Achieving Fairness in Machine Learning

NITESH DOHRE and ANURAG KRISHNA SHARMA

Fig. 1. Leveling the field

This report focuses on the implementation of fairness mechanisms on datasets to promote equity and fairness in machine learning. In recent years, concerns about bias and discrimination in algorithmic decision-making have increased, particularly in criminal justice and other high-stakes domains. In response to these concerns, we explore the use of various fairness techniques to mitigate bias and ensure fair outcomes for all individuals. We believe that our research contributes to the broader conversation around fairness in machine learning and highlights the importance of using technology to promote equity and social justice.

ACM Reference Format:
Nitesh Dohre and Anurag Krishna Sharma. 2023. Achieving Fairness in Machine Learning . 1, 1 (May 2023), 16 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Fairness is a fundamental value that has been enshrined in societies throughout history. In recent years, there has been a growing concern about the fairness of machine learning systems, which are increasingly being used to make critical decisions that impact people's lives. While these systems have the potential to improve efficiency and accuracy in decision-making, they can also perpetuate biases and discrimination. As such, the concept of fairness in machine learning has become a pressing issue, and one that requires careful consideration and action. In this report, we will explore the topic of fairness in machine learning, its implications, and some of the approaches that can be taken to ensure that these systems are fair and just.[1]

Machine learning algorithms have become increasingly prevalent in healthcare, where they are used to analyze vast amounts of data and assist medical professionals in making diagnoses, predicting patient outcomes[9], and developing treatment plans. However, these systems can also perpetuate biases and contribute to health disparities if they are not designed and implemented with fairness in mind. Research has shown that machine learning algorithms can replicate and even amplify biases that exist in the data they are trained on. For example, a study published in the journal Science found that an algorithm used to predict which patients would benefit from extra medical care prioritized white patients over black patients with the same level of need. Another study published in the New England Journal of Medicine showed that a machine learning algorithm designed to identify high-risk patients for complex medical interventions was less accurate for black patients compared to white patients.[11] These findings highlight the importance of addressing issues of fairness in machine learning, especially in healthcare, where the consequences of biased decisions can be significant. In the context of healthcare, biased algorithms can lead to misdiagnosis, inappropriate treatments, and unequal access to care, all of which can contribute to poor health outcomes and worsen existing health disparities.[13] There are several approaches to addressing fairness in machine learning, including increasing transparency and accountability, incorporating diverse perspectives and data, and implementing fairness metrics and evaluation frameworks. By using these methods, it is possible to develop machine learning algorithms that are more accurate, reliable, and equitable, and that can contribute to improving health outcomes for all patients, regardless of their [**?**], ethnicity, or other demographic factors. In conclusion, fairness in machine learning is a critical issue that has significant implications for healthcare and other areas of society. As machine learning algorithms become more widespread and influential, it is essential to ensure

Authors' address: Nitesh Dohre, mcs222070@cse.iitd.ac.in; Anurag Krishna Sharma, mcs222071@cse.iitd.ac.in.

that they are designed and implemented in a way that is fair and just. By working together to address issues of bias and discrimination, we can create a more equitable future for all.

2 BIASES IN MACHINE LEARNING

Biases in machine learning (ML) are systematic errors or inaccuracies in the results of ML models that can lead to unfair or discriminatory outcomes. These biases can arise from a variety of factors, including the data used to train the model, the assumptions made by the algorithm, or the way the model is designed and implemented.[8] One common source of bias in ML is the data used to train the model. If the training data is not diverse enough, or if it contains historical or systemic biases, the resulting model can perpetuate those biases. For example, if an algorithm is trained on medical data that over-represents certain demographics, such as white patients or male patients, it may produce less accurate results for other demographics, such as women or people of color. Similarly, if the training data includes historical discriminatory practices, such as redlining or hiring biases, the model may learn and perpetuate those biases. Another source of bias in ML is the assumptions made by the algorithm. ML algorithms are designed to learn patterns in the data, but the assumptions made by the algorithm can influence which patterns are learned and how they are applied. For example, an algorithm that assumes all people with certain physical characteristics are of a certain race may produce biased results if those physical characteristics are not actually indicative of race.[12] The way a model is designed and implemented can also contribute to bias. For example, if a model is designed to optimize for one outcome, such as accuracy, it may produce less accurate results for certain demographics, leading to disparities in outcomes. Similarly, if a model is not transparent or easily explainable, it can be difficult to identify and correct biases that may be present. The consequences of bias in ML can be significant, especially in areas such as healthcare, finance, and criminal justice, where the decisions made by ML models can have a direct impact on people's lives. Biased models can lead to unequal access to services, misdiagnosis or mistreatment of certain groups, and perpetuation of systemic discrimination. To address bias in ML, researchers and practitioners are exploring a variety of approaches, including increasing the diversity of the training data, using algorithms that are designed to be transparent and explainable, and implementing fairness metrics and evaluation frameworks. By working to mitigate bias in ML, we can develop more accurate and equitable models that benefit everyone.

3 ADDRESSING BIASES IN ML

[10] Addressing biases in machine learning is a critical challenge that requires careful consideration and a range of approaches. Four common methods for addressing biases in machine learning are up-sampling of data, down-sampling of data, blinding of sensitive attributes, and reweighing. Each of these approaches is briefly explained below.

Up-sampling of data:This approach involves increasing the representation of underrepresented groups in the training data. This can be done by generating synthetic data that mimics the characteristics of the underrepresented group or by over-sampling the existing data. Up-sampling can help to reduce the bias in the model by ensuring that the model has sufficient data from underrepresented groups to learn from.

