MEPS Data: Examining Utilization

Michael Garcia-Perez mig009@ucsd.edu
Christine Deng cydeng@ucsd.edu
Shivani Suthar ssuthar@ucsd.edu
Anika Garg agarg@ucsd.edu
Joshua Brusewitz jbrusewi@ucsd.edu
Emily Ramond eramond@deloitte.com

Abstract

This study delves into reducing biases in healthcare machine learning models by leveraging the MEPS (Medical Expenditure Panel Survey) dataset. It targets disparities in healthcare delivery and decision-making, particularly examining healthcare utilization among diverse patient demographics. While prior research has highlighted biases in these models, a gap exists in understanding utilization patterns across demographics. Our project addresses this by analyzing the MEPS dataset, seeking to refine healthcare cost predictions and resource allocation strategies for improved patient outcomes and equity in the healthcare system. By using the MEPS dataset, which contains comprehensive information on healthcare utilization and expenditure, our methods focus on uncovering how different demographics utilize healthcare, contributing to a more nuanced understanding of biases in machine learning models and their impact on healthcare disparities.

Code: https://github.com/christine-deng/DSC180A-Capstone

1 Introduction

In this paper, we delve into techniques aimed at addressing inequalities within healthcare machine learning models by analyzing the MEPS dataset. Understanding and mitigating healthcare system inequalities, particularly within machine learning, is crucial as these models significantly impact healthcare policies, distribution, and patient outcomes. For example, consider the ripple effect of an error within a predictive model estimating a patient's healthcare costs; such a miscalculation could leave the patient financially unprepared for their treatment, potentially depriving them of essential care. Moreover, extensive evidence, as highlighted in works like Williams et al. (2002), highlights the persistent presence of discrimination within healthcare systems, often rooted in factors like race and gender, magnifying the urgency for reform. It's essential to recognize that the impact of healthcare systems extends beyond patients alone; administrative staff, nurses, developers, and policymakers are integral stakeholders influenced by these systems because the accuracy of these models significantly shapes their operational strategies, negotiations, and overall effectiveness. Our approach centers on forecasting healthcare utilization, enabling healthcare providers to pinpoint patients needing more care, thus optimizing resource allocation—such as staffing and inventory—towards locations expected to serve a higher proportion of high-utilization patients. In the process of developing several predictive models, our final recommendation for building a 'fair' classification model for utilization is a logistic regressor, with adversarial training as the bias-mitigation technique. Ultimately, our project seeks to rectify bias in healthcare models, fostering fairness in resource allocation and ultimately enhancing patient well-being and quality of life.

1.1 Literature Review

The rise of machine learning model usage in healthcare systems has sparked growing concern regarding its potential impact on fairness in medical care. Chen et al. (2022) offer a comprehensive overview of the complex issue of algorithm fairness in the development and implementation of artificial intelligence (AI) systems in healthcare, as discussed in "Algorithm Fairness in AI for Medicine and Healthcare." The paper delves into the critical problem of ensuring equitable care, particularly in AI-driven healthcare, where recent assessments of AI models have exposed disparities in diagnosis, treatment recommendations, and healthcare billing across different racial subpopulations. Within the context of current healthcare challenges, this perspective piece synthesizes the multifaceted realm of fairness in machine learning, shedding light on the origins of algorithmic biases in today's clinical workflows. These biases, ranging from variations in image acquisition to genetic diversity and inconsistencies in how healthcare professionals label data, significantly contribute to healthcare disparities. Furthermore, the article takes a close look at the cutting-edge technologies and approaches aimed at mitigating bias within AI systems, including federated learning, disentanglement techniques, and model explainability.

Similarly, Chen, Pierson, Rose, Joshi, Ferryman, and Ghassemi (2021) discuss the ethical considerations for each step in the pipeline of model development in the context of health-

care in their work "Ethical machine learning in healthcare." Machine learning models may exacerbate the existing health disparities and inequities in our current society, through biases present in the different pipeline steps. In particular, data collection is relevant to the MEPS data because marginalized populations may be less represented in the dataset and face lowered model performance as a result. Another step discussed that is important for the replication project is the post-deployment considerations; the paper stresses regular audits and advises evaluating model performance across distinct groups and outcomes to uncover possible issues. Given potential underrepresentation of marginalized populations in the dataset, focusing audits on these subgroups proves crucial for spot-checking.

These inequalities in healthcare have even led to differences in life-expectancy among different demographic groups, as highlighted in "Understanding and Mitigating Health Inequities — Past, Current, and Future Directions" published in the New England Journal of Medicine. Furthermore, "Factors That Affect Health-Care Utilization" discusses the multifaceted factors that affect healthcare utilization and access in the United States. These articles demonstrate how there are clear racial and ethnic disparities, with minority populations often facing lower socioeconomic status, limited access to healthful environments, and higher rates of chronic illnesses. The Affordable Care Act (ACA) has made strides in reducing the uninsured population but disparities still persist (National Academies of Sciences). Additionally, unconscious bias among healthcare providers can perpetuate inequalities in care. Socioeconomic factors, including income and poverty, are closely linked to risk factors for chronic diseases. Income influences access to healthcare resources, and those with lower income levels often experience barriers to receiving needed medical care. Geographic disparities are also notable, with rural areas typically having fewer healthcare providers and transportation challenges that impact access to care (Lavizzo-Mourey et al. 1681). Finally, individuals with disabilities face unique challenges in accessing healthcare, including physical limitations, lack of accommodations, and discrimination. Despite having greater healthcare needs, these individuals may encounter difficulties in obtaining the services they require. Addressing these complex issues is essential to achieving healthcare equity and ensuring that all Americans have access to quality care.

Inequalities in healthcare have impacts beyond just patient health; the issue significantly impacts the economy as well, as explained in "The financial toll of health disparities in the United States." The article examines how health disparities by race, ethnicity, and education level pose a significant economic burden statewide and nationally in the United States (LaVeist et al., 2023). Overall, current literature addresses many of the growing concerns over the impact of machine learning in healthcare and ways to mitigate potential harms, and further research is imperative in the area of machine learning in healthcare to ensure fairness in medical care moving forward.

1.2 Description of Relevant Data

The MEPS dataset details comprehensive information on healthcare patients in the United States. This data was collected on a nationally representative sample of the civilian noninstitutionalized population of the United States by the Agency for Healthcare Research and

Quality in 2015. The data collection process involved a combination of household interviews and surveys completed by individuals and families, as well as information provided by employers, medical providers, and health insurance companies.The original dataset contains 35427 rows and 1831 columns before preprocessing scripts are applied.

The data it contains ranges from demographic information on patients such as age and race, full details of their healthcare situation ranging down to the extremely minute, and financial data relating to expenditure and personal wealth. The dataset is incredibly granular in its parameters, with information on even obscure subjects, such as whether or not a doctor has advised the patient to wear a bike helmet. This granularity means that anyone working with the data will have to greatly pare down the features of the dataset in order to make it usable, as well as mitigate a large amount of potential bias deriving from the great amount of demographic information it contains.

