UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE INFORMÁTICA
CURSO DE CIÊNCIA DA COMPUTAÇÃO

DIEGO DIMER RODRIGUES

Assessing pre-training bias in Health data and estimating its impact on machine learning algorithms

Work presented in partial fulfillment of the requirements for the degree of Bachelor in Computer Science

Advisor: Profa. Dra. Mariana Recamonde Mendoza

Porto Alegre
April 2023

ACKNOWLEDGEMENTS

I am forever grateful to my advisor, Profa. Dra. Mariana Mendoza, who has helped me through this journey and shared her knowledge with me. She was there when I wrote my first line of code, and my interest in the machine learning topics is greatly due to her classes. I am also grateful to all the teachers at INF-UFRGS, who had participated in my journey as an undergraduate student. I would also like to express my sincere gratitude to my parents and my siblings, they have been there for me every time I needed and were nothing but supportive when I decided my career, supporting my every decision since. If it was not for your sacrifices and examples, I would not have completed my graduation. To my colleagues at SAP, who have provided the support and time I needed to step away from work to conclude this work, and contributed with many valuable discussions. To Rodrigo, who has listened to me talking about almost every line I wrote in this work, and have not really understood many of them, I am sure you will become a skilled computer scientist one day.

ABSTRACT

Machine learning (ML) is a rapidly growing field of computer science that has found many fruitful applications in several domains, including Health. However, ML is also highly susceptible to bias, which introduces concerns regarding their ability to inflict harm. Bias can come from various sources, such as the design of the algorithm, the selection of data, and the strategies underlying data collection. Thus, data scientists must be vigilant in ensuring that the developed models do not perpetuate social disparities based on gender, religion, sexual orientation, or ethnicity. This work aims to explore pre-training bias metrics to investigate the existence of bias in Health data. The metrics also analyze how protected attributes and their correlated features are distributed for the predicted class against the target attributes, giving insight into how the trained model may produce biased predictions. Our goal is to evaluate pre-training bias metrics in three different health datasets and assess the impact of bias on the performance of ML algorithms. Our experiments involve artificially modified versions of the dataset to increase the values of the pre-training bias metrics to favor privileged classes as well as to lower the values of these metrics to reduce the discrepancy in the data and the risk of bias. We trained models using four supervised learning algorithms: Logistic Regression, Decision Tree, Random Forest, and K-Nearest Neighbors. Each algorithm was tested on six to ten different training sets with varying random seeds to split the data in each iteration. We evaluated the performance of the trained models using the same test sets for every dataset variation, reporting the Accuracy and F1-Score. By analyzing pre-training metric bias and the predictive performance of models, this study demonstrates that performance can be significantly affected by skewed data distribution and that the performance metrics may sometimes mask the bias incorporated by the algorithm. In some cases, classification errors may be more pronounced in one group (*e.g.,* the disadvantaged group), accentuating specific errors such as false positives and false negatives, which may have different implications depending on the clinical prediction problem under analysis

Keywords: Machine learning. bias. pre-training bias metrics. model evaluation. health.

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

ACM FAccT	Association for Computing Machinery Conference on Fairness, Accountability, and Transparency
CI	Class Imbalance
CDDL	Conditional Demographic Disparity in Labels
DT	Decision Tree
FN	False Negative
FP	False Positive
KL	Kullback-Leibler
KNN	K-Nearest Neighbors
KS	Kolmogorov-Smirnov
ML	Machine Learning
TN	True Negative
TP	True Positive

CONTENTS

1 INTRODUCTION

Machine learning (ML) is an increasingly popular field of computer science that is attracting a lot of interest from both students and professionals. As the volume of collected data grows, it becomes more difficult to extract insights and detect and analyze patterns manually or with traditional statistical methods, making ML algorithms indispensable tools in many domains. In Health, in particular, in which data has always been a central part of decision-making, we have witnessed a growing increase in the application of ML algorithms to assist in the definition of diagnosis, prognosis or treatment (GHASSEMI et al., 2020; MIOTTO et al., 2018). Applications of ML algorithms have become abundant in most clinical fields and are expected to grow steadily due to the increasing availability of complex health datasets, even motivating the development of guidelines specifically aimed at communicating the development and results of ML models in Health (STEVENS et al., 2020; NAVARRO et al., 2022).

However, the use of ML algorithms also introduces concerns regarding their ability to inflict harm, especially in sensitive domains such as Health (CHEN et al., 2021). As the algorithms become more and more sophisticated, it also becomes more difficult to understand exactly how the input data is being handled and how the model is learning from it. Thus, a major concern, and also a current challenge in the field, is how to guarantee that the model is not introducing or perpetuating social or historical context related to the domain that can cause the system to be subject to systematic errors in their ability to classify subgroups of patients, estimate risk levels, or make predictions, especially in a way considered to be unfair (CHEN et al., 2021). This phenomenon of systematic errors is commonly referred to as bias, meaning that the decisions made by the algorithm are skewed toward a particular group of people (MEHRABI et al., 2019).

Many real-world examples of how bias negatively affected the decisions of the ML algorithms and perpetuated social health disparities already exist. For instance, Obermeyer et al. (2019) analyzed an algorithm widely used by US Hospitals and insurers to allocate health care to patients and identified that the algorithm has been systematically discriminating against black people. Authors found that the algorithm was less likely to refer black people than white people who were equally sick to programms that aim to help patients with complex health needs. Lower social economic status was also associated with worse predictive model performance in the study conducted by Juhn et al. (2022). According to the authors, this may be due to the fact that people with lower so-

cial economic status usually have less comprehensive medical registers due to less access to health care resources, which in turn may impact on the models since their predictive success relies on data completeness and quality.

Thus, analyzing the existence of bias in the data allow scientists to identify gaps in data collection and in the representation of specific groups of individuals and anticipate how this bias can perpetuate to training and test sets and affect ML model development. Moreover, bias analysis is also a useful resource to explain why a model is performing in a specific way and probably different from expected for a particular group of individuals with common characteristics.

One of the approaches to investigate bias is using pre-training bias metrics. If we look deeper into the data that will be used for model development before training takes place, we can evaluate the existence of pre-training bias with specific metrics, which will give us an overview, or at least some insights, on how our data carries the information about the population we are analyzing and how fair it seems to be regarding all the possible subgroups. The ideal scenario is to have a data that provides enough and equal representation for the different subgroups that exist in the population. In health-related applications, where a wrong diagnostic might cause harm to the individual[1], it is essential to expend extra time inspecting the data and seeking for potential biases, and understanding how the underlying patterns are influencing the developed model to avoid errors that might lead to critical and harmful issues.

Thus, this work aims to evaluate the risk of bias in Health datasets using pre-training bias metrics and to assess the impact of pre-training bias in the performance of supervised machine learning algorithms. We analyze four pre-training bias metrics over three different datasets collected from different sources and related to different prediction tasks. We train prediction models with four different ML algorithms, computing and reporting the pre-training bias metrics from the original dataset and also from two variations created artificially: one intended to aggravate the existing bias and penalize more the unprivileged or underrepresented groups, and the other intended to mitigate the existing bias by manually balancing the distribution of the protected attribute (*i.e.,* the sensitive characteristics, such as race, gender, age, etc). We run multiple rounds of model training and testing for each dataset, and report the average accuracy of F1-Score for each model, as well as the values for the pre-training bias metrics calculated over the dataset variations.

Our goal is to show that in a real scenario, it is possible to evaluate the pre-training

[1] https://www.nytimes.com/2022/12/15/health/medical-errors-emergency-rooms.html

bias metrics before deploying the model, and that this analysis can provide important insights about how the model trained upon this data will perform, specially for the historically unprivileged groups. We concluded that the pre-training bias metrics provided important information for two out of the three scenarios investigated, and that the risk for poor performance was confirmed by the results from the performance evaluation of models. We also observed the importance of a correct understanding of the domain and how choosing the right protected attributes to analyze can provide valuable information.

The remainder of this work is organized as follows. Chapter 2 provides the theoretical foundations of the work, defining the notion of bias and protected attributes, and presenting the pre-training bias metrics and the supervised machine learning algorithms that are used. Chapter 3 reviews previous works that aimed at identifying, mitigating or acknowledging bias in ML models, with special interest in the Health domain. Chapter 4 explains the methodology applied in this work, presenting the details about data collection, the modifications applied to data in order introduce or reduce bias artificially, and the steps involved in model training and validation. Chapters 5 reports and discusses our experimental results for the three datasets analyzed in this work. Finally, Chapter 6 summarize our findings and conclusions and points out directions for future works.

2 THEORETICAL BACKGROUND

The use of machine learning (ML) algorithms is increasingly present in our lives. From movie recommendations to clinical prediction in healthcare, machine learning models are used to detect potential issues and guide humans in their decisions. As the use of machine learning models increases, especially in sensitive domains such as Health, it raises the concern regarding its ability to inflict harm due to biased data used to train the models (CHEN et al., 2021). In this chapter, we present the definition of bias and how it can be identified and classified. Then, we review some key concepts and the definition of the pre-training bias metrics that are used in this work, along with an overview of each ML algorithm used.

2.1 Definition of bias and protected attributes

Bias refers to an inclination or prejudice for or against one person or group, especially in a way considered to be unfair. In machine learning (ML), bias occurs when an algorithm makes systematic errors due to faulty assumptions during the model development cycle, leading to discriminatory or unjust decisions. Bias can exist in many shapes and forms. In this work, we focus on bias that arises in the data rather than in the algorithm, as existing biases in the training data can affect algorithms, leading to biased outcomes and ultimately harming specific groups.

According to Mehrabi et al. (2019), biases that stem from data to algorithms can be categorized into the following types:

- *Measurement Bias*: This bias is related to the selection, utilization, and measurement of particular features. The use of unrelated features as proxies to determine the outcome of a model may exclude certain minorities based on attributes[1] that do not define their experience.
- *Omitted Attribute Bias*: This bias arises when one or more essential attributes are left out of the model.
- *Representation Bias*: This bias is related to how a sample is drawn from a population during data collection. The under-representation of some groups leads to inaccurate predictions due to a lack of information about individuals in these groups. It

[1] We note that the words attributes and features are used interchangeably throughout the text.

may be caused by external factors (*e.g.,* lack of diversity in the population used) or the data collection method.

- *Aggregation Bias*: arises from false conclusions drawn about individuals from observing the entire population. When a model defines the outcome based on the entirety of the population, not taking into consideration factors that might differentiate the individual from the majority of the population, like gender and ethnicity.
- *Sampling Bias*: Similar to representation bias, this bias arises from non-random sampling of subgroups. The consequence of this type of bias is that the model may not be general enough to work for a new population.
- *Longitudinal Data Fallacy*: This bias arises when researchers analyze temporal data and fail to use longitudinal analysis to track cohorts over time. Cross-sectional analysis, which combines cohorts at a single time point, may lead to bias compared to the longitudinal analysis.
- *Linking Bias*: This bias arises when network attributes are obtained from user connections, activities, or interactions that differ and misrepresent the true behavior of the users.

In Suresh and Guttag (2019), bias is defined as having five different sources:

- *Historical*: Bias can be caused by historical reasons as the data may be collected at a point in time that does not reflect the current state of society, leading to biased results.
- *Representation*: This bias arises from a lack of representation from a part of the input space, similar to what was described in Mehrabi et al. (2019).
- *Measurement*: This bias arises from how data is measured, where sometimes proxies for ideal features and labels are used. This can occur in several ways, such as the granularity of data across different groups, the quality of data, and if the defined classification is an oversimplification.
- *Aggregation*: This bias arises from the assumption that the mapping from inputs to labels can be inconsistent across different groups.
- *Evaluation*: This bias arises when the benchmark data for an algorithm does not represent the target population.

Another important concept in investigating bias is the notion of protected attributes. These attributes, such as race, gender, and religion, are crucial components of a dataset, and models should produce equitable outcomes for all groups. The determination

of which attributes are protected is application-specific, and there is no universal rule for identifying them. Depending on the application, it is possible to determine which group is privileged and which is unprivileged.

2.2 Pre-training metrics for measuring bias

In this work, we use four pre-training metrics to assess bias. The definition and formulation for these metrics are extracted from Hardt et al. (2021). In this section, we give an overview of the pre-training metrics for measuring bias, with some examples for each metric. The following notations are used throughout the definitions:

- facet a represents the feature value that defines a demographic that bias favors (*i.e.,* the overrepresented or advantaged group)
- facet d represents the feature value that defines a demographic that bias disfavors (*i.e.,* the underrepresented or disadvantaged group)

2.2.1 Class Imbalance (CI)

When an attribute has little representation of a specific class or category, there can be bias for the underrepresented group. This applies not only for the target feature (*i.e.,* the standard class imbalance problem), but also for predictive features used in the modeling process. For example, if 80% of a data set has gender MALE, when predicting for a new entry with gender FEMALE, the model might not be able to generate an accurate prediction due to data bias introduced while training. In the equation, 2.1, n_a represents the number of values in facet a and n_d represents the number of values in facet d. CI values range from -1 to 1. Positive values mean that the facet a has more representation than facet d, with 1 meaning data only contains value on facet a. Negative values mean the data has more representation for facet d, -1 meaning only values in facet d. An ideal CI value is near zero, which means that both groups are equally represented in the data set.

$$CI = (n_a - n_d)/(n_a + n_d) \tag{2.1}$$

2.2.2 Kullback-Leibler (KL) Divergence

Kullback-Leibler (KL) Divergence measures the divergence between the label distributions for two facets, a and d, represented by $P_a(y)$ and $P_d(y)$, respectively. However, it is important to note that KL Divergence is not a true distance metric, as it is asymmetric and does not satisfy the triangle inequality. The most commonly used implementation involves using natural logarithms, resulting in KL Divergence being measured in nats.

To illustrate the concept, let's consider a dataset related to predicting heart disease. Suppose that for the female facet (facet d), 80% of individuals have a risk of heart disease. On the other hand, for the male facet (facet a), only 10% of individuals have a risk of heart disease. Since the favorable outcome is not having heart disease, facet a would have 90% for $P_a(y)$, while facet d would have 20% for $P_d(y)$. Using the equation for KL Divergence (Equation 2.2), we can calculate the KL divergence between the two distributions

$$KL(P_a||P_d) = \sum_{Y} P_a(y) * log[P_a(y)/P_d(y)] \tag{2.2}$$

which, applied to our example, would result in the Equation 2.3

$$KL = 0,9 * Ln(0,9/0,2) + 0,1 * Ln(0,1/0,8) \tag{2.3}$$

Since the label is binary, we would have two terms in the formula, but KL divergence also works when the output is not binary, for example in the scenario of college admissions, where the outcome can be Rejected, Wait-listed or Accepted. In this case, the formula would have three terms, but the protected attribute will always be binary (it has the privileged and unprivileged classes, general cases can have some classes as privileged and the others as unprivileged). The range of values for this metric is between 0 and $+\infty$, with a value near zero meaning the outcomes are similarly distributed for different facets, and a positive value means divergence, with the larger the value, the bigger the divergence. In the example shown in Equation 2.3, the outcome would be 1.145, meaning that for individuals with gender female, there is a divergence in distribution comparing to the outcomes of individuals from the gender male.

2.2.3 Kolmogorov-Smirnov (KS)

This metric is equal to the maximum divergence between labels in the distribution for different facets of a dataset. It finds the most unbalanced label. The formula is as follows, where $P_a(y)$ is the number of members in facet a and $P_d(y)$ refers to the members in facet d:

$$KS = max(|P_a(y) - P_d(y)|) \tag{2.4}$$

This metric can be used for multicategorical targets as well, since the number of terms in the equation is related to the number of possible values for target y. As an example, for the problem of predicting heart disease, in a case where 20% of women would have risk of heart disease (facet d), and 30% of men would present the same risk, using the formula in 2.4 we can calculate the KS as in Equation 2.5.

$$KS = max(|0.3 - 0.2|, |0.7 - 0.8|) \tag{2.5}$$

KS results in values ranging between [0, +1], where values near zero indicate that the labels are evenly distributed between the facets (in this example, if both men and women had a 20% chance of having heart disease, the terms would all be 0). Values near one indicate that the labels are unbalanced, with one meaning all outcomes are in one facet (100% of women with heart disease and 0% of men with heart disease, resulting in $1 - 0$ in the formula term).

2.2.4 Conditional Demographic Disparity in Labels (CDDL)

This metric aims to provide information about the proportion of negative outcomes in a specific facet of a dataset. The formula for this metric can be seen in Equation 2.6, where n in the number of observations in the dataset, and i is for the different outcomes in the correlated attribute.

$$CDDL = \frac{1}{n} * \sum_i n_i * DD_i \tag{2.6}$$

where DD_i is defined according to Equation 2.7.

$$DD_i = \frac{n_d^{(0)}}{n^{(0)}} - \frac{n_d^{(1)}}{n^{(1)}} \tag{2.7}$$

For example, for the college admissions case, let's assume that in a dataset with 20 observations, 10 men and 10 women, 50% of each group is more than 20 years old. For men over 20 years old, 80% of them are accepted (4 men accepted, 1 man rejected), and for the others the acceptance rate is 40% (2 approved, 3 rejected). For women over 20 years old, 60% are accepted (3 accepted, 2 rejected) and for women below 20 years 20% are accepted (1 accepted, 4 rejected). Using the formulas in Equation 2.6 and Equation 2.7, we have the result in Equation 2.8.

$$DD_i = \frac{1}{20} * 10 * (\frac{4}{7} - \frac{1}{3}) + 10 * (\frac{2}{3} - \frac{3}{7}) = 0.238 \tag{2.8}$$

The result for this metric ranges between [-1, +1], with 1 meaning there is no rejection in facet a or subgroup and no acceptance in facet d or subgroup (in the example, subgroup means being over 20 years old). Positive values indicate there is a demographic disparity. In our example, it indicates that the rate at which man over 20 years old are accepted are higher than the acceptance rates for women in the same subgroup. On the other hand, negative values indicates no demographic disparity for the facet d or subgroup, as they have higher acceptance rates than the privileged group, with -1 meaning no rejections on the unprivileged class or subgroup and no acceptance on the privileged group or subgroup.

2.3 Supervised learning algorithms

In this work, we use four supervised learning algorithms: Logistic Regression, Decision Tree, Random Forest, and K-Nearest Neighbors. This section provides an overview of each algorithm applied. Our methodology is based on algorithms implementations provided by Scikit-learn library (PEDREGOSA et al., 2011), an open source machine learning library for Python.

2.3.1 Logistic Regression

Logistic regression is a simple and efficient classification model for both binary and linear classification problems. It is also known as logit regression, maximum-entropy classification, or the log-linear classifier. This model uses a logistic function to model the probabilities of possible outcomes (PEDREGOSA et al., 2011).

In this model, the idea of odds can be used to represent the probability of a certain event occurring. The odds can be defined using Equation 2.9 where p represents the probability of a positive event (in this case 1).

$$\frac{p}{(1-p)} \tag{2.9}$$

We can then define the logit function, which is the logarithm of the odds, as in Equation 2.10.

$$logit(p) = log\frac{p}{(1-p)} \tag{2.10}$$

Using the logit function, we can build a linear expression to represent the relationship between feature values and log-odds, as seen in Equation 2.11.

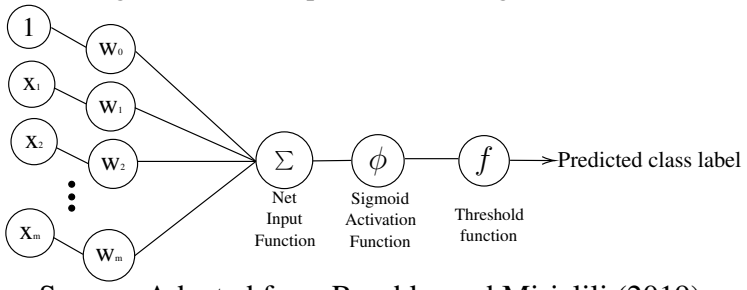
$$logit(p(y=1|x)) = w_0x_0 + w_1x_1 + ... + w_mx_m = \sum_{i=0}^{m} w_ix_i = w^tx \tag{2.11}$$

In Equation 2.11, $p(y=1|x)$ is the conditional probability that a particular example belongs to class 1 given its features **x**. Since we are interested in predicting the probability that an example belongs to a particular class, we look at the inverse form of the logit function in Equation 2.12, which is also known as the sigmoid function due to its characteristic S-shape.

$$\phi(z) = \frac{1}{1+e^{-z}} \tag{2.12}$$

In this equation, z is the combination of weights and the inputs, as described in Equation 2.11. The sigmoid function maps input values into a range of [0,1], with an intercept at $\phi(z) = 0.5$. To understand the execution of the linear regression, an example of the flow can be seen in Figure 2.1.

Figure 2.1 – Example of a linear regression flow.



Source: Adapted from Raschka and Mirjalili (2019).