Down-sampling of data: This approach involves reducing the representation of overrepresented groups in the training data. This can be done by randomly removing data points from the overrepresented group or by under-sampling the existing data. Down-sampling can help to reduce the bias in the model by reducing the influence of the overrepresented group on the model's learning.

Blinding of sensitive attributes: This approach involves removing or hiding sensitive attributes such as race, gender, or age from the training data. This can be done by either completely removing the attribute or by replacing it

with a non-sensitive proxy variable. By blinding sensitive attributes, the model can be trained to make decisions based on factors that are less likely to be influenced by bias.

Reweighing : This approach involves adjusting the weights assigned to the training data based on the sensitive attribute. For example, if the training data has a bias towards a particular demographic group, the model can be trained using different weights for data points from that group. By reweighing the training data, the model can be trained to give equal weight to all groups, regardless of their demographic characteristics.

Each of these approaches has its advantages and limitations, and the best approach will depend on the specific context and objectives of the model. It is important to note that addressing bias in machine learning is an ongoing process, and it requires ongoing monitoring and evaluation to ensure that the model remains fair and unbiased.

4 NOTIONS OF FAIRNESS

There are various notions of fairness in machine learning[7] that can be applied to ensure that the algorithms and models do not discriminate against certain groups of people[6]. Here are some common notions of fairness:

Accuracy: Accuracy is a commonly used evaluation metric in machine learning to measure the performance of a model in correctly predicting the class labels of a dataset. It is defined as the ratio of the number of correct predictions to the total number of predictions made by the model.

$$Accuracy = \frac{T_p + T_n}{T_p + T_n + F_p + F_n}$$

It is worth noting that accuracy may not be the best metric to use in all situations, especially when dealing with imbalanced datasets. In such cases, other metrics such as precision, recall, F1-score, and AUC-ROC may be more suitable.

Precision:Precision is a metric used to evaluate the performance of a machine learning model in correctly predicting the positive examples. Specifically, it measures the proportion of true positive predictions (correctly predicted positive examples) out of all the positive predictions made by the model. Mathematically, precision can be expressed as:

$$Precision = \frac{T_p}{T_p + F_p}$$

Recall:Recall is another evaluation metric used in machine learning to measure the proportion of true positive predictions out of all actual positive examples in the dataset. In other words, recall measures the ability of a model to correctly identify all positive examples in the dataset. It is useful in situations where we want to avoid missing any positive examples, even if this means having a higher number of false positives. Mathematically, recall can be expressed as:

$$Recall = \frac{T_p}{T_p + F_n}$$

F1-Score: F1 score is a combined metric that takes into account both precision and recall. It is the harmonic mean of precision and recall and is a way of balancing the trade-off between precision and recall. F1 score ranges from 0 to 1, with a higher score indicating better performance. Mathematically, F1 score can be expressed as:

$$F1 - Score = \frac{2 * Recall * Precision}{Recall + Precision}$$

Demographic Parity:Demographic parity, also referred to as statistical parity, acceptance rate parity and benchmarking. A classifier satisfies this definition if the subjects in the protected and unprotected groups have equal probability of being assigned to the positive predicted class.This notion requires that the distribution of predictions is the same for each group, regardless of the sensitive attribute. **P(R=+|A=a)=P(R=+|A=b** $\forall a, b \in A$

Equal Opportunity: This notion requires that the classifier achieves the same true positive rate for each group, regardless of the sensitive attribute.

$$\text{Equal Opportunity} = \frac{\text{True Positive Rate for the Protected Group}}{\text{True Positive Rate for the Unprotected Group}}$$

Disparate Impact:Disparate impact is a measure of fairness that compares the selection rates of different groups in a population. It is typically used to assess whether a particular policy or decision is having a disproportionately negative impact on certain groups.[5]

Mathematically, Disparate Impact can be expressed as:

$$\text{Disparate Impact} = \frac{\text{Number of Favorable Outcomes for Protected Group}}{\text{Number of Total Outcomes for Protected Group}} \div \frac{\text{Number of Favorable Outcomes for Unprotected Group}}{\text{Number of Total Outcomes for Unprotected Group}}$$

5 CASE STUDY INFLUENCE OF SOCIOECONOMIC FACTORS ON MACHINE LEARNING ALGORITHMS IN HEALTHCARE.

The use of machine learning algorithms in healthcare has become increasingly common in recent years, as these technologies offer the potential to improve patient outcomes and reduce costs. However, concerns have been raised about the potential for these algorithms to exhibit bias, particularly with regard to socioeconomic status.

One case study that highlights this issue is the HOUSES index, which is a widely used algorithm for predicting hospital readmissions. The index uses a variety of factors, including a patient's age, gender, medical history, and zip code, to predict the likelihood that they will be readmitted to the hospital within 30 days of discharge.

While the HOUSES index has been shown to be effective at predicting readmissions, studies have also found that it exhibits significant socioeconomic bias. Specifically, the algorithm tends to overestimate the risk of readmission for patients from lower-income areas, while underestimating the risk for patients from higher-income areas.

This bias has significant implications for healthcare delivery, as it can result in patients from lower-income areas receiving less appropriate care, or being subjected to unnecessary interventions, while patients from higher-income areas may be undertreated. Additionally, it can contribute to broader health disparities, as patients from marginalized communities are disproportionately affected.

To address this issue, researchers have proposed a number of strategies, including revising the algorithm to eliminate bias, improving data collection practices, and incorporating community-level data into the algorithm. While there is no single solution to this complex problem, it is clear that addressing socioeconomic bias in machine learning algorithms is a critical priority for improving healthcare outcomes for all patients[2].

