture also discussed the harm disabled individuals. Specifically, it discussed how assumptions are often made when creating these AI tools which makes them inaccessible for disabled individuals. For example, individuals who are hard of hearing may struggle in a video interview due to a lack of captions, but the AI tools cannot account for that. This isn't necessarily due to a data-related or technical reason. Rather, it is the result of a poorly designed and exclusive AI tool. The lecture also delved into ways in which the AI tools can discriminate against disabled individuals. One example they discuss is how some tools which judge the pleasantness of an applicant may misinterpret the facial effects of a stroke as signs of displeasure. This most likely occurred because the model wasn't trained to recognize pleasantness on individuals who have suffered strokes. Therefore, biased data was most likely the cause of this discrimination.

4 Problem 4

4.1 Part A

Overall, the baseline random forest model didn't perform terribly. However, there is certainly room for improvement in both performance and fairness. Looking at the evaluation metrics, we see the accuracy of the model is only ˜60%.

The low precision (58%) indicates that this model struggles to make accurate positive predictions. The high FNR (˜50%) indicates this model also often incorrectly predicts positive instances to be negative.

The fairness metrics show that the performance of this model differs based on gender. The ideal values for FNR Difference, FPR Difference, and SRD are 0%. For this model, these values hover around 10%. The ideal values for DPR and EOR are 1. For this model, they're around 70-75%.

In many cases, these evaluation and fairness metrics may be acceptable for the model at hand. However, this model theoretically will be used to predict whether or not a patient needs additional medical care. Any inaccuracy or bias can have fatal effects. Therefore, I believe the current performance and fairness metrics are not yet acceptable.

4.2 Part B

In this section, I tuned the hyperparameters of the baseline random forest model from part A to maximize accuracy. Specifically, I tuned hyperparameters 'max_depth', or the maximum depth of the tree, and 'n_estimators', the number of the trees in the forest. The possible maximum depths were 5, 10, 15, and 20. For most of the 10 models trained in this section, the maximum depth which produced the most accurate model was 10. I believe this indicates the prediction being made by the model was somewhat complex. Therefore, a max depth of 10 was able to produce an accurate tree without causing overfitting. The possible number of trees were 50, 100, 150, and 200. For all 10 models trained in this section, the number of trees in the forest that resulted in the most accurate model was 200. This indicates that more decision trees led to more accurate predictions. I hypothesize that this was a complex prediction that required many decision trees to accurately execute. I wonder if increasing

the possible number of trees in the forest could have improved performance. However, I am wary of increasing the number of decision trees too drastically, as it could lead to overfitting.

The performance of the models were judged on 5 evaluation metrics. Below, plots illustrate the performance of the baseline untuned model versus the tuned model for those metrics.

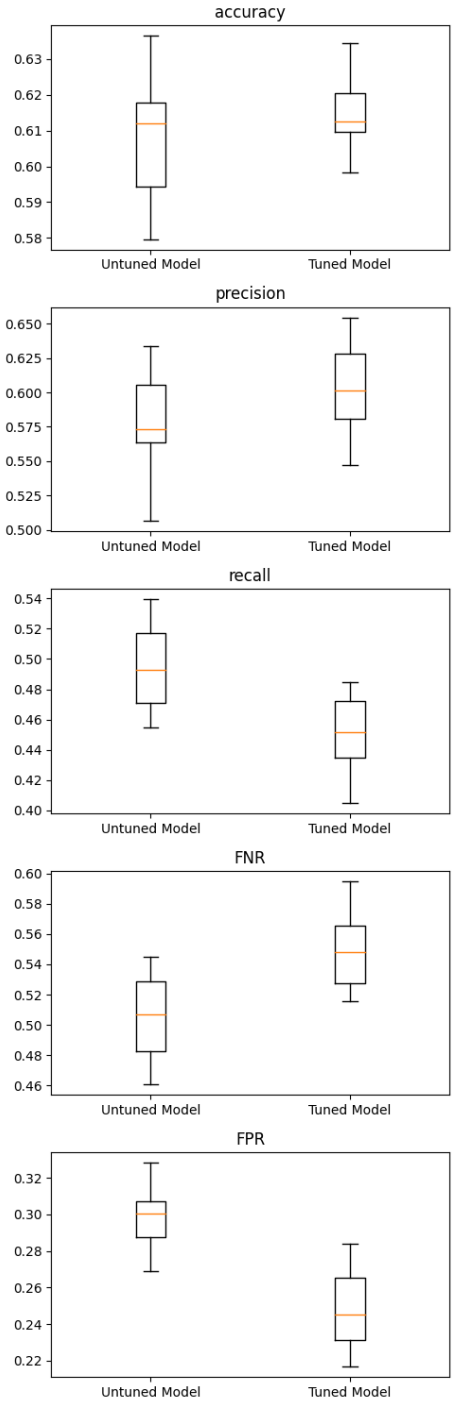

Figure 1: Untuned VS. Tuned Model's Performance on Evaluation Metrics

Tuning the model resulted in about the same median accuracy as the untuned model. However, the range of accuracies decreased and the minimum accuracy was much higher than before, indicating tuning the model still improved the accuracy. The median precision also increased, and the median FPR decreased. This indicates tuning the model resulted in an overall improved performance.

