University of Genova

Double Master's Degree in Interdisciplinary and Innovative Engineering and Computer Engineering

Mitigating Regressiveness in Accuracy and Fairness in Machine Learning

by

Anna Pallarès López

Thesis submitted for the degree of *Computer Engineering* under the Mobility Framework together with the Polytechnic University of Catalonia.

September 2024

Luca Oneto	Supervisor
Irene Buselli	Supervisor
Raul Benítez	UPC Supervisor

Dibris

Department of Informatics, Bioengineering, Robotics and Systems Engineering

To my grandmother, who passed away while I was abroad in Genova working on the present dissertation. As you always told me "*No one will ever be able to take the knowledge you acquire, and this is the most important thing in life*".

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements. This dissertation contains fewer than 65,000 words including appendices, bibliography, footnotes, tables and equations and has fewer than 150 figures.

Anna Pallarès López
September 2024

Acknowledgements

I would like to express my deepest gratitude to Prof. Luca Oneto for providing me with the opportunity to join his research group by developing this work, which has clearly marked a turning point in my professional career. His unwavering belief in my potential and continuous support have been fundamental to my progress.

Additionally, I would like to extend my appreciation to Dr. Irene Buselli for her strong support throughout the development of my work. I am sure I could not have undertaken this journey without her since Irene's willingness to share her expertise so generously and patiently has been invaluable to my research. Her meticulous attention to detail and profound knowledge have greatly enriched my understanding of the subject, and her constructive feedback has pushed me to work toward excellence in all aspects of my work.

To Raul Benitez, my home university supervisor, I would like to express my deepest appreciation for opening the doors for me to pursue this period abroad and for giving my career a significant boost. Also many thanks to Yolanda Vidal for dedicating her time to supervise and correct my work.

I am also thankful for the personal support that has been equally vital. My friends, Alessandro, Igor, Elia, Mendy, Giacomo, Matthieu, Lorenzo, Michelle, and Leonardo, have not only been exceptional classmates but have also made my time in Genova unforgettable. Their friendship and shared experiences have been a cornerstone of my journey.

My partner, Eduard, deserves special recognition for his unconditional support; for being both a partner and a colleague. He has always empowered me to overcome challenges and planted in me the belief that I can achieve anything. Not only has he been always there willing to empower me and provide support, but Eduard's role in my journey goes further; his intellectual contributions and critical insights have also enriched my academic work.

Together, we’ve tackled complex problems, transforming the daunting task of research into a collaborative and more rewarding experience.

I would also like to acknowledge my brother, Roger, whose brilliance and expertise in the field of machine learning have been a constant source of help and inspiration. His ability to solve complex problems effortlessly has greatly aided my research.

Lastly, I must thank my parents, without whom none of this would have been possible. Their assistance in all aspects allowed me to study abroad and pursue my dreams, which is an enormous privilege for which I am profoundly grateful. They have always believed in my potential and supported every decision I’ve made, providing wise counsel and invaluable perspectives.

Without the collective support of all these people, I would not be here today presenting this work. I am deeply grateful to each of them for their roles in shaping my path.

Abstract

The rapid advancement of Artificial Intelligence (AI) and Machine Learning (ML) technologies has positively transformed numerous industries but also introduced challenges regarding their trustworthiness and ethicality. A critical issue arises when newly updated models correct past wrong predictions but simultaneously disrupt already-correct ones leading to a perceived regression in performance, namely *regressiveness*. Especially, when this degradation in performance of some samples is not balanced with respect to a sensitive attribute (e.g. gender or ethnicity), a discriminatory model emerges. Therefore, the intersection between regressiveness and unfairness in machine learning is foremost studied in this work by presenting novel methodologies to alleviate this phenomenon.

The study proposes two novel mitigation strategies within the framework of our new fairness metric, unfair regression, based on the difference in the negative flips phenomenon, which quantifies the disparity in disrupted predictions between sensitive and non-sensitive groups in an updated model. By focusing on minimizing this metric, we introduce the first mitigation approach that only affects the model selection phase. We implement a double-step cross validation algorithm that accounts for both accuracy and unfair regression minimization. The second mitigation algorithm directly affects the learning phase, adding the unfair regression metric as a constraint within the Support Vector Machine (SVM) framework, yielding to our Unfair-Regression-Free SVM (URFSVM). These algorithms contribute on reducing bias and enhance fairness in predictive performance. Both approaches ensure that model updates do not disproportionately affect any particular group, promoting more equitable and trustworthy machine learning systems.

The objectives of this study include conducting an in-depth review of current bias mitigation techniques as well as regression in performance, studying the intersection between them, developing novel in-processing methods to minimize regressiveness, evaluating these methods between them and against standard and fairness-enhanced models, and integrating the proposed approaches into existing machine learning frameworks without compromising

performance. On the whole, this work intends to contribute to the development of AI technologies that align with ethical standards and international human rights, through our novel strategies, ensuring that no individual is discriminated against based on gender, ethnicity, disabilities, or social status.

Abstract

El ràpid avenç de la Intel·ligència Artificial (IA) i l'aprenentatge automàtic, o *Machine Learning* (ML), ha transformat positivament nombroses indústries, però també ha plantejat reptes significatius pel que fa a la fiabilitat i ètica de presa de decisions d'aquests models. Un problema crític es presenta quan l'actualització d'un model corregeix prediccions que anteriorment eren errònies, però simultàniament modifica aquelles que ja eren correctes. Aquest fenomen, anomenat regressió en el rendiment, o *regressiveness*, es dona a través de girs negatius, o *Negative Flips*, generat així, una percepció de no haver millorat el model després de l'actualització. Específicament, quan aquesta reducció de precisió no es dona d'igual forma respecte atributs sensibles, com pot ser el gènere o l'ètnia, es crea un model discriminatori. Així doncs, en aquest treball s'estudia la intersecció de dos fenomens més o menys estudiats en el món de recerca, que no han estat estudiats en conjunt: la degradació en el rendiment i l'equitat en els models d'aprenentatge automàtic.

En aquest estudi es proposen dues noves estratègies per a mitigar aquest fenomen dins del marc de la nova mètrica d'avaluació que proposem, *Unfair Regression*, que es basa en els *Negative Flips*. En minimitzar aquesta mètrica, s'introdueix la primera estratègia de mitigació que només afecta la fase de selecció del model, o *model selection*. S'implementen dos passos en el procés de *cross validation*, fent que l'algorisme tingui en compte la minimització d'ambdós mètriques: la precisió i l'equitat. En la segona estratègia de mitigació, l'algorisme afecta directament a la fase d'aprenentatge, incorporat com a restricció la mètrica desenvolupada, *Unfair Regression*, dins del marc de Support Vector Machines (SVMs), creant així el model anomenat *Unfair-Regression-Free SVM (URFSVM)*. Ambdues estratègies contribueixen en reduir el biaix i en millorar l'equitat en l'aprenentatge automàtic, assegurant que les actualitzacions del model no afectin negativament de forma desproporcionada a cap grup en particular, promovent així resultats més fiables, robustos i justos.

Els objectius d'aquest estudi inclouen realitzar una revisió exhaustiva de les tècniques actuals de mitigació de biaixos així com de la regressió en el rendiment, estudiar la intersecció entre aquests fenòmens, desenvolupar nous mètodes del tipus *in-processing* per minimitzar aquesta regressió, avaluar aquests mètodes entre ells i en comparació amb els models estàndards existents, i integrar els mètodes proposats en els marcs existents d'aprenentatge automàtic sense comprometre el rendiment. En conjunt, aquest treball té el propòsit de contribuir al desenvolupament de tecnologies d'IA que s'alineïn amb els estàndards ètics i els drets humans internacionals, assegurant que cap individu sigui discriminat per raons de gènere, ètnia, discapacitat o estatus social.

Abstract

El rápido adelanto de la Inteligencia Artificial (IA) y el aprendizaje automático, o *Machine Learning* (ML), ha transformado positivamente numerosas industrias, pero también ha planteado retos significativos en cuanto a la fiabilidad y ética de toma de decisiones de estos modelos. Un problema crítico se presenta cuando la actualización de un modelo corrige predicciones que anteriormente eran erróneas, pero simultáneamente modifica aquellas que ya eran correctas. Este fenómeno, denominado regresión en el rendimiento, o *regressiveness*, se da a través de giros negativos, o *Negative Flips*, generado así, una percepción de no haber mejorado el modelo después de la actualización. Específicamente, cuando esta reducción de precisión no se da de igual forma respeto atributos sensibles, como puede ser el género o la etnia, se crea un modelo discriminatorio. Así pues, en este trabajo se estudia a intersección de dos fenomenos más o menos estudiados en el mundo de investigación, que no han sido estudiados en conjunto: de la degradación en el rendimiento y la equidad en los modelos de aprendizaje automático.

En este estudio se proponen dos nuevas estrategias para mitigar este fenómeno dentro del marco de la nueva métrica de evaluación que propongamos, *Unfair Regression*, que se basa en los *Negative Flips*. Al minimizar esta métrica, se introduce la primera estrategia de mitigación que solo afecta la fase de selección del modelo, o *model selection*. Se implementan dos pasos en el proceso de *cross validation*, haciendo que el algoritmo tenga en cuenta la minimización de ambos métricas: la precisión y la equidad. En la segunda estrategia de mitigación, el algoritmo afecta directamente a la fase de aprendizaje, incorporado como restricción la métrica desarrollada, *Unfair Regression*, dentro del marco de *Support Vector Machines* (SVMs), creando así el modelo llamado *Unfair-Regression-Free SVM (URFSVM)*. Ambas estrategias contribuyen al reducir el sesgo y al mejorar la equidad en el aprendizaje automático, asegurando que las actualizaciones del modelo no afecten negativamente de forma desproporcionada a ningún grupo en particular, promoviendo así resultados más fiables, robustos y justos.

Los objetivos de este estudio incluyen realizar una revisión exhaustiva de las técnicas actuales de mitigación de sesgos así como de la regresión en el rendimiento, estudiar la intersección entre estos fenómenos, desarrollar nuevos métodos del tipo *in-processing* para minimizar esta regresión, evaluar estos métodos entre ellos y en comparación con los modelos estándares existentes, e integrar los métodos propuestos en los marcos existentes de aprendizaje automático sin comprometer el rendimiento. En conjunto, este trabajo pretende contribuir al desarrollo de tecnologías de IA que se alineen con los estándares éticos y los derechos humanos internacionales, asegurando que ningún individuo sea discriminado por razones de género, etnia, discapacidad o estatus social.

Table of contents

List of figures xiii

List of tables xv

Nomenclature xvi

I Section One 1

1 Introduction 2

1.1 Background 3

1.2 Significance 5

1.3 Objectives 5

2 Literature Review 7

2.1 Related Work 7

2.2 Gap in Literature 9

3 Theoretical Background 11

3.1 Fundamentals of Machine Learning 11

3.1.1 The Supervised Learning Problem 12

3.1.2 The Classification Problem: Support Vector Machines 13

3.1.3 Model Selection and Error Estimation 15

3.2 Bias and Fairness in Machine Learning 17

3.2.1 Definitions and Forms of Bias 18

3.2.2 Sensitive Variables: the Unprivileged Groups 19

3.2.3 Metrics for Predictive Analysis and Fairness 20

3.3 Regression in Model Updates 24

3.3.1 Unfair Regression 25
3.4 Approaches to Mitigate Bias: Fair Algorithms 26
3.5 SVM Under Fairness Constraints: Unfair-Regression-Free SVM 27
3.5.1 Optimization Problem 29
3.6 Challenges in Bias Mitigation: the Trade-off Between Accuracy and Fairness 30

II Section Two 32

4 Methodology 33
4.1 Research Design Study Overview 33
4.1.1 Study Questions and Hypothesis 34
4.2 Data Collection and Preparation 35
4.2.1 Data Sources 35
4.2.2 Data Preparation 37
4.3 Mitigation Strategies 37
4.3.1 Mitigation (i): Double-Step Cross Validation 39
4.3.2 Mitigation (ii): Unfair-Regression-Free Support Vector Machine . . 39
4.4 Evaluation Metrics 40

5 Results 42
5.1 Reference Models: Standard SVM 42
5.2 Mitigation Strategy (i): Double-Step Cross Validation Method 45
5.3 Mitigation Strategy (ii): Unfair-Regression-Free SVM Method 46
5.4 Comparison through Methods 47

6 Discussion 49
6.1 Interpretation of Findings 49
6.2 Comparison with Prior Studies 50
6.3 Limitations of the Study 50

7 Conclusions 52

References 54

List of figures

3.1 Graphical representation of the supervised learning problem. 12

3.2 Linear Support Vector Machine classification based on two predictors (x_1, x_2). Classification hyperplane (-) and maximal margins (- - -) are shown. . . 13

3.3 Double step cross validation exemplification. This method is employed to determine the hyperparameters by taking into account both accuracy and fairness metrics. A filtered list of the best accuracies is used for a second cross validation where the negative flip metric is considered. 17

3.4 A famous example of two criminals that were scored using the COMPAS algorithm. Bernard Parker, left, was rated high risk; Dylan Fugett was rated low risk despite of his critical history. Source: Josh Ritchie for ProPublica [1]. 18

3.5 Confusion matrix and classification metrics for a binary classification problem. 21

3.6 Regression in model update: when updating an old classifier (-) to a new one (- -), we can introduce errors that the old classifier did not make (negative flips, bottom-left, red). 25

3.7 Overview of the fair approaches in the machine learning pipeline. 27

4.1 Machine learning sysems-based pipeline. 34

5.1 Fairness metrics for the standard SVM. On the left, the normalized difference in equal opportunity is plotted against the normalized misclassification error. On the right, the same is done for the difference in demographic parity. Note that, the closer a point to the origin is, the better the results are in both accuracy and fairness. 43

5.2 Normalized percentage of lost accuracy in model updates and unfair regression (UR). The closer the data point to the origin, the better performance in both accuracy and fairness. Note that the subscript numeration for the model notation designates the model configuration in terms of dataset size and kernel function, as listed in Table 4.2. 45

5.3 Normalized UR (the smallest the better) and inverted accuracy gain (the smallest the better) for each dataset and updating scenario. Gray lines join the results of the different methods on the same dataset. Note that, for meaningful plots, the inverted accuracy gain has been computed by subtracting the max value from each accuracy gain value and then normalized. Results retrieved from Table 5.2. 48

List of tables

2.1 Highlighted works in the current literature on fairness in ML within in-processing techniques. 8

2.2 Highlighted works on regression in model updates of ML systems. 9

3.1 Classification metrics. 22

4.1 Dataset characteristics overview with the related statistics, and the sensitive features involved. Gender considers the two groups as male and female; ethnicity considers the ethnic group as white and other ethnic groups; foreign considers being or not being a foreign person. 37

4.2 Model configurations varying subset sizes and types of kernel functions. . . 38

4.3 Considered scenarios for model updating. The new model is the enhanced classifier, while the old model performs as a reference. 38

5.1 Performance and fairness metrics for the datasets with different kernel methods and dataset sizes. 43

5.2 Accuracy gain (ACCG) and unfair regression (UR) tested on a model update scenario, for the different cases considered going from f^*_{old} to f^*_{new}. 44

5.3 Overall accuracy (ACC), balanced accuracy (B-ACC), unfair regression (UR), and accuracy gain (ACCG) obtained when using the 2-step CV method on the three different updating scenarios for various datasets with $f^*_{old} \rightarrow f^*_{new}$. 46

5.4 Overall accuracy (ACC), balanced accuracy (B-ACC), unfair regression (UR), and accuracy gain (ACCG) obtained when using the unfair-regression-free SVM (URFSVM) method on the three different updating scenarios for various datasets with $f^*_{old} \rightarrow f^*_{new}$. 47