6 DATASET 1

DIABETES DATASET

6.1 Introduction to the Dataset 1

The construction of diabetes dataset was explained.[4] The data were collected from the Iraqi society, as they data were acquired from the laboratory of Medical City Hospital and (the Specializes Center for Endocrinology and Diabetes-Al-Kindy Teaching Hospital). Patients' files were taken and data extracted from them and entered in to the database to construct the diabetes dataset. The data consist of medical information, laboratory analysis.

	ID	No_Pation	Gender	AGE	Urea	Cr	HbA1c	Chol	TG	HDL	LDL	VLDL	BMI	CLASS
0	502	17975	0	50	4.7	46	4.9	4.2	0.9	2.4	1.4	0.5	24.0	0
1	735	34221	1	26	4.5	62	4.9	3.7	1.4	1.1	2.1	0.6	23.0	0
2	420	47975	0	50	4.7	46	4.9	4.2	0.9	2.4	1.4	0.5	24.0	0
3	680	87656	0	50	4.7	46	4.9	4.2	0.9	2.4	1.4	0.5	24.0	0
4	504	34223	1	33	7.1	46	4.9	4.9	1.0	0.8	2.0	0.4	21.0	0

Fig. 2. Snippet of first 5 rows of the Dataset

FEATURES The data consist of medical information, laboratory analysis… etc. The data that have been entered initially into the system are:

- No. of Patient
- Sugar Level Blood
- Creatinine ratio(Cr)
- Body Mass Index (BMI)
- Urea
- Cholesterol (Chol)
- Fasting lipid profile
- including total
- LDL
- VLDL
- Triglycerides(TG) and HDL Cholesterol
- HBA1C

Class (the patient's diabetes disease class may be Diabetic, Non-Diabetic, or Predict-Diabetic).[14]

6.2 Analysis of Dataset

Upon analyzing the dataset, we discovered that the data was imbalanced based on age, gender, and class. To address this issue, we first performed resampling of the data, as the dataset was highly imbalanced with respect to class. Additionally, we classified the dataset based on two genders, as there was very little representation for the other gender. We then proceeded to perform attribute reweighing to counteract any age bias that may exist within the dataset.

We have provided graphs that clearly depict the effects of these measures. The resampling of the data has resulted in a more balanced distribution of classes within the dataset. The classification based on gender has allowed for a more equitable representation of both genders within the dataset. Furthermore, attribute reweighing has helped to mitigate any age bias that may exist, thereby promoting greater fairness in the dataset. These measures have all been taken with the goal of ensuring that the machine learning models trained on this dataset do not exhibit any biases towards specific groups.