As a result of this, models built off of the dataset make for an excellent case study on AI fairness. Common use cases of this data will relate to either deriving insight into social issues in America or creating a better experience for patients or hospitals, and these are situations wherein bias absolutely must be mitigated as much as possible to ensure an effective and equitable outcome, to a degree far greater than other tasks with other datasets. The complexity of the dataset also means that tools that can effectively mitigate bias within it have a great chance to prove their worth if they can work well here. Ultimately, refining AI methods to responsibly leverage the MEPS dataset's depth can pave the way for more equitable and precise healthcare analytics, setting a standard for the responsible use of AI in sensitive and multifaceted domains.

2 Methods

The methodology of our project was done in the several steps outlined below. Here, we will provide an explanation of each step and its purpose.

2.1 Exploratory Data Analysis (EDA)

2.1.1 Pre-processing

To measure utilization, we created the feature, 'UTILIZATION' to measure an individual's total number of trips requiring a form of medical care. This composite feature was created by summing up the individual's number of office based visits, outpatient visits, ER visits, inpatient nights, and home health visits.

The model classification task is to predict whether an individual will have high utilization. To define "high", we set the threshold for the 'UTILIZATION' feature to be 10, which is approximately the average utilization for the considered population. Any individual with a 'UTILIZATION' feature of 10 or greater is considered to have high utilization.

To pre-process the data, scripts were adapted from IBM's AIF360 package. In particular,

the following steps were applied:

1. Create a new column called ”RACE” that takes on the value ”White” if the individual is reported as White or Non-Hispanic White, and ”non-White” otherwise.
2. Rename columns that are panel/round specific, generalizing the variable names and removing extraneous numbers specific to that panel/round.
3. Drop rows with missing data, where the features have values < -1
4. Compute the ”UTILIZATION” feature mentioned above, and binarize it to 0 if it is < 10 and 1 otherwise.

Additionally, outlier detection and handling of nulls was done through an investigation of the dataset, seen more thoroughly in the code.

2.1.2 Distributions of Features

We decided to examine the distribution of key features in our dataset, mostly demographics about the individual. For concision, the figures reflect only the subsection of the MEPS dataset using panel number 19, the figures using panel number 20 are omitted due to the distributions being largely similar. Refer to the code for the full figures. A more in-depth examination of features in the dataset was performed, but the following will outline key features important in correlation analysis. Refer to the code for an investigation on other features.

Sex

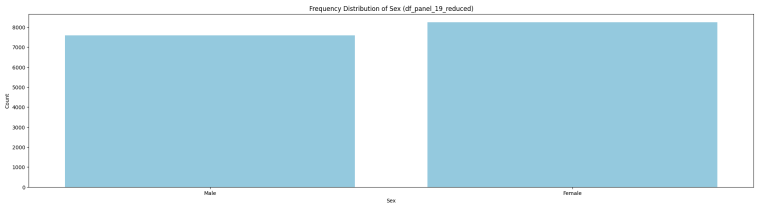

Figure 1: Frequency distribution of Sex

Our dataset features more female entries than male entries. The dataset features samples drawn from a nationally representative subsample of U.S. households, and the percentage of female population in the U.S. is slightly higher than the male population, but not by a large percentage. As such, the sex distribution in our dataset is consistent with it being a nationally representative subsample of the U.S. sex distribution.

Age

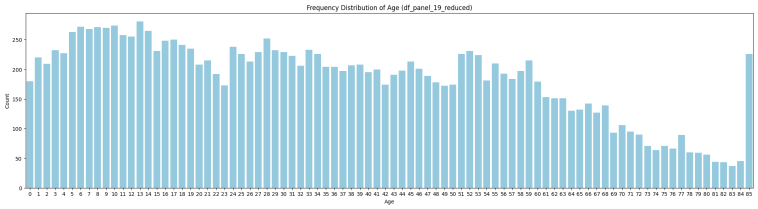

Figure 2: Frequency distribution of Age

The age distributions for individuals is mostly uniform from ages 0 through 55. There are slight dips in the distribution at certain ages but no significant peaks. After the age 55, there is a downward trend where there are less people for each year increase in age. There is, however, a spike at age 85. 85 is the greatest age to be eligible for the dataset, and 0 is the minimum age.

Race

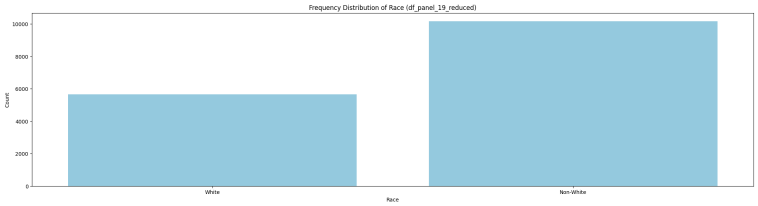

Figure 3: Frequency distribution of Race

The dataset features a greater number of Non-White individuals than White Individuals. Looking at our pre-processing methods, all races other than White were grouped into the singular category “Non-White”. This includes individuals classified as Black, Asian, American Indian/Alaska Native, Native Hawaiian/Pacific Islander, or Multiple Races Reported, leading to a grouped count of Non-White individuals.

2.1.3 Correlation Analysis

Next, we looked at the proportion of different demographics in our dataset who had high utilization. Notably, we examined some key demographic information about individuals such as their race, sex, and age. We believe these features will contribute to an individual’s utilization status given the context that in healthcare, certain populations (e.g., older people) are more at-risk and will require greater care. Additionally, certain populations (e.g., wealthier individuals) will have greater access to care, allowing for more utilization.

Sex vs. Utilization

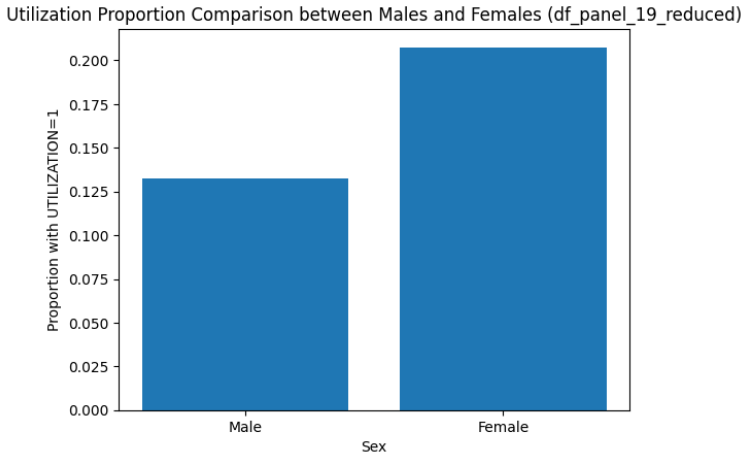

Figure 4: Utilization Proportion Comparison between Males and Females

A greater proportion of females had high utilization compared to males.

