UNIVERSIDAD ADOLFO IBÁÑEZ

MASTER'S THESIS

Fairness and Transparency in ML: A Methodological Framework for Ethical Model Evaluation

Author:
Nelson Salazar Valdebenito

Supervisors:
Gonzalo Ruz Heredia
Reinel Tabares Soto

Examination Committee:
Rolando de la Cruz
Mauricio Valle

Thesis conducted in accordance with the requirements for the Master of Science in Data Science degree

from the

Faculty of Engineering and Sciences

January 31, 2024

UAI FACULTAD DE INGENIERÍA Y CIENCIAS

UNIVERSIDAD ADOLFO IBÁÑEZ

Abstract

Faculty of Engineering and Sciences

Master of Science in Data Science

by Nelson Salazar Valdebenito

In the times when AI and ML applications are more widely spread across our society, new challenges arise when these are being used by public organisations. Ethical concerns may emerge for various reasons, but especially when these models inherit certain biases that could eventually lead up to hurting and/or punishing people with no reason whatsoever. In this study, a comprehensive methodology is introduced to address ethical considerations in machine learning, particularly focusing on biases and disparities. This thesis showcases the application of this methodology through two experiments, each involving two models: a base model without ethical considerations and an improved one. The primary objective was to evaluate and mitigate inequalities and biases in machine learning models, in this case, using data from Chile's Public Criminal Defence Office (DPP), which needed a tool that could help in predicting the outcome of a criminal trial. The results were promising, with improvements observed in both models. Specifically, for one of our experiments, there was a reduction in the false negative rate from 40.54% to 31.63%. Additionally, disparities between attributes were reduced by an average of 44.37%. The methodology not only aids in understanding these ethical challenges but also offers tools to optimize and address them, all while maintaining performance.

Keywords: Ethical AI, Ethical Algorithms, Fairness in AI, Responsible AI

Acknowledgements

I want to thank my family: Natalia, Nelson and Francisco, for shaping me into the man and professional I am today. This is also to my friends and colleagues (both from Buk and Pisapapeles), for always supporting me during this period.

Additionally, I would like to thank my supervisors, Reinel and Gonzalo. It was a privilege working with such talented researchers, and I'm thankful for their support, advise and compromise with my work. Thanks to the Algoritmos Éticos team for letting me work on my thesis with them, as well as to Chile's Public Defence Office (DPP) for providing access to the data used in this study.

This research was supported by IDB Lab, innovation laboratory of the Inter-American Development Bank Group [Project ATN/ME-18240-CH], National Agency for Research and Development, ANID + Applied Research Subdirection/ IDeA I+D 2023 grant [folio ID23I10357], ANID PIA/BASAL FB0002, ANID/PIA/ANILLO ACT210096, CH-T1246 : Oportunidades de Mercado para las Empresas de Tecnología - Compras Públicas de Algoritmos Responsables, Éticos y Transparentes.

Finally, I would like to thank my cats, Mila and Zapallo, for helping me endure on those long nights of work.

Contents

Chapter 1

Introduction

Artificial Intelligence (AI) and Machine Learning (ML) models are no longer the future, but the present; and even if we do not realize it, they are already part of our daily lives through various things, such as the algorithm that recommends movies to us or what our social media feed displays.

However, the application of these tools is not always the same. There are several diferences between a model that predicts which movie one would want to see, as to have one that, for example, is designed to predict the outcome of a criminal trial.

Whilst a movie recommendation algorithm can make mistakes without having a significant impact on our lives (that is, it suggests a movie we really do not like), the situation changes dramatically when these algorithms are applied in contexts with serious consequences for individuals. For example, in the judicial system, an algorithm attempting to predict the outcomes of a criminal trial could influence a person's sentence, potentially altering their life significantly.

Ethical issues can arise when we consider the biases present in the data with which these algorithms are trained. If the judicial system has historically had a bias towards certain demographic groups, an algorithm trained with this data may perpetuate, or even intensify, this bias.

If such technology is implemented in crucial sectors such as healthcare, finance or law enforcement, the ramifications can be considerable and harmful. In these industries, an erroneous forecast or partial verdict could result in an incorrect diagnosis, economic adversity or false imprisonment. It is not exclusively about receiving inaccurate recommendations; rather, it concerns the compromise of fundamental rights, fairness and impartiality.

In addition, another ethical problem is the opacity of these models. Since they are often considered "black boxes", it can be difficult for those affected, or even for those charged with implementing them, to understand exactly how the algorithm arrives at a particular decision. This makes it difficult to challenge a decision or to ensure that the decision has been made fairly.

The use of ML applications in public contexts also raises questions about liability and consent. Who is responsible when an algorithm makes a harmful decision? How can we ensure that these models do not possess risks to our society? How can we be sure that the individuals affected by these decisions were properly informed about the use of such tools?

As mentioned in the text *Analysis of Ethical Development for Public Policies in the Acquisition of AI-Based Systems*[53], the growth of artificial intelligence and its applications

in various areas of society present new challenges for the public sector, mainly in terms of regulation. In June 2023, the European Union approved its first AI regulatory act, which sets out harmonised rules for AI systems. The regulation outlines specific bans on certain AI practices, mandates for high-risk systems, and emphasises transparency in AI-human interactions and media manipulation. It also provides a comprehensive framework for market surveillance, covering both providers and users of AI systems in the EU [14]. Governments are not the only ones interested in this, industry is too. OpenAI wants regulation for what it calls "frontier AI" [1], i.e. models whose capabilities could pose serious risks to public safety.

In addition, AI cannot replicate many human capabilities. In the field of law, a lawyer must combine abstract thinking with skills to solve problems in situations of high uncertainty, in both the legal and factual domains; something a classification model cannot yet achieve [52]. This is, models cannot "reason" in the same manner as a human being. Even the large language models — such as OpenAI's ChatGPT or Google's Bard — that have become so popular recently are not capable of doing so, or at least not independently [25, 57]. This is due to the fact that behind that black box, there are only probabilities and computations among vectors and matrices. Hence, models should be "ethically trained" in order to avoid these problems.

It is crucial to recognise that while the potential benefits of AI and ML are enormous, the risks they pose, especially if left unchecked, can have profound implications for individuals and society as a whole.

1.1 Key Definitions

Several terms have become crucial when considering the ethical and practical implications of models and their decision-making capabilities. These terms shed light on how models process information, how they can potentially introduce inequalities, and how to interpret their results. Some of these fundamental concepts that are going to be used in this thesis are as follows:

1. **Transparency:** This refers to the comprehensibility of an algorithm's decision-making process. In essence, algorithms should not operate as "black boxes" where decisions are made without clarity. Ideally, we should have a clear understanding, or at least a basic idea, of how the algorithm works.

2. **Bias:** This refers to any systematic bias in the algorithm's results that favours certain groups over others. For example, an algorithm that consistently predicts that people from a certain demographic group are likely to underperform in a job - regardless of reality - is biased.

3. **Fairness:** This ensures that each group of people is treated fairly by the algorithm. In practical terms, the results of the algorithm should not unfairly discriminate against any particular group.

4. **Disparities:** This term addresses the differences in outcomes or effects that models produce across various demographic groups. Disparities can arise even in the absence of explicit bias, highlighting the nuanced nature of fairness. It underscores the importance of evaluating the broader impacts of algorithms across different chapters of the population to ensure equitable results.

5. **SHAP Value**: Comes from SHapley Additive exPlanations. Is a game-theoretic approach to explaining the outputs of machine learning models. Inspired by

Shapley values, they assign the contribution of each feature to a prediction, ensuring a fair distribution. This provides a better understanding of model interpretability and individual predictions [47].

1.2 Research Context & Opportunity

This thesis was produced as part of the Ethical Algorithms (*Algoritmos Éticos*) project, led by the GobLab of the University Adolfo Ibáñez and developed in collaboration with the Inter-American Development Bank (IDB), which aims to incorporate ethical standards into the development of tools based on ML and AI models. These ethical standards aim to improve the levels of transparency, fairness, privacy, explainability, and accountability of ML models. It also seeks to establish new purchasing standards for these tools in public tenders.

This project focuses primarily on those performed by or for public institutions and/or state agencies. Currently, UAI is providing advice and technical assistance to the Public Criminal Defense Office, FONASA, the Ministry of Science and Technology, among other entities.

In particular, this thesis is part of the work carried out for the Public Defence Office (DPP) of Chile, a public body that provides defence services to anyone who does not have a lawyer to defend them in a criminal trial. The DPP needed a predictive system to help predict the outcome of a trial in order to carry out audit and quality control tasks on its defence lawyers.

Due to the nature of the predictions these models will make, ethical concerns may arise. That is why there are multiple tools that can be used to minimise these problems. Ethical assessment tools can serve several purposes, including identifying disparities and biases in models. These tools, such as What-If-Tool [58], SHAP [36], AI Fairness 360 [5], Fairlearn [56], and Aequitas [49], are effective for analysing model performance, optimising models, and auditing their results. For documentation purposes, there are different tools available such as Model Cards [39] and Datasheets for Datasets [22], among others.

This thesis provides a comprehensive walkthrough of the entire process underlying our proposed methodology, which is used to evaluate inequalities and biases in machine learning models. This framework aims not only to help analyse and understand these problems, but also to mitigate them by using different tools to optimise and address them. This begins with an exploratory data analysis, which lays the foundation for understanding the nature of the data we are working with. This is followed by the training phase, which is critical to the construction of the initial model. Then, as an iterative step, the model gets refined to ensure that ethical considerations are appropriately incorporated. Finally, when no major biases or disparities are found (or there is balance between performance and fairness), the resulting model gets documented using a comprehensive yet simple documentation framework that ensures transparency in the qualities and functioning of the model.

The methodology was put into test by doing two experiments using the datasets provided by the DPP; where each of them has two models: a base one with no ethical considerations, and an improved one. In the end, both models were enhanced in both disparities and biases, whilst having a relatively minimal impact on performance. Our best results include a false negative rate reduction 40.54% to 31.63%,

and its disparities between attribute reduced by an average of 44.37% for one of the experiments.

1.3 Thesis Structure

This thesis is structured as follows. Chapter 2 presents the research question as well as the objectives proposed for this project. Chapter 3 contains the state of the art, reviewing relevant literature and previous works in the field of ethical AI. Chapter 4 presents the proposed methodology, detailing the techniques and approaches used, as well as the datasets and tools that were used, and the experiments that were done. In Chapter 5, the experiments, models and results are shown; whilst in Chapter 6 those results are discussed. In Chapter 7 contains the conclussions and future work. Finally, Appendix A and B contains the documentation/model report for the final (enhanced) models.

Chapter 2

Research Question and Objectives

2.1 Research Question

How can the fairness levels of a machine learning model be improved through the use of ethical tools?

2.2 General Objective

Design an ethical evaluation methodology for machine learning models, based on the use of tools that allow the analysis, optimisation, and visualisation of the characteristics and performance of these models, in order to improve their levels of fairness and transparency.

2.3 Specific Objectives

1. Identify and assess the ethical tools currently available for enhancing the fairness and transparency of machine learning models.
2. Incorporate these ethical tools into an evaluation methodology tailored specifically for machine learning models.
3. Design and develop machine learning models using DPP's datasets, applying some mild feature engineering and hyperparameter optimisations.
4. Apply this methodology on these models, analysing the impact on their fairness and transparency levels.

Chapter 3

State of the Art

In this chapter, we outline the state of the art in ethical evaluation for ML models. We begin by examining how AI has been applied in the legal sector, serving as a foundational context due to the thesis project's alignment with a penal entity. From there, we provide an overview of existing work methodologies and their core principles. Finally, we discuss the practical applications of notable ethical evaluation tools, like Aequitas and the What If Tool, in various areas. The aim of this chapter is to set the stage for the development of a methodology that can be universally applied across diverse contexts.

3.1 ML & AI applied in law

The application of ML in the field of law is not yet widespread. This is due to the fact that law and legislation in general heavily rely on human reasoning and cognitive abilities, which clearly cannot always be replicated with artificial systems.

As mentioned in Surden (2014)[52], machines lack the ability to replicate many of the intellectual capabilities of humans, especially the more advanced ones (such as the ability to reason through the use of analogies). Surden gives an example where many lawyers must combine abstract thinking with problem-solving skills in situations of high uncertainty, both in legal and factual fields; something that an AI system obviously cannot do at present.

However, we limit ourselves by thinking that the only potential application of AI and ML in the field of law lies in replacing humans in making decisions on such delicate issues. As in many other areas and industries, it is best to see how these systems can assist in other aspects of law or legislation, such as performing predictive and/or prescriptive analyses on what is going to happen or what could happen in a trial based on the evidence at hand; it can also be used to understand certain phenomena, such as corruption.

The application of ML in law does not necessarily seek to replace human reasoning as such, but rather aims to complement its practice. Thus, in the case of the project to be developed, these models will not only aim to assist in defense processes, but also in quality control of the work carried out by defenders.

In Mahfouz & Kandil (2012)[37], several ML models are proposed to predict the outcome of various litigation processes related to disputes over construction sites. Among the models used, since the goal was to classify into one class, Support Vector Machines (SVM) [13], a naive Bayesian classifier, and various neural network

models (specifically, decision trees) were utilized, with the former showing the best performance.

To gauge corruption levels across different countries, Melo & Delen (2020)[34] employed various ML algorithms to understand the factors explaining corruption levels (such as education levels, government integrity, effectiveness of the judicial system, and so on). In this case, being a multiclass classification model, three types of models were used: Random Forest [28, 10] (which proved to be the most effective), neural networks, and SVMs.

To predict the outcome of judicial decisions — particularly in divorce cases — Li et al (2019)[33] introduced a concept distinct from traditional ML models, leaning more towards cognitive computing as a means to model and understand the reasoning of the Chinese judicial system. This approach achieved better results than a SVM or a neural network model.

In the field of crime and criminal law, there are plenty of examples that use ML and AI for classifaction and predictive purposes, such as in Mitchell et al (2020)[38] and Xu et al (2022)[60]. In the former, various ML models were used to predict the outcome of criminal trials, using the presented evidence as a training method. In this case, a naive Bayesian classifier was chosen to establish a simple and effective model. In the latter article, deep learning models were used to classify and assess four common crimes in China, with the aim of identifying cases that may have been wrongly adjudicated as guilty.

Continuing on the topic, crime prediction stands out as one of the most prevalent applications of ML and AI in this area. This approach takes advantage of the large amount of data on crimes, their locations, and other factors to predict potential criminal activity, thereby aiding law enforcement agencies and policymakers in their decision-making processes, such as in preventive patrols or in investigative processes carried out by police forces. Multiple examples can be found, such as in Shah et al. (2021)[50], where advanced techniques like deep learning and computer vision were utilised to create a system for law enforcement designed to monitor crime hotspots and recognize individuals through voice notes. Similar applications can be found in Rummens et al. (2017) [48], as well as in Rani & Rajasree (2014)[46] and in Kim et. al (2018)[31], where different approaches were proposed to do predictive analysis for crimes in different contexts.

Finally, AI and ML techniques have also been applied to predict trial outcomes and the potential charges faced by defendants. Ye et al. (2018)[61] explored this by developing a method to generate judicial explanations from the details of criminal cases, providing a clearer understanding of how charges are determined. Similarly, Zhong et al. (2017)[64] introduced a framework that focuses on predicting judgment results, emphasizing the potential of AI to assist in legal decision-making processes. Predictions can also be used to support the judge, as Feng et al. (2021)[19] proposed a system for statute recommendation, which uses neural networks to predict relevant legal statutes based on case facts that could be leveraged by the judge to make its job more efficient and accurate when analysing multiple cases at the same time.

3.2 Assessing ethical concerns

Ethical ML/AI is not a new concept, as it has been previously assessed by legal systems as well as in academia, deliberating on issues like data privacy, algorithmic bias, and transparency, emphasizing the need for guidelines that ensure fairness and accountability. Interpretability and explainability are one of the most important aspects of ML and AI systems, as the nature of the decisions that these models can make can be critical in certain situations, especially when applied in the public sector.

Studies such as Bibal et al. (2021) [7] and Doshi-Velez (2017) [18] delve into the role of explainability from the legal point of view, and they both conclude that there exists a distinct misalignment between the legal understanding of "explainability" and that of ML. On the one hand, the technical community primarily seeks to unravel the intricate workings of the machine itself, aiming for a deep comprehension of its internal mechanics and logic. On the other hand, the legal perspective emphasizes the explicability of the decisions and outcomes produced by the machine, focusing on transparency and justification rather than the nitty-gritty technical details. This divergence in views can lead to challenges in AI development, as the pursuit of explainability might introduce constraints that hamper model complexity or optimization. Consequently, aiming for higher levels of explainability can inadvertently escalate the cost of model creation, both in terms of resources and time.