Nomenclature

Roman Symbols

$\mathbf{a}$	Coefficient vector for fairness constraint in SVM
b	Bias term in the SVM equation, determining the offset of the hyperplane
C	Regularization parameter in SVM
F	Set of functions f
f	Function mapping inputs to outputs in the supervised learning problem
I	Set of indices for samples with positive labels
K	Gaussian Kernel function in SVM
L	Generalization error
l	Loss function
N	Total count of i samples in a dataset
S	System or rule in supervised learning
S	Binary sensitive attribute
$\mathbf{w}$	Weight vector in SVM
X	Input space
Y	Output space
Z	Cartesian product of X and Y

Greek Symbols

α	Lagrange multiplier in SVM
ε	Slack variable in SVM
γ	Parameter that defines the with of the Gaussian kernel in a SVM

Superscripts

a	Index for samples in a dataset corresponding to a sensitive attribute
b	Index for samples in a dataset corresponding to a non-sensitive attribute
i	Index for samples in a dataset
j	Index for samples in a dataset

Other Symbols

$\mathscr{A}$	An algorithm that designs the learning machine $\mathscr{R}$
$\mathscr{A}^*$	Best algorithm that designs the learning machine $\mathscr{R}$
D_n	Dataset of n examples
f^*	Best function mapping inputs to outputs in the supervised learning problem
$\mathscr{H}$	Set of hyperparameters that characterize an algorithm
$\mathscr{H}^*$	Best set of hyperparameters that characterize an algorithm
$\hat{L}$	Empirical error
$\mathbb{R}$	Set of real numbers
$\mathscr{R}$	Set of rules
$\mathfrak{R}$	Learning machine, specific rule of a set of rules $\mathscr{R}$
$\hat{y}$	Predicted output

Acronyms / Abbreviations

ACC	Accuracy
$ACCG$	Accuracy Gain due to an update of the model

AI	Artificial Intelligence
$B-ACC$	Balanced Accuracy
COMPAS	Correctional Offender Management Profiling for Alternative Sanctions
CV	Cross Validation
DDP	Difference on Demographic Parity Oportunity
DEO	Difference on Equal Oportunity
DL	Deep Learning
DP	Demographic Parity
EO	Equal Oportunity
FN	False negative
FP	False positive
$LGTBIQA+$	Collective of lesbians, gays, bisexuals, transgenders, queer, questioning, intersex, asexuals, and others.
ML	Machine Learning
NFD	Negative Flips Difference
NF	Negative Flips
NFR	Negative Flip Rate
NN	Neural Networks
PC	Positive Congruent
PCT	Positive Congruent Training
RBF	Radial Basis Function
SL	Supervised Learning
SVM	Support Vector Machine

TN	True negative
TP	True positive
TPR	True positive rate
URFSVM	Unfair-Regression-Free Support Vector Machine
UR	Unfair Regression

Part I

Section One

Chapter 1

Introduction

Over the past few years, Artificial Intelligence (AI) has experienced extraordinary advances, which have caused an increase in its worldwide use. Because AI-based algorithms enable the automation of decision-making processes using data, valuable resources such as time and money are substantially cut down compared to traditional methods that rely heavily on human intervention.

AI encompasses a broad range of technologies, with machine learning (ML) being a key subset that enables AI systems to learn from data and improve their performance over time without explicit programming. However, since ML algorithms strongly rely on human-generated data, biases present in these data are unintentionally transferred to the algorithm's decisions, leading to unfair outcomes. As a result, along with these advancements comes a growing concern regarding the fairness of machine learning algorithms. Several studies have revealed biases in machine learning applications, specifically in areas such as facial recognition [2], candidate ranking [3], and hiring decisions [4]. For example, if an AI-based algorithm is used for hiring in a men-dominant job, in most of the cases discrimination against women will be perpetuated despite having the exact same features as the other class.

Despite the efforts to mitigate these concerns, there remains an important need to address the issue of regression in machine learning algorithms. This occurs when a model is updated and begins to incorrectly predict previous accurate examples, leading to a drop in performance for a specific class. The problem we address in this research focuses on the intersection between unfairness and regression in performance. This particular case is given when regressiveness is unbalanced with respect to the sensitive attribute, which can be perceived as discriminatory behavior. Such phenomenon will be mentioned several times across this document as unfairness in regressiveness, unfair regression, unbalanced regression or UR.

This work addresses this critical issue by proposing novel methodologies to mitigate unbalanced regressiveness in both accuracy and/or fairness within machine learning systems. By doing so, we attempt to advance in the field and contribute to the development of more equitable and reliable machine learning models.

The study is structured as follows: first, the topic overview establishes the motivation, goals, and significance of the research. Following, the literature review is set, highlighting the existing research gaps. Our new metric is introduced by defining the unbalance between negative flips (NF) in model updates, which is referred to as unfair regression. After that, two mitigation strategies are defined to palliate this unbalance in predictive analysis. The first one affects only the model selection phase while the second one is built within the framework of Support Vector Machines (SVM). In both scenarios, we use our newly introduced metric (i.e. Unfair Regression) set as a constraint in the problem optimization that aims to be minimized. Lastly, the results showcase a comparative analysis between the standard and enhanced models, discussing how these new methods perform in terms of fairness. In addition, a summary of the most significant findings, their implications, and potential directions for further investigation completes the thesis.

1.1 Background

The rapid advancement of AI-based technologies has transformed numerous industries and processes. From Alan Turing's pioneering contributions to computer science through the development of the perceptron and the subsequent initial neural network architecture (the multilayer perceptron) in the mid-20th century, to the recent boom in deep learning (DL) algorithms that have driven remarkable advancements in fields like natural language processing (NLP) and computer vision (CV), machine learning has become an integral part of our daily lives.

The 21st century brought rapid advancements as AI has significantly evolved with improvements in deep learning, natural language processing, and AI's integration into various industries, demonstrating profound impacts on technology and society [5], such as the well-known large language model named GPT-3 [6]. According to a study conducted by Gartner, the adoption of AI and machine learning technologies has increased by 270 % in the past four years [7]. As a result, concerns about the awareness of the trustworthiness of these systems, particularly regarding bias and fairness, has grown in recent years. Researchers must ensure that these systems do not perpetuate negative social impacts on minority groups discriminated against based on gender, ethnicity, or disability, among others.

Furthermore, in machine learning there is also the unrelated problem of regressiveness in performance, which comes from the term *regression*, very well known in classical software development. This concept refers to a decline in software performance or functionality followed by an update. Similarly, in ML-based systems, updates are also required for various reasons, such as the availability of new data or models, and the need to optimize different technical or ethical metrics. The updated versions of ML models are designed to improve the average performance not taking into account the sample-wise performance (i.e. performance on specific predictions). Consequently, in classification tasks, an update may decrease the average number of misclassifications while introducing misclassification on samples that were correctly predicted in the old model. These newly introduced misclassifications are called negative flips (NF) [8] and the need for reducing them is a challenge in different applications [9], [10].

While there is significant literature on *algorithmic unfairness*, the concept of regression in ML is relatively new in the literature. Research has focused on developing methods to handle new data more effectively, making the regression in model updates a less prominent area of research focus. Accordingly, ongoing and future investigation needs to continue exploring and refining such strategies that consider the multifaceted nature of bias in data and that minimize the regression of accuracy and/or fairness in machine learning models aligned with the existing laws and international human rights standards according to the EU AI Act [11]. This study reviews existing research dedicated to fairness in machine learning, as well as techniques and metrics for identifying and reducing biases [12]. Specifically, new methodologies to deal with unbalanced regression within the SVM framework are proposed, allowing us to focus for the first time on the intersection between unfairness and regressiveness.

1.2 Significance

The awareness of bias and (un)fairness, and the consequent research emerged after the 1964 US Civil Rights Act [13], [14] marking a significant turning point in the fight against discrimination. Furthermore, discriminating based on certain criteria was illegal as it explicitly prohibited the unfair treatment of individuals based on their protected attributes (e.g. gender, race, disability [7], [15]). However, as machine learning relies on human-produced data, biased and unfair algorithms are still present nowadays. One clear example is seen in the COMPAS dataset, used for the prediction of recidivism in the criminal justice system, which has been found to disproportionately classify black defendants as high risk compared to white defendants with similar backgrounds [1]. Beyond this example, there is still a wide range of unfair examples, including hiring practices [4], credit scoring [16], and healthcare decision-making [17].

In addition to these concerns, the issue of regression in machine learning systems, while less studied than unfairness, also requires attention. This work addresses the intersection of technical and ethical debt, focusing specifically on the mitigation and interaction between regression (i.e., the tendency of updated models to fail predictions that were correctly performed by older versions of the model [8]) and unfairness (i.e., the tendency of algorithms to perpetuate or amplify historical biases against sensitive groups [18]).

The significance of this study lies in its potential to tackle the problem of unfair regression in machine learning algorithms, which can have serious implications across various fields since it perpetuates bias in the predicted outcomes. Therefore, by studying the intersection between regression and fairness we are able to ease this problematic behavior in the continuous updates in ML. On the whole, this study contributes to the development of more equitable and ethical machine learning systems, ensuring no human is discriminated against for its gender, ethnicity, (dis)abilities, or social status [19].

1.3 Objectives

The main study objective is to develop novel methodologies to mitigate unbalanced regressiveness in model updating, with resulting accuracy and/or fairness increase within machine learning systems. To achieve the general goal of enforcing regressiveness to be equally distributed with respect to the sensitive attribute, more specific objectives are outlined:

1. Conduct deep research on the existing state-of-the-art techniques for bias mitigation and identify the main gaps in the literature.

2. Define the phenomenon of Unfair Regression and show its existence by testing it on standard SVM models and relative updates.

3. Propose a first mitigation acting only on the model selection phase, i.e., taking into account the unfair regression in the hyperparameter tuning process.

4. Develop a novel in-processing technique introduced as the second mitigation strategy aimed at minimizing the occurrence of negative flips difference between discriminated and non-discriminated groups, i.e. unfair regression, within the SVM framework.

5. Evaluate the effectiveness of the new methodologies in mitigating unfair regression during model updates.

6. Determine how this introduced metric correlates with already existing fairness metrics in the present literature.

7. Assess how these proposed approaches can be integrated into existing machine learning frameworks without compromising performance.

8. Explore the vast implications of reducing unfair regression on the trustworthiness and ethical considerations of AI applications.

Chapter 2

Literature Review

This literature review highlights the importance of fairness in machine learning models by examining techniques for mitigating bias in such algorithms. More specifically, this section examines how fairness and regressiveness are currently approached in machine learning, discusses relevant studies, and points out areas that need further research.

2.1 Related Work

Algorithmic bias and its effects on social equality have become more of a concern as machine learning becomes increasingly common in major sectors like finance [16], healthcare [20], and criminal justice [1]. As a result, several fairness evaluations and comparative works have been introduced in the last decade. Methodologies can be categorized into *pre-processing* of data [21, 22, 23, 24], *in-processing* in the model training, [15, 25, 26, 27, 28, 29], and *post-processing* on the model outputs [30, 31, 32]. For an extensive review of the mentioned methodologies refer to Section 3.4).

In the present work we focus on the techniques that incorporate fairness constraints during the training phase, resulting in the already defined in-processing techniques. Complementary to the literature reviewed, some other works are presented as a constrained optimization [33, 34]. However, this methods present certain limitations, as highlighted by Kamishima et al. work [35], where the lack of a convex objective function yield solutions of logistic regression trapped in local minima. In Table 2.2, leading research of in-processing techniques for fair algorithm development is highlighted together with their main findings and limitations.

On the other hand, regressiveness in machine learning systems is still a less-explored topic. Some works are presented in the literature such as the research of Angioni et. al. [9] which deals with the fact that model updates may not only induce a perceived regression of classification accuracy via negative flips but also a regression of other trustworthiness-related metrics, such as adversarial robustness. Another example of efforts made on regressiveness mitigation is the work of Yan et. al. on *Positive-Congruent Training: Towards Regression-Free Model Updates* [8], in which a simple approach for positive-congruent (PC) training is proposed which enforces congruence with the reference model by giving more weights to samples that were correctly classified. A brief overview of the mentioned research can be found in Table 2.2.

Table 2.1 Highlighted works in the current literature on fairness in ML within in-processing techniques.

Author(s)	Research Title	Main Findings	Limitations
Agarwal et al., 2018	A Reductions Approach to Fair Classification	Introduced a reduction approach to fair classification by turning the problem into a sequence of cost-sensitive problems.	May not generalize well across all types of data distributions.
Celis et al., 2019	Classification with Fairness Constraints: A Meta-Algorithm with Provable Guarantees	Provided a meta-algorithm with provable guarantees, capable of handling multiple fairness constraints.	Complexity increases significantly with the addition of multiple constraints.
Goel et al., 2018	Non-Discriminatory Machine Learning through Convex Fairness Criteria	Proposed convex fairness criteria that facilitate non-discriminatory learning in a convex optimization framework.	Primarily theoretical; real-world application and effectiveness can vary.
Manisha and Gujar, 2018	A Neural Network Framework for Fair Classifier	Developed a neural network framework for fair classification, aiming at fairness in neural network training.	May require significant computational resources; effectiveness is dependent on network architecture and size.
Zhang et al., 2018	Mitigating Unwanted Biases with Adversarial Learning	Applied Adversarial Learning to mitigate unwanted biases by training a predictor and an adversary simultaneously.	Adversarial training can be unstable and may lead to reduced model accuracy.
Zafar et al., 2017	Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment	Introduced fairness definitions beyond disparate treatment and impact, focusing on avoiding disparate mistreatment.	Balancing fairness with accuracy can be challenging, especially with strict fairness constraints.
Zafar et al., 2017 (Fairness Constraints)	Fairness Constraints: A Flexible Approach for Fair Classification	Provided mechanisms to incorporate fairness constraints directly into classifier training processes.	The constraints can limit the classifier's predictive performance and scalability.

Table 2.2 Highlighted works on regression in model updates of ML systems.

Author(s)	**Research Title**	**Main Findings**	**Limitations**
Toneva et. al., 2018	An Empirical Study of Example Forgetting during Deep Neural Network Learning	Defines a "forgetting event" to have occurred when an individual training example transitions from being classified correctly to incorrectly over the course of learning.	Studies implicit negative flips during training of a single model.
Yan et al., 2019	Positive-Congruent Training: Towards Regression-Free Model Updates	Proposes a simple approach for positive-congruent training, Focal Distillation, which enforces congruence with the reference model by giving more weights to samples that were correctly classified.	It is not generalized for tracking both accuracy and fairness metrics in the regression of model updates.

2.2 Gap in Literature

State-of-the-art literature offers a wide variety of fairness metrics to be used for bias measurement. However, there is still no consensus on which is "the best" definition of fairness since it is typically impossible to achieve multiple definitions simultaneously [36, 37]. As a result, the choice of the metric remains a debated topic. Moreover, there is a significant research gap regarding a formal and comparative study of each metric’s strengths and limitations. The literature further outlines algorithms in the three categories aforementioned: pre-processing, in-processing, and post-processing. Each category has its advantages and disadvantages, and different fairness metrics are used to address bias in each. Concurrently, restrictions and recommendations exist on which types of algorithms to use. However, a systematic approach is still missing that allows the research community to choose the optimal technique for their specific application. More importantly, the intersection between fairness and regressiveness in machine learning is still an unexplored topic.

As a result, in this work, we propose a new methodology to mitigate unfair regressiveness in machine learning systems, by introducing a novel fairness metric based on the so-called *negative flips*, named *Unfair Regression*. Considering positive NF or total NF it is also possible, however, these two expressions of the same metrics are analogous to considering the difference in equal opportunity (DEO) or difference in demographic parity (DDP) in standard definitions of fairness.

