POLITECNICO DI TORINO

Master's Degree
in Mathematical Engineering

Master's Degree Thesis

A Synthetic Data Generation Approach for Subgroup-Based Bias Mitigation in Structured Data

Supervisors
Prof.ssa Eliana Pastor
Prof. Flavio Giobergia

Candidate
Maria Antonietta Longo

Academic Year 2024-2025

To my family and to Filippo.

Summary

Nowadays, it is increasingly common to entrust decisions to Artificial Intelligence through Machine Learning algorithms, especially in fields such as medical diagnosis, social networks, smart cities, and finance.

Since these decisions directly impact people, it is essential to assess their reliability and trustworthiness. Accuracy provides an indication of a model's performance but is insufficient to determine how much one can truly rely on its predictions. A key issue is that models depend on data, which is often unevenly represented, potentially leading to unfair predictions that disproportionately affect smaller or less represented populations. This phenomenon, known as *Representation Bias*, arises when the sample used for model development does not adequately capture certain segments of the population, resulting in poor generalization for those groups.

When a model systematically misclassifies specific feature value pairs, problematic subgroups, it exhibits bias against the affected populations. Existing bias mitigation methods for tabular data often require prior knowledge of biases rather than identifying them automatically, which may be limiting when misclassifications stem from complex social contexts. Additionally, some approaches rely on a held-out dataset, which is not always available.

This thesis proposes a new model-agnostic bias mitigation method for tabular data, which uses an algorithm for the automatic identification of problematic subgroups and generates new representative data using an interpolation model. This improves model predictions for instances containing problematic subgroups and, most importantly, enhances fairness.

Acknowledgements

I would like to thank my advisor, Prof. Eliana Pastor, for guiding me in the development of this thesis with expertise and professionalism, but above all, for inspiring my passion for the topics of explainability and trustworthiness in AI. A sincere thank you also to my co-advisor, Prof. Flavio Giobergia, for helping me refine every detail and encouraging me to think critically about every aspect of this work. Thank you both for your patience and kindness toward me.

A heartfelt thanks to my family, who taught me never to take anything for granted; to my brother Giuseppe, for always being by my side and sharing with me these years away from home; and to my little sister Lucia, for bringing me joy and unconditional love. Thank you to my grandmother Porsietta, for her endless love and support, which I still feel even though she has not been with me for a year now, and to my grandfather Antonio, for always being a loving and playful presence. A special thank you to Uncle Francesco because, without him, I wouldn't be here today.

Finally, my deepest gratitude to Filippo, for always being by my side and believing in me. From the very first day of this master's program, he has brightened my days with his smile. I am also immensely grateful to my friends for their unwavering support throughout this journey.

Turin, March 2025.

Acronyms

RB	Representation Bias
BM	Bias Mitigation
ML	Machine Learning
AI	Artificial Intelligence
PnD	Partition-and-Debias
FP	False Positive
FN	False Negative
FPR	False Positive Rate
FNR	False Negative Rate
FPM	Frequent Pattern Mining
SMOTE	Syntetic Minority Over-sampling TEchnique
SMOTE-NC	Syntetic Minority Over-sampling TEchnique-Nominal Continuous
DT	Decision Tree
GB	Gradient Boosting
KNN	K-Nearest Neighbors
LG	Logistic Regression
RF	Random Forest

Contents

List of Tables

List of Figures

Chapter 1

Introduction and Motivation

This first chapter introduces the fundamental concepts related to *Representation Bias* and its possible solutions, along with a general overview of the thesis.

Specifically, Section 1.1 provides a definition of *Representation Bias*, explaining its causes and the contexts in which it can occur. Real-world examples of situations where this phenomenon has been observed are presented in 1.1.1.

Next, Section 1.2 introduces the concept of *Bias Mitigation* and presents a classification of the main bias mitigation techniques. Section 1.2.1 delves deeper into a specific approach that involves the identification of subgroups, also providing a general definition of subgroups.

Finally, Section 1.3 offers a comprehensive overview of the thesis structure, outlining the content of each chapter.

1.1 *Representation Bias* in Machine Learning

Nowadays, it is becoming increasingly common for certain decisions to be entrusted to Artificial Intelligence through Machine Learning algorithms. This is happening more frequently in areas such as medical diagnosis, social network, smart cities, or finance: more broadly, machine learning is being

applied in any field where decisions are required. Since these types of decisions directly impact people, it is essential to assess both the reliability and trustworthiness of the predictions. In particular, when decisions are based on data, the Accuracy of such predictions depends on the data used to make them. The system essentially learns the patterns that are present within the data provided. Ideally, we would like these data to accurately reflect the underlying distribution from which future production data will come. However, this alone is insufficient, as it only offers a broad view of the general performance of the model. A model that performs well overall may still be unreliable in areas where the data has sparse representation, potentially impacting smaller or less common subgroups. Such areas often correspond to minority populations or rare yet critical cases where accurate predictions are essential for important decisions. Consequently, if the data do not adequately capture all relevant groups within a population, the results of the decision-making system for those groups may be lacking reliability. This kind of bias that arises when the sample used for model development does not adequately capture certain segments of the population, leading to poor generalization for specific groups within the intended user population, is called ***Representation Bias*** [1].
In summary, Representation Bias can occur when the sampling process captures only part of the population or when the target population has shifted or differs from the population used in model training.

1.1.1 Examples of Representation Bias

To better understand how the bias issue unfolds and emerges, it can be helpful to examine examples that demonstrate the impact of bias on individuals and society. This bias can lead to actual discrimination if one relies solely on the ML model and its potentially high Accuracy, without conducting further in-depth analysis of the data.
Due to Representation Bias, data-driven ML models can learn unfair and discriminatory patterns as happens in:

COMPAS score case

Here the goal is to use AI to assist judges to make decisions.
The COMPAS score (Correctional Offender Management Profile for Alternative Sanctions) is a tool developed by the company Northpointe to assess the risk of criminal recidivism. Journalists at ProPublica analyzed data from 7,000 arrests in Broward County, Florida, made in 2013 and 2014, examining the risk scores assigned to individuals alongside their actual recidivism, defined as whether they were charged with new crimes within two years. The analysis revealed notable racial disparities: the algorithm was more likely to incorrectly label Black defendants as high risk, resulting in nearly double the False Positiverate compared to white defendants. Conversely, white defendants were often mislabeled as low risk more frequently than Black defendants [2]. The issue here is that the model is trained on data in which African-American individuals with a high risk of recidivism are overrepresented compared to the number of Caucasian individuals who reoffend. This results in a clear racial bias in the model's predictions.

Amazon recruitment case

Here the goal is to use AI for the examination of job applications for Amazon.
The AI tool designed to assist recruiters with tech field applications exhibited bias against women. Specifically, it penalized candidates from all-women's colleges because the model tended to downgrade applications that included the word "women's" [3].
The issue lies in the AI system that learns to make decisions based on historical data, which means it can reinforce existing biases. In this case, the discrimination is directed at women because the tech industry is predominantly male.

United Kingdom exam results in 2020 pandemic case

During the COVID-19 pandemic, the UK government implemented an algorithm to determine A-Level results after exams were canceled. This system relied on students' prior academic performance to predict their grades. However, the algorithm produced significant issues, particularly disadvantaging students from underrepresented and disadvantaged backgrounds. As a result, many students received lower-than-expected grades [4]. In this case,

the model leads to discrimination against students from lower-income backgrounds, as it relies heavily on historical academic performance, which may not reflect their true potential due to external factors like limited access to resources. This causes students from disadvantaged groups to be unfairly penalized.

These three real-world examples share an important commonality: in each case, the discrimination arises from the underrepresentation of certain population groups in the training data provided to the model. This lack of diverse and inclusive data leads to biased outcomes, where marginalized or minority groups are unfairly impacted by the model's decisions.

1.2 Bias Mitigation as Solution for Bias Problems

The term "***Bias Mitigation***" refers to the collection of all methods aimed at reducing, and ideally eliminating, discriminatory effects generated by a model's predictions. These techniques are designed to address and correct for biases that may arise in the decision-making process, ultimately striving to ensure fairer and more equitable outcomes across different demographic or sensitive groups.
Although the biases can emerge from both ***structured*** (tabular) data and ***unstructured*** (e.g. images, text, graph) data, this thesis will focus on the bias mitigation applied to the former.
Drawing on the concept of "trustworthy AI", bias mitigation solutions can be categorized into two types: ***model-dependent*** and ***model-agnostic***. For an ML model making decisions for humans, a certain level of fairness is expected: if we can trust the prediction and understand the model's functionality, we can assess whether it relies on sensitive or protected information or makes decisions based on discriminatory factors.
Once these biases are identified, they can be mitigated through two main strategies. The first involves modifying the model itself, such as by including constraints or by adding penalties to the loss function; in this case, the bias mitigation solution is model-dependent. The second strategy leaves the model unchanged and instead addresses the available data. This might include pre-processing steps such as re-balancing classes, performing data

augmentation, or reweighing, as well as post-processing methods that adjust model outputs rather than the model itself; in this case, the solution is model-agnostic.

1.2.1 What *Subgroup-based* Bias Mitigation means

In the case of model-agnostic bias mitigation achieved through data augmentation, additional data could be acquired externally. However, this inevitably incurs costs and requires targeted efforts to identify data that align with specific objectives to minimize expenses. Alternatively, additional data can be generated using techniques and strategies that leverage existing data. This second approach offers significant advantages, as it eliminates the costs associated with acquiring new data and avoids any ethical concerns related to data collection.

In any case, for this thesis, subgroup-based mitigation refers to a bias mitigation approach where additional data are incorporated into the training set only after identifying which specific data to include. Subgroups are defined as feature-value pairs (e.g ., age=30 and gender=female) [5] that represent specific segments of the dataset. These data are typically selected because they are associated with problematic subgroups, such as those responsible for a high number of False Positives, False Negatives, or for which the ML model tends to make more incorrect predictions.

1.3 Thesis Overview

The aim of this thesis is to propose innovative, model-agnostic bias mitigation methodologies specifically designed for structured data. The proposed techniques are based on data augmentation and do not require the acquisition of new data. Instead, the additional data are either generated from the existing dataset or drawn directly from a "reserved" subset of data that is kept separate from both the training and testing sets.

The structure of this thesis is as follows. Chapter 2 explores various bias mitigation studies applied to both structured and unstructured datasets, using techniques that either involve or exclude an initial automatic subgroup identification step. This section compares these approaches, highlighting the necessity of subgroup identification for achieving the objectives of this

thesis. Additionally, it emphasizes that the proposed data augmentation methods do not incur extra costs for acquiring new data, as they rely on augmenting or generating data from the existing dataset. Chapter 3 presents the prerequisites necessary to describe and understand the proposed solution. Chapter 4 provides the general method designed to address the bias problem, while Chapter 5 showcases the results obtained by applying these methods to the data, along with a description of the datasets. Finally and Chapter 6 contains the corresponding conclusions and future works.

Chapter 2

Related works

The Representation Bias, as previously defined and described, can be considered an intrinsic property of the dataset itself, independent of how the data will be used downstream. This type of bias is not tied to the machine learning model applied to the data: if it exists, it remains a characteristic of the dataset, regardless of the chosen model.

While the primary objective of this thesis is to propose mitigation solutions for structured data, with a particular focus on tabular datasets, analyzing techniques applicable to unstructured data can be valuable. Such methodologies may provide useful insights or encompass concepts and approaches that could be adapted and applied to tabular data as well.

To this end, Section 2.1 provides a description of the specific causes of Representation Bias, while Section 2.2 examines the techniques for addressing Representation Bias presented in various papers. Specifically, Section 2.2.1 discusses methods that require domain experts to identify potential biased categories in the data, whereas Section 2.2.2 explores techniques that do not rely on experts, as the subgroups are automatically identified by specialized algorithms. Finally, Section 2.3 offers a commentary on the related works presented in the previous section.

2.1 Possible causes of Representation Bias

To better understand what ***Rapresentation Bias*** means, let's analyze three possible reasons that cause it. In practical terms, its origin can be understood by analyzing the various types of biases that contribute to it. For example, ***Historical Bias*** [6] reflects the socio-technical issues in the

world. One example of this is the Google search for "CEO United States," where the results predominantly show images of male CEOs. This is not surprising, as, according to the latest data, women hold the position of CEO in 10.4% of Fortune 500 companies. While there has been progress toward gender equality, this percentage remains lower than the overall representation of women in the workforce. Another cause could be the ***Underlying Distribution Skew*** [6] since the data may be imbalanced across subpopulations without discriminatory intent. For example, according to the US Census Bureau [7], 7% of the American population is of Asian descent, but if a sample were collected, this subgroup would be underrepresented simply due to chance. This can lead to unintentional discrimination in some applications. Finally, another possible cause could be ***Self-Selection Bias*** [6], which occurs when a subset of the population voluntarily chooses not to participate or is unable to participate in a specific experiment. For example, if an online survey about the benefits of technology is distributed only via email, people without internet access or those who do not check their inbox frequently may be excluded, skewing the results in favor of participants with greater access to technology.

In this chapter, some of the existing techniques for mitigating this type of bias will be examined. These techniques are typically categorized into two main groups: those designed for structured data (e.g., images, text, graphs) and those developed for unstructured data (e.g., tabular data). Recall that these types can further be categorized into two groups again: those that do not involve intervening on the model, known as model-agnostic methods, and those that require modifying the model itself, referred to as model-dependent methods.

2.2 Representation-Bias Mitigation Techniques

2.2.1 Subgroups Identified by Domain Experts

Some studies propose solutions to address bias without automated subgroup identification. One such approach, applied to image data, is the Partition-and-Debias (PnD) [8] method, which handles multiple unknown biases without relying on predefined subgroups. The PnD method implicitly divides the bias space into distinct subspaces, using a set of experts specialized for each type of bias, with each expert focusing on a specific aspect of the problem. In other words, partitioning the space into subspaces allows the model to

"specialize" in handling different facets of the bias, with each expert dealing with a particular dimension, thereby improving the precision of bias mitigation. A gating module is then employed to combine the contributions of the experts, deciding which expert to activate and how to integrate their respective information to produce a final bias-free classification. The gating module acts as a "selector", directing data to the most appropriate expert for processing, and generating a final decision based on an aggregation of the experts' responses. This approach stands out for its ability to tackle scenarios where the type and extent of biases are unknown, without the need to explicitly define subgroups. As a result, the method proves particularly effective and adaptable in addressing complex biases that arise in real-world contexts, where the specific manifestations of bias in the data cannot always be predicted in advance.
Another methodology that does not require subgroup identification is FairDo, [9] a method designed to reduce bias in tabular datasets by transforming them to mitigate discrimination across multiple protected attributes such as nationality, age, and gender. Rather than removing data, FairDo uses preprocessing techniques to adjust feature distributions, ensuring fairness without explicitly defining subgroups. This approach minimizes disparities in model predictions, enhancing fairness without compromising performance, as demonstrated on real-world datasets.

Another group of bias mitigation states can be found in various synthetic data generation methods that share the common goal of improving fairness without compromising the overall performance of the model. For example, several synthetic data generation techniques have been developed and compared [10], which certainly allow data augmentation, but often do not involve careful selection of the data to be generated, as they typically focus only on balancing datasets that are biased due to class imbalance. One of the techniques presented in the discussed survey is SMOTE, which will be mathematically described in detail in the *Background* section. Unlike other works that use SMOTE to generate data, in this thesis the process will first involve searching for subgroups where the model is under-performing.

Even though the data is generated from scratch, no additional costs for acquiring external data are required. This represents a significant advantage. Another advantage is that generating new data allows for the removal of some of the causes underlying Representation Bias. Creating new data points with feature-value pairs that may be difficult to extract using traditional acquisition techniques logically helps eliminate at least three main

causes of bias already discussed: *historical bias, underlying distribution bias, and self-selection bias*, as discussed earlier in this chapter.