Numerous studies have proposed a range of guidelines and methodologies related to AI ethics. These range from technical guidelines and analyses [4, 9, 40] to their practical implementation in healthcare [11, 55], and further on how these systems can successfully integrate into our society [21, 62]. However, a consensus on a unified standard, both technically and ethically speaking, remains elusive [20, 63], as many of these studies are task-specific and focused on AI development rather than on ethical considerations [3, 41].

Jobin et al. (2019) [30] suggests that while there are some convergence in some topics (such as transparency, fairness, explainability and privacy), there are multiple differences in the conceptual and procedural aspects of these guidelines, such as the importance and interpretation of said principles, who they pertain to and how they should be implemented. On the other hand, Hagendorff (2020)[24] highlights the prevailing gaps in AI ethics, emphasizing the lack of tangible consequences for ethical deviations. He argues that while ethics is often touted in institutional setups, it predominantly serves promotional purposes, with software developers frequently perceiving it as a supplementary, non-binding element, often overshadowed by economic motives. This disconnect underscores the misalignment between AI applications and foundational societal values. Another problem, as it is noted in Lo Piano (2020)[35], is that there are some ML models that are constantly learning, meaning that these guidelines and methodologies could prove unsuccessful to explain the same models in the future.

3.3 Ethical tools applied

Ethical assessment tools can serve several purposes, including identifying disparities and biases in models. These tools, such as What-If-Tool [58], SHAP [36], AI Fairness 360 [5], Fairlearn [56], and Aequitas [49], are effective for analysing model

performance, optimising models, and auditing their results. For documentation purposes, there are different tools available such as Model Cards [39] and Datasheets for Datasets [22], among others.

The use of ethical tools and techniques for ML algorithms is not widespread when developing such models. This is largely due to its recent emergence and a lack of obligation for their use by private companies. However, in the public sector, this becomes a more pressing issue.

For instance, a common use case is in models that classify taxpayers, debtors, and financial and tax-related issues. Specifically, Black et al. (2022) [8] conducted a study on income fairness in the tax audit models used by the US' Internal Revenue Service (IRS). The study examined vertical equity, that is, how much the model represents important individual differences, given their relevance to public tax policies. Using various ethical tools, the study discovered that prioritising return on investment (ROI) from auditing lower-income taxpayers by the Internal Revenue Service (IRS) compromises vertical fairness in an effort to decrease the expenses associated with such audits.

Healthcare is another crucial field for ML models. Rajkomar et al. (2018) [45] investigated the effects of these algorithms on medical care, particularly how biases and disparities could put more vulnerable groups at risk of incorrect diagnoses or treatments. The research employed distributive justice strategies to guarantee impartial outcomes for patients and ensures that resource allocation is as equitable as possible.

Singh and Joachims (2018) [51] proposed a methodology to ensure fairness in rankings. The study emphasized the prevalence of rankings in various contexts, including books, movies, jobs, and individual comparisons. Therefore, it is essential for classification algorithms to prioritize not only the benefit of the end-user but also the fairness of the ranked entity. The research has presented various probabilistic approaches that act as constraints in diverse situations, like ensuring demographic parity or addressing disparities in data.

Assessing fairness has also been done in the field of criminal justice. Chouldechova (2016) [12] applies this to recidivism prediction instruments, finding that there are disparate impact when recidivism prevalence differs across groups, meaning that even when an instrument is free from predictive bias, its application can still lead to unintended disproportionate adverse impacts on specific groups, especially when individuals assessed as high risk face stricter penalties. In Berk et al. (2017) [6] and Kleinberg et al. (2016) [32] it is discussed that fairness in risk assessment tools cannot be maximised along with accuracy, and that there will be tradeoffs between performance and fairness, as the base rates can significantly influence the outcomes and predictions. These base rates refer to the fundamental probabilities or initial rates of certain events or outcomes. For instance, in a criminal justice context, a base rate might pertain to the likelihood of an individual on parole either failing or succeeding. Adjustments to these base rates can lead to potential fairness and accuracy challenges, emphasizing the intricate balance between ensuring fairness and maintaining predictive accuracy.

Ethical algorithms encompass not only the models' operations but also their documentation, specifically their transparency. Heger et al. (2022) [27] examined the needs and perspectives of individuals constructing ML models on data documentation, using datasheets for datasets. The study uncovered the absence of standardised methodologies for comprehensive and user-friendly documentation. Furthermore,

there remains a disparity in users' attitudes towards documentation, often seen as an extra burden rather than a tool for facilitating accountable employment of ML models.

Data also have important implications for model training. According to a recent study on the use of ML models to detect COVID-19 in X-ray images from various databases [2], researchers discovered that data imbalances and biases in patient age and sex can lead to underperformance of the algorithm, which affects the accuracy and generalisability of predictions.

Furthermore, Davidson et al. (2019) [15] determined that numerous Twitter/X datasets exhibit biases towards African American English speakers/writers. This is particularly problematic when applied in technologies for detecting abusive language, as these systems tend to classify these tweets as abusive more frequently than those written in standard English, resulting in discrimination against groups who are often the targets of such language abuse. According to Wiegand et al. (2019) [59], these models' biases are closely related to the way the data was sampled.

Chapter 4

Methodology

The purpose of this chapter is to present the methodology that is going to be used for developing this thesis, as well as the databases and tools that are going to be applied.

As previously mentioned, this thesis is part of the "Algoritmos Éticos" project made by the GobLab of the University Adolfo Ibáñez, and in particular, has been developed within the project made for the Public Criminal Defense Office (Defensoría Penal Pública, DPP). Hence, the approach will mainly focus on the application of ethical tools within the judicial context, and specifically for the needs of the DPP. However, the idea is that this methodology can be applied to similar projects and in other cases where it could be necessary to conduct an ethical analysis in ML models.

4.1 Methodology

The framework proposed by this project applies multiple tools that can be used to analyse and improve ML models in terms of fairness and biases. These tools are completely agnostic to the technique used for creating a model (e.g. gradient boosting machines, neural networks, among others), and have different use cases, such as for analysing, optimising and documenting the models.

The idea is that this framework can answer questions such as:

- How is our model performing in terms of fairness and biases?
- How can we improve it?
- How this model works?

The execution of this will follow a similar approach presented in Beretta (2023) [23], which consists of a two-part experiment: the first one, using a base/normal model; whilst the second one uses a fairer model, that takes into considerations some ethical improvements.

However, our approach is not specifically focused on the training itself, but rather on the pre and post training of the model; and it also considers other aspects such as the analytical and documentation part for a complete ethical model evaluation.

In particular, the framework follows these steps:

1. Do some basic exploratory data analysis (EDA) in order to identify the nature of the data, including potential disparities and class imbalances.

2. Take the base model, and evaluate it in terms of its performance, biases (throught counterfactual analysis) and fairness (disparities).
3. With the information gathered in the previous section, we create an enhanced model for fairness and ethical considerations. For this, we want to:
    - Mitigate potential disparities in the training and testing datasets.
    - Balance sample and class weights.
    - Optimise model hyperparameters.
    - Select the optimal classification threshold for a prediction.
4. Evaluate this new model, and compare it to the base iteration.
5. Apply documentation frameworks and other tools that can help to make a

Figure 4.1 shows a detailed flowchart for the proposed framework.

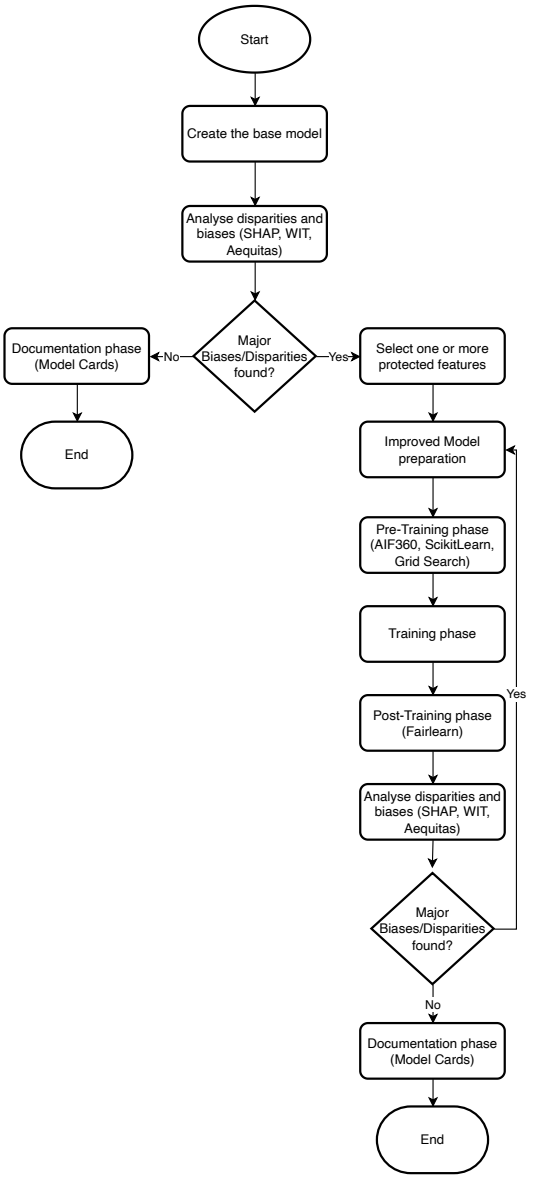


FIGURE 4.1: Proposed framework flowchart.

4.2 Problem description

As it was previously mentioned, the problem comes from Chile's Public Criminal Defense Office (*Defensoría Penal Pública*, DPP). This entity needed a tool based on ML techniques that could help in predicting the outcome of a trial, so that they could conduct a series of audits and quality control tests on its lawyers.

Since this implies the use of sensitive data from people processed by different crimes, it requires some kind of ethical analysis of the models developed for this purpose, specially since they are trying to predict an outcome that could potentially change the life of the defendant. In particular, every model developed has one class to predict: if a person has a favourable (1) or unfavourable (0) outcome to its trial. An unfavourable outcome occurs when the defendant is imprisoned and/or its sentence gets increased. A favourable outcome is anything but the previous mentioned – i.e.: sentence reduced, paroled or acquitted.

One of the main concerns that can arise is the way in which the model(s) handles the predictions. This is because of the potential biases that may arise due to the nature of the crime - i.e. its prevalence in certain areas of the country, or by whom it is committed. In that regard, it is important to determine whether one or more models (one per crime) should be developed. A single model would facilitate the development process, but because the crimes may be so different, it may have problems of accuracy, which would mean that it would not be very effective for its purpose. On the other hand, with multiple models, this shall not be a problem, as each would be a specialist in one crime. This can also facilitate ethical analysis, as it allows for the model to inherit any biases and differences that might be found within a given crime.

Similarly, another concern that may arise is the way in which these models are used. While they are intended to be used for auditing purposes, it cannot be ruled out that a lawyer might use them to determine whether or not it is in his or her interest to defend an accused person. It is important to have clear documentation that correctly defines the purpose of these tools, how they work and a clear definition of the potential risks associated with their use.

4.3 Databases & Tools

All the information that will be used for the development this thesis will come from the database of crimes processed by the DPP between the years 2017 and 2022. Specifically, two files provided in a CSV format will be used, which correspond to two types of crimes: drug trafficking, and petty theft.

Each dataset contains multiple records of individuals accused of said crimes, and both of them have a mixture of defendant-related and trial-related columns. The second ones are directly related to the outcome of the trial, so those cannot be used for training purposes (as it can introduce noise and biases). Hence, we are using those related to the defendant: region (geographical location; "Region"), level of development of the crime ("Desarrollo"), hearings ("Audiencias Efectivas") and the lawyer ("Defensor").

However, there are many critical variables (such as the age, sex, and previous cases of the defendant, among others) that were not available in these sets. The absence of these features obviously hurts the performance that a model can have, but it also

limits the capabilities of analysing potential biases and disparities. Hence, some feature engineering should be done, and will be explained in Chapter 5.

4.3.1 Development & ethical tools

In this section, we look at the essential development and ethical tools utilised throughout our research. Those in the first category are as follows:

1. **Python**: A versatile and widely-used programming language, Python offers extensive libraries and frameworks for ML and data analysis, facilitating the design and implementation of our algorithms [54].
2. **Data processing and visualisation**: This includes libraries such as Pandas [42] and Numpy [26] for data for data manipulation, analysis, and transformation, providing a foundation for our data preprocessing needs. For visualisations, we use Matplotlib [29] for basic graphs and plots.

Optimisation tools:

1. **Scikit Learn**: An open-source ML library for Python, Scikit Learn provides simple tools for data analysis and modelling, making it easier to implement ML algorithms [43].
2. **AI Fairness 360**: An extensible open-source toolkit that offers a comprehensive set of fairness metrics for datasets and ML models, as well as algorithms to mitigate biases.
3. **Fairlearn**: A tool that focuses on assessing and improving fairness in ML models, offering mitigation algorithms and fairness metrics.

Analytical and documentation tools:

1. **What-If-Tool**: An interactive visual interface developed by Google researchers that provides in-depth exploration capabilities for ML models, allowing users to understand their predictions more profoundly. This includes the ability to look at the distribution of the data, facilitates the values of certain metrics, review counterfactuals (similar data points with which the model delivers different results) and analyse the overall performance of the models.
2. **SHAP**: Offers a game-theoretic approach to model explanations, providing clarity on feature contributions to individual predictions, enabling more transparent model interpretability.
3. **Aequitas**: An open-source bias audit toolkit, has a set of tools that allows an ethical analysis to the results generated by the model, where the data and the results are contrasted to measure the levels of bias that a model may have. The purpose of this tool is to improve the fairness within attributes, providing metrics and visualisations that allow decisions to be made regarding the quality of the model in terms of biases. Aequitas facilitates the process of understanding and interpreting disparities in ML models across various groups.
4. **Model Cards**: A framework developed by Google Cloud researchers that is designed to provide comprehensive and standardised reporting on ML model performances, ensuring transparency in their capabilities and limitations across diverse operational environments and user groups. Just like the way the nutritional information label that comes on food packaging works, Model Cards

seeks to contain all the information associated with a model: its inputs, outputs, type of model (classifier, regression), ML method used (multilayer perceptron, random forest, XGBoost, etc.), among other data.

4.4 Experiments

Two experiments were conducted, one with the drug trafficking set, and the second one with the petty theft set. Each experiment was divided into two parts: the base situation and the improved situation.

As it was previously mentioned, the purpose of these models is to predict if a defendant will a have a favourable or unfavourable outcome to its trial. A set can have one or several outcomes (i.e. "Outcome 1", "Outcome 2", etc), which can ocurr in different instances. For the sake of simplicity, the models developed in here will always take the last outcome available. For the drug traffiking set, it will be the second outcome; for the petty theft dataset, it will be the first (and only) one.

1. The base situation consists of a basic model that can predict the previous mentioned classes. Once the model is created, the results are analysed using SHAP values [36], the What If Tool [58] and Aequitas [49], so that the biases and disparities get identified.
2. In the improved situation, as its name suggests, the model gets refined upon the insights gathered in the first part. This also considers the selection of a protected feature. In this case, for both experiments, it is going to be the "Region" feature, since it is the only variable that can be a potential source of bias, and that comes directly from datasets.
   (a) For the pre-training phase, we use AI Fairness 360 [5] and its Reweighting tool to assign new weights to the sample data, as well as Scikit-Learn's [43] *compute_class_weight* function for optimising the class weights. Some basic hyperparmeter tunning was also applied for some gains in model performance.
   (b) For the post-training phase, we use Fairlearn's *ThresholdOptimizer* function for optimising the prediction threshold of our model. The constraint used for this is the false negative rate parity, since we are interested in minimising the amount of people that has a non-favourable ending to its trial. After that, the model gets analysed using the same tools mentioned before.
3. When the final model is selected (i.e. there were no major biases/disparities found in the enhanced model, as stated in Figure 4.1), it is documented using Google's Model Cards framework for model documentation. A model card can be made on any platform or format, and in this case, these were made in Canva.

For the protected variable, we want to select those regions that are under-represented in the dataset. There are a total of 16 regions in Chile, and since we want to ensure a fairly equal representation of these places, in both experiments the regions will be sorted by number of records (highest to lowest), and half of them will be selected (8) as favored and/or unfavored.

