ASSESSING BIAS IN A PREDICTIVE MODEL FOR HEART DISEASE DIAGNOSIS

INDIVIDUAL EXPERIMENTAL WORK

A report written by

VICTOR ODOH

Date: 3rd May 2023

CONTENTS

Abstract

This project aims to assess bias in a predictive model for heart disease diagnosis, and to evaluate the effectiveness of applying fairness criteria to the model to address any observed biases. The Gradient Boosting Model was built using Scikit-learn library, and fairness criteria were applied to investigate racial bias in the model. The study found that while adjusting the model to ensure equal opportunity and demographic parity improved the true positive rate for both the white and non-white groups, it led to a significant decrease in overall accuracy and an increase in false positives, which could lead to unnecessary medical procedures and harm to patients. The study highlights the complexity of ensuring fairness in predictive models and emphasizes the need for careful consideration of specific contexts and trade-offs involved. Ultimately, using a dataset with a relatively even distribution of instances among protected groups can further ensure that a model is generally fair in its predictions. This project contributes to efforts to reduce health disparities and improve the quality of care for all patients, regardless of race or ethnicity.

1.0 Introduction

Artificial intelligence (AI) has become increasingly prevalent in healthcare, with machine learning models being used to improve diagnosis, treatment selection, and health system efficiency. However, there is growing concern about the potential for these models to perpetuate or exacerbate health disparities by encoding biases present in historical data. In particular, certain populations that have experienced human and structural biases in the past, also known as protected groups, are vulnerable to harm by incorrect predictions or withholding of resources (Rajkomar et al., 2018).

In the context of disease diagnosis, biases in AI models have the potential to perpetuate inequities in access to healthcare and exacerbate health outcomes disparities. Heart disease is one of the leading causes of death worldwide, and accurate and timely diagnosis is critical for effective treatment and management of the disease. However, there is evidence that heart disease diagnosis and treatment may be subject to racial and ethnic disparities, with some groups receiving lower quality care and experiencing worse outcomes (Kumar et al., 2016).

This project aims to assess bias in a predictive Gradient Boosting Model for heart disease diagnosis, and to evaluate the effectiveness of applying fairness criteria to the model to address any observed biases. Specifically, the objectives of this project include developing and evaluating the model using appropriate data, analyzing the extent to which the model may be biased toward a racial group as represented in the data, applying certain fairness criteria to the model to handle such bias, and analyzing the consequent effects of adjusting the model to satisfy the associated fairness constraints.

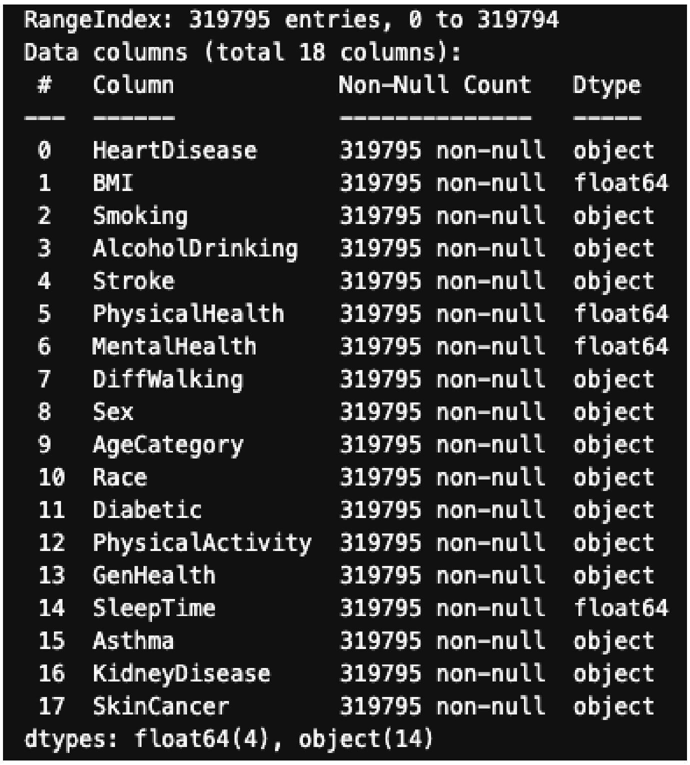
By identifying and mitigating bias in the predictive model for heart disease diagnosis, this project aims to contribute to efforts to reduce health disparities and improve the quality of care for all patients, regardless of race or ethnicity.