The algorithm developed in this work is intended to minimize the difference between negative flips in sensitive and non-sensitive attributes (i.e. unfair regression) while preserving, as much as possible, the accuracy during the optimization of a classification problem using a support vector machine (SVM). Therefore, this method aims to present a cutting-edge approach that allows us to measure and minimize untrustworthiness and unfair regression in machine learning systems.

Chapter 3

Theoretical Background

3.1 Fundamentals of Machine Learning

Artificial Intelligence can be defined as the capability of a computer system and machine to perform tasks that traditionally require human intelligence. Furthermore, what makes a machine intelligent is the ability to memorize things. Nowadays, when we talk about artificial intelligence, in most cases we refer to machine learning systems. As aforementioned, ML is a branch of AI that allows computers to learn from data and improve their performance over time without explicit programming. These algorithms can identify patterns and relationships within data, extract meaningful insights, and make predictions or decisions based on those patterns. This ability to learn from data enables ML systems to automate tasks, recognize speech, classify images, and more.

Within ML systems, we can distinguish three fundamental approaches: supervised learning, unsupervised learning, and reinforcement learning. In supervised learning (SL) the algorithm learns from labeled examples to make predictions or classifications. On the other hand, unsupervised learning involves learning from unlabeled data, where the algorithm is not provided with any target label and it aims to extract meaningful hidden patterns. Reinforcement learning is a distinct branch of ML where the computer learns to interact with an environment and takes actions to maximize a reward signal. ML becomes Deep Learning (DL) when a neural network architecture is employed to perform tasks of any of the three aforementioned branches.
Along all these approaches, there exist multiple algorithms, and the choice depends on the task's main objective and the data's nature.

In the present dissertation, bias mitigation techniques are built and achieved within the SL framework for classification tasks, using the well-known learning algorithm named support vector machine.

3.1.1 The Supervised Learning Problem

Supervised learning aims to classify an unknown system, or rule, to a specific outcome. From a mathematical notation point of view, by observing the system or rule, $S : X \rightarrow Y$, which maps a point x from an input space X, into a point y of an output space Y, we can build a rule (i.e. a learning machine) $\mathfrak{R} : X \rightarrow Y$, that similarly maps a point $x \in X$, into a point $\hat{y} \in Y$, as displayed in Fig. 3.1. We can define the space Z as the cartesian product between the input and the output space, $Z = X \times Y$, being $z \in Z$ a point in this space. From the system S, a series of n examples can be obtained, which compound the dataset D_n. Therefore, the goal of supervised learning is to project $D_n \in Z^n$ into a rule $\mathfrak{R}$ selected from a predefined set of possible rules, $\mathscr{R}$, during the learning phase [38].

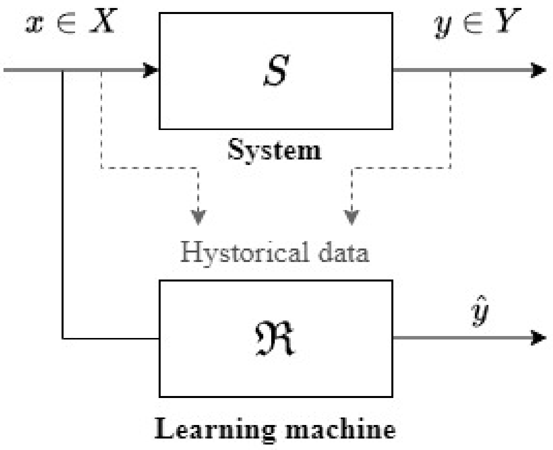


Figure 3.1 Graphical representation of the supervised learning problem.

Given a labeled dataset $D_n = \{(\mathbf{x}_i, y_i), \ldots, (\mathbf{x}_n, y_n)\}$ that consists of n examples, we can consider that there is some specific unknown function f within a set of functions F, $f \in F$, that best represents the mapping from an input space $x \in X$ to an output space $y \in Y$ of our machine $\mathfrak{R}$: $\hat{y} = f(\mathbf{x})$.

Then, the quality of the learning machine, $\mathfrak{R}$, that mimics the system's behavior, S, for a determined set of data D_n, can be measured through the loss function ℓ, mathematically described as $\ell(\hat{y}, y) : Z \rightarrow \mathbb{R}$, which quantifies the difference between the predicted output $\hat{y} = f(\mathbf{x})$ and the actual output y for a given input $\mathbf{x}$.

3.1.2 The Classification Problem: Support Vector Machines

SVMs as classifiers became well-known in the 1990s and early 2000s for their excellent performance in a variety of contexts [39]. SVMs are particularly well-suited for binary classification tasks, where the objective is to identify the optimal hyperplane that separates two classes in an n-dimensional feature space. This hyperplane is selected to maximize the margin, which is the distance between the hyperplane and the closest data points from each class, known as support vectors. The SVM aims to minimize classification errors by balancing the trade-off between a wider margin and the penalty for misclassification.

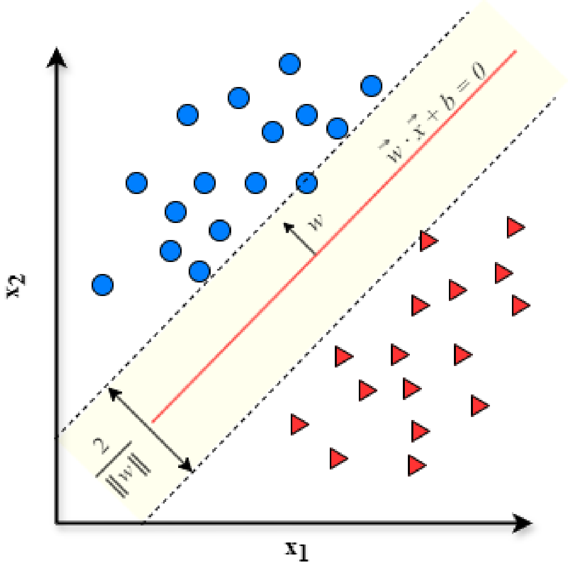


Figure 3.2 Linear Support Vector Machine classification based on two predictors (x_1, x_2). Classification hyperplane (-) and maximal margins (- - -) are shown.

The SVM approach is fundamentally grounded in the concept of finding this optimal separating hyperplane, mathematically defined in Eq. 3.1. Those points that lie immediately next to the separating hyperplane, namely the support vectors, are the most critical elements when defining the decision boundary. However, by focusing on these support vectors rather than on the entire dataset, SVMs can handle high dimensional spaces efficiently and are less vulnerable to overfitting.

$$\vec{w} \cdot \vec{x} + \mathrm{b} = 0 \tag{3.1}$$

The Kernel-Trick

Theoretically, when data is not linearly separable in the original future space, a transformation can be applied to project the data into a higher-dimensional space where a linear separation is possible [40]. However, applying such transformation directly may be too computationally expensive, especially with high-dimensional data. To address this issue, SVMs take advantage of the "kernel-trick", which enables them to efficiently operate in high-dimensional spaces without explicitly computing the transformation [41].

The kernel function computes the dot product of the data points in the transformed feature space directly, allowing the SVM to construct a non-linear decision boundary in the original space [42]. Different kernel functions can be employed depending on the problem, each specifying a different kind of decision boundary. Commonly used kernels include the linear kernel, which is effective when data is approximately linearly separable, and non-linear kernels such as polynomial and radial basis function (RBF) kernels. The choice of kernel significantly impacts the performance and complexity of the SVM model. Non-linear kernels, such as the Gaussian kernel, are more versatile and particularly powerful in capturing complex decision boundaries but require careful tuning of additional hyperparameters.

Hyperparameters

The primary hyperparameter in SVMs is the regularization parameter C, often referred to as the box constraint. The parameter C controls the trade-off between achieving a low training error and a large margin. A smaller value of C allows for a wider margin at the cost of some misclassifications, promoting better generalization to unseen data. Conversely, a larger C prioritizes reducing training errors but may lead to overfitting, as it results in a narrower margin.

When using non-linear kernels, an additional hyperparameter must be optimized, such as the γ (*gamma*)parameter in the Gaussian kernel. The γ parameter determines the influence of individual training samples on the decision boundary. A low γ value implies that the influence extends far, resulting in smoother decision boundaries, while a high γ value leads to a more complex model with decision boundaries that closely follow the training data [43]. Proper calibration of these hyperparameters is crucial to the performance of the SVM, as they directly influence the classifier's capacity to generalize to new data.

3.1.3 Model Selection and Error Estimation

Training an algorithm and evaluating its performance on the same data usually yields overfitting, the event when an algorithm becomes too specific for the trained data and fails in generalizing and predicting to new, unseen data. This phenomenon was first explored in the early 30s by Larson et. al., [44] and, consequently, many efforts were made to fix this issue. As a result, Cross Validation (CV) was raised in the 70s [45, 46, 47] proving that testing the output of an algorithm on a new set of data – the so-called test set– brought to more reliable performance estimates.

At this point, we can define model selection as the process of choosing the best-performing model among a set of candidate models and/or a set of hyperparameter ranges based on their relative performance. In classification tasks, CV is a popular strategy for model selection, based on splitting data into multiple subsets: training the model on the subset of data called the training set, and evaluating its performance on another subset called the validation set. Then, CV selects the configuration with the smallest estimated risk. The main foundation of CV relies on the assumption that data are identically distributed and the training and validation sets are independent, *i.i.d.* This makes this method well suited to *almost* any algorithm in *almost* any framework, for instance in classification as demonstrated in previous works [48, 49]. Furthermore, Arlot et al. in 2010 [50], proved through empirical experiments that in the framework of binary classifications CV yielded *almost* always to the best performance. In the present work, the CV strategy and its modification are used to determine the optimal hyperparameters to train the support vector machines.

Then, an algorithm $\mathscr{A}_{\mathscr{H}}$ characterized by a set of hyperparameters, $\mathscr{H}$, permits the design of a rule $\Re \in R$ whose performance can be measured through the loss function ℓ. The quantity we want to measure is the generalization error, which is the error that the model will perform on new unseen data.

$$L(\Re) = L(f) = \mathbb{E}_z \ell(f, z) \tag{3.2}$$

Since the probability distribution over the set of points in the space Z is unknown, $L(\Re)$ cannot be computed and therefore, must be estimated. This yields to the empirical error expression:

$$\hat{L}(\Re, D_n) = \frac{1}{n} \sum_{z \in D_n} \ell(f, z) \tag{3.3}$$

In addition, to select the best algorithm and hyperparameters configurations, $\mathscr{A}^*{}_{\mathscr{H}^*}$, in a set of possible ones, we will define f^* as the model built with the algorithm $\mathscr{A}^*$ and set of hyperparameters $\mathscr{H}^*$, which allows achieving a performance close to the optimal one.

Summarizing, the CV technique is intended to select the best algorithm with the optimal hyperparameters in a set of possible algorithms with a set of possible hyperparameters. Since we assume data is *i.i.d*, the optimal algorithm should be able to achieve a small error on a dataset that is independent of the training set.

Double-step Cross Validation

A novel validation procedure, first introduced in the research of Donini M., Oneto L., et al. [51], is used for the hyperparameter selection. This procedure, in two steps using 10 folds, is expected to improve not only the hyperparameter selection in terms of accuracy but also in fairness. The first step of the double-step CV focuses on maximizing the accuracy by evaluating different hyperparameter configurations and identifying the set of values that yield the best metric values, which correspond to a pre-set value of above 97% of the best accuracy. Then, in the second step, the CV procedure is repeated, taking into account only the earlier best-selected hyperparameters and choosing the optimal values that generate the lowest negative flip rate difference between the sensitive and non-sensitive groups. An exemplification of its foundation is shown in Figure 3.3.

The double-step CV algorithm is implemented in the present work to examine which is the effect of enhancing the model selection phase on the accuracy and fairness metrics we assess. This approach is critical in scenarios where high accuracy can sometimes lead to unintended biases against certain groups. By incorporating the double-step CV, we aim to create a more balanced model that not only performs well overall but also adheres to fairness principles. This methodology is tested on all the selected datasets, comparing traditional single-step CV results with those obtained through the double-step CV.

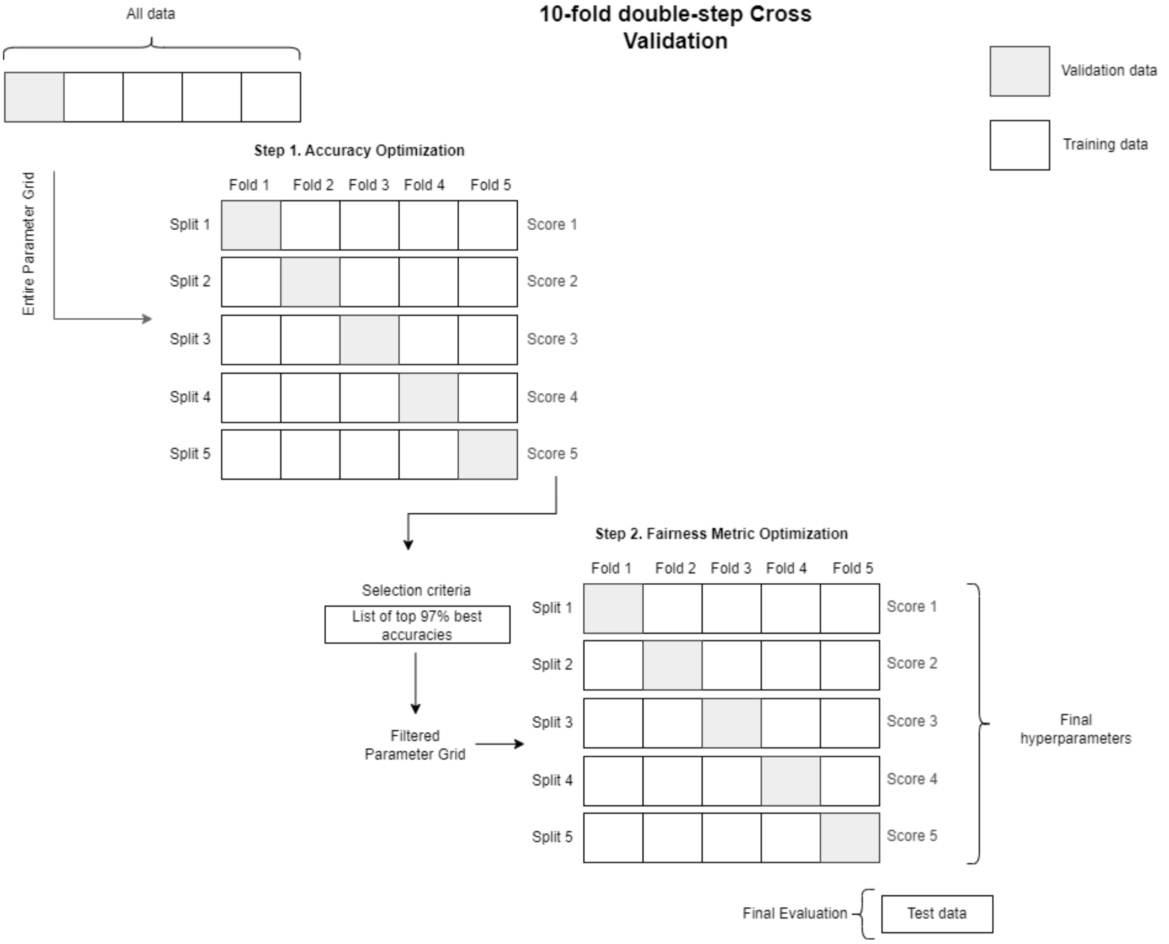


Figure 3.3 Double step cross validation exemplification. This method is employed to determine the hyperparameters by taking into account both accuracy and fairness metrics. A filtered list of the best accuracies is used for a second cross validation where the negative flip metric is considered.