Additionally, a new variable was added just for analysis purposes (therefore, it was not used for the training phases). This is "Zona", or the zone of the country, as Chile can be divided into three different zones according to its region: north, central and south. This was done by assigning each region to its respective zone, and it can be useful for analysing the nature of the data and the overall impact of the model in a broader way than "Region".

Chapter 5

Results

In this chapter, we present the results of two experiments. For each one, we compare the base model to its improved version. There are two sections – one for each experiment – which are composed by three subsections. These are the following:

1. ***Subsection 5.x.1***: Shows some basic EDA, the feature selection and feature engineering done for the models.
2. ***Subsection 5.x.2***: Shows the architecture, hyperparameters and training phase for both the base and improved versions.
3. ***Subsection 5.x.3***: The actual results of the models, divided into four parts:
   (a) Model Performance
   (b) SHAP Values
   (c) Fairness & Disparities (using What-If-Tool and Aequitas)

Where x corresponds to the first (1) or second (2) experiment.

5.1 Experiment 1: Drug Traffiking

5.1.1 Data Analysis & Feature Selection

Using the dataset "Trafico Drogas 2017 - 2022", we have the following characteristics:

- There are 16,690 non-null rows, which provides a good basis for training models.
- The columns we are interested in are those associated with the defendant, ignoring any type of name or identifier. Thus, the columns "Region", "Desarrollo", "Audiencias Efectivas" and "Defensor" are relevant to the analysis. Two new variables were added to provide a better analysis and contexto to the models we are going to train: "Edad" (age) and "Extranjero" (foreigner).
- For the first one, we use the information from the Fourteenth National Drug Survey in the General Population of Chile, made by the National Service for the Prevention and Rehabilitation of Drug and Alcohol Consumption (SENDA) [16] in Chile, which states that people between the ages of 18 and 34 years old concentrates the majority of the drug consumption in the country. Hence, for this experiment, we take a normal distribution with a mean of 26 (the average between 18 and 34) and a standard deviation of 8.

- For the foreigner variable, we use the column "P.S. Expulsión" as the basis for it, as it indicates whether the defendant was sentenced to deportation. This variable is not the most ideal, as it may not take into account cases where there has been no such sentence; and being deported is considered a favourable outcome, so there may be a bias in this variable.
- There are no visible class imbalances in this set.

Figure 5.1 and Figure 5.2 shows the distribution of outcomes by region and zone, respectively:

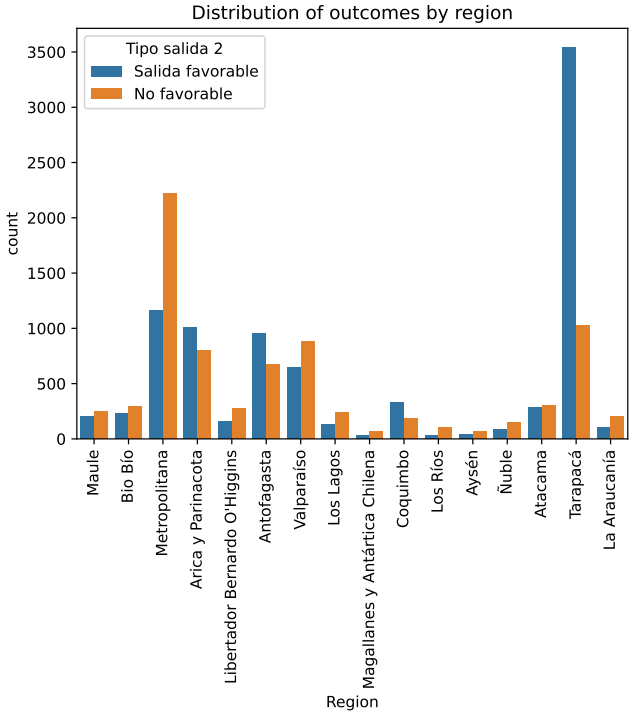


FIGURE 5.1: Outcome distribution by region.

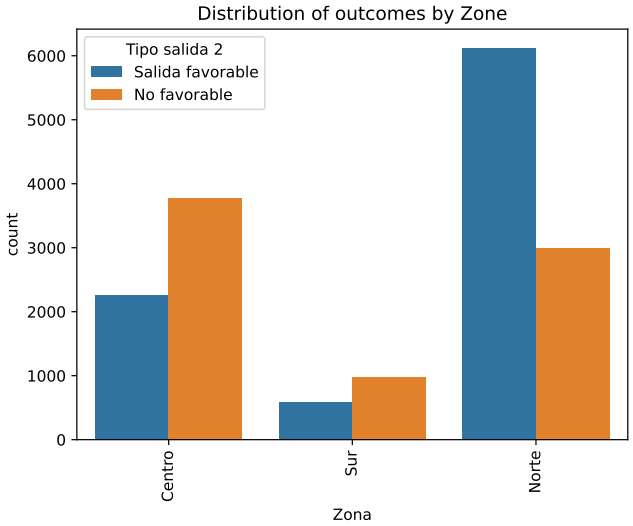


FIGURE 5.2: Outcome distribution by zone.

From these figures it can be seen that the prevalence of this crime is mainly on the northern regions of the country, followed by the central zone, and finally the southern zone, which, compared to the other two, has least amount of cases related to this crime.

Note that the northern zone is the only one where there is a majority of favourable outcomes, whilst on the other ones the unfavourable outcomes are more prevalent, showing that the severity of the courts across the country can vary by region. This could potentially generate biases, and may particularly favour cases from the northern regions of the country.

5.1.2 Models & Training

For this experiment, a multi-layer perceptron was used. The details of the architecture and training for the base situation are as follows:

- **Architecture**: Multi-Layer Perceptron (MLP), developed with Tensorflow, and trained to minimise the binary crossentropy.
- **Type**: Classification. In this case, to predict if a person has a favourable outcome (1) or not (0).
- **Datasets**: 70% split for the training set, 30% for the testing set, and a 1000 rows were removed from the training dataset to create a validation dataset. In total, there are 10683 rows for training, and 5007 for testing.
- **Hyperparameters**:
  - 70 training epochs.
  - Learning rate: 0.01
  - Batch size: 1024.
  - Layers: 4 - One input layer, two hidden layers and one output layer.
  - Neuron configuration: 6, 13, 5, 1.
  - Activation functions: Sigmoid for each layer.

For the second part of this experiment, the improved situation, there is a similar configuration with a few changes, those being the amount of neurons for the hidden layers: 14 for the first one, and 12 for the second one, just for testing purposes; as well as both hidden layers now using a ReLU activation function.

This new model also uses the improved sample weights that were calculated by AIF360's reweighting function. The class weight were also balanced using Scikit-Learn's *compute_class_weight* method, even though it was not really necessary to do so. The Model Card (documentation) for this enhanced version can be found in Appendix 1.

5.1.3 Model results

In this subsection, the actual results of the models are shown, divided into four parts (Performance, SHAP Values, Predictions, and Fairness & Disparities).

Model Performance

Model performance for both models are shown in Table 5.1:

Model/Dataset	Accuracy	Precision	Recall	F1 Score
Base Train	80%	100%	62%	77%
Base Test	78%	100%	59%	75%
Improved Train	67%	67%	72%	70%
Improved Test	65%	67%	68%	68%

TABLE 5.1: Model performance.

SHAP Values

SHAP values were calculated by using the data from the training set for both models. Table 5.2 presents the absolute SHAP values, indicating the overall effect on the model, whereas Table 5.3 provides information about its impact on the predictions (if it contributes towards positive or negative predictions).

Model	Lawyer	Age	Region	Development	Hearings	Foreigner
Base	0.006392	0.001494	0.016019	0.000132	0.000645	0.330375
Improved	0.051984	0.004848	0.046972	0.000172	0.002176	0.217328

TABLE 5.2: Average absolute SHAP Values by feature.

Model	Lawyer	Age	Region	Development	Hearings	Foreigner
Base	-0.000258	0.000114	0.000420	0.000132	0.000057	-0.027403
Improved	0.003826	-0.001084	0.009636	0.000172	0.000040	0.017798

TABLE 5.3: Average SHAP Values by feature.

Figure 5.3 and Figure 5.4 should be interpreted in the following way. The X-axis represents the real SHAP value and its effect on predictions. Negative SHAP values, on the left side of zero, indicate that the variable contributes to a negative prediction by the model. Positive SHAP values, on the right side of zero, indicate that the variable contributes to positive predictions. The feature value is indicated by the colour, where blue denotes lower values and red higher ones.

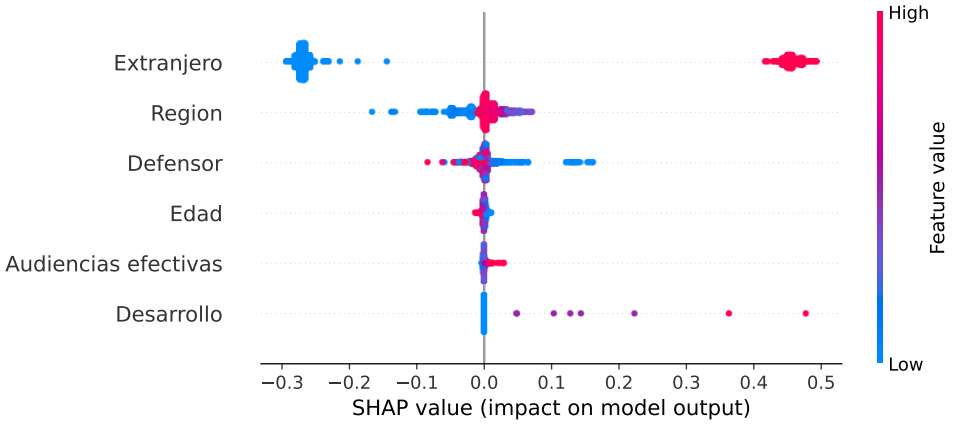


FIGURE 5.3: SHAP values distribution by feature (base model).

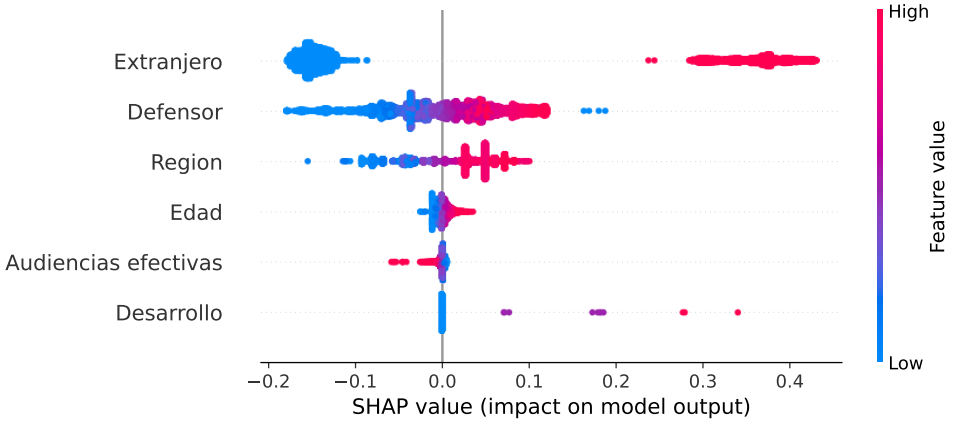


FIGURE 5.4: SHAP values distribution by feature (optimised model).

Fairness & Disparities (WIT + Aequitas)

The overall metrics for the False Positive Rate (FPR), False Negative Rate (FNR), False Discovery Rate (FDR) and False Omission Rate (FOR) between attributes are shown in Table 5.4. The confusion matrix for both models can be found on Table 5.5.

Model	**FPR**	**FNR**	**FDR**	**FOR**
Base	0.13%	40.54%	0.19%	31.88%
Improved	39.25%	31.63%	33.25%	37.51%

TABLE 5.4: Bias metrics between attributes.

Model	**True Positives**	**True Negatives**	**False Positives**	**False Negatives**
Base	1594	2323	3	1087
Improved	1833	1413	913	848

TABLE 5.5: Confussion Matrix.

Figure 5.5 and Figure 5.6 shows us the distribution of predictions made by the base and improved models, binned by the actual value of the attribute “Tipo de salida 2”.

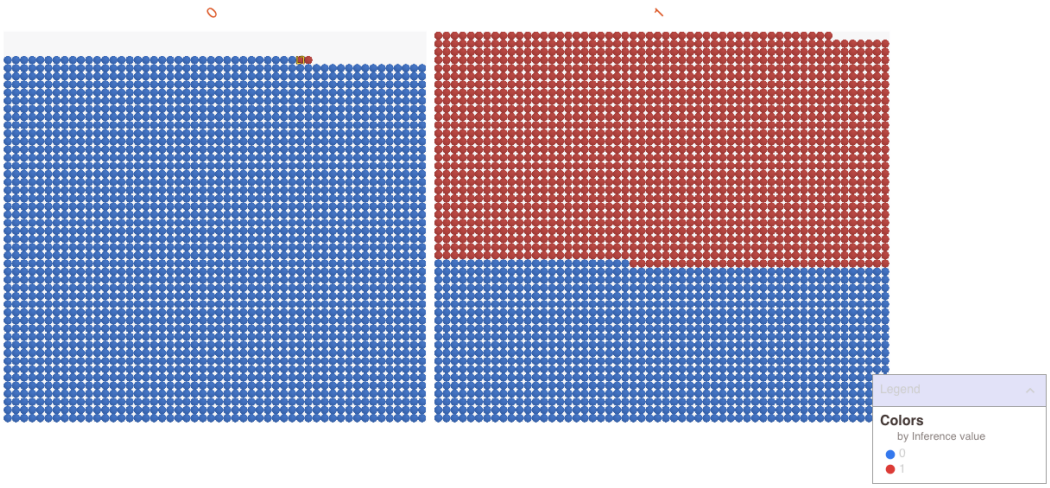


FIGURE 5.5: Distribution of predictions from the base model, binned by the actual outcome.

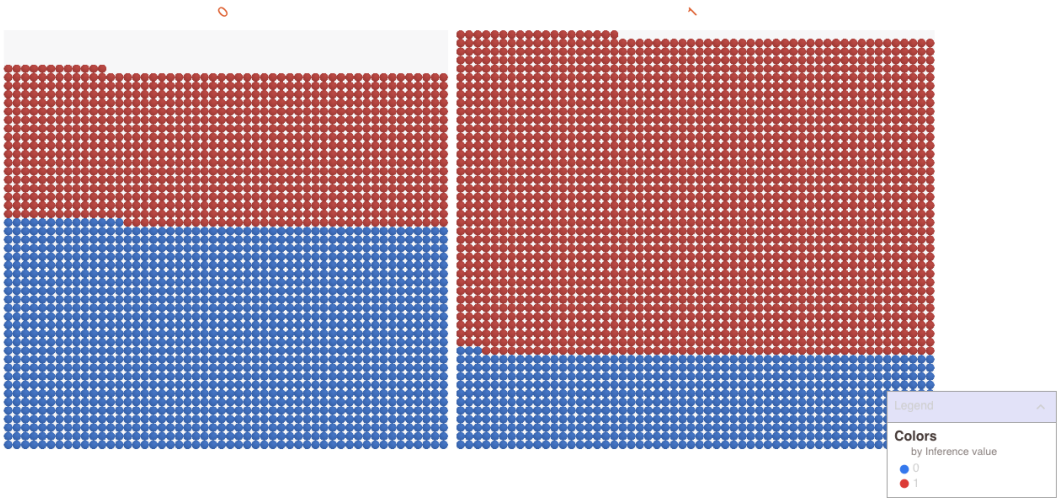


FIGURE 5.6: Distribution of predictions from the improved model, binned by the actual outcome.

In terms of disparties, three assessments were made using the different reference groups that Aequitas admits. These are the following: the major group for each feature, the group with the lowest metrics, and a predifined group. In this case, it was set to a defendant from the Metropolitan Region (central zone), aged between 18 and 28 years old, chilean (not foreigner), whose felony was consummated.

These assesments were made considering the True Positive Rate, the False Omission Rate and the False Negative Rate. The TPR is of interest to us to see how positive predictions are distributed within the groups of each feature; FOR is useful to see where the model is failing to make correct predictions; while the FNR aids in understanding the proportion of false negatives over the total actual positives. Finally, our disparity tolerance (τ) is set to 1.5 times larger or smaller than the size of each reference group.

Table 5.6, as well as Figures 5.7, 5.8 and 5.9 are the assessment results for the base model; while Table 5.7 is the crosstab for the variable "Region", broken down by

attribute value. Columns “TPR”, “FNR”, “FOR” and “FDR” are the previously mentioned bias metrics.