The selection of these studies as examples of bias mitigation techniques highlights an important limitation: bypassing automated subgroup identification relies heavily on prior knowledge of the problem and the categories potentially subject to discrimination. While this approach can be effective, it carries the risk of overlooking problematic categories that were not anticipated, leaving certain sources of bias unaddressed and potentially reducing the comprehensiveness of the mitigation strategy. Not identifying subgroups for bias mitigation simplifies implementation by avoiding the need to manually define and manage them. It also makes the approach more scalable, as it can be applied to a broader range of datasets without requiring prior knowledge of all possible subgroups. Furthermore, this approach is flexible and adaptable to situations where bias may arise from complex, unknown interactions, rather than from easily identifiable subgroups.

2.2.2 Automated Subgroup Identification

It is worth noting that methodologies have been developed which work even when subgroups are not explicitly defined. However, not searching for subgroups can lead to lower precision in addressing complex biases, as such approaches may miss subtle aspects of the issue. Generalizing to undefined groups may be less effective in cases where specific subgroups need focused attention. Furthermore, in situations where bias is linked to interactions between variables, a targeted subgroup analysis would be necessary. For example, if the bias is only present among young women of a specific nationality, not exploring such subgroups may fail to detect it.
For these reasons, the objective of the thesis, bias mitigation in structured data, is achieved through the identification of subgroups. This section does not detail how the subgroup identification is performed in some related works, as the algorithm used - DivExplorer [5] - is thoroughly explained in the *Background* section. Therefore, for the purposes of this part, it is assumed that the method for identifying subgroups is already established and understood as a foundational aspect of the approach.

As previously mentioned, analyzing bias mitigation techniques which are effective for specific data types can broaden the understanding of the problem and inspire adaptable methods for other contexts. In this thesis, the bias mitigation approach applied to speech data [11] has played a pivotal

role. This method identifies problematic subgroups where the ML model performs poorly.

Here, data are initially split into four sets: train, test, validation, and holdout (or held-out). Subgroup discovery occurs in the validation set, focusing on the top K problematic subgroups with the highest divergence, defined as *"a measure of different classification behavior on data subgroups"* [5]. Matching data is then retrieved from the held-out set and incorporated into the train set.

A critical advantage of this process is its cost-free nature: new data are not acquired but they are strategically selected from a reserved subset. This approach improves model performance, evaluated on the test set, while maximizing the effective use of available data; moreover it benefits from the advantages of an initial analysis of subgroups where the model is underperforming.

2.3 Discussion

In conclusion, this chapter has provided an overview of various bias mitigation techniques, emphasizing the distinction between approaches that require and do not require the automatic identification of subgroups. While methods that do not explicitly define subgroups, such as PnD and FairDo, offer flexibility and scalability, they may fall short in addressing complex, subtle biases that arise in specific subgroups. By contrast, approaches that involve the identification of problematic subgroups offer a more targeted and precise solution, particularly in cases where bias is linked to interactions between variables or underperforming model segments.

This thesis adopts a hybrid approach, combining the subgroup identification strategy used in speech data with synthetic data generation techniques such as SMOTE. By first identifying subgroups where the model exhibits underperformance, followed by data augmentation, this approach ensures that bias mitigation is both effective and efficient. The method leverages existing data, avoiding additional acquisition costs, and enhances fairness by addressing multiple causes of Representation Bias. Furthermore, it guarantees that the approach is grounded in the specific characteristics of the data, ensuring that no significant biases go unnoticed.

Ultimately, the combination of subgroup identification and data augmentation represents a powerful strategy for mitigating bias in structured

datasets, offering a more comprehensive and cost-effective solution to the complex issue of fairness in machine learning models.

Chapter 3

Background

In this chapter, the prerequisites necessary to fully understand the bias mitigation approaches developed throughout this thesis are presented. The mathematical theory and general notation used in the subsequent chapters are thoroughly examined, with particular attention to the operation of DivExplorer for subgroup identification.

Specifically, 3.1 outlines the main features of DivExplorer that make it particularly suitable for our purposes, highlighting how this algorithm will be used to identify problematic subgroups. Section 3.2 introduces key definitions such as *itemset, support set, outcome function, support, divergence*, and *f-divergence*, which are essential for understanding the algorithm's functioning. A general overview of the algorithm is provided in 3.3. Finally, the last section, 3.4, presents a mathematical explanation of why and when divergence is statistically significant and not merely the result of random statistical fluctuations.

3.1 Problematic Subgroup Identification via DivExplorer

The purpose of this section is to describe and analyze the functioning of DivExplorer, an algorithm for identifying problematic subgroups that leverages frequent pattern mining (FPM) techniques. This tool will be used to identify subgroups that pose challenges for a given ML model. The choice of DivExplorer over other existing approaches is motivated by its unique features, which make it particularly suited to the objectives of this study.

Specifically, the algorithm allows for the identification of problematic subgroups while ensuring they are **sufficiently represented** in the dataset; it also allows for a **complete exploration** of the dataset, which is a critical feature for a thorough and comprehensive analysis.
Another extremely important feature of the algorithm is that it can search for subgroups without altering the data in any way, thereby **ensuring the interpretability** of the results. An additional important property is that it is **model agnostic**. This is particularly useful for subgroup search because it allows the approach to be applied to any classification model, making it flexible and adaptable. Whether someone is working with decision trees, k-nearest neighbors, or neural networks, the approach remains consistent, enabling broad applicability across different model types without needing specific adjustments for each.

However, complete exploration is essential because the metrics used to evaluate performance differences between subgroups do not adhere to the property of monotonicity. This means that, given two subsets D_1 and D_2 of the main dataset D, where

$$D_1 \subset D_2 \subset D$$

it is not possible to establish a priori whether the divergence of D_1 is greater than, less than, or equal to that of D_2, also if it is known that D_2 is a superset for D_1. Thus, a full exploration of the dataset is necessary to ensure the identification of all relevant subgroups.

Additionally, DivExplorer stands out for its ability to set a representativeness threshold for subgroups and its computational efficiency. These characteristics make it the ideal tool for this study. The details of the algorithm's functioning, along with its definitions and properties presented in this chapter, will be entirely derived from the paper *"Looking for Truble: Analyzing Classifier Behavior via Pattern Divergence"* [5], where DivExplorer is introduced and described in detail.

3.2 Notation and Preliminary Definitions

As previously mentioned, divergence can be generally defined at a theoretical level as a measure of differing classification behavior on data subgroups. Here, subgroups refer to itemsets composed of multiple feature-attribute pairs.
The main formal definitions useful for these concepts are presented below.

Definition 3.2.1 (Itemset)**.** Given an n-dimensional dataset D, it consists of a set of instances on a set of A attributes, so that $|A| = n$ is the number of attributes; every attribute is discrete and $a \in A$ can take finite and discrete set D_a of values and m_a is such that $m_a = |D_a|$. An $item\alpha$ is an attribute equality $a = c$ for $a \in A$ and $c \in D_a$. An instance x is covered by the item $\alpha : a = c$, written $x \models a$, if $x(a) = c$.
Finally, an ***itemset*** is a set of items $I = \{\alpha_1, ..., \alpha_k\}$ such that

$$attr(\alpha_i) \neq attr(\alpha_j), \qquad \forall 1 \leq i < j \leq k$$

The itemset I can be also represented as the conjunction $\alpha_1 \wedge \cdots \wedge \alpha_k$ of its items.

Definition 3.2.2 (Support-set and Support)**.** The ***support-set*** of the itemset I is defined as $D(I) = \{x \in D \mid x \models I\}$ and consists of the instances that satisfy I.
The ***support*** of I is given by

$$\text{supp}(I) = \frac{|D(I)|}{|D|}$$

Definition 3.2.3 (Length of an Itemset)**.** The ***length of an itemset*** is the number of elements it contains, ranging from 0 (empty itemset) to n (number of attributes), attr(I) denotes the set of attributes in an itemset I. For a subset $B \subseteq A$, I_B represents the itemsets over the attributes in B. Specifically, I_A consists of the itemsets that contain all the attributes of the dataset.

Definition 3.2.4 (Itemset f-divergence)**.** Consider a dataset D as before, be $f : 2^D \to R$ a function, such function represents a statistic that can be computed over subsets of the dataset, such as False Positive or negative classification rates.
The notation $f(I)$ is used to denote f evaluated on the set of instances that satisfy I.

The ***f-divergence*** of an itemset I is defined as the difference between the statistic f computed on I and the statistic f computed on the entire dataset. The f-divergence of itemset I is given by the following expression:

$$\Delta_f(I) = f(I) - f(D)$$

Instead of passing f directly to DivExplorer, f is specified as an outcome rate of an outcome function, this enables the efficient calculation of itemset divergences.
If f is generic or can be understood from the context, then $\Delta_f(I) \equiv \Delta(I)$.

Property 3.2.1 (a finer discretization never hides *f-divergence*)**.** *Let X be a set of instances, and let $X_1, ..., X_m$ be such that $\bigcup_{i=1}^{m} X_i = X$ and $X_i \bigcap X_j = 0$, $\forall\ 1 \leq i < j \leq m$, so that $X_i, ..., X_j$ is a partition for X, it holds that: $\exists i \in \{1, 2, \ldots, m\}$ such that*

$$|\Delta_f(X)| <= |\Delta_f(X_i)|$$

so for any f-divergence measure, there is at least one subset X_i, $1 \leq i \leq m$, with f-divergence equal or greater than the f-divergence of X in absolute value.

The demonstration follows from the definition of the overall f-divergence: the f-divergence of X is a weighted average of $X_i, ..., X_m$.
Mathematically,

$$\Delta_f(X) = \sum_{i=1}^{m} \frac{|X_i|}{|X|} \Delta_f(X_i)$$

The implication of this property is that if a discretization is refined further, for every divergent itemset in the coarser discretization, there is at least one finer itemset that has equal or greater divergence.

Definition 3.2.5 (Outcome Function and Positive Outcome Rate)**.** As in the previous definitions, let D be a dataset, it is defined as
outcome function the function

$$o : D \to \{T, F, \bot\},$$

where the letters T and F stand, respectively, for **True** and **False** and the symbol $\bot$ stands for **Ignored**.
In the same context and under the same notation, it is defined as ***posive outcome rate*** of o over a set of instances $X \subseteq D$ the ratio that measures the fraction of instances in X for which the outcome function o yields a positive result T relative to the total number of instances in X for which the outcome function is defined. Matematically,

$$f_o(X) := \frac{|\{x \in X \mid o(x) = T\}|}{|\{x \in X \mid o(x) \neq \bot\}|}$$

it is possible to notice that the instances x such that $o(x) = \perp$ are not counted in the number that represents the positive outcome rate.

An outcome function like the following:

$$o_{fpr}(x) = \begin{cases} T & \text{if } u(x) \wedge \neg v(x), \\ F & \text{if } \neg u(x) \wedge \neg v(x), \\ \perp & \text{if } v(x). \end{cases}$$

is an example of outcome function suitable for this study, indeed when the function $v : D \to \{T, F\}$ is the ground truth and $u : D \to \{T, F\}$ is the classification[1] outcome, o_{fpr} is suitable to study the False Positiverate. Indeed, going into the details:

- $o_{fpr}(x)$ is True when $u(x) \wedge \neg v(x)$, here the classifier predicts positive $(u(x) = True)$, but the ground truth is negative $(v(x) = False)$;
- $o_{fpr}(x)$ is False when $\neg u(x) \wedge \neg v(x)$, here the classifier predicts negative $(u(x) = False)$, and the ground truth is also negative $(v(x) = False)$;
- $o_{fpr} = \perp$, here the ground truth is positive $(v(x) = True)$ this case excludes the instance from consideration, as it concerns instances with positive ground truth. These may correspond to true positives or False Negatives.

In this case the related outcome rate is the following:

$$f_{o_{fpr}}(X) := \frac{|\{x \in X \mid o_{fpr}(x) = T\}|}{|\{x \in X \mid o_{fpr}(x) \neq \perp\}|}$$

Another appropriate example of an outcome function is the next one:

$$o_{fnr}(x) = \begin{cases} T & \text{if } \neg u(x) \wedge v(x), \\ F & \text{if } u(x) \wedge \neg v(x), \\ \perp & \text{if } v(x). \end{cases}$$

This function computes the number of False Negatives in fact:

[1] under the hypothesis that dealing with classifiers is the focus

- $o_{\text{fnr}}(x)$ is True when $\neg u(x) \wedge v(x)$, that is the classifier predicts negative ($u(x) = \text{False}$) but the ground truth is positive ($v(x) = \text{True}$).
- $o_{\text{fnr}}(x)$ is false when $u(x) \wedge \neg v(x)$, meaning the classifier predicts positive ($u(x) = \text{True}$) but the ground truth is negative ($v(x) = \text{False}$).
- $o_{\text{fnr}}(x) = \perp$ when $\neg v(x)$, that is the ground truth is negative, in which case the instance is excluded from consideration.

Similarly to before, the related outcome rate measures the fraction of instances in X for which the outcome function yields a False Negative result (i.e., True) relative to the total number of instances in X for which the outcome function is defined (i.e., when the ground truth is positive), mathematically:

$$f_{o_{fnr}}(X) := \frac{|\{x \in X \mid o_{fnr}(x) = T\}|}{|\{x \in X \mid o_{fnr}(x) \neq \perp\}|}$$

From this discussion, it is evident that the subgroup search approach is model agnostic. In fact, the classification result u represents the result of a generic classification function, not a specific one. Therefore, the approach is applicable to any classification model, making it versatile and independent of the particular classifier used. In other words, the method does not rely on a specific classification model and can be applied to various models without modification to apply to the model itself.

3.3 DivExplorer Algorithm Overview

The algorithm described in [5] focuses on efficiently identifying subgroups within a dataset where the behavior of a classification model significantly deviates from its overall behavior. This approach is implemented in the DivExplorer tool, which leverages FPM techniques to extract meaningful patterns and evaluate their divergence, making it possible to uncover issues like bias or systematic Errorsin the model.

The process begins by preparing the dataset, discretizing continuous attributes, and encoding the classifier's outcomes (such as FP or FN) into a format that allows for efficient computation. The encoded outcomes are used to compute metrics like the positive outcome rate for any subset of data. This step ensures that the algorithm is flexible and applicable to various metrics of interest.

The algorithm then performs a systematic exploration of frequent patterns in the dataset using methods like Apriori or FP-growth, which are well-established techniques in FPM. For each candidate subgroup (itemset) identified, it calculates its frequency (support) in the dataset and evaluates its divergence. As stated before, divergence is computed as the difference between the metric observed within the subgroup and the metric observed globally across the entire dataset. Only itemsets that meet a user-defined support threshold are retained, ensuring that the analysis focuses on statistically significant subgroups.

What sets this algorithm apart is its integration of divergence computation directly within the pattern mining process. This approach eliminates the need for additional dataset scans, significantly improving efficiency. Furthermore, the algorithm is designed to support multiple metrics, such as FPR and FNR, allowing for a comprehensive evaluation of the classifier's performance.

DivExplorer focuses on frequent itemsets, defined by a user-specified threshold, to avoid the impact of statistical fluctuations in low-support groups and prioritize divergences that affect substantial portions of the dataset. To streamline results, it employs post-exploration pruning: a pattern is discarded if adding an attribute contributes only marginally to divergence below a threshold ϵ:

$$|\Delta_f(I) - \Delta_f(I - \{\alpha\})| \leq \epsilon,$$

the modification assumes that the pattern $I - \alpha$ captures the f-divergence of pattern I. This ensures the output highlights only the most meaningful and informative patterns, eliminating redundancy and noise.