However, the median recall decreased and the median FNR increased, indicating the model's ability to correctly classify positive instances decreased. This indicates there was a tradeoff when tuning the model. In search of improved accuracy, the model became more conservative when classifying positive instances.

The models' performance were also judged on 5 fairness metrics. The performance of the untuned vs tuned models for the 5 fairness metrics are plotted below.

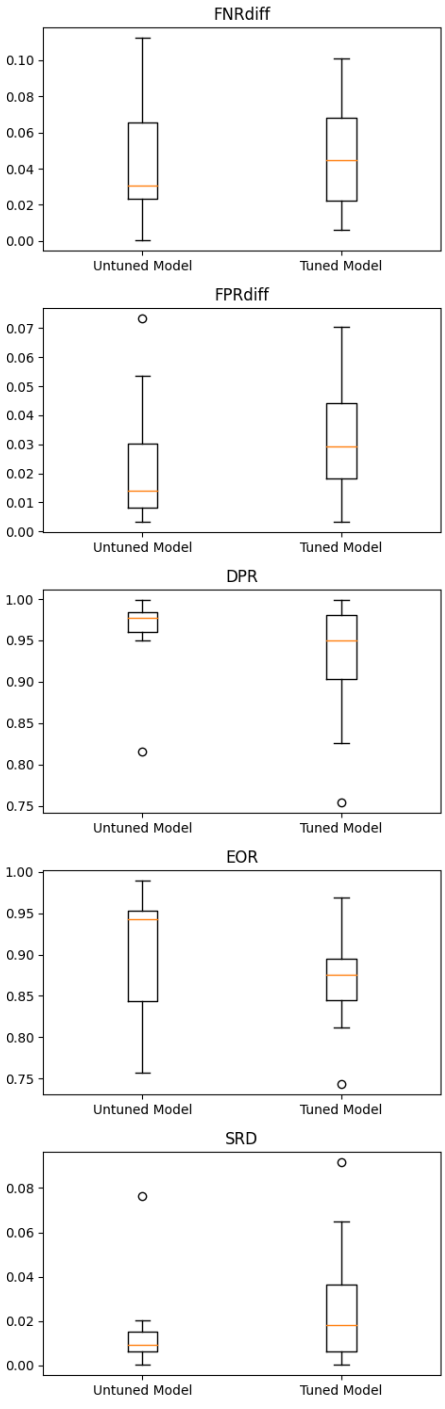

Figure 2: Untuned VS. Tuned Model's Performance on Fairness Metrics

In general, looking at these plots, it appears the fairness of the untuned model wasn't terrible. One metric that definitely warranted improvement was EOR. While the median was fairly high (˜95%), the minimum was ˜75%. There also was a large difference between the maximum and minimum for EOR, indicating the untuned model's performance for this metric was a bit unreliable. After tuning, the median decreased to ˜87%. However, the range of possible EOR's also decreased, indicating the tuned model's performance on this metric is more reliable. However, the decrease in the median value for EOR is still troublesome.

The tuned model has shifted the median of all the other fairness metrics away from their ideal values as well. This indicates that the tuned model likely sacrificed fairness for improved accuracy. This goes back to the concept of tradeoffs. It's very difficult to have both a perfectly accurate and perfectly fair model. In pursuit of one, the other often suffers, which we observed in this experiment.

4.3 Part C

In this section, I applied the Adversarial Fairness Classifier, an in-processing intervention, to help balance optimizing both fairness and accuracy. I struggled with this in the previous section. The Adversarial Fairness Clarifier takes a parameter 'alpha' which dictates the tradeoff between fairness and accuracy. A smaller alpha will prioritize accuracy and a larger alpha will prioritize fairness. Below, plots illustrate the performance of the model for four different values of alpha for 5 evaluation metrics.