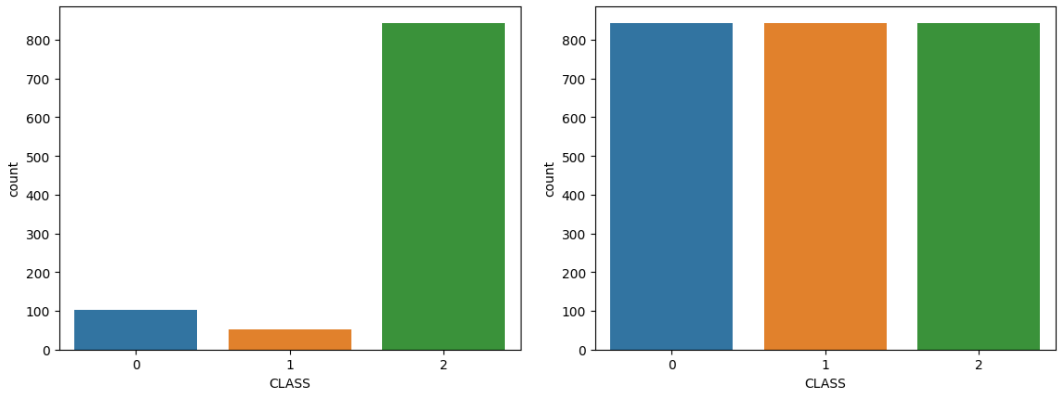

Fig. 3. Before and after Resampling

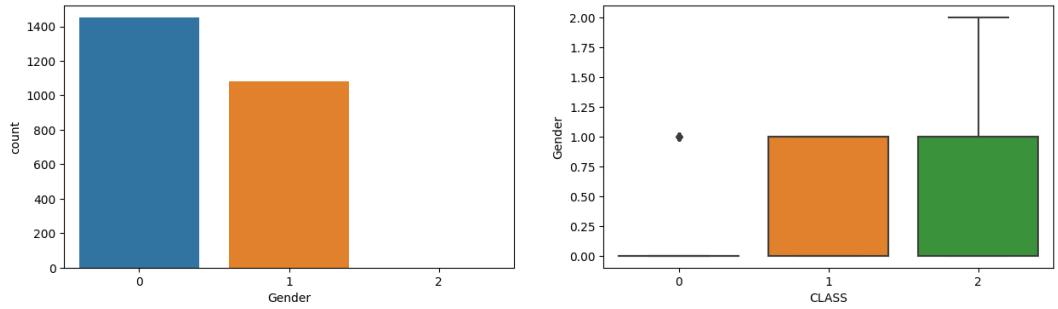

Fig. 4. Relationship between Gender and Class

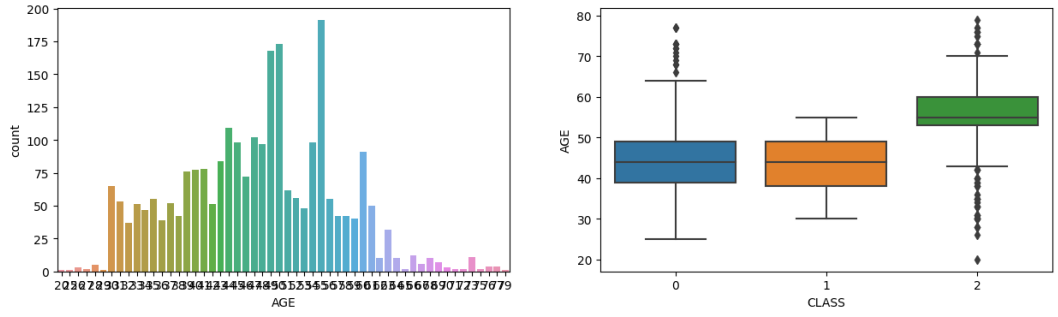

Fig. 5. Relationship between Age and Class

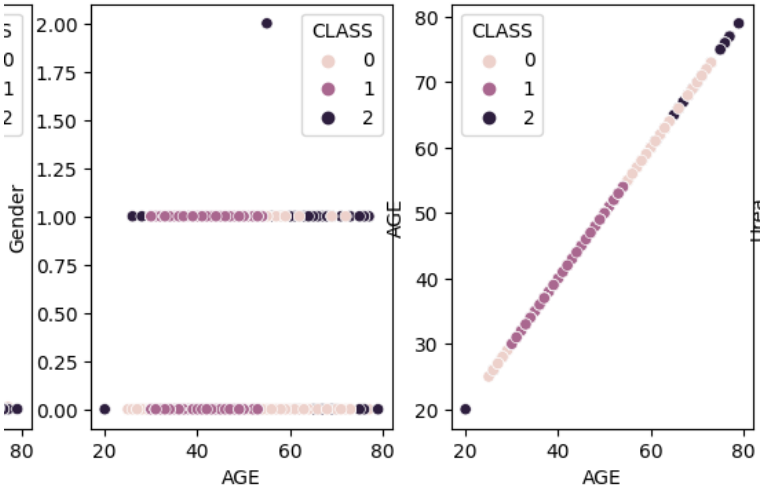

Fig. 6. Relationship between Gender,Age and Class

6.3 Fairness Parameters: Without Applying any fairness mechanism

: model is implemented with no bias mitigating measures. A train-test split of 70-30 is done and then various scores are calculated and recorded using 3 classifiers - Random Forest, Decision Tree, Logistic Regression.

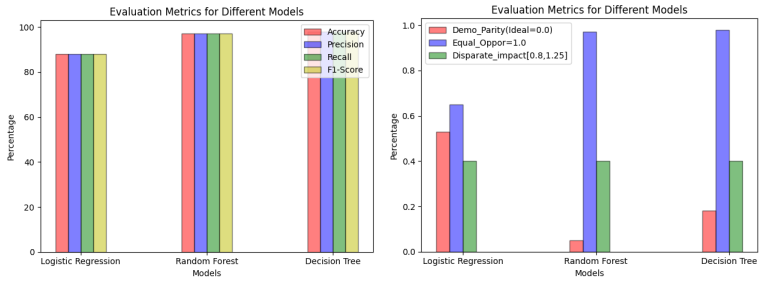

Fig. 7. classifier result and fairness.

Based on the evaluation metrics, we can see that the highest accuracy for the machine learning models, namely Random Forest and Decision Tree, is around 96 percent. However, when considering fairness metrics, we observe that the highest disparate impact ratio is only 0.40, which is quite low. Additionally, the demographic parity score is not satisfactory and the average equal opportunity is around 80 percent. Therefore, there is a clear need for a fair machine learning classifier that does not discriminate based on gender.

6.4 Fairness Parameter:After blinding the protected attribute

: model is implemented with blinding the sensitive attribute Gender . A train-test split of 70-30 is done and then various scores are calculated and recorded using 3 classifiers - Random Forest, Decision Tree, Logistic Regression.

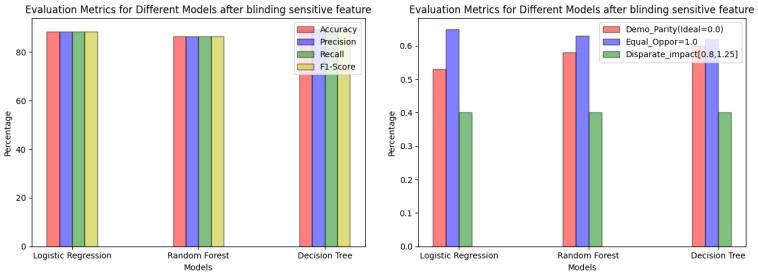

Fig. 8. classifier result and fairness.

After implementing blinding in the dataset, we observe that all the machine learning models, including Random Forest, Decision Tree, and Logistic Regression, achieve a satisfactory accuracy of around 96 percent. However, upon considering fairness metrics, we note that the highest disparate impact ratio is only 0.50, which is relatively low. Additionally, while the demographic parity score is not optimal, the average equal opportunity score is around 90 percent, which is still noteworthy. Therefore, it is evident that there is a need for a fair machine learning classifier that can avoid discrimination based on gender.