The goal in the logistic regression algorithm is to find the weights (vector of values

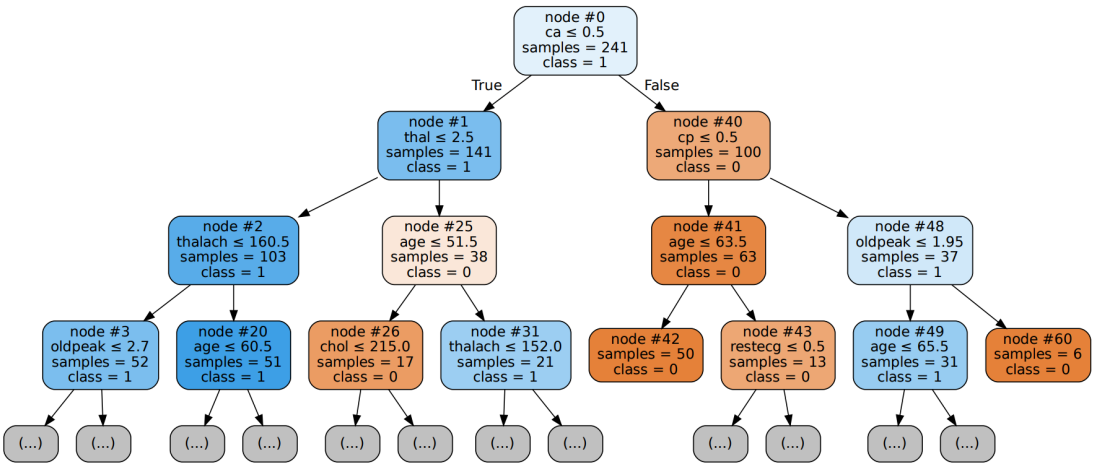
w) that satisfies certain criteria. In Pedregosa et al. (2011) scenario, which is used in this work, the goal is to minimize the cost function in Equation 2.12 with regularization term $r(w)$.

$$min_w C \sum_{i=1}^{n} (-y_i log(\hat{p}(X_i)) - (1 - y_i) log(1 - \hat{p}(X_i))) + r(w) \tag{2.13}$$

2.3.2 Decision Tree

Decision Trees are supervised learning algorithms that are commonly used for both classification and regression tasks. The primary objective of this algorithm is to learn decision rules from the available training data. However, decision trees can suffer from overfitting, which occurs when the generated rules are too specific to the training data, and therefore, fail to generalize well to new data. While the generated tree may perform perfectly on the training set, any outlier data will have a significant impact, leading to poor performance on the test set. To illustrate, Figure 2.2 shows an example of a decision tree generated in this work, with a depth of 3.

Figure 2.2 – Example of decision tree generated for the dataset related to heart attack prediction.



Source: The Author

At each node in the tree, an attribute is selected, and the data is split based on whether the condition is true or false, until the tree reaches the terminal nodes. For instance, node #42 in the figure is a terminal node that predicts the class 0. It is worth noting that the depth of the tree is a hyperparameter that needs to be tuned during the model training process. A tree that is too deep may capture noise in the training data, which can lead to overfitting, whereas a tree that is too shallow may fail to capture important patterns in the data, leading to underfitting. Thus, finding the optimal tree depth

is crucial for achieving good performance on the test set.

Decision trees require a metric to select in which attribute we will split the tree at a given point. In this work, we use the concept of Entropy, also known as cross-entropy or multinomial deviance, which is equivalent to minimizing the log loss. The evaluation of the log loss is based on Equation 2.14. Here, T is the tree model computed on the dataset D, which comprises n pairs, and m is a node.

$$LL(D,T) = \sum_{m \in T} \frac{n_m}{n} H(Q_m) \tag{2.14}$$

The definition for $H(Q_m)$ can be found in Equation 2.15. Here, k is the predicted class, Q_m is the data at node m, and p_{mk} is the proportion of class k observations in node m.

$$H(Q_m) = \sum_{k} p_{mk}(1 - p_{mk}) \tag{2.15}$$

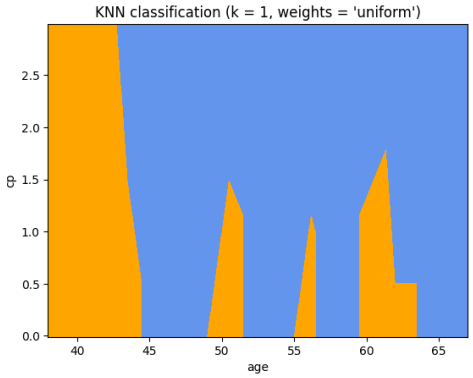
2.3.3 Random Forest

Random Forests are a popular ensemble method that uses Decision Trees to solve classification and regression tasks. As Decision Trees are prone to overfitting and can be unstable, Random Forests employ a perturb-and-combine strategy to create a collection of trees that can combine their predictions and yield more accurate predictions on new data (BREIMAN, 1998). In this work, we use bagging (bootstrapping and aggregation), which involves training each decision tree in the ensemble on a sample drawn with replacement from the training set. The implementation used in this work differs from the classic implementation, as it combines the classifiers by averaging their predictions, instead of letting each tree vote for a single class (BREIMAN, 2001).

2.3.4 K-Nearest Neighbors

The K-Nearest Neighbors (KNN) algorithm is a machine learning algorithm that can be used both for classification and regression problems. When making a prediction, the algorithm identifies the K closest labeled data points in the training set and assigns the label or predicts the value based on the most common label or average value among

Figure 2.3 – Decision Boundary for the KNN algorithm



Source: The Author

those K neighbors. The algorithm establishes a decision boundary in the space such that all points within the boundary belong to the output label that it represents. This can be visualized in a plot by comparing two features, as shown in Figure 2.3, which presents the decision boundary for a modified version of the Heart Attack Dataset used in this work, based on two features (*i.e.,* age and cp). The predicted label for any given combination of these features is either 0 (blue) or 1 (orange), as illustrated in the plot.

To determine the K-Nearest neighbors, we need to establish the value of K, *i.e.,* the number of neighbors to consider when predicting the label, and the metric used to calculate the distance between each point. In this work, we empirically calculate the error and accuracy for each value of neighbors, ranging from 1 to the number of samples in the dataset to determine the optimal number of neighbors. We select the value of K that minimizes the error and maximizes the accuracy. To compute the distance between two points in $\mathbb{R}^q$, we use the Minkowski metric (also known as p-norm) with p=2, which corresponds to the Euclidean distance. The formula for the distance is shown in 2.16, where x and y are two points in $\mathbb{R}^q$. (KRAMER, 2013)

$$||x' - x_j||^p = (\sum_{i=1}^{q} |(x_i)' - (x_i)_j|^p)^{1/p} \tag{2.16}$$

3 RELATED WORKS

In recent years, various libraries and open-source tools have been developed to address machine learning fairness. For instance, the AIF 360 library (BELLAMY et al., 2018) provides different techniques to identify and mitigate bias, with over five fairness metrics and support for more than ten algorithms for bias mitigation. Zehlike et al. (2017) offers fairness metrics such as difference of means, disparate impact, and odds ratio, which utilize trained data to quantify bias. FairML (ADEBAYO, 2016) uses statistical parity to measure bias and presents its advantages and disadvantages. FairTest (TRAMER et al., 2015) is another open-source tool that detects bias by examining the association between the model outcome and protected attributes, with eight different metrics.

Studies have shown that word embedding algorithms may encode marginalized populations differently, perpetuating social bias. For example, Zhang et al. (2020) demonstrated that such models associated African Americans and Blacks with prison and Whites and Caucasians with hospital in a fill-in-the-blank style course of action. The study used post-training metrics to verify demographic disparity and equality of opportunity, concluding that biased data would generate/amplify biased results. Similarly, Júnior et al. (2022) analyzed the COMPAS dataset on criminal recidivism, comparing the accuracy, false and true positive and false and true negative rates for models generated using each cohort of the original data. The study used race as a sensitive attribute and found that the classifier was more accurate when predicting recidivism for Black people and when predicting that it would not happen for every other ethnicity. Alelyani (2021) used a general dataset with information about the census to detect representational bias that might be generating inaccurate predictions for women and non-white people. They used the KL Divergence metric to evaluate the bias, trained the model, and performed a swap on the protected attributes to analyze the impact of the bias in the dataset. The work concludes that data is biased by nature as it reflects the cognition of human brains. It suggests that we should think about why we have bias in the first place and what would be the consequences of swapping the privileged classes. The objective is to have interpretable models that can justify and handle existing bias.

AI models trained on health data can be biased due to protected attributes, such as ethnicity and race. To address this issue, various studies have explored different approaches for evaluating and mitigating bias in these models.

For instance, Noseworthy et al. (2020) evaluated the performance of a convolu-

tional neural network trained on a dataset of different cohorts containing ethnicity and race for protected attributes, and recommended that AI tools report the model's performance for diverse ethnic, racial, age and sex groups. Park et al. (2021) used disparate impact and equal opportunity difference to assess bias in predictive models for postpartum depression and found that bias mitigation algorithms outperformed removing the attribute that might cause bias. Pfohl, Foryciarz and Shah (2021) analyzed two datasets and characterized the impact of penalizing group fairness violations on model performance and group fairness using different performance and fairness metrics. Mandhala et al. (2022) used pre-training bias metrics to evaluate bias in models trained on three datasets, focusing on the model's performance and defining a function to augment data to make it balanced while reducing disparity, minimizing a loss function while maximizing the model performance, this work found that mitigation strategies can be used improved model performance. Finally, Fletcher, Nakeshimana and Olubeko (2021) focused on predicting pulmonary diseases over a set of patients from India and used equality of odds to tune the model and analyze the bias induced by the gender attribute. The study highlighted the challenges of fairness criteria in developing countries, where there is no legal framework to regulate or enforce discrimination prevention and economic disparities present serious obstacles to fair access and benefits from technologies.

While related work has focused on various aspects of bias in health data, there is a gap in the literature regarding pre-training bias metrics for this domain. Table 3.1 presents a summary of the related work. In this work, we will address this gap, evaluating the impact of pre-training bias metrics in the performance of ML algorithms applied to Health datasets.

Table 3.1 – Related work summary

	Post-training	Pre-training	Health Data	General Data
Zhang et al. (2020)	x			x
Júnior et al. (2022)	x			x
Noseworthy et al. (2020)	x		x	x
Park et al. (2021)	x		x	x
Pfohl, Foryciarz and Shah (2021)	x		x	x
Alelyani (2021)		x		x
Mandhala et al. (2022)		x	x	x

4 METHODOLOGY

This chapter describes the methodology applied in the present work to assess the impact of pre-training bias in ML algorithms trained with datasets from the Health domain. Our methodology involves the following steps:

- Collecting datasets related to the Health domain to conduct experiments;
- Adjusting bias manually to create highly unbalanced versions and equally balanced versions of each dataset;
- Measuring pre-training bias with the metrics described in Section 4.1: class imbalance, KL divergence, Kolmogorov-Smirnov and Conditional Demographic Disparity in Labels;
- Optimizing hyperparameters for the ML algorithms selected (*i.e.,* Logistic Regression, Decision Tree, Random Forest, K-Nearest Neighbors);
- Training and evaluating the ML models using the four different algorithms and the different versions of each dataset: the original one, the highly unbalanced, with two variations for the Depression in Medical Students Dataset, and the equally balanced.

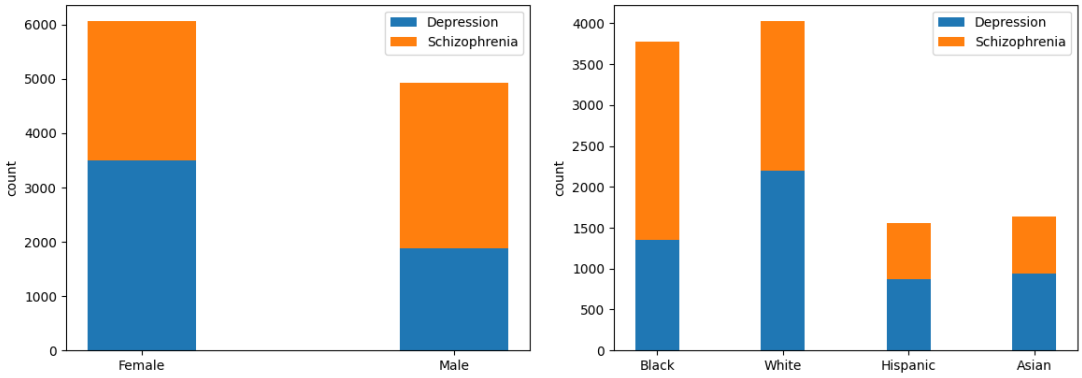
4.1 Data Collection

4.1.1 Intersectional-Bias Dataset

The first dataset used in this study is the Intersectional-Bias Dataset, presented in the ACM FAccT 2022 conference (MASLEJ et al., 2022). The dataset was artificially generated and contains demographic and clinical features that can be used to train a classifier to predict a diagnosis of schizophrenia or depression. This dataset contains two protected attributes, **race** and **sex**. The distribution for each attribute along the different target labels can be found in Figure 4.1. The original dataset contains 11000 instances, with Diagnosis being the target attribute. In the dataset analysis, we see that the target variable (Diagnosis) is highly correlated with Rumination and Tension, and Sex is highly correlated with Rumination as well (values above 0.5). The full heatmap for this dataset can be seen in Figure 4.2. From the original study, a False Positive is the most hurtful result (Diagnostic predicted as 1, schizophrenia), as it is agreed between clinicians that it is better to be misdiagnosed with affective disorder than schizophrenia, so when analyzing

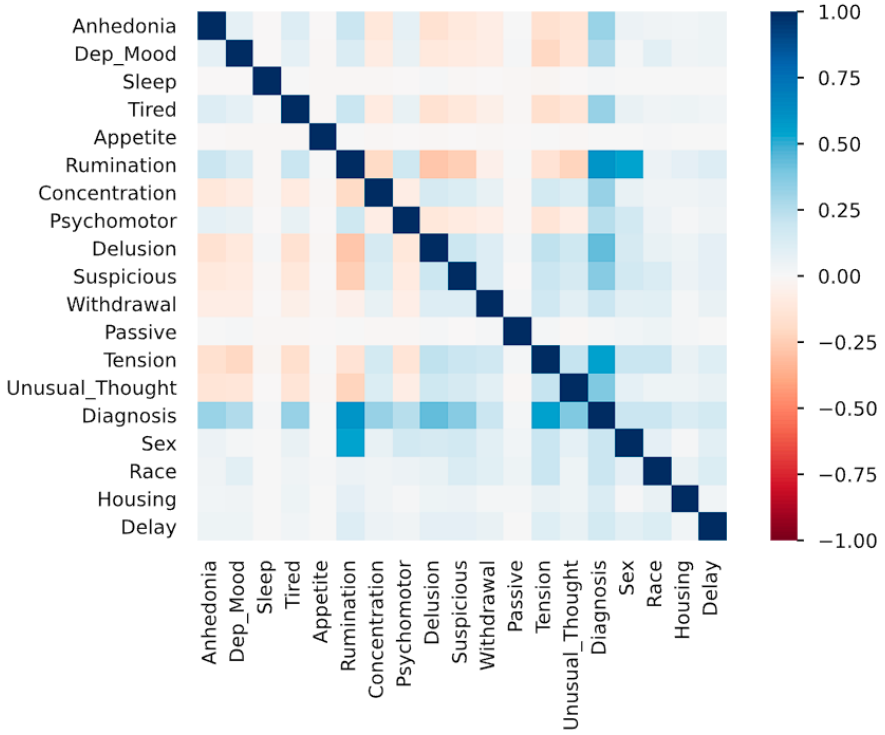
this dataset we will also pay attention to the false positive rate.

Figure 4.1 – Class distribution for protected attributes in the Intersectional-Bias Dataset



Source: The Author

Figure 4.2 – Heatmap for Intersectional-Bias Dataset



Source: The Author

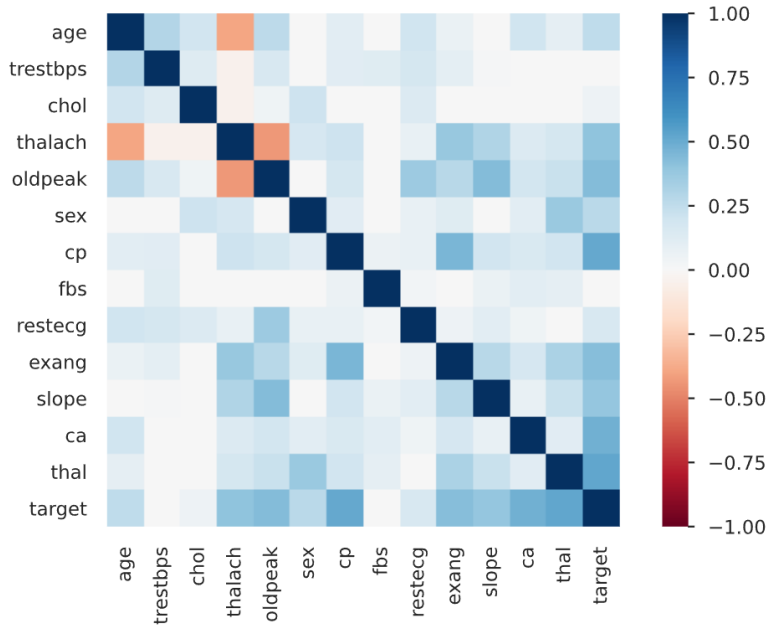
4.1.2 Heart Attack Dataset

The second dataset used in the work is the Heart Attack Analysis & Prediction Dataset extracted from Newman et al. (1998). This dataset contains 303 instances and 14 features, which are described in Table 4.1. The dataset is associated with classification tasks, where the goal is to predict whether a patient has heart disease or not. For the protected attribute in this dataset, we have the feature **sex**, with female (0) being the unprivileged value. The heatmap for feature correlations can be found in Figure 4.3, in which we can see that **cp** and **thal** are highly correlated with **target** (0.509 and 0.521, respectively).

Table 4.1 – Features in the Heart Attack Dataset.

	Type	Definition
age	Numeric	The age of the patient in years
sex	Categorical	The gender of the patient (1 = male; 0 = female)
cp	Categorical	The type of chest pain experienced by the patient (1 = typical angina; 2 = atypical angina; 3 = non-anginal pain; 0 = asymptomatic)
trestbps	Numeric	The resting blood pressure (in mm Hg on admission to the hospital)
chol	Numberic	The serum cholesterol level of the patient in mg/dl
fbs	Binary	Fasting blood sugar >120 mg/dl (1 = true; 0 = false)
restecg	Categoric	The resting electrocardiograph results (0 = normal; 1 = having ST-T wave abnormality; 2 = hypertrophy)
thalach	Numeric	The maximum heart rate achieved
xang	Binary	Exercise induced angina (1 = yes; 0 = no)
oldpeak	Numeric	ST depression induced by exercise relative to rest
slope	Categorical	The slope of the peak exercise ST segment (1 = upsloping; 2 = flat; 0 = downsloping)
ca	Numeric	Number of major vessels colored by flourosopy
thal	Categorical	A blood disorder called thalassemia (0 = normal; 1 = fixed defect; 2 = reversable defect)
target	Binary	Diagnosis of heart disease (angiographic disease status) (0 =<diameter narrowing; 1 =>50% diameter narrowing)

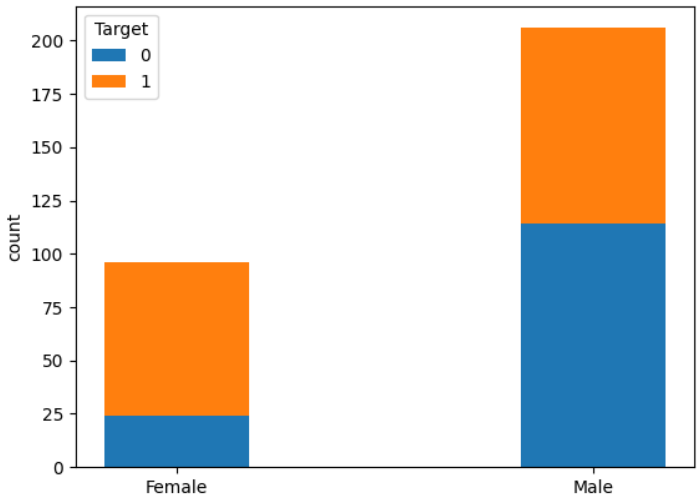
Figure 4.3 – Heatmap for feature correlations in Heart Attack Dataset.



Source: The Author

For the protected attribute, we see that the majority of instances are related to

Figure 4.4 – Target x Sex for Heart Attack Dataset.



Source: The Author

men: almost 70% of the total instances (*i.e.,* 207) correspond to men, while near 30% of the data (*i.e.,* 96) correspond to women. We also notice that in this dataset, women are more likely to have heart disease, with 75% of them having the target value set to 1, against 45% of men linked to the same target value. The comparison between gender and target is in Figure 4.4.

4.1.3 Depression in Medical Students Dataset

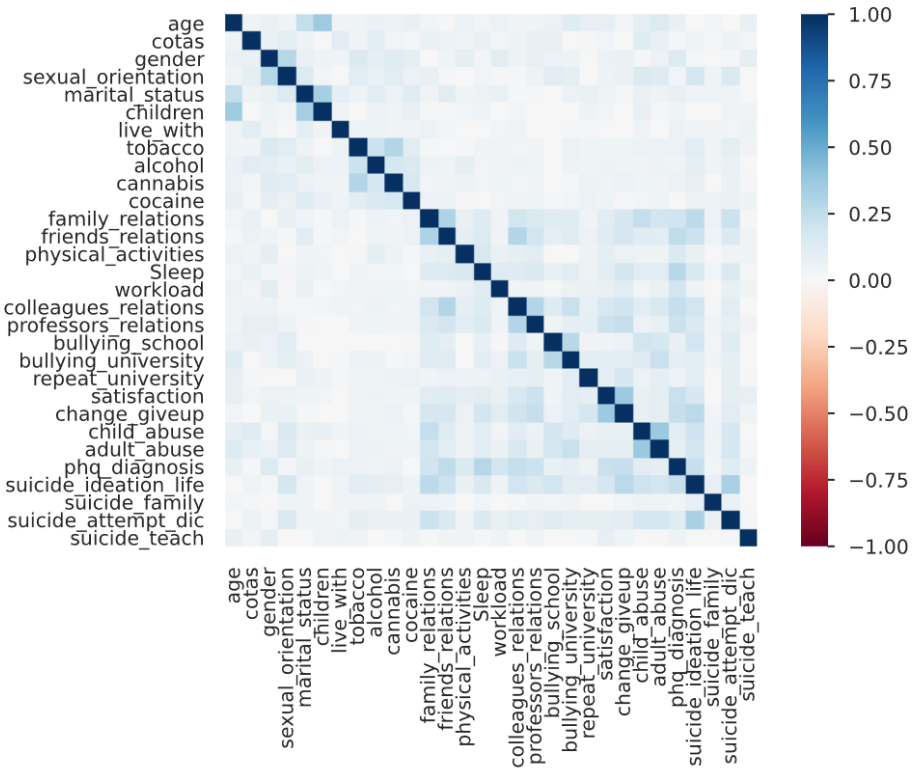
The third dataset used in this study is the Depression in Medical Students Dataset. It contains personal information, university status and mental health indicators from 4840 medical students from Brazil. The dataset was collected in Marcon et al. (2020) and further analyzed in the study by Pereira (2020), in which 43 features were used to identify pathological patterns about the consumption of alcohol among medical students in Brazil. For this study, we have removed some features to simplify the dataset and make it easier to visualize the key concept we wanted to understand about the bias in the data. We used 29 features from the study conducted by Pereira (2020), which are described in Table 4.2, and we used the diagnosis of depression (*i.e.,* **phq_diagnosis** feature) as the target in our classification models.