Assessment	**FNR Disparities**	**FOR Disparities**	**TPR Disparities**
Reference Group	0.81	1.04	21.005
Major Group	7.53	1.61	19.96
Min. Metrics Group	9.15	2.04	28.12

TABLE 5.6: Average disparities for the base model.

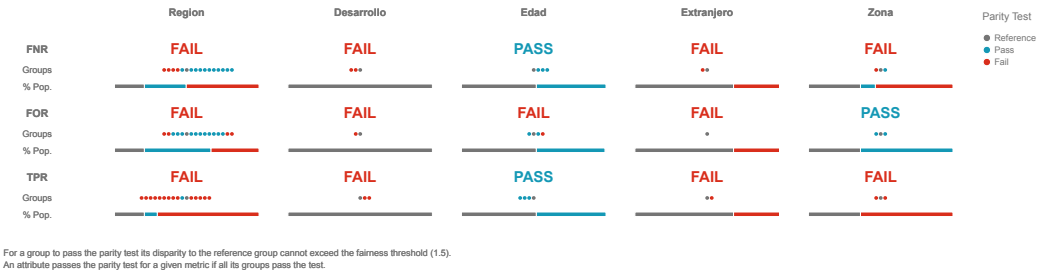


FIGURE 5.7: Model assessment for the reference group.

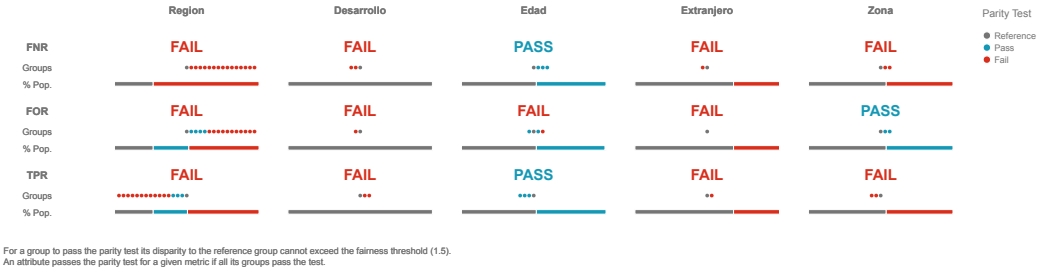


FIGURE 5.8: Model assessment for the major group.

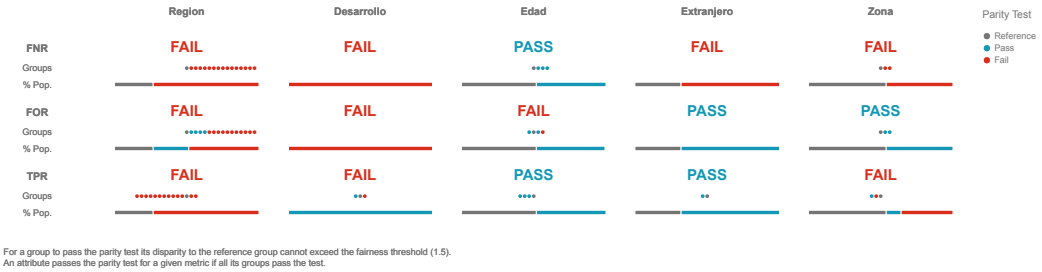


FIGURE 5.9: Model assessment for the min metrics group.

Attribute Value	TPR	FNR	FOR	FDR
Antofagasta	0.8213	0.1787	0.2167	0.0000
Arica y Parinacota	0.7276	0.2724	0.2562	0.0000
Atacama	0.6782	0.3218	0.2456	0.0000
Aysén	0.0000	1.0000	0.3889	-
Bio Bío	0.0139	0.9861	0.4057	0.0000
Coquimbo	0.2547	0.7453	0.5852	0.0000
La Araucanía	0.0556	0.9444	0.3238	0.0000
Libertador Bernardo O'Higgins	0.0566	0.9434	0.3759	0.2500
Los Lagos	0.0227	0.9773	0.3554	0.0000
Los Ríos	0.0000	1.0000	0.1818	-
Magallanes y Antártica Chilena	0.0909	0.9091	0.3448	0.0000
Maule	0.0156	0.9844	0.5000	0.0000
Metropolitana	0.1525	0.8475	0.3067	0.0357
Tarapacá	0.9377	0.0623	0.1720	0.0000
Valparaíso	0.1117	0.8883	0.3911	0.0000
Ñuble	0.0833	0.9167	0.4000	0.0000

TABLE 5.7: Average bias metrics between attributes for the base model.

Table 5.8, along with Figures 5.10, 5.11 and 5.12 are the assessment results for the enhanced model. Table 5.9 is the crosstab for the variable "Region".

Assessment	FNR Disparities	FOR Disparities	TPR Disparities
Reference Group	1.05	1.21	0.99
Major Group	1.06	0.87	0.99
Min. Metrics Group	2.12	2.20	1.78

TABLE 5.8: Average disparities for the improved model.

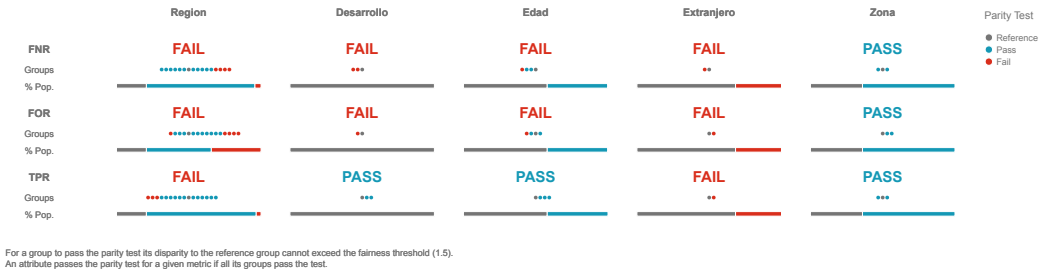


FIGURE 5.10: Model assessment for the reference group.

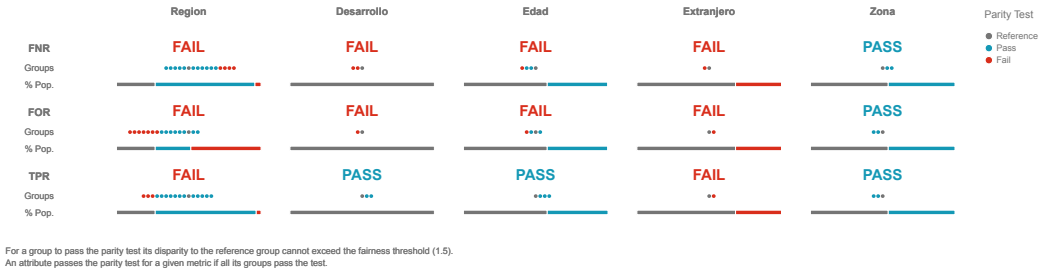


FIGURE 5.11: Model assessment for the major group.

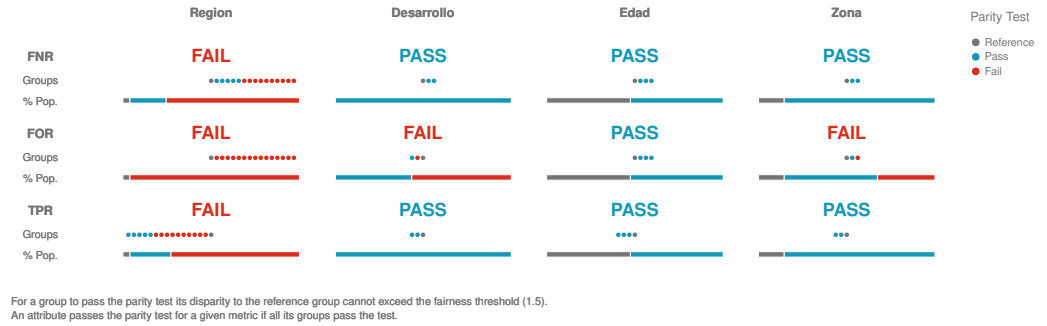


FIGURE 5.12: Model assessment for the min metrics group.

Attribute Value	**TPR**	**FNR**	**FOR**	**FDR**
Antofagasta	0.7113	0.2887	0.3088	0.0000
Arica y Parinacota	0.7276	0.2724	0.2562	0.0000
Atacama	0.7126	0.2874	0.2427	0.1143
Aysén	0.3571	0.6429	0.4091	0.6429
Bio Bío	0.7083	0.2917	0.4286	0.5984
Coquimbo	0.6226	0.3774	0.6154	0.3196
La Araucanía	0.4167	0.5833	0.4286	0.7414
Libertador Bernardo O'Higgins	0.7925	0.2075	0.2895	0.5758
Los Lagos	0.5682	0.4318	0.3800	0.6528
Los Ríos	0.5000	0.5000	0.1818	0.8182
Magallanes y Antártica Chilena	0.2727	0.7273	0.3333	0.5000
Maule	0.5938	0.4062	0.5098	0.5000
Metropolitana	0.6949	0.3051	0.2935	0.6306
Tarapacá	0.6955	0.3045	0.5040	0.0000
Valparaíso	0.6383	0.3617	0.3598	0.5367
Ñuble	0.6250	0.3750	0.5000	0.6150

TABLE 5.9: Average bias metrics between attributes for the improved model.

5.2 Experiment 2: Petty Theft

5.2.1 Data Analysis & Feature Selection

Using the dataset "Hurto Falta 2017 - 2022", we have the following characteristics:

- There are 39,954 non-null rows, which provides a good basis for training models.
- The columns we are interested in are those associated with the defendant, ignoring any type of name or identifier. Thus, the columns "Region", "Desarrollo", "Audiencias Efectivas" and "Defensor" are relevant to the analysis. In this case, the variable "Edad" (age) was the only one added to this set, as the column "P.S. Expulsión" was not present in this file.
- To create this new variable, we use the data from the Centre for Crime Studies and Analysis (CEAD) for the crime of theft between the years 2017 and 2022 [17]. This variable also considers a normal distribution with a mean of 31 (average age between 18 and 44, which accounts for the majority of those charged with this crime) and a standard deviation of 8.
- There are significant class imbalances in this set, as there are 28,819 unfavourable outcomes versus 11,135 favourable outcomes. This means that there are 2.5 times more unfavourable outcomes than favourable ones.

Figure 5.13 and Figure 5.14 shows the distribution of outcomes by region and zone, respectively.

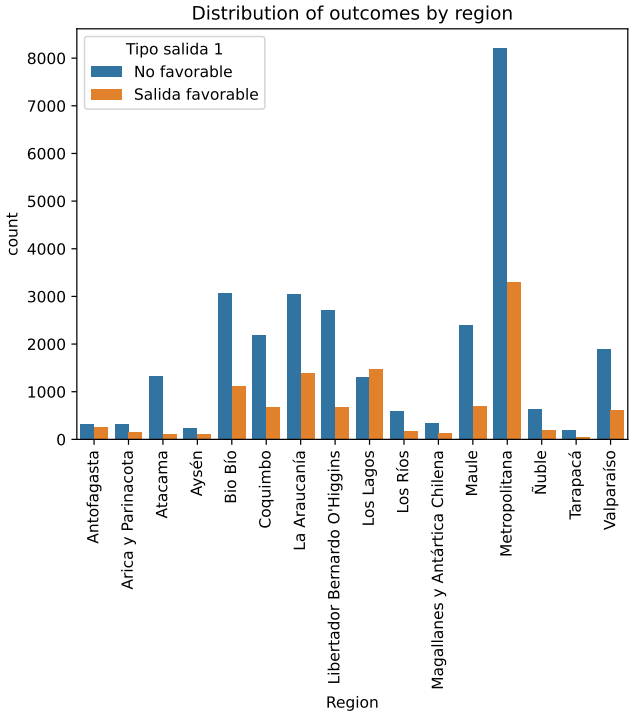


FIGURE 5.13: Outcome distribution by region.

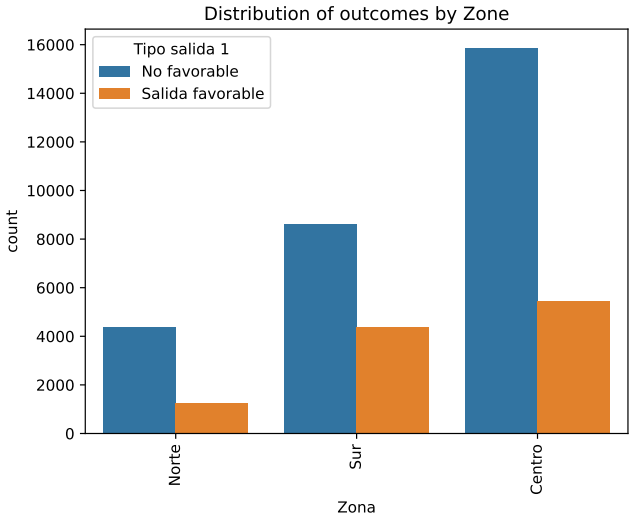


FIGURE 5.14: Outcome distribution by zone.

It can be seen that this crime/felony is more prevalent in the Metropolitan Region and, overall, in the central part of the country; followed by southern and northern zones, respectively.

However, compared to the previous experiment, the severity of the courts is more consistent across regions. This is in line with what was mentioned above, where there is a clear prevalence of the unfavourable outcome for this crime. So, in principle, the main source of bias between regions may be the higher concentration of cases in the Metropolitan Region, with other regions being underrepresented as a result.

5.2.2 Models & Training

The second experiment has a different model architecture, in this case using a gradient-boosting framework. The details for the base situation are as follows:

- **Architecture**: Light Gradient-Boosting Machine (LGBM), using the *lightgbm* library.
- **Type**: Classification. In this case, to predict if a person has a favourable ending (1) or not (0).
- **Datasets**: 70% split for the training set, 30% for the testing set. In total, there are 28967 rows for training, and 11987 for testing.
- **Hyperparameters**:
  - Boosting type: gbdt
  - 100 estimators.
  - 63 leaves per estimator (tree).
  - Unlimited maximum depth.
  - Learning rate: 0.01

For the improved model, the amount of leaves per tree goes up to 120 after doing a basic grid search optimisation. It also considers the new sample weights calculated by AIF360, as well as the new class weights from the *compute_class_weight* function. The Model Card for this improved version can be found in Appendix 2.

5.2.3 Model Results

In this subsection, the actual results of the models are shown, divided into four parts (Performance, SHAP Values, Predictions and Fairness & Disparities):

Performance

Model performance is shown in Table 5.10.

Model/Dataset	Accuracy	Precision	Recall	F1 Score
Base Train	78%	77%	33%	46%
Base Test	76%	67%	27%	38%
Improved Train	77%	62%	43%	51%
Improved Test	74%	53%	35%	42%

TABLE 5.10: Model performance.

SHAP Values

SHAP Values were calculated using the data from the training set for both models. The averages are laid out in Table 5.11 and Table 5.12.

Model	Lawyer	Age	Region	Development	Hearings
Base	0.060050	0.004756	0.186943	0.167577	0.377383
Improved	0.088094	0.009539	0.183160	0.188467	0.374409

TABLE 5.11: Average absolute SHAP Values by feature.

Model	Lawyer	Age	Region	Development	Hearings
Base	-0.019104	0.000492	0.023519	0.010751	-0.027672
Improved	-0.016245	-0.000893	0.010753	0.012887	-0.031332

TABLE 5.12: Average SHAP Values by feature.

The distribution of these values can be found in Figures 5.15 and 5.16.

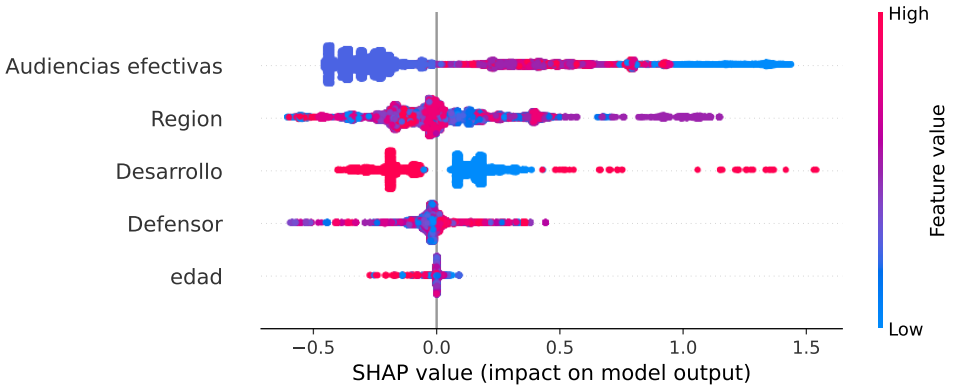


FIGURE 5.15: SHAP values distribution by feature (base model).