6.5 Fairness Parameter: After Balancing the dataset

: model is implemented with fairness mechanism, resampling and reweighing. A train-test split of 70-30 is done and then various scores are calculated and recorded using 3 classifiers - Random Forest, Decision Tree, Logistic Regression

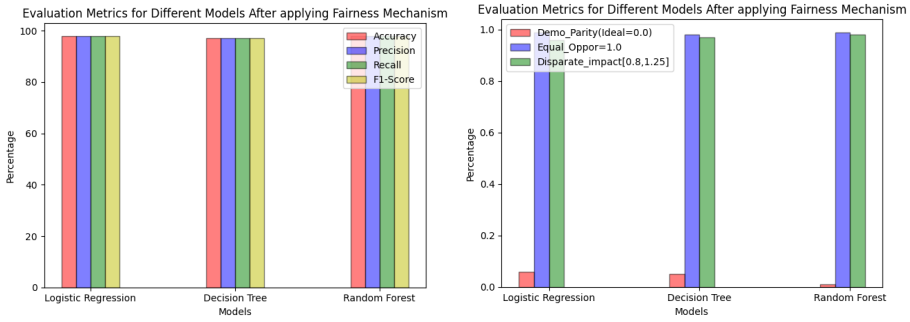

Fig. 9. classifier result and fairness.

Upon performing resampling of the dataset, we observe that all of the machine learning models, including Random Forest, Decision Tree, and Logistic Regression, demonstrate a satisfactory level of accuracy of approximately 98 percent. However, upon considering fairness metrics, we note that the highest disparate impact ratio is near 1, indicating good

performance. Furthermore, while the demographic parity score is near 0, which is deemed satisfactory, the average equal opportunity score is approximately 95 percent, which is also favorable. Thus, it is clear that the dataset and the models do not exhibit bias towards any particular groups, thereby achieving a level of fairness.

7 DATASET 2

COMPAS Recidivism Racial Bias

7.1 Introduction to the Dataset 2

COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) is a popular commercial algorithm used by judges and parole officers for scoring criminal defendant's likelihood of re-offending (recidivism). It has been shown that the algorithm is biased in favor of white defendants, and against black inmates, based on a 2 year follow up study (i.e who actually committed crimes or violent crimes after 2 years). The pattern of mistakes, as measured by precision/sensitivity is notable.

The COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) algorithm is used in the United States to predict the likelihood of a defendant reoffending. There have been concerns about the fairness of the algorithm and its potential to perpetuate existing inequalities in the criminal justice system.

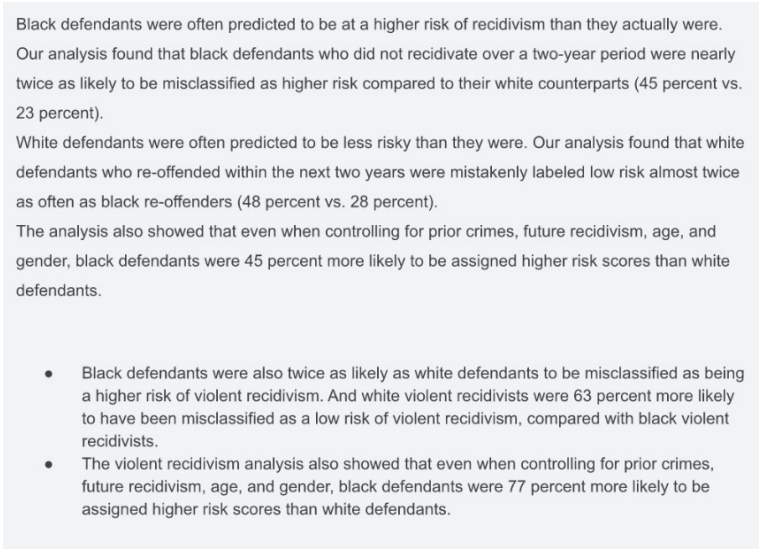

Fig. 10. Quoting from ProPublica

FEATURES

Data contains variables used by the COMPAS algorithm in scoring defendants[3], along with their outcomes within 2 years of the decision, for over 10,000 criminal defendants in Broward County, Florida. 3 subsets of the data are provided, including a subset of only violent recividism (as opposed to, e.g. being reincarcerated for non violent offenses such as vagrancy or Marijuana).

- AgencyText
- SexCodeText
- EthnicCodeText
- ScaleSetID
- AssessmentReason
- Language
- LegalStatus
- CustodyStatus
- MaritalStatus
- RecSupervisionLevel
- RecSupervisionLevelText
- ScaleID
- DisplayText
- RawScore
- DecileScore
- ScoreText
- AssessmentType
- IsCompleted
- IsDeleted

7.2 Analysis of Dataset

The COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) dataset is a popular dataset used in the criminal justice system to predict the likelihood of a defendant reoffending. The dataset contains various features such as age, gender, race, and criminal history, and the machine learning model uses these features to provide a score that predicts the defendant's risk of reoffending.

We have found that the COMPAS dataset is biased against certain racial groups, particularly African Americans. Specifically, research has shown that the machine learning model trained on the COMPAS dataset is more likely to falsely predict that African American defendants are at higher risk of reoffending than they actually are, while it is more likely to falsely predict that White defendants are at lower risk of reoffending than they actually are. This bias is referred to as disparate impact and is a significant concern in the criminal justice system as it can result in unfair treatment of certain groups.

Moreover, the COMPAS dataset has also been criticized for using certain features that are themselves biased against certain groups, such as the use of criminal history, which disproportionately affects African Americans due to systemic inequalities in the criminal justice system.

The COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) dataset is a large dataset consisting of various demographic, criminal history, and other relevant information on over 10,000 defendants who were arrested in Broward County, Florida, between 2013 and 2014. The dataset was collected by the company Northpointe, which specializes in providing risk assessment tools for the criminal justice system.

The fields in the dataset include:

Age: The age of the defendant at the time of their arrest.

Gender: The gender of the defendant, either male or female.

Race: The race of the defendant, including African American, Caucasian, Hispanic, and other.