From the transformed dataset, we have three protected attributes: **cotas**, which

Table 4.2 – Depression in Medical Students Dataset Features

	Type	Definition
age	Numeric	The age of the student in years
cotas	Binary	If the student used racial/economic quota in his application
gender	Categorical	Student gender identity (1=Female, 2=Male)
sexual_orientation	Categorical	Student sexual orientation (1=Heterossexual, 2=Homossexual, 3=Bissexual, 4=Other)
marital_status	Binary	If the student is married (1=yes, 0=no)
children	Binary	If the student has children
live_with	Categorical	Student living situation (1=Alone, 2=With parents/spouse,3=With friends/students,4=Pension/Republic)
tobacco	Categorical	Student tobacco use (1=Not in the last 3 months,2=Once or twice,3=Monthly,4=Weekly,5=Daily)
alcohol	Categorical	Student alcohol frequence of ingestion (0=Never, 1=Monthly,2=2-4 times a month,3=2-4 times a week, 4=Four+ times a week)
cannabis	Categorical	Student cannabis use (1=Not in the last 3 months,2=Once or twice,3=Monthly,4=Weekly,5=Daily)
cocaine	Categorical	Student cocaine use (1=Not in the last 3 months,2=Once or twice,3=Monthly,4=Weekly,5=Daily)
family_relations	Categorical	Student's view on the relationship with their family (1=Bad,2=Regular,3=Good,4=Great,5=Excelent)
friends_relations	Categorical	Student's view on the relationship with their friends (1=Bad,2=Regular,3=Good,4=Great,5=Excelent)
physical_activities	Binary	If the student practices physical activities
Sleep	Categorical	Student's sleeping habits (1=Bad,2=Regular,3=Good,4=Great,5=Excelent)
workload	Categorical	Student's current workload on college (1=Light,2=Moderate,3=Heavy,4=Very Heavy)
colleagues_relations	Categorical	Student's view on the relationship with their colleagues (1=Bad,2=Regular,3=Good,4=Great,5=Excelent)
professors_relations	Categorical	Student's view on the relationship with their professors (1=Bad,2=Regular,3=Good,4=Great,5=Excelent)
bullying_school	Binary	If the student experienced bullying in school
bullying_university	Binary	If the student experienced bullying in university
repeat_university	Categorical	Student grade retention history (1=Never,2=1 time,2=2 times,3= 3 times, 4=4 or more times)
satisfaction	Categorical	Student current satisfaction with their course (1=Deeply unsatisfied,2=Unsatisfied,3=Indifferent,4=Satisfied,5=Very satisfied)
change_giveup	Binary	If the student has ever though of giving up on college
child_abuse	Binary	If the student has experienced abuse in their childhood
adult_abuse	Binary	If the student has experienced abuse in their adulthood
phq_diagnosis	Binary	If the student presents a depressive disorder
suicide_ideation_life	Binary	If the student has ever attempted suicide
suicide_family	Binary	If the student has a family member who ever attempted suicide

Figure 4.5 – Heatmap for feature correlations in Depression in Medical Students Dataset



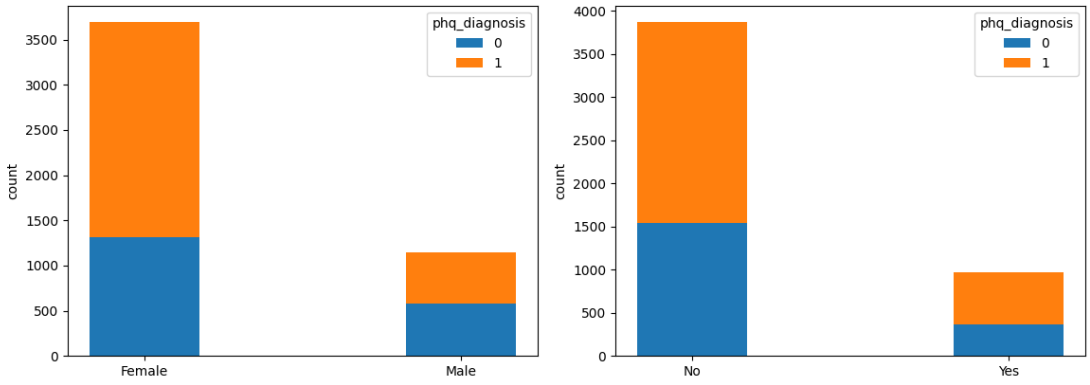
Source: The Author

indicates if the student has used social or racial quotas to enter in college, with "Yes" being the unprivileged group, **gender**, with female being the unprivileged group and **age**, with age below 18 years old being the unprivileged group, as the study was conducted

related to alcohol abuse and this can be considered the age at which people are allowed to buy alcohol in Brazil (where the study was conducted).

Moreover, this dataset does not show any high correlation between its features, as it can be seen in the complete heatmap of Figure 4.5. Thus, we use the highest correlation found for with the target attribute (phq_diagnosis) as the correlated attribute in CDDL, which will be *change_giveup*. In the dataset, we see that 76% of the students are female against 24% of male, but the female students in the dataset are more likely to have a diagnosis of depression, as 65% of women belongs to the class 1, against 49% from men with the same outcome. For the *cotas* attribute, almost 80% of the students have not used the quotas' system to get into college, and among the two classes, the proportion of students with or without quota with the diagnosis positive is very similar (37% and 40% respectively). The analysis of class distribution for the attributes *gender* and *cotas* is shown in Figure 4.6, and the class distribution for age can be found in Figure 4.7, showing all ages (left) and grouping instances by ages over 18 (right).

Figure 4.6 – Class distribution for *gender* and *cotas* on Depression in Medical Students Dataset.

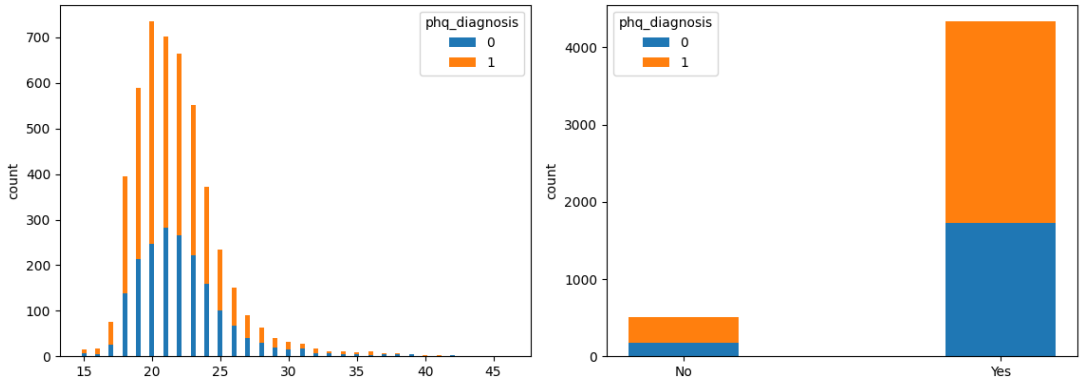


Source: The Author

4.2 Manual bias introduction or reduction

To generate and evaluate different effects in the trained model in terms of the presence of bias, we manually modified the original datasets to change the representation for each protected attribute. Modifications were made in two directions, either adding more representation bias artificially to create a highly unbalanced dataset, or reducing representation bias artificially to create an equally balanced dataset. To achieve that, we

Figure 4.7 – Class distribution for *age* on Depression in Medical Students Dataset, showing all ages (left) and grouped by ages over 18 or not (right).



Source: The Author

manually remove instances from the original data in order to affect the metrics results, analyzing over the variations and the original dataset for each of the selected datasets. Since this operation depends on the dataset being analyzed, more details will be provided in Chapter 5, which is divided in three sections per dataset that will detail how each dataset was transformed to introduce or reduce the bias, and then the experimental results are shown for each transformed dataset.

4.3 Measurement of pre-training bias

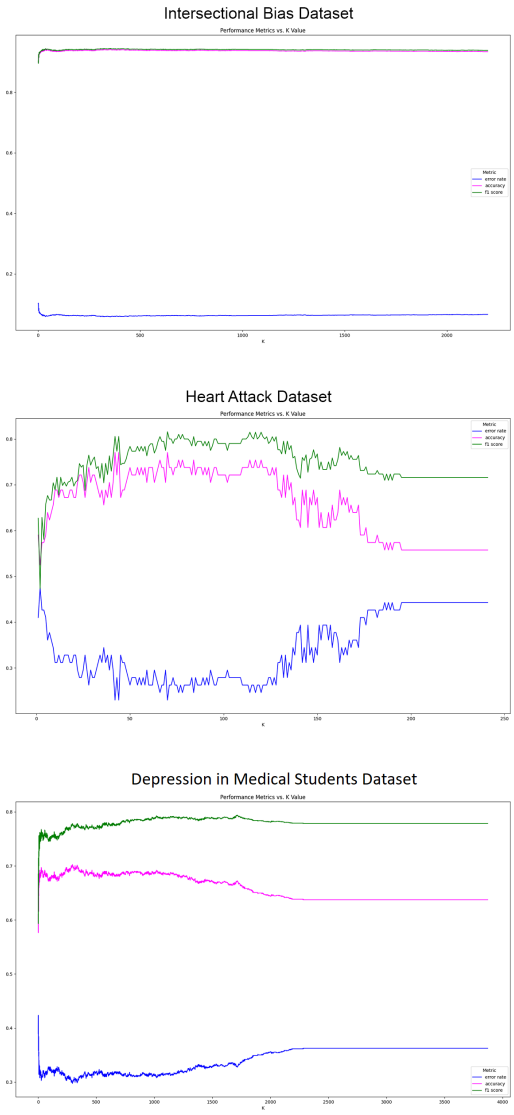
The risk of bias was evaluated with the pre-training bias metrics presented in Section 2.2: class imbalance, KL divergence, Kolmogorov-Smirnov, and Conditional Demographic Disparity in Labels. For Kolmogorov-Smirnov, we used the pandas profiling tool (BRUGMAN, 2019) to find what are the correlated attributes for each dataset. The measurement of pre-training bias was conducted for the original dataset, as well as for the generated datasets with artificial changes aiming to create highly unbalanced and an equally balanced datasets.

4.4 Model hyperparameters configuration

Each of the machine learning algorithms used in this work has its specific hyperparameters, as discussed in Section 2.3, that need to be configured. For the logistic

regression, the maximum number of iterations was initially set to 1000, and increased at each run to reach a limit where the algorithm would converge naturally. For the decision tree and random forest, the criterion was set to entropy, as it is the most commonly used in the literature. The maximum depth of each tree and the number of trees in the ensemble were set empirically by observing the results in test runs.

Figure 4.8 – Analysis of the K value for the KNN algorithm.



Source: The Author

For the number of neighbors in the KNN algorithm, a script was developed to test different values of K and get the model performance (*i.e.,* accuracy, F1-score, and error rate). The initial value was set to one (only the "closest" neighbor) and was increased at each iteration. For the Heart Attack and the Depression in Medical Students Dataset, the maximum value allowed was the number of entries in the dataset, which would mean

"every neighbor". For the Intersectional Bias dataset, the maximum value allowed was the number of instances divided by 4, as it gave acceptable performance metrics when executed. The final values of K for the Intersectional Bias Dataset, Heart Attack Dataset, and Depression in Medical Students Dataset were defined as 318, 42, and 296, respectively. The complete graphs for the K value analysis can be found in Figure 4.8. For the weight function used in the KNN prediction, the *uniform* method was chosen, so all points in the neighborhood are weighted equally.

4.5 Model training and performance evaluation

To evaluate the model performance, our goal is to have the metrics calculated over a number of variations of the training datasets. Thus, we split the dataset to be used for model development and evaluation into training and test sets using the *train_test_split* function from Scikit-learn (PEDREGOSA et al., 2011), with the test sets corresponding to 20% of the original dataset. For each dataset, we create the variations mentioned in Section 4.2 by adjusting the proportions on the training sets, leaving the test sets with the same original proportion for the protected attributes across all variations. For the Intersectional Bias dataset and the Depression in Medical Students Dataset, the data split is performed 6 times with a different random seed, while for the Heart Attack Dataset, it is repeated 10 times. The pre-training bias metrics and the model performance for the datasets are calculated as the average between the different values calculated over each split. This way, it is possible to compare the experimental results among different variations of training sets, using varied test sets with distributions similar to the original data and investigating the impact of bias in the performance of machine learning algorithms.

Performance evaluation was carried out with the following metrics: Accuracy and F1-Score (GOUTTE; GAUSSIER, 2005).

4.5.1 Accuracy

Accuracy is the base metric used for model evaluation. It can be calculated as the ratio of correct predictions over all predictions made for a given dataset. The formula can

be found in Equation 4.1.

$$ACC = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}} \tag{4.1}$$

4.5.2 F1-Score

F1 score can be calculated as the harmonic average of precision and recall, as in Equation 4.4. Precision estimates the proportion of positive instances predicted correctly by the model, while Recall measures the ability of a model to correctly identify positive instances. The definition of precision and recall can be found in equations 4.2 and 4.3, respectively, where TP stands for True Positive (values that were predicted correctly with the positive output), FP stands for False Positive (values predicted positive but were actually negative) and FN stands for False Negative (values predicted as negative but were actually positive). The analysis of the confusion matrix enables a visual analysis of these factors, as shown in Table 4.3.

$$precision = \frac{TP}{TP + FP} \tag{4.2}$$

$$recall = \frac{TP}{TP + FN} \tag{4.3}$$

$$F1 = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4.4}$$

Table 4.3 – Confusion Matrix definition

		Actual Values	
		Positive	Negative
Predicted Values	Positive	TP (True Positive)	FP (False Positive)
	Negative	FN (False Negative)	TN (True Negative)

5 EXPERIMENTS AND RESULTS

To assess the impact of the pre-training bias on machine learning algorithms, we have artificially introduced bias in three datasets, aiming to increase or reduce the values of the pre-training bias metrics analyzed. In the following sections, we report how each variation was generated from the original dataset, along with the results obtained for the pre-training bias metrics and for model performance evaluation considering all the original and modified datasets.

5.1 Intersectional-Bias Dataset

In the first dataset analyzed, two variations were created: one with high imbalance, where we tried to increase the value of the pre-training metrics for all the metrics at the same time, and one where the goal was to have the minimum possible value for all the pre-training bias metrics. This dataset had two protected attributes, **Sex**, which had male as the privileged class and female as the unprivileged class, and **Race**, which had white as the privileged class and Asian, Black and Hispanic as unprivileged classes (which were grouped into "Non-White" to ease the visualization in the figures). The feature used for the correlated attribute on CDDL was Rumination.

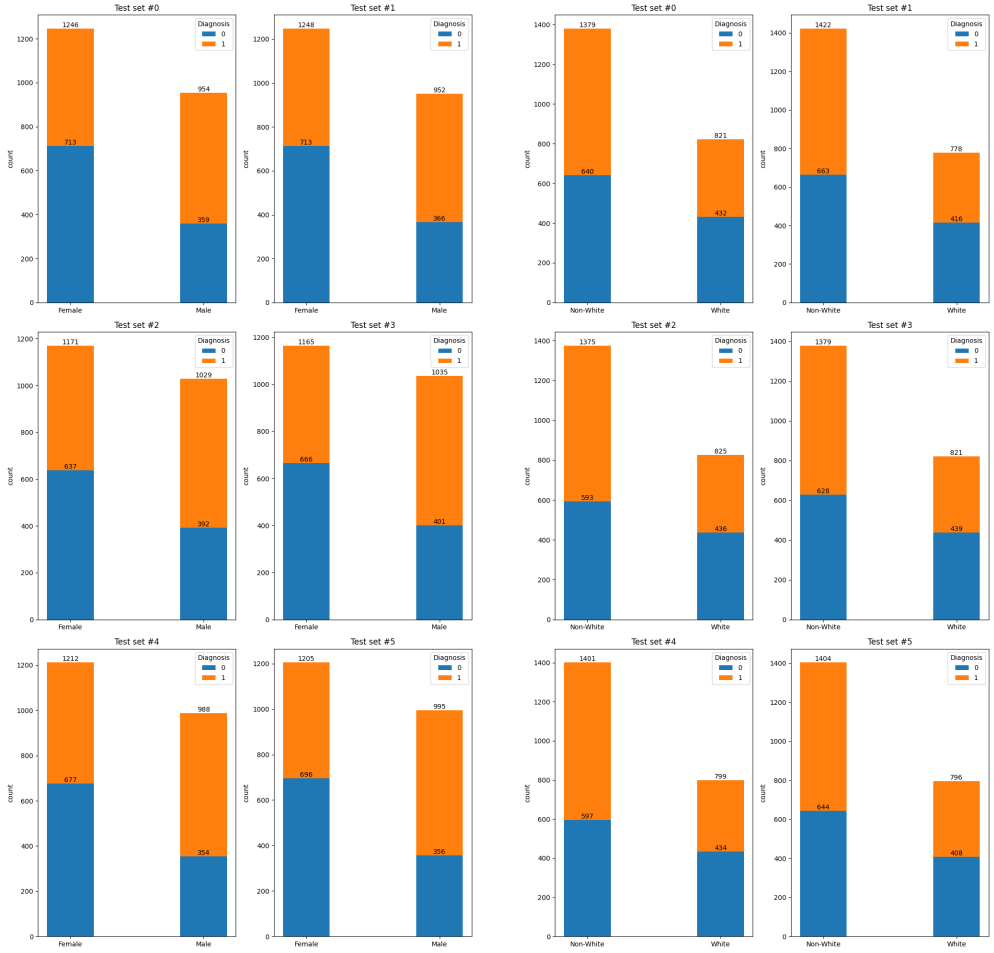
The split into training and test data was executed 6 times, each time with a different random seed. The distributions for the test set were kept the same for all dataset variations, changing only the distribution for the training sets. The test sets for this dataset can be found in Figure 5.1.

To identify which features might have contributed to the predicted class, we also calculated the feature importance for each algorithm when this analysis is embedded in the process of model training: for Random Forest and Decision Tree, Scikit-learn already provides the feature importance as evaluated during model training (PEDREGOSA et al., 2011), and for Logistic Regression, feature importance can be estimated based on the weights of the equation (the w values in Figure 2.1). Feature importance was not analyzed for the KNN algorithm.

The table containing the full report of the pre-training bias metrics for this dataset can be found in Table 5.1, and the table with the report of the performance for the four machine learning algorithms can be found in Table 5.2. The next sections will discuss these results in more details.

Figure 5.1 – Test sets for the Intersectional-Bias Dataset. For each test set, the distributions for the two protected attributes, sex and race, are shown.

(a) Sex (b) Race



Source: The Author

Table 5.1 – Pre-training bias metrics values for the Intersectional-Bias Dataset.

Metric name	Original Dataset	High Imbalance	Equal Balance
Class Imbalance (Sex)	-0.103	0.755	0.000
KL Divergence (Sex)	0.077	0.474	0.000
KS (Sex)	0.195	0.459	0.000
CDDL (Sex, Rumination)	-0.184	-0.207	0.005
Class Imbalance (Race)	-0.268	0.235	0.000
KL Divergence (Race)	0.018	0.938	0.000
KS (Race)	0.096	0.503	0.000
CDDL (Race, Rumination)	0.079	0.500	-0.007

Table 5.2 – Performance results for the Intersectional-Bias Dataset.

Metric	Training Algorithm	Original Dataset	High Imbalance	Equal Balance
Accuracy	Logistic Regression	89.061	84.917	88.508
	Decision Tree	84.386	78.553	83.462
	Random Forest	88.705	80.545	87.894
	KNN	88.402	81.227	87.318
F1-Score	Logistic Regression	89.437	86.464	89.050
	Decision Tree	84.808	80.151	84.262
	Random Forest	89.146	83.431	88.679
	KNN	89.010	84.428	88.388

5.1.1 Original dataset

The original dataset for the Intersectional-Bias had more representation for the unprivileged classes on the protected attributes. The full training set (the sum of the 6 splits) used in the original dataset can be found in Figure 5.2.