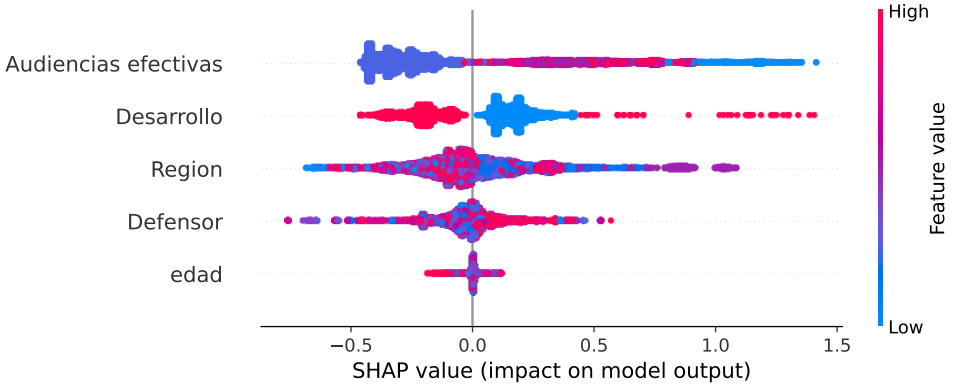


FIGURE 5.16: SHAP values distribution by feature (improved model).

Fairness & Disparities (Aequitas + WIT)

The overall metrics for the False Positive Rate (FPR), False Negative Rate (FNR), False Discovery Rate (FDR) and False Omission Rate (FOR) are shown in Table 5.13. The confusion matrix for both models can be found on Table 5.14.

Model	FPR	FNR	FDR	FOR
Base	4.96%	73.20%	32.98%	22.46%
Improved	12.21%	64.16%	47.54%	21.56%

TABLE 5.13: Bias metrics between attributes.

Model	True Positives	True Negatives	False Positives	False Negatives
Base	878	8279	432	2398
Improved	1174	7647	1064	2102

TABLE 5.14: Confussion Matrix.

Figure 5.17 and Figure 5.18 shows us the distribution of predictions made by the base and improved models, binned by the actual value of the attribute "Tipo de salida 1":

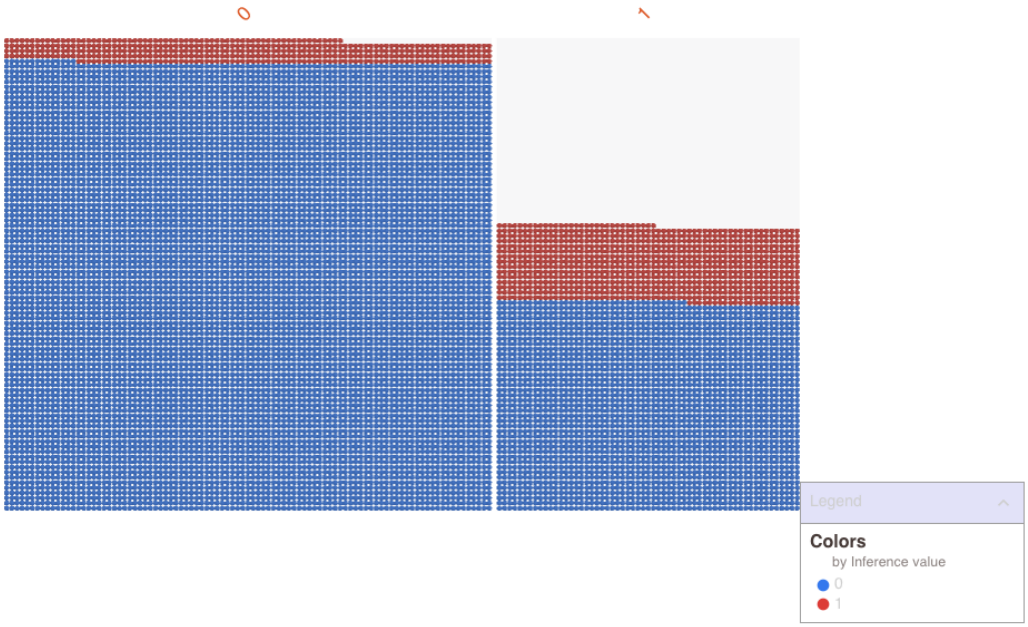


FIGURE 5.17: Distribution of predictions from the base model, binned by the actual outcome.

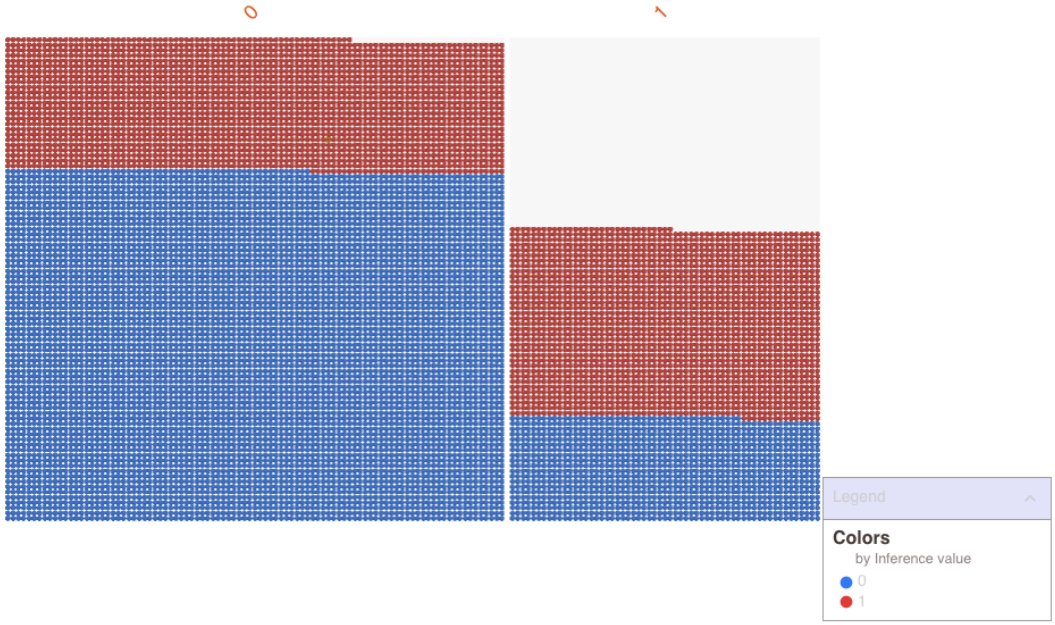


FIGURE 5.18: Distribution of predictions from the improved model, binned by the actual outcome.

The assessments made for this part of the experiment were the same – one with the major group for each feature, the group with the lower measures and a reference group. The metrics used for this task were also the same: FNR, FOR and TPR disparities; and the results for the base model can be found in Table 5.15 and in Figures 5.19, 5.20 and 5.21. Table 5.16 is the crosstab for the variable "Region'.

Assessment	**FNR Disparities**	**FOR Disparities**	**TPR Disparities**
Reference Group	1.273	1.165	0.644
Major Group	1.273	1.165	0.644
Min. Metrics Group	2.170	2.899	6.272

TABLE 5.15: Average disparities for the base model.

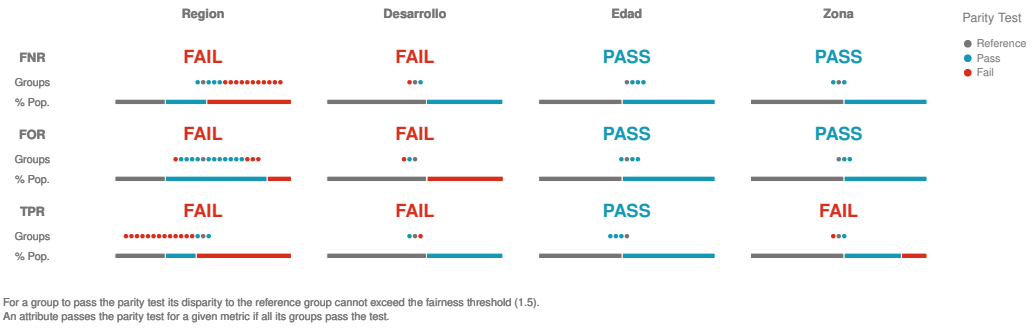


FIGURE 5.19: Model assessment for the reference group.

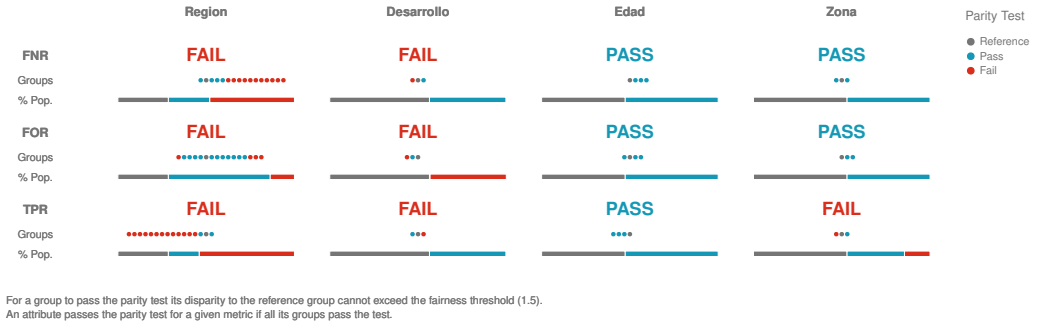


FIGURE 5.20: Model assessment for the major group.

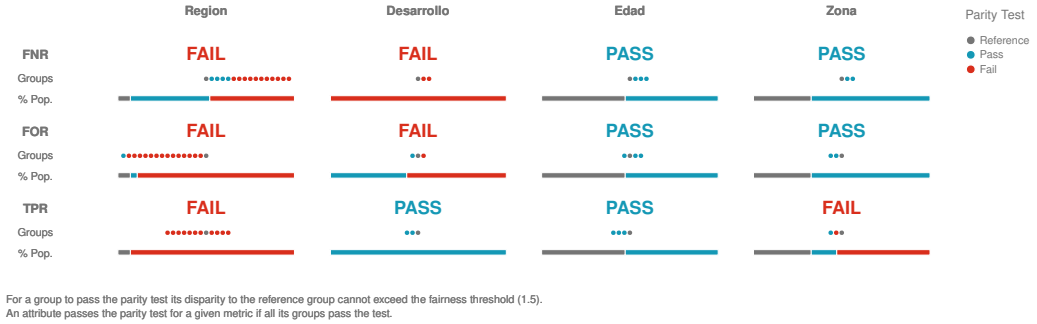


FIGURE 5.21: Model assessment for the min metrics group.

Attribute Value	**TPR**	**FNR**	**FOR**	**FDR**
Antofagasta	0.1235	0.8765	0.4201	0.3333
Arica y Parinacota	0.0806	0.9194	0.3540	0.2857
Atacama	0.2432	0.7568	0.0625	0.4706
Aysén	0.0000	1.0000	0.2700	1.0000
Bio Bío	0.3642	0.6358	0.1750	0.2903
Coquimbo	0.0890	0.9110	0.2180	0.5854
La Araucanía	0.1256	0.8744	0.2815	0.1905
Libertador Bernardo O'Higgins	0.0841	0.9159	0.1943	0.2800
Los Lagos	0.4921	0.5079	0.4945	0.4312
Los Ríos	0.2642	0.7358	0.1674	0.2632
Magallanes y Antártica Chilena	0.0294	0.9706	0.2538	0.8333
Maule	0.0735	0.9265	0.2190	0.1176
Metropolitana	0.4225	0.5775	0.1952	0.2644
Tarapacá	0.0000	1.0000	0.1728	-
Valparaíso	0.0000	1.0000	0.2411	1.0000
Ñuble	0.0000	1.0000	0.2231	-

TABLE 5.16: Bias metrics between attribute values for the base model.

Table 5.17 and Figures 5.22, 5.23 and 5.24 are the assessment results for the improved model. Table 5.18 is the crosstab for the variable "Region".

Assessment	**FNR Disparities**	**FOR Disparities**	**TPR Disparities**
Reference Group	1.001	1.075	0.999
Major Group	1.004	1.078	0.993
Min. Metrics Group	1.362	4.302	1.619

TABLE 5.17: Average disparities for the improved model.

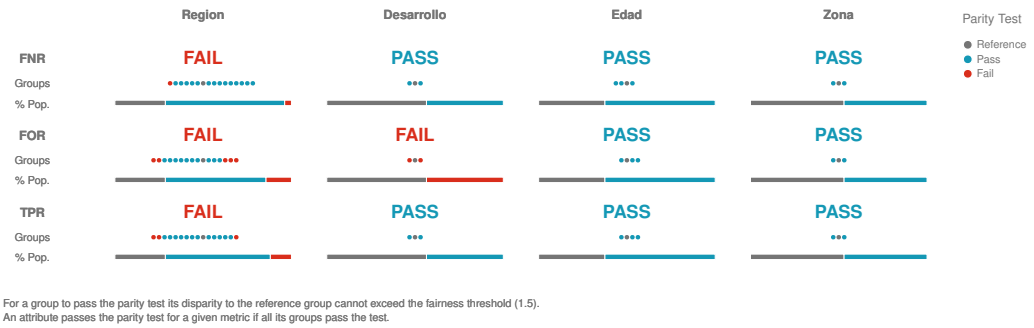


FIGURE 5.22: Model assessment for the reference group.

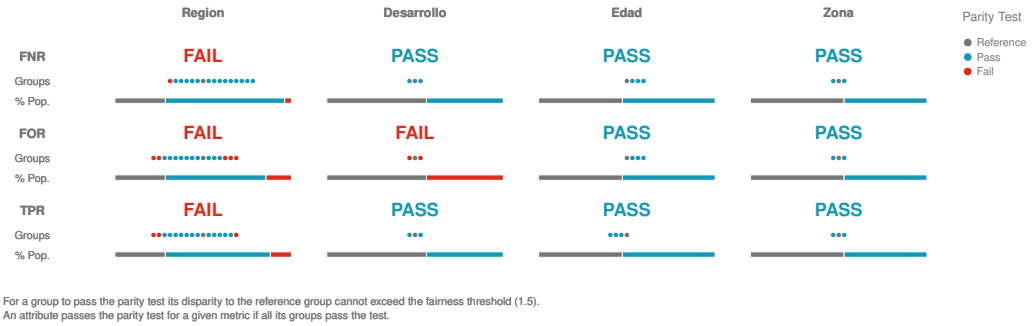


FIGURE 5.23: Model assessment for the major group.

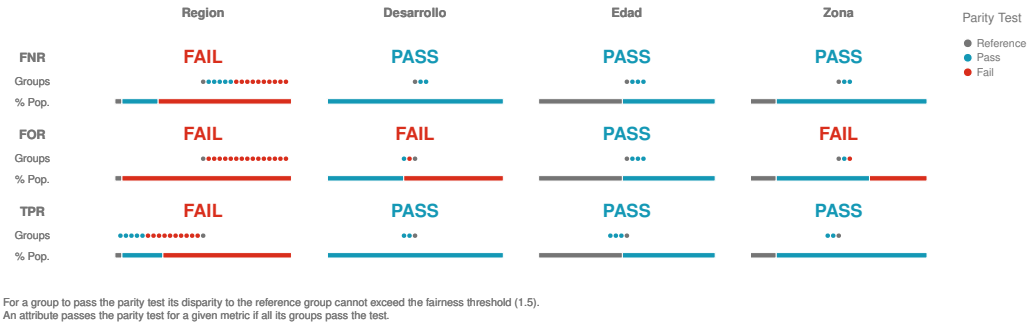


FIGURE 5.24: Model assessment for the min metrics group.

Attribute Value	**TPR**	**FNR**	**FOR**	**FDR**
Antofagasta	0.4938	0.5062	0.3154	0.2593
Arica y Parinacota	0.2581	0.7419	0.3407	0.5152
Atacama	0.5946	0.4054	0.0380	0.6857
Aysén	0.5556	0.4444	0.2000	0.6512
Bio Bío	0.4768	0.5232	0.1535	0.3543
Coquimbo	0.4136	0.5864	0.1750	0.6030
La Araucanía	0.3571	0.6429	0.2549	0.5167
Libertador Bernardo O'Higgins	0.3318	0.6682	0.1592	0.4779
Los Lagos	0.2292	0.7708	0.4964	0.3245
Los Ríos	0.3585	0.6415	0.1604	0.5250
Magallanes y Antártica Chilena	0.2059	0.7941	0.2411	0.7083
Maule	0.3039	0.6961	0.2243	0.7490
Metropolitana	0.3853	0.6147	0.2043	0.2657
Tarapacá	0.5714	0.4286	0.1091	0.6923
Valparaíso	0.2802	0.7198	0.2040	0.5526
Ñuble	0.3571	0.6429	0.1989	0.7143

TABLE 5.18: Average bias metrics between attributes for the improved model.