…………………MORE DETAILS IN THE COMPLETE REPORT………………

3.0 Methodology

3.1 Data Collection and Pre-processing.

The dataset is the "Personal Key Indicators of Heart Disease" data. It was retrieved from kaggle.com (https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease). It initially consisted of 18 attributes (4 numerical and 14 categorical) and about 320k instances before pre-processing.



```
RangeIndex: 319795 entries, 0 to 319794
Data columns (total 18 columns):
 #   Column            Non-Null Count   Dtype  
---  ------            --------------   -----  
 0   HeartDisease      319795 non-null  object 
 1   BMI               319795 non-null  float64
 2   Smoking           319795 non-null  object 
 3   AlcoholDrinking   319795 non-null  object 
 4   Stroke            319795 non-null  object 
 5   PhysicalHealth    319795 non-null  float64
 6   MentalHealth      319795 non-null  float64
 7   DiffWalking       319795 non-null  object 
 8   Sex               319795 non-null  object 
 9   AgeCategory       319795 non-null  object 
 10  Race              319795 non-null  object 
 11  Diabetic          319795 non-null  object 
 12  PhysicalActivity  319795 non-null  object 
 13  GenHealth         319795 non-null  object 
 14  SleepTime         319795 non-null  float64
 15  Asthma            319795 non-null  object 
 16  KidneyDisease     319795 non-null  object 
 17  SkinCancer        319795 non-null  object 
dtypes: float64(4), object(14)
```

Figure 1.0 Initial Features of the Dataset

For the purpose of this project, *HeartDisease* was chosen as the target variable to build a predictive model for heart disease…

…………………MORE DETAILS IN THE COMPLETE REPORT……………….

3.3 Investigation of Bias

…………………MORE DETAILS IN THE COMPLETE REPORT……………….

The model was adjusted using some calculated decision thresholds, to satisfy some fairness criteria constraints.

3.3.1 Applying the 'Equal Accuracy' Criterion

The "equal accuracy" fairness criterium ensures that predictive models perform equally well for each protected group, regardless of their demographic characteristics. In this regard, it means that the model should achieve similar levels of accuracy for the white and non-white groups, without creating bias or discrimination towards any one of them. There was only a fractional difference between the predictive accuracies for the white and Non-white Race groups, hence there may be no need to adjust the model to ensure Equal Accuracy.

However, to illustrate the application of the Equal Accuracy Criterion, I further attempted to equalize both accuracies by adjusting the model through some calculated decision thresholds for each race group.

3.3.2 Applying the 'Demographic Parity' Criterion

This requires that the proportion of positive predictions should be equal for different groups, regardless of their demographic characteristics. In other words, it ensures that the rate at which the model diagnoses a patient with heart disease (whether the diagnosis is right or wrong) is the same either for a white or non-white patient so that it is not systematically favoring or disfavoring any of the concerned groups.

There are several metrics that can be used to evaluate whether demographic parity has been satisfied. In this project, the metrics/approaches used are Disparate Impact Ratio (DIR), and Predictive Accuracy. DIR is the ratio of the proportion of positive predictions for one group to the proportion of positive predictions for the other group. The proportion of positive predictions for each demographic group was computed and compared for any significant difference. To satisfy demographic parity, we want the DIR to be close to 1, which indicates that the model treats both groups equally. The model was adjusted using a calculated decision. For the Accuracy, we want it to be high, but to also ensure that it is similar across both groups (Kamiran et al.,2012).