At the end of the process, the algorithm outputs a ranked list of subgroups, sorted by their divergence values. This allows users to quickly identify the most critical patterns where the classifier's behavior diverges from its overall performance. By providing insights into subgroup-specific behaviors, the algorithm facilitates fairness analysis, Error detection, and model debugging, enabling a deeper understanding of the model's limitations and potential biases. It is important to recall that the combination of efficiency, scalability, and model-agnostic design makes this algorithm a powerful tool for exploring and improving ML models.

3.4 Statistical Significance of Divergence

Once it is established that DivExplorer identifies problematic subgroups by ranking them based on their divergence, one might question whether the observed divergence is statistically significant or merely the result of random statistical fluctuations due to the finite size of the dataset.
Since the outcome function is boolean, it is possible to adopt an approach based on Bayesian statistics [12]; in particular, one can assume that **the goal is to estimate the precision in the knowledge of the positive rate**. Under this hypothesis, each instance in the itemset can be viewed as a **Bernoulli trial** where:

- the outcome T represents a success with probability Z
- the outcome F represents a failure with probability $1 - Z$

The goal is therefore to estimate the parameter Z which corresponds to the positive success rate as said before.
A set of n instances observed in the itemset produces k_+ successes (T) and k_- failures (F), with $k_+ + k_- = n$.

Then:

$$k^+ = |\{x | x \models I \wedge o(x) = T\}|, \qquad k^- = |\{x | x \models I \wedge o(x) = F\}|$$

This scenario can be modeled as a series of independent trials, each characterized by a common parameter Z, typical of the Bernoulli distribution. Initially, before observing any trial, the value of Z is unknown.
To reflect this lack of knowledge, a uniform distribution is assumed as the prior for Z, such that

$$P(Z) = 1, \qquad Z \in [0, 1]$$

This implies that every value of Z in the range [0, 1] is considered equally probable.

Using Bayes' rule, after observing k^+ successes and k^- failures, the knowledge about Z can be updated as follows:

$$P(Z \mid \text{data}) \propto P(\text{data} \mid Z) P(Z)$$

where data refers to the subset of instances $D(I)$ from the main dataset D that satisfy the conditions defined by a specific itemset I. Each instance in "data" contributes to the calculation of success and failure rates based on an outcome function $o(x)$, which categorizes the instances (e.g., as true positive, False Positive , or ignored) according to the classifier's predictions and the ground truth labels.

Here, $P(\text{data} \mid Z)$ represents the likelihood of the observed data given Z.

For independent Bernoulli trials, this likelihood follows a binomial distribution[13]:

$$P(\text{data} \mid Z) = Z^{k^+}(1-Z)^{k^-},$$

since $P(Z)$ is the uniform prior assumed initially.
Combining these, we get:

$$P(Z \mid \text{data}) \propto Z^{k^+}(1-Z)^{k^-},$$

but

$$P(Z \mid \text{data}) \propto Z^{k^+}(1-Z)^{k^-} = \text{Beta}(k^+ + 1, k^- + 1)(Z)^2.$$

For the previous Beta distribution, the mean is:

$$\mu_I = \frac{k^+ + 1}{k^+ + k^- + 2}$$

and the variance is:

$$\nu_I = \frac{(k^+ + 1)(k^- + 1)}{(k^+ + k^- + 2)^2(k^+ + k^- + 3)}$$

the advantages of this form of mean and variance is the numerical stability when $k^+ + k^- = 0$, so when the outcome function is $\perp$ on the itemset considered.

---

[2]The Beta distribution [13]is defined as:

$$\text{Beta}(\alpha, \beta)(Z = z) = kz^{\alpha-1}(1-z)^{\beta-1}, \quad z \in [0,1]$$

where k is a normalization constant and $\alpha, \beta > 0$. The expected value and variance are:

$$E[\text{Beta}(\alpha, \beta)] = \frac{\alpha}{\alpha + \beta}, \quad \text{Var}(\text{Beta}(\alpha, \beta)) = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}.$$

Finally with these values, it is possible to compare the positive rate observed in the itemset with the global rate using a Welch's t-test, which accounts for the estimated variances.

This approach allows for a robust estimation of the significance of the divergence, even for itemsets with limited data, leveraging the properties of the Beta distribution to rigorously represent the uncertainty associated with the estimation of the parameter Z.

Chapter 4

Proposed Solution

This chapter provides a detailed analysis of the method developed for Bias Mitigation in tabular data. The proposed method, referred to throughout this work as *SMOTE-NC Data Generation*, represents an innovative approach and will be examined here from a theoretical perspective, while its practical application, along with experimental results, will be discussed in the next chapter.

Specifically, 4.1 outlines the main steps of the bias mitigation proposal presented in this thesis. The proposed approach can ideally be divided into three main steps, which are described in 4.2, 4.3, and 4.4. These sections cover, respectively, the identification of subgroups, the generation of new data, and the retraining phase.

Within section 4.3, the subsection 4.3.1 explains the general functioning and use of SMOTE and SMOTE-NC, while sections 4.3.2 and 4.3.3 provide a mathematical description of these two methods.

4.1 SMOTE-NC Data Generation method

The method developed in this thesis work, described in detail in this section, is based on the premise that bias mitigation in tabular data requires targeted interventions to address underrepresented or problematic subgroups. These subgroups, if neglected, can compromise the fairness and accuracy of machine learning models, making it essential to adopt a systematic approach to mitigate disparities and improve the representativeness of the data, thereby enhancing predictive performance.

The *SMOTE-NC Data Generation Method* is an innovative approach designed to tackle these challenges by systematically augmenting the training set with synthetic data. This method leverages data augmentation to improve the representation of minority subgroups and, finally, to reduce bias in model predictions. After the preprocessing and the splitting of the data into training, test and validation sets, and after training the model, the proposed solution is structured into three main phases, as shown in the figure 4.1.

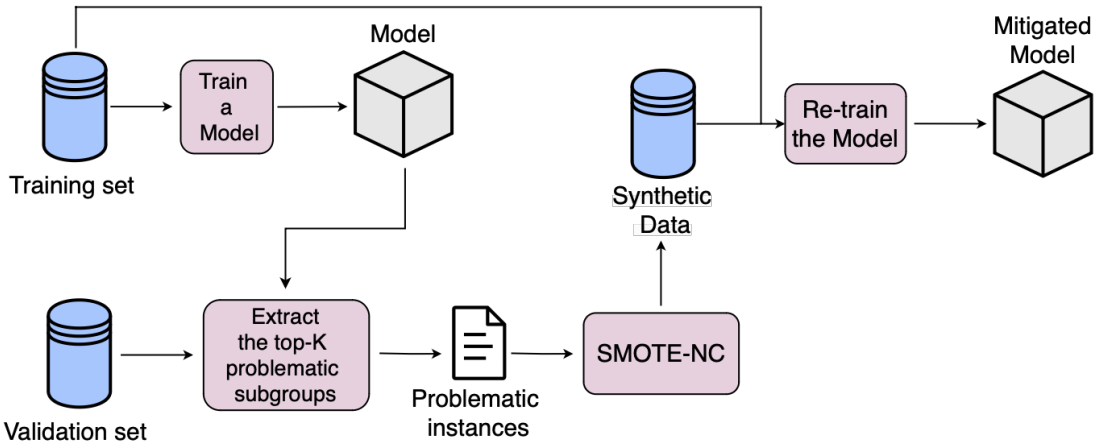


Figure 4.1: Visual description of the SMOTE-NC generation method. After preprocessing, a mitigation metric and model are chosen, and problematic subgroups are identified in the validation set using DivExplorer. Instances from the validation that match these subgroups (Problematic instances) are then selected, and new data points are generated based on these instances. The generated samples are added to the training set, followed by model retraining to enhance performance and fairness.

In particular, the first phase focuses on identifying problematic subgroups within the dataset, specifically in the validation set. These subgroups are defined as subsets of data that exhibit bias, are underrepresented, or demonstrate disparities in model performance, posing a risk to the overall fairness of the system.

The second phase focuses on generating synthetic samples that replicate the characteristics of the identified problematic subgroups. This is achieved by selecting instances from the validation set (here called problematic instances)that match the problematic subgroups detected in the validation set. Synthetic data is then generated using techniques such as **SMOTE** (for numerical variables) or **SMOTE-NC** (a more general approach suited

for datasets containing categorical variables). This process ensures that the generated samples preserve the statistical properties of the subgroups while mitigating their underrepresentation.

The third phase incorporates the newly created samples into the training set, enhancing its diversity and mitigating the imbalance or bias present in the original dataset. The model is then retrained on this augmented training set to improve its performance and fairness.

4.2 Problematic Subgroup Identification

The identification of problematic subgroups represents the first step in the methodology proposed in this Bias Mitigation study.
This process is carried out using DivExplorer, a tool whose theoretical foundations and functionality have been detailed in the *Background* chapter. Specifically, in this work, DivExplorer is employed to detect problematic subgroups, following the divergent subgroup search approach outlined in [11].

After selecting a Machine Learning model, defining a metric for identifying underperforming subgroups (e.g., False Positives, False Negatives, or Error rate), and splitting the dataset into three subsets: train set, test set, and validation set, the search for problematic subgroups is performed on the validation set. Thanks to the model predictions on the validation set and the corresponding ground truth labels, DivExplorer can be applied to the validation set itself to identify problematic subgroups. Then, using these subgroups, the problematic instances from the validation set—i.e., instances that contain a certain number of problematic subgroups—are used to generate new data. By the way, during the analysis, both the reference metric and the support threshold—representing the minimum percentage of subgroup presence in the considered dataset portion—can be specified. By varying these parameters, it is possible to identify subgroups and determine which ones are problematic or divergent.

4.3 New samples Generation

The core idea of this Bias Mitigation approach is that increasing the presence of samples in the train set that match the identified problematic subgroups

can significantly improve the model's predictions for these subgroups. Initially, the model's performance on these subgroups may be poor, but by incorporating a sufficient number of representative examples into the train set, the model becomes better at generalizing, leading to more accurate predictions for these specific cases.

The methodology developed in this work involves, after identifying the problematic subgroups as described in the previous section, generating new samples that match these subgroups using a linear interpolation method called **SMOTE**, which is applicable when all variables are numerical, or **SMOTE-NC**, which is used when there are also categorical variables.

This approach provides greater flexibility, allowing for:

- the selection of how many of the problematic subgroups identified by DivExplorer should be considered for sample generation;
- the total number of synthetic samples to be generated;
- the distribution of these synthetic samples across the different classes.

In the next sections, the mathematical formulation and the explanation of SMOTE and SMOTE-NC will be provided.

4.3.1 SMOTE and SMOTE-NC

Datasets are defined as imbalanced when the samples belonging to a certain class are significantly fewer than those of another class. This condition poses a significant challenge for ML models, as the uneven distribution of classes can compromise the model's performance. Specifically, classifiers tend to underestimate the importance of the minority class, resulting in difficulties in correctly identifying the examples that belong to it. For this reason, the use of specific techniques designed to handle imbalanced datasets proves particularly valuable, especially in contexts where data subgroups are underrepresented but crucial for analysis.

One of the most widely used techniques to address the issue of class imbalance is SMOTE (Synthetic Minority Over-sampling Technique), introduced by Chawla et al. (2002)[14]. This technique, designed for datasets with numerical features, generates new synthetic examples for the minority class by performing linear interpolation between pairs of existing points belonging to the same class.

An extension of SMOTE presented in the same publication is SMOTE-NC (Synthetic Minority Over-sampling Technique for Nominal and Continuous features)[14], which allows the method to be applied to datasets containing both numerical and categorical features. SMOTE-NC combines linear interpolation for numerical features with a probabilistic approach for categorical features, where category values are chosen randomly based on the most frequent categories among the nearest neighbors.

Both approaches stand out for their ability to enhance the representation of the minority class without simply duplicating the original data, thereby reducing the risk of overfitting and improving the model's performance.

This section will delve into and mathematically formalize the synthetic data generation technique employed by the two methods, with a particular focus on both numerical and categorical values. For simplicity, the generated instance will be referred to as a *synthetic data point, a synthetic point, a synthetic instance*, or simply *data, instance or point*, whenever it is clear from the context that it refers to a simulated instance rather than an existing one.

While this chapter focuses solely on the theoretical and mathematical formalization of data generation, the following chapters will explore how SMOTE is applied to address the problem of bias mitigation.

4.3.2 SMOTE Theoretical Formulation

As will become evident from the theoretical discussion presented in this section, one of the distinctive features of SMOTE is that the generation of synthetic examples occurs in the **feature space**, rather than in the data space. This means that synthetic data is created by **operating directly on the numerical or categorical variables** that describe the observations in the dataset, without manipulating the raw data, such as images, text, or signals.

Working in the feature space makes the method less constrained by the specifics of a given application, enabling its use in a wide variety of contexts. For instance, in the case of numerical variables, SMOTE generates new data through linear interpolation between nearby points belonging to the minority class, avoiding mere replication of existing examples.
Formally, let D be an imbalanced dataset composed of n classes, and let $P_{class_1} \subset D, \ldots, P_{class_n} \subset D$, where P_z for $z \in \{1, \ldots, n\}$ is the set of data points belonging to class z.

Under the hypothesis that i is the minority class in D, let P_i represent the set of data points that belong to this class. The objective is to generate new synthetic points specifically for P_i, thereby increasing its representation in the dataset.

To achieve this, the procedure begins with the random selection of a point

$$P_i^{\text{chosen}} \in P_i$$

For a fixed parameter k, the k-nearest neighbors of P_i^{chosen} within P_i are identified.
From these neighbors, one is randomly selected, denoted as k_j^{chosen}.

Given the two data points P_i^{chosen} and k_j^{chosen}, a new synthetic data point

$$P_{ij}^{\text{synthetic}}$$

is generated via linear interpolation as follows:

$$P_{ij}^{\text{synthetic}} = P_i^{\text{chosen}} + r \cdot (k_j^{\text{chosen}} - P_i^{\text{chosen}})$$

where r is a random scalar uniformly sampled from the interval $[0, 1]$.

This interpolation is performed component-wise , that is to say feature by feature, ensuring that each feature of the synthetic point is calculated as a linear combination of the corresponding features of P_i^{chosen} and k_j^{chosen}. Repeating this procedure multiple times generates a desired number of synthetic samples for the minority class P_i.

The figure 4.2. illustrates a simple and stylized example of the generation of a synthetic sample in a small unreal dataset consisting of only two features. The synthetic point generated, $P_{ij}^{\text{synthetic}}$, lies on the line segment connecting the two points P_i^{chosen} and k_j^{chosen}. This is because the generation method relies on linear interpolation. Linear interpolation calculates an intermediate point along the segment joining two points in a vector space, using a weighted combination of their coordinates, where the weight is determined by a random value r chosen in the interval [0,1].