3.2 Bias and Fairness in Machine Learning

As prediction-based decision algorithms are increasingly adopted by several organizations, concerns about the bias and (un)fairness in the models used arise. Systems that have an impact on people's lives create ethical responsibilities about making fair and unbiased judgments regardless of social aspects such as race, gender, class, and the like [52]. Recognizing and reducing bias and unfairness can be a tough undertaking task since different notions may be perceived between cultures [53]. Consequently, (un)fairness criteria are influenced by many factors such as user experience, as well as cultural, social, historical, political, legal, and ethical factors [54]. However, in this dissertation, state-of-the-art definitions and notions are used to address the main goal of building a fair algorithm under newly introduced fair metrics.

One of the most relevant and orotund cases of racial bias is found on the COMPAS dataset (Correctional Offender Management Profiling for Alternative Sanctions) from which judges and parole officers use a popular commercial algorithm for scoring criminal defendant's likelihood of reoffending (recidivism). It has been shown that the algorithm is biased in favor of white defendants, and against black inmates, based on a 2-year follow-up study (i.e who actually committed crimes or violent crimes after 2 years) [1]. In such analysis, Larson et. al. showed a notable pattern of mistakes measured by precision and sensitivity.

Figure 3.4 A famous example of two criminals that were scored using the COMPAS algorithm. Bernard Parker, left, was rated high risk; Dylan Fugett was rated low risk despite of his critical history. Source: Josh Ritchie for ProPublica [1].

To understand the basics and main causes for the present bias in machine learning and the consequent unfair algorithms, in the following subsections various sources of bias and their impact on fairness are explored.

3.2.1 Definitions and Forms of Bias

Bias can be defined as a systematic error that, under the context of fairness, places the privileged groups at the advantage of having a positive outcome. We understand a positive outcome as a favorable prediction to the recipient, such as receiving a loan, being hired for a job, or not being arrested.

The ways in which human bias can affect a dataset used to train a ML model are, unfortunately, still copious [7, 12, 36]. Between these several means in how human misperception can be transferred to the algorithms, we can distinguish between different ways through which bias in machine learning models is introduced, according to [7]:

- **Training data**: since machine learning models learn from training data, if this is corrupted, the resulting model will be so.

- **Label definitions**: when the target label contains unclear information about the outcome, incorrect predictions may result, leading to biased decisions.

- **Feature selection**: when using features that are not significant for the real-world model application, bias against protected groups can occur.

- **Proxies**: proxies are variables that are not explicitly sensitive variables such as gender or ethnicity, but that may represent them. For instance, using height and body weight as proxies for gender can introduce bias in certain applications.

- **Masking**: even when deleting any sensitive attributes or proxies, to achieve a new representation of the data, new features, known as masked features, might be created to take the place of the other attributes. As a consequence, bias will still be present.

Overall, bias is a significant problem to address and can manifest in various ways, as previously mentioned. This issue is particularly concerning because it can be subtle and not easily detected without a thorough analysis of the data and models. Additionally, individuals are not always aware of their own prejudices, which can inadvertently contribute to unfairness in machine learning.

3.2.2 Sensitive Variables: the Unprivileged Groups

The sensitive variable concept, also often referred to as protected attribute, was first introduced when researchers began to study the field of fairness in machine learning in the early 2010s [55]. The nature of the problem of proving models to be biased and unfair, developed the need to identify and categorize those variables that were found to be discriminated, so that received an unfavorable treatment. Not as a surprise, those attributes for which predictions were biased, corresponded to unprivileged classes such as colored-skin people, women, the LGTBIQA+ collective, foreigners, disabled people, and/or aged individuals.

One pioneer work discussing sensitive variables' role in machine learning was Moritz Hardt, Eric Price, and Nathan Srebro's paper, "Equality of Opportunity in Supervised Learning," presented at the 2016 Neural Information Processing Systems (NeurIPS) conference [30], where a criterion for fairness in machine learning models that consider sensitive attributes was introduced to ensure equal opportunity. From now on, several studies have been carried out that aim to formalize and establish a detailed methodology to address this issue, as well as to discuss what can be considered a protected attribute.

In the present work, we will focus on the groups based on ethnicity, gender, and origin, the sensitive variables being racial people, women, and foreigners, respectively. Our goal is to mitigate any kind of bias towards any group, removing unfair predictions. The methodology through which this is achieved is based on the implementation of a *fair* constraint in the Support Vector Machine formulation, making it an in-processing approach to bias mitigation.

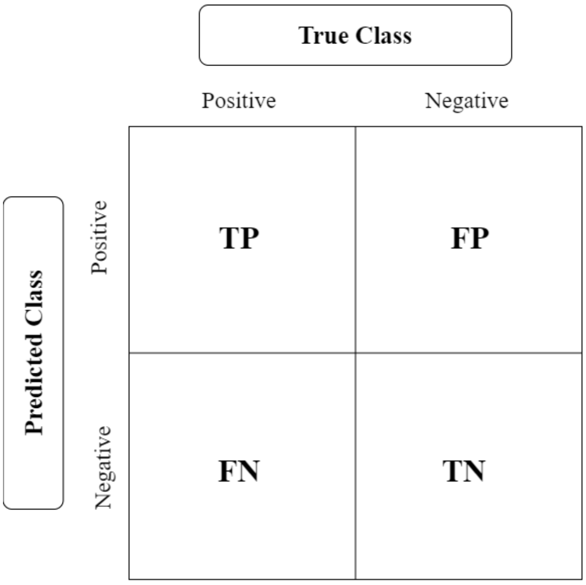
3.2.3 Metrics for Predictive Analysis and Fairness

Defining fairness has become a huge topic in machine learning since is a complex and multifaceted aspect and thus many notions and perspectives may be considered. Far from now, there is no unique, comprehensive definition of fairness but a set of proposed metrics that measure fairness instead. According to previous works [52, 56], more than twenty different notions of fairness have been proposed, and which to use in each circumstance is still up for debate [51, 57]. Nevertheless, a general notion of fairness can be defined as a quantification of how much undesired bias exists in training data or a model. As a result, researchers have been working on defining mathematical expressions that serve as metrics to assess (un)fairness. As mentioned, more than 20 expressions are found in the literature, however, all of them fall into two larger categories that allow us to represent different perspectives on what means for an algorithm to be fair. The first category is denoted as group fairness, which is also known as statistical fairness, and aims to achieve equality by focusing on group membership, such as gender or ethnicity. Meanwhile, the second group, individual fairness, aims to achieve equitably at the level of individuals, regardless of their group. Further remark, is that most of the fairness definitions rely on the well-known statistical metrics retrieved from the confusion matrix.

From this table, the different definitions can be taken out:

- **True positive (TP)**: when the predicted and true outcomes both correspond to the positive class.

- **False positive (FP)**: an outcome predicted to be in a positive class when the true outcome belongs to the negative class.
- **False negative (FN)**: an outcome predicted to be in the negative class when the true outcome belongs to the positive class.
- **True negative (TN)**: when the predicted and true outcomes both correspond to the negative class.



Predicted Class	True Class: Positive	True Class: Negative
Positive	TP	FP
Negative	FN	TN

Figure 3.5 Confusion matrix and classification metrics for a binary classification problem.

From the definitions retrieved from the confusion matrix, several statistical metrics can be defined, such as the ones summarized in Table 3.1. While these metrics are often considered to evaluate the statistical performance of the classifiers, they can also be used to study the performance in fairness as well as potential trade-offs in algorithmic fairness. In fact, since the increasing popularity of algorithmic fairness, most of the metrics fairness recently introduced rely on some of the definitions presented.

In this work, a couple of the fairness notions already introduced in the current literature are explained as well as used to assess the model performance. Furthermore, we will prove that it is not possible to satisfy different fairness metrics simultaneously. The metrics chosen include demographic parity [22, 35, 58] and equal opportunity [30, 32, 59], which belong to the larger category of group fairness.

Table 3.1 Classification metrics.

Metric	**Formula**	**Description**
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$	Overall percentage of correct classifications
Sensitivity, Recall, True Positive Rate	$\frac{TP}{TP+FN}$	Percentage of true label 1 observations that were classified as label 1
Specificity, True Negative Rate	$\frac{TN}{TN+FP}$	Percentage of true label 0 observations that were classified as label 0
False Positive Rate	$\frac{FP}{FP+TN}$	Percentage of true label 0 observations that were classified as label 1
False Negative Rate	$\frac{FN}{FN+TP}$	Percentage of true label 1 observations that were classified as label 0
Precision	$\frac{TP}{TP+FP}$	Percentage of predicted label 1 observations that were correctly classified
F1 score	$2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$	Measures accuracy via a combined recall and precision metric

For the following mathematical definitions, the problem of binary classification is considered and the following notation is used: $y \in \{0,1\}$: target variable (e.g. the applicant deserves or not to be hired, where 1 is the advantageous outcome), $\hat{y} = 1$ denotes a prediction with a positive outcome, $S \in \{a,b\}$: a binary sensitive attribute where a is the *unprivileged* class and b is the *privileged* class.

Demographic Parity

DP was one of the first fairness metrics suggested in the fairness literature [22, 35, 58] and one of the most well-known [36]. This metric can be defined as the probability of having a positive outcome regardless of the group membership. This can be also expressed as the probability of achieving a favorable label for the protected class equal to the probability of achieving a favorable label for the unprotected class:

$$P(\hat{y}_i = 1|S_i = a) = P(\hat{y}_j = 1|S_j = b), \ \forall a,b,i \neq j \tag{3.4}$$

Therefore, an algorithm is considered fair if it accomplishes this equality. In particular, the metric of difference in demographic parity is defined as the difference between the demographic parity computed on the two different groups, which is desired to be as close to zero as possible:

$$DDP = |DP_{S=a} = DP_{S=b}| = 0 \tag{3.5}$$

Equal Opportunity

While demographic parity only focuses on the percentage of observations with favorable predictions, the introduced metric of equal opportunity, suggested by several authors [30, 32, 59], considers two different classification metrics: False Negatives and True Positive. The first one, means predicting a positive label when the true label is negative (i.e. $P(\hat{y} = 0|y = 1)$), whereas the second one focuses on predicting the positive labels correctly (i.e. $P(\hat{y} = 1|y = 1)$. From these quantities, we can compute the True Positive Rate, defined as follows:

$$TPR = \frac{TP}{FN} \tag{3.6}$$

For a binary classification problem, where 1 is the positive outcome (i.e. getting a loan or not being arrested), Equal Opportunity can be defined as:

$$P(\hat{y}_i = 1|S_i = a, y_i = 1) = P(\hat{y}_j = 1|S_j = b, y_j = 1), \ \forall a,b,i \neq j \tag{3.7}$$

Specifically, a classifier is considered fair under equal opportunity if the true positive rate matches both sensitive and non-sensitive attributes, according to:

$$DEO = |TPR_{S=a} - TPR_{S=b}| \tag{3.8}$$

Unexplored Metrics

Most of the fairness metrics, such as those previously introduced, are predictive analysis-based. Instead, this research introduces an unexplored notion of fairness based on changes in a model's predictions from correct to incorrect outcomes, or rather, NF. Before we proceed further, it is important to understand the background and theoretical framework of the newly introduced metric we will be discussing. To simplify this understanding, the phenomenon of regression in model updates is first introduced, which yields to explore the concept of the negative flip.

3.3 Regression in Model Updates

Technical debt in software engineering refers to the long-term consequences of taking shortcuts or making sub-optimal decisions during development, which can lead to increased maintenance costs and difficulties in the future [60]. While the expectation of updating a model is that only improvements will occur, a decline in performance may happen on some occasions, leading to the well-known regression in machine learning. Regression and technical debt are closely related since the more technical debt is accumulated, the higher the likelihood of regression during model updates. This is because the system becomes more fragile and harder to manage, leading to unexpected issues when changes are made.

In particular, regression in model updates can be manifested through negative flips, a phenomenon exemplified by an increase in overall accuracy, but with the misclassification of some predictions that were correctly predicted in the old version of the model [8], which is naturally felt as a step backward. Mathematically, we can define a negative flip if the following event is observed:

$$\hat{y}_i^{old} = y_i \quad \text{and} \quad \hat{y}_i^{new} \neq y_i \tag{3.9}$$

More specifically, we can define the negative flip rate (NFR) which measures the fraction of samples that are affected by a negative flip, according to the following expression, where N is the total number of i samples.

$$\text{NFR} = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i^{\text{new}} \neq y_i, \hat{y}_i^{\text{old}} = y_i) \tag{3.10}$$

As aforementioned, when a machine learning model is updated, it’s expected to perform better or at least maintain its performance across all metrics and groups. However, if the new model shows a decline in accuracy specifically for the sensitive variables (like gender or race), it creates an imbalance, leading to unfair regression.

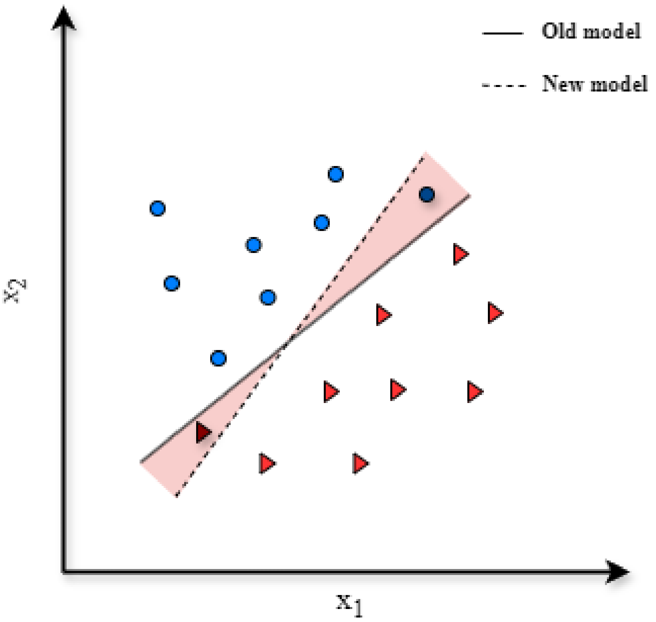


Figure 3.6 Regression in model update: when updating an old classifier (-) to a new one (- -), we can introduce errors that the old classifier did not make (negative flips, bottom-left, red).

3.3.1 Unfair Regression

Bias is introduced through regression when an updated model's accuracy decreases only for a specific class, maintaining or even increasing the other class's performance. This decrease in accuracy can be due to the negative flips presence, in the way that a new model incorrectly predicts the output for a test sample that the old model correctly classified. If these incorrect predictions disproportionately affect one group (e.g., female users), it results in unfair treatment, where the new model is systematically worse for a specific group. This imbalance in negative flips between classes of the same attribute (e.g., male vs. female) is a critical fairness issue, leading to unfair regression (UR).

Given the regression metric defined in Eq. 3.10, it is possible to assess how fair a model is, by quantifying the difference of negative flips that occur for a specific class through UR, which ideally should be as close to zero as possible.

$$\mathrm{UR} = |\mathrm{NFR}_{S=a} - \mathrm{NFR}_{S=b}| \tag{3.11}$$

In an optimization scenario, we can generalize this equation by setting a threshold, ε, of *unfairness* that one is willing to pay. For the particular case study of this research, ε is set to 0.

$$\text{UR} \leq \varepsilon, \quad \text{being } \varepsilon = 0 \tag{3.12}$$

Summarizing, our state-of-the-art methodology aims to mitigate unbalanced regressiveness in accuracy and fairness by means of an optimization problem that encompasses a trade-off between accuracy and negative flips during the model training.

3.4 Approaches to Mitigate Bias: Fair Algorithms

In the current state-of-the-art literature there exist many algorithms that assist in improving fairness [7, 22, 31, 51, 61, 62, 63, 64, 65, 66]. Almost all of them belong to three different approaches depending on the stage of the machine learning pipeline where fairness is imposed, see Fig. 4.1. In short, these different strategies involve pre-processing algorithms, if fairness is set before the training; in-processing algorithms if fairness is imposed during the training phase; and finally post-processing algorithms for those that modify the outcome to enforce fair predictions.