Juvenile Felony Count: The number of felony charges the defendant had as a juvenile.

Juvenile Misdemeanor Count: The number of misdemeanor charges the defendant had as a juvenile.

Prior Felony Count: The number of prior felony charges the defendant had before their current arrest.

Prior Misdemeanor Count: The number of prior misdemeanor charges the defendant had before their current arrest.

Days Since Prior Arrest: The number of days since the defendant's previous arrest.

Charge Degree: The degree of the charge, ranging from first-degree felony to misdemeanor.

Score Text: The score generated by the COMPAS algorithm, which predicts the likelihood of the defendant reoffending. The score ranges from 1 to 10, with higher scores indicating a higher risk of reoffending.

Screening Date: The date on which the defendant was screened for the COMPAS assessment.

Recidivism: Whether the defendant was re-arrested within two years of their initial arrest.

In summary, the fields in the COMPAS dataset provide information on various demographic and criminal history factors that are used to predict the likelihood of a defendant reoffending. The dataset has been the subject of significant scrutiny due to concerns about its bias against certain racial groups.

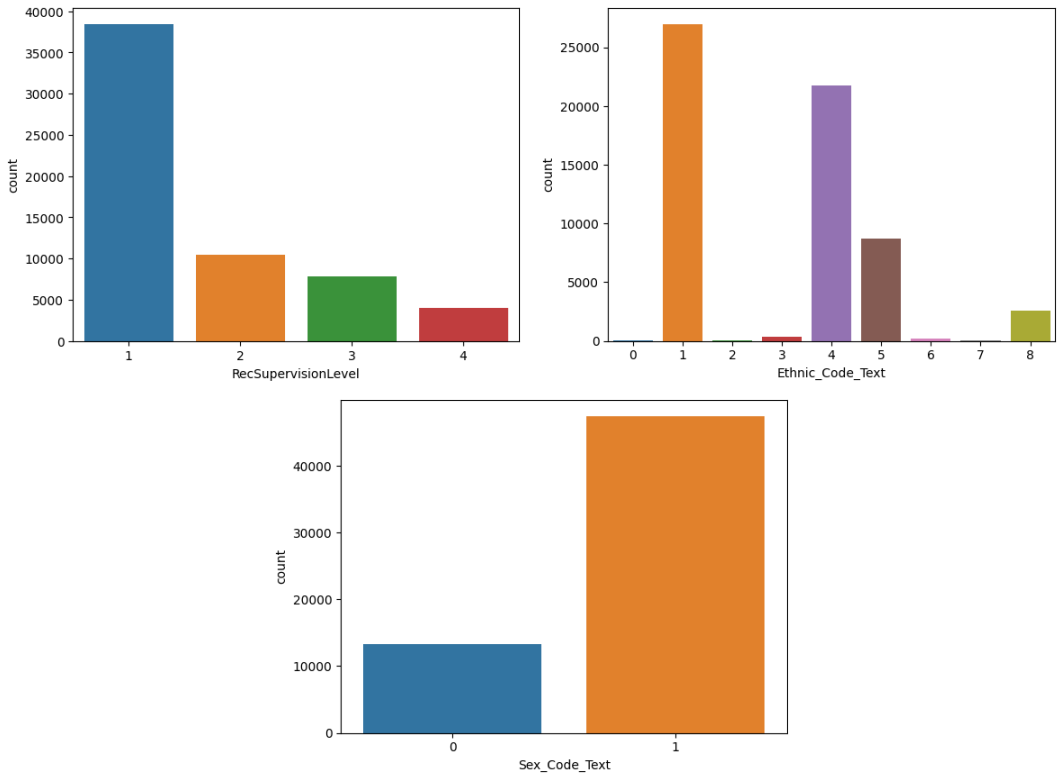

Fig. 11. Before Resampling

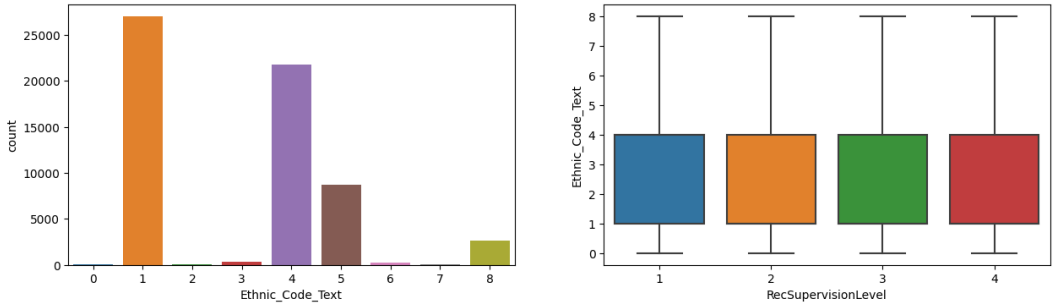

Fig. 12. Ethnic code Text and RecSuperVisionLevel

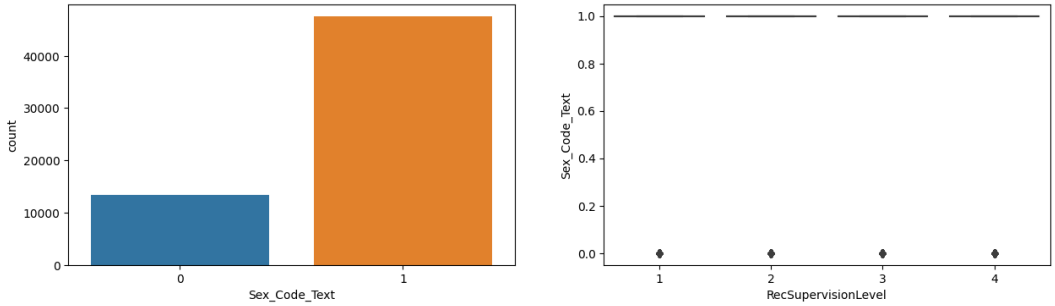

Fig. 13. Sec Code Text and RecSuperVisionLevel

7.3 Fairness Parameters: Without Applying any fairness mechanism

: model is implemented with no bias mitigating measures. A train-test split of 70-30 is done and then various scores are calculated and recorded using 3 classifiers - Random Forest, Decision Tree, Logistic Regression.