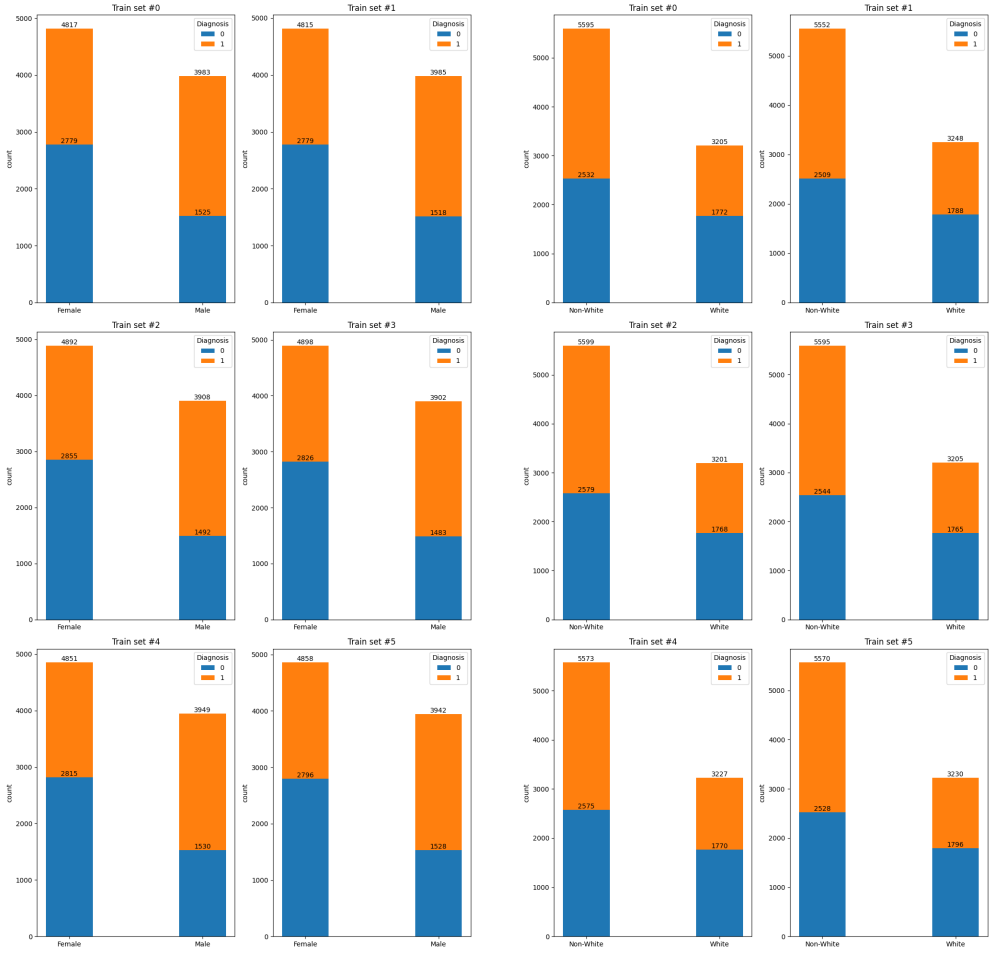
Thus, as expected, the values seen in Table 5.1 for class imbalance were negative for both sex and race. In general, the values for the pre-training bias metrics for the original dataset were close to zero, which indicates that the dataset was not free of bias, but the distribution across the protected attributes was close to being equally distributed.

The performance for the original dataset, shown in Table 5.2, was considered high for all the algorithms. Accuracy values ranged from 84.386 to 89.061, while F1-score ranged from 84.808 to 89.437. However, if we look at the rates at which the models performed in relation to the value of the Sex attribute, we may notice that the algorithms are more likely to make wrong predictions for female than male, and the false positive rate was higher on the unprivileged class, which was considered worse than having a false negative for this scenario. The complete chart for the sex attribute can be seen in Figure 5.3. Looking at the race attribute, the unprivileged value was also more prune to be mislabeled, with a higher error rate and a higher false positive rate. The complete Chart for the Race attribute can be seen in Figure 5.4.

Looking at how each feature contributed to the models, we see that sex was considered important for Decision Tree and Logistic Regression, for the Random Forest the protected attributes were not among the most important attributes, but the performance was similar to the logistic regression. The complete graphs with the feature importance values can be seen in Figure 5.5

Figure 5.2 – Train sets for the original Intersectional-Bias Dataset. For each train set, the distributions for the two protected attributes, sex and race, are shown.

(a) Sex (b) Race

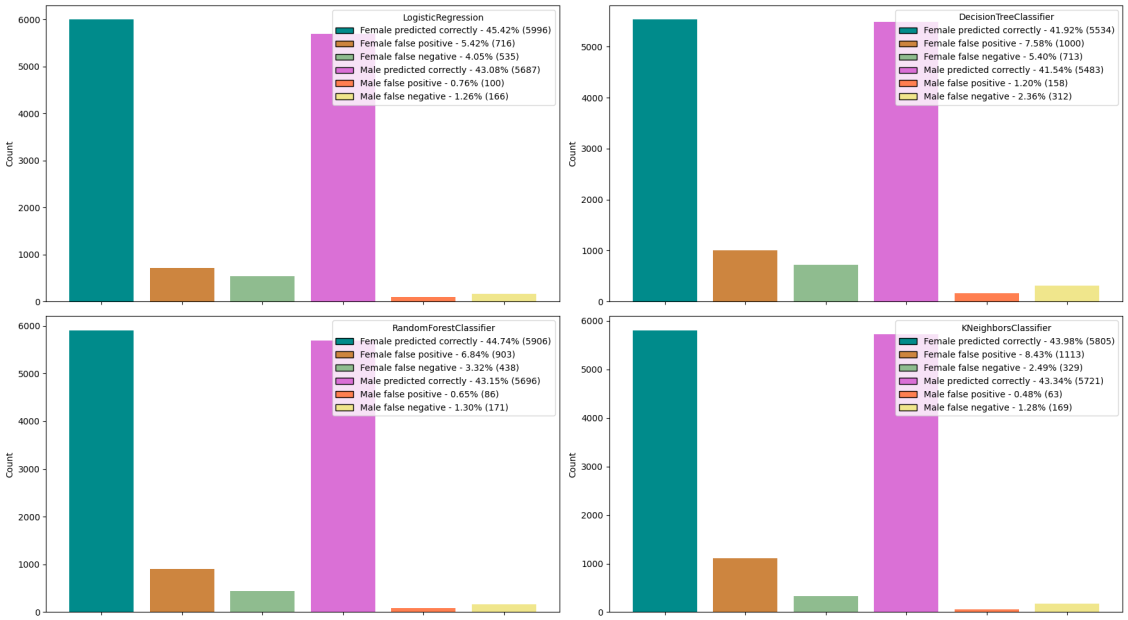


Source: The Author

5.1.2 Highly unbalanced

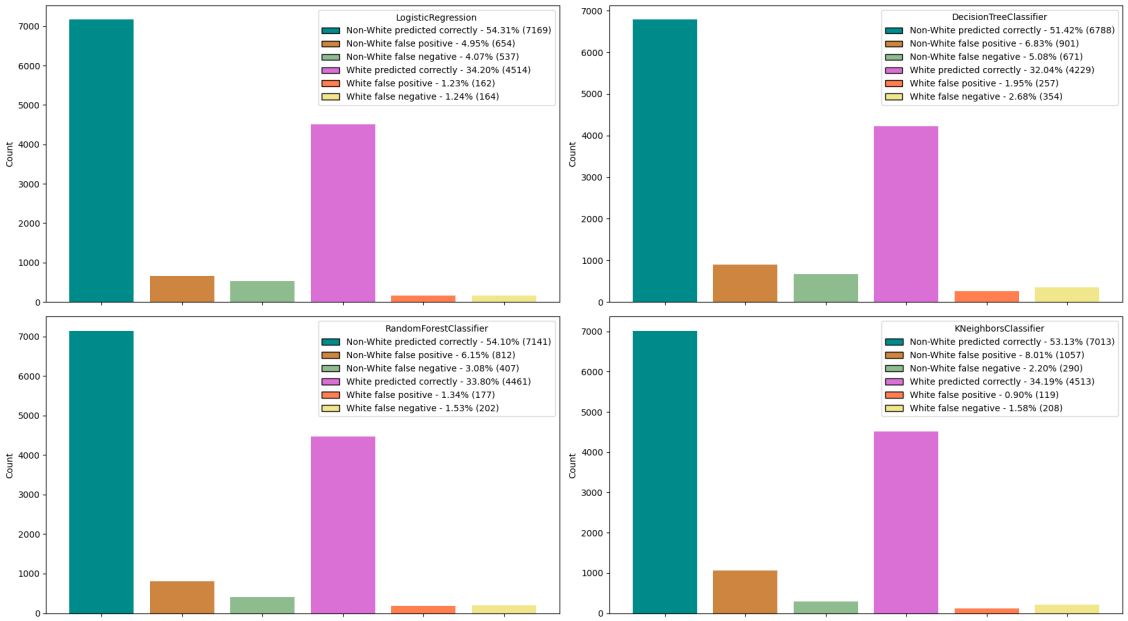
To create the highly unbalanced dataset, we needed to remove both female and non-white instances from the dataset to increase the values for the pre-training bias metrics. To decide the proportion at which we would remove from each class and what is the target for the dropped attribute, a script was developed to combine the possible values for diagnosis and the proportion of instances that would be randomly dropped, so at each iteration, we could calculate the different pre-training bias metrics values. At the end, we had to empirically decide which metrics we would favor in the dataset, so the chosen values were the ones that had the most number of pre-training bias metrics close to be considered high.

Figure 5.3 – Chart for the sex attribute in the original Intersectional-Bias Dataset



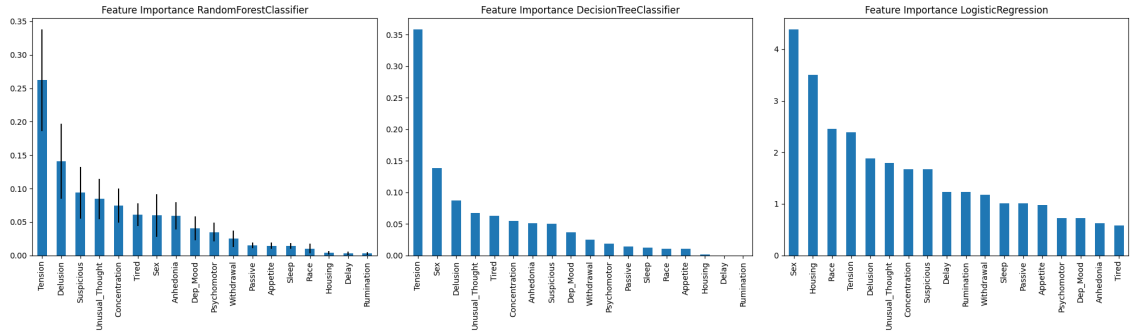
Source: The Author

Figure 5.4 – Chart for the race attribute in the original Intersectional-Bias Dataset



Source: The Author

Figure 5.5 – Feature importance for the original Intersectional-Bias Dataset



Source: The Author

The complete train sets for this dataset variations (the sum of the six iterations) can be found in Figure 5.6, for a) sex and b) race attributes.

Analyzing the pre-training metrics values provided in Table 5.1, only CDDL for Sex and CI for Race were considered low values. But the overall values obtained suggest that we reached the goal of introducing a distribution that differs from the original dataset.

For the Sex attribute, we saw a high increase in the FP rate in all the models, reaching an increase of over 19% on the KNN algorithm for the unprivileged class. The increase, however, was much less prominent in the privileged class. For Male, the largest differences were seen in the Random Forest and Decision Tree, with an approximate 3% increase in the FP rates. The complete chart is in Figure 5.7.

In the Race attribute, the models had an increase in the GP rates for Non-White individuals for all the trained models. The higher increase was observed in the Random Forest classifier, which has an FP rate around 2% in the original dataset and reached over 24% in the highly unbalanced dataset. For the privileged attribute, we only saw an increase in the FP rate for KNN (below 5%), while for the other algorithms, the values were similar or decreased. The complete chart for the Race attribute can be found in Figure 5.8.

The feature importance for the highly unbalanced variation was very similar to the original dataset, except for the Random Forest, where Race became the second most important attribute. Tension was the most important attribute for the Decision Tree and the Random Forest, and Sex was the most important for Logistic Regression.

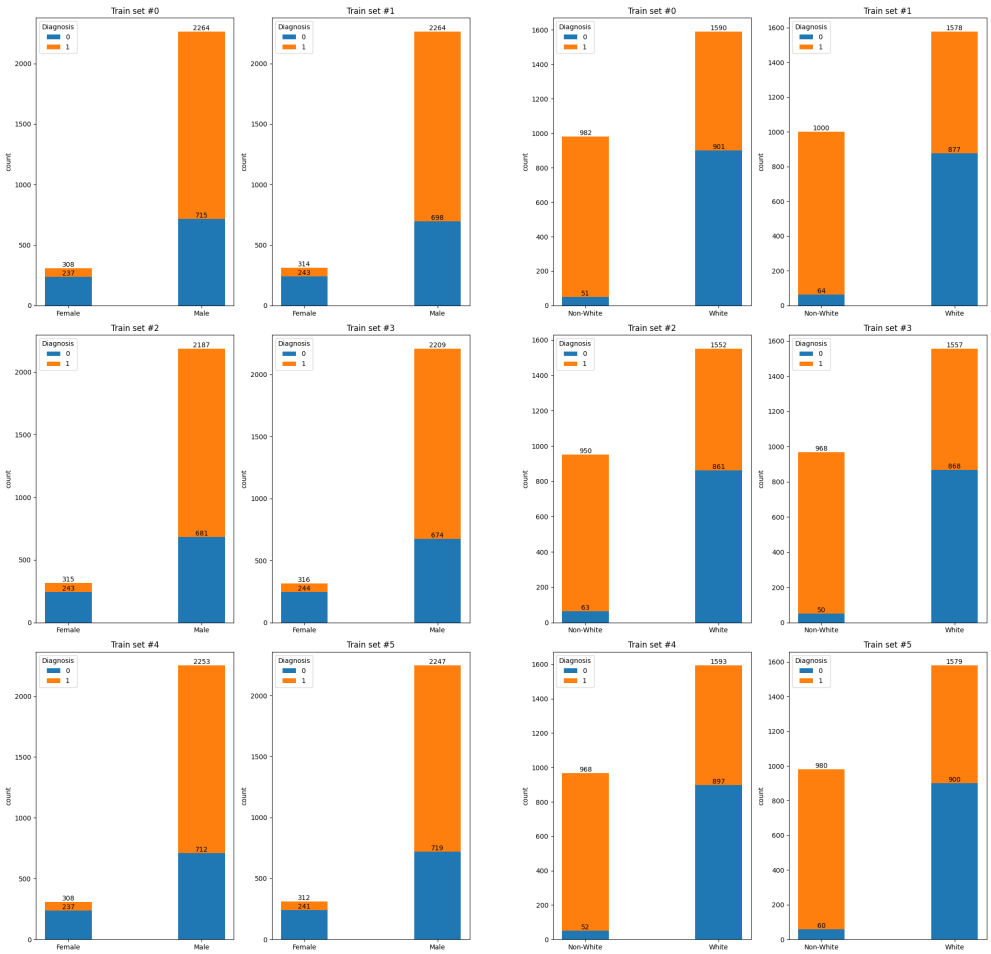
5.1.3 Equally balanced

In order to force the same proportions for all protected attributes on the dataset, we iterated through all the possible combinations of values for the protected attributes and the target, removing from the original set the difference between them on the subset with higher number of individuals. This approach resulted in the training sets shown in Figure 5.10 (the sum of the 6 iterations). The impact of removing these instances in the training sets caused each train iteration to have, in average, 3600 fewer instances than for the original dataset. The pre-training bias metrics calculated over this dataset variations, shown in Table 5.1, were close to zero.

Analyzing the sex attribute with focus in the FP rate, the performance was sim-

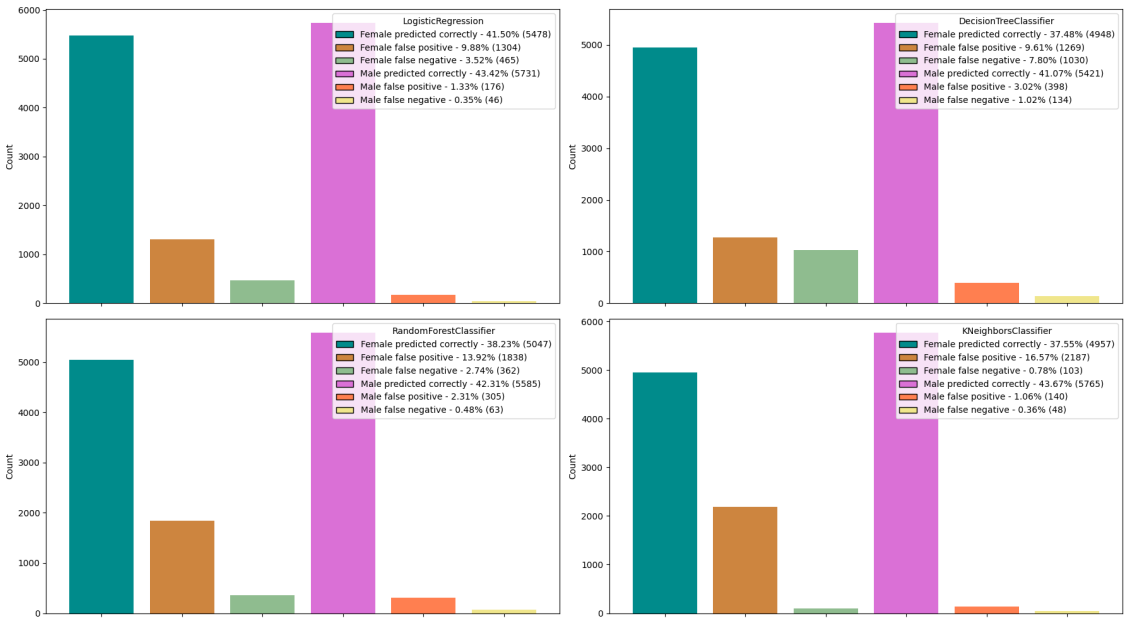
Figure 5.6 – Train sets for the Highly unbalanced Intersectional-Bias Dataset. For each train set, the distributions for the two protected attributes, sex and race, are shown.

(a) Sex (b) Race



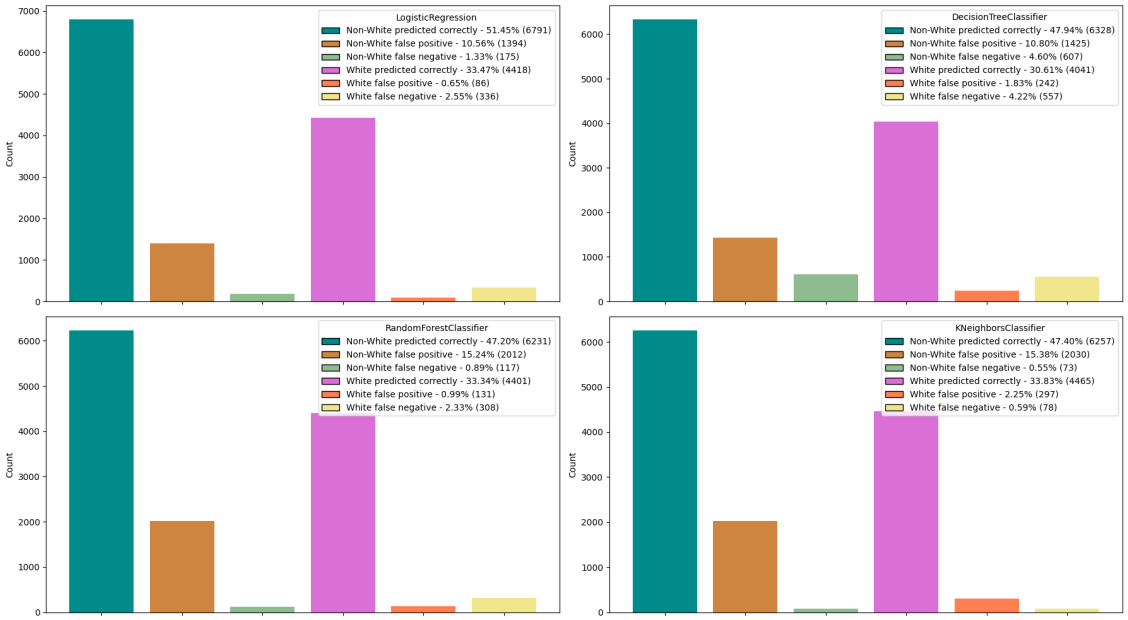
Source: The Author

Figure 5.7 – Chart for the sex attribute in the highly unbalanced Intersectional-Bias Dataset.



Source: The Author

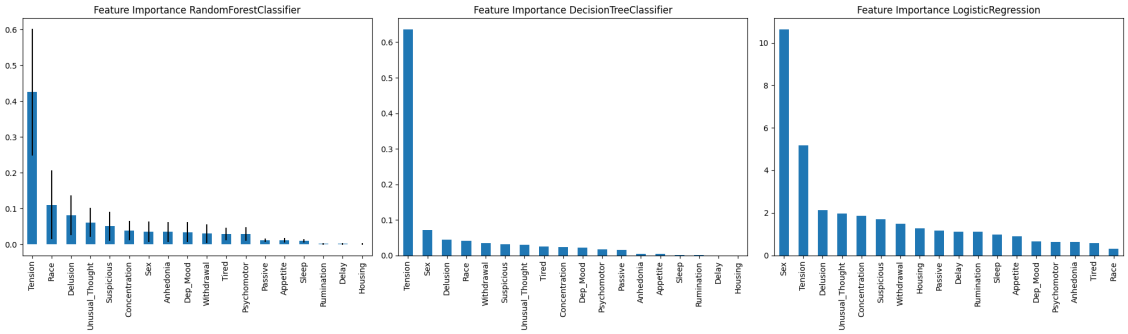
Figure 5.8 – Chart for the race attribute in the highly unbalanced Intersectional-Bias Dataset.