4.3.3 SMOTE-NC Theoretical Formulation

As reminder, SMOTE-NC is an extension of SMOTE designed to handle datasets that contain both numerical and categorical features. This method

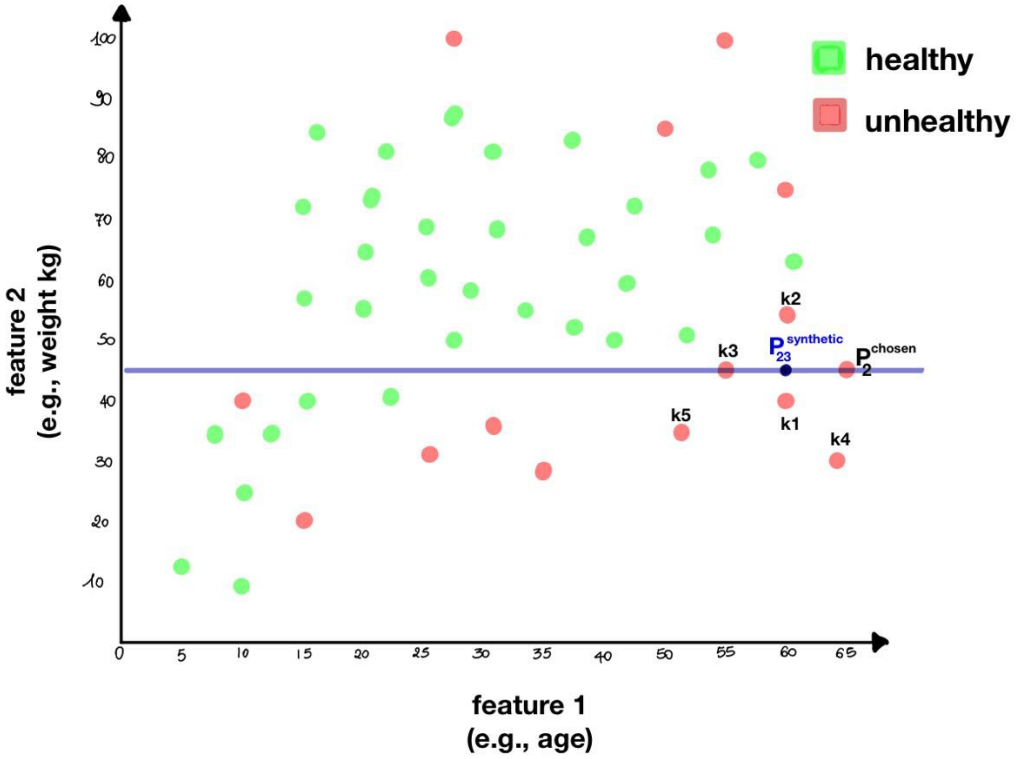


Figure 4.2: Example of SMOTE synthetic data generation, when the number of features is 2, $r = 0.5$, $k_j^{\text{chosen}} = k_3 = (55, 45)$, $P_i^{\text{chosen}} = P_2^{\text{chosen}} = (65, 45)$, so $P_{ij}^{\text{chosen}} = P_{23}^{\text{chosen}} = (60,45)$.

adapts the synthetic data generation process to consider the distinct nature of categorical variables, ensuring meaningful interpolation across both feature types.

As with SMOTE, SMOTE-NC operates in the **feature space**, generating synthetic examples directly on the variables that describe the observations in the dataset. For numerical features, SMOTE-NC uses linear interpolation, while for categorical features, the new synthetic values are chosen based on the mode (most frequent category) among the neighbors. This hybrid approach enables the generation of realistic synthetic examples without altering the raw data, making the method versatile and applicable across various domains.

Formally,

let D represent an imbalanced dataset composed of n classes, and let $P_{class_1} \subset D, \ldots, P_{class_n} \subset D$, where P_z for $z \in \{1, \ldots, n\}$ is the set of data points belonging to class z.

Assuming i is the minority class in D, let P_i represent the set of data points that belong to this class. The goal is to generate new synthetic points for P_i while respecting the nature of both numerical and categorical features.

To achieve this, the process begins by selecting a point

$$P_i^{\text{chosen}} \in P_i$$

at random.

For a fixed parameter k, the k-nearest neighbors of P_i^{chosen} within P_i are identified. From these neighbors, one is randomly selected, denoted as

$$k_j^{\text{chosen}}.$$

To identify the k-nearest neighbors of a point $P_i^{\text{chosen}} \in P_i$, SMOTE-NC uses a mixed distance metric that combines numerical and categorical distances. Let the feature set of the dataset be divided into numerical components $\mathcal{N}$ and categorical components $\mathcal{C}$. The distance d between two points P and Q is defined as:

$$d(P, Q) = \sqrt{\sum_{x \in \mathcal{N}} (P_x - Q_x)^2 + \sum_{y \in \mathcal{C}} \Delta(P_y, Q_y) \cdot \text{Med}^2},$$

where P_x and Q_x are the values of the numerical feature x for points P and Q, respectively and $\Delta(P_y, Q_y)$ is 0 if the categorical feature y matches ($P_y = Q_y$), and 1 otherwise ($P_y \neq Q_y$), finally Med is the median of the standard deviations of the numerical features in P_i. This term scales the contribution of categorical mismatches relative to numerical distances, ensuring balanced weighting.

This mixed distance metric allows SMOTE-NC to consider both feature types effectively when identifying neighbors.

With the same distinction between categorical and numerical features, so that $\mathcal{N}$ (numerical) and $\mathcal{C}$ (categorical), for numerical features $x \in \mathcal{N}$, the synthetic data point $P_{ij}^{\text{synthetic}}$ is calculated using linear interpolation:

$$P_{ij,x}^{\text{synthetic}} = P_{i,x}^{\text{chosen}} + r \cdot (k_{j,x}^{\text{chosen}} - P_{i,x}^{\text{chosen}}),$$

where r is a random scalar uniformly sampled from the interval $[0, 1]$.

For categorical features $y \in \mathcal{C}$, the synthetic value is determined by selecting the most frequent category among P_i^{chosen}, k_j^{chosen}, and the other neighbors:

$$P_{ij,y}^{\text{synthetic}} = \text{Mode}(\{P_{i,y}^{\text{chosen}}, k_{j,y}^{\text{chosen}}, \dots, k_{max,y}\}).$$

By combining these rules for numerical and categorical features, SMOTE-NC produces a hybrid synthetic point:

$$P_{ij}^{\text{synthetic}} = \{P_{ij,x}^{\text{synthetic}} \text{ for } x \in \mathcal{N},\ P_{ij,y}^{\text{synthetic}} \text{ for } y \in \mathcal{C}\}.$$

It is possible to conclude that both SMOTE and SMOTE-NC represent a general and flexible approach that reduce the complexity associated with handling raw data by focusing on feature-space operations. This makes these methods particularly well-suited for structured or tabular datasets with numerical or categorical features.

A key advantage of both techniques is their independence from the data's domain or origin: they can be applied across diverse fields, such as healthcare, finance, image analysis, or natural language processing, as long as the data is represented as vectors of features. This flexibility ensures broad applicability regardless of the specific context or application requirements.

These characteristics make them particularly well-suited for addressing bias mitigation in structured data contexts, where an adequate representation of classes is crucial for improving model fairness and performances.

4.4 Model Retraining

The synthetic records generated using this technique are then incorporated into the training set, significantly enriching it with additional data points that have been created through interpolation.
This process involves generating new samples that share the same statistical properties as the identified problematic subgroups, ensuring that the model is exposed to a more diverse and balanced set of training examples. The newly generated samples not only enhance the representativeness of underrepresented subgroups but also reduce the bias in the model by addressing the disparities that existed in the original dataset.

Once these synthetic samples are added to the training set, the next step involves retraining the selected ML model on this augmented dataset. This retraining process ensures that the model is exposed to a broader range of data, ideally improving its ability to generalize across previously underrepresented subgroups. The key advantage of this phase is that the model is no longer trained on a dataset with imbalanced subgroup representation, but on one that has been adjusted to provide a different distribution of subgroups.

To evaluate the effectiveness of this bias mitigation strategy, the model's performance is compared before and after the augmentation phase. This comparison can be done both qualitatively and quantitatively.
Qualitative evaluation involves examining how well the model now handles the previously problematic subgroups by looking at its predictions on those specific groups.
Quantitatively, performance metrics such as Accuracy, False Positives, False Negatives, and Error rates can be measured and compared, providing a clear picture of how much improvement has been made. This comprehensive evaluation allows for a thorough understanding of the impact of the bias mitigation technique and its ability to enhance the fairness and performance of the ML model.

The next chapter will present the experiments conducted and the results obtained by applying the strategies outlined in the baselines. Practical details of the implementation and an in-depth evaluation of the experimental results will be analyzed.

Chapter 5

Experimental Setting and Results

This chapter describes the experimental setting and presents the results of the experiments conducted to test the proposed bias mitigation strategy. The code used in this thesis can be found in [15]. In particular, 5.1 explains the criteria used for describing the datasets, highlighting which aspects are emphasized and which are omitted. Subsections 5.1.1 and 5.1.2 focus on the Adult and COMPAS datasets, respectively, including only the elements relevant to the proposed Bias Mitigation solution. Section 5.2 focuses on the experiments and results. In particular,it presents the competitor strategies and the metrics used to assess bias mitigation. Its subsections—5.2.1, 5.2.2, 5.2.3, and 5.2.4—describe, respectively, the evaluation metrics used to determine whether and which strategy is the most effective, the models trained and used for predictions, the results obtained on the Adult, and those obtained on the COMPAS datasets. The last two subsections are further divided into six sub-subsections each. For the Adult dataset, 5.2.3.1 presents the results of BM when the chosen metric is False Positives, with a general overview of the outcomes provided in 5.2.3.2. Similarly, 5.2.3.3 focuses on the results when False Negatives are the selected metric, with the main outcomes summarized in 5.2.3.4. Finally, 5.2.3.5 extends the analysis to the more general metric of total Errors, with the main outcomes in 5.2.3.6. For the COMPAS dataset, the same structure is followed. 5.2.4.1 presents the results for False Positives, with a summary in 5.2.4.2. 5.2.4.3 focuses on False Negatives, with the main findings outlined in 5.2.4.4. Lastly, 5.2.4.5 examines total Errors, concluding the key outcomes in 5.2.4.6.

5.1 Datasets Description

In this chapter, the concepts explained theoretically so far will be put into practice by presenting the results of bias mitigation experiments. Particular attention will be given to the preprocessing phase of the dataset, focusing on its attributes, values, and discretization. These steps are crucial for implementing the strategy, especially as this work involves the injection of new data points designed to address problematic subgroups. Discretization plays a fundamental role, as tools like DivExplorer are highly sensitive to how data is grouped. By managing variable granularity, it becomes possible to identify fairness issues more effectively and provide explainable solutions.

5.1.1 Adult Dataset Description

The ***Census Income Dataset***, more commonly known as ***Adult***, is a dataset provided by the U.S. Census Bureau through the UCI Machine Learning Repository. It is widely used for classification problems and predictive analysis, specifically to determine whether an individual's annual income exceeds $50,000 based on economic and socio-demographic attributes.

As shown in Table 5.1, the dataset includes sensitive variables such as **gender** and **race**, making it particularly suitable for analyzing and mitigating bias in tabular data. The target variable, **income**, serves as an ideal case study for evaluating the fairness of predictive models. Additionally, it demonstrates how the identification of problematic subgroups can highlight disparities observable in real-world scenarios. This underscores the importance of adopting methods that not only mitigate bias but also ensure high interpretability and explainability of results.

The original dataset, as downloaded and before preprocessing, consists of 32,561 instances and 15 attributes, including the target variable income.

During the preprocessing phase, duplicate entries were removed, resulting in a total of 32,537 available instances, with 24,698 belonging to class 0 and 7,839 to class 1. This indicates a strong class imbalance toward class 0;values with similar meanings were standardized. For example, values such as *'?'* and *'Unknown'* in the *workclass* feature were unified under the same label. Similarly, for the *native-country* feature, categories such as *'Germany', 'England', 'Scotland', 'France', 'Italy', 'Ireland', 'Greece', 'Poland', 'Portugal', 'Yugoslavia',* and *'Hungary'* were grouped under a single value,

Table 5.1: Description and Type of attributes in the Adult dataset.

Attribute	Description	Values
age	Age of the individual	Numeric
workclass	Employment type	Categorical
fnlwgt	Final weight (sampling weight)	Numeric
education	Highest level of education	Categorical
education-num	Years of education	Numeric
marital-status	Marital status	Categorical
occupation	Type of occupation	Categorical
relationship	Relationship to the household	Categorical
race	Race of the individual	Categorical
sex	Gender of the individual	Categorical
capital-gain	Capital gains	Numeric
capital-loss	Capital losses	Numeric
hours-per-week	Hours worked per week	Numeric
native-country	Country of origin	Categorical
income	Income level of the individual	<50K, >50K

'Caucasian/White'. Features like *educational-num* and *age* were discretized to enhance interpretability and facilitate the analysis. Only after completing these steps was label encoding applied to convert categorical variables into numerical representations.

This is just one example of the preprocessing steps applied to the dataset. More details on the number of possible values for each feature, both before and after the full preprocessing, can be found in Table 5.2, this table provides only general information; if necessary, further details will be provided regarding the possible values. It is important to note that DivExplorer is sensitive to discretization choices, and passing features to the tool prior to encoding ensures explainability, maintaining the interpretability of both the input data and the results, this concept will be more clear in the next subsection.

Table 5.2: Number of attributes in the Adult dataset before and after preprocessing.

Attribute	**# original distinct values**	**# preprocessed distinct values**
age	73	6
workclass	9	6
fnlwgt	21648	only normalized
education	16	4
education-num	16	12
marital-status	7	5
occupation	15	6
relationship	6	6
race	5	5
sex	2	2
capital-gain	119	only normalized
capital-loss	92	only normalized
hours-per-week	94	3
native-country	42	6
income	2	2

5.1.2 COMPAS Dataset Description

The ***COMPAS Dataset*** is derived from the COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) risk assessment tool, which is used in the American criminal justice system to predict the likelihood that an individual will reoffend. The dataset, sourced from the *ProPublica* investigation and available through Kaggle, includes information on individuals assessed by COMPAS, such as demographic attributes, prior criminal history, and risk scores. It is widely used in fairness and bias analysis within machine learning, particularly to study algorithmic decision-making and potential disparities across different demographic groups. The

original dataset consists of 52 features and 18,316 instances and includes sensitive information such as **race** and **gender**, making it particularly suitable for the objectives of this thesis. During the preprocessing phase, redundant or less informative features—such as first name, last name, middle name, and arrest date—were removed. The target variable used for the analysis is the feature renamed "**violent recidivist**", which takes the value 1 if an individual is considered a repeat offender in the case of violent crime and 0 otherwise. After eliminating irrelevant features and instances with missing values, the dataset comprises 10 features, including the target variable, and a total of 18,293 instances, with 16,954 belonging to class 0 and 1,339 to class 1. This indicates a strong class imbalance toward class 0. Table 5.3 provides a concise description of the variables along with their respective types.

Table 5.3: Description and Type of Attributes in the COMPAS Dataset.

Attribute	**Description**	**Values**
sex	Gender	Categorical
race	Race/ethnicity	Categorical
recidivism risk	Risk score for recidivism	Numeric
risk level	Risk category w.r.t. recidivism score	Categorical
violent recidivism risk	Predicted risk score for violent recidivism	Numeric
violent risk level	Risk category w.r.t. violent recidivism score	Categorical
juvenile offenses	# Prior juvenile offenses	Numeric
age	Age at the assessment time	Numeric
prior offenses	# Prior Offenses	Numeric
violent recidivist	Whether the individual reoffended violently	Binary (0 = No, 1 = Yes)

The feature juvenile offenses does not exist in the original dataset; it was created during the preprocessing phase by summing the values of three features: juv_fel_count, juv_misd_count, and juv_other_count. These features represent, respectively, the number of juvenile felony offenses, juvenile

misdemeanor offenses, and other juvenile offenses recorded for an individual.

Table 5.4: Number of attributes in COMPAS dataset before and after preprocessing.

Attribute	**# original distinct values**	**# preprocessed distinct values**
sex	2	2
race	6	6
recidivism risk	10	10
violent recidivism risk	10	10
violent risk level	3	3
juvenile offenses	-	14
age	65	6 - discretized
prior offenses	39	7 - discretized
violent recidivist	2	2

For this reason, in Table 5.4, which details the number of distinct values each feature could assume before and after preprocessing, a dash ("-") is placed in the corresponding row for juvenile offenses under the "Before Preprocessing" column, indicating that it was not originally present in the dataset.

Recall that it is important to note that DivExplorer is sensitive to discretization choices, and passing features to the tool prior to encoding ensures explainability, maintaining the interpretability of both the input data and the results, this concept will be more clear in the next subsection.