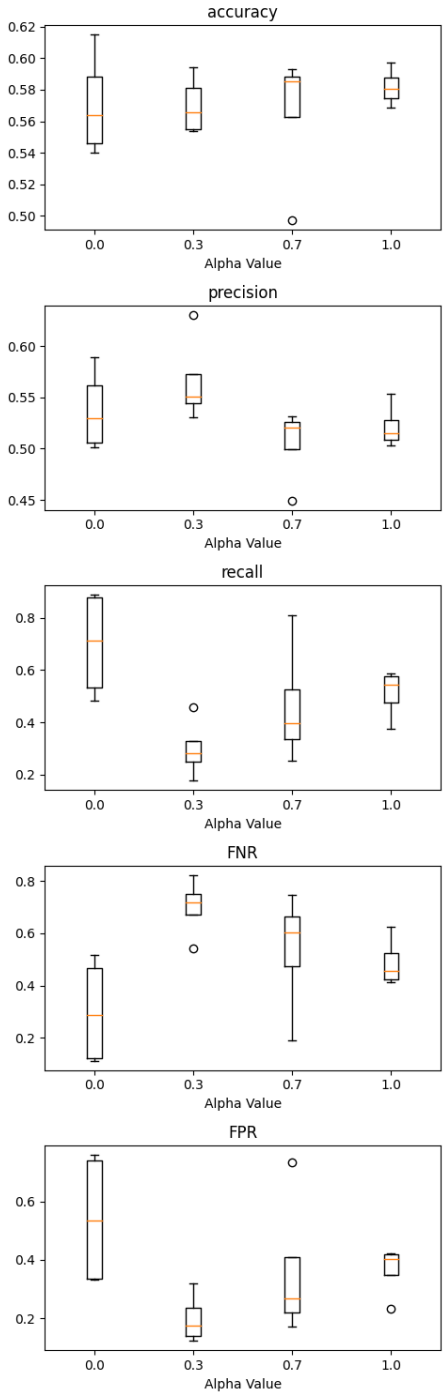

Figure 3: Model Performance on Evaluation Metrics for Different Alphas

The median accuracy stayed relatively stable for the different values of alpha. It's also very close to the median accuracy of the tuned model from the previous section. In this section, variability decreased as accuracy increased, showing the model became more predictable with higher alphas. This may appear confusing since a higher alpha should lead to prioritizing fairness over accuracy. However, this could indicate that strengthening the fairness constraints actually led to more stable predictions. In the variability also decreased significantly after tuning.

Precision's median also stayed relatively stable across different alpha variables. However, it is about 5% lower than the tuned model's precision. Variability also decreased with the increased model, strengthening the hypothesis that strengthened fairness constraints lead to stable predictions.

Recall's median was over 20% higher than the tuned model's median when alpha was 0. This indicates that prioritizing accuracy when in-processing improves the model's ability to identify actual positive cases. The median experiences a significant decrease when alpha increases to .3, but gradually climbs back up as alpha continues to increase.

The model's FNR has a similar pattern to recall. It increases sharply when alpha increases to .3, and then slowly decreases as alpha continues growing. It is possible that the first introduction of fairness constraints causes a sharp decrease in performance. However, as alpha continues to grow, the model may learn to better balance performance and fairness. The median for FNR when alpha is 0 is 30%, about 25% lower than the previous tuned model's median. However, it is important to note that there is far more variability in this model's FNR. One possible explanation is that this model is more sensitive to how the data is split than the previous tuned model.

FPR decreases sharply when alpha increases to .3, and then slowly increases

as alpha continues growing. The lowest median value of FPR recorded is ˜20%, which is pretty close to the median FPR value for the previous tuned model. The variability for this model is a bit higher but not by too much. Therefore, I believe this model has a similar FPR to the previous tuned model when alpha=.3.

The model was also judged on 5 fairness metrics. They are plotted below for the different values of alpha.