Race vs. Utilization

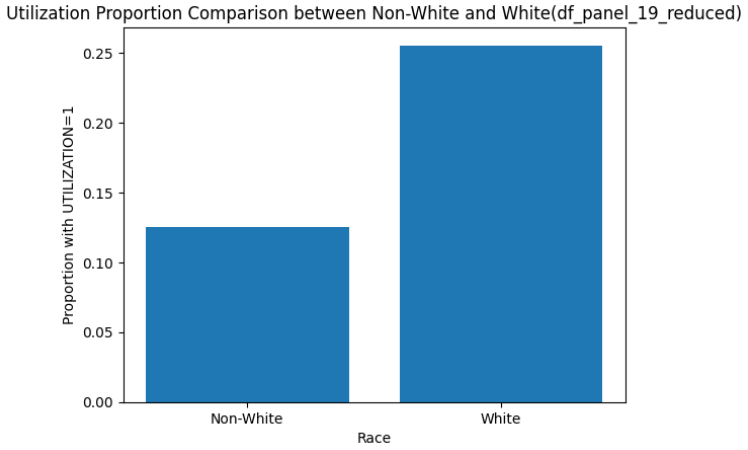

Figure 5: Utilization Proportion Comparison between non-White and White

A greater proportion of white individuals had high utilization compared to their non-white counterparts.

Using these important features of sex and race to determine how they correlate with utilization, we performed a correlation analysis by computing a matrix.

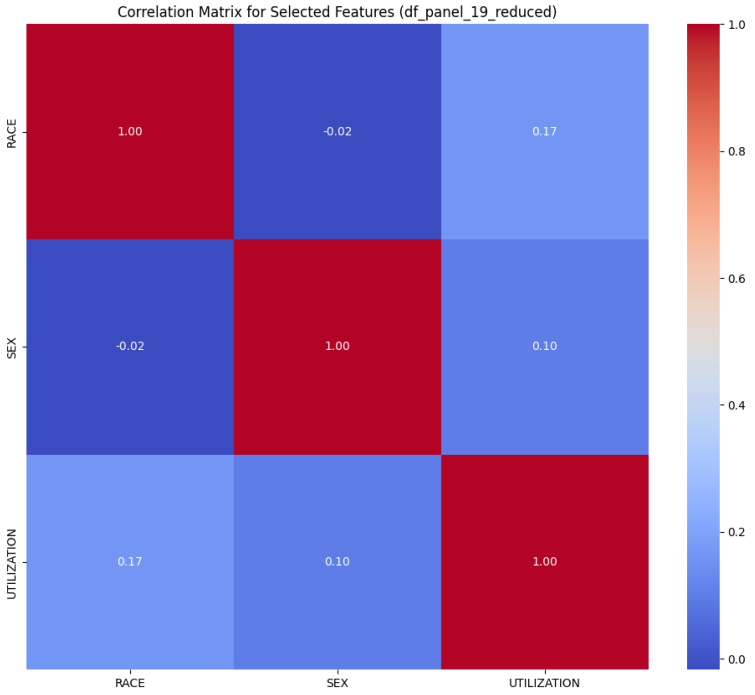

Figure 6: Correlation Matrix for Utilization, Sex, Race

It can be seen that race and utilization have a positive correlation of 0.17 and sex and utilization have a positive correlation of 0.10. This can potentially indicate that one's race is more determinant of their utilization than their sex.

Age vs. Utilization

We also binned the ages to determine how utilization looked for the different age groups.

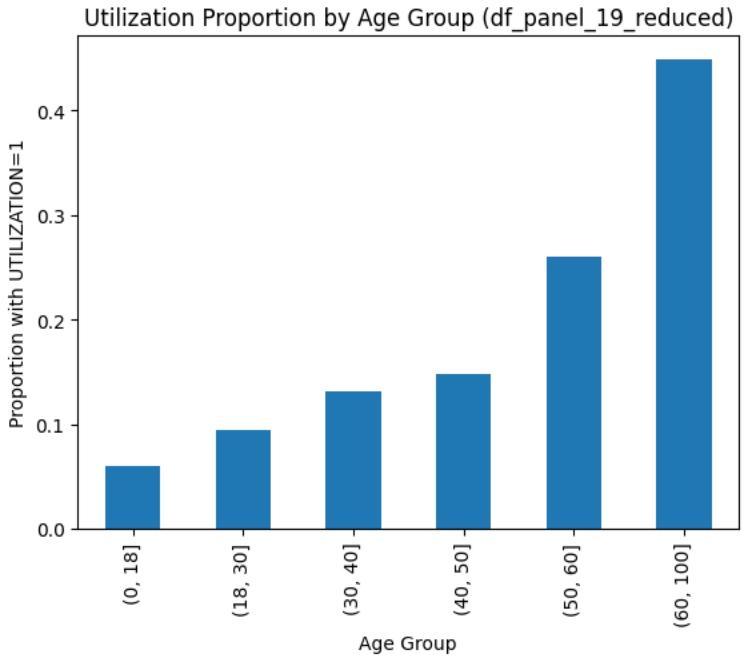

Figure 7: Utilization Proportion by Age Group

The distribution was left skewed, meaning older individuals had higher utilization. The bin with individuals aged 60-100 (the highest age in the dataset is 85) had the highest proportion of individuals with high utilization. This makes sense given the context of healthcare since the older one is, the more health conditions they will have and require more utilization of medical services.

We decided to examine the correlation between age and utilization by plotting the trend between the ages and utilization. We saw a clear upward trend where the utilization proportion increased with age.

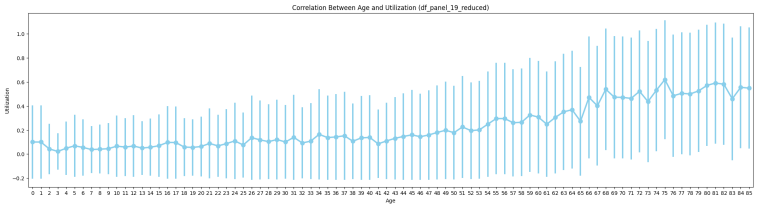

Figure 8: Correlation between Age and Utilization

In our analysis, we carefully chose features to assess multicollinearity with utilization, aiming to gain insights into potential relationships and influences. The rationale behind these choices is rooted in a comprehensive analysis strategy. We considered demographic factors, health status indicators, and specific health conditions to ensure a holistic view of potential multicollinearity with utilization. Additionally, we incorporated confidence intervals to provide a statistical context to our findings, enhancing the robustness of our analysis.

2.1.4 Additional Analysis

Lastly, we did some additional analysis by creating a pair plot for certain features pertaining to one's health. More specifically, we chose age, weight, mental health, and physical health all of which are detailed further below in our analysis of the pair plot.