5.2 Experiments and Results

In this section, the general method described in the previous chapter will be applied to the data from the tabular datasets previously described. Specifically, for the datasets described, the effect of the data generation technique on **False Positives, False Negatives, and overall number of Errors** will be examined.
The results obtained with this method will be compared to those from two

alternative methods: the ***Random Data Acquisition Method***, which involves acquiring a specific number of random instances from the held-out set, which are then added to the training set for retraining, and the ***Targeted Data Acquisition Method***, which follows the approach proposed in the paper [11], where, after identifying the problematic subgroups, instances matching these subgroups are searched for in the held-out set and added to the training set for retraining. Finally, the method described in this thesis, known as the ***SMOTE-NC Data Generation Method*** or simply, the Data Generation Method, will be used for comparison.

5.2.1 Evaluation Metrics

In the following sections, concerning the metrics used for bias mitigation, recall that the mitigation will be performed by considering the **number of False Positives**, the **number of False Negatives**, and the **total number of Errors** made by the model. Regarding the evaluation metrics for the applied methods, **Accuracy** and **F1-Score** will be considered to assess the overall performance before and after mitigation. However, to assess the effectiveness of the mitigation, divergence values will be analyzed both before and after the process. Specifically, $\mathbf{\Delta_{avg}}$ represents the absolute mean divergence across all subgroups, $\mathbf{\Delta_{max}}$ is the worst-case absolute divergence for any subgroup, and $\mathbf{\Delta_i}$ denotes the absolute mean divergence for the top i most problematic subgroups.

5.2.2 Experimental Setting

The diverging subgroups are identified in the validation set. The experiments are conducted using the following machine learning models: Decision Tree, Gradient Boosting, Logistic Regression, and Random Forest. It is important to emphasize that the methodology adopted is model-agnostic, meaning it is independent of the specific model. These models are employed solely to enable comparisons and to evaluate the consistency of the proposed methodology. Indeed, the models were used with their default parameters, without any hyperparameter tuning or optimization. However, if any of these four models exhibit zero False Positives, zero False Negatives, or both, they will be replaced by the k-Nearest Neighbors algorithm to ensure a more balanced evaluation.

With the DivExplorer tool, it is possible to identify all subgroups with a minimum support in the dataset. In this context, **subgroups with positive divergence and a t-statistic $>$ 2 are considered problematic or divergent**, as they exhibit behavior that significantly deviates from the overall dataset.

Moreover, the following experiments will present the results of the mitigation process by adjusting the **min_sup** that is the minimum support: the minimum percentage of instances containing a specific subgroup. As this threshold changes, the number of problematic subgroups varies.

K% represents the percentage of problematic subgroups, relative to the total number of problematic subgroups found by varying the metric and support, that are used to perform the mitigation. Another parameter influencing the number of identified problematic subgroups is the redundancy or **pruning parameter**. This parameter defines the threshold below which a pattern is discarded if its contribution to the overall divergence is smaller than the defined threshold.

The following sections present the results of the mitigation applied to the previously described datasets, varying the problematic subgroups injected into the training set.

5.2.3 Adult Experiments and Results

The preprocessed Adult Dataset consists of a total of 32,537 instances. These are divided as follows: 40% (13,014 instances) is allocated to the train set, while the remaining 60% is evenly split among the test set, validation set, and held-out set, with 20% each (6,507 or 6,508 instances). Specifically, the train set contains 13,014 instances, the validation set has 6,507 instances, the holdout set includes 6,508 instances, and the test set comprises 6,508 instances.

Table 5.5 shows how the number of subgroups found with DivExplorer varies as the minimum support changes, while keeping the redundancy parameter fixed at $\epsilon = 0.01$ and the model used is the DT.

From this table, it is clear that the number of original subgroups found remains the same regardless of the reference metric. What changes according to the reference metric is the divergence, and indeed, it can be observed that the number of subgroups varies when the pruning parameter is adjusted between different metrics. This happens because the number of subgroups with a given support is an inherent characteristic of the dataset, independent

of the model and its performance.

Table 5.5: Number of subgroups before and after post-exploration pruning for DT model and fixed $\epsilon = 0.01$

Metric	Minimum Support	# Subgroups	# Pruned Subgroups
FP	10%	3793	224
	15%	1655	96
	20%	893	58
FN	10%	3793	658
	15%	1655	270
	20%	893	146
ER	10%	3793	271
	15%	1655	133
	20%	893	83

However, when pruning is applied, it is done based on divergence, which instead varies depending on the model and the chosen metric.

The Table 5.6 shows the most divergent subgroups with respect to the False Positive, False Negative and Error Metrics, along with their respective divergences for the aforementioned Decision Tree Model when redundancy parameter is $\epsilon = 0.01$ and the minimum support varies from 10% to 20% of the validation set.

Such table provides a detailed overview of the problematic subgroups for the various metrics. From this point, it is possible to understand what interpretable subgroups mean: intuitively, it can be observed that, when the metric is False Positive , the problematic subgroups with feature-value pairs like *race = 'White'* or *relationship = 'Husband'* reflect the social belief that a white male (since "Husband") might more easily be associated with a higher income.

It is important to notice that, after preprocessing the data, the value names differ from those in the original dataset.

Specifically: {race='White'} encompasses the ethnicities: 'Germany', 'England', 'Scotland', 'France', 'Italy', 'Ireland', 'Greece', 'Poland', 'Portugal', 'Yugoslavia', 'Hungary'; {education ='Non Graduated'} includes: 'Preschool', '1st-4th', '5th-6th', '7th-8th', '9th', '10th', '11th', 'HS-grad',

'Some-college', '12th'; {hours = 'Overtime'} refers to more of 41 working hours per week.

Table 5.6: Number of subgroups, number of problematic subgroups, the most divergent one with different minimum support and different the metrics for DT model.

Metric	**Minimum Support**	**# Pruned Subgroups**	**#Divergent Subgroups**	**Most Divergent Subgroup**
FP	10%	224	169	{hours=Overtime, marital-status=Married, education=Bachelor's Degree}
	15%	96	60	{hours=Overtime, race= White, native-country=United-States, relationship= Husband}
	20%	58	34	{hours=Overtime, race= White, native-country=United-States, relationship= Husband}
FN	10%	658	426	{capital-gain=0.0,capital-loss=0.0 relationship= Not-in-family, workclass=Private, education=Non Graduated}
	15%	270	194	{capital-gain=0.0, capital-loss=0.0, education=Non Graduated, marital-status=Never-married}
	20%	146	104	{capital-gain=0.0, capital-loss=0.0, education=Non Graduated, marital-status=Never-married}
ER	10%	271	178	{education=Bachelor's Degree, capital-gain=0.0, marital-status=Married}
	15%	133	77	{marital-status=Married, native-country=United-States, capital-loss=0.0, capital-gain=0.0, hours=Overtime}
	20%	83	50	{marital-status=Married, native-country=United-States, capital-loss=0.0, capital-gain=0.0, hours=Overtime}

5.2.3.1 Adult False Positive Mitigation

As a preliminary analysis of the solution's impact on False Positives, Figure 5.1 is examined.

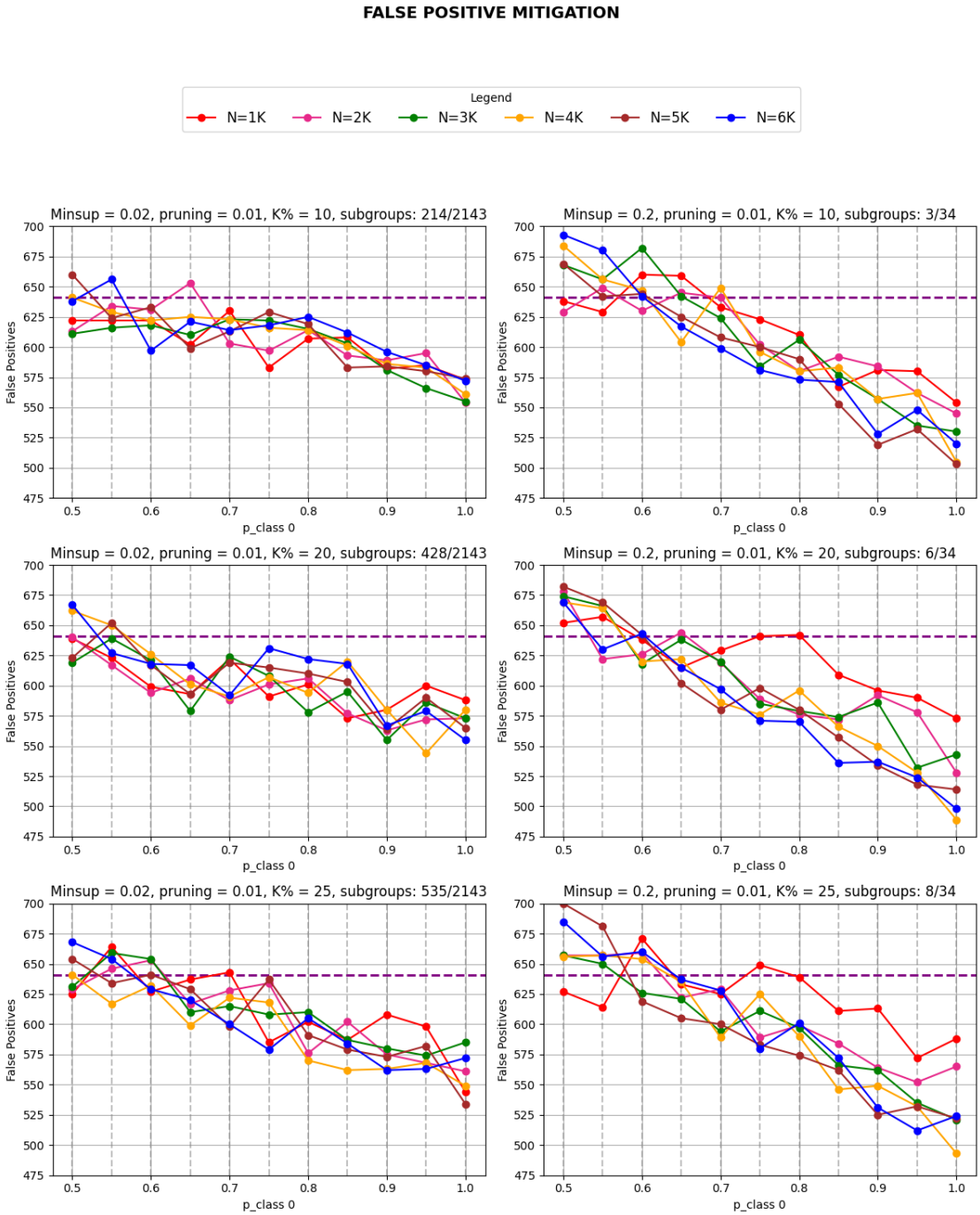
This image illustrates the impact of bias mitigation on False Positives using SMOTE-NC, applied to problematic subgroups identified with DivExplorer, and for a Decision Tree Model. The x-axis represents the probability that synthetic points belong to class 0, while the y-axis indicates the number of False Positives. To perform the mitigation, synthetic points (1K-6K) are injected into the training set aiming to balance the data and reduce False Positives. The experiments are conducted by varying the number of problematic subgroups used to identify problematic instances in the validation set, allowing an analysis of how different subgroup selections influence the effectiveness of the mitigation strategy.

On the left side of the figure, the minimum support threshold is lower, allowing for the identification of a larger number of problematic subgroups, while on the right, the higher support threshold results in fewer detected subgroups. From top to bottom, the number of subgroups selected for data augmentation increases, leading to a more gradual and stable reduction in False Positives, regardless of the number of synthetic points injected.

The overall trend shows that, for a lower support threshold, the number of False Positives drops below the initial level (before mitigation) as soon as the probability of a synthetic point belonging to class 0 exceeds 70%. For a higher support threshold, this occurs only when the probability reaches 80%. This difference arises because lower support captures more subgroups and more divergences, allowing mitigation to act on a broader range of problematic instances.

In contrast, a higher minimum support selects only the most representative and frequent patterns, reducing noise and preventing overfitting to specific subgroups. As a result, the number of False Positives reaches the lowest absolute value for this experimental setting when the probability of a sample belonging to class 0 is maximized (p=1). Recall that the probability of a sample belonging to class 0 is increased as part of a False Positive Bias Mitigation strategy.

Figure 5.1: False Positives trend generated with SMOTE-NC (1K-6K) as $p_{\text{class }0}$ varies for a Decision Tree. On the left, min_sup = 2%; on the right, min_sup = 20%; for both pruning parameter = 1%. Each row compares results for the same percentage of problematic subgroups used in mitigation.



However, this greater coverage can introduce more variability and noise in the synthetic data, leading to a less effective reduction in False Positives.

Although the previous image may provide an idea of the effects of mitigation, it says nothing about the progression of the divergence. Therefore, a more accurate view of how the evaluation metrics described at the beginning of the chapter vary can be found in Tables 5.7 and 5.8.

Such Tables compare the three data acquisition strategies -described at the beginning of the section- for False Positive Mitigation in a Decision Tree model, considering an increasing number of problematic subgroups (%K) to consider for mitigation the Representation Bias, and so more additional training samples.

In the first Table, the minimum support is set to 0.02, while in the second, it is set to 0.2.

By comparing the two tables, it can be observed that, given the same number of samples included in the training set, the Generation strategy, proposed in this thesis, is the most effective in reducing divergence and the number of False Positives when the number of inserted samples is sufficiently high.

On the other hand, when evaluating divergence performance with a smaller number of samples, the Targeted Acquisition approach performs better. In this case, all holdout points matching the problematic subgroups and belonging to class 0 were included, aligning with the objective of False Positive Mitigation . This confirms that the targeted inclusion of samples from problematic subgroups is particularly beneficial in the early stages of data acquisition, where the availability of new examples is limited, and each added sample has a greater impact on reducing divergence.

Another noteworthy observation is that in the second table, which considers a minimum support of 20%, the divergence values are overall lower compared to the case with a minimum support of 2%.

Table 5.7: Comparison of Results for Targeted Data Acquisition, Random Data Acquisition, and SMOTE-NC Data Generation Approaches for Decision Tree Model, metric: False Positive , MinimumSupport:2%, and Pruning:0.01.
Note: For each *% K* (10, 20, 25), the best results for each metric within the same sample size are marked with underline, while the overall best results are marked in **bold**.

% K	# Sub-groups	# Sam-ples	Approach	Accuracy	F1-Score	# FP	Δ_{avg}	Δ_{max}	Δ_{10}	Δ_{20}
-	-	-	Original	0.803	0.593	641	0.171	0.631	0.608	0.593
10	214	406	Random Acquisition	**0.805**	**0.595**	632	0.163	0.640	0.619	0.604
			Targeted Acquisition	**0.805**	0.586	600	0.152	0.545	0.531	0.512
			Generation $p = 1$	***0.805***	0.584	593	0.160	0.514	0.506	0.497
		5000	Random Acquisition	***0.805***	**0.595**	654	0.176	0.586	0.574	0.563
			Generation $p = 1$	0.800	0.564	**574**	**0.104**	**0.472**	**0.449**	**0.436**
20	428	553	Random Acquisition	0.805	**0.598**	650	0.167	0.636	0.614	0.602
			Targeted Acquisition	**0.811**	0.595	567	0.159	0.573	0.547	0.527
			Generation $p = 1$	0.806	0.585	583	0.160	0.649	0.613	0.590
		5000	Random Acquisition	0.805	**0.598**	654	0.176	0.586	0.574	0.563
			Generation $p = 1$	0.799	0.558	**565**	**0.126**	**0.552**	**0.510**	**0.497**
25	535	604	Random Acquisition	0.806	**0.603**	656	0.164	0.607	0.581	0.570
			Targeted Acquisition	0.808	0.579	576	0.139	0.548	0.537	0.529
			Generation $p = 1$	**0.811**	0.592	585	0.177	0.629	0.613	0.604
		6000	Random Acquisition	*0.804*	0.597	651	0.164	0.633	0.606	0.591
			Generation $p = 1$	0.797	0.553	**572**	**0.108**	**0.551**	**0.530**	**0.513**

This suggests that when problematic subgroups are less fragmented and more representative, the False Positive issue is more contained, and divergence reduction occurs more effectively, regardless of the acquisition strategy adopted.