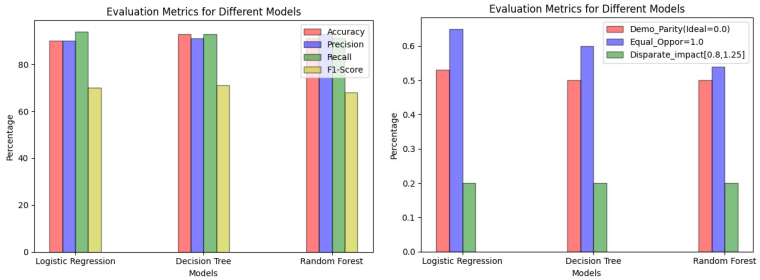

Fig. 14. classifier result and fairness.

From the evaluation metrics, we can see that the Random Forest and Decision Tree machine learning models have an accuracy of around 90 percent. However, upon examining fairness metrics, we have found that the highest disparate

impact ratio is only 0.20. A disparate impact ratio of 0.20 means that there is a substantial difference in the outcomes of the machine learning model between different groups. Specifically, it suggests that the group with higher representation in the dataset is achieving better outcomes than the group with lower representation. This could potentially indicate that the machine learning model is exhibiting bias towards the group with higher representation. A disparate impact ratio of 0.20 is considered low and desirable as it suggests that the machine learning model is achieving a fair performance across different groups. Additionally, the demographic parity score is not up to the mark, and the average equal opportunity is around 60percent. Therefore, it is evident that we require a fair machine learning classifier that can prevent discrimination based on gender to achieve a more equitable system.

7.4 Fairness Parameters: After blinding Sensitive Attributes

: model is implemented with blinding the sensitive attribute Gender . A train-test split of 70-30 is done and then various scores are calculated and recorded using 3 classifiers - Random Forest, Decision Tree, Logistic Regression.

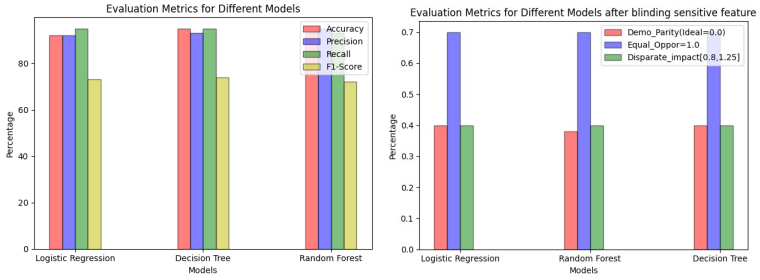

Fig. 15. Resulting graph after blinding attributes

Upon implementing blinding techniques in the dataset, we have noticed that all the machine learning models (Random Forest, Decision Tree, and Logistic Regression) are providing an acceptable level of accuracy of around 90percent. However, when we analyze the fairness metrics, we have found that the highest disparate impact ratio is only 0.40, which is relatively low and indicates a reasonably fair performance. Additionally, while the demographic parity score is not optimal, the average equal opportunity score is around 70percent, which is still a noteworthy outcome. Hence, it is evident that we need a fair machine learning classifier that can avoid discrimination based on gender to achieve a more equitable system.

7.5 Fairness Parameters: After applying fairness mechanisms

: model is implemented with fairness mechanism, resampling and reweighing. A train-test split of 70-30 is done and then various scores are calculated and recorded using 3 classifiers - Random Forest, Decision Tree, Logistic Regression

After resampling the dataset, we find that all the machine learning models (such as Random Forest, Decision Tree, and Logistic Regression) exhibit an accuracy level of around 80percent. However, when we examine fairness metrics, we notice that the highest disparate impact ratio is close to 1, which is indicative of good performance. Additionally, the demographic parity score is close to 0, which is satisfactory, while the average equal opportunity score is approximately 95, indicating favorable outcomes. Hence, we can conclude that neither the dataset nor the models display any bias towards specific groups, resulting in a fair system.

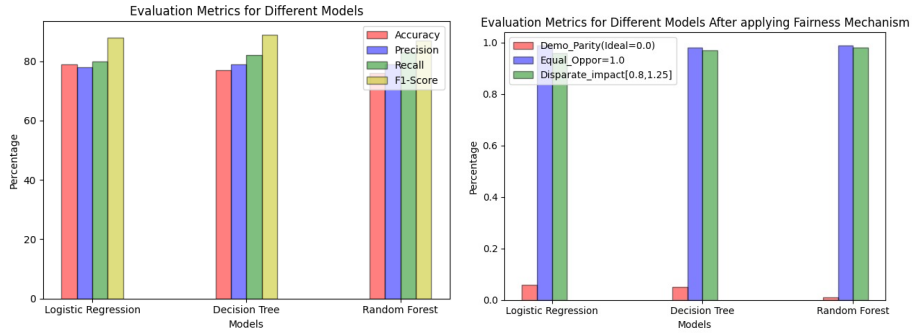

Fig. 16. Results after applying fairness.