Pre-processing The first technique includes pre-processing algorithms, based on altering the original dataset before it is used for training. In this way, the algorithm applied makes it possible to readjust the features and labels in the original data to ensure that the training data is diverse and representative of the population. In general, this approach can only be used for optimizing demographic parity or individual fairness as it does not contain the information of the target [36]. Some of these algorithms comprise techniques like fair representation learning [24], resampling, and reweighing [23], among others.

In-processing The second technique aims to incorporate fairness constraints or regularization terms during the training process and is the one used in the present work. In this case, the model is optimized subject to such constraints so that the model output is not biased towards any particular group. Most of the works in the literature fall into this category [36], as such methods can be used to optimize any fairness definition [15, 59, 67]. Some approaches include fair optimization, explainable AI, and adversarial training. However, this approach has a drawback since it may not be applicable if the classifier is not accessible.

Post-processing Lastly, post-processing techniques are those that allow fairness to be met without modifying the classifier [30], as the mitigation attack is done after the model is trained, i.e. by modifying the outcome decision. Such methods can be used to optimize most of the fairness constraints except for counterfactual fairness. Some techniques are thresholding and calibration.

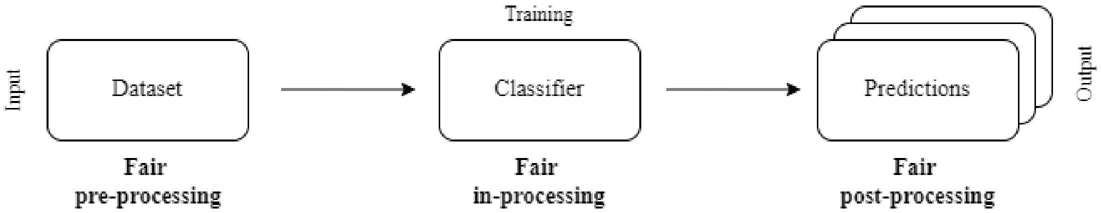


Figure 3.7 Overview of the fair approaches in the machine learning pipeline.

The new methodologies presented in this work are part of in-processing techniques, which modify directly the learning phase of the machine learning system.

3.5 SVM Under Fairness Constraints: Unfair-Regression-Free SVM

With the novel metric introduced to address fairness, unfair regression, a new algorithm emerges for the support vector machine under fairness constraints.
Subject to the UR constraint:

$$UR = |\mathrm{NFR}_{S=a} - \mathrm{NFR}_{S=b}| = 0 \tag{3.13}$$

Let I be the cardinality of samples of a dataset with a correct predicted outcome in the old model:

$$I = \{i : f^{\text{old}}(\mathbf{x}_i) = y_i\} \tag{3.14}$$

Where $I_{S_i=a}$ is the number of samples that belong to the sensitive attribute:

$$I_{S_i=a} = I_a = \{i : f^{\text{old}}(\mathbf{x}_i) = y_i \; \& \; S_i = a\} \tag{3.15}$$

And similarly for the non-sensitive attributes $S_i = b$:

$$I_{S_i=b} = I_b = \{i : f^{\mathrm{old}}(\mathbf{x}_i) = y_i \;\&\; S_i = b\} \tag{3.16}$$

Then, our linear support vector machine problem becomes:

$$\frac{1}{n_{I_a}} \sum_{i \in I_a} f(\mathbf{x}_i) - \frac{1}{n_{I_b}} \sum_{i \in I_b} f(\mathbf{x}_i) = 0 \tag{3.17}$$

Since $f(\mathbf{x}_i)$ is defined by

$$f(\mathbf{x}_i) = \sum_i \mathbf{w}\mathbf{x}_i + \mathrm{b} \tag{3.18}$$

Combining 3.17 and 3.17 we obtain:

$$(\frac{1}{n_{I_a}} \sum_{i \in I_a} \mathbf{w}\mathbf{x}_i - \frac{1}{n_{I_b}} \sum_{i \in I=b} \mathbf{w}\mathbf{x}_i) + \mathrm{b} - \mathrm{b} = 0 \tag{3.19}$$

Finally, reorganizing terms,

$$\mathbf{w}(\frac{1}{n_{I_a}} \sum_{i \in I_a} \mathbf{x}_i - \frac{1}{n_{I_b}} \sum_{i \in I_b} \mathbf{x}_i)) = 0, \tag{3.20}$$

where $\mathrm{a} = \frac{1}{n_{I_a}} \sum_{i \in I_a} \mathbf{x}_i - \frac{1}{n_{I_b}} \sum_{i \in I_b} \mathbf{x}_i$

Similarly, the SVM problem using the Gaussian kernel, can be defined as:

$$\frac{1}{n_{I_a}} \sum_{i \in I_a} (\sum_{j=1}^{n} \alpha_j y_j K(\mathbf{x}_i, \mathbf{x}_j) + \mathrm{b}) - \frac{1}{n_{I_b}} \sum_{i \in I_b} (\sum_{j=1}^{n} \alpha_j y_j K(\mathbf{x}_i, \mathbf{x}_j) + \mathrm{b}) = 0 \tag{3.21}$$

Again, reorganizing and simplifying terms:

$$\sum_{j=1}^{n} \alpha_j y_j (\frac{1}{n_{I_a}} \sum_{i \in I_a} K(\mathbf{x}_i, \mathbf{x}_j) - \frac{1}{n_{I_b}} \sum_{i \in I_b} K(\mathbf{x}_i, \mathbf{x}_j)) = 0 \tag{3.22}$$

To conclude, the adaptation of the Support Vector Machine framework to incorporate fairness constraints (i.e. UR) is a determinant step in developing more equitable machine learning models. We can consistently reduce prediction bias by explicitly including measures that consider deviations among sensitive variables. This adaptation not only maintains the algorithm's predictive performance while improving its fairness but also aligns with legal and ethical requirements. The mathematical formulations presented here illustrate that it is feasible to address fairness by changing classic SVM methodologies, and thereby provide the foundation for future study and development in this field. In the following subsection, the optimization problem under fairness constraints set up in *Python* is presented.

3.5.1 Optimization Problem

To reach our goal, the support vector machine method must be implemented with fairness constraints under an optimization problem. This optimization problem is solved with the Gurobi Optimizer, and in a nutshell, it is implemented by the functions below.

The objective function to be optimized is the regularized loss function, commonly used in Support Vector Machines (SVMs). For the linear case, we use the following expressions.

$$\min_{w,\mathrm{b},\xi} \frac{||w||^2}{2} + C\sum_{i=1}^{N} \xi_i \tag{3.23}$$

Within the optimization problem, the constraints include the SVM margin constraints and the non-negativity of the slack variables ξ, as well as a fairness constraint that adjusts the decision boundary, as described below.

The margin constraints:

$$y_i(\mathbf{w}^T\mathbf{x}_i + \mathrm{b}) \geq 1 - \xi_i, \quad \forall i \tag{3.24}$$

Non-negativity constraints for the slack variables:

$$\xi_i \geq 0, \quad \forall i \tag{3.25}$$

And the fairness constraints:

$$\mathbf{w}^T \left(\frac{1}{|S_{i=a}|} \sum_{i \in S_{i=a}} \mathbf{x}_i - \frac{1}{|S_{i=b}|} \sum_{i \in S_{i=b}} \mathbf{x}_i \right) = 0 \tag{3.26}$$

For the non-linear scenario, we have to minimize the following expression:

$$\min \quad \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i\alpha_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j) - \sum_{i=1}^{N} \alpha_i \tag{3.27}$$

Subject to the following constraints:

$$\sum_{i=1}^{N} \alpha_i y_i = 0 \tag{3.28}$$

$$\sum_{i=1}^{N} \alpha_i y_i f_i = 0 \tag{3.29}$$

$$0 \leq \alpha_i \leq C, \quad \text{for } i = 1, \ldots, N \tag{3.30}$$

Where f_i is defined as:

$$f_i = \left(\frac{1}{|S_{i=a}|} \sum_{i \in S_{i=a}} K(\mathbf{x}_i, \mathbf{x}_j) - \frac{1}{|S_{i=b}|} \sum_{i \in S_{i=b}} K(\mathbf{x}_i, \mathbf{x}_j) \right) \tag{3.31}$$

By creating the algorithm that solves this optimization problem, we can conduct a support vector machine under fairness constraints that maximize the space between the two classes while minimizing the errors we allow to occur in our problem subject to the defined constraints. As a result, we hypothesize we will obtain an unbiased and trustworthy classifier.

3.6 Challenges in Bias Mitigation: the Trade-off Between Accuracy and Fairness

The impact of using fairness metrics on the accuracy performance strongly relies on the fairness definition used, on the nature of the dataset, and on the algorithms used. Either way, previous studies on algorithmic fairness showed that in most cases, imposing fairness can conflict with accuracy: increasing the fairness of solutions concerning subgroups often leads to a decrease in overall accuracy [21, 59, 64]. This is not a surprise since the objective is redirected from accuracy to both accuracy and fairness. Therefore, determining potential trade-offs between both goals is needed. Nowadays, it is still not well understood which is the actual trade-off to achieve fairness while maintaining the good performance of the predictor, which is closely related to the regression in both metrics when a model is updated. Nevertheless, some present works in the current literature allow a systematic evaluation of potential trade-offs between accuracy and fairness metrics, such as the research done by Haas et al., 2019, named *The Price of Fairness*, see [57]. In the cited study, Haas et al. propose a framework that in a nutshell, calculates the Pareto fronts to optimize hyperparameters and uses these trade-offs to establish the most appropriate level of fairness for each algorithm.

Despite that, finding the optimal trade-off between accuracy and fairness metrics is out of the scope of this research, instead, we focus on evaluating the performance by setting our fairness constraint to a fixed value of zero. This way, we will not only be able to reduce unfairness, but also assess how the accuracy is affected by the new SVM configuration.

Part II

Section Two

Chapter 4

Methodology

4.1 Research Design Study Overview

When new data becomes available or new algorithms are proposed in the literature, it is mandatory to ensure the optimal quality of the deployed model by updating (or upgrading) it, [9], [8]. For this reason, in our methodology, we define f^*_{old} the "old" model (i.e. the model trained with the original data and algorithms), and f^*_{new} the "new" model (i.e. the model trained with the updated data and upgraded algorithms). Then, two different ways to make updates [1] are studied, going from f^*_{old} to f^*_{new}, in such way to continue to improve the accuracy of f^*_{new} with respect to f^*_{old}, but also minimizing the unfair regression. To do so, a couple of strategies have been defined: (i) adjusting only the tuning phase, i.e. model selection phase, by using a refined cross validation method (double-step cross validation); or (ii) modifying the learning algorithm to account for the unfair regression phenomenon as well as adjusting the tuning phase.

In both cases, performance and fairness metrics are assessed, and the same set of hyperparameters for the cross validation are selected. Additionally, to ensure reliable and robust outcomes, we trained the classifiers using 30 different random states for splitting the data. The final metric results are computed as the average across these multiple runs together with the standard deviations, allowing us to tackle the variability of the results.

In the following subsections, the methodology is carefully set out. Data sources and preparation are exposed in the first place, followed by the model configurations chosen to

[1] From now on, we will use the word *update* for both update and upgrade scenarios for language simplicity

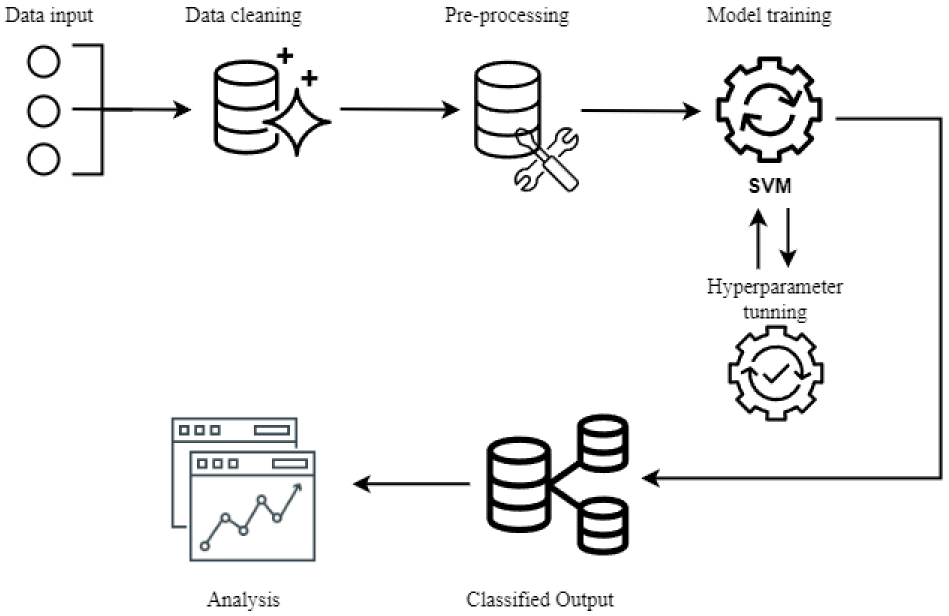


Figure 4.1 Machine learning sysems-based pipeline.

conduct the analysis. Next, the algorithm design is described and finally, the evaluation metrics used are defined.

4.1.1 Study Questions and Hypothesis

In this section we list the questions and hypotheses made as a starting point.

Questions

1. Does the type of model update, (i.e. the subset size and kernel function) affect the model performance when an update is done?
2. What is the mitigation strategy's impact on the accuracy and fairness metrics?
3. Can we obtain *fair* classifiers without sacrificing too much accuracy?

Hypothesis

1. Larger subset sizes will generally lead to better model performance due to the increased amount of training data available, and the non-linear kernels (like Gaussian kernel) are expected to perform better on datasets with complex, non-linear relationships. However, within the same dataset, resulting metrics are expected to be more or less

equal within the different types of updates, indicating the robustness of the mitigation strategy applied.

2. An enhancement of the fairness metrics (UR) when applying a mitigation strategy is expected. However, a trade-off between both measures must be studied as we add a second constraint to the problem.

3. The overall performance can be compromised when the optimization problem is double-constraint set. However, since we are assessing metrics in model updates which are hypothesized to perform better, a reasonable increase in accuracy while imposing UR is expected within the proposed strategies.

4.2 Data Collection and Preparation

This section outlines the sources of data used, providing a detailed overview of the most important characteristics of each dataset. Note that, the data used to train our models was not collected but retrieved from pre-existing datasets. Furthermore, data cleaning, preprocessing, and normalization steps are described.

4.2.1 Data Sources

To conduct the research, four well-known datasets used in fair machine learning (Adult, Arrhythmia, COMPAS, and German Credit) were employed to evaluate our algorithms. Further description of each set of data is done next and highlighted specifications are summarized in Table 4.1.

Adult Dataset

Adult dataset [68] is a database from the UCI repository that contains 12 features of demographic characteristics for 32,560 examples, where 2,399 are missing values. The prediction task is to determine whether a person earns over 50K $ a year. Due to time and computational resources, the dataset has been randomly downsized 30 times in order to make it possible to achieve the task.

Arrhythmia Dataset

The Arrhythmia dataset [17] from the UCI repository, contains 279 attributes for 452 instances gathered from the study of H. Altay Guvenir, intending to distinguish between the presence and absence of cardiac arrhythmia and to classify them in one of the 16 groups. In the our work, the classification task is reduced to a binary classification problem: to determine whether there is or not an arrhythmia disease. Therefore the target variable is binarized in our algorithm to "no illness" against "arrhythmia illness" outcome for any of the 15 different arrhythmia categories.