Chapter 6

Discussion

6.1 Drug Trafficking

The performance of the model shows clear differences between the base and improved configurations. As it can be seen in Table 5.1, the base model excels in accuracy and precision, but the recall is quite low, indicating a tendency towards negative predictions. The improved version, on the other hand, opts for a compromise, favouring a balanced distribution of false negatives and positives, although at the expense of some accuracy and precision.

Examining the base model SHAP values in Table 5.2, it becomes evident that the variables "Foreigner" and "Region" predominantly influence the model, with "lawyer" and "Age" trailing them. Particulary, when looking at Table 5.3, it can be seen that "Foreigner" has on average a negative impact towards predictions, whilst "Region" has a mainly positive one, although given its value, its impact is virtually negligible. "lawyer" and "Age" also have mainly a negligible impact on the predictions. In the improved configuration, the influence of "Region" and "lawyer" equilibrates, but their influence becomes more pronounced than in the baseline setup, as their SHAP values went up from 0.016019 and 0.006392, to 0.046972 and 0.051984, respectively – which is a 193.17% increase for the variable "Region", and a 713.23% for "lawyer". On the other hand, "Foreigner" reduces its impact by 34.15%, and now it has a positive influence towards predictions when looking at Table 5.3, as its average value is now positive. "Region" still has a net positive impact in the predictions, and its value is now 22.9 times bigger than in the base model.

Analysis of Figures 5.3 and 5.4 reveals the dual nature of the "Foreigner" variable, which could positively or negatively influence the model. In the base model, lower value regions leans towards negative predictions, a likely reflection of overrepresentation of certain regions, especially from the northern areas. The enhanced model, however, moderates this, resulting in a more balanced influence of this variable on predictions. The variable "lawyer" also shifts in influence, moving from a negligible negative impact in the base model to a still neutral but positive influence post-optimisation, which goes in hand with what was shown in Table 5.3.

Insights from Table 5.4 highlight the base model's proficiency in positive predictions but its struggle with negative ones. Specifically, its high false negative rate (FNR) stands out at 40.54%, which can be visualised in Figure 5.5, where almost half of the positive bin was actually predicted as an unfavourable outcome.

The improved version addresses this but experiences an uptick in the false positive rate (FPR) — from 0.13% to 39.25% — in contrast to a decrease in FNR, from 40.54%

down to 31.63%. This change can be seen in Figure 5.6, where the positive bin has a reduced amount of false negatives compared to the base model, but at the cost of having almost 40% of the negative bin filled with false positives. Table 5.5 complements this, as the number of true positives went up from 1594 to 1833, while false negatives were reduced from 1087 records to 848 – an aproximate 22% reduction; by contrast, the amount of false positives is roughly 304 times bigger than in the base model, as the it rose from 3 to 913 records.

In the base model, as evidenced by Table 5.6 and corresponding Figures 5.7 to 5.9, notable disparities emerge. Its performance metrics indicate consistent failure, with pronounced disparities, especially in the latter metrics. The "Foreigner" variable is a lone exception, achieving FOR parity. Focusing on the critical "Region" variable (Table 5.7), only a few regions meet the benchmark, but the average FNR per attribute is 74.86%, which means that the model tends to incorrectly predict true positives as negative depending on the region of the defendant in 74% of the cases. The average TPR is 25.13%, which means that the model also has problems in correctly identifying true positives. Finally, the FOR is 34%, meaning that of all the negative predictions made by the model, 34% were actually positive and were erroneously missed. When extending the analysis to the "Zone" variable, disparities in FNR and TPR become even more apparent, underscoring significant regional discrepancies.

Contrastingly, the enhanced model shows important improvements, as seen in Table 5.8 and its associated figures (5.10, 5.11 and 5.12). It exhibits a marked reduction in FNR disparities by an average of 44.37%, FOR disparities by 8.50%, and TPR by 94.67%. Although the model hasn't attained complete fairness or statistical parity, it represents a significant improvement. The "Region" variable, while still problematic, displays improvement as more regions clear the assessment, as its average FNR was reduced to only 39.76%, while the true positive rate was improved from 25.13% to 60.23%. The average FOR was slightly impacted, as its average value went up to 37.75%. Considering the zones, all now align with the set fairness criteria. In terms of parity, the "Foreigner" variable achieves statistical parity, and "Zone" accomplishes FOR and TPR parity.

6.2 Petty Theft

The results of this experiment, as described in Table 5.10, show a suboptimal performance for both models, due to the lack of key variables in the dataset that are essential for effective training. However, the improved model shows a clear advantage, outperforming the base model with superior recall and F1 values. While accuracy and precision show marginal fluctuations, the overall performance of the second one makes it superior.

Insights from Table 5.11 reveals that the base model predominantly relies on the number of hearings to drive decisions. The regional aspects and the development grade of the crime come secondary, whereas factors like age and the barrister contributes the less. Table 5.12 shows that both "Hearings" and "lawyer" have a negative influence towards predictions, as its average SHAP Value is negative. The improved model, according to Table 5.11, displays subtle alterations, with the "Development" variable slightly edging over "Region", as its impact is now 1.12 times bigger than in the base model, while the second variable actually reduced its effect by 2.02%. However, the amount of hearings is still the dominant variable, having a negligible change compared to the base model (a reduction of 0.79%). When looking

at Table 5.12, "Region" now has a more neutral effect, as its SHAP value was reduced by 54.29% (0.023519 to 0.010753), while, on the other hand, "Hearings" is now 1.13 times more influential towards negative predictions. The overall results shown in this table tells us that the optimisations made the variables have a more "negative" effect on predictions, as every feature, excepting "Development", had a reduction or maintained a negative SHAP value.

Figures 5.15 and 5.16 basically confirms what was discussed previously, although it can be seen that "Hearings" had plenty of cases with big positive effect on the predictions, this makes sense as the minimum SHAP value for this variable was -0.455377 for the base model and -0.464773 for the enhanced one, while the maximum value was 1.438100 and 1.414085, respectively. It can also be seen that an increased number of hearings tends to enhance predictive results. As for the variable "Region", there is no visible tendency or skew towards a kind of prediction, independently of the region.

Table 5.13 shows similarities between the situations observed in this and the prior experiment. The base model has problems with false negatives – as its FNR is at 73.20% –, predicting many outcomes as negatives that should have been positives. This manifests in the model's high recall, and it can be seen in Figure 5.17, where the positive bin has most of its predictions labeled as an unfavourable outcome. The improved model, while attempting to rectify these issues, still struggles with accurate predictions, likely due to the dataset's absence of crucial features needed for training. In this case, the FNR went down to 64.16%, as Table 5.14 indicates that the number of false negatives were reduced from 2398 to 2102 records (a 12.34% reduction). The false positive rate, on the other hand, went up from 4.96% in the base model, to 12.21% in the second one, as the amount of false positives rose from 432 to 1174 predictions (i.e. almost tripled). Both changes in FNR and FPR can be visualised in Figure 5.18, where in the positive bin, the number of false negatives was slightly reduced (but is still predominant), whilst the same thing happens in the negative one.

A comparative analysis using the table 5.15 and related figures 5.19 to 5.21 suggests that the metrics of the base model are a bit better than in the previous experiment. Despite this, it still registers failure on most tests. However, "Zone" and "Age" achieve FNR parity, and the former also achieves TPR parity. The protected variable "Region" underperforms in many areas. When looking at Table 5.16, it can be seen that the FNR between attribute values is also high, as its average is at 85%, meaning that more than 8 out of 10 predictions that were classified as negatives were actually positive, denoting big accuracy problems in between regions. TPR is also low, as it is on average at 14.05%, meaning that the model also has problems when classifying positives. The false omission rate, however, stayed relatively low with an average of 24.64%.

With the enhanced model, disparities do improve, but the magnitude of the change is not as profound as in the previous experiment. Specifically, according to Table 5.17 and Figures 5.22 to 5.24, FNR disparities were decreased by an average of 25.40%, while FOR disparities decreased by 7.04%. However, TPR disparities actually increased by 11.71%. The model is still unable to attain overall fairness or statistical parity. Nevertheless, every variable, with the exception of "Region", aligns with FNR parity. "Zone" and "Age" are the sole achievers of TPR parity, with "Age" being the only one that achieved FOR parity. Despite improvements, "Region" does

not have any parity alignment, even though its average bias metrics between attributes were also improved overall, as the FNR decreased from 85% to 61.41%, its TPR went up to 38.58%, and its FOR went down to 21.72%, therefore aligning with the statement made previously: the model is not only fairer, but it is also more accurate than the first one.

Chapter 7

Conclusion & Future Work

The improvements made using AI Fariness 360 and Fairlearn were mainly positive, as both experiments improved their overall fairness. Biases and disparities by attribute were reduced, although these changes were not sufficient to achieve statistical parity and/or fairness.

The changes also improved the parity levels between our protected variable: Region, which means that the inequalities between the different regions of the country were reduced, thus minimising potential biases.

Most importantly, these changes did not have an overall negative impact on model performance. In the first experiment we obtained a slight reduction in both accuracy and precision, but in the second experiment the model actually improved in terms of performance.

But this prompts the question: to what extent do the models contribute to addressing these biases and inequalities? The response is likely not much in comparison to the data that is used for training. This is the primary conclusion drawn from these two experiments: the level of fairness depends more on the data itself than on the trained model. Essentially, as the model learns from the data patterns, it is impossible to train something that disregards these issues.

The work done by this and other frameworks is mainly to mitigate these problems, not to avoid them altogether. If something like this is wanted or needed, it can be done in two ways: one is to adapt the way in which data that is already stored in a database is collected. This mainly involves creating more complex queries that already take these ethical aspects into account, although this is usually impractical. On the other hand, it can be beneficial to utilise data cleansing, quality checks, and data augmentation methods to minimise ethical concerns. One such approach would be to implement sampling techniques or generate synthetic data (e.g. SMOTE), although the obvious problem with this is that it can introduce noise into the model.

However, this is just one aspect of the issue. Quantity is crucial, as it is usually needed a dataset with a substantial number of entries for training, as well as having multiple variables that can potentially be utilised for these purposes, including those for doing some feature engineering. In this instance, we lacked features that may have been advantageous for improving the prediction accuracy of the models in both experiments. In particular, the analysis lacks critical variables, including the accused's sex, real age, previous cases, and other relevant case information such as days, expert reports, and degree of participation.

Ultimately, while some improvement techniques can enhance models, data remains the primary factor in developing a responsible and impartial model. Thus, altering hyperparameters or adjusting prediction thresholds may prove insufficient when the available data is lacking, as has been demonstrated by these experiments.

Finally, as we continue to develop and improve our methodology for identifying and addressing biases in machine learning models, a number of possibilities emerge for future investigation and enhancement of our current framework.

1. **Additional Analysis**: Due to the fact that we had a lack of variables to review and train, there are plenty of analyses that could be made in the future with more data available. One of such examinations that could be made is a counterfactual evaluation – a prediction that mirrors or is similar to a specific data point but with an opposite outcome. This can be made with What-If-Tool, and it could help to elucidate further biases present in the models, as well as to compare the baseline and improved versions with specific points.
2. **Optimisation Techniques in Model Training**: There is potential in using regularization methods that take into account protected variables [44]. By incorporating fairness considerations into the optimisation process, it is expected that models will be able to be trained to intrinsically decrease bias and more accurately represent all aspects of the data. This would have a twofold benefit, improving prediction precision while also guaranteeing that the model does not unintentionally continue societal biases.
3. **Data Disparity Sampling Techniques**: Data forms the fundamental basis upon which Machine Learning models are built. However, discrepancies in data may create a notable source of prejudice, particularly when certain groups are underrepresented. To overcome this hurdle, we could examine and utilise advanced sampling techniques customised to the distinct needs of each use case. By guaranteeing that the training data is a more representative sample of the wider population, we can create models that are both more precise and impartial. These techniques may vary from oversampling underrepresented groups to more advanced methods of creating synthetic data, guaranteeing that models have enough data to learn from even if the historical datasets are biased.

Bibliography

[1] Markus Anderljung, Joslyn Barnhart, Anton Korinek, Jade Leung, Cullen O'Keefe, Jess Whittlestone, Shahar Avin, Miles Brundage, Justin Bullock, Duncan Cass-Beggs, Ben Chang, Tantum Collins, Tim Fist, Gillian Hadfield, Alan Hayes, Lewis Ho, Sara Hooker, Eric Horvitz, Noam Kolt, Jonas Schuett, Yonadav Shavit, Divya Siddarth, Robert Trager, and Kevin Wolf. Frontier ai regulation: Managing emerging risks to public safety, 2023.

[2] Daniel Arias-Garzón, Reinel Tabares-Soto, Joshua Bernal-Salcedo, and Gonzalo A. Ruz. Biases associated with database structure for covid-19 detection in x-ray images. *Scientific Reports*, 13(1):3477, Mar 2023.

[3] Jacqui Ayling and Adriane Chapman. Putting ai ethics to work: are the tools fit for purpose? *AI and Ethics*, 2(3):405–429, Aug 2022.

[4] Benjamin Baron and Mirco Musolesi. Interpretable machine learning for privacy-preserving pervasive systems, 2019.

[5] Rachel K. E. Bellamy, Kuntal Dey, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, Seema Nagar, Karthikeyan Natesan Ramamurthy, John Richards, Diptikalyan Saha, Prasanna Sattigeri, Moninder Singh, Kush R. Varshney, and Yunfeng Zhang. AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias, October 2018.

[6] Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. Fairness in criminal justice risk assessments: The state of the art, 2017.

[7] Adrien Bibal, Michael Lognoul, Alexandre de Streel, and Benoît Frénay. Legal requirements on explainability in machine learning. *Artificial Intelligence and Law*, 29(2):149–169, Jun 2021.

[8] Emily Black, Hadi Elzayn, Alexandra Chouldechova, Jacob Goldin, and Daniel Ho. Algorithmic fairness and vertical equity: Income fairness with irs tax audit models. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 1479–1503, New York, NY, USA, 2022. Association for Computing Machinery.

[9] Paula Boddington. *Towards a Code of Ethics for Artificial Intelligence*. Springer, Nov 2017.

[10] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, Oct 2001.

[11] Irene Y. Chen, Emma Pierson, Sherri Rose, Shalmali Joshi, Kadija Ferryman, and Marzyeh Ghassemi. Ethical machine learning in healthcare. *Annual Review of Biomedical Data Science*, 4(1):123–144, Jul 2021.

[12] Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments, 2016.

[13] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, Sep 1995.

[14] Council of European Union. Council regulation (EU) no 52021pc0206, 2021. `https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206`.

[15] Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. Racial bias in hate speech and abusive language detection datasets, 2019.

[16] Observatorio Nacional de Drogas. Décimo cuarto estudio nacional de drogas en población general de chile, 2020. URL: https://www.senda.gob.cl/wp-content/uploads/2022/01/Estudio-PG2020.pdf.

[17] Centro de Estudios y Análisis del Delito. Portal cead, estadísticas delictuales. URL: https://cead.spd.gov.cl/estadisticas-delictuales/.

[18] Finale Doshi-Velez, Mason Kortz, Ryan Budish, Chris Bavitz, Sam Gershman, David O'Brien, Stuart Schieber, James Waldo, David Weinberger, and Alexandra Wood. Accountability of AI under the law: The role of explanation. *CoRR*, abs/1711.01134, 2017.

[19] Yi Feng, Chuanyi Li, Jidong Ge, Bin Luo, and Vincent Ng. Recommending statutes: A portable method based on neural networks. *ACM Trans. Knowl. Discov. Data*, 15(2), jan 2021.

[20] Jessica Fjeld, Nele Achten, Hannah Hilligoss, Adam Nagy, and Madhulika Srikumar. Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for ai. *Berkman Klein Center Research Publication*, 2020(1), January 2020.