Table 5.8: Comparison of Results for Targeted Data Acquisition, Random Data Acquisition, and SMOTE-NC Data Generation Approaches for Decision Tree Model, metric: False Positive , MinimumSupport:20%, and Pruning:0.01.
Note: For each % K (10, 20, 25), the best results for each metric within the same sample size are marked with underline, while the overall best results are marked in **bold**.

% K	# Sub-groups	# Sam-ples	Approach	Accuracy	F1-Score	# FP	Δ_{avg}	Δ_{max}	Δ_{10}	Δ_{20}
-	-	-	Original	0.803	0.593	641	0.054	0.190	0.172	0.153
10	3	1252	Random Acquisition	0.809	**0.605**	634	0.038	0.177	0.154	0.129
			Targeted Acquisition	***0.813***	0.592	533	0.009	0.123	0.079	0.049
			Generation $p = 1$	0.812	0.598	569	0.018	0.152	0.106	0.069
		8000	Random Acquisition	0.809	0.605	653	0.045	0.193	0.168	0.142
			Generation $p = 1$	0.806	0.559	**489**	**0.008**	**0.086**	**0.066**	**0.047**
20	6	1440	Random Acquisition	0.805	**0.600**	647	0.045	0.189	0.165	0.138
			Targeted Acquisition	0.811	0.589	547	0.012	0.135	0.100	0.061
			Generation $p = 1$	0.813	0.596	549	0.014	0.136	0.102	0.070
		2000	Random Acquisition	0.805	**0.600**	664	0.013	0.173	0.115	0.072
			Generation $p = 1$	***0.815***	0.596	**528**	**0.008**	**0.128**	**0.083**	**0.053**
25	8	1715	Random Acquisition	0.808	**0.606**	643	0.024	0.168	0.138	0.091
			Targeted Acquisition	***0.816***	0.599	522	0.013	0.136	0.094	0.061
			Generation $p = 1$	0.808	0.587	573	0.020	0.133	0.100	0.073
		4000	Random Acquisition	0.808	0.606	634	0.034	0.171	0.145	0.115
			Generation $p = 1$	0.813	0.582	**493**	**0.005**	**0.112**	**0.071**	**0.042**

This confirms that the overall improvement of the model should not be assessed solely based on aggregate metrics but also by considering the impact on critical subgroups. In general, the best performance is achieved with the Generation Acquisition Method when $p = 1$, provided that the number of inserted data points is sufficiently high.

As the proposed solution is model-agnostic, it is expected that the observations made for the Decision Tree can be generalized to any model. To confirm this, an analysis is conducted on Gradient Boosting, Logistic Regression, and Random Forest.

Table 5.9 presents a comparative analysis of different models—Gradient Boosting (GB), Logistic Regression (LR), and Random Forest (RF)—to evaluate the impact of bias mitigation strategies.

Table 5.9: Comparison of results for different models. Metric: False Positive , K:20%, Minimum Support: 20%, and Pruning: 1%.
Note: For each model type (GB, LR, RF), the best results for each metric are marked in **bold**.

Model	**# Samples**	**Approach**	**Accuracy**	**F1-Score**	**# FP**	Δ_{avg}	Δ_{max}	Δ_{10}	Δ_{20}
GB	-	Original	0.865	0.681	247	0.022	0.117	0.092	0.055
	2000	Random Acquisition	**0.866**	**0.684**	237	0.022	0.113	0.087	0.053
		Generation $p = 1$	0.861	0.645	158	0.007	0.069	0.038	0.007
	4000	Random Acquisition	**0.866**	**0.684**	242	0.018	0.112	0.079	0.041
		Generation $p = 1$	0.858	0.622	**109**	**0.006**	**0.048**	**0.029**	**0.006**
LR	-	Original	**0.809**	**0.474**	234	0.018	0.144	0.077	0.052
	2000	Random Acquisition	**0.809**	**0.474**	246	0.018	0.144	0.077	0.052
		Generation $p = 1$	0.800	0.369	117	0.005	0.036	0.020	0.009
	6000	Random Acquisition	0.808	**0.474**	181	0.019	0.162	0.080	0.055
		Generation $p = 1$	0.791	0.304	**12**	**0.004**	**0.030**	**0.010**	**0.004**
RF	-	Original	**0.843**	**0.655**	426	0.081	0.204	0.179	0.167
	2000	Random Acquisition	0.842	0.653	420	0.071	0.183	0.164	0.149
		Generation $p = 1$	**0.843**	0.632	335	0.026	0.134	0.085	0.056
	6000	Random Acquisition	0.842	0.653	415	0.073	0.200	0.178	0.164
		Generation $p = 1$	0.840	0.611	**251**	**0.008**	**0.099**	**0.044**	**0.014**

The table reports accuracy, F1-score, False Positives, and divergence metrics (Δ_{avg}, Δ_{max}, Δ_{10}, Δ_{20}) before and after mitigation while maintaining a minimum support of 20%, a 20% threshold for problematic subgroups, and

pruning at 1%.

The results confirm that increasing the number of generated samples consistently enhances bias mitigation. Across all models, the Generation strategy ($p = 1$) achieves the most significant reduction in False Positives and divergence metrics. For GB, using 4000 generated samples reduces False Positives from 247 to 109, with Δ_{max} decreasing from 0.117 to 0.048. Similarly, for LR, False Positives drop from 234 to 12, and Δ_{max} is minimized to 0.030 with 6000 generated samples. In RF, the number of False Positives decreases from 426 to 251, and Δ_{max} is reduced to 0.099 with the same approach.

While Random Acquisition achieves moderate reductions, its effectiveness is lower than that of Generation, particularly at higher sample sizes. The results reinforce that synthetic data augmentation through SMOTE-NC is the most effective strategy for mitigating bias, minimizing subgroup disparities, and improving overall fairness across different models.

5.2.3.2 Adult FP Mitigation Main Outcomes

The analysis demonstrates that bias mitigation using SMOTE-NC effectively reduces False Positives, particularly when synthetic points are predominantly assigned to class 0. The extent of this reduction depends on the number of problematic subgroups considered and the minimum support threshold used.

Key observations:

1. The Generation strategy, proposed in this study, proves to be the most effective approach when a sufficiently large number of synthetic samples are injected into the training set.

2. Targeted Acquisition outperforms other methods when the number of additional samples is limited, highlighting its suitability for early-stage data augmentation.

3. A higher minimum support threshold leads to a more effective reduction in False Positives and divergence, suggesting that addressing less fragmented, more representative subgroups enhances mitigation effectiveness.

4. While Accuracy and F1-score provide a general performance overview, they do not fully capture fairness improvements at the subgroup level. A slight decrease in F1-score in some cases aligns with a stronger reduction in divergence, emphasizing the need for fairness-aware evaluation metrics.

5. The best overall performance is achieved using the Generation Acquisition strategy when a sufficiently large number of samples are included.

This confirms that bias mitigation strategies must be carefully tailored based on the dataset characteristics and the balance between fairness and model Accuracy.

5.2.3.3 Adult False Negative Mitigation

In this section, the experiments from the previous section are repeated, but with a focus on a different metric. Specifically, the objective here is to mitigate False Negatives.

In the previous section, the experiments were conducted by comparing results between a low and a high support. This comparison can be replicated in this section; however, the number of subgroups with low support is extremely high in this case. For instance, when $min\_sup = 0.02$, the identified subgroups are 49,692. After applying pruning with $\epsilon = 0.01$, this number is reduced to 11,554 and 4,102 subgroups have positive divergence and a t-test > 2. This large number of problematic subgroups makes the mitigation process unstable if we aim to replicate the exact approach used previously. The instability arises because for example, the divergence values for the first 20 subgroups and the first 40 subgroups are very similar. This similarity in divergence makes it challenging to assess the differences between subgroups or to rank them meaningfully, complicating the comparison and the mitigation process.

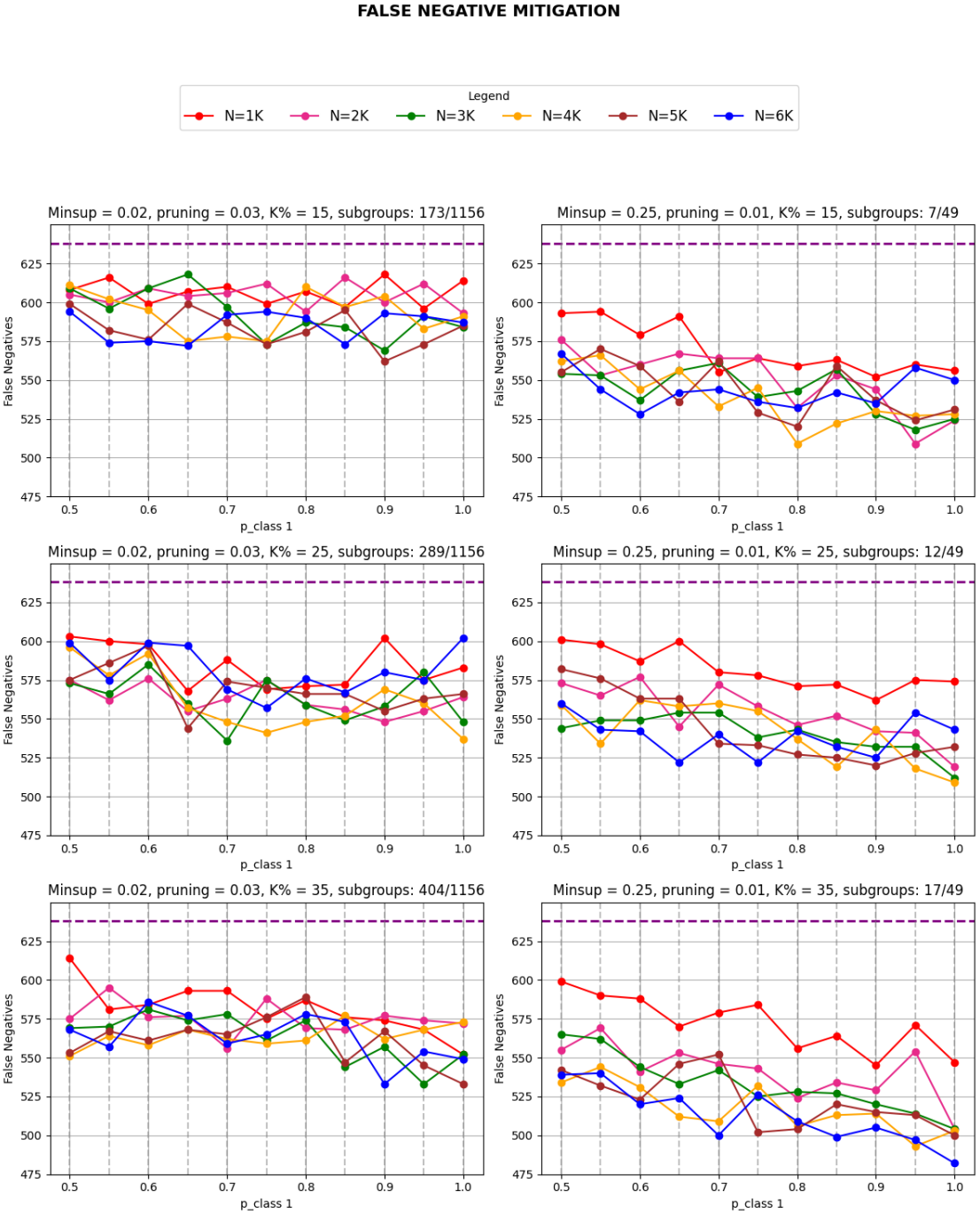
Therefore, there are two possible ways to reduce the number of problematic subgroups: adjusting the pruning parameter, epsilon, or performing the mitigation starting from subgroups identified with a higher minimum support threshold.
If the first strategy is adopted, bias mitigation does not occur effectively. Specifically, considering that in this case, mitigation is expected when problematic instances are included in the training set, provided their label is 1, it is observed that for an epsilon value equal or greater than 0.01 and a support of 2%, the problematic instances with label 1, from which SMOTE-NC should generate new class 1 data, are not sufficient in number to represent an adequate variety of feature-value pairs, thus rendering the mitigation ineffective. The results of the mitigation attempt with low support can be visualized in Table 5.10, from this table, it can be observed that it is not possible to determine which method is the most effective for mitigation. In fact, the divergence performance often worsens after adding new data. To analyze the trend of False Negatives as the previously described parameters vary, a pattern similar to that observed for False Positives in the previous subsection emerges. The Figure 5.2 illustrates the impact of bias mitigation on False Negatives using SMOTE-NC, applied to problematic subgroups identified with DivExplorer, for a Decision Tree model.

The x-axis represents the probability that synthetic points belong to class 1, while the y-axis indicates the number of False Negatives. To perform the mitigation, synthetic points (1K-6K) are injected into the training set to balance the data and reduce False Negatives.

Table 5.10: Comparison of Results for Targeted Data Acquisition, Random Data Acquisition and SMOTE-NC Data Generation Approaches for Decision Tree Model, metric: False Negative, MinimumSupport:2%, and Pruning:0.03.
Note: For each *% K* (15, 20, 25), the best results for each metric within the same sample size are marked with underline, while the overall best results are marked in **bold**.If nothing is in bold, then the metric is worse than the initial one.

% K	# Sub-groups	# Samples	Approach	Accuracy	F1-Score	# FN	Δ_{avg}	Δ_{max}	Δ_{10}	Δ_{20}
-	-	-	Original	0.803	0.593	638	0.236	0.593	0.593	0.593
15	173	219	Random Acquisition	0.804	**0.596**	627	0.219	0.600	0.600	0.600
			Targeted Acquisition	0.797	0.586	633	0.224	0.596	0.596	0.596
			Generation $p = 1$	**0.805**	0.595	634	0.234	0.594	0.594	0.594
		5000	Random Acquisition	0.804	**0.596**	631	0.212	0.598	0.598	0.598
			Generation $p = 1$	0.786	0.586	**585**	**0.201**	0.627	0.627	0.627
20	231	320	Random Acquisition	**0.810**	**0.604**	626	0.218	0.601	0.601	0.601
			Targeted Acquisition	0.794	0.587	618	**0.214**	0.606	0.606	0.606
			Generation $p = 1$	0.808	0.598	635	0.223	0.595	0.595	0.595
		6000	Random Acquisition	**0.810**	**0.604**	606	0.235	0.614	0.614	0.614
			Generation $p = 1$	0.782	0.584	**574**	0.222	0.634	0.634	0.634
25	289	373	Random Acquisition	**0.811**	**0.606**	622	0.229	0.603	0.603	0.603
			Targeted Acquisition	0.792	0.587	607	0.216	0.613	0.613	0.613
			Generation $p = 1$	0.801	0.585	654	0.226	**0.583**	**0.583**	**0.583**
		6000	Random Acquisition	**0.811**	**0.606**	606	0.235	0.614	0.614	0.614
			Generation $p = 0.8$	0.778	0.572	**602**	**0.205**	0.616	0.616	0.616

Figure 5.2: False Negative trend generated with SMOTE-NC (1K-6K) as $p_{\text{class 1}}$ varies for a Decision Tree. On the left, min_sup = 2% and pruning parameter = 3% ; on the right, min_sup = 25% and pruning parameter = 1%. Each row compares results for the same percentage of problematic subgroups used in mitigation.



The experiments are conducted by varying the number of problematic subgroups used to identify problematic instances in the validation set, allowing an analysis of how different subgroup selections influence the effectiveness of the mitigation strategy. On the left side of the figure, the minimum support is lower (2%), whereas on the right side, it is higher (25%). From top to bottom, the number of subgroups considered in the problematic instances increases as a larger percentage of subgroups are incorporated into the mitigation strategy.