Source: The Author

ilar to the original dataset. On average, we observed 3% increase pf FP rates for the unprivileged class, and in the privileged class the FP rate variation was below 1% for all algorithms, with the FP rate being slightly higher in the original dataset. The KNN model was an exception in this analysis, for which the FP rate was 0.07% higher in the equally

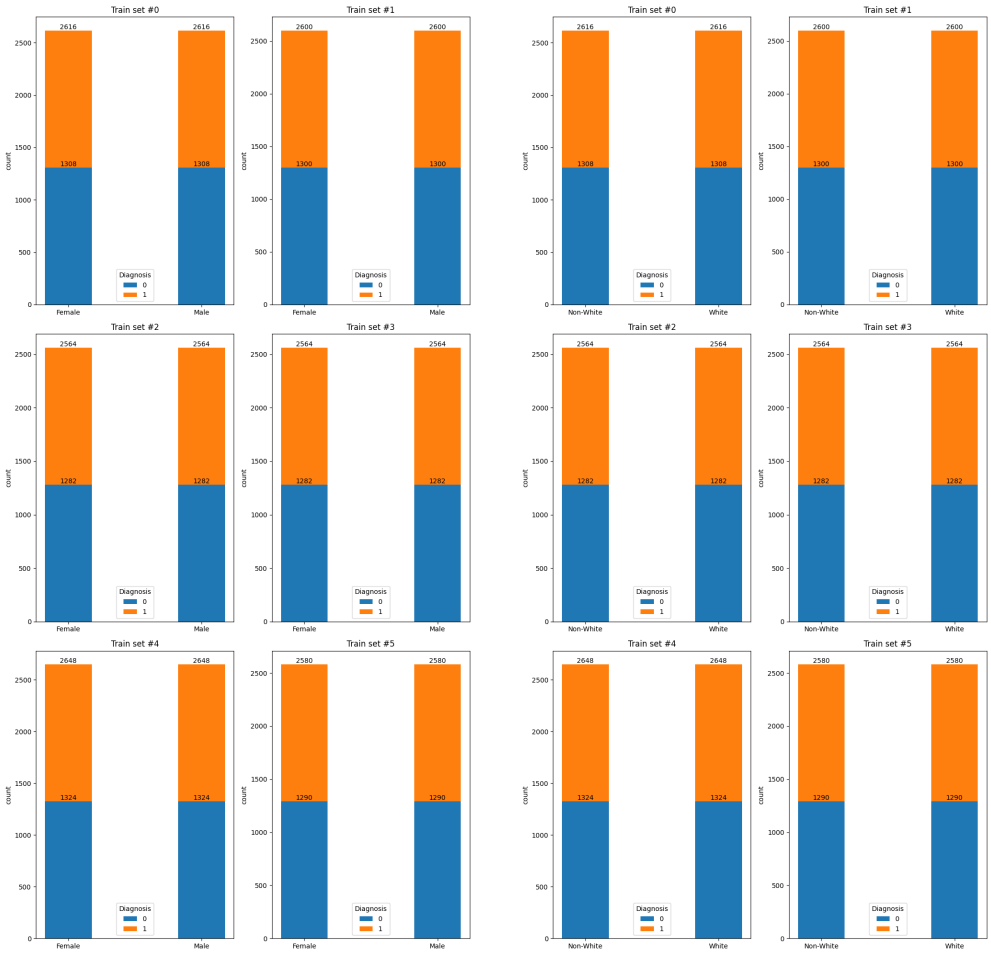
Figure 5.9 – Feature importance for thehighly unbalanced Intersectional-Bias Dataset.



Source: The Author

Figure 5.10 – Train sets for the equally balanced Intersectional-Bias Dataset. For each train set, the distributions for the two protected attributes, sex and race, are shown.

(a) Sex (b) Sex

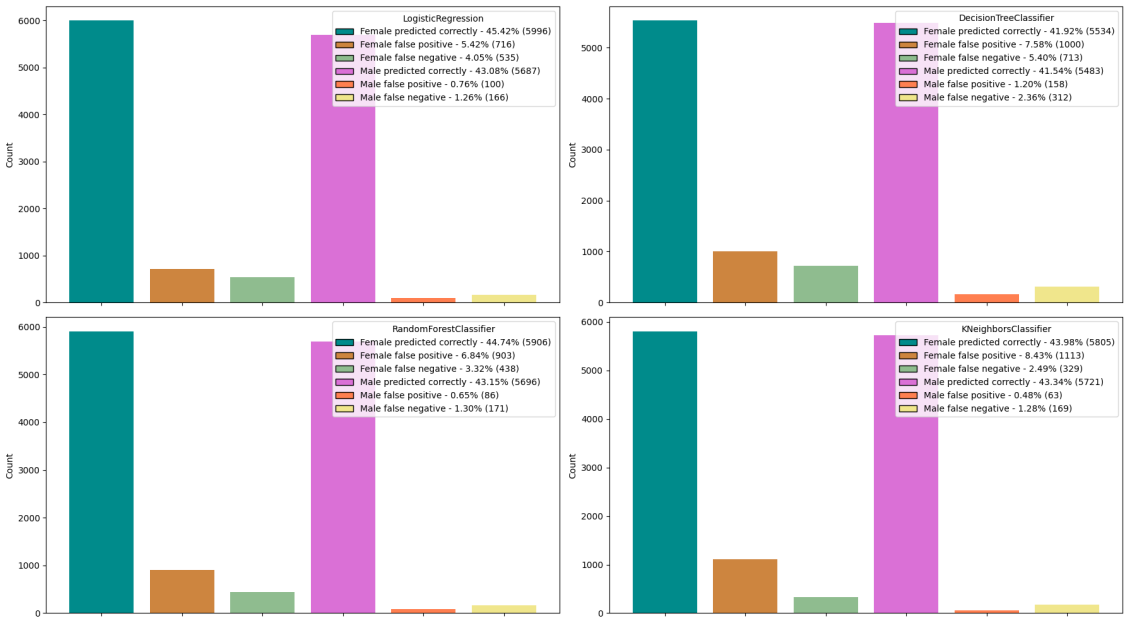


Source: The Author

balanced dataset. The complete chart for the sex attribute can be found in Figure 5.11. In the Race attribute, the protected attribute also had a similar increase in the FP rate, on

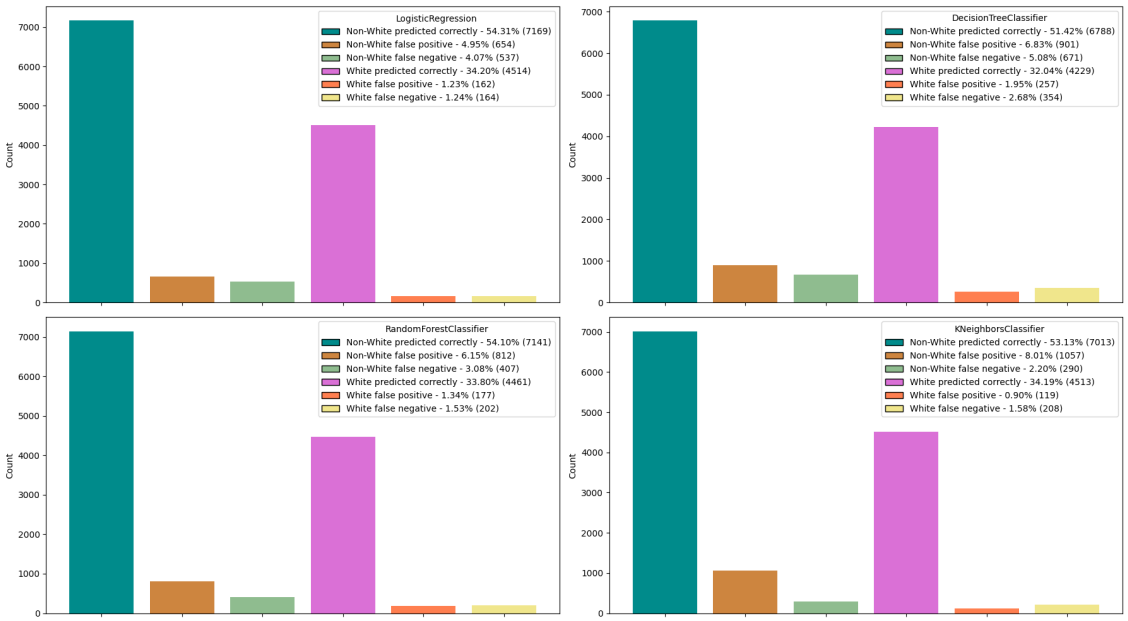
average 2% in the unprivileged class for each model. For the privileged class, the FP rates were also similar to the original dataset, with variations under 1% for each model. The chart for the Race attribute can be found in Figure 5.12.

Figure 5.11 – Chart for the sex attribute in the equally balanced Intersectional-Bias Dataset



Source: The Author

Figure 5.12 – Chart for the race attribute in the equally balanced Intersectional-Bias Dataset

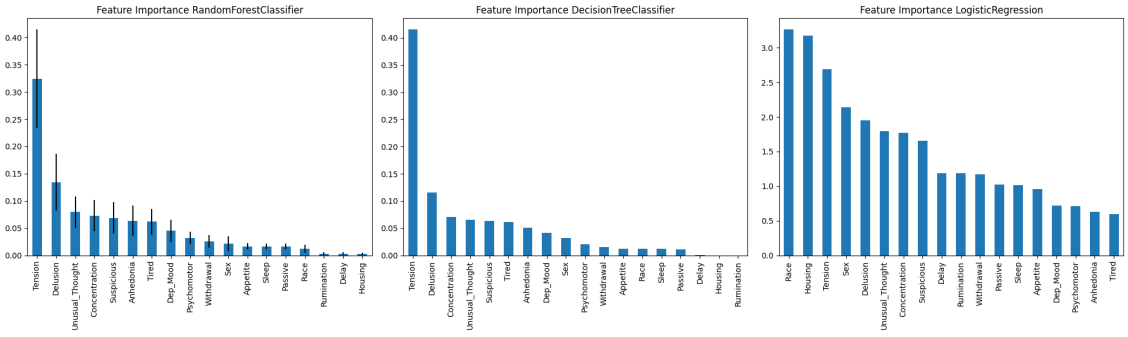


Source: The Author

The feature importance of this dataset variation were similar to the original ver-

sion, but we see some changes, specially in the protected attribute Sex, that lost its importance in the three algorithms, and Race becomes the most important feature in the Logistic Regression. The full graphs can be seen in Figure 5.13.

Figure 5.13 – Feature importance for the equally balanced Intersectional-Bias Dataset.



Source: The Author

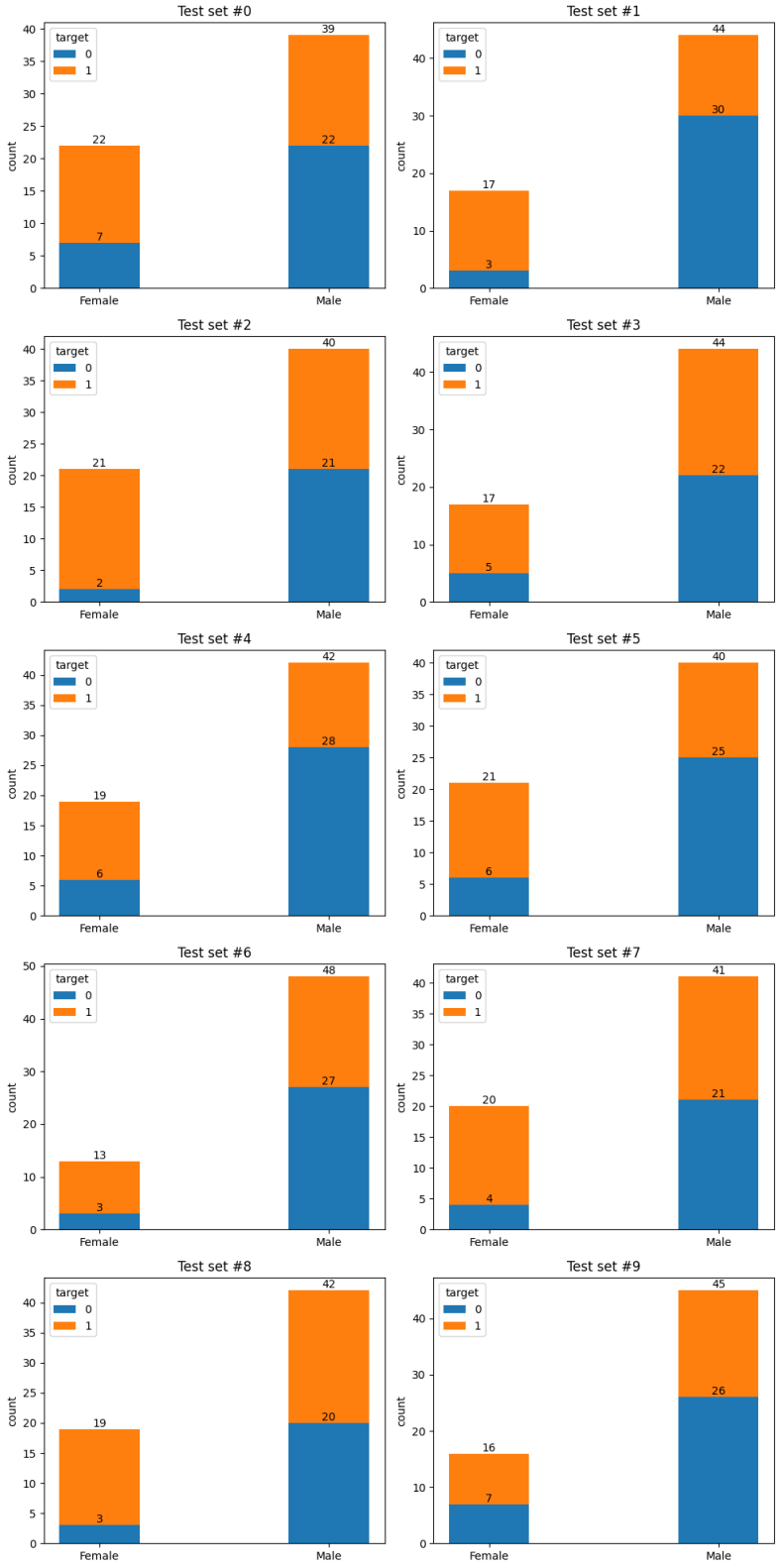
5.2 Heart Attack Analysis & Prediction Dataset

In this dataset, we introduced bias by changing the proportions of the protected attribute **sex** in the training sets, creating two variations over the original dataset: one highly unbalanced and one equally balanced. Each algorithm was run 10 times, varying the seed supplied to the algorithm responsible for generating random numbers used to split the data into training and test sets. The test sets have a common distribution across all the datasets variations, and can be found in Figure 5.14. The table with the pre-training bias metrics calculated over all the variations of the dataset can be found in Table 5.3, and the performance for each algorithm can be found in Table 5.4.

Table 5.3 – Heart Attack Dataset pre-training bias metrics values

Metric name	Original Dataset	High Imbalance	Equal balance
Class Imbalance (sex)	0.357	0.548	0.000
KL Divergence (sex)	0.202	1.346	0.000
KS (sex)	0.299	0.524	0.000
CDDL (sex, cp)	0.291	0.375	0.086
CDDL (sex, thal)	0.108	0.275	-0.123

Figure 5.14 – Test sets for Heart Attack Dataset



Source: The Author

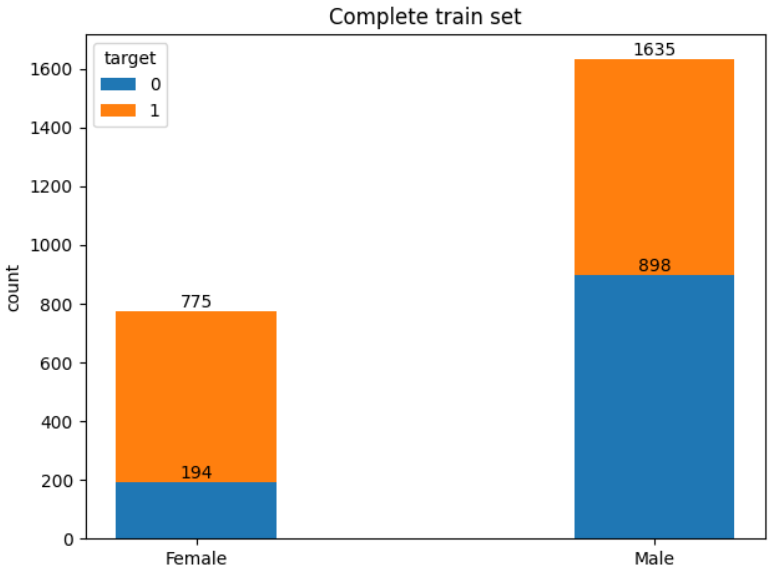
Table 5.4 – Performance results for Heart Attack Dataset

Metric	Training Algorithm	Original Dataset	High Imbalance	Equal Balance
Accuracy	Logistic Regression	84.590	83.934	80.656
	Decision Tree	76.885	79.344	73.115
	Random Forest	83.607	83.443	79.672
	KNN	67.377	68.197	53.115
F1-Score	Logistic Regression	85.826	85.518	83.226
	Decision Tree	77.977	80.720	76.938
	Random Forest	84.564	85.021	82.961
	KNN	71.017	72.486	69.019

5.2.1 Original dataset

As we have seen in Figure 4.4, this dataset has more representation of male individuals than females. This unbalance is reflected in the test sets (Figure 5.14) and in the training set, which is shown in Figure 5.15 (the sum of the 10 executions). We can see that men correspond to around 68% of the dataset, whereas women are only 32% of instances.

Figure 5.15 – Train sets for the original Heart Attack Dataset



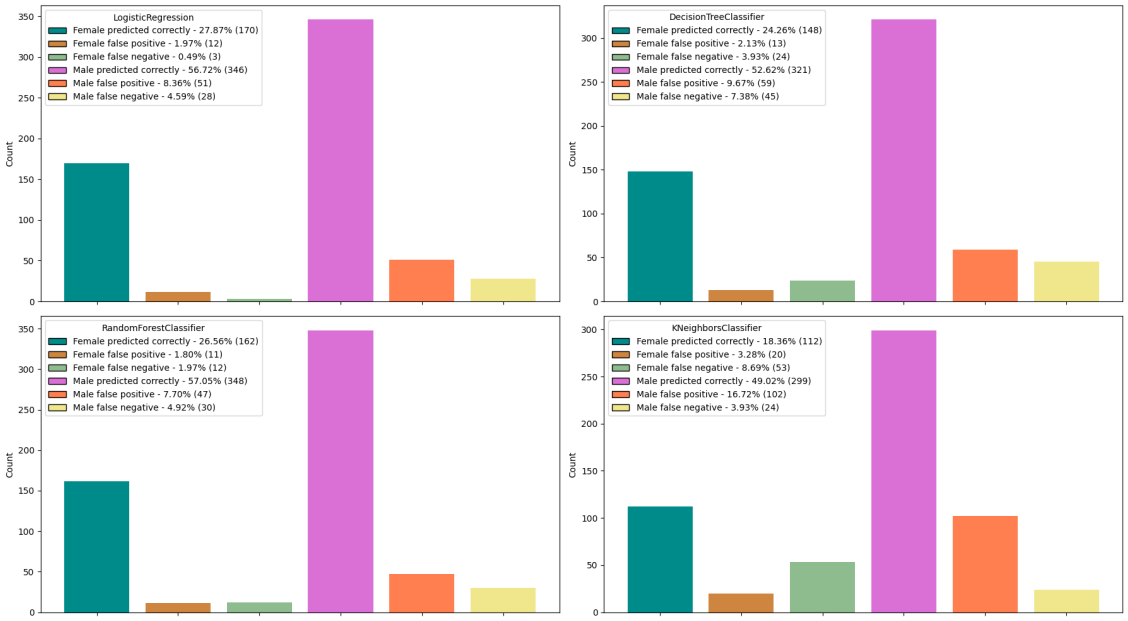
Source: The Author

As we can see in the Figure 5.16, the models are prone to achieve more accurate predictions for men (in absolute quantity), which is expected since this value for the sex attribute has more representation both in training and test data. Moreover, this behavior can be seen in all four algorithms analyzed. By the proportions of instances correctly predicted for women and men, we might infer that the unbalancing was not heavily affecting

individuals by their gender. If we look at the rate at which the algorithms classified men and women correctly, the differences are under 12% for all algorithms, and the proportion at which the models correctly predicted the class for women was higher than for men.

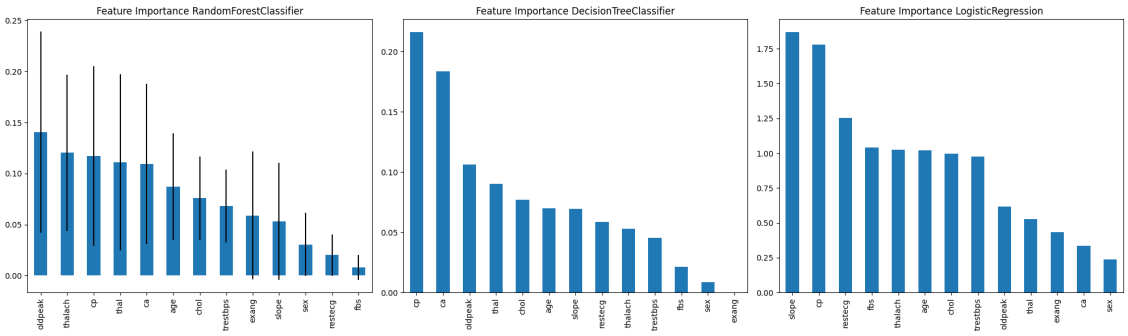
To understand what is the impact of the features on the trained models, the feature importance analysis for the three possible algorithms was calculated and the graphs reporting the values can be found in Figure 5.17. We can see that the **sex** attribute contribution was very little to the final class prediction in all the three algorithms.

Figure 5.16 – Chart for the original Heart Attack Dataset.