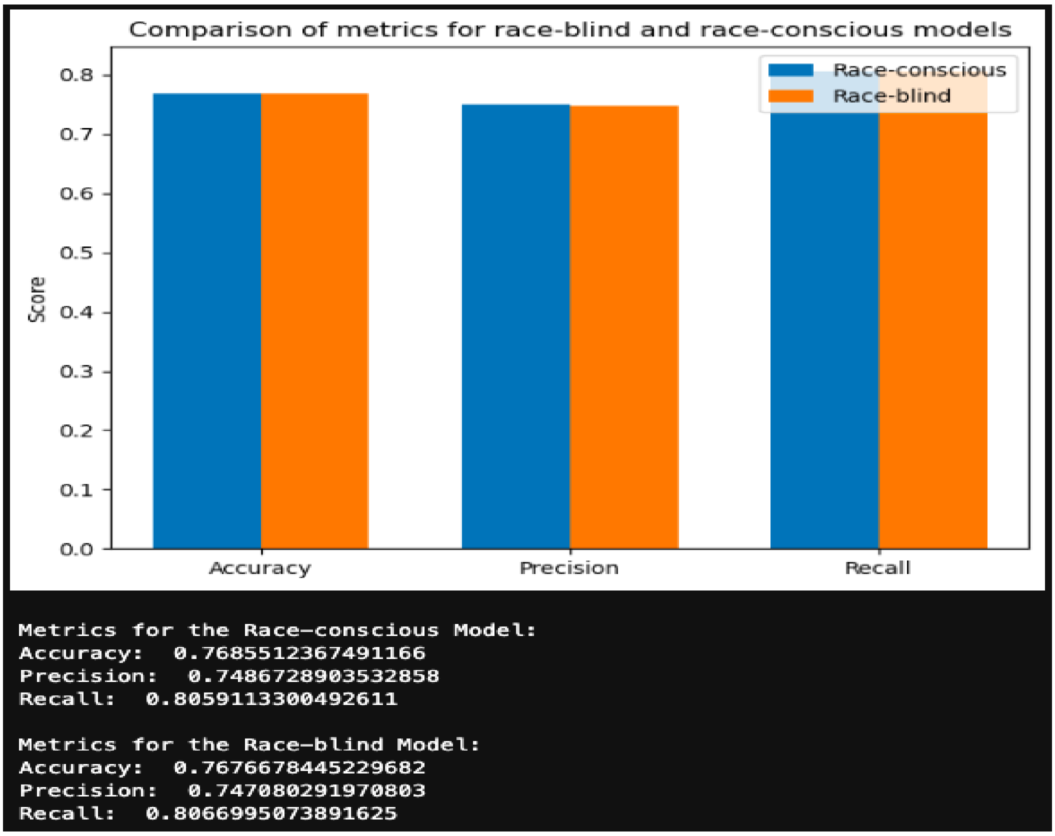
3.3.3 Applying the 'Equal Opportunity' Criterion

The "Equal Opportunity" fairness criterion requires that the true positive rate (TPR)/Recall is the same for both protected groups. In other words, it requires that the rate at which the model correctly diagnoses a patient with heart disease must be the same for either a white or non-white patient so that the model does not discriminate against any group in this regard.

To apply the Equal Opportunity criterion, I modified the decision threshold so that the true positive rate (Recall) is equal for both groups. One way to do this is to calculate the true positive rate separately for each demographic group and then set the decision threshold to ensure that these rates are equal.

4.0 Findings and Discussions

4.1 Fairness Through Unawareness



```
Metrics for the Race–conscious Model:
Accuracy:  0.7685512367491166
Precision:  0.7486728903532858
Recall:  0.8059113300492611

Metrics for the Race–blind Model:
Accuracy:  0.7676678445229682
Precision:  0.747080291970803
Recall:  0.8066995073891625
```

Figure 2.0 Metrics For Race-blind and Race-Conscious Models

Both models maintained similar results for Accuracy, Precision, and Recall although there was only a fractional decrease of Accuracy and Precision for the race-blind Model compared to that of the Race-conscious. This implies that there was not enough correlation between the *Race* attribute and the non-protected proxy parameters to have had a significant impact on the model's predictive accuracy, hence, further investigation was required.

From an ethical standpoint, it is important to ensure that the model is not biased against any specific group and that it does perpetuate or amplify existing inequalities in society. In this case, the race-conscious model appears to have slightly higher recall than the race-blind model, which means it is better at identifying positive cases for the protected group. However, this comes at the cost of slightly lower precision, which means that there may be more false positives (i.e., cases that are predicted to be positive but are negative) for the protected group.

From a practical standpoint, the differences in performance between the two models are relatively small, and it may not be worth the additional complexity and potential ethical concerns of using a race-conscious model.