These variations in support and subgroup selection significantly impact the number of False Negatives. When the support is low and fewer subgroups are considered, the number of False Negatives tends to be higher. This is because a broader coverage, resulting from a lower minimum support, introduces more variability and noise into the synthetic data, making the mitigation process less effective. Conversely, a higher minimum support focuses on the most representative and frequent patterns, reducing noise and minimizing the risk of overfitting to specific subgroups. As a result, the number of False Negatives tends to decrease as the probability of generating a class 1 instance increases. This aligns with the False Negative bias mitigation strategy, where the probability of assigning a sample to class 1 is intentionally increased to counteract the bias.

To gain a clearer understanding of the impact on other metrics, particularly divergence, the strategy of fixing the pruning parameter at 0.01 is adopted, and the performance are compared while varying the percentage of subgroups considered in the mitigation process for two relatively high values of minimum support.

It is important to note that keeping the pruning parameter at a higher value and setting the minimum support excessively high can lead to a significant imbalance between the number of instances labeled as class 0 and those labeled as class 1 in the validation set, specifically when filtering for instances that match problematic subgroups.

In fact, during experiments focused on False Negative Mitigation, it was observed that for certain (low) values of support and subgroup selection, the number of problematic class 1 instances was several orders of magnitude lower than that of class 0 instances. However, the generation strategy developed in this thesis relies on a sufficient number of original data points to generate synthetic ones effectively, ensuring that the mitigation process remains impactful.

Table 5.11: Comparison of Results for Targeted Data Acquisition, Random Data Acquisition and SMOTE-NC Data Generation Approaches for Decision Tree Model, metric: False Negative, MinimumSupport:25%, and Pruning:0.01.
Note: For each % K (15, 20, 25), the best results for each metric within the same sample size are marked with underline, while the overall best results are marked in **bold**.

% K	# Sub-groups	# Sam-ples	Approach	Accuracy	F1-Score	# FN	Δ_{avg}	Δ_{max}	Δ_{10}	Δ_{20}
-	-	-	Original	0.803	0.593	638	0.111	0.444	0.329	0.255
15	7	534	Random Acquisition	**0.807**	0.602	616	0.128	0.437	0.339	0.291
			Targeted Acquisition	0.799	**0.606**	564	0.095	0.422	0.277	0.225
			Generation $p = 0.85$	0.799	0.596	602	0.112	0.467	0.343	0.266
		4000	Random Acquisition	**0.807**	0.602	609	0.125	0.441	0.311	0.240
			Generation $p = 1$	0.786	0.599	**528**	**0.060**	**0.405**	**0.234**	**0.185**
20	9	557	Random Acquisition	**0.805**	0.599	649	0.129	0.455	0.341	0.290
			Targeted Acquisition	0.800	**0.609**	557	0.093	0.427	0.276	0.224
			Generation $p = 1$	0.799	0.601	586	0.079	0.408	0.267	0.223
		8000	Random Acquisition	**0.805**	0.609	601	0.134	0.423	0.329	0.276
			Generation $p = 0.8$	0.784	0.592	**548**	**0.054**	**0.409**	**0.240**	**0.179**
25	12	648	Random Acquisition	**0.805**	0.599	617	0.134	0.479	0.363	0.295
			Targeted Acquisition	0.800	**0.609**	553	0.088	0.389	0.256	0.207
			Generation $p = 1$	0.798	0.598	588	0.094	0.370	0.274	0.236
		5000	Random Acquisition	**0.805**	0.599	631	0.111	0.372	0.275	0.218
			Generation $p = 1$	0.782	0.594	**532**	**0.054**	**0.333**	**0.242**	**0.194**

For this reason, Tables 5.11 and 5.12 compare performance for those specific values of minimum support, pruning parameter, and percentage of subgroups considered, here the support threshold is set to 25% and 35%, respectively, with the pruning parameter fixed at $\epsilon = 0.01$.

By analyzing the two tables, it is evident that, given the same number of samples in the training set, the Generation strategy, when a sufficiently

large number of samples is added, is the most effective in reducing both divergence and the number of False Negatives.

Table 5.12: Comparison of Results for Targeted Data Acquisition, Random Data Acquisition and SMOTE-NC Data Generation Approaches for Decision Tree Model, metric: False Negative, MinimumSupport:35%, and Pruning:0.01.
Note: For each *% K* (15, 20, 25), the best results for each metric within the same sample size are marked with underline, while the overall best results are marked in **bold**.

% K	# Sub-groups	# Sam-ples	Approach	Accuracy	F1-Score	# FN	Δ_{avg}	Δ_{max}	Δ_{10}	Δ_{20}
-	-	-	Original	0.803	0.593	638	0.071	0.298	0.175	0.134
15	3	511	Random Acquisition	**0.806**	0.599	625	0.100	0.313	0.220	0.171
			Targeted Acquisition	0.800	0.606	564	0.067	0.234	0.195	0.159
			Generation $p = 1$	0.804	0.606	587	0.071	0.234	0.197	0.160
		3000	Random Acquisition	**0.806**	0.599	610	0.108	0.316	0.219	0.162
			Generation $p = 1$	0.797	**0.618**	**499**	**0.038**	**0.167**	**0.147**	**0.128**
20	4	534	Random Acquisition	**0.807**	0.602	616	0.101	0.306	0.246	0.190
			Targeted Acquisition	0.801	**0.608**	**561**	0.078	0.241	0.187	0.150
			Generation $p = 0.95$	0.799	0.598	599	0.063	0.225	0.199	0.157
		6000	Random Acquisition	**0.807**	0.602	606	0.077	0.313	0.204	0.134
			Generation $p = 0.8$	0.788	0.579	618	**0.062**	**0.179**	**0.143**	**0.106**
25	5	615	Random Acquisition	**0.806**	0.602	613	0.111	0.289	0.219	0.169
			Targeted Acquisition	0.799	**0.606**	561	0.069	0.230	0.161	0.126
			Generation $p = 1$	0.797	0.597	590	0.068	0.247	0.194	0.152
		6000	Random Acquisition	**0.806**	0.602	606	0.077	0.313	0.204	0.134
			Generation $p = 1$	0.783	0.598	**519**	**0.033**	**0.119**	**0.094**	**0.084**

However, when fewer samples are introduced, the Targeted Acquisition approach performs better in reducing False Negatives. This aligns with the objective of False Negative Mitigation, as all holdout points belonging to the

problematic subgroups and classified as class 1 are included. This confirms that a focused inclusion of samples from these problematic subgroups is especially advantageous in the early stages of data acquisition when new examples are limited, and each additional sample has a more significant impact on divergence reduction.

Another key observation is that, in the second table—where a minimum support of 35% is applied—divergence values are consistently lower than in the case with a minimum support of 25%. This suggests that when problematic subgroups are less fragmented and more representative, the False Negative issue is more contained, and divergence reduction is more effective, regardless of the adopted acquisition strategy.

Additionally, Accuracy and F1-score provide a broad measure of model performance but do not fully capture subgroup-level behavior. A high Accuracy or F1-score does not necessarily imply a better mitigation of divergence in problematic subgroups. Consequently, again, it is not surprising that strategies such as Generation Acquisition, particularly when a large number of samples is added, show a significant reduction in divergence while experiencing a slight decrease in F1-score compared to other approaches. This highlights the importance of evaluating model improvements not solely through aggregate metrics but also through their impact on critical subgroups.

In general, the Generation Acquisition Method achieves the best performance when , provided that a sufficiently large number of samples is introduced. However, Targeted Acquisition offers strong results when fewer samples are available, demonstrating its effectiveness in early-stage data augmentation strategies.

As before, since the proposed solution is model-agnostic, it is reasonable to assume that the observations made for the Decision Tree can be generalized to other models.

To confirm this hypothesis, an analysis was conducted on Gradient Boosting, Logistic Regression, and Random Forest. The key findings are summarized in Table 5.13, which presents the performance of different models while keeping the Minimum Support, the percentage of problematic subgroups in the holdout set, and pruning fixed, with specific values for each model.

Table 5.13: Comparison of results for different models, metric: False Negative. For GB K:10%, MinimumSupport: 2%, and Pruning: 5%. For LR K:15%, MinimumSupport: 40%, and Pruning:1%. For RF K:35%, MinimumSupport: 25%, and Pruning:1%. *Note:* For each model type (GB, LR, RF), the best results for each metric are marked in **bold**. If nothing is in bold, then the metric is worse than the initial one.

Model	**# Samples**	**Approach**	**Accuracy**	**F1-Score**	**# FP**	Δ_{avg}	Δ_{max}	Δ_{10}	Δ_{20}
GB	-	Original	0.865	0.681	630	0.281	0.598	0.547	0.475
	3000	Random Acquisition	**0.867**	0.686	624	0.279	0.602	0.529	0.453
		Generation $p = 0.8$	0.841	**0.690**	410	0.253	0.696	0.514	0.384
	6000	Random Acquisition	**0.867**	0.686	612	0.291	0.610	0.550	0.470
		Generation $p = 1$	0.812	0.663	**365**	**0.222**	0.658	**0.461**	**0.298**
LR	-	Original	0.809	0.474	1008	0.106	0.345	0.246	0.173
	2000	Random Acquisition	0.809	0.476	1005	0.104	0.347	0.245	0.175
		Generation $p = 0.8$	0.787	0.524	779	0.072	0.286	0.217	0.183
	4000	Random Acquisition	0.809	0.476	987	0.104	0.354	0.257	0.189
		Generation $p = 0.8$	0.754	**0.532**	**612**	**0.015**	**0.178**	**0.127**	**0.105**
RF	-	Original	0.843	0.655	598	0.127	0.490	0.373	0.328
	2000	Random Acquisition	0.843	0.655	594	0.146	0.494	0.413	0.362
		Generation $p = 1$	0.832	**0.661**	500	0.100	0.439	0.297	0.256
	6000	Random Acquisition	0.843	0.655	428	0.141	0.481	0.407	0.357
		Generation $p = 1$	0.821	0.653	**305**	**0.062**	**0.424**	**0.283**	**0.231**

The results indicate that increasing the number of generated samples leads to a substantial reduction in False Negatives across all models. For GB, using 6000 generated samples decreases False Negatives from 630 to 365, with Δ_{avg} reducing from 0.281 to 0.222. In LR, False Negatives drop from 1008 to 612 with 4000 generated samples, and Δ_{max} is reduced from 0.345 to 0.178. Similarly, for RF, False Negatives decrease from 598 to 305, while Δ_{avg} is minimized from 0.127 to 0.062.

Unlike the results observed for False Positives, the mitigation strategy in this case also leads to an improvement in the overall F1-score. This is

likely due to the initial dataset imbalance, where the minority class (label 1) was underrepresented, causing the models to struggle with recall. The introduction of synthetic samples helps address this imbalance, leading to better overall predictive performance.

These findings reinforce that increasing the number of synthetic samples effectively mitigates False Negatives and reduces subgroup discrepancies. Furthermore, the observed improvement in F1-score suggests that the mitigation strategy not only enhances fairness but also improves the model's ability to correctly classify minority class instances.

5.2.3.4 Adult FN Main Outcomes

The analysis demonstrates that bias mitigation using SMOTE-NC effectively reduces False Negatives, particularly when synthetic points are predominantly assigned to class 1. However, the extent of this reduction is strongly influenced by data imbalance, the number of problematic subgroups considered, and the minimum support threshold. Key observations:

1. Data imbalance poses a major challenge: when the number of class 1 instances is significantly lower than class 0 (e.g., for support = 30%, epsilon = 1% ,and 5% of problematic subgroup considered, 52 instances of class 1 vs. 1,895 of class 0), SMOTE-NC struggles to generate effective synthetic samples, limiting the mitigation effectiveness.

2. A high number of problematic subgroups, especially when using a low support threshold, reduces the effectiveness of the mitigation due to increased variability and noise in the synthetic data.

3. For higher support values, the Generation strategy achieves better mitigation, even when a lower number of synthetic samples are added, as it focuses on the most representative and frequent patterns.

4. As in the case of False Positives, Accuracy and F1-score are not reliable indicators for evaluating bias mitigation. Instead, divergence metrics should be prioritized, as they better capture subgroup-level behavior and fairness improvements.

The best overall performance is achieved using the Generation Acquisition strategy when a sufficiently large number of samples are included. However, Targeted Acquisition remains effective in early-stage data augmentation, particularly when fewer samples are available.

5.2.3.5 Adult Error Mitigation

At the end, for the Adult Dataset, the mitigation in this section is performed with respect to the Error Metric.

In this case, the analysis is conducted by exploring the entire range of probabilities for class 0, from 0 to 1. This methodological choice differs from other approaches where False Positives and False Negatives are examined separately. In those cases, the probability of class 0 is varied from 0.5 to 1 for False Positives, while the same interval is applied to class 1 probability for False Negatives.

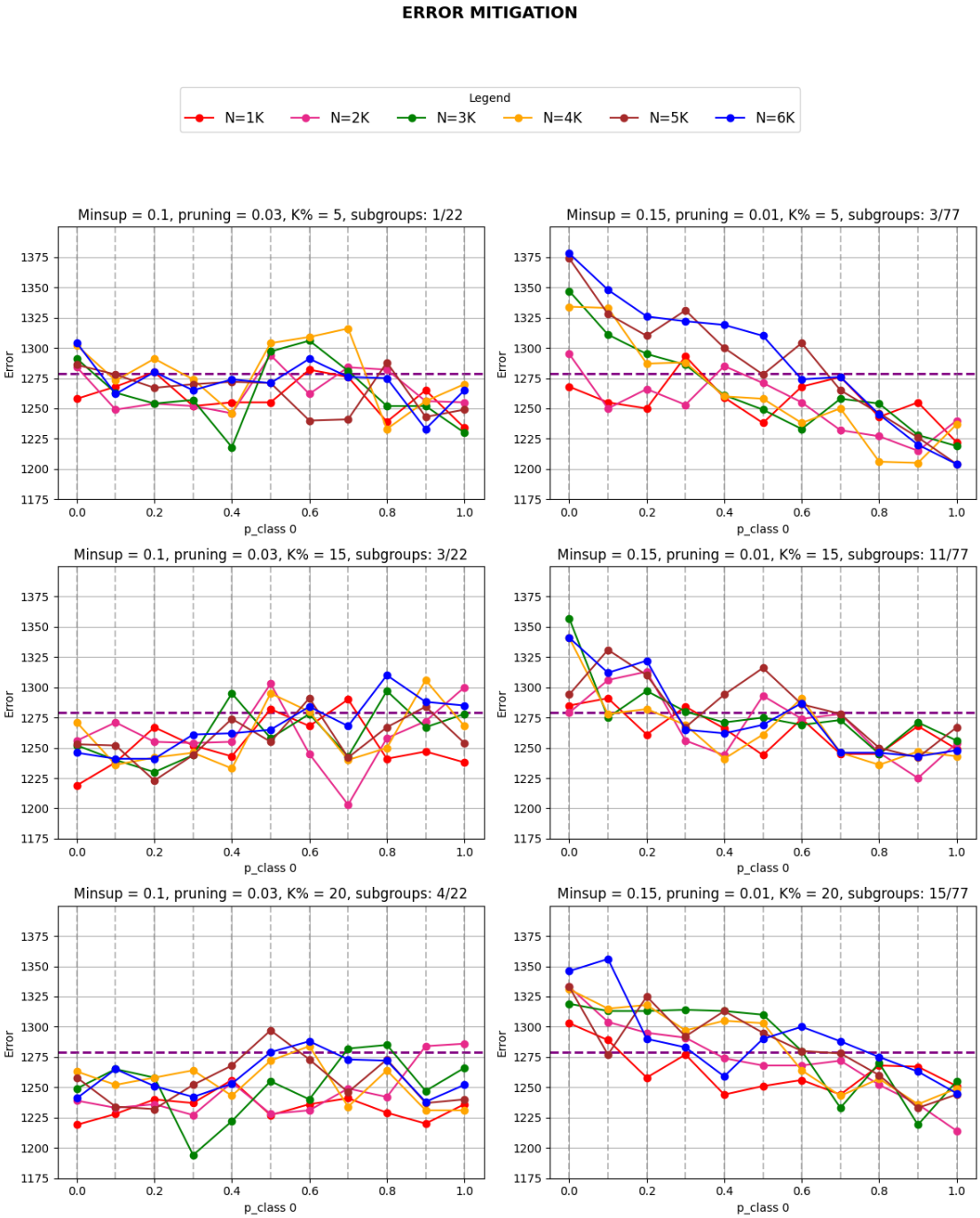
However, in this part of study, the focus is not on distinguishing between False Positives and False Negatives but rather on analyzing problematic subgroups as a whole—those in which the model makes a high number of Errors, regardless of their nature. Since it is not known in advance whether the predominant type of Error in these subgroups arises from overestimation or underestimation of the true class, it is necessary to explore the full probability spectrum. This ensures a comprehensive coverage of all possible Error configurations.