Source: The Author

Figure 5.17 – Feature importance for the original Heart Attack Dataset

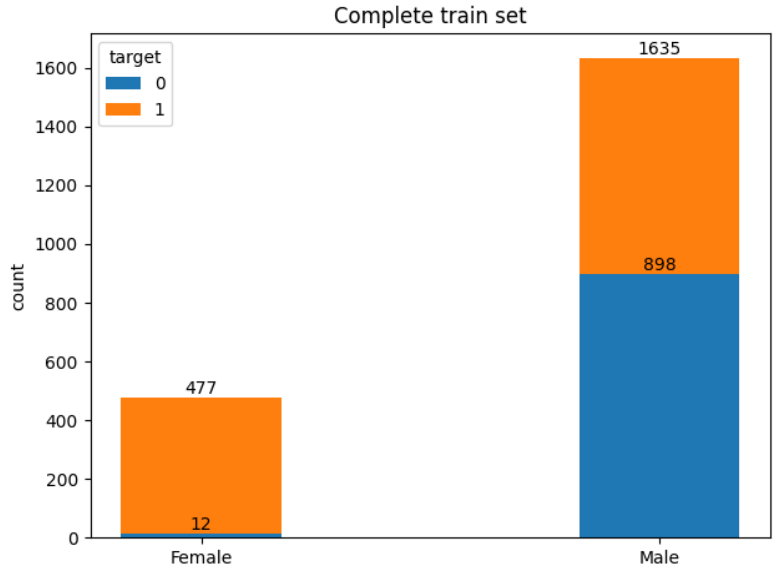


Source: The Author

5.2.2 Highly unbalanced

In order to generate a highly unbalanced dataset, we firstly looked at the distribution for each of the correlated features in relation to the sex attribute, tested the possible combinations for the values and evaluated the pre-training bias metrics reported on each iteration. To generate the final unbalanced dataset, we removed 85% and 80% of women with *thal* equals to 2 and 3 and target equals to 0, respectively. From the resulting dataset, 80% of women with *cp* equals to 2 or 0 and target equals to 0 were removed as well, and finally 20% of women with target equal to one were removed. The complete train set (with the sum of the 10 splits) can be observed in Figure 5.18.

Figure 5.18 – Train sets for the highly unbalanced Heart Attack Dataset.

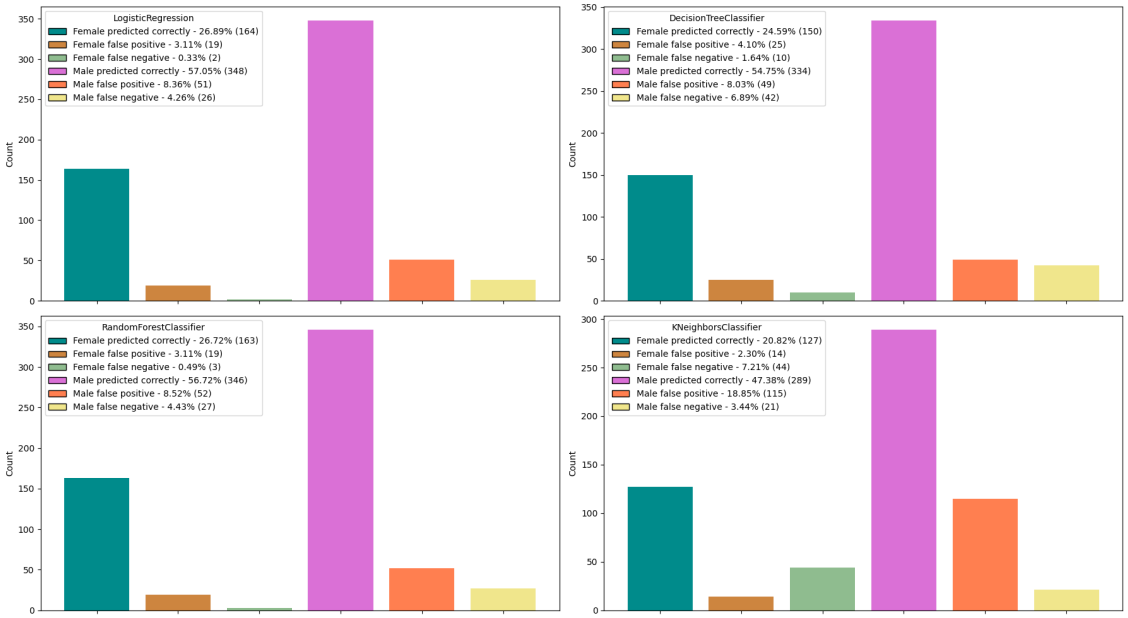


Source: The Author

The numbers in Table 5.3 related to the pre-training bias metrics evaluated in the highly unbalanced dataset show that all metrics are affected with the unbalance, meaning the dataset has more male representation. The distributions for classes were highly affected (percentage of women with positive output higher than the percentage of men with positive output) and there was a disparity in the labels for women with thalassemia and chest pain generated by the removals. The performance for the models can be found in Table 5.4.

By analyzing the proportions of correct and incorrect predictions for male and female, we can see that although the accuracy and F1-score might have actually improved for some models (*e.g.,* the Decision Tree). The rate at which the model is getting correct

Figure 5.19 – Chart for the highly unbalanced Heart Attack Dataset.

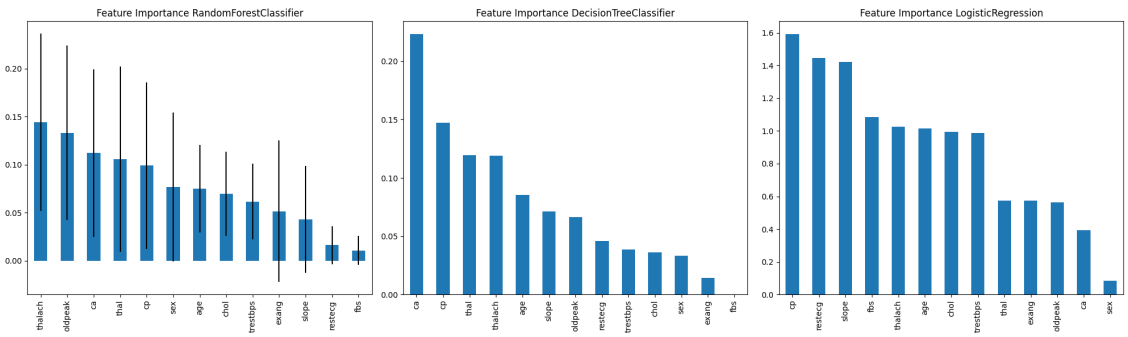


Source: The Author

predictions for the unprivileged classes has decreased for Logistic Regression. These results are shown in the charts of Figure 5.19.

If we look at the importance of each feature for this dataset, as shown in Figure 5.20, we can see that for the Decision Tree, the ones affected by changes in its proportions contributed significantly to the result, whereas in other models, like the Random Forest, these variables contributed less, and the performance for women in test set was slightly affected. Finally, in the Logistic Regression model, where *cp* was the most significant variable, *thal* and *sex* contributed very little. For the privileged class, the accuracy reached by all models were almost the same (varying around 1% for each model).

Figure 5.20 – Feature importance for the highly unbalanced Heart Attack Dataset

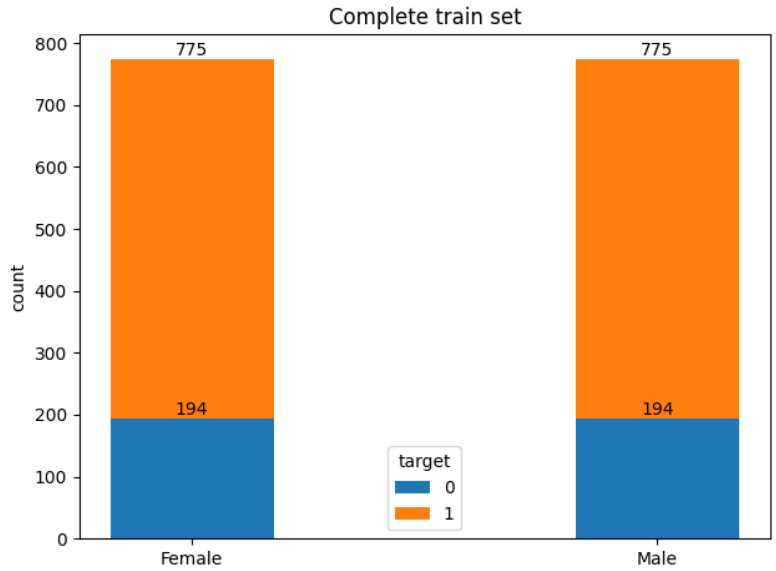


Source: The Author

5.2.3 Equally balanced

To generate equal balance in the dataset, we firstly need to look at the correlated attributes in the dataset (*cp* and *thal*). Based on the original distribution, it is not possible to achieve the perfect balance for these features, but the values were low enough to be acceptable. The values can be seen in Table 5.3. To balance the *sex* attribute, we simply removed men from the dataset, computing the difference in the absolute number of women and men for each possible output and removing this number from the train set. The complete dataset used for training (the sum of 10 repetitions) can be seen in Figure 5.21.

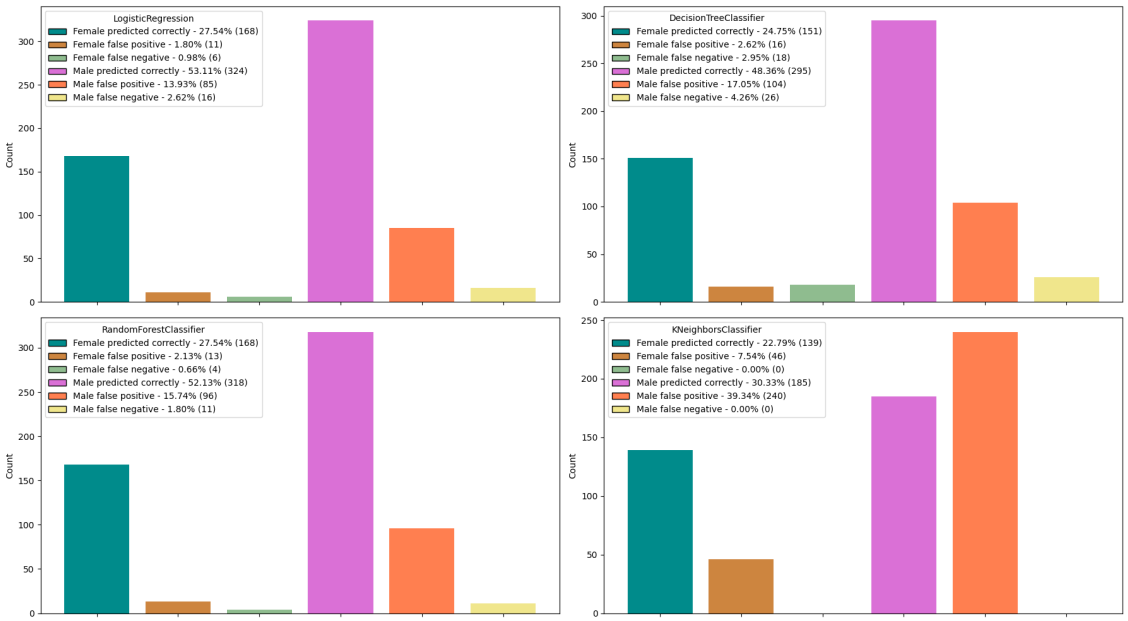
Figure 5.21 – Train sets for the equally balanced Heart Attack Dataset.



Source: The Author

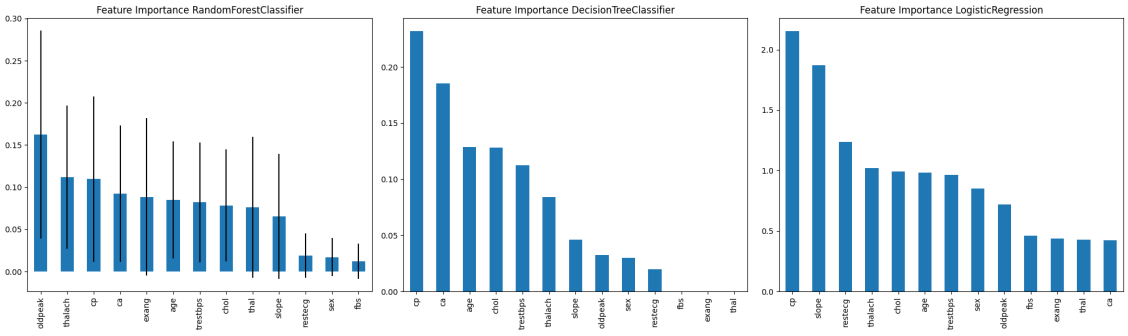
Analyzing the Figure 5.22, we can see that the pattern of having a higher percentage of women predicted correctly (when compared to the total number of women in the testing sets) was kept the same as the other datasets. However, in this specific case, we can notice that comparing with the correct prediction for women with the original dataset, KNN had an improvement of almost 15% in the overall accuracy, and its false negative rate decreased substantially, achieving almost zero for the unprivileged class. In the privileged class, all the algorithms had worse performance when comparing with the original dataset, with the higher difference being in the KNN, where the correct predictions went from 70% to 43%. If we look at the feature importance for the equally balanced dataset, in Figure 5.23, we can see that the protected attribute had no significant impact for any algorithm.

Figure 5.22 – Chart for the equally balanced Heart Attack Dataset.



Source: The Author

Figure 5.23 – Feature importance for the equally balanced Heart Attack Dataset.



Source: The Author

5.3 Depression in Medical Students Dataset

In this dataset, we focus mainly on the analysis of the **age** attribute, with some discussion on the **gender** attribute, as we want to see the differences between changing the pre-training bias metrics for values that contribute differently to our models predictions. The tests were run for the three protected attributes, but since the scope of experiments got too wide and the analysis of feature importance for the algorithms (*i.e.*, Random Forest, Decision Tree and Logistic Regression) showed that not all attributes were highly contributing to the outcome, we decided to reduce the scope of the analysis and focus on

attributes suggests having a more important impact on the outcome.

For this dataset, we generate two highly unbalanced variations, one related to the **age** attribute and one related to the **gender** attribute. For the equally balanced variation, we attempt to level the proportions of both attributes on the dataset. The experiment was repeated with 6 different random seeds for the dataset split into train and test data. Similar to the previous experiments, when varying the attribute distributions in dataset, we only modified in the training data and kept the original distribution in the test set. Figure 5.24 shows the 6 different datasets used for testing, with the distribution for the protected attributes. The tables with the pre-training bias metrics and models performance for all the dataset variations can be found in Table 5.5, and in Table 5.6, respectively.

Table 5.5 – Depression in Medical Students Dataset pre-training bias metrics values

Metric name	Original Dataset	High Imbalance (gender)	High Imbalance (age)	Equal Balance
Class Imbalance (gender)	-0.528	0.668	-0.433	0.000
KL Divergence (gender)	0.049	0.786	0.029	0.000
KS (gender)	0.154	0.452	0.093	0.000
CDDL (gender, change_giveup)	0.106	0.251	0.122	0.004
Class Imbalance (age>18)	0.794	0.802	0.577	0.000
KL Divergence (age>18)	0.004	0.003	0.741	0.000
KS (age>18)	0.044	0.040	0.575	0.000
CDDL (age>18, change_giveup)	0.025	0.015	0.632	0.012

5.3.1 Original dataset

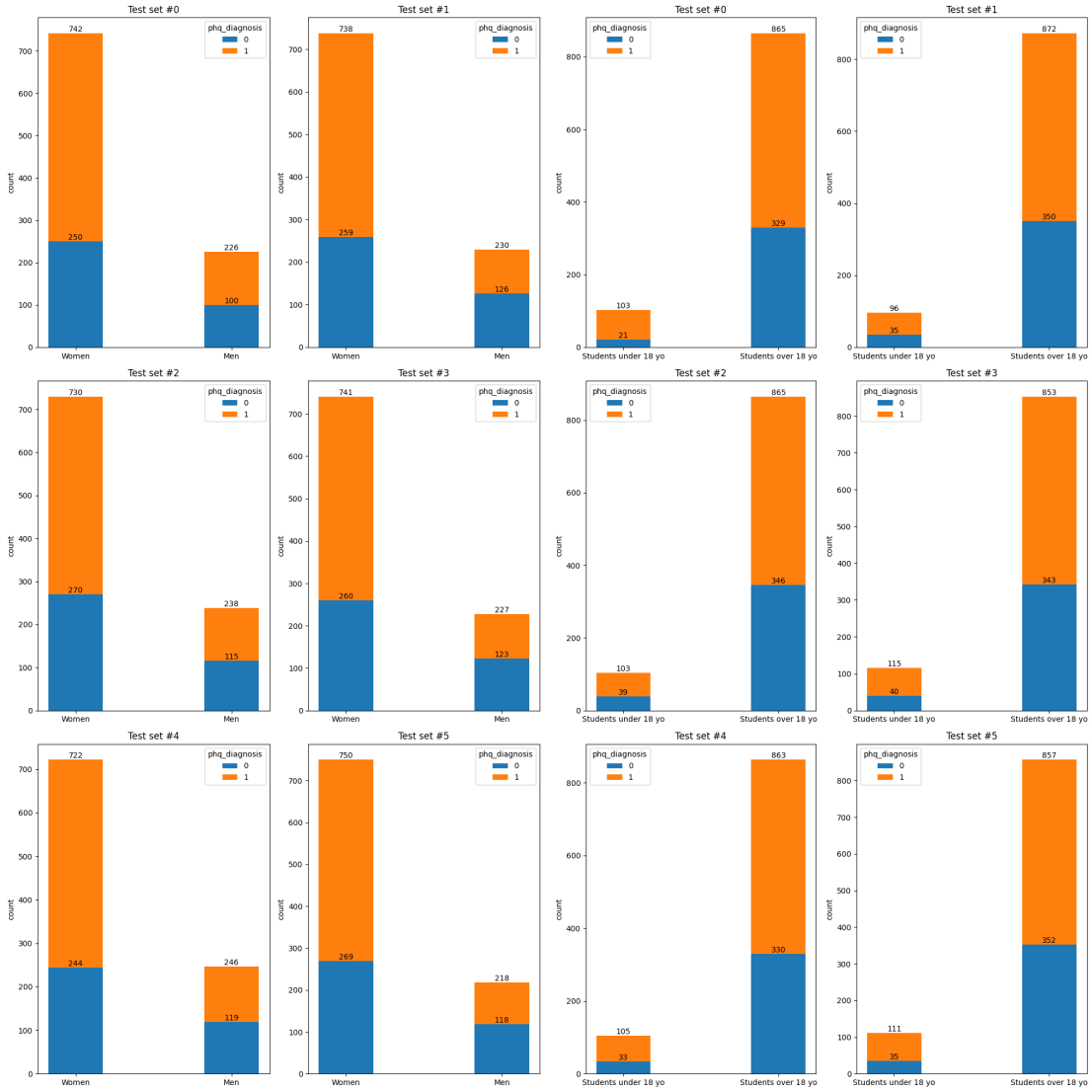
The original dataset in this case is inherently unbalanced for the protected attribute **gender** and for the attribute **age**, as it has more female representation than male and more representation of individuals over 18 years old, so we have a negative value for CI in gender and a positive value for CI in age. The KL, KS and CDDL values indicate that even though the dataset has more representation from a certain class, the outputs are fairly distributed across the difference facets, since none of these metrics reported a value above 0.2 (the complete table can be found in Table 5.5). The complete train set can be found in

Table 5.6 – Performance results for Depression in Medical Students Dataset

Metric	Training Algorithm	Original Dataset	High Imbalance (gender)	High Imbalance (age)	Equal Balance
Accuracy	Logistic Regression	71.212	65.599	48.123	66.271
	Decision Tree	61.398	64.704	48.140	56.990
	Random Forest	69.542	65.582	43.578	65.031
	KNN	69.025	68.165	39.652	57.989
F1-Score	Logistic Regression	77.522	76.828	28.900	71.004
	Decision Tree	68.094	75.139	35.217	62.477
	Random Forest	76.338	76.742	18.206	70.198
	KNN	76.853	75.024	3.045	61.309

Figure 5.24 – Test sets for the Depression in Medical Students Dataset. For each test set, the distributions for the two protected attributes, gender and age, are shown.

(a) Gender (b) Age

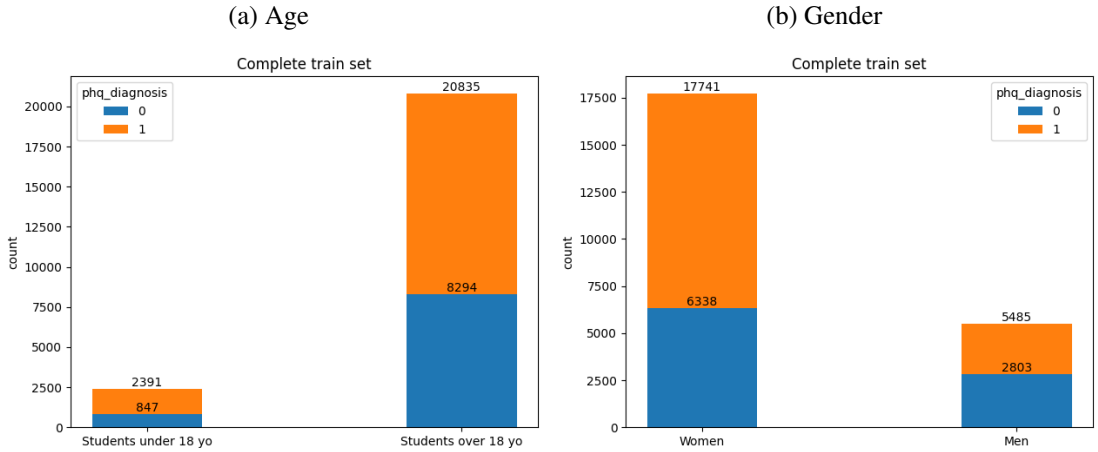


Source: The Author

Figure 5.25.

Analyzing the predictions for this dataset, we can see that the model was slightly more prone to get correct predictions for women than for men, both in absolute number (which is expected due to the higher number of individuals in this class on the test sets) and in proportion (72%, 62%, 70% and 70% of correct instances for women against 67%, 58%, 67% and 64% from men on Logistic Regression, Decision Tree, Random Forest and KNN, respectively). The complete chart for the original dataset concerning the gender attribute can be found in Figure 5.26.

Figure 5.25 – Train sets for the original Depression in Medical Students Dataset.