4.2 The 'Equal Accuracy' Criterion

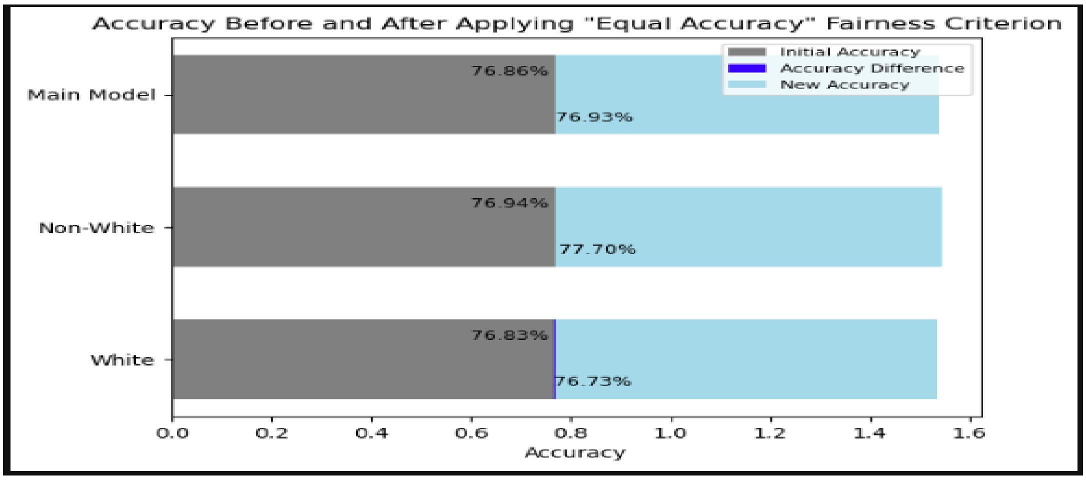


Figure 2.1 Effect of applying the Equal Accuracy Criterion.

As expected, the new results are almost the same as the initial accuracies of both groups and the model (77% approx for each). This means that the fractional difference between the initial accuracies for both groups is very small, and hence, the modification to the decision thresholds did not make a significant difference and was not necessary. Apparently, it may be difficult to achieve perfect equality in accuracy between the two protected groups as there exists just a minute difference between their accuracies. It can be fair to say that the model already satisfies Equal Accuracy across both race groups, hence the model generalizes fairly for both groups. This implies that for either a white or non-white patient, the model maintains the same predictive accuracy when predicting the presence or absence of heart disease.

4.3 Demographic Parity Criterion

From Figure 2.2 we see that after adjusting the model, the DIR moved very close to 1 (from 0.73 to 0.91) indicating that the adjusted model treats both groups almost equally.