COMPAS Datset

The COMPAS dataset, an acronym for Correctional Offender Management Profiling for Alternative Sanctions, is a popular algorithm used in the US criminal justice system and has recently faced critical examination for being proven biased towards ethnicity. In particular, the algorithm is used for scoring criminal defendant's probability of recidivism. The dataset used for this algorithm contains 10 binary variables and over 6,000 observations. In this dataset as well, we randomly downsized by six times the number of samples to 1,028 examples due to limitations on time and computational resources.

German Credit Dataset

The German Credit Dataset [69] is a commonly used dataset for evaluating credit risk and is also employed to assess the presence of bias in machine learning models. It contains information on about 1,700 loan applicants and includes 20 attributes (7 numerical and 13 categorical) that describe each applicant, such as the purpose of the loan, amount requested, marital status, age, gender, job, and housing status. Moreover, the target attribute is included to describe the classification prediction: whether the applicant should be granted with a credit or not.

Summary

In table 4.1 an overview of the data's most relevant specifications is done. Note that for Adult and COMPAS datasets a significant size drop is made, and the final number of features is 1,025 and 1,028 respectively. An important aspect to be highlighted is the sensitive data representation. The most balanced dataset regarding the protected attribute, is the Arrythmia dataset, being equally represented. Adult and COMPAS are shown to be quite unbalanced: Adult has a higher representation of the non-protected group (66%), while COMPAS has more data for the sensitive attribute (66%). While this unbalancing is not significant in these last two datasets, for the German dataset we can report several proportion issues, as up to 96 % of the observations correspond to the protected group of foreigners.

Table 4.1 Dataset characteristics overview with the related statistics, and the sensitive features involved. Gender considers the two groups as male and female; ethnicity considers the ethnic group as white and other ethnic groups; foreign considers being or not being a foreign person.

Dataset	Ref.	# samples	# features	Sensitive Attribute
Adult	UCI	1,025	12	Gender
Arrhythmia	UCI	452	279	Gender
COMPAS	ProPublica	1,028	10	Ethnicity
German	UCI	1000	20	Foreign

4.2.2 Data Preparation

Raw data was cleaned, preprocessed and normalized so that it was suitable for further training and analysis. First, we addressed the problem of missing values consistently according to the dataset characteristics. COMPAS and German datasets were complete while Adult and Arrhythmia had some gaps. For these two last, different approaches were adopted: in the Adult dataset, we dropped the samples containing any missing values in any feature, as data was both numerical and categorical, while in the Arrhythmia dataset, which is based on numerical features, missing values were replaced by the mean of the corresponding column.

To deal with categorical features, we used the so-called *get_dummies* function build-in *Pandas* library to convert those attributes to numerical data suitable for the classification algorithm, except for the COMPAS dataset, where data is binarized.

Finally, the *Scikit Learn* preprocessing normalize function was used to normalize data so that the feature vectors had a unitary norm. We used the default 'L2' norm, known as the *Euclidean* norm, which scaled the input vector so that the sum of the squares of each element is equal to a unit. The normalization process ensures the features are on a similar scale, improving the performance and training stability of the model.

4.3 Mitigation Strategies

In this section, we describe how our proposed mitigation strategies are implemented. In the first place, to simulate model updates, different subset sizes and kernel functions are considered going from f^*_{old} to f^*_{new}, where this last model is supposed to be a more capable one. The linear kernel is typically well-suited for linearly separable data, while the Gaussian kernel is chosen for its ability to handle non-linear data separations, potentially improving model flexibility and fairness in more complex datasets. Moreover, in prior experiments, we observed that models trained on the entire dataset and employing a non-linear kernel achieved superior accuracy. Consequently, configurations using 100% of the data are expected to have higher overall accuracy due to the larger training set. Similarly, Gaussian kernels are anticipated to enhance performance due to their capability to manage complex data. Based on these observations, we define f^*_{new} as a model that uses more data for the training and/or the Gaussian Kernel.

Table 4.2 Model configurations varying subset sizes and types of kernel functions.

Model ID	Subset Size	Kernel Function
1	20 %	Linear
2	100 %	Linear
3	20 %	Gaussian
4	100 %	Gaussian

We considered two updating scenarios: (i) Data extension, and (ii) Data extension plus change on the kernel function. In the first one, we consider the case when the initial set of data is trained with f^*_{old}, and then more data become available and we retrain the same model f^*_{new} holding the additional data, which means that data of the old model is present on the new one. On the other side, the second updating scenario does the same as the first type of model update plus that the kernel function used for generating f^*_{new} has been changed with respect to the one used for f^*_{old} for a more capable one. In our experiments, we used Linear SVM for f^*_{old} and Gaussian kernel SVM for f^*_{new}.

Table 4.3 Considered scenarios for model updating. The new model is the enhanced classifier, while the old model performs as a reference.

ID	f^*_{old}		f^*_{new}		Improvement
	Subset (%)	Kernel	Subset (%)	Kernel	
Update Scenario 1	20	Linear	100	Linear	Data increase
Update Scenario 2	20	Gaussian	100	Gaussian	Data increase
Update Scenario 3	20	Linear	100	Gaussian	Change of kernel and data increase

Two strategies are explored to estimate and mitigate unfair regression in the selected model updates. The first mitigation strategy focuses on the adjustment of the model selection through the implementation of double-step cross validation, carefully described in Subsection 3.1.3. Instead, the second strategy adds a modification in the learning phase to account for unfair regression in the SVM algorithm.

4.3.1 Mitigation (i): Double-Step Cross Validation

Mitigating unfair regression requires careful tuning of hyperparameters, and within our objectives, we cannot rely only on one metric, such as accuracy, since it would neglect the unfair regression aspect. Having this aspect in mind, this strategy incorporates in the cross validation algorithm an added criteria for hyperparameter selection based on unfair regression, as already explained in Section 3.1.3. This way, using the SVM algorithm used from the scikit-learn library, we hypothesize to obtain a fair classifier.

The hyperparameter range for the tuning phase was chosen based on preliminary cross validation results [51]: for the C parameter, a range of 8 logarithmically spaced values between 10^{-4} and 10^{3}, and the additional parameter for the non-linear cases, γ, a four-value range from 0.001 to 1 was set. The hyperparameter optimization was performed through the *GridSearchCV* function from scikit-learn for the standard SVM method, while the methods accounting with a mitigation strategy used our two-step cross validation, where both accuracy and UR were considered. In all scenarios, the models were subjected to a 10-fold cross validation to ensure the robustness and reliability of the performance metrics.

Algorithm 1 Simplified Algorithm Snippet for the 2-Step CV: Mitigation Strategy (i)

```
Require: $X\_train, y\_train, C\_values, gamma\_values, kfold, CVmetric, f^{*}_{old}, p = 0.03$
Step 1: CV to maximize accuracy
Perform Grid Search on SVM with the specified kernel over $C\_values$
Record accuracies for each $C$ and select values with accuracy above threshold, according
    to $(1 - p)$
Step 2: CV to minimize UR
For each $C$ selected in Step 1, perform CV to minimize UR
Return best C and best $\gamma$ (in case of Gaussian Kernel) based on UR minimization
```

4.3.2 Mitigation (ii): Unfair-Regression-Free Support Vector Machine

In our second mitigation strategy, we add a constraint on the SVM algorithm, modifying the learning algorithm, according to Eq. 3.20 and Eq. 3.22. The advantage of this action is that the learning phase stays aware of our desire to mitigate the unfair regression while relying on both metrics in the hyperparameter tuning phase. Therefore, in this new mitigation, we combine the first strategy with the modification of the learning algorithm. This new model was implemented using the Gurobi optimizer.

Algorithm 2 Simplified Algorithm Snippet for URFSVM: Mitigation Strategy (ii)

```
Require: $X, y, C, boolS0, boolfold$
    Step 1: Initialize Model and Variables
    Create optimization model $m$ and initialize weights $w$ and slack variables $\varepsilon$
    Add bias term $b$ to the model
    Step 2: Define Objective Function
    Set the objective function to minimize $0.5 \cdot w^T w + C \cdot \sum \varepsilon$
    Step 3: Add Constraints
    Add SVM margin constraints and UR constraint based on $boolS0$ and $boolfold$
    Step 4: Optimize
    Optimize the model and extract the solution for $w$ and $b$
    Return optimized weights $w$ and bias $b$
```

4.4 Evaluation Metrics

The performance metrics used to evaluate the models are defined in this section. Our criteria to ensure success, focus on the balance between maintaining accuracy while improving fairness, and ensuring that the updated model does not disproportionately impact negatively against any group.

Different types of metrics for evaluating both accuracy and fairness are used. To evaluate the performance of the classifiers, we use the overall accuracy disaggregated by sensitive attributes as well as the balanced accuracy. Note that, the accuracy disaggregated by sensitive attributes also gives us a rough idea of the fairness of the model: if the accuracy is significantly unbalanced towards any group, it means the classifier may be unfair.

Performance Metrics

- Overall accuracy: gives us a rough idea of how good the classifier is, giving us the percentage of correct predictions made by the model.

- Disaggregated accuracy: gives us the accuracy separated by attribute.

- Balanced accuracy: very useful in imbalanced data, gives us the arithmetic mean of the sensitivity and specificity.

On the other hand, in the standard SVM we use the difference in demographic parity and the difference in equal opportunity to have a rough overview of the behavior of the classifiers in terms of fairness. Then, our newly introduced metric, UR, is computed in all models to examine the effectiveness of our strategies in the different methods.

Fairness metrics

- Difference in Demographic Parity (DDP): computes as the absolute difference of positive outcomes between groups. This metric is crucial for models that require equal treatment across demographics. Mathematically defined in Eq. 3.5.

- Difference in Equal Opportunity (DEO): mainly focuses on the true positive rate and it is calculated as the ratio of correctly predicted positives to actual positives. A notion of fairness can be yielded by comparing the ratio across the groups. Mathematically defined in Eq. 3.7.

- Unfair Regression: based on the negative flips difference that happens when a model update is done in the way that some examples are misclassified in the new model and were correctly predicted in the old one. Unfair regression stands for the scenario when regressiveness occurs in an unbalanced way. Mathematically described in 3.11

These metrics are chosen based on their relevance to the regressiveness problem in fairness and accuracy, providing a balanced view of model performance across different perspectives. For further description and detailed mathematical notation of the listed metrics, refer to section 3.2.3.

Chapter 5

Results

In this chapter we test how the methodology presented in the previous section performs in real-world datasets. It is important to recall our main objective: mitigate unbalanced regressiveness in accuracy and unfairness within model updates. The different scenarios to simulate updates are: (i) enlarging the dataset size using Linear SVM, (ii) enlarging the dataset size using Gaussian kernel SVM, and (iii) enlarging the dataset size plus upgrading the model modifying the Kernel function of the SVM. Such model updates have been chosen according to previous results: enhancements were shown in terms of performance when using rather more data or the Gaussian kernel.

Results are presented in the following order: first, the method using the standard SVM is presented, proving that unbalanced regression occurs. After that, the results of the models trained with the first mitigation strategy, namely the 2-step CV, are gathered. Finally, the results of our second method are displayed, using the URFSVM algorithm as a mitigation strategy. To ensure statistical significance, we repeated the experiments 30 times.

5.1 Reference Models: Standard SVM

The results of training our data with the *SVC* algorithm from the scikit-learn library as well as with the standard cross validation are presented in this section. Accuracy and several fairness metrics are displayed. In Figure 5.1 the misclassification error is plotted against both DDP and DEO, showing that the classifier is biased if we are limited to the use of standard methods.

Table 5.1 Performance and fairness metrics for the datasets with different kernel methods and dataset sizes.

Dataset	ACC	B-ACC	DEO	DDP
f^*_{old}: 20% of dataset size with linear kernel				
Adult	0.80 ± 0.03	0.67 ± 0.05	0.14 ± 0.10	0.10 ± 0.05
Arrhythmia	0.67 ± 0.04	0.65 ± 0.04	0.46 ± 0.11	0.17 ± 0.03
COMPAS	0.72 ± 0.03	0.69 ± 0.03	0.23 ± 0.15	0.22 ± 0.10
German	0.69 ± 0.03	0.63 ± 0.04	0.34 ± 0.20	0.20 ± 0.13
f^*_{old}: 20% of dataset size with Gaussian kernel				
Adult	0.79 ± 0.03	0.67 ± 0.06	0.17 ± 0.10	0.11 ± 0.06
Arrhythmia	0.68 ± 0.05	0.65 ± 0.05	0.40 ± 0.13	0.25 ± 0.04
COMPAS	0.72 ± 0.03	0.70 ± 0.04	0.23 ± 0.16	0.24 ± 0.10
German	0.69 ± 0.03	0.64 ± 0.04	0.42 ± 0.20	0.22 ± 0.14
f^*_{old}: 100% of dataset size with linear kernel				
Adult	0.71 ± 0.07	0.78 ± 0.04	0.10 ± 0.08	0.31 ± 0.07
Arrhythmia	0.74 ± 0.04	0.74 ± 0.03	0.72 ± 0.06	0.15 ± 0.04
COMPAS	0.72 ± 0.03	0.73 ± 0.02	0.12 ± 0.08	0.22 ± 0.05
German	0.74 ± 0.02	0.64 ± 0.04	0.35 ± 0.13	0.17 ± 0.07
f^*_{old}: 100% of dataset size with Gaussian kernel				
Adult	0.74 ± 0.06	0.79 ± 0.03	0.12 ± 0.09	0.29 ± 0.05
Arrhythmia	0.69 ± 0.04	0.71 ± 0.04	0.84 ± 0.08	0.24 ± 0.05
COMPAS	0.72 ± 0.02	0.73 ± 0.02	0.15 ± 0.09	0.25 ± 0.06
German	0.74 ± 0.02	0.62 ± 0.03	0.37 ± 0.15	0.11 ± 0.07

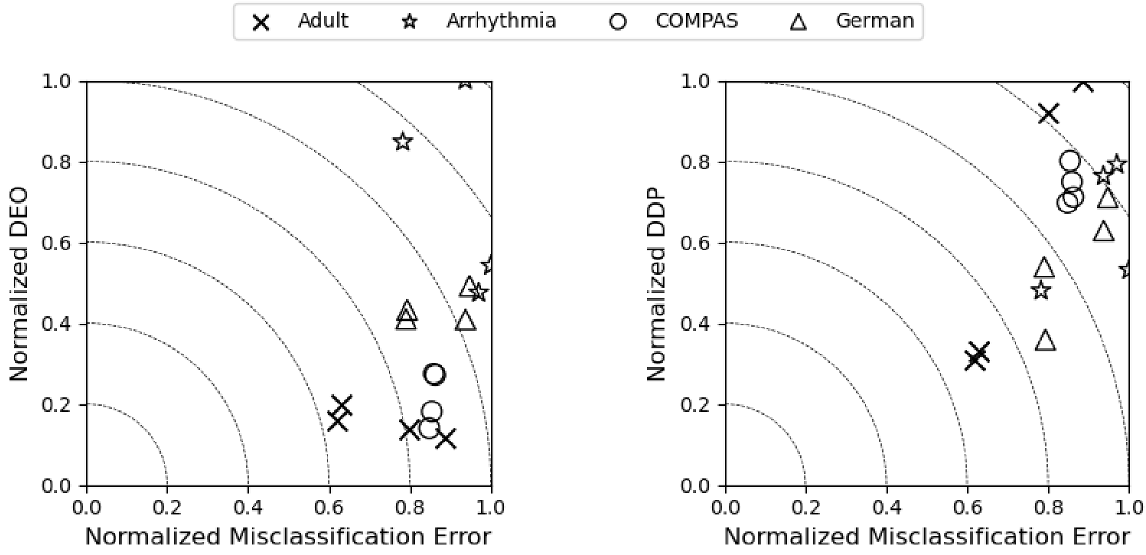


Figure 5.1 Fairness metrics for the standard SVM. On the left, the normalized difference in equal opportunity is plotted against the normalized misclassification error. On the right, the same is done for the difference in demographic parity. Note that, the closer a point to the origin is, the better the results are in both accuracy and fairness.