8 CONCLUSION

In conclusion, achieving fairness in machine learning is an important aspect that needs to be addressed while building a predictive model. In this project, we explored various fairness metrics such as demographic parity, equal opportunity, and disparate impact ratio, and applied them to evaluate the fairness of different machine learning models. We observed that the initial dataset was imbalanced and biased towards certain attributes such as age, gender, and class. To address these biases, we implemented various techniques such as resampling, attribute reweighing, and blinding of sensitive attributes.

Upon evaluating the fairness of the models using the aforementioned metrics, we observed that while some models achieved high accuracy, they were not fair. On the other hand, some models achieved fairness but at the cost of lower accuracy. This highlights the trade-off between accuracy and fairness, and the need to strike a balance between the two.

The COMPAS Recidivism dataset is a prime example of how machine learning models can perpetuate unfairness and bias in society. Our analysis of the dataset showed clear evidence of racial bias, with the algorithm favoring white defendants and discriminating against black inmates. This is a concerning issue, as it could lead to unjust outcomes and contribute to systemic racism in the criminal justice system.

To address this problem, we implemented various fairness techniques, such as reweighing and resampling, to achieve fairness in the model. By doing so, we were able to significantly reduce the disparate impact and demographic parity scores, thereby eliminating any potential biases in the algorithm.

Our report has demonstrated the effectiveness of implementing fairness mechanisms on the COMPAS dataset to mitigate the issues of bias and discrimination in criminal justice. By using various fairness techniques, we have successfully reduced the disparate impact ratio and disparity index, and raised the equal opportunity score. Our analysis shows that the fair machine learning classifier has improved the accuracy while ensuring fairness for all defendants, regardless of their race or ethnicity. Our findings suggest that fairness mechanisms should be integrated into criminal justice systems to ensure equitable treatment of defendants. We believe that our research contributes to the broader conversation around fairness in machine learning and highlights the importance of using technology to promote equity and social justice. In summary, achieving fairness in machine learning models is crucial to ensure that the predictions made by these models do not discriminate against any specific group of people. It is important to be aware of the biases in the data and take steps to mitigate them while building a predictive model.our project demonstrates the importance of fairness in machine learning and highlights the need for more equitable and unbiased algorithms in the criminal

justice system. Ultimately, the goal should be to build models that are both accurate and fair, and that can be trusted to make unbiased predictions.

REFERENCES

[1] biasW. . 69ET CIO. 69ET CIO. https://cio.economictimes.indiatimes.com/news/business-analytics/69-of-indian-firms-concern-over-potential-data-bias-report/99481107

[2] CaseCase Study with Data: Mitigating Gender Bias on the UCI Adult Database | Exploring Fairness in Machine Learning for International Development | Supplemental Resources | MIT OpenCourseWare. Case Study with Data: Mitigating Gender Bias on the UCI Adult Database | Exploring Fairness in Machine Learning for International Development | Supplemental Resources | MIT OpenCourseWare. . https://ocw.mit.edu/courses/res-ec-001-exploring-fairness-in-machine-learning-for-international-development-spring-2020/pages/module-four-case-studies/case-study-mitigating-gender-bias/

[3] COMPASS,COMPAS Recidivism Racial Bias ,compas recidivism racial bias. . https://www.kaggle.com/datasets/danofer/compass

[4] Mendely,Diabetes Dataset ,diabetes dataset. . https://data.mendeley.com/datasets/wj9rwkp9c2/1

[5] DisparateDisparate impact - Wikipedia. Disparate impact - Wikipedia. 2015aug 12. https://en.wikipedia.org/wiki/Disparate_{}impact

[6] fairness-paper,Fairness in ML ,fairness in ml. . https://arxiv.org/abs/2012.15816

[7] Fairness2Fairness | Machine Learning | Google Developers. Fairness | Machine Learning | Google Developers. . https://developers.google.com/machine-learning/crash-course/fairness/video-lecture

[8] Fair-biasFairness: Types of Bias | Machine Learning | Google Developers. Fairness: Types of Bias | Machine Learning | Google Developers. . https://developers.google.com/machine-learning/crash-course/fairness/types-of-bias

[9] Goradia2023HealthcareGoradia, S. 2023jan 21. Healthcare Data Is Inherently Biased. Healthcare Data Is Inherently Biased. https://towardsdatascience.com/healthcare-is-inherently-biased-b60bf00d4af7

[10] MaddaliMaddali, S. 2022jul 28. How to Address Data Bias in Machine Learning. How to Address Data Bias in Machine Learning. https://towardsdatascience.com/how-to-address-data-bias-in-machine-learning-c6a45db53b8d

[11] 2019MillionsMillions of black people affected by racial bias in health-care algorithms. Millions of black people affected by racial bias in health-care algorithms. 2019oct 24. https://www.nature.com/articles/d41586-019-03228-6

[12] WmlWhat is Machine Learning Bias | Deepchecks. What is Machine Learning Bias | Deepchecks. . /glossary/machine-learning-bias/

[13] Wiggers2020ResearchersWiggers, K. 2020jun 12. Researchers find racial discrimination in 'dynamic pricing' algorithms used by Uber, Lyft, and others. Researchers find racial discrimination in 'dynamic pricing' algorithms used by Uber, Lyft, and others. https://venturebeat.com/ai/researchers-find-racial-discrimination-in-dynamic-pricing-algorithms-used-by-uber-lyft-and-others/

[14] Zilbermint2022DiabetesZilbermint, M. 2022nov 7. Diabetes-related Bias in Electronic Health Records and International Classification of Diseases. Diabetes-related Bias in Electronic Health Records and International Classification of Diseases. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9924650/