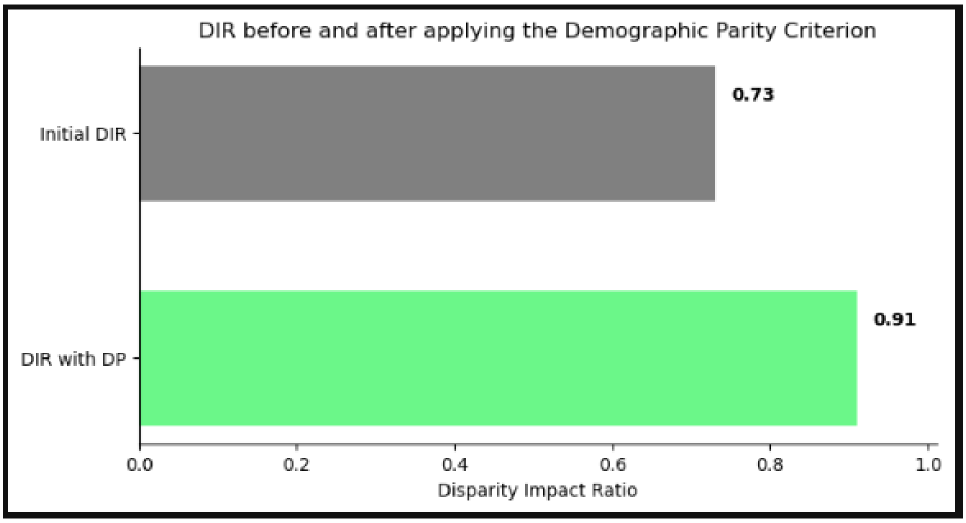


Figure 2.2 DIR Values Before and After Demographic Parity

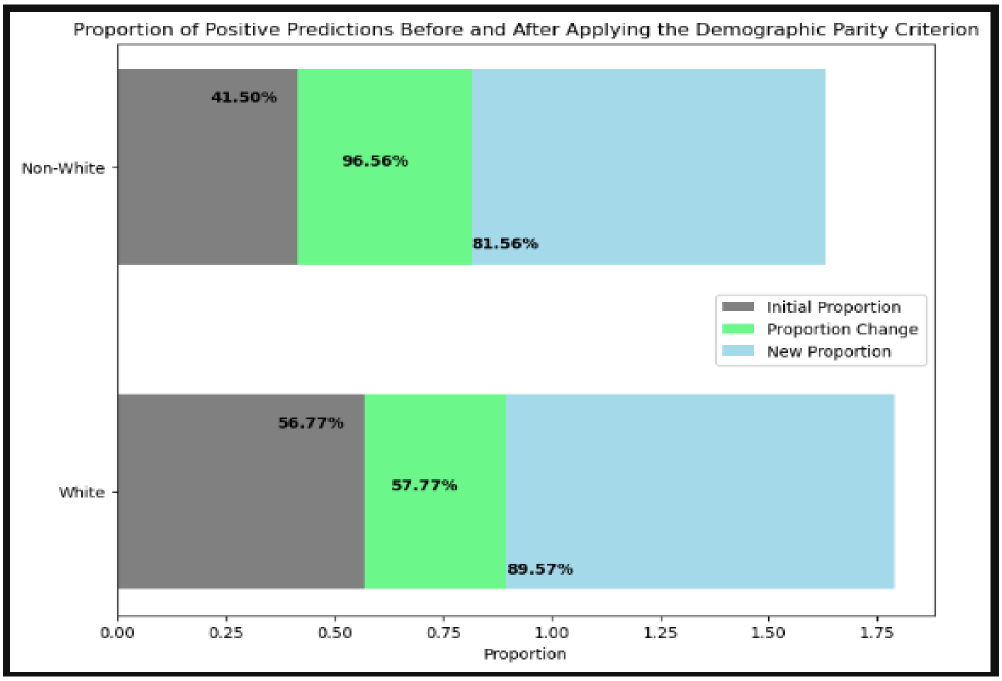


Figure 2.3 Changes in Proportions of Positive Predictions For Both Groups

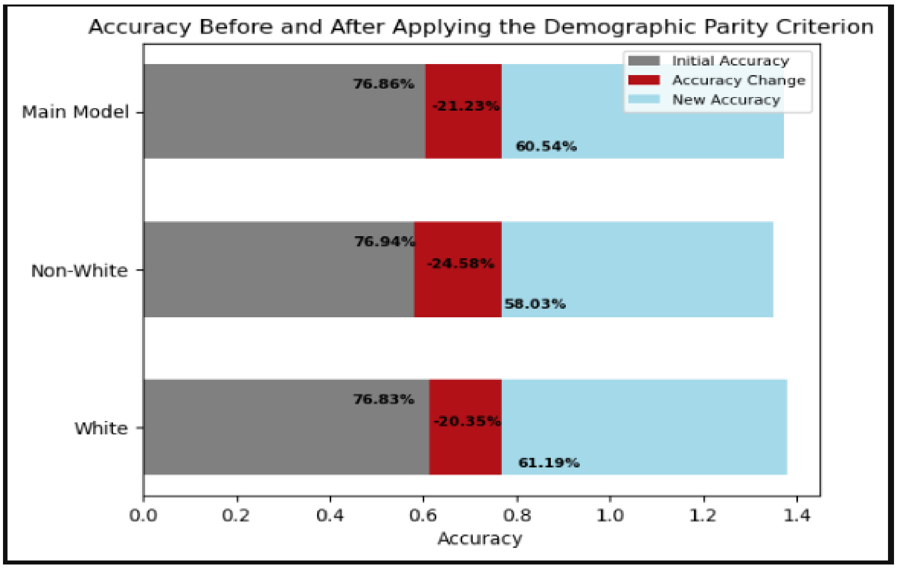


Figure 2.4 Effect of Demographic Parity on the Predictive Accuracy

However, ensuring Demographic Parity came at the cost of a huge depreciation in the model’s accuracy from 77% (approx. value for each group i.e. White race group, Non-white Race group, main model) to 60% (approx. value for each group). Even though the new accuracy is similar across the groups, it is rather too poor.