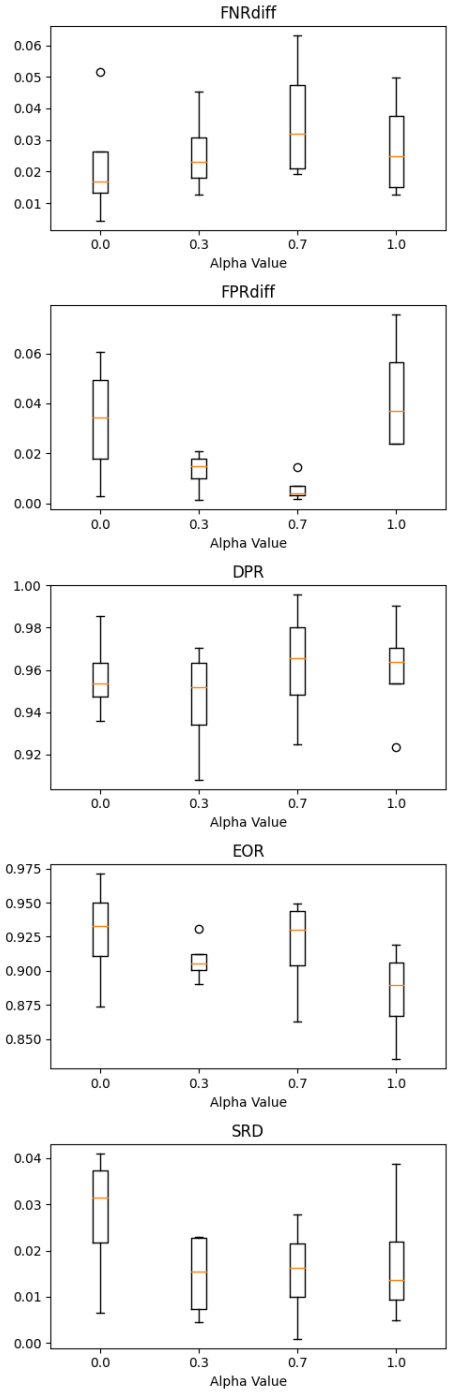

Figure 4: Model Performance on Evaluation Metrics for Different Alphas

The value of FNR and FPR difference, EOR and SDR for this model is fairly consistent across all alphas. This could indicate that the model doesn't doesn't have much of the bias tracked by these metrics, so strengthening fairness constraints doesn't affect the values much. They are also similar to the previous tuned model's respective metric values. Therefore, the in-processed model has a similar FNR and FPR difference, EOR and SDR to the tuned model for all alphas.

DPR also stays pretty consistent across all alphas. The median is fairly consistent to the medion of the tuned model. However, the variability is far less, indicating this model has more consistent outcomes across protected groups regardless of data splits.

4.4 Part D

In this section, I used Threshold Optimizer, a post-processing algorithm. I chose equalized odds as the fairness constraint under which Threshold Optimizer would perform. By optimizing equalized odds, I ensure the model is able to correctly identify individuals in need of extra care across all protected groups, ensuring all individuals who need help can receive it. Additionally, the model won't incorrectly flag individuals for extra care, ensuring precious resources aren't being divulged to those who don't actually need them.

Below, plots illustrate how an untuned model performed on 5 evaluation metrics versus how the post processed model performed.