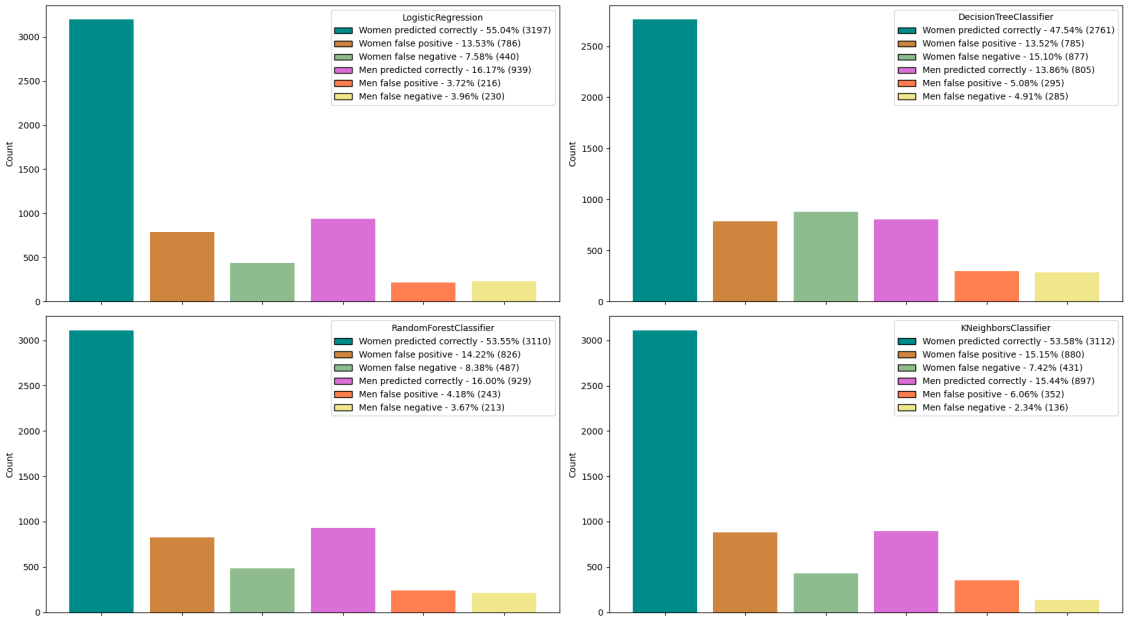


Source: The Author

In the age attribute, we see that even with the distribution across the labels being very disproportional, the model was able to get accurate predictions for both individuals under and over 18 years old, achieving similar values for accuracy in all models. We observed the privileged class (under 18 years old) having a slightly higher accuracy. The complete chart for the age attribute can be found in Figure 5.27.

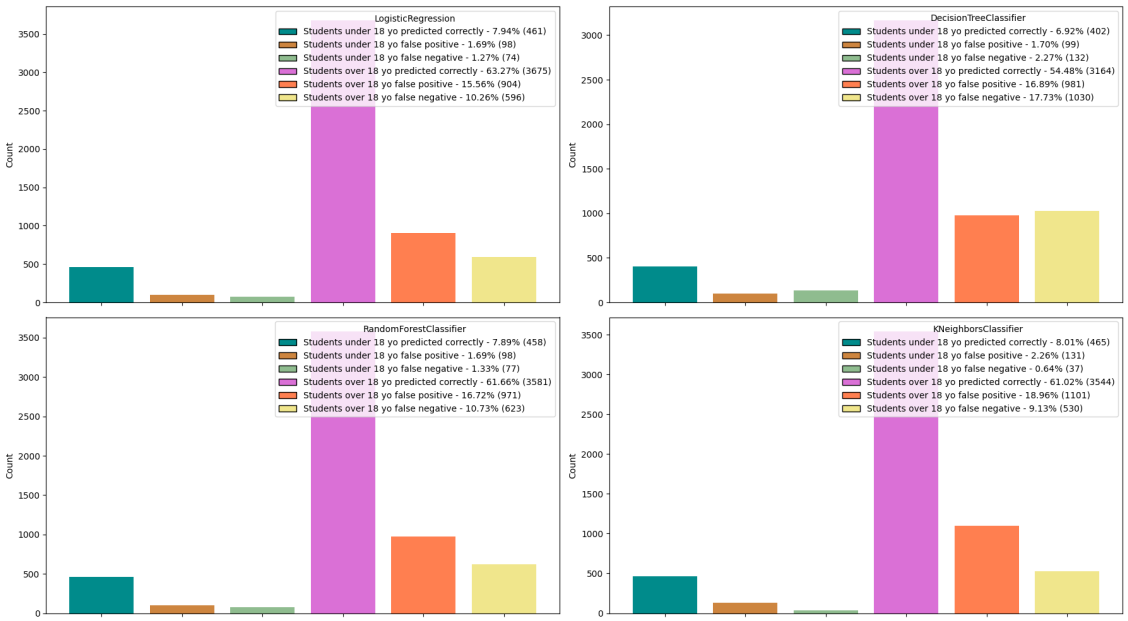
The feature importance chart (in Figure 5.28) shows that for both Random Forest

Figure 5.26 – Chart for gender in the original Depression in Medical Students Dataset



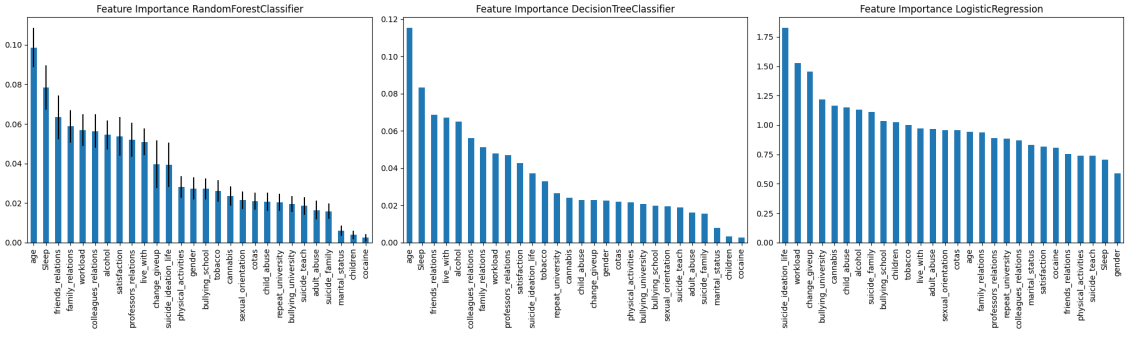
Source: The Author

Figure 5.27 – Chart for age in the original Depression in Medical Students Dataset.



Source: The Author

Figure 5.28 – Feature importance for the original Depression in Medical Students Dataset



Source: The Author

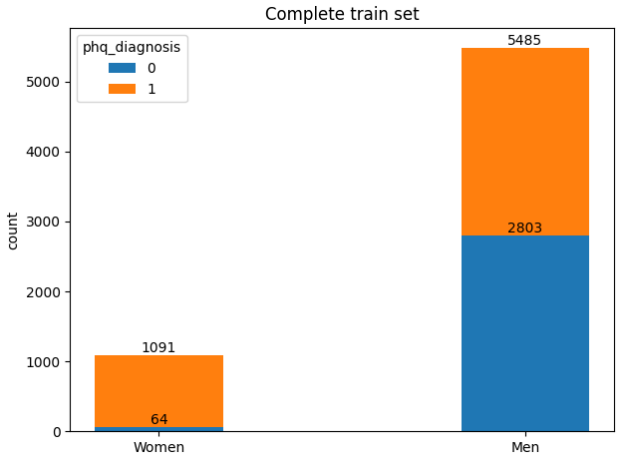
and Decision Tree, the **age** feature highly contributed to the predicted class, while **gender** is not really contributing to the models predictions. The **change_giveup** (which was the correlated attribute chosen for CDDL) contributed more for the logistic regression than for the other algorithms.

5.3.2 Highly unbalanced in relation to gender

To make the dataset highly unbalanced, it was necessary to drop instances from the class with the majority of individuals in the training sets, which made the training sets

to be around one third the size of the original training sets. We note that this reduction in the number of training instances can impact the models' performance. We removed 99% of the women in the original dataset with negative output and 91% of women with positive output. Figure 5.29 shows the complete training set (sum of 6 iterations) used for this variation.

Figure 5.29 – Train sets for the highly unbalanced (gender) Depression in Medical Students Dataset.

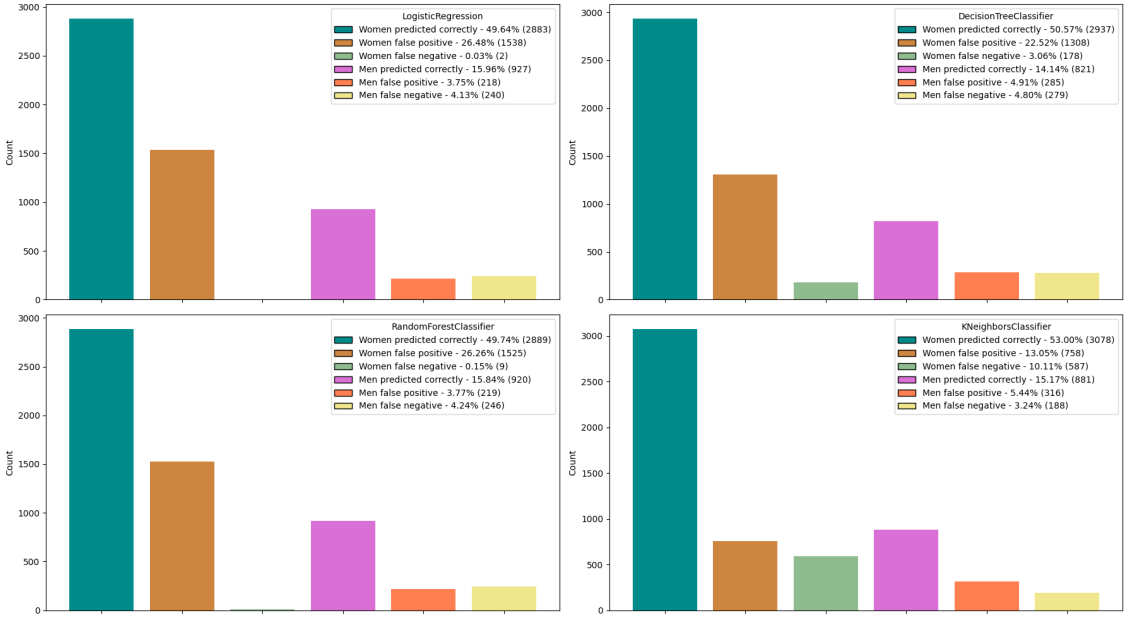


Source: The Author

Removing this percentage increased the values of all pre-training bias metrics, as we see more instances in the privileged class (which affects directly the CI), and the proportion for the target was also affected in these groups, since we kept more instances with the positive output (affecting KS and KL). For CDDL, it was not possible to increase the value above 0.251 as it would negatively impact the other pre-training bias metrics.

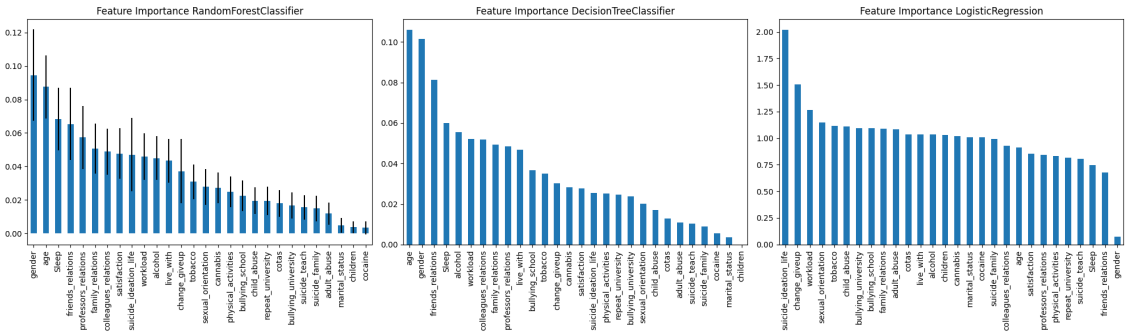
In this variation, we see that the accuracy for men was kept almost the same as the original dataset, as it can be seen in Figure 5.30, varying around 2% between the algorithms. For women, we see that the performance was slightly affected when comparing to the predictions on the original dataset, with logistic regression having around 7% less accuracy on this variation. We also see that some models had clear variations on the false positive rates, as it would be expected by the proportion of women with negative output in this dataset variation. The KL and KS values for this dataset (0.786 and 0.452) also indicate that the training set has more representation in one facet for the unprivileged group. Figure 5.31 shows the feature importance for this dataset variation. We see that gender was important in both Random Forest and Decision Tree, whilst in Logistic Regression it did not make significant impact.

Figure 5.30 – Chart for gender on the highly unbalanced (gender) Depression in Medical Students Dataset.



Source: The Author

Figure 5.31 – Feature importance for the highly unbalanced (gender) Depression in Medical Students Dataset.



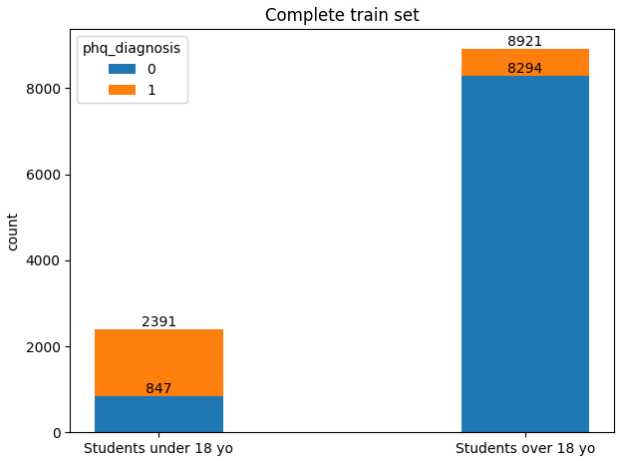
Source: The Author

5.3.3 Highly unbalanced in relation to age

To generate a highly unbalanced dataset for the **age** attribute, it was necessary to drop instances for the positive outcome from the privileged group for this attribute. So, 95% of the students over 18 years old with the target value equals to one were removed from the dataset. The complete train set used for this variation can be found in Figure 5.32.

In this variation, it was possible to increase the values for the pre-training metrics for all metrics, with all pre-training bias metrics related to this feature reporting values

Figure 5.32 – Train sets for the highly unbalanced (age) Depression in Medical Students Dataset.



Source: The Author

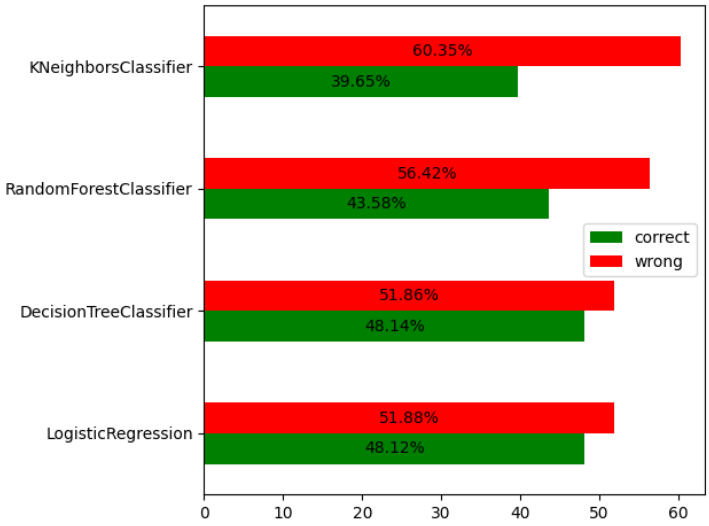
above 0.5. The full table with the metrics for this variation can be found in Table 5.5, and the performance for the models can be found in Table 5.6. The feature importance analysis for this dataset can be found in Figure 5.35, with age being the most important feature for Random Forest and Decision Tree. For the Logistic Regression model, age was not among the most important attributes. Each training iteration had an average of 1885 instances, while the original dataset average training size was around 3871.

For students under 18 years old, we see that when comparing to the original dataset, KNN showed the biggest difference in accuracy, reporting a great portion of values with false negative, from 5.85% in the original dataset to 59.24% in this variation. It is worth noticing that the training set has more individuals related to the positive output for this class. This might indicate that the decision boundary in the KNN algorithm was not using the age attribute for the decision process as much as the other algorithms. For all other algorithms, the accuracy was on average 3% lower than the original dataset.

Analyzing students over 18 years old, the performance for all algorithms was highly impacted. Comparing with the original dataset, we see that all models were affected. If we look at the accuracy, we see that all models labeled fewer instances correctly than the original model, with an average of 24% less accuracy across the models, in Figure 5.33, we report for each model the number of correct predictions compared with the number of wrong predictions. The false negative rate in these models was also increased, with most of the errors being false negatives. The complete chart for this model variation can be found in Figure 5.34.

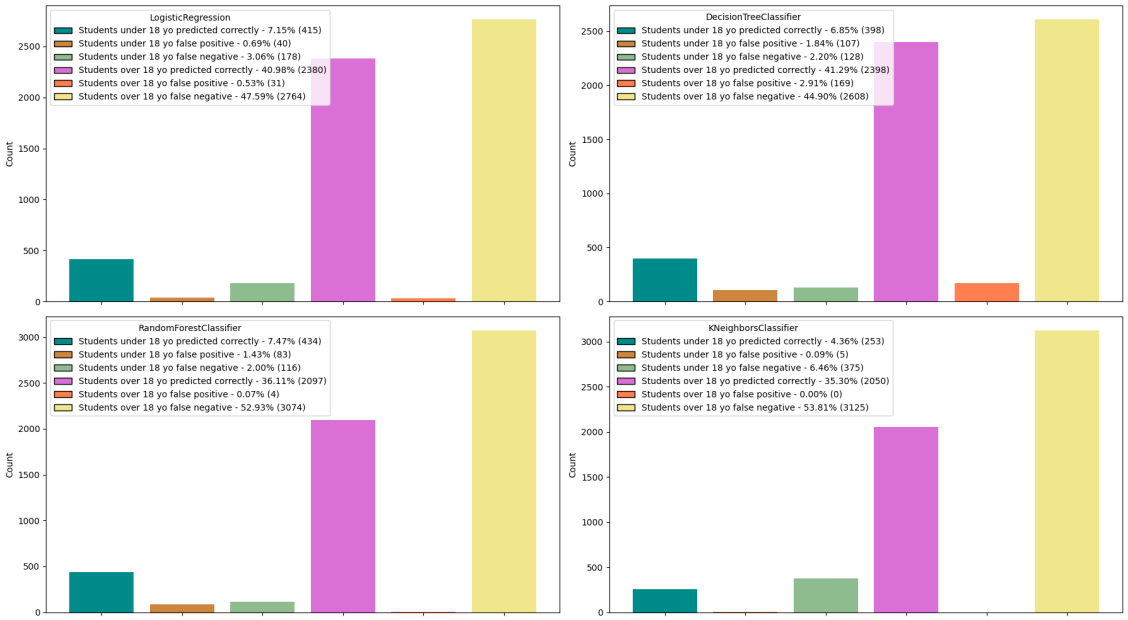
Analyzing the feature importance results for this variation of dataset, we observed

Figure 5.33 – Percentage of correct predictions for each algorithm in the highly unbalanced High (age) Depression in Medical Students Dataset.



Source: The Author

Figure 5.34 – Chart for age on the Highly unbalanced (age) Depression in Medical Students Dataset



Source: The Author

that age was by far the most important feature for the Random Forest and Decision Tree classifiers. For Logistic Regression, age was the least important feature. The most important feature for this model was marital_status and change_giveup. The results are shown in Figure 5.35.

5.3.4 Equally balanced

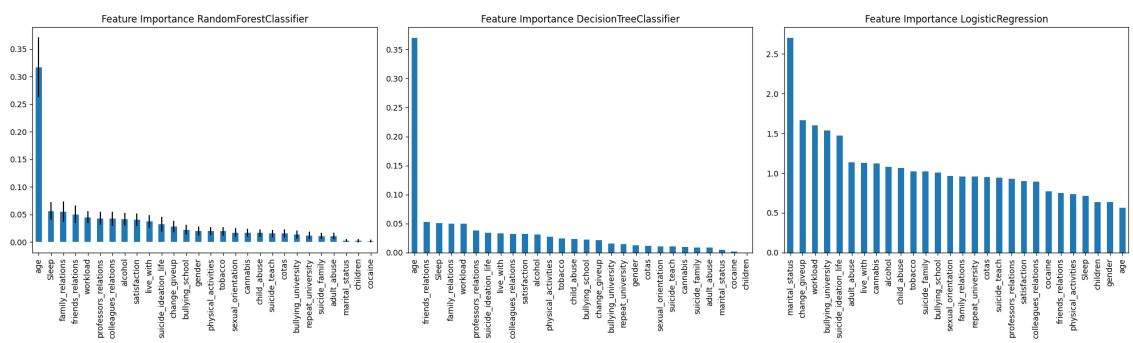
To equally balance the original dataset, we had to remove students over 18 years old and women, checking all the possible combinations for these features and the target attribute and then removing instances from the training set. The complete train set can be found in Figure 5.36. The metrics values in Table 5.5 show that for this variation, all pre-training bias metrics were very close to zero. The downside of this variation is that the number of entries in each training set was drastically reduced, from around 3871 instances in the original dataset to an average of 176 on each training iteration for this variation. The accuracy and F1-Score for this dataset variation can be found in Table 5.5.

The performance for this variation decreased in all training algorithms, with an increase of over 10% in the error rate when compared to the original dataset for the KNN algorithm for the unprivileged class in the gender attribute. The average decrease in accuracy for the gender attribute in the unprivileged class was around 6.5% for this metric when compared to the original dataset, and around 5% for the privileged class.

In the age attribute, we see that the privileged class decreased in accuracy for around 6% for the privileged class, while the unprivileged class decreased over 8% across the four models. The chart for the complete dataset for the protected attributes age and gender can be found in figures 5.38 and 5.37, respectively.