[21] Luciano Floridi, Josh Cowls, Monica Beltrametti, Raja Chatila, Patrice Chazerand, Virginia Dignum, Christoph Luetge, Robert Madelin, Ugo Pagallo, Francesca Rossi, Burkhard Schafer, Peggy Valcke, and Effy Vayena. Ai4people—an ethical framework for a good ai society: Opportunities, risks, principles, and recommendations. *Minds and Machines*, 28(4):689–707, Dec 2018.

[22] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna M. Wallach, Hal Daumé III, and Kate Crawford. Datasheets for datasets. *CoRR*, abs/1803.09010, 2018.

[23] Ornella Beretta Gutiérrez. Metodología para la evaluación de modelos regularizados bajo condiciones de fairness. Master's thesis, Universidad Adolfo Ibáñez, January 2023.

[24] Thilo Hagendorff. The ethics of ai ethics: An evaluation of guidelines. *Minds and Machines*, 30(1):99–120, Mar 2020.

[25] Shibo Hao, Yi Gu, Haodi Ma, Joshua Jiahua Hong, Zhen Wang, Daisy Zhe Wang, and Zhiting Hu. Reasoning with language model is planning with world model, 2023.

[26] Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van

Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. Array programming with NumPy. *Nature*, 585(7825):357–362, September 2020.

[27] Amy K. Heger, Liz B. Marquis, Mihaela Vorvoreanu, Hanna Wallach, and Jennifer Wortman Vaughan. Understanding machine learning practitioners' data documentation perceptions, needs, challenges, and desiderata, 2022.

[28] Tin Kam Ho. Random decision forests. In *Proceedings of 3rd International Conference on Document Analysis and Recognition*, volume 1, pages 278–282 vol.1, 1995.

[29] J. D. Hunter. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3):90–95, 2007.

[30] Anna Jobin, Marcello Ienca, and Effy Vayena. The global landscape of ai ethics guidelines. *Nature Machine Intelligence*, 1(9):389–399, Sep 2019.

[31] Suhong Kim, Param Joshi, Parminder Singh Kalsi, and Pooya Taheri. Crime analysis through machine learning. In *2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*, pages 415–420, 2018.

[32] Jon M. Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. *CoRR*, abs/1609.05807, 2016.

[33] Jiajing Li, Guoying Zhang, Longxue Yu, and Tao Meng. Research and design on cognitive computing framework for predicting judicial decisions. *Journal of Signal Processing Systems*, 91(10):1159–1167, Oct 2019.

[34] Marcio Salles Melo Lima and Dursun Delen. Predicting and explaining corruption across countries: A machine learning approach. *Government Information Quarterly*, 37(1):101407, 2020.

[35] Samuele Lo Piano. Ethical principles in machine learning and artificial intelligence: cases from the field and possible ways forward. *Humanities and Social Sciences Communications*, 7(1):9, Jun 2020.

[36] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc., 2017.

[37] Tarek Mahfouz and Amr Kandil. Litigation outcome prediction of differing site condition disputes through machine learning models. *Journal of Computing in Civil Engineering*, 26(3):298–308, 2012.

[38] Jane Mitchell, Simon Mitchell, and Cliff Mitchell. Machine learning for determining accurate outcomes in criminal trials. *Law, Probability and Risk*, 19(1):43–65, 03 2020.

[39] 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. *CoRR*, abs/1810.03993, 2018.

[40] Brent Mittelstadt, Chris Russell, and Sandra Wachter. Explaining explanations in AI. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. ACM, jan 2019.

[41] Jessica Morley, Luciano Floridi, Libby Kinsey, and Anat Elhalal. From what to how: An initial review of publicly available ai ethics tools, methods and research to translate principles into practices. *Science and Engineering Ethics*, 26(4):2141–2168, Aug 2020.

[42] The pandas development team. pandas-dev/pandas: Pandas, February 2020.

[43] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[44] Flavien Prost, Hai Qian, Qiuwen Chen, Ed H. Chi, Jilin Chen, and Alex Beutel. Toward a better trade-off between performance and fairness with kernel-based distribution matching. *CoRR*, abs/1910.11779, 2019.

[45] Alvin Rajkomar, Michaela Hardt, Michael D. Howell, Greg Corrado, and Marshall H. Chin. Ensuring fairness in machine learning to advance health equity. *Annals of Internal Medicine*, 169(12):866–872, Dec 2018.

[46] Anchal Rani and S Rajasree. Crime trend analysis and prediction using mahanolobis distance and dynamic time warping technique. *Int J Comput Sci Inf Technol*, 5(3):4131–4135, 2014.

[47] Benedek Rozemberczki, Lauren Watson, Péter Bayer, Hao-Tsung Yang, Olivér Kiss, Sebastian Nilsson, and Rik Sarkar. The shapley value in machine learning, 2022.

[48] Anneleen Rummens, Wim Hardyns, and Lieven Pauwels. The use of predictive analysis in spatiotemporal crime forecasting: Building and testing a model in an urban context. *Applied Geography*, 86:255–261, 2017.

[49] Pedro Saleiro, Benedict Kuester, Abby Stevens, Ari Anisfeld, Loren Hinkson, Jesse London, and Rayid Ghani. Aequitas: A bias and fairness audit toolkit. *CoRR*, abs/1811.05577, 2018.

[50] Neil Shah, Nandish Bhagat, and Manan Shah. Crime forecasting: a machine learning and computer vision approach to crime prediction and prevention. *Visual Computing for Industry, Biomedicine, and Art*, 4(1):9, Apr 2021.

[51] Ashudeep Singh and Thorsten Joachims. Fairness of exposure in rankings. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery Data Mining*, KDD '18, page 2219–2228, New York, NY, USA, 2018. Association for Computing Machinery.

[52] Harry Surden. Machine learning and law. *Wash. L. Rev.*, 89:87, 2014.

[53] Reinel Tabares-Soto, Joshua Bernal-Salcedo, Zergio Nicolás García-Arias, Ricardo Ortega-Bolaños, María Paz Hermosilla, Harold Brayan Arteaga-Arteaga, and Gonzalo A. Ruz. *Analysis of Ethical Development for Public Policies in the Acquisition of AI-Based Systems*, pages 184–212. Exploring Ethical Problems in Today's Technological World. IGI Global, Hershey, PA, USA, 2022.

[54] Guido Van Rossum and Fred L. Drake. *Python 3 Reference Manual*. CreateSpace, Scotts Valley, CA, 2009.

[55] Effy Vayena, Alessandro Blasimme, and I. Glenn Cohen. Machine learning in medicine: Addressing ethical challenges. *PLOS Medicine*, 15(11):e1002689, Nov 2018.

[56] Hilde Weerts, Miroslav Dudík, Richard Edgar, Adrin Jalali, Roman Lutz, and Michael Madaio. Fairlearn: Assessing and improving fairness of ai systems, 2023.

[57] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2023.

[58] James Wexler. The what-if tool: Code-free probing of machine learning models, Sep 2018.

[59] Michael Wiegand, Josef Ruppenhofer, and Thomas Kleinbauer. Detection of Abusive Language: the Problem of Biased Datasets. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 602–608, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[60] Ke Xu, Hangyu Liu, Fang Wang, and Hansheng Wang. 'This Crime is Not That Crime'—Classification and evaluation of four common crimes. *Law, Probability and Risk*, 20(3):135–152, 07 2022.

[61] Hai Ye, Xin Jiang, Zhunchen Luo, and Wenhan Chao. Interpretable charge predictions for criminal cases: Learning to generate court views from fact descriptions. *CoRR*, abs/1802.08504, 2018.

[62] Han Yu, Zhiqi Shen, Chunyan Miao, Cyril Leung, Victor R. Lesser, and Qiang Yang. Building ethics into artificial intelligence. *CoRR*, abs/1812.02953, 2018.

[63] Yi Zeng, Enmeng Lu, and Cunqing Huangfu. Linking artificial intelligence principles, 2018.

[64] Haoxi Zhong, Zhipeng Guo, Cunchao Tu, Chaojun Xiao, Zhiyuan Liu, and Maosong Sun. Legal judgment prediction via topological learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3540–3549, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

Appendix A

Model Card for the Drug Traffiking Experiment

A.1 Text content from Figure A.6:

Model Details:

- **Context**: Developed by Nelson Salazar as part of the "Algoritmos Éticos" project made by the GobLab of the University Adolfo Ibañez, and in particular, has been developed within the project made for the Public Criminal Defense Office (Defensoría Penal Pública, DPP).
- **Model Objectives**: Predict the outcome for defendants associated with drug trafficking crimes.
- **Version**: 2.0, october 2023.
- **Type**: Classification. In this case, to predict if a person has a favourable outcome (1) or not (0).
- **Architecture**: Multi-Layer Perceptron (MLP), developed with Tensorflow, and trained to minimise the binary crossentropy.
- **Hyperparameters**:
  - 70 training epochs.
  - Learning rate: 0.01
  - Batch size: 1024.
  - 4 - One input layer, two hidden layers and one output layer.
  - Neuron configuration: 6, 14, 12, 1.
  - Activation functions: ReLU for hidden layers, Sigmoid for output layer.

Intended Use:

- Developed for the DPP for testing purposes.
- It can be used to audit and control the processes carried out by the DPP's lawyers, in order to improve the quality of the service provided by them.
- It can be used to support decision-making in relation to individual cases.

Factors:

- Due to the nature of the model, among the relevant factors of the model, we find thoses variables that are related to the defendant: age, whether the defendant is a foreigner or not, and the geographical region in which they are being prosecuted.
- Other factors includes those related to the defendant's case, such as the degree of development of the crime, the amount of hearings, and the barrister that is defending them.

Metrics:

- **Model Performance**: Accuracy (67%), F1 (70%), Precision (67%), Recall (72%).
- **Bias Metrics**: False Positive Rate (FPR, 39.25%), False Negative Rate (FNR, 31.63%), False Discovery Rate (FDR, 33.25%) and False Omission Rate (FOR, 37.51%)
- Fairness: Does it meet Statistical Parity criteria? – No; Equitable Model? – No.
- **Average Disparities between attributes**:
  - FNR: 1.41x
  - FOR: 1.43x
  - TPR: 1.25x

Training and Testing Data:

- **Dataset**: Tráfico de Drogas 2017 - 2022.
- **Source**: Public Criminal Defense Office (Defensoría Penal Pública, DPP).
- **Format**: CSV/XLSX.
- **Details**: 16,690 non-null rows.
- **Training set**: 70% split. (10,683 rows).
- **Validation**: 1000 rows deducted from the training set.
- **Testing set**: 30% split (5,007 rows).

Ethical Considerations:

- The model does not consider the use of sensitive data of the defendant, such as name, RUT (ID) and/or the case-accused ID.
- It was trained with ethical considerations in mind. This includes using the geographical region of the defendant as a protected variable.
- Other considerations:
  - Sample weights optimised to ensure a fairer distribution of data across different classes in the training process.
  - Class weights were also balanced.
  - The prediction threshold was also adjusted to prioritise FNR parity, as we are interested in minimising the amount of people that has a non-favourable ending to its trial.

Caveats and Recommendations:

- Since this model has been trained with ethical considerations in mind, model performance is not the most optimal.
- One relevant group that was not considered in the model – because it was not in the dataset – is the sex of the accused. There is the risk that some biases may exist in this area, so this should be added in future training.
- Another risk: it could be used by the barristers as a way to predict the outcome of a trial, to then see whether or not it is appropiate to take a defendant.

Quantitative Analysis:

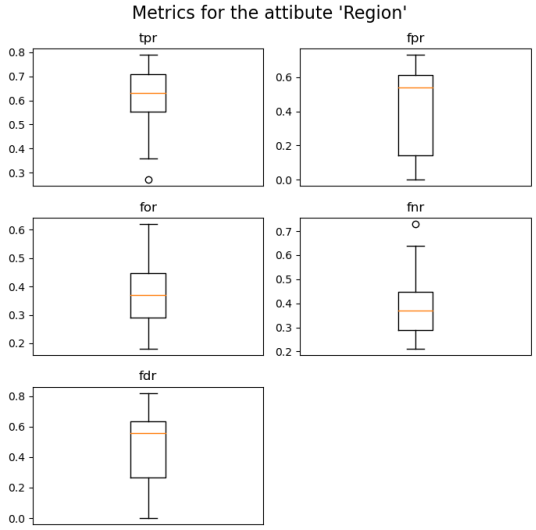


FIGURE A.1: Bias metrics for the attribute "Region".

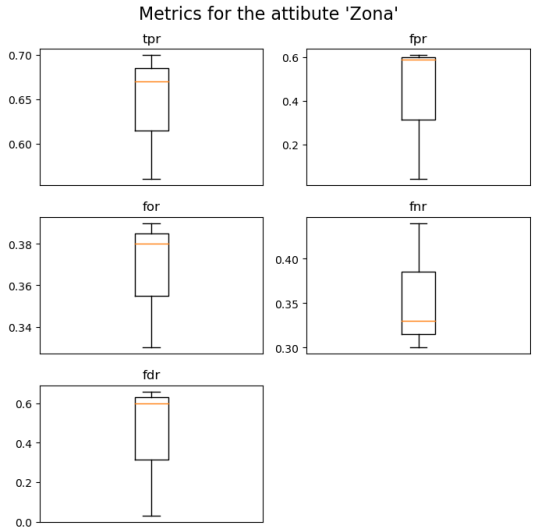


FIGURE A.2: Bias metrics for the attribute "Zona".

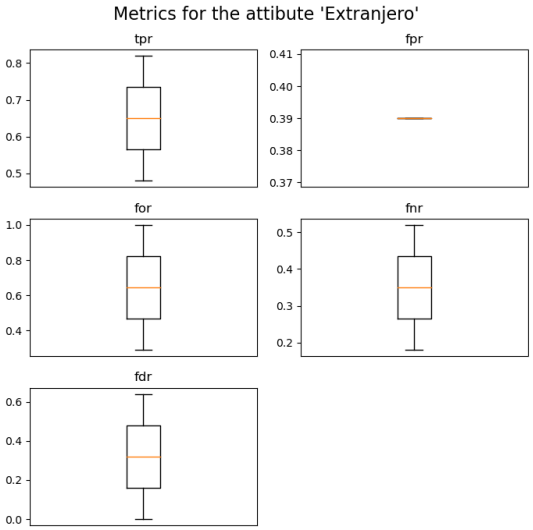


FIGURE A.3: Bias metrics for the attribute ”Extranjero”.

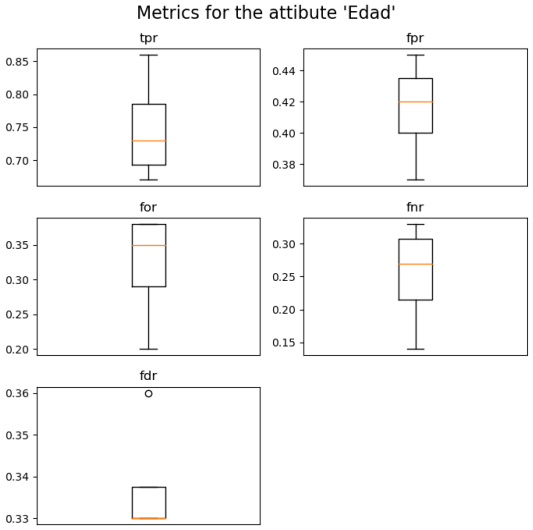


FIGURE A.4: Bias metrics for the attribute ”Edad”.

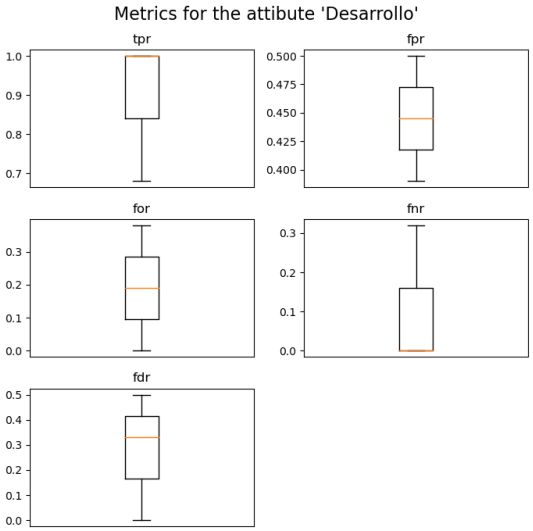


FIGURE A.5: Bias metrics for the attribute ”Desarrollo”.