Next, we simulate model updates according to Table 4.3 with no mitigation strategy applied. The key aspect of this analysis is to observe whether such negative flips occur in an unbalanced way toward a specific class, i.e. if regression is disproportionately unbalanced. In such a case, we could say that we are in front of a model update that is promoting bias and providing unfair predictions.

Table 5.2 Accuracy gain (ACCG) and unfair regression (UR) tested on a model update scenario, for the different cases considered going from f^*_{old} to f^*_{new}.

Dataset	**UR**	**ACCG**
Update Scenario 1: from f^*_{old} to f^*_{new} enlarging dataset size maintaining Linear SVM		
Adult	0.28 ± 0.02	0.01 ± 0.07
Arrhythmia	0.14 ± 0.06	0.01 ± 0.07
COMPAS	0.03 ± 0.02	0.00 ± 0.04
German	0.12 ± 0.06	0.00 ± 0.05
Update Scenario 2: from f^*_{old} to f^*_{new} enlarging dataset size maintaining Gaussian kernel SVM		
Adult	0.28 ± 0.02	0.03 ± 0.09
Arrhythmia	0.27 ± 0.09	-0.05 ± 0.06
COMPAS	0.03 ± 0.02	0.00 ± 0.03
German	0.12 ± 0.06	0.00 ± 0.03
Update Scenario 3: from f^*_{old} to f^*_{new} enlarging dataset size plus upgrading the kernel function		
Adult	0.28 ± 0.02	-0.06 ± 0.07
Arrhythmia	0.25 ± 0.09	0.02 ± 0.06
COMPAS	0.03 ± 0.02	0.00 ± 0.04
German	0.13 ± 0.07	0.05 ± 0.04

Figure 5.2 shows how disparities between sensitive and non-sensitive groups happen within model updates. Specifically, on the German, Adult, and Arrhythmia datasets, such differences are bigger than on the COMPAS dataset. From these results, we have a starting point for a forthcoming comparison and to examine potential trade-offs between accuracy and fairness.

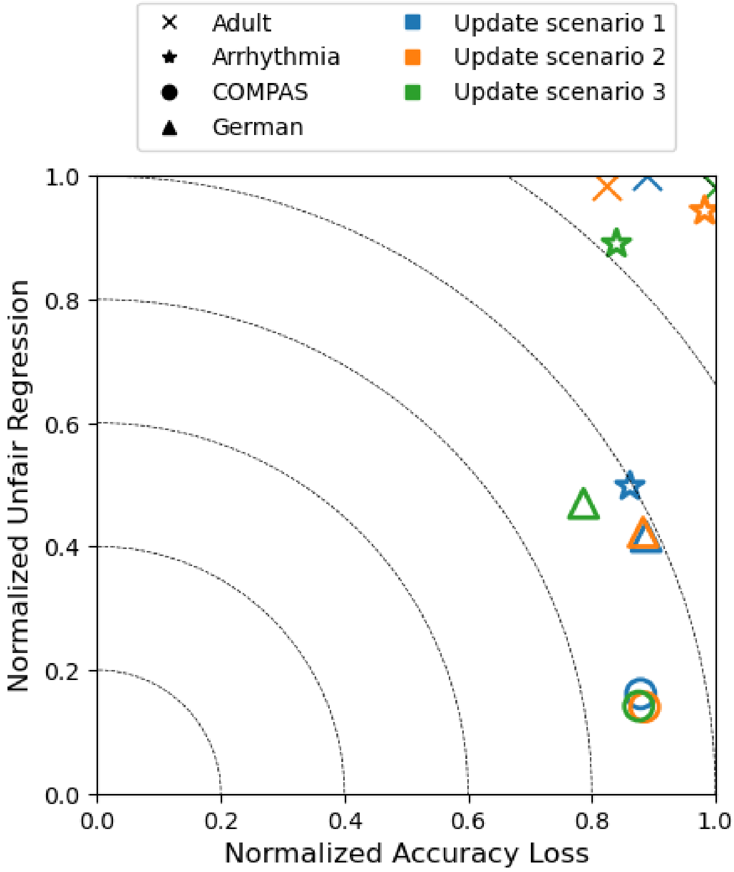


Figure 5.2 Normalized percentage of lost accuracy in model updates and unfair regression (UR). The closer the data point to the origin, the better performance in both accuracy and fairness. Note that the subscript numeration for the model notation designates the model configuration in terms of dataset size and kernel function, as listed in Table 4.2.

5.2 Mitigation Strategy (i): Double-Step Cross Validation Method

Enhanced cross validation procedures play a crucial role in our study, allowing us to observe and examine how refinements in the tuning phase impact both accuracy and UR. The novel methodology of double-step cross validation, recently introduced in the literature [51], has been shown to provide notable improvements in model performance.

While these improved models help fine-tune the model selection process and improve the robustness of results, it is important to emphasize that they do not incorporate fairness constraints within the SVM algorithm but in the model selection task. As a drawback, the learning phase does not know our intentions to incorporate fairness.

Table 5.3 Overall accuracy (ACC), balanced accuracy (B-ACC), unfair regression (UR), and accuracy gain (ACCG) obtained when using the 2-step CV method on the three different updating scenarios for various datasets with $f^*_{old} \rightarrow f^*_{new}$.

Dataset	**ACC**	**B-ACC**	**UR**	**ACCG**
Update Scenario 1: from f^*_{old} to f^*_{new} enlarging dataset size maintaining Linear SVM				
Adult	0.81 ± 0.01	0.61 ± 0.01	0.12 ± 0.01	0.16 ± 0.01
Arrhythmia	0.67 ± 0.04	0.65 ± 0.04	0.11 ± 0.05	0.01 ± 0.01
COMPAS	0.73 ± 0.01	0.72 ± 0.01	0.05 ± 0.04	0.01 ± 0.02
German	0.69 ± 0.04	0.62 ± 0.05	0.10 ± 0.06	0.01 ± 0.06
Update Scenario 2: from f^*_{old} to f^*_{new} enlarging dataset size maintaining Gaussian kernel SVM				
Adult	0.81 ± 0.01	0.61 ± 0.02	0.10 ± 0.01	0.18 ± 0.06
Arrhythmia	0.68 ± 0.04	0.66 ± 0.04	0.14 ± 0.07	0.05 ± 0.06
COMPAS	0.74 ± 0.01	0.72 ± 0.01	0.06 ± 0.04	0.00 ± 0.00
German	0.70 ± 0.03	0.64 ± 0.05	0.10 ± 0.06	0.01 ± 0.01
Update Scenario 3: from f^*_{old} to f^*_{new} enlarging dataset size plus upgrading the kernel function				
Adult	0.81 ± 0.01	0.61 ± 0.01	0.10 ± 0.01	0.18 ± 0.01
Arrhythmia	0.67 ± 0.04	0.65 ± 0.04	0.14 ± 0.06	0.05 ± 0.03
COMPAS	0.73 ± 0.01	0.72 ± 0.01	0.06 ± 0.04	0.00 ± 0.00
German	0.69 ± 0.04	0.62 ± 0.05	0.10 ± 0.06	0.01 ± 0.03

5.3 Mitigation Strategy (ii): Unfair-Regression-Free SVM Method

Our Unfair-Regression-Free Support Vector Machine (URFSVM) model is the newly developed algorithm, used to mitigate the unfair regression phenomena. Such an algorithm, accounts for the fairness constraint within the SVM framework, by minimizing the unfair regression occurrences. In other words, this method incorporates a modification in the learning phase so that the model is aware of our desire to mitigate the unfair regression. Moreover, the refinement in the model selection phase is also considered.

As a result, this method is not only expected to perform well but also must not exhibit any unfair disparities towards a specific group when doing model updates. Table 5.4 gathers the results of accuracy, balanced accuracy, unfair regression, and accuracy gained from training the new model on the specified updating scenario.

Table 5.4 Overall accuracy (ACC), balanced accuracy (B-ACC), unfair regression (UR), and accuracy gain (ACCG) obtained when using the unfair-regression-free SVM (URFSVM) method on the three different updating scenarios for various datasets with $f^*_{old} \rightarrow f^*_{new}$.

Dataset	**ACC**	**B-ACC**	**UR**	**ACCG**
Update Scenario 1: from f^*_{old} to f^*_{new} enlarging dataset size maintaining Linear SVM				
Adult	0.80 ± 0.04	0.78 ± 0.03	0.05 ± 0.04	0.00 ± 0.04
Arrhythmia	0.69 ± 0.09	0.71 ± 0.08	0.08 ± 0.05	0.38 ± 0.12
COMPAS	0.72 ± 0.05	0.73 ± 0.05	0.03 ± 0.03	0.44 ± 0.07
German	0.74 ± 0.02	0.62 ± 0.04	0.11 ± 0.11	0.43 ± 0.04
Update Scenario 2: f^*_{old} to f^*_{new} enlarging dataset size maintaining Gaussian kernel SVM				
Adult	0.78 ± 0.06	0.67 ± 0.14	0.05 ± 0.03	-0.02 ± 0.06
Arrhythmia	0.58 ± 0.07	0.62 ± 0.06	0.04 ± 0.03	0.27 ± 0.08
COMPAS	0.69 ± 0.06	0.70 ± 0.07	0.04 ± 0.03	0.41 ± 0.08
German	0.73 ± 0.03	0.55 ± 0.07	0.09 ± 0.07	0.43 ± 0.04
Update Scenario 3: f^*_{old} to f^*_{new} enlarging dataset size plus upgrading the kernel function				
Adult	0.77 ± 0.05	0.70 ± 0.13	0.05 ± 0.03	-0.03 ± 0.04
Arrhythmia	0.55 ± 0.05	0.58 ± 0.04	0.08 ± 0.06	0.27 ± 0.05
COMPAS	0.69 ± 0.07	0.70 ± 0.08	0.04 ± 0.02	0.41 ± 0.08
German	0.73 ± 0.03	0.55 ± 0.07	0.08 ± 0.07	0.42 ± 0.04

5.4 Comparison through Methods

Standard classifiers have been proven to have unfairness issues since inequity performance can be seen through metrics such as DEO and DDP, as reported in Figure 5.1. From this initial analysis, we demonstrated that fairness issues go beyond the representation of the sensitive class in the data. While the Arrhythmia dataset (which indeed is extremely unbalanced being 96% of the data representing the sensitive class) has the highest ratio of DEO in the Gaussian kernel models, the Adult dataset (which is exemplary balanced), has a high rate of unfairness as well when referring to the DDP metric. Contrarily, the linear examples of Arrhythmia have a similar value of DEO as the German dataset, in which the unbalance is not as notable.

Therefore, we can confirm fairness issues in ML are more complex than just considering the imbalance in data (a detailed explanation is outlined in Section 3.2).

On the other side, higher occurrences of UR clearly appear when no mitigation strategy is used, as expected (see Figure 5.2). When performing model updates on the standard SVM and going from f^*_{old} to f^*_{new} with no mitigation strategy applied, it is seen that while very little or no accuracy improvement is done, the regression on fairness is disproportionately unbalanced towards a particular group. Instead, after the mitigation strategies are applied, a significant decrease in the misclassification error and unbalanced regression is seen.

Unfair regression across the accuracy loss is plotted (see Figure 5.3) for the three updating scenarios and the three studied methods. Red-colored data points represent the Standard-SVM-trained classifiers, yellow-colored points the models trained with the mitigation strategy (i), and green ones stand for the mitigation strategy (ii).

A significant improvement is observed in the models trained under fairness constraints since they are generally clustered towards the lower half of the y-axis, which represents a low ratio of UR. This suggests that both of the proposed mitigation strategies effectively reduce disparities in standard SVM models.

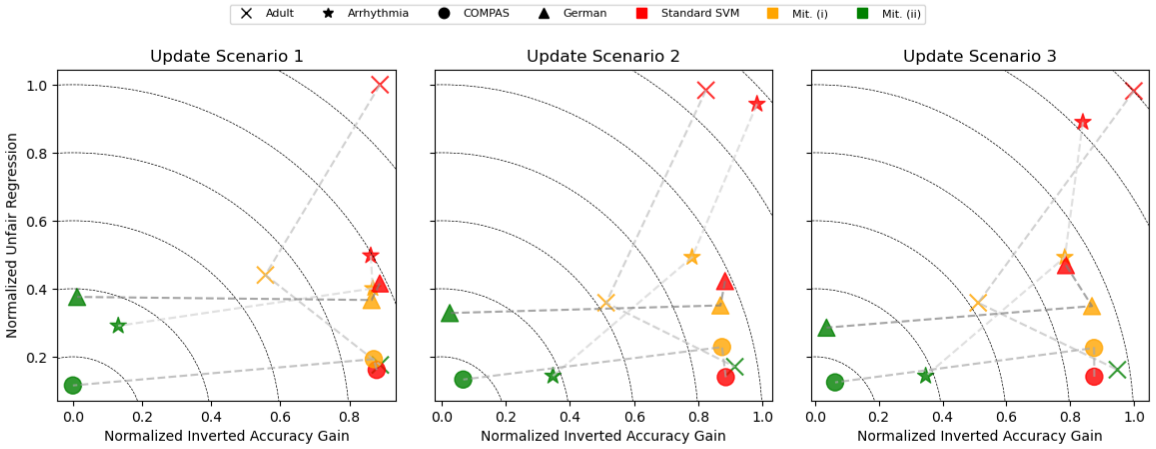


Figure 5.3 Normalized UR (the smallest the better) and inverted accuracy gain (the smallest the better) for each dataset and updating scenario. Gray lines join the results of the different methods on the same dataset. Note that, for meaningful plots, the inverted accuracy gain has been computed by subtracting the max value from each accuracy gain value and then normalized. Results retrieved from Table 5.2.

Chapter 6

Discussion

6.1 Interpretation of Findings

The results clearly show that performing model updates without any fairness mitigation leads to unbalanced regressiveness in accuracy and fairness. This disproportionate behavior affecting particular groups is evident in datasets like Arrhythmia and Adult where Standard SVM models exhibited a high degree of bias, leading to unfair outcomes.

However, the incorporation of the Double-Step Cross Validation (2-step CV) and the Unfair-Regression-Free SVM (URFSVM) strategies effectively reduced unbalanced regressiveness both in accuracy and fairness. This last mitigation strategy, URFSVM, consistently performed better in reducing unfair regression while maintaining or even increasing the overall performance. While the Gaussian kernel SVM is proven to yield better results in terms of performance, it resulted in poorer fairness metrics when compared to the Linear SVM. We hypothesize this is happening due to the simplicity of the Linear SVM, which can indeed promote more equitable treatment in some cases, leading to better fairness outcomes.

Results further demonstrate the importance of the 2-step CV method in balancing fairness and accuracy during the hyperparameter tuning, showing significant improvements in both aspects for all datasets. In particular, the Adult dataset marked significant improvements in fairness when this single first mitigation strategy was applied. On the other side, when the second mitigation strategy was introduced, the Adult dataset did not perform as good in accuracy metrics, but the Arrhythmia, COMPAS, and German datasets significantly increased their classification accuracy.

6.2 Comparison with Prior Studies

Several studies have been done in the field of fairness in machine learning, focusing on bias mitigation through the different processing techniques. For example, the work of Agarwal et. al., *A Reduction Approach to Fair Classification*, introduces a new methodology to deal with unfair classification by reducing a fair classification to a sequence of cost-sensitive classification problems, achieving fair outcomes. However, some limitations arise since the method cannot be applied if the protected attribute is not accessible during training-time, making it adequate for any type of data.

Zafar et. al. introduced through the research *Fairness Constraints: A Flexible approach for Fair Classification*, a flexible constraint-based framework to enable the design of fair margin-based classifiers. Despite the efforts, the strategy may not be extendable to every scenario since it does not work well if data is unbalanced, as contrarily seen in our work.