This implies that while ensuring an equal proportion of positive predictions for both groups, the chance that the model diagnoses a patient with heart disease (whether correctly or wrongly) increased by 97% approx. for the non-white group and by 58% approx. for the white group, but at the same time, increasing the chance of an erroneous diagnosis (More False positives).

4.4 **Equal Opportunity Criterion**

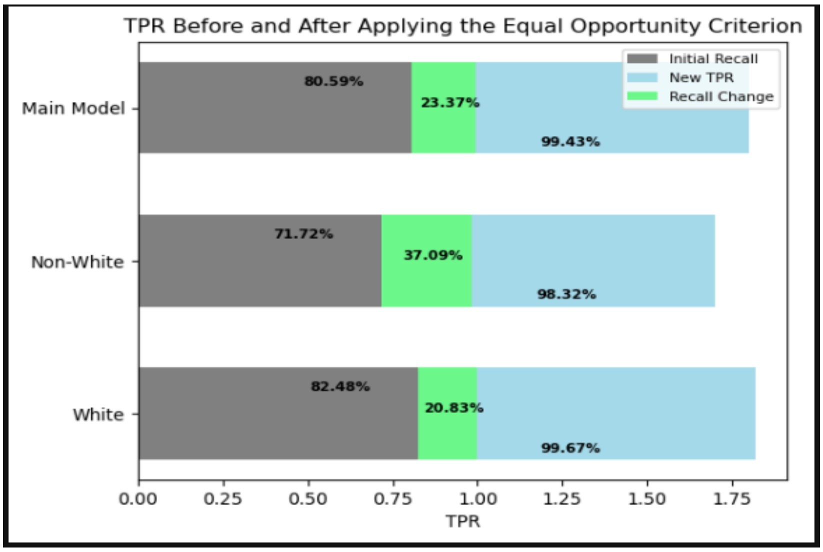


Figure 2.5 TPRs Before and After Equal Opportunity

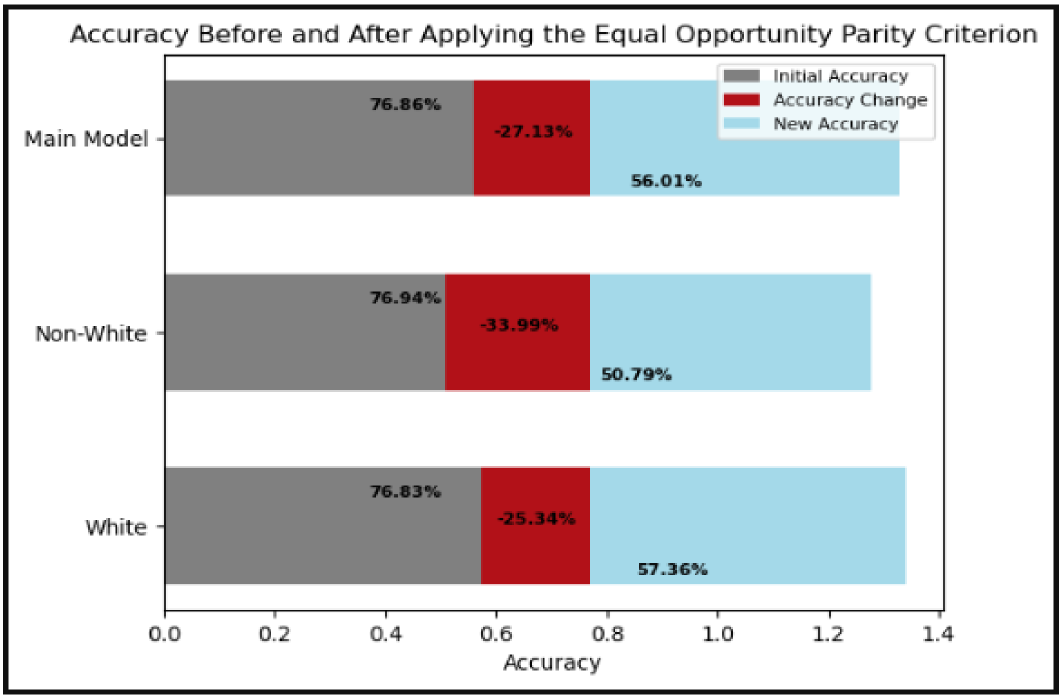


Figure 2.6 Effect of Equal Opportunity on Predictive Accuracy

The new TPR for both the White and Non-White groups improved, which suggests that the model is more effective in identifying heart disease in these groups. This is an important ethical implication as it ensures that the model is not biased against any particular group based on their race.

The unintended consequences of applying equal opportunity criteria to this predictive model for heart disease diagnosis are that although the TPR for both the white and non-white groups has significantly improved, the accuracy for both groups and overall accuracy has significantly decreased. This means that while the model is now better at identifying positive cases for both groups, it is now misclassifying a larger number of negative cases, leading to more false positives. This can result in unnecessary testing and medical procedures for patients who are actually healthy, leading to increased costs and potential harm to the patients. It can also lead to distrust in the healthcare system and the predictive model itself, particularly among non-white patients who may be disproportionately affected by the increased number of false positives.

5.0 Conclusion

This project further expanded my horizon on different approaches to the assessment of bias in a model. It has been observed that satisfying one fairness constraint can have adverse effects on other metrics. Determining the appropriate decision threshold for a model is a complex task that depends on various factors, including the specific characteristics of the data and the desired trade-offs between different types of errors and disparities.