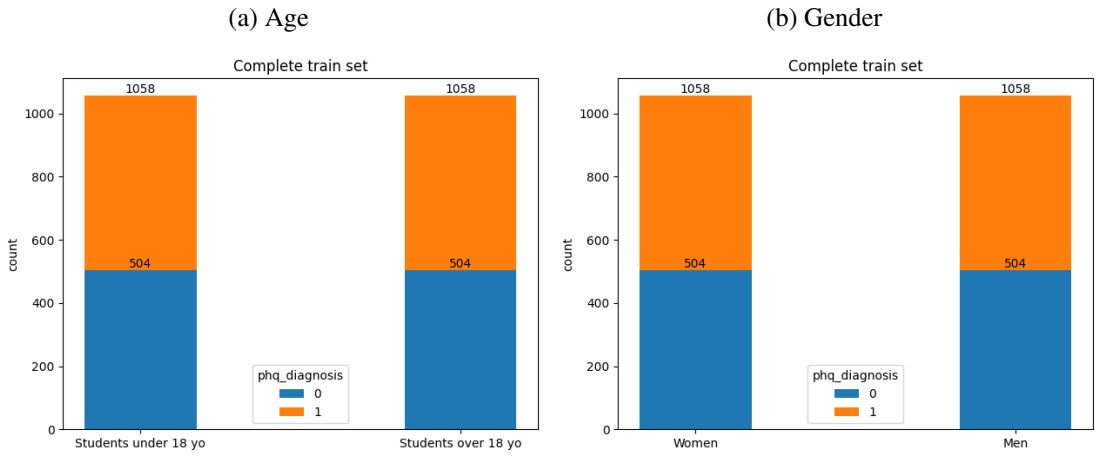
By looking at the feature importance for this dataset, it is possible to see that age is still among the most relevant attributes for Random Forest and Decision Tree, but the same behavior is not present in Logistic Regression. Comparing the results with the unbalanced datasets, the performance in this variation was closer to the highly unbalanced

Figure 5.35 – Feature importance for the highly unbalanced (age) Depression in Medical Students Dataset



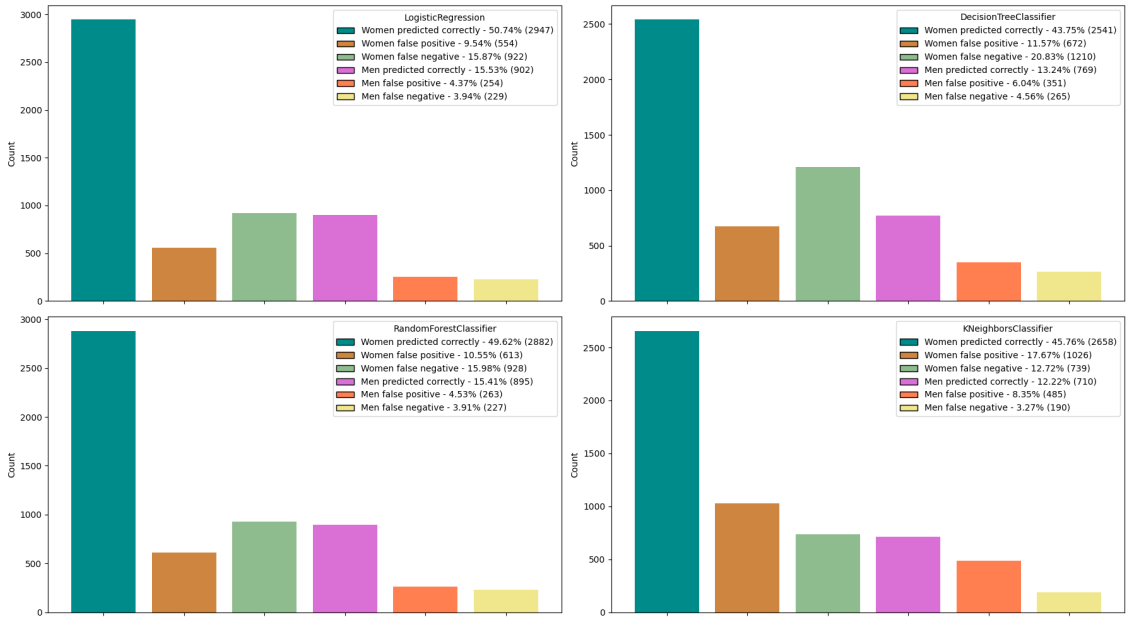
Source: The Author

Figure 5.36 – Train sets for the Equally balanced Depression in Medical Students Dataset.



Source: The Author

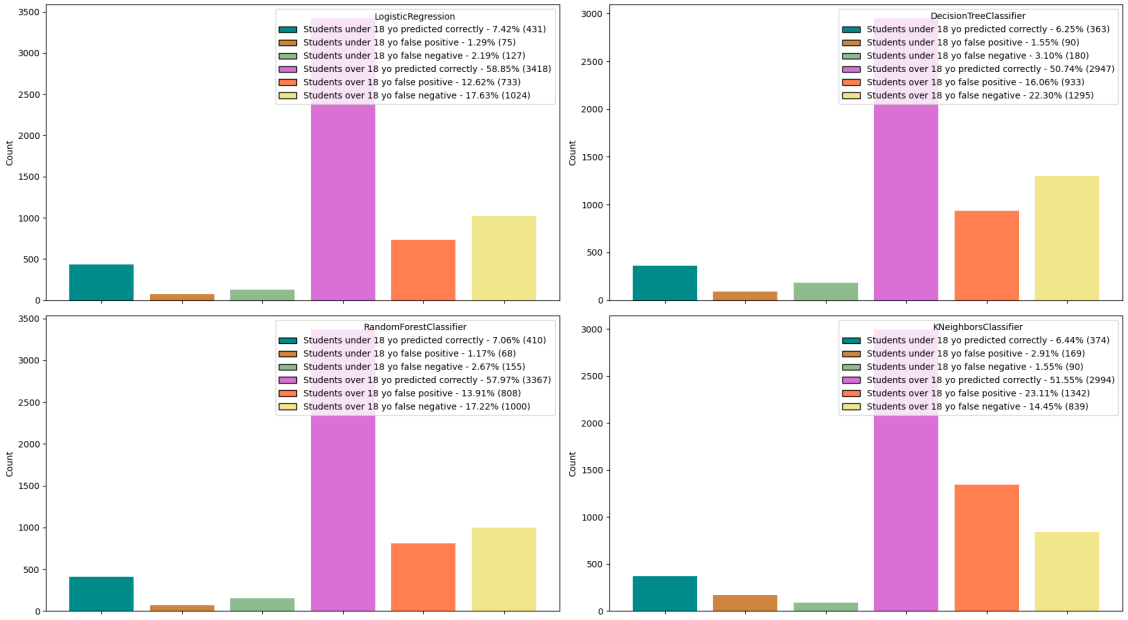
Figure 5.37 – Chart for gender in the equally balanced Depression in Medical Students Dataset.



Source: The Author

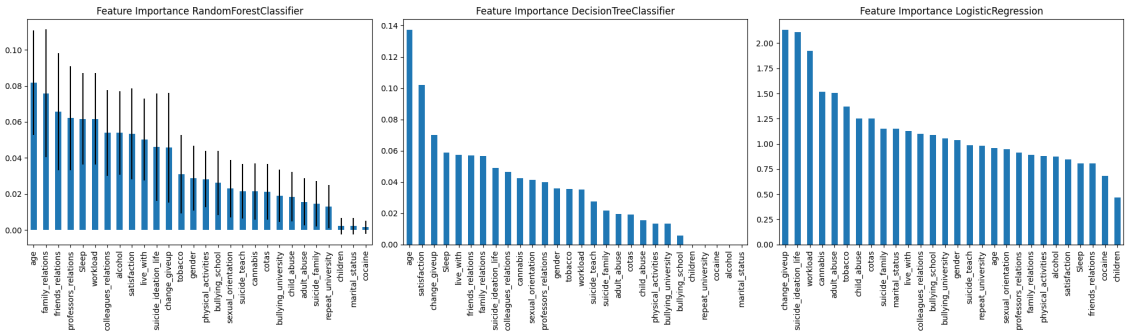
on gender than to the highly unbalanced on age. The graph containing the complete feature importance for this variation can be found in Figure 5.23.

Figure 5.38 – Chart for age in the equally balanced Depression in Medical Students Dataset.



Source: The Author

Figure 5.39 – Feature importance for the Equally Balanced Depression in Medical Students Dataset



Source: The Author

6 CONCLUSION

Fairness in Machine Learning has been a hot topic in recent years. The challenge of automating human decisions without incorporating inherent biases has been the focus of several works, and is a challenge that needs to be addressed before we are able to rely on any decision made by an algorithm. Defining algorithm fairness itself is complex and varies by location and domain. The notion of algorithm fairness itself is not something easily defined, as it can vary according to the domain in which the models are deployed, and what is valid in one country may not be the same in other countries (WACHTER; MITTELSTADT; RUSSELL, 2020).

Pre-training bias metrics can quantify how datasets are positioned in a spectrum of distribution for different protected attributes. These values can represent the population from which the data was collected or reveal issues with access and equality to certain groups if the values do not reflect how society is in general or in reality. Although protected attributes may be relevant to a study, such as in gender-related disease research, relying solely on these attributes can lead to inaccurate models. In these cases, it makes sense to have more instances from the gender more affected by the disease in question. However, when the trained model is deployed in another hospital or location where this disease is not common or where the prevalence is not so high for that specific gender, the model may generate wrong predictions due to inherent bias.

The analysis conducted in this work intends to provide preliminary evidence that health data might cause harm depending on the way it was acquired for the study. Highly unbalanced datasets can result in mislabeling, which in turn can lead to misdiagnosing a patient's health status. Table 6.1 shows the average accuracy and F1-Score for all models used in this study, along with the average accuracy for all the protected attributes (values are separated by comma for the datasets with multiple protected attributes, (*i.e., sex, race* for the Intersectional-bias, and *gender, age* for the Depression in Medical Students Dataset).

Table 6.1 – Comparative table for all the experimental results obtained in this work.

	Dataset Variation	Average Accuracy	Average F1-score	Average Accuracy Unprivileged Class	Average Accuracy Privileged Class
Intersectional-Bias	Original Dataset	87.638	88.100	81.310, 85.846	95.343, 90.733
	Highly Unbalanced	81.311	83.618	70.477, 76.576	94.499, 89.489
	Equally Balanced	86.795	87.595	80.175, 84.064	94.856, 91.513
Heart Attack	Original Dataset	78.115	79.846	80.000	77.294
	Highly Unbalanced	78.730	80.936	81.622	77.471
	Equally Balanced	71.639	78.036	84.595	66.000
Depression in medical students	Original Dataset	67.794	74.702	68.885, 70.537	64.440, 67.459
	Highly Unbalanced (gender)	66.012	75.933	66.623, 70.774	64.061, 65.43
	Highly Unbalanced (age)	44.873	21.342	42.624, 59.242	52.058, 43.116
	Equally Balanced	61.570	66.247	62.333, 62.322	59.134, 61.478

In the Intersectional-Bias Dataset, changing the proportion of protected attributes resulted in lower accuracy on the unprivileged classes (*i.e.,* a decrease of about 10%) and small variation of about 1% for the privileged class. The models trained for this dataset showed smaller values for the highly unbalanced variation of the dataset. However, these values alone do not indicate how protected classes were affected, nor do they provide information about how the dataset performed on the unprivileged classes regarding false positives or false negatives, which depending on the context might mean a different form of harm for that class. When we applied manual variations to the dataset aiming at lowering the pre-training bias metrics values to reduce the bias from the dataset, we were able to observe higher accuracy on the unprivileged classes, and even some increase on performance for the privileged classes on the dataset.

In the Heart Attack Dataset, when the bias was introduced and the pre-training bias metrics values increased, we did not observe a high impact on the performance of models. On the contrary, we noted a curious improvement, with the accuracy for both the privileged and unprivileged classes showing a slight increase in relation to the original dataset. This observation can be related to the impact caused by randomly removing instances from the dataset, or to how the data was gathered in the first place. Also, the fact that the test sets have less female representation could imply in higher performance. The analysis of features importance for this dataset did not return high values for the protected attributes, which may corroborate the results of no significant changes in performance for a highly unbalanced dataset. When we aimed to reduce the values of pre-training bias metrics, the performance of the models was impacted. We observed an increase in the accuracy for the unprivileged value, while the privileged class seemed to be negatively impacted with the manual modification of the dataset distribution. Thus, we raise the question whether this dataset needs the imbalance to actually perform well on both classes.

For the Depression in Medical Students Dataset, two different situations emerged from the introduced bias. In the analysis related to the gender attribute, we had to remove instances from the class with the majority of values, such that the training sets were highly reduced when compared to the original dataset (*i.e.,* approximately 28% of the original dataset). So, the expectation was that the performance would be the worst of the study. However, this expectation did not translate into the experimental results, and the average accuracy for the complete dataset and for the protected attributes were very close to the original dataset variation. In the second highly unbalanced variation for this dataset, re-

lated to the age attribute, we focused in keeping all the metrics high at the same time and not worrying about the imbalance for the attribute itself, as the original variation already had a high class imbalance. The result for the second variation showed that unbalancing the labels by removing the positive output representation from the privileged class caused a huge impact in the accuracy for the protected attributes, both in the privileged and unprivileged classes. This variation also significantly impacted the overall performance of models, especially in terms of F1-score, allowing us to conclude that for this dataset, age could be a source of bias when predicting the outcome (*i.e.,* diagnostic of depression). The equally balanced dataset in this scenario showed that even with fewer instances in the training set (around 352 per algorithm iteration), if the data is balanced and represents equally different portions of the population, it can reach satisfactory results. This is especially true when comparing with the highly unbalanced on age version, since the equally balanced had fewer instances and reported higher accuracy and F1-score.

Through the experiments conducted in this work, we conclude that analyzing the pre-training bias metrics for a dataset before deploying the trained model can be really beneficial to the study. The pre-training bias metrics can indicate issues in the quality or in the distribution of the data, avoiding us to spend a large time or too many computational resources in training models with biased data. Moreover, the analysis of the pre-training bias can help explain some results obtained with the trained model, especially if the performance is not satisfactory, suggesting direction in which the data collection should be improved if possible (*e.g.,* attributes values that should be added to the dataset to enhance the predictive performance). Even though we had some results showing how the pre-training bias metrics can reflect on the overall performance of models, or more specifically on the false positive and false negative rates, some questions remain open and deserve further investigation. For instance, how can we mitigate the bias inherent to the data and reduce the original values of the pre-training bias metrics without removing instances from the dataset? How can we judge if an attribute should be protected or not? Is there a pattern to be identified in the dataset, or can we expect that the protected attributes are the same for all datasets considering a common domain (*e.g.,* Health)? Recent studies have attempted to address some of these questions, emphasizing the need for data scientists to take responsibility for the machine learning models they develop. It is crucial to ensure that these models accurately represent our society and that no group is excluded or harmed by the limited predictive power of models due to their personal characteristics used as model inputs.

REFERENCES

ADEBAYO, J. Fairml: Toolbox for diagnosing bias in predictive modeling. 2016.

ALELYANI, S. Detection and evaluation of machine learning bias. **Applied Sciences (Switzerland)**, MDPI AG, v. 11, 7 2021. ISSN 20763417.

BELLAMY, R. K. E. et al. Ai fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. 10 2018. Available from Internet: <http://arxiv.org/abs/1810.01943>.

BREIMAN, L. Arcing classifiers 1. **The Annals of Statistics**, v. 26, p. 849, 1998.

BREIMAN, L. **Random Forests**. [S.l.: s.n.], 2001. 5-32 p.

BRUGMAN, S. **pandas-profiling: Exploratory Data Analysis for Python**. 2019. <https://github.com/pandas-profiling/pandas-profiling>. Version: 2.X, Accessed: 12/25/2022.

CHEN, I. Y. et al. Ethical machine learning in healthcare. **Annual Review of Biomedical Data Science Annu. Rev. Biomed. Data Sci**, v. 2021, p. 123–144, 2021. Available from Internet: <https://doi.org/10.1146/annurev-biodatasci-092820->.

FLETCHER, R. R.; NAKESHIMANA, A.; OLUBEKO, O. Addressing fairness, bias, and appropriate use of artificial intelligence and machine learning in global health. **Frontiers in Artificial Intelligence**, Frontiers Media S.A., v. 3, 4 2021. ISSN 26248212.

GHASSEMI, M. et al. A review of challenges and opportunities in machine learning for health. **AMIA Summits on Translational Science Proceedings**, American Medical Informatics Association, v. 2020, p. 191, 2020.

GOUTTE, C.; GAUSSIER, E. A probabilistic interpretation of precision, recall and f-score, with implication for evaluation. In: LOSADA, D. E.; FERNÁNDEZ-LUNA, J. M. (Ed.). **Advances in Information Retrieval**. Berlin, Heidelberg: Springer Berlin Heidelberg, 2005. p. 345–359. ISBN 978-3-540-31865-1.

HARDT, M. et al. Amazon sagemaker clarify: Machine learning bias detection and explainability in the cloud. 9 2021. Available from Internet: <http://arxiv.org/abs/2109.03285http://dx.doi.org/10.1145/3447548.3467177>.

JUHN, Y. J. et al. Assessing socioeconomic bias in machine learning algorithms in health care: a case study of the houses index. **Journal of the American Medical Informatics Association**, Oxford University Press, v. 29, n. 7, p. 1142–1151, 2022.

JúNIOR, R. L. I. et al. Justiça nas previsões de modelos de aprendizado de máquina:um estudo de caso com dados de reincidência criminal. 2022.

KRAMER, O. Dimensionality reduction with unsupervised nearest neighbors. **INTELLIGENT SYSTEMS REFERENCE LIBRARY**, v. 51, 2013. Available from Internet: <http://www.springer.com/series/8578>.

MANDHALA, V. N. et al. Detecting and mitigating bias in data using machine learning with pre-training metrics. **Ingenierie des Systemes d'Information**, International Information and Engineering Technology Association, v. 27, p. 119–125, 2 2022. ISSN 21167125.

MARCON, G. et al. Who attempts suicide among medical students? **Acta Psychiatrica Scandinavica**, Wiley Online Library, v. 141, n. 3, p. 254–264, 2020.

MASLEJ, M. et al. **Intersectional-Bias-Assessment**. 2022. Available from Internet: <http://openml1.win.tue.nl/dataset45040/dataset_45040.pq>.

MEHRABI, N. et al. A survey on bias and fairness in machine learning. 8 2019. Available from Internet: <http://arxiv.org/abs/1908.09635>.

MIOTTO, R. et al. Deep learning for healthcare: review, opportunities and challenges. **Briefings in Bioinformatics**, Oxford University Press, v. 19, n. 6, p. 1236–1246, 2018.

NAVARRO, C. L. A. et al. Completeness of reporting of clinical prediction models developed using supervised machine learning: a systematic review. **BMC Medical Research Methodology**, Springer, v. 22, p. 1–13, 2022.

NEWMAN, D. et al. **UCI Repository of machine learning databases**. 1998. Available from Internet: <http://www.ics.uci.edu/~mlearn/MLRepository.html>.

NOSEWORTHY, P. A. et al. Assessing and mitigating bias in medical artificial intelligence: The effects of race and ethnicity on a deep learning model for ecg analysis. **Circulation: Arrhythmia and Electrophysiology**, Lippincott Williams and Wilkins, 3 2020. ISSN 19413084.

OBERMEYER, Z. et al. Dissecting racial bias in an algorithm used to manage the health of populations. **Science**, American Association for the Advancement of Science, v. 366, n. 6464, p. 447–453, 2019.

PARK, Y. et al. Comparison of methods to reduce bias from clinical prediction models of postpartum depression. **JAMA Network Open**, American Medical Association, v. 4, 4 2021. ISSN 25743805.

PEDREGOSA, F. et al. Scikit-learn: Machine learning in Python. **Journal of Machine Learning Research**, v. 12, p. 2825–2830, 2011.

PEREIRA, F. de Á. **Analisando padrões no consumo de álcool entre estudantes de Medicina brasileiros: uma abordagem de aprendizado de máquina**. Monografia (TCC) — Universidade Federal do Rio Grande do Sul, 2020.

PFOHL, S. R.; FORYCIARZ, A.; SHAH, N. H. An empirical characterization of fair machine learning for clinical risk prediction. **Journal of Biomedical Informatics**, v. 113, p. 103621, 2021. ISSN 1532-0464. Available from Internet: <https://www.sciencedirect.com/science/article/pii/S1532046420302495>.

RASCHKA, S.; MIRJALILI, V. **Python machine learning : machine learning and deep learning with python, scikit-learn, and tensorflow 2**. [S.l.: s.n.], 2019. 741 p. ISBN 9781789955750.

STEVENS, L. M. et al. Recommendations for reporting machine learning analyses in clinical research. **Circulation: Cardiovascular Quality and Outcomes**, Am Heart Assoc, v. 13, n. 10, p. e006556, 2020.

SURESH, H.; GUTTAG, J. V. A framework for understanding sources of harm throughout the machine learning life cycle. 1 2019. Available from Internet: <http://arxiv.org/abs/1901.10002http://dx.doi.org/10.1145/3465416.3483305>.

TRAMER, F. et al. Fairtest: Discovering unwarranted associations in data-driven applications. **arXiv preprint arXiv:1510.02377**, 2015.

WACHTER, S.; MITTELSTADT, B.; RUSSELL, C. **WHY FAIRNESS CANNOT BE AUTOMATED: BRIDGING THE GAP BETWEEN EU NON-DISCRIMINATION LAW AND AI**. 2020.

ZEHLIKE, M. et al. Fairness measures: Datasets and software for detecting algorithmic discrimination. 7 2017. Available from Internet: <http://fairness-measures.org>.

ZHANG, H. et al. Hurtful words. In: . [S.l.]: Association for Computing Machinery, Inc, 2020. p. 110–120. ISBN 9781450370462.