UNIVERSIDAD ADOLFO IBÁÑEZ

ALGORITMOS ÉTICOS | RESPONSABLES Y TRANSPARENTES

Model Card: Drug Traffiking

Model Details

- **Context:** Developed by Nelson Salazar as part of the "Algoritmos Éticos" project made by the GobLab of the University Adolfo Ibañez, and in particular, has been developed within the project made for the Public Criminal Defense Office (Defensoría Penal Pública, DPP).
- **Model Objectives:** Predict the outcome for defendants associated with drug trafficking crimes.
- **Version:** 2.0, october 2023.
- **Type:** Classification. In this case, to predict if a person has a favourable outcome (1) or not (0).
- **Architecture:** Multi-Layer Perceptron (MLP), developed with Tensorflow, and trained to minimise the binary crossentropy.
- **Hyperparameters**:
  - 70 training epochs.
  - Learning rate: 0.01
  - Batch size: 1024.
  - Layers: 4 - One input layer, two hidden layers and one output layer.
  - Neuron configuration: 6, 14, 12, 1.
  - Activation functions: ReLU for hidden layers, Sigmoid for output layer.

Intended Use

- Developed for the DPP for testing purposes.
- It can be used to audit and control the processes carried out by the DPP's lawyers, in order to improve the quality of the service provided by them.
- It can be used to support decision-making in relation to individual cases.

Factors

- Due to the nature of the model, among the relevant factors of the model, we find thoses variables that are related to the defendant: age, whether the defendant is a foreigner or not, and the geographical region in which they are being prosecuted.
- Other factors includes those related to the defendant's case, such as the degree of development of the crime, the amount of hearings, and the barrister that is defending them.

Metrics

- **Model Performance**: Accuracy (67%), F1 (70%), Precision (67%), Recall (72%).
- **Bias Metrics**: False Positive Rate (FPR, 39.25%), False Negative Rate (FNR, 31.63%), False Discovery Rate (FDR, 33.25%) and False Omission Rate (FOR, 37.51%)
- **Fairness**: Does it meet Statistical Parity criteria? -- No; Equitable Model? -- No.
- **Average Disparities between attributes**:
  - FNR: 1.41x
  - FOR: 1.43x
  - TPR: 1.25x

Training and Testing Data

- **Dataset:** Tráfico de Drogas 2017 - 2022.
- **Source**: Public Criminal Defense Office (Defensoría Penal Pública, DPP).
- **Format**: CSV/XLSX.
- **Details**: 16,690 non-null rows.
- **Training set**: 70% split. (10,683 rows).
- **Validation**: 1000 rows deducted from the training set.
- **Testing set**: 30% split (5,007 rows).

Ethical Considerations

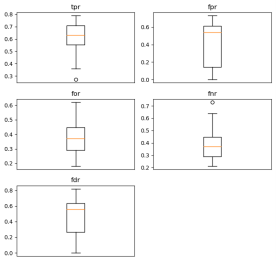
- The model does not consider the use of sensitive data of the defendant, such as name, RUT (ID) and/or the case-accused ID.
- It was trained with ethical considerations in mind. This includes using the geographical region of the defendant as a protected variable.
- Other considerations:
  - Sample weights optimised to ensure a fairer distribution of data across different classes in the training process.
  - Class weights were also balanced.
  - The prediction threshold was also adjusted to prioritise FNR parity, as we are interested in minimising the amount of people that has a non-favourable ending to its trial.

Caveats and Recommendations

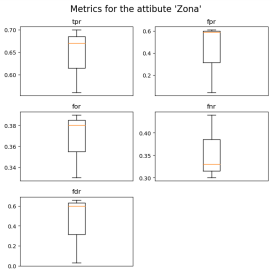
- Since this model has been trained with ethical considerations in mind, model performance is not the most optimal.
- One relevant group that was not considered in the model -- because it was not in the dataset -- is the sex of the accused. There is the risk that some biases may exist in this area, so this should be added in future training.
- Another risk: it could be used by the barristers as a way to predict the outcome of a trial, to then see whether or not it is appropiate to take a defendant.

Quantitative Analysis

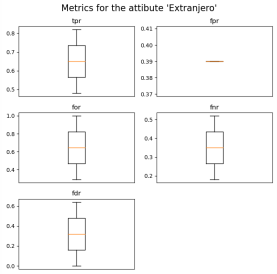
Metrics for the attibute 'Region'



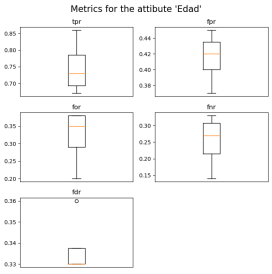
Metrics for the attibute 'Zona'



Metrics for the attibute 'Extranjero'



Metrics for the attibute 'Edad'



Metrics for the attibute 'Desarrollo'

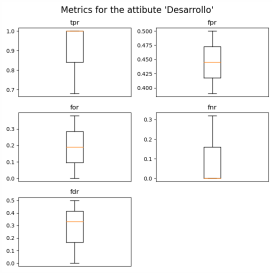


FIGURE A.6: Model Card.

Appendix B

Model Card for the Petty Theft Experiment

B.1 Text content from Figure B.5:

Model Details:

- **Context**: Developed by Nelson Salazar as part of the "Algoritmos Éticos" project made by the GobLab of the University Adolfo Ibañez, and in particular, has been developed within the project made for the Public Criminal Defense Office (Defensoría Penal Pública, DPP).
- **Model Objectives**: Predict the outcome for defendants associated with petty theft offences.
- **Version**: 2.0, october 2023.
- **Type**: Classification. In this case, to predict if a person has a favourable outcome (1) or not (0).
- **Architecture**: Light Gradient-Boosting Machine (LGBM), using the *lightgbm* library.
- **Hyperparameters**:
  - Boosting type: gbdt.
  - 100 estimators
  - 120 leaves per estimator (tree).
  - Unlimited maximum depth.
  - Learning rate: 0.01

Intended Use:

- Developed for the DPP for testing purposes.
- It can be used to audit and control the processes carried out by the DPP's lawyers, in order to improve the quality of the service provided by them.
- It can be used to support decision-making in relation to individual cases.

Factors:

- Due to the nature of the model, among the relevant factors of the model, we find thoses variables that are related to the defendant: the age and the geographical region in which they are being prosecuted.
- Other factors includes those related to the defendant's case, such as the degree of development of the crime, the amount of hearings, and the barrister that is defending them.
- In this case, we do not have a variable that indicates whether the defendant is foreigner or not.

Metrics:

- **Model Performance**: Accuracy (77%), F1 (51%), Precision (62%), Recall (43%).
- **Bias Metrics**: False Positive Rate (FPR, 12.21%), False Negative Rate (FNR, 64.16%), False Discovery Rate (FDR, 47.54%) and False Omission Rate (FOR, 21.56%).
- Fairness: Does it meet Statistical Parity criteria? – No; Equitable Model? – No.
- **Average Disparities between attributes**:
  - FNR: 1.12x
  - FOR: 2.15x
  - TPR: 1.20x

Training and Testing Data:

- **Dataset**: Hurto Falta 2017 - 2022.
- **Source**: Public Criminal Defense Office (Defensoría Penal Pública, DPP).
- **Format**: CSV/XLSX.
- **Details**: 39,954 non-null rows.
- **Training set**: 70% split. (28,967 rows).
- **Testing set**: 30% split (11,987 rows).

Ethical Considerations:

- The model does not consider the use of sensitive data of the defendant, such as name, RUT (ID) and/or the case-accused ID.
- It was trained with ethical considerations in mind. This includes using the geographical region of the defendant as a protected variable.
- Other considerations:
  - Sample weights optimised to ensure a fairer distribution of data across different classes in the training process.
  - Class weights were also balanced.
  - The prediction threshold was also adjusted to prioritise FNR parity, as we are interested in minimising the amount of people that has a non-favourable ending to its trial.

Caveats and Recommendations:

- Since this model has been trained with ethical considerations in mind, model performance is not the most optimal.
- One relevant group that was not considered in the model – because it was not in the dataset – is the sex of the accused. There is the risk that some biases may exist in this area, so this should be added in future training.
- Another risk: it could be used by the barristers as a way to predict the outcome of a trial, to then see whether or not it is appropiate to take a defendant.
- Multiple critical columns were missing in this dataset, hence the poor performance metrics. More training variables are needed in order to have a proper model for predicting these cases.

Quantitative Analysis:

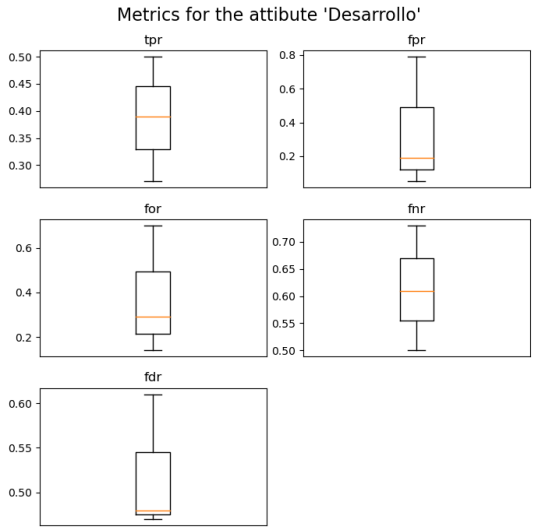


FIGURE B.1: Bias metrics for the attribute "Desarrollo".

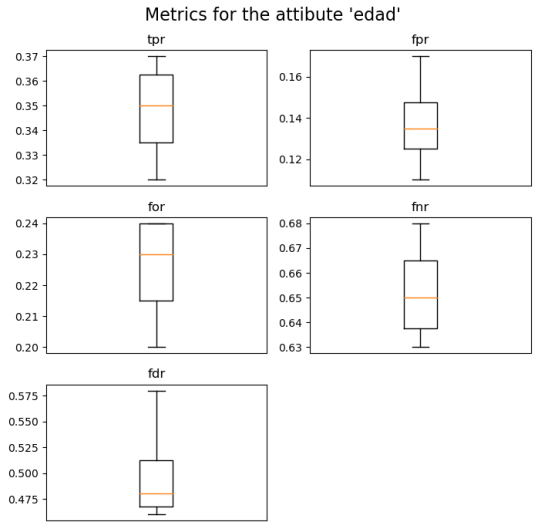


FIGURE B.2: Bias metrics for the attribute "Edad".

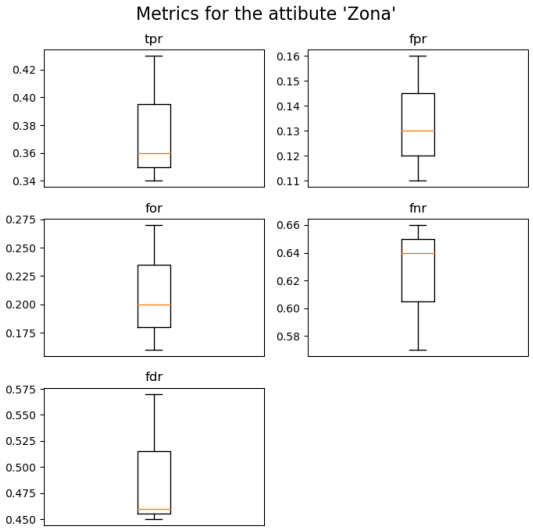


FIGURE B.3: Bias metrics for the attribute "Zona".

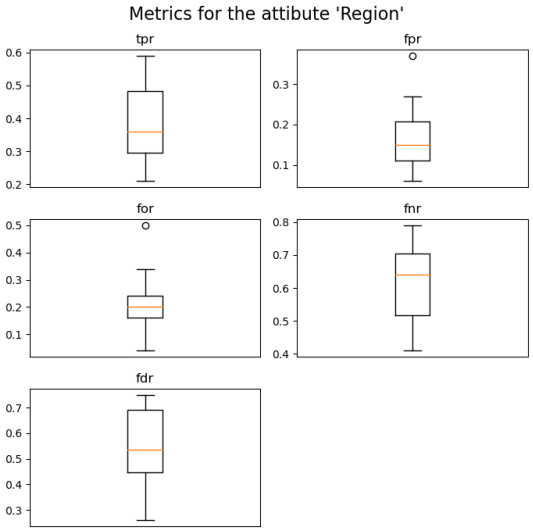


FIGURE B.4: Bias metrics for the attribute "Region".

UNIVERSIDAD ADOLFO IBÁÑEZ

ALGORITMOS ÉTICOS | RESPONSABLES Y TRANSPARENTES

Model Card: Petty Theft

Model Details

- **Context:** Developed by Nelson Salazar as part of the "Algoritmos Éticos" project made by the GobLab of the University Adolfo Ibañez, and in particular, has been developed within the project made for the Public Criminal Defense Office (Defensoría Penal Pública, DPP).
- **Model Objectives:** Predict the outcome for defendants associated with petty theft offences.
- **Version:** 2.0, october 2023.
- **Type:** Classification. In this case, to predict if a person has a favourable outcome (1) or not (0).
- **Architecture:** Light Gradient-Boosting Machine (LGBM), using the *lightgbm* library.
- **Hyperparameters**:
  - Boosting type: gbdt.
  - 100 estimators
  - 120 leaves per estimator (tree).
  - Unlimited maximum depth.
  - Learning rate: 0.01

Intended Use

- Developed for the DPP for testing purposes.
- It can be used to audit and control the processes carried out by the DPP's lawyers, in order to improve the quality of the service provided by them.
- It can be used to support decision-making in relation to individual cases.

Factors

- Due to the nature of the model, among the relevant factors of the model, we find thoses variables that are related to the defendant: the age and the geographical region in which they are being prosecuted.
- Other factors includes those related to the defendant's case, such as the degree of development of the crime, the amount of hearings, and the barrister that is defending them.
- In this case, we do not have a variable that indicates whether the defendant is foreigner or not.

Metrics

- **Model Performance**: Accuracy (77%), F1 (51%), Precision (62%), Recall (43%).
- **Bias Metrics**: False Positive Rate (FPR, 12.21%), False Negative Rate (FNR, 64.16%), False Discovery Rate (FDR, 47.54%) and False Omission Rate (FOR, 21.56%)
- **Fairness**: Does it meet Statistical Parity criteria? -- No; Equitable Model? -- No.
- **Average Disparities between attributes**:
  - FNR: 1.12x
  - FOR: 2.15x
  - TPR: 1.20x

Training and Testing Data

- **Dataset:** Hurto Falta 2017 - 2022.
- **Source**: Public Criminal Defense Office (Defensoría Penal Pública, DPP).
- **Format**: CSV/XLSX.
- **Details**: 39,954 non-null rows.
- **Training set**: 70% split. (28,967 rows).
- **Testing set**: 30% split (11,987 rows).

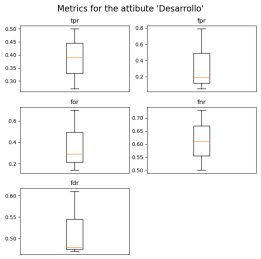
Ethical Considerations

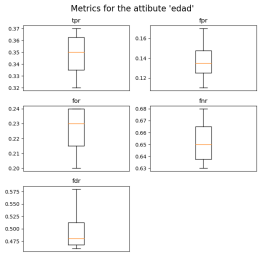
- The model does not consider the use of sensitive data of the defendant, such as name, RUT (ID) and/or the case-accused ID.
- It was trained with ethical considerations in mind. This includes using the geographical region of the defendant as a protected variable.
- Other considerations:
  - Sample weights optimised to ensure a fairer distribution of data across different classes in the training process.
  - Class weights were also balanced.
  - The prediction threshold was also adjusted to prioritise FNR parity, as we are interested in minimising the amount of people that has a non-favourable ending to its trial.

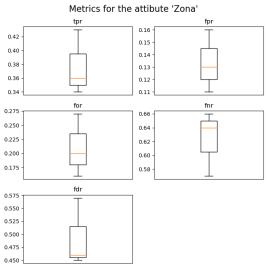
Caveats and Recommendations

- Since this model has been trained with ethical considerations in mind, model performance is not the most optimal.
- One relevant group that was not considered in the model -- because it was not in the dataset -- is the sex of the accused. There is the risk that some biases may exist in this area, so this should be added in future training.
- Another risk: it could be used by the barristers as a way to predict the outcome of a trial, to then see whether or not it is appropiate to take a defendant.
- Multiple critical columns were missing in this dataset, hence the poor performance metrics. More training variables are needed in order to have a proper model for predicting these cases.

Quantitative Analysis







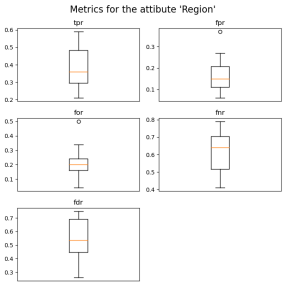


FIGURE B.5: Model Card.