It is worth noting that the metrics reported in the study are only a partial measure of the overall performance of the models, and that other factors such as computational efficiency, interpretability, and scalability may also be important considerations. Ultimately, the decision of whether to settle for any particular versions of the adjusted model should be based on a careful analysis of the specific context and trade-offs involved. Ultimately, using a dataset with a relatively even distribution of instances among protected groups can further ensure that a model is generally fair in its predictions.

Further research can aim to determine more optimal decision thresholds that further maximize the overall accuracy while satisfying fairness criteria constraints, and also, to further investigate the bias among the other minority race groups within the Non-white group.

References

Chen, Z., Liu, X., Yang, Q., Wang, Y.-J., Miao, K., Gong, Z., Yu, Y., Leonov, A., Liu, C., Feng, Z., & Chuan-Peng, H. (2023). Evaluation of Risk of Bias in Neuroimaging-Based Artificial Intelligence Models for Psychiatric Diagnosis: A Systematic Review. JAMA Network Open, 6(3), e231671. https://doi.org/10.1001/jamanetworkopen.2023.1671

Garb, H. N. (2021). Race bias and gender bias in the diagnosis of psychological disorders. Clinical Psychology Review, 90, 102087. https://doi.org/10.1016/j.cpr.2021.102087

Kamiran, F., & Calders, T. (2012). Data preprocessing techniques for classification without discrimination. Knowledge and information systems, 33(1), 1-33.

Kumar, A., Fonarow, G. C., Eagle, K. A., Hirsch, A. T., Califf, R. M., & Albert, M. A. (2016). Association of race and sex with risk of incident acute coronary heart disease events. JAMA, 316(20), 2115-2125.

Mosteiro, P., Kuiper, J., Masthoff, J., Scheepers, F., & Spruit, M. (2022). Bias Discovery in Machine Learning Models for Mental Health. Information, 13(5), Article 5. https://doi.org/10.3390/info13050237

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. Science, 366(6464), 447–453. https://doi.org/10.1126/science.aax2342

Paul, S., Maindarkar, M., Saxena, S., Saba, L., Turk, M., Kalra, M., Krishnan, P. R., & Suri, J. S. (2022). Bias Investigation in Artificial Intelligence Systems for Early Detection of Parkinson's Disease: A Narrative Review. Diagnostics, 12(1), Article 1. https://doi.org/10.3390/diagnostics12010166

Rajkomar, A., Hardt, M., Howell, M. D., Corrado, G., & Chin, M. H. (2018). Ensuring fairness in machine learning to advance health equity. Annals of Internal Medicine, 169(12), 866-872.