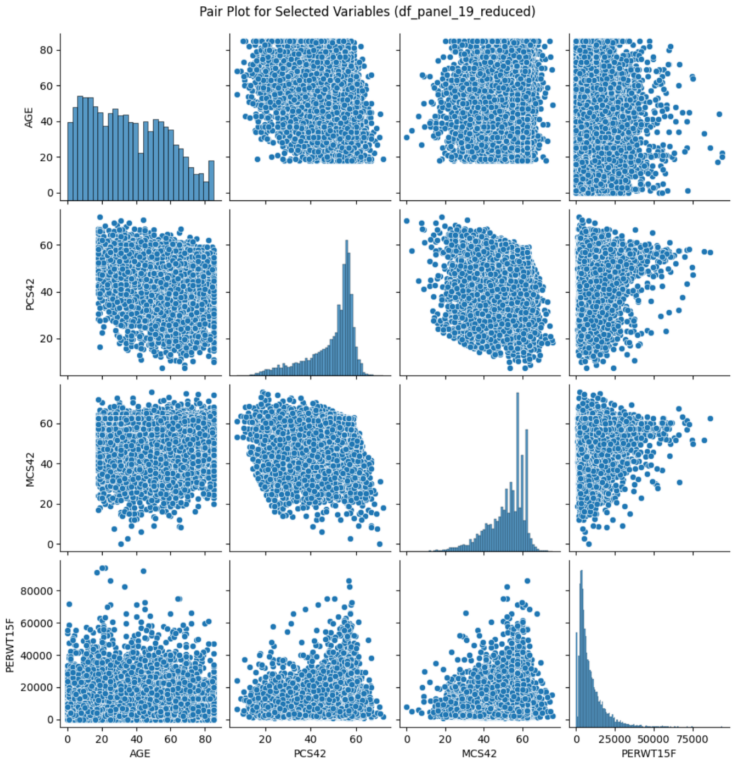

Figure 9: Pair Plot of Age, Weight, Mental Health, and Physical Health

The features used in the pair plot are defined as such:

- Age: Age of the individual.
- Pcs42 - Officially defined as "PHY COMPONENT SUMMRY SF-12V2 IMPUTED". In essence, it is the score someone got on the physical component of a health-related quality-of-life questionnaire.
- Mcs42 - Officially defined as "MNT COMPONENT SUMMRY SF-12V2 IMPUTED". In essence, it is the score someone got on the mental component of a health-related quality-of-life questionnaire.
- Perwt15f - Officially defined as "person-level weight". In essence, it is the weight of a person calculated by a certain algorithm.

As seen from the plot above, while the distribution of age seems to be closer to uniform, the distribution of pcs42 and mcs42 seem to be left skewed, and the distribution of perwt15f seems to be right skewed. Additionally, the relationships between the variables all look pretty clustered around one area and don't seem to have a clear shape to them indicating

that there is likely no significant relationship between any of these variables.

In summary, our feature selection was guided by the need to capture diverse aspects of health and demographics, enabling us to identify potential multicollinearity and understand the complex interplay between various factors and healthcare utilization.

2.2 Model Development and Fairness Evaluation

2.2.1 Training Models Without Debiasing

We trained two baseline models for our classifying utilization task: a Logistic Regression model, and a Random Forest model.

Our protected attribute is Race, with our privileged group being White and our unprivileged group being non-White. The features used in predicting utilization are as follows:

- 'AGE': Age of the individual.
- 'RACE': Race of the individual.
- 'PCS42': Physical Component Summary (PCS) score at wave 4 and wave 2.
- 'MCS42': Mental Component Summary (MCS) score at wave 4 and wave 2.
- 'K6SUM42': Kessler 6 (K6) non-specific psychological distress scale at wave 4 and wave 2.
- 'REGION': Geographic region of the individual.
- 'SEX': Gender of the individual.
- 'MARRY': Marital status of the individual.
- 'FTSTU': Full-time student status.
- 'ACTDTY': Activity status.
- 'HONRDC': Honor and recognition status.
- 'RTHLTH': Overall health.
- 'MNHLTH': Mental health.
- 'HIBPDX': High blood pressure diagnosis.
- 'CHDDX': Coronary heart disease diagnosis.
- 'ANGIDX': Angina diagnosis.
- 'MIDX': Myocardial infarction diagnosis.
- 'OHRTDX': Other heart disease diagnosis.
- 'STRKDX': Stroke diagnosis.
- 'EMPHDX': Emphysema diagnosis.
- 'CHBRON': Chronic bronchitis diagnosis.
- 'CHOLDX': High cholesterol diagnosis.
- 'CANCERDX': Cancer diagnosis.
- 'DIABDX': Diabetes diagnosis.
- 'JTPAIN': Joint pain diagnosis.
- 'ARTHDX': Arthritis diagnosis.
- 'ARTHTYPE': Arthritis type.

- 'ASTHDX': Asthma diagnosis.
- 'ADHDADDX': ADHD/ADD diagnosis.
- 'PREGNT': Pregnancy status.
- 'WLKLIM': Walking limitation.
- 'ACTLIM': Activity limitation.
- 'SOCLIM': Social limiation.
- 'COGLIM': Cognitive limitation.
- 'DFHEAR42': Difficulty hearing at wave 4 and wave 2.
- 'DFSEE42': Difficulty seeing at wave 4 and wave 2.
- 'ADSMOK42': Smoking status at wave 4 and wave 2.
- 'PHQ242': Patient Health Questionnaire-2 (PHQ-2) score at wave 4 and wave 2.
- 'EMPST': Employment status.
- 'POVCAT': Poverty status.
- 'INSCOV': Insurance coverage.

To develop our baseline models, a Logistic Regression classifier was employed as the predictive model for high utilization. To enhance the model's robustness, a StandardScaler was applied to standardize the features. The Logistic Regression model was fitted to the training data using a pipeline, incorporating the necessary preprocessing steps.

A Random Forest classifier was also trained, using the hyperparameters of 500 decision trees, each with a minimum of 25 samples per leaf. As with logistic regression, feature standardization was applied, and instance weights were considered during training.

A validation set was developed to find the best threshold hyperparameter corresponding to the highest best balanced accuracy. Thresholding allowed us to evaluate the models' sensitivity to different classification thresholds used as hyperparameters when training both the Logisitic Regression and Random Forest classifiers. The threshold values ranged from 0.01 to 0.5, with 50 evenly spaced points, allowing us to determine the trade-off between accuracy and fairness.

2.2.2 Training Models Without Debiasing - Additional Model Development

In our additional model development, we performed Feature Selection by selecting specific columns during model fitting. We selected 5 columns in our dataset, which include the the individuals' region, age, sex, race, and marital status. The training process remains exactly the same as the baseline model, the only difference is which features are being used in the prediction task (i.e., the dataset the models are trained on contains less columns).

2.3 Bias Mitigation Techniques

2.3.1 Bias Mitigation in Preprocessing Stage Using Reweighing

To mitigate bias, the Reweighing technique from IBM's AIF360 toolkit was applied to our baseline Logistic Regression and Random Forest models. This pre-processing technique

adjusts the weights of instances in the dataset to ensure fairness across different groups. As seen in the EDA, there is a difference in the proportion of demographics who have high utilization, versus those who do not. Reweighing can ensure our models are not biased towards the majority class, which improves the training process and overall performance for the minority class. In addition to class imbalance, Reweighing also factors in protected attributes and adjusts instance weights to mitigate biases related to that attribute, which in our dataset, is Race.

The Reweighing technique transforms the dataset the models are trained on, so the training process for the Logistic Regression and Random Forest models remains the same as described above, in the baseline methodology. The only difference is that the models are trained on the transformed data.