As done previously, the first step is to observe how the total number of Errors changes after applying the mitigation strategy proposed in this thesis.

For this reason the Figure 5.3 illustrates the trend of Errors on the test set following the addition of a variable number of new instances to the training set, specifically targeting subgroups where the classification model exhibits the most difficulty.

Recall that the dashed line in each panel represents the initial number of Errors, given by the sum of False Positives and False Negatives. The x-axis indicates the probability that a synthetic point belongs to class 0; since this is a binary classification problem, the complementary value represents the probability of belonging to class 1. The number of newly added instances increases progressively from left to right, while the number of problematic subgroups considered increases from top to bottom.

Figure 5.3: Error trend generated with SMOTE-NC (1K-6K) as $p_{\text{class }0}$ varies for a Decision Tree. On the left, min_sup = 10% and pruning parameter = 3% ; on the right, min_sup = 15% and pruning parameter = 1%. Each row compares results for the same percentage of problematic subgroups used in mitigation.



Two additional parameters influence the process: the minimum support, which defines the threshold of data required for a subgroup to be considered relevant, and pruning, which removes subgroups whose divergence from the model's overall behavior falls below a certain threshold.

On the left side of the figure, the minimum support is lower (10%) and the pruning parameter is higher (3%) whereas on the right side, the minimum support is higher (15%) and the pruning paramerer is lower (1%). From top to bottom, the number of subgroups considered in the problematic instances increases as a larger percentage of subgroups are incorporated into the mitigation strategy. Analyzing the graph reveals several key trends. In general, adding data from problematic subgroups contributes to a reduction in Errors compared to the baseline indicated by the dashed line. This effect becomes more pronounced in panels where a greater number of subgroups are considered, suggesting that broader coverage of the areas where the model struggles leads to improved performance. Another clear trend is that a higher minimum support results in a more significant reduction in Errors. This indicates that including more representative subgroups—those with a larger amount of data—has a stronger impact on the model's ability to generalize effectively to the test set instances.

An interesting aspect concerns the effect of pruning. In the panels on the right, where a higher pruning threshold is applied, the Error reduction appears more consistent and pronounced compared to the panels on the left. This suggests that removing subgroups with low divergence can be beneficial, preventing the addition of instances to the training set that do not contribute meaningfully to Error mitigation. However, excessive pruning may lead to the loss of relevant information, potentially reducing the effectiveness of the method.

Finally, the analysis of the x-axis shows that Errors tend to decrease more consistently in the rightmost panels as the probability of class 0 increases. This suggests that generating new data with a more pronounced distribution toward one of the two classes can positively impact the model's ability to correct its Errors.

In summary, the addition of synthetic instances belonging to problematic subgroups is an effective strategy to improve the performance of the model, as in the case of Error Metric, provided that the selection of these subgroups is made in a targeted manner, prioritizing those with sufficient support and a meaningful divergence from the general behavior of the classifier.

The newly described image provides insights into how the number of Errors in the test set changes when new synthetic data is added. However, they do not offer any information about divergence performance. To investigate this aspect, Tables 5.14 and 5.15 are used.

Table 5.14: Comparison of Results for Targeted Data Acquisition, Random Data Acquisition and SMOTE-NC Data Generation Approaches for Decision Tree Model, metric: Errors, MinimumSupport:10%, and Pruning:3%.
Note: For each % K (5, 15, 20), the best results for each metric within the same sample size are marked with underline, while the overall best results are marked in **bold**.If nothing is in bold, then the metric is worse than the initial one.

% K	# Sub-groups	# Samples	Approach	Accuracy	F1-Score	# ER	Δ_{avg}	Δ_{max}	Δ_{10}	Δ_{20}
-	-	-	Original	0.803	0.593	1279	0.026	0.201	0.167	0.123
5	1	704	Random Acquisition	0.806	**0.601**	1262	0.019	0.198	0.160	0.120
			Targeted Acquisition	0.808	**0.601**	1248	0.014	0.206	0.156	0.113
			Generation $p = 0.1$	**0.809**	0.599	1245	0.044	0.225	0.170	0.132
		4000	Random Acquisition	0.806	**0.601**	**1243**	0.013	0.189	0.150	0.111
			Generation $p = 0.5$	0.800	0.589	1304	**0.002**	0.214	**0.138**	**0.097**
15	3	2551	Random Acquisition	0.810	0.603	1280	0.025	0.196	0.154	0.118
			Targeted Acquisition	0.803	0.595	**1235**	0.015	0.209	0.146	0.100
			Generation $p = 0.8$	0.805	0.586	1270	0.026	0.224	0.166	0.121
		4000	Random Acquisition	0.803	0.595	1243	0.013	**0.189**	0.150	0.111
			Generation $p = 0.5$	0.801	0.593	1295	**0.003**	0.207	**0.122**	**0.076**
20	4	2817	Random Acquisition	**0.814**	**0.617**	**1209**	0.018	0.197	0.136	0.100
			Targeted Acquisition	0.810	0.607	1239	0.023	0.191	0.145	0.108
			Generation $p = 0.8$	0.803	0.592	1283	0.023	0.223	0.160	0.119
		4000	Random Acquisition	**0.814**	**0.617**	1243	0.013	0.189	0.150	0.111
			Generation $p = 0.2$	0.807	0.607	1258	**0.001**	**0.180**	**0.127**	**0.095**

Table 5.15: Comparison of Results for Targeted Data Acquisition, Random Data Acquisition and SMOTE-NC Data Generation Approaches for Decision Tree Model, metric: Errors, MinimumSupport:15%, and Pruning:1%.
Note: For each % K (5, 15, 20), the best results for each metric within the same sample size are marked with underline, while the overall best results are marked in **bold**.

% K	# Sub-groups	# Sam-ples	Approach	Accuracy	F1-Score	# ER	Δ_{avg}	Δ_{max}	Δ_{10}	Δ_{20}
-	-	-	Original	0.803	0.593	1279	0.079	0.403	0.370	0.360
5	227	1710	Random Acquisition	0.808	0.606	1250	0.085	0.422	0.380	0.367
			Targeted Acquisition	0.806	0.601	1265	0.080	0.397	0.375	0.362
			Generation $p = 0.8$	0.808	0.596	1249	0.077	0.382	0.374	0.365
		3000	Random Acquisition	**0.808**	**0.606**	**1242**	0.076	0.415	0.361	0.346
			Generation $p = 0.5$	0.806	0.603	1264	**0.075**	**0.363**	**0.345**	**0.335**
15	681	2750	Random Acquisition	0.811	**0.611**	1229	0.074	0.372	0.351	0.339
			Targeted Acquisition	0.812	0.607	1224	0.088	0.403	0.375	0.364
			Generation $p = 1$	0.805	0.577	1268	0.081	0.411	0.383	0.368
		2000	Random Acquisition	0.808	**0.611**	1275	0.079	0.384	0.364	0.358
			Generation $p = 1$	**0.813**	0.594	**1216**	**0.077**	**0.374**	**0.351**	**0.341**
20	909	2962	Random Acquisition	0.808	0.602	1250	**0.074**	**0.384**	**0.352**	**0.342**
			Targeted Acquisition	**0.810**	**0.608**	1235	0.089	0.449	0.419	0.403
			Generation $p = 1$	**0.810**	0.592	1238	0.084	0.400	0.377	0.363
		5000	Random Acquisition	0.808	0.602	**1026**	0.089	0.378	0.369	0.363
			Generation $p = 0.5$	0.802	0.603	1058	0.089	0.388	0.356	0.345

These Tables compare the three data acquisition strategies described at the beginning of the chapter when using the Decision Tree model. The comparison considers an increasing number of problematic subgroups, leading to a progressively larger training set. In the first table, the minimum support is set to 10%, while the pruning parameter is fixed at 3%. In the second table, the minimum support is increased to 15%, while pruning is set to 1%.

The results indicate that the use of Generation Acquisition with SMOTE-NC has a complex impact on Errors, which must be carefully analyzed. In particular, the reduction of total Errors is less evident compared to the decrease in False Positives or False Negatives individually. This occurs because total Error is a combination of both components: improving one aspect can worsen the other, making the overall balance less effective.

One key factor to consider is the effect of increasing the percentage of problematic subgroups (% K). When K increases, the Generation Acquisition strategy introduces synthetic samples into a broader range of subgroups, enhancing the dataset's diversity. However, this diversification can have contrasting effects: in some cases, the model benefits from the enriched dataset and improves its generalization capabilities, while in others, the introduction of synthetic data leads to variations that the model struggles to interpret correctly, increasing divergence from the original distribution. The net effect depends on the model's ability to adapt to the new information, and in many scenarios, adding generated data reduces Errors within specific subgroups but does not significantly decrease overall Errors.

A similar mechanism occurs when increasing the minimum support threshold for selecting problematic subgroups. A higher threshold means that the model focuses on larger, more representative subgroups of the overall data distribution, excluding rarer ones. While this can enhance learning stability, it also reduces the model's ability to correct Errors in minority subgroups, which may negatively impact the overall Error mitigation capability. Specifically, if the model is already biased toward a particular type of Error (e.g., producing more False Negatives than False Positives), using SMOTE-NC may amplify this tendency, improving one metric while worsening another.

This phenomenon explains why, in many cases, Generation Acquisition with SMOTE-NC does not necessarily lead to a reduction in total Error, despite improving False Positives or False Negatives individually. Unlike Random Acquisition, which introduces new points uniformly, Targeted Acquisition selects specific samples from the holdout set that belong to problematic subgroups.

However, this strategy can have unintended consequences: if the added samples have labels that do not help the model correct its decisions, they

can reinforce existing divergences and degrade overall performance.
Generation Acquisition, on the other hand, actively modifies the data distribution by generating new points, influencing the model's behavior in unpredictable ways. For Accuracy and F1-Score, the same considerations apply as those made in the False Negative and False Positive Mitigation.

These results suggest that Error mitigation through data augmentation techniques like SMOTE-NC requires careful attention to the balance between False Positives and False Negatives. The choice of K and the minimum support threshold can significantly impact the effectiveness of the intervention, and an overly aggressive optimization of a single aspect may compromise the overall classification balance.

To verify whether the observations made for the Decision Tree hold for other models, Table 5.16 presents a comparative analysis of different models—Gradient Boosting (GB), Logistic Regression (LR), and Random Forest (RF)—to evaluate the impact of data augmentation strategies on overall classification Errors. The table reports accuracy, F1-score, total Errors, and divergence metrics (Δ_{avg}, Δ_{max}, Δ_{10}, Δ_{20}) before and after mitigation, using model-specific parameters for K, minimum support, and pruning.

The results highlight that data augmentation has varying effects depending on the model and the chosen augmentation method. In the case of GB, Random Acquisition with 2000 samples yields slight improvements in Accuracy and F1-Score, while marginally reducing the total number of Errors. Similarly, using Generation Acquisition with 5000 samples ($p = 0.5$) decreases Δ_{avg} and Δ_{10}, suggesting a positive impact on subgroup discrepancies.
However, the overall number of Errors remains stable, indicating that augmentation does not always directly enhance all metrics.
For LR, the use of Generation Acquisition with 5000 samples increases the F1-Score from 0.474 to 0.578, but at the cost of a higher number of total Errors. This suggests that while the model becomes more effective at capturing the minority class, it also introduces more misclassifications overall.

Table 5.16: Comparison of results for different models, metric: Error.
For GB K:25%, MinimumSupport: 10%, and Pruning: 1%. For LR K:20%, MinimumSupport: 10%, and Pruning:3%. For RF K:5%, MinimumSupport: 15%, and Pruning:1%.
Note: For each model type (GB, LR, RF), the best results for each metric are marked in **bold**. If nothing is in bold, then the metric is worse than the initial one.

Model	**# Samples**	**Approach**	**Accuracy**	**F1-Score**	**# ER**	Δ_{avg}	Δ_{max}	Δ_{10}	Δ_{20}
GB	-	Original	0.865	0.681	877	0.063	0.195	0.176	0.164
	2000	Random Acquisition	**0.866**	**0.684**	**868**	0.066	0.195	0.180	0.169
		Generation $p = 0.2$	0.860	0.704	910	0.062	**0.186**	0.174	0.167
	5000	Random Acquisition	0.865	0.681	874	0.063	0.195	0.176	0.164
		Generation $p = 0.5$	0.864	0.699	883	**0.056**	0.189	**0.169**	**0.160**
LR	-	Original	0.809	0.474	1242	0.040	0.245	0.206	0.166
	2000	Random Acquisition	0.809	0.478	1251	0.034	0.242	0.202	0.158
		Generation $p = 0.2$	0.803	0.547	1284	0.030	0.216	0.152	0.123
	5000	Random Acquisition	0.809	0.478	1250	0.040	0.248	0.204	0.165
		Generation $p = 0.2$	0.789	**0.578**	1371	**0.011**	**0.214**	**0.129**	**0.103**
RF	-	Original	0.843	0.655	1024	0.069	0.195	0.182	0.174
	2000	Random Acquisition	0.839	0.648	**1014**	0.070	**0.191**	0.178	0.170
		Generation $p = 1$	0.838	**0.623**	1047	**0.063**	0.199	0.182	0.172
	6000	Random Acquisition	0.839	0.648	1043	0.067	0.197	0.186	0.177
		Generation $p = 1$	0.838	0.604	1058	0.064	0.195	**0.175**	**0.165**

Similarly, for RF, Random Acquisition with 2000 samples slightly reduces Errors, whereas Generation Acquisition ($p = 1$) leads to variations in Error distribution, with improvements in Δ_{10} and Δ_{20}, but no clear overall gain.

These findings reinforce that the effectiveness of synthetic sample generation depends on the model and the specific parameter configuration.

Unlike the case of False Negatives, where augmentation consistently improved results, the impact on total Errors is less predictable. This is likely

due to the initial class imbalance: before augmentation, the dataset contained a limited number of positive instances (label 1), making the models prone to underfitting the minority class. As a result, while augmentation helps rebalance the dataset, it can also introduce new classification Errors, emphasizing the need for careful tuning to maximize fairness without compromising overall model performance.

5.2.3.6 Adult Error Main Outcomes

The analysis demonstrates that bias mitigation using SMOTE-NC effectively impacts overall Error reduction, particularly in relation to the balance between False Positives and False Negatives. The effectiveness of this approach depends on the number of synthetic samples generated, the subgroup selection criteria, and specific parameter configurations.
Key observations:

1. The Generation Acquisition through SMOTE-NC has complex effects on Errors. It improves False Positives or False Negatives separately, but it doesn't always lead to an overall reduction in total Error. This is because improving one type of Error may worsen the other.

2. Increasing the number of subgroups (% K) and adjusting the minimum