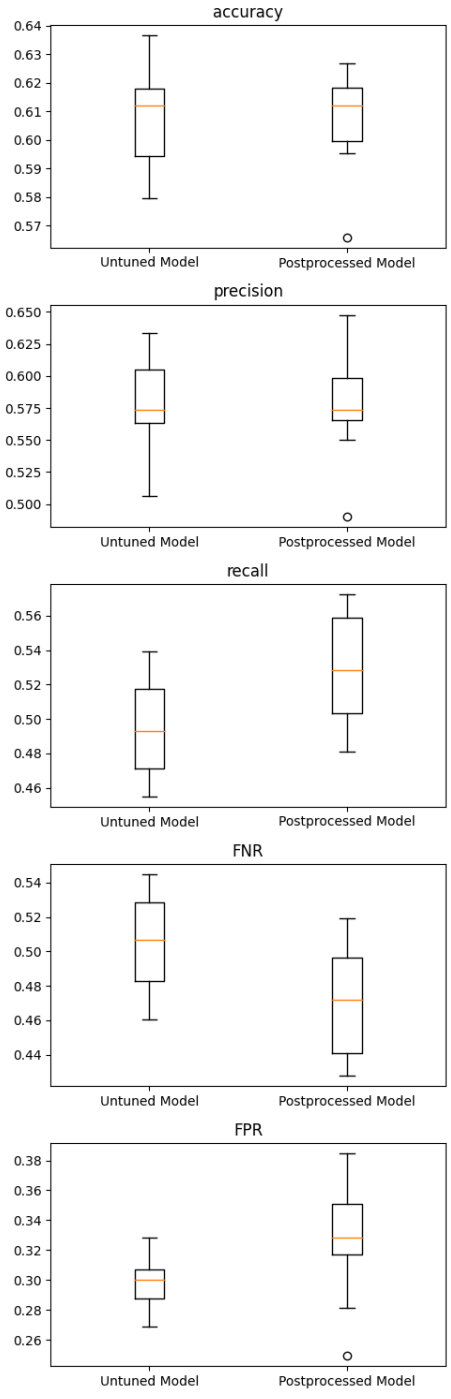

Figure 5: Untuned VS. Post-processed Model's Performance on Evaluation Metrics

The post-processed model and the tuned model have fairly similar accuracy and precision. However, there are some other differences. Recall is ˜8% higher, indicating this model is better at identifying actual positive cases than the tuned model. FNR is ˜10% lower, indicating the post-processed model can more accurately predict actual positives. However, the post-processed model's FPR is 10% higher and more varied. Therefore, when choosing between the post-processed model and the previous tuned model, it's crucial to consider the trade-off. Is it more important that all individuals in need of extra care are accurately identified? Is it more important that healthy people aren't being incorrectly given extra care, taking resources away from those in need? The post-processed model works if you believe the first statement, the tuned model is better if you believe the latter statement.

The post-processed model and the in-processed model have similar accuracies and precisions. The post-processed model's recall, FNR, and FPR is less than the in-processed model's most optimal value for these respective metrics. They are, however, similar to the in-processed model's metrics when alpha=1.0, or when fairness is being optimized over accuracy. This indicates that the post-processed model, which optimizes a fairness constraint, performs similarly to the in-processed model when it is also optimizing fairness constraints.

The performance of the post-processed model was also evaluated based on 5 fairness metrics. Below, plots illustrate how an untuned model performed on the fairness metrics metrics versus how the post processed model performed.

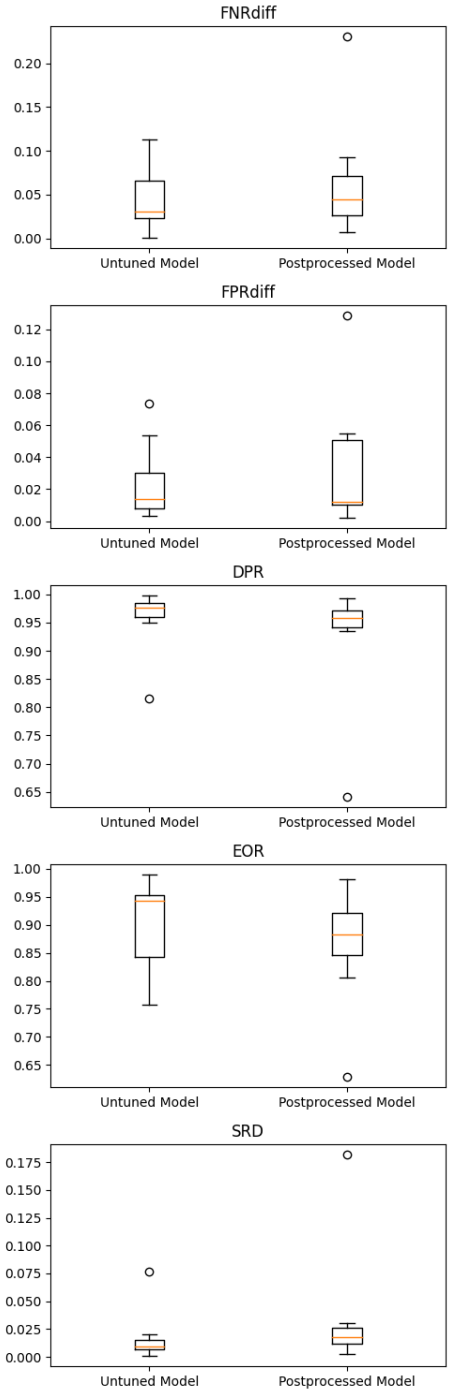

Figure 6: Untuned VS. Post-processed Model's Performance on Fairness Metrics

The post-processed and tuned model have similar medians and variability for FNR difference, FPR difference, and EOR. For DPR and SRD, the medians are similar but the post-processed model is much less varied. This indicates that the post-processed model can more accurately balance the proportion and probability of positive predictions across groups than the tuned model.

The post-processed and in-processed model have similar medians and variabilities for FNR difference, FPRdifference, and SRD. For DPR, the post-processed model has a similar variability to the in-processed model when alpha=0 or 1. The post-processed model's variability for EOR is also a bit higher than the in-processed model's variability for EOR. This indicates the post-processed model is less consistent than the in-processed model at achieving equalized odds when dealing with different data splits.

In general, I observed that an increase of fairness typically leads to a decrease in accuracy, and vice versa. For example, hyperparameter tuning lead to a more accurate model, but measurements of fairness decreased as a result. However, I also observed instances in which the pursuit of fairness actually improved some measures of accuracy. Specifically, the in-processed model showed me that increasing fairness can actually decrease variability in our models' measurements of performance. I also observed how hyperparameter tuning can