Pre-processing techniques aim to mitigate bias inherent in the dataset before model training. This approach is useful when the bias is rooted in the dataset and needs to be addressed prior to model learning. In-processing techniques, on the other hand, mitigate bias during the training of the machine learning model. Our models do not utilize in-processing techniques. A combined approach using both pre-processing and in-processing techniques, can, however, allow for a comprehensive strategy for mitigating bias, addressing issues in the dataset while refining the model training process.

2.3.2 Bias Mitigation in Postprocessing Stage Using Disparate Impact Remover

Lastly, a Disparate Impact Remover was used to mitigate biases at various repair levels to transform the data. The Disparate Impact Remover is designed to reduce the influence of protected attributes (in our dataset, Race) in the decision-making process of the models. The type of bias we are trying to mitigate in our models is disparate impact, which is the ratio of favorable outcomes for the unprivileged group to that of the privileged group.

We used several different repair levels, ranging from 0 to 1 in increments of 0.1. The repair levels controls the level at which the Disparate Impact Remover is applied, and the hyperparameter adjusts the balance between fairness and accuracy.

The Disparate Impact Remover modifies the features of the input data to reduce the correlation between the protected attributes and the model's output. This attempts to mitigate biases caused by the protected attributes in the dataset. Because the only difference is that the models are trained on the transformed data, the training process for the Logistic Regression and Random Forest models remains the same as described above, in the baseline methodology.

3 Results

Here are the results of each of the model development stages of our methodology section outlined above.

Additionally, here are the definitions of each of the fairness metrics we used to evaluate our models:

- Threshold corresponding to Best balanced accuracy: The threshold corresponding to the best balanced accuracy is a hyper-parameter used to assign the data to labels/outputs, and these values correspond to the thresholds that result in the highest balanced accuracy for the models, meaning they are the optimal thresholds.

- Best balanced accuracy: This metric refers to the overall accuracy of the models and does not consider the privileged vs. unprivileged group.

- Corresponding 1-min(DI, 1/DI) value: Disparate Impact (DI) is the probability of success given the unprivileged group, divided by the probability of success given the privileged group, but is rewritten as 1 - min(DI, 1/DI) here to convert it to a proportion, since the DI can be greater than 1 if the privileged group has a less chance of success.

- Corresponding average odds difference value: Computed as the average difference of false positive rate (false positives / negatives) and true positive rate (true positives / positives) between unprivileged and privileged groups.

- Corresponding statistical parity difference value: Computed as the difference of the rate of favorable outcomes received by the unprivileged group to the privileged group.

- Corresponding equal opportunity difference value: This metric is computed as the difference of true positive rates between the unprivileged and privileged groups. The true positive rate is the ratio of true positives to the total number of actual positives for a given group.

- Corresponding Theil index value: Computed as the generalized entropy of benefit for all individuals in the dataset, with alpha = 1. It measures the inequality in benefit allocation for individuals.

3.1 Model Development and Fairness Evaluation

3.1.1 Training Models Without Debiasing

Logistic Regression (LR) Model

After we build a logistic regression model without debiasing, here were our results:

First, we created a plot comparing the 1 - min(DI, 1/DI) value and balanced accuracy of

our logistic regression model:

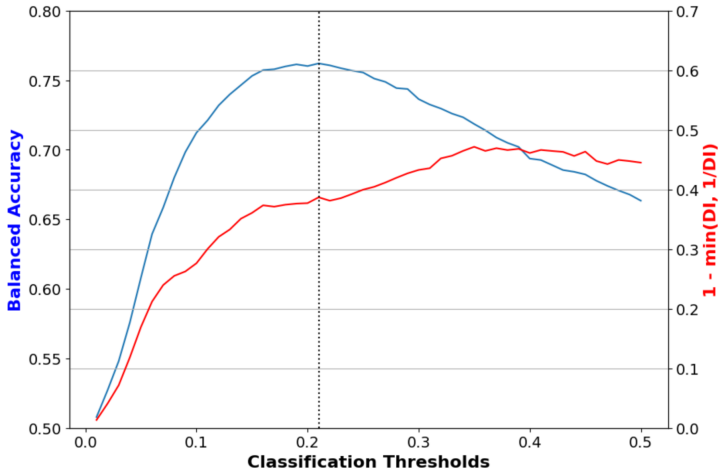

Figure 10: 1 - min(DI, 1/DI) and Balanced Accuracy of LR Model Without Debiasing

Second, we created a plot comparing the average odds difference and balanced accuracy of our logistic regression model:

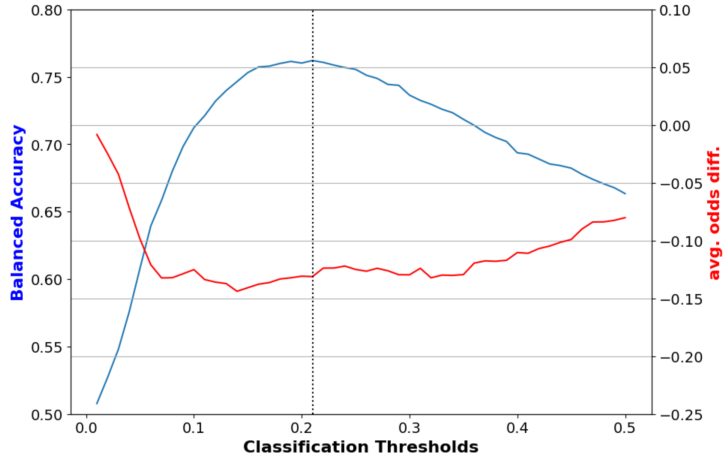

Figure 11: Average Odds Difference and Balanced Accuracy of LR Model Without Debiasing

Third, we generated the following fairness metrics after testing our LR model:

- Threshold corresponding to Best balanced accuracy: 0.2100
- Best balanced accuracy: 0.7505
- Corresponding 1-min(DI, 1/DI) value: 0.4000
- Corresponding average odds difference value: -0.1457
- Corresponding statistical parity difference value: -0.1839
- Corresponding equal opportunity difference value: -0.1496
- Corresponding Theil index value: 0.0957

Random Forest (RF) Model

After we build a random forest model without debiasing, here were our results:

First, we created a plot comparing the 1 - min(DI, 1/DI) value and balanced accuracy of our random forest model:

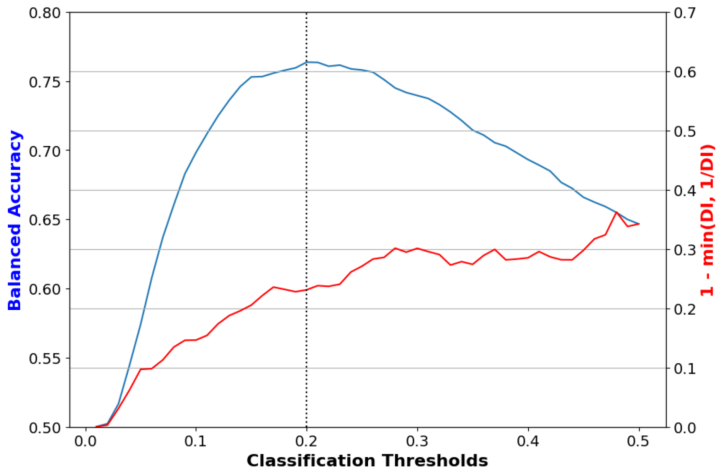

Figure 12: 1 - min(DI, 1/DI) and Balanced Accuracy of RF Model Without Debiasing

Second, we created a plot comparing the average odds difference and balanced accuracy of our random forest model:

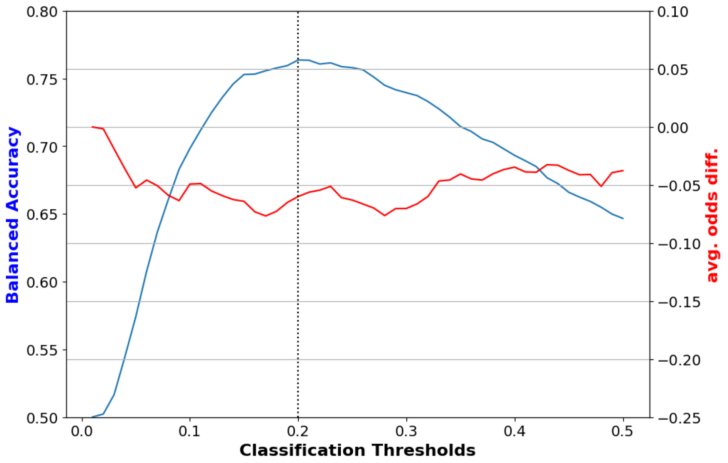

Figure 13: Average Odds Difference and Balanced Accuracy of RF Model Without Debiasing

Third, we generated the following fairness metrics after testing our RF model:

- Threshold corresponding to Best balanced accuracy: 0.2000
- Best balanced accuracy: 0.7558
- Corresponding 1-min(DI, 1/DI) value: 0.2901

- Corresponding average odds difference value: -0.0894
- Corresponding statistical parity difference value: -0.1393
- Corresponding equal opportunity difference value: -0.0729
- Corresponding Theil index value: 0.0914

3.1.2 Training Models Without Debiasing - Additional Model Development

Logistic Regression (LR) Model
After developing our logistic regression model through the additional model development steps mentioned above, these were our results on the fairness metrics after testing:

- Threshold corresponding to Best balanced accuracy: 0.2800
- Best balanced accuracy: 0.7001
- Corresponding 1-min(DI, 1/DI) value: 0.4888
- Corresponding average odds difference value: -0.1758
- Corresponding statistical parity difference value: -0.1886
- Corresponding equal opportunity difference value: -0.1981
- Corresponding Theil index value: 0.1166

Random Forest (RF) Model
After developing our logistic regression model through the additional model development steps mentioned above, these were our results on the fairness metrics after testing:

- Threshold corresponding to Best balanced accuracy: 0.1800
- Best balanced accuracy: 0.7196
- Corresponding 1-min(DI, 1/DI) value: 0.2826
- Corresponding average odds difference value: -0.1008
- Corresponding statistical parity difference value: -0.1381
- Corresponding equal opportunity difference value: -0.0831
- Corresponding Theil index value: 0.1012

3.2 Bias Mitigation Techniques

3.2.1 Bias Mitigation in Preprocessing Stage Using Reweighing

Logistic Regression (LR) Model
After developing our logistic regression model using the bias mitigation technique of reweighing, here were our results:

First, we created a plot comparing the 1 - min(DI, 1/DI) value and balanced accuracy of our logistic regression model:

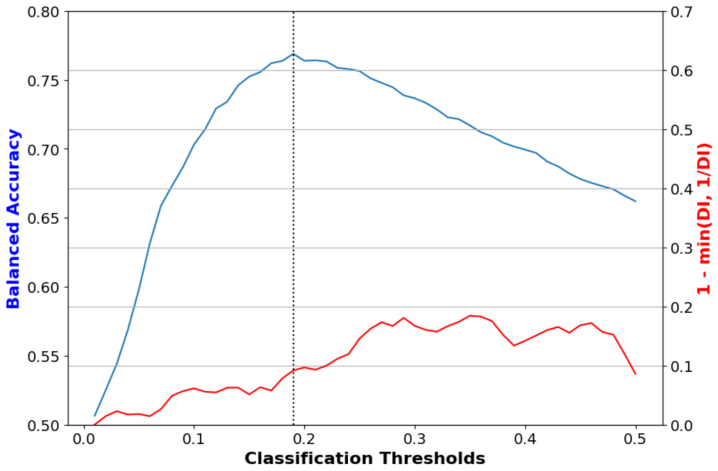

Figure 14: 1 - min(DI, 1/DI) and Balanced Accuracy of LR Model After Reweighing

Second, we created a plot comparing the average odds difference and balanced accuracy of our logistic regression model:

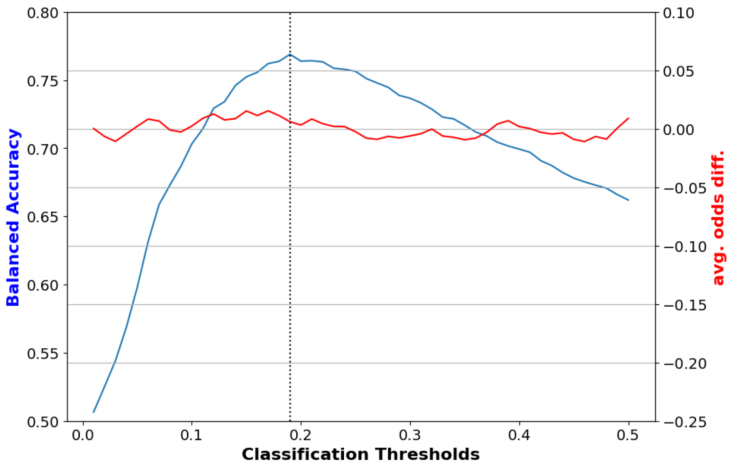

Figure 15: Average Odds Difference and Balanced Accuracy of LR Model After Reweighing

Third, we generated the following fairness metrics after testing our LR model:

- Threshold corresponding to Best balanced accuracy: 0.1900
- Best balanced accuracy: 0.7544
- Corresponding 1-min(DI, 1/DI) value: 0.1153
- Corresponding average odds difference value: -0.0077
- Corresponding statistical parity difference value: -0.0466
- Corresponding equal opportunity difference value: -0.0137
- Corresponding Theil index value: 0.0912

Random Forest (RF) Model

After developing our random forest model using the bias mitigation technique of reweigh-

ing, here were our results:

First, we created a plot comparing the 1 - min(DI, 1/DI) value and balanced accuracy of our random forest model:

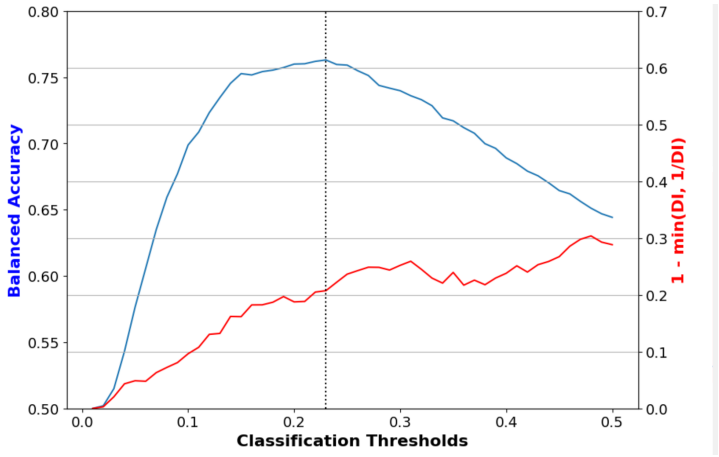

Figure 16: 1 - min(DI, 1/DI) and Balanced Accuracy of RF Model After Reweighing

Second, we created a plot comparing the average odds difference and balanced accuracy of our random forest model:

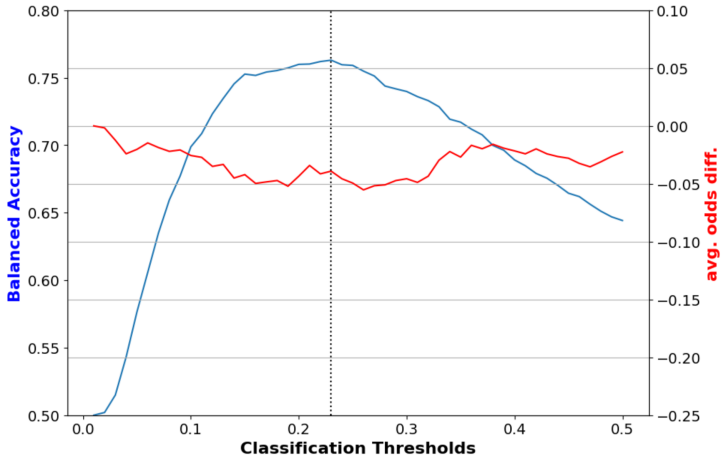

Figure 17: Average Odds Difference and Balanced Accuracy of RF Model After Reweighting

Third, we generated the following fairness metrics after testing our RF model:

- Threshold corresponding to Best balanced accuracy: 0.2300
- Best balanced accuracy: 0.7566
- Corresponding 1-min(DI, 1/DI) value: 0.2743
- Corresponding average odds difference value: -0.0692
- Corresponding statistical parity difference value: -0.1150
- Corresponding equal opportunity difference value: -0.0616

- Corresponding Theil index value: 0.0944

3.2.2 Bias Mitigation in Postprocessing Stage Using Disparate Impact Remover

Logistic Regression (LR) Model
After developing our logistic regression model using the disparate impact (DI) remover, here were our results:

First, we created a repair level vs. balanced accuracy plot:

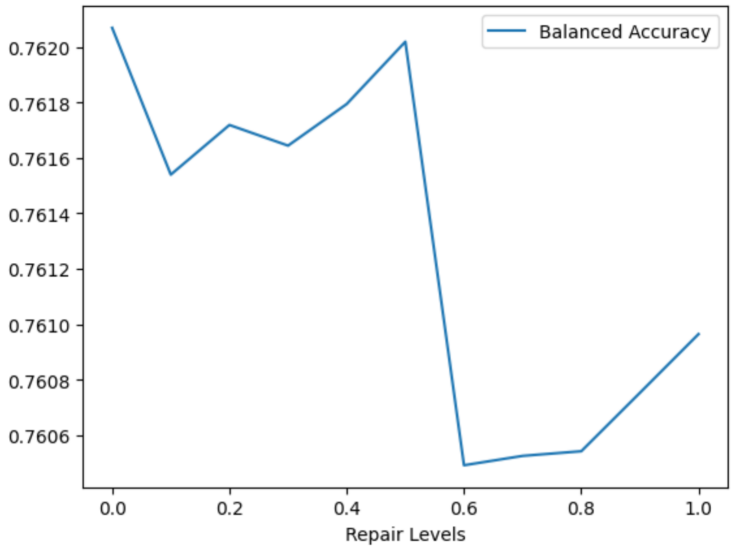

Figure 18: Repair Level vs. Balanced Accuracy for LR After DI Remover

Second, we created a repair level vs. disparate impact plot:

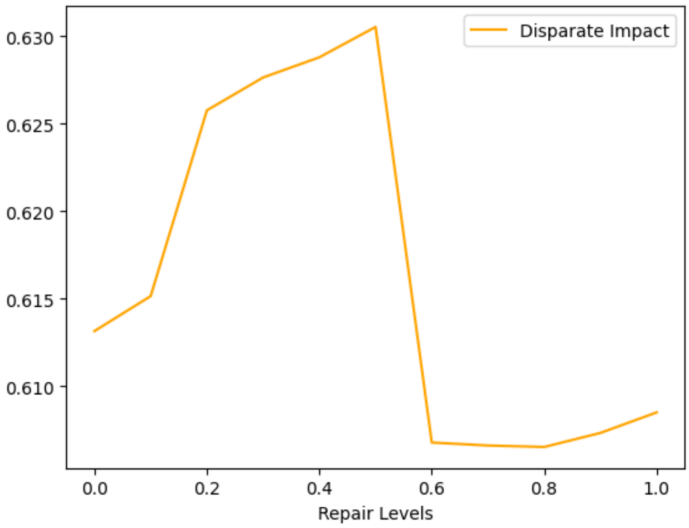

Figure 19: Repair Level vs. Disparate Impact for LR After DI Remover

Random Forest (RF) Model
After developing our random forest model using the disparate impact remover, here were our results:

First, we created a repair level vs. balanced accuracy plot:

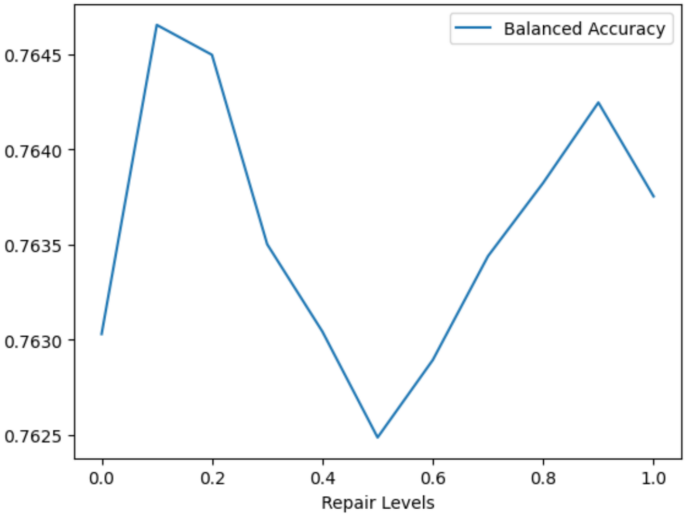

Figure 20: Repair Level vs. Balanced Accuracy for RF After DI Remover

Second, we created a repair level vs. disparate impact plot:

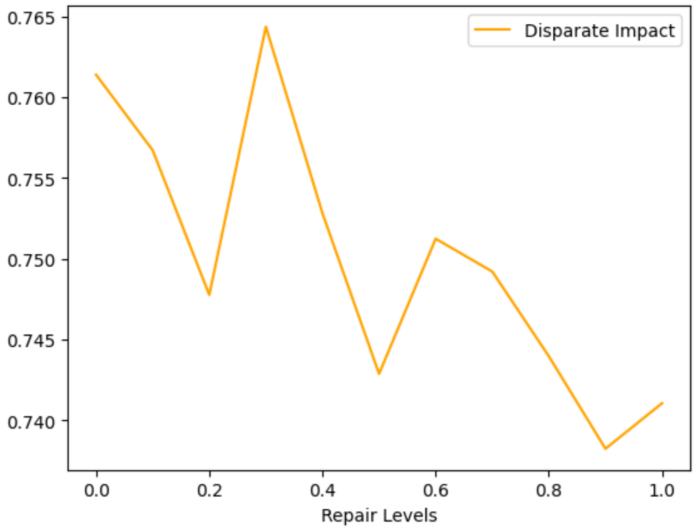

Figure 21: Repair Level vs. Disparate Impact for RF After DI Remover

When comparing the bias mitigation steps of the reweighing process and disparate impact removal process, we saw the greatest success when looking at the fairness metric values with the disparate impact remover. More specifically, for the disparate impact remover, it can be seen that from the plots above, for our random forest model, our highest balanced accuracy would be around 0.7648 and our highest disparate impact would be around 0.764. While not perfect, these values are still relatively close to 1 displaying a moderate level of accuracy and fairness. For our logistic regression model, our highest balanced accuracy would be around 0.762 and our highest disparate impact would be around 0.632. Again, while not perfect, these values are pretty close to 1 displaying a pretty good level of accuracy and fairness and it seems like overall, our random forest model could potentially be a bit better. However, it is difficult to determine one specific model that would be the best since as we saw throughout our prior analysis, both logistic regression and random forest seem to demonstrate trade-offs between fairness and accuracy.

4 Discussion

In our exploration of bias mitigation in healthcare machine learning models, particularly through the utilization of IBM's aif360 toolkit, we encountered both advancements and limitations in our endeavor to refine models to a deployable standard. While aif360 provided an effective framework for identifying and addressing bias, the journey to achieving a deployable level of fairness in our models proved to be complex and multi-dimensional.

The use of aif360's Reweighing and Disparate Impact Remover techniques allowed for significant strides in reducing bias within our logistic regression and random forest models. These methods effectively adjusted the model's learning process, balancing the representation of diverse demographic groups and reducing the disparity in predictions between privileged and unprivileged groups. However, even with these changes, the models did

not begin to approach a level of fairness wherein they could be deployed in a real-world context. A balanced accuracy below 0.8 is not even close enough to trust with the health of real patients.

The persistent presence of residual bias in our models underscores a fundamental challenge in AI-driven healthcare: the inherent complexities of real-world data. Our study illuminated the intricate dynamics of demographic factors and their multifaceted impact on healthcare utilization patterns. The MEPS dataset, rich in detail, presented both an opportunity and a challenge, revealing the depths of inherent biases in healthcare data.

Furthermore, our findings highlight the need for continual evolution in bias mitigation techniques. While tools like aif360 offer a starting point, the pursuit of deployable, fair models necessitates ongoing research, development, and innovation. This includes exploring advanced algorithms, incorporating a wider range of fairness metrics, and integrating domain expertise to ensure a comprehensive understanding of the underlying healthcare dynamics. In addition, when working with critical tasks like patient care, we must consider whether ML is even appropriate at all.

5 Conclusion

In this study, we embarked on a comprehensive exploration of bias mitigation techniques applied to logistic regression and random forest models, utilizing reweighing during the preprocessing stage. This technique, differentiating example weights based on (group, label) combinations, aimed to rectify disparate impact, a measure reflecting the ratio of favorable outcomes between unprivileged and privileged groups.

The evaluation of fairness and accuracy metrics revealed intriguing insights. The logistic regression model demonstrated commendable fairness metrics, boasting a balanced accuracy of 75.44%, a 1-min(DI, 1/DI) value of 0.1153, and an average odds difference of -0.0077. On the other hand, the random forest model exhibited slightly higher accuracy at 75.81%, with a 1-min(DI, 1/DI) value of 0.2511 and an average odds difference of -0.0635. Notably, the elusive balance between accuracy and fairness remained a challenge, with neither model emerging as an ideal 'fair' classifier for predicting 'high' utilization.

Our overarching discussions underscore valuable lessons learned in AI model development. Critical considerations, including representative data collection, thoughtful feature selection, and vigilant monitoring during deployment, have emerged as paramount. Measuring the downstream impact of AI requires a multifaceted approach, incorporating fairness metrics, expert evaluations, transparency, and stakeholder feedback.

The study highlighted the concept of disparate impact, elucidating the potential biases originating from data type, dimension, collection methods, and representation. Furthermore, we recognize the often-overlooked social factors that can impede access and quality of care for underserved populations.

In conclusion, our recommendation for a 'Fair' classifier leans towards logistic regression with adversarial training as the bias-mitigation technique. This decision prioritizes fairness

over a marginal reduction in accuracy, aligning with ethical imperatives to prevent the reinforcement of existing health disparities. The journey through bias mitigation techniques has provided profound lessons, emphasizing the commitment to equitable healthcare outcomes as AI continues to play a pivotal role in shaping our healthcare landscape.

6 References

Chen, I. Y., Pierson, E., Rose, S., Joshi, S., Ferryman, K., Ghassemi, M. (2021). Ethical machine learning in healthcare. *Annual review of biomedical data science*, 4, 123-144.

Chen, Richard J., et al. "Algorithm Fairness in AI for Medicine and Healthcare." Last revised 24 Mar 2022 (v2).

National Academies of Sciences. Factors That Affect Health-Care Utilization. *National Academies Press (US)*; 2019. https://www.ncbi.nlm.nih.gov/books/NBK500097/

LaVeist TA, Pérez-Stable EJ, Richard P, et al. The Economic Burden of Racial, Ethnic, and Educational Health Inequities in the US. *JAMA.* 2023. DOI: 10.1001/jama.2023.

Lavizzo-Mourey, Risa J., Richard E. Besser, and David R. Williams. "Understanding and Mitigating Health Inequities —Past, Current, and Future Directions." *New England Journal of Medicine, vol. 384, no. 18, 2021, pp. 1681-1684.* doi: 10.1056/NEJMp2008628.

Williams, Lawrence A., et al. "Discrimination in Health Care: How Bad Is It?" *Health Affairs, vol. 21, no. 4, 2002, pp. 22-32.*

Appendices

A.1 Training Details

A.2 Additional Figures

A.3 Additional Tables