A similar work, made by Kamishima et. al., named *Fairness-Aware Classifier with Prejudice Remover Regularizer*, proposed a regularization approach applicable to any prediction algorithm however, the problem lacked convexity of the objective function and consequently, the method was trapped on the local minima.

Yet, the listed studies differ from our investigation since they focus on model outputs and not on the regressiveness observed in continuous updates. Instead, our approach goes beyond traditional fairness adjustments by embedding fairness constraints directly in the learning problem and considering unfair regression in model updates. Accordingly, our method would be more comparable to the research of Yan et. al. on positive congruent training. However, we extend the concept by considering sensitive attributes during model updates, tackling both accuracy and fairness.

6.3 Limitations of the Study

While our method resulted in satisfactory results, we also consider some limitations. On the first hand, the focus on the Support Vector Machines architecture as the primary machine learning method may limit the generalization to other learning structures. However, this is already planned to be expanded in our next research. On the other hand, our algorithm can perform its task only if the sensitive attribute is properly labeled and separable from the data since it is needed to start the optimization problem. Lastly, the computational cost of the proposed strategies, particularly on the URFSVM, is relatively high, which could restrict their application in real-time systems.

Chapter 7

Conclusions

This work contributes to the advancement of fairness in ML by developing and validating a novel fairness metric, namely Unfair Regression (UR), designed to quantify the unbalanced occurrences of negative flips in ML model updates. This fairness metric allowed us to develop two mitigation strategies to address the technical issues of unbalanced regressiveness in accuracy and fairness that occur when using standard methodologies.

The first method, Double-Step Cross Validation (2-step CV), which integrates fairness constraints into the hyperparameter tuning process, demonstrated a significant fairness improvement by dramatically decreasing the occurrences of UR. This strategy not only confirmed an enhancement in performance but also provided evidence of the importance of considering fairness metrics during the model selection phase. The second approach, Unfair-Regression-Free SVM (URFSVM), an algorithm created within the standard SVM framework for a binary classification problem, modifies the learning algorithm itself to minimize UR directly during the training process. The second set of experiments in real-world data, which used both of the strategies proposed, showed a further improvement in fairness but above all, a significant gain in accuracy. Adding this intermediate step of evaluating the model's performance by only modifying the tuning phase has given us an insight into the importance of incorporating fairness constraints both in the model selection and learning process.

In addition, by systematically evaluating the impact of such novel methodologies in real-world datasets, this research provides empirical evidence that improving fairness is indeed compatible with the maintainability of the model's performance. Altogether, this work contributed to the development of trustworthy AI by addressing both technical and ethical debt in ML systems.

Future work should focus on extending this study to other learning systems such as Neural Networks so that the applicability to other machine learning architectures can be assessed. Moreover, since most of the AI-based algorithms used currently in the industry belong to deep learning systems, it would significantly help promote the use of our methodology in real-world applications. In addition, testing our algorithms on other real-world datasets with different data distributions would also help us understand the applicability and limitations of such models. Furthermore, incorporating other fairness metrics such as equalized odds or demographic parity could help understand how these methods behave across different fairness definitions.

References

[1] Kirchner Lauren Angwin Julia Larson Jeff, Mattu Surya. How we analyzed the compas recidivism algorithm. https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm, 2016. Accessed: 2024-04-15.

[2] Mustafa Atay, Hailey Gipson, Tony Gwyn, and Kaushik Roy. Evaluation of gender bias in facial recognition with traditional machine learning algorithms. In *2021 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 1–7, 2021.

[3] Deepak Kumar, Tessa Grosz, Elisabeth Greif, Navid Rekabsaz, and Markus Schedl. Identifying words in job advertisements responsible for gender bias in candidate ranking systems via counterfactual learning. In *RecSys in HR'23: The 3rd Workshop on Recommender Systems for Human Resources, in conjunction with the 17th ACM Conference on Recommender Systems*, Singapore, Singapore, 2023. CEUR Workshop Proceedings, CEUR-WS.org. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

[4] Alessandro Fabris, Nina Baranowska, Matthew J. Dennis, David Graus, Philipp Hacker, Jorge Saldivar, Frederik Zuiderveen Borgesius, and Asia J. Biega. Fairness and bias in algorithmic hiring: a multidisciplinary survey. 2024.

[5] McKinsey & Company. How competition is driving ai's rapid adoption. https://www.mckinsey.com/mgi/overview/in-the-news/how-competition-is-driving-ais-rapid-adoption#, 2024. Accessed: 2024-04-10.

[6] Factored. The evolution of machine learning since the 20th century. https://factored.ai/machine-learning-engineering/, 2024. Accessed: 2024-03-20.

[7] Dena F. Mujtaba and Nihar R. Mahapatra. Ethical considerations in ai-based recruitment. In Miriam Cunningham and Paul Cunningham, editors, *2019 IEEE International Symposium on Technology in Society (ISTAS)*, East Lansing, Michigan, USA, 2019. IEEE.

[8] Sijie Yan, Yuanjun Xiong, Kaustav Kundu, Shuo Yang, Siqi Deng, Meng Wang, Wei Xia, and Stefano Soatto. Positive-congruent training: Towards regression-free model updates. *AWS/Amazon AI*, May 2021.

[9] Daniele Angioni, Luca Demetrio, Maura Pintor, Luca Oneto, Davide Anguita, Battista Biggio, and Fabio Roli. Robustness-congruent adversarial training for secure machine learning model updates. *Journal of Latex Class Files*, 18(9), 2020.

[10] Luca Oneto, Simone Minisi, Andrea Garrone, Renzo Canepa, Carlo Dambra, and Davide Anguita. Simple non regressive informed machine learning model for predictive maintenance of railway critical assets. In *30th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, ESANN 2022, Bruges, Belgium, October 5-7, 2022*, 2022.

[11] European Commission. Proposal for a regulation of the european parliament and of the council laying down harmonized rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts. COM/2021/206 final, 2021. Accessed: 2024-09-11.

[12] Ramya Srinivasan and Ajay Chander. Biases in ai systems. *Communications of the ACM*, 64(8):44–49, 2021.

[13] Simon Caton and Christian Haas. Fairness in machine learning: A survey. 2023.

[14] Civil Rights Act of 1964, Title VII, Equal Employment Opportunities. Pub. L. No. 88-352, §701 et seq., 1964. [Online]. Available: https://www.eeoc.gov/statutes/title-vii-civil-rights-act-1964.

[15] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez-Rodriguez, and Krishna P Gummadi. Fairness constraints: A flexible approach for fair classification. *Journal of Machine Learning Research*, 20:1–42, 2019.

[16] Graziele Teles, Joel J. P. C. Rodrigues, Kashif Saleem, Jalal Al-Muhtadi, and Muhammad Imran. Machine learning and decision support system on credit scoring. *Neural Computing and Applications*, 32:9809–9826, 2020.

[17] Acar Burak Muderrisoglu Haldun Guvenir, H. and R. Quinlan. Arrhythmia. UCI Machine Learning Repository, 1998. DOI: https://doi.org/10.24432/C5BS32.

[18] Dana Pessach and Erez Shmueli. A review on fairness in machine learning. *ACM Comput. Surv.*, 55(3), feb 2022.

[19] Milena A. Gianfrancesco, Suzanne Tamang, Jinoos Yazdany, and Gabriela Schmajuk. Potential biases in machine learning algorithms using electronic health record data. *JAMA Internal Medicine*, 178(11):1544–1547, 2018.

[20] Amir Masoud Rahmani, Efat Yousefpoor, Mohammad Sadegh Yousefpoor, Zahid Mehmood, Amir Haider, Mehdi Hosseinzadeh, and Rizwan Ali Naqvi. Machine learning (ml) in medicine: Review, applications, and challenges. *Mathematics*, 9(22), 2021.

[21] Flavio P. Calmon, Dennis Wei, Karthikeyan Natesan Ramamurthy, and Kush R. Varshney. Optimized data pre-processing for discrimination prevention, 2017.

[22] Michael Feldman, Sorelle Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact, 2015.

[23] Faisal Kamiran and Toon Calders. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 33(1):1–33, 2012.

[24] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In Sanjoy Dasgupta and David McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 325–333, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR.

[25] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudik, John Langford, and Hanna Wallach. A reductions approach to fair classification. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 60–69. PMLR, 10–15 Jul 2018.

[26] L. Elisa Celis, Lingxiao Huang, Vijay Keswani, and Nisheeth K. Vishnoi. Classification with fairness constraints: A meta-algorithm with provable guarantees, 2020.

[27] Naman Goel, Mohammad Yaghini, and Boi Faltings. Non-discriminatory machine learning through convex fairness criteria. In *AAAI Conference on Artificial Intelligence*, pages 3029–3036. AAAI, 2018.

[28] Padala Manisha and Sujit P. Gujar. A neural network framework for fair classifier. *arXiv preprint arXiv:1811.00247*, 2018.

[29] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '18, page 335–340, New York, NY, USA, 2018. Association for Computing Machinery.

[30] Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, pages 3315–3323, 2016.

[31] Faisal Kamiran, Asim Karim, and Xiangliang Zhang. Decision theory for discrimination-aware classification. In *2012 IEEE 12th International Conference on Data Mining*, pages 924–929, 2012.

[32] Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q. Weinberger. On fairness and calibration. In *Advances in Neural Information Processing Systems*, pages 5680–5689, 2017.

[33] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P. Gummadi. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th International Conference on World Wide Web*, WWW '17, page 1171–1180, Republic and Canton of Geneva, CHE, 2017. International World Wide Web Conferences Steering Committee.

[34] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rogriguez, and Krishna P. Gummadi. Fairness Constraints: Mechanisms for Fair Classification. In Aarti Singh and Jerry Zhu, editors, *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 962–970. PMLR, 20–22 Apr 2017.

[35] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. Fairness-aware classifier with prejudice remover regularizer. In Peter A. Flach, Tijl De Bie, and Nello Cristianini, editors, *Machine Learning and Knowledge Discovery in Databases*, pages 35–50, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg.

[36] Jianhui Gao, Benson Chou, Zachary R. McCaw, Hilary Thurston, Paul Varghese, Chuan Hong, and Jessica Gronsbell. A tutorial on fairness in machine learning in healthcare, 2024.

[37] Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning: Limitations and Opportunities*. fairmlbook.org, 2023. Accessed: 2023-12-13.

[38] Luca Oneto. *Model Selection and Error Estimation in a Nutshell. Modeling and Optimization in Science and Technologies*. Springer Nature Switzerland, 2019.

[39] Bernhard Schölkopf and Alexander J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, 2001.

[40] Maria Han Veiga and François Gaston Ged. *The Mathematics of Machine Learning: Lectures on Supervised Methods and Beyond*. De Gruyter Textbook. De Gruyter, 2024.

[41] Shuzhan Fan. Understanding the mathematics behind support vector machines, May 2018. Accessed: 2024-09-04.

[42] Tammy Jiang, Jaimie L. Gradus, and Anthony J. Rosellini. Supervised machine learning: A brief primer. *Behavior Therapy*, 51(5):675–687, 2020.

[43] Scikit-learn developers. Understanding the parameters of the rbf kernel for svm classification. https://scikit-learn.org/stable/auto_examples/svm/plot_rbf_parameters.html, 2024. Accessed: 2024-04-15.

[44] Sara C. Larson. The shrinkage of the coefficient of multiple correlation. *Journal of Educational Psychology*, 22:45–55, 1931.

[45] Frederick Mosteller and J. W. Tukey. Data analysis, including statistics. In G. Lindzey and E. Aronson, editors, *Handbook of Social Psychology, Vol. 2*. Addison-Wesley, 1968.

[46] Mervyn Stone. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society: Series B (Methodological)*, 36:111–147, 1974.

[47] Seymour Geisser. The predictive sample reuse method with applications. *Journal of the American Statistical Association*, 70:320–328, 1975.

[48] Luc Devroye and T. Wagner. Distribution-free performance bounds for potential function rules. *Information Theory, IEEE Transactions on*, IT-25:601 – 604, 10 1979.

[49] Peter L. Bartlett, Stephane Boucheron, and Gábor Lugosi. Model selection and error estimation, Oct 2002.

[50] Sylvain Arlot and Alain Celisse. A survey of cross-validation procedures for model selection. *Statistics Surveys*, 4(none), January 2010.

[51] Michele Donini, Luca Oneto, Shai Ben-David, John Shawe-Taylor, and Massimiliano Pontil. Empirical risk minimization under fairness constraints. In *32nd Conference on Neural Information Processing Systems (NIPS 2018)*, Montréal, Canada, 2018. Neural Information Processing Systems Foundation.

[52] Sahil Verma and Julia Rubin. Fairness definitions explained. In *IEEE/ACM International Workshop on Software Fairness (FairWare'18)*, pages 1–7, Gothenburg, Sweden, 2018. ACM.

[53] Tiago P. Pagano, Rafael B. Loureiro, Fernanda V. N. Lisboa, Rodrigo M. Peixoto, Guilherme A. S. Guimarães, Gustavo O. R. Cruz, Maira M. Araujo, Lucas L. Santos, Marco A. S. Cruz, Ewerton L. S. Oliveira, et al. Bias and unfairness in machine learning models: A systematic review on datasets, tools, fairness metrics, and identification and mitigation methods. *Big Data and Cognitive Computing*, 7:15, 2023.

[54] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. ACM, January 2019.

[55] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard S. Zemel. Fairness through awareness. *CoRR*, abs/1104.3913, 2011.

[56] Pratik Gajane and Mykola Pechenizkiy. On formalizing fairness in prediction with machine learning. *arXiv preprint arXiv:1710.03184v3*, 2018.

[57] Christian Haas. The price of fairness - a framework to explore trade-offs in algorithmic fairness. In *Fortieth International Conference on Information Systems*, Munich, Germany, 2019. University of Nebraska at Omaha, Information Systems and Quantitative Analysis, Omaha, NE, USA.

[58] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In Sanjoy Dasgupta and David McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 325–333, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR.

[59] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P. Gummadi. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th International Conference on World Wide Web*, pages 1171–1180, Perth, Australia, 2017. International World Wide Web Conferences Steering Committee.

[60] D. Sculley, Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips, Dietmar Ebner, Vinay Chaudhary, Michael Young, Jean-François Crespo, and Dan Dennison. Hidden technical debt in machine learning systems. In *Advances in Neural Information Processing Systems 28 (NIPS 2015)*, pages 2494–2502. Curran Associates, Inc., 2015.

[61] Ai fairness 360. https://aif360.mybluemix.net/. Accessed: 2024-04-10.

[62] Niels. Bantilan. Themis-ml: A fairness-aware machine learning interface for end-to-end discrimination discovery and mitigation. *Journal of Technology in Human Services*, 36(1):15–30, 2018.

[63] Matt J. Kusner, Joshua R. Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness, 2018.

[64] Faisal Kamiran and Toon Calders. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 33(1):1–33, 2012.

[65] Sorelle A. Friedler, Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan P. Hamilton, and Derek Roth. A comparative study of fairness-enhancing interventions in machine learning. http://arxiv.org/abs/1802.04422, Feb 2018. Accessed: 2024-04-10. Code available at https://github.com/algofairness/fairness-comparison.

[66] Toon Calders and Sicco Verwer. Three naive bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery*, 21(2):277–292, 2010.

[67] Toon Calders, Faisal Kamiran, and Mykola Pechenizkiy. Building classifiers with independency constraints. In *2009 IEEE International Conference on Data Mining Workshops*, pages 13–18, 2009.

[68] Barry Becker and Ronny Kohavi. Adult. UCI Machine Learning Repository, 1996. DOI: https://doi.org/10.24432/C5XW20.

[69] Hans Hofmann. Statlog (German Credit Data). UCI Machine Learning Repository, 1994. DOI: https://doi.org/10.24432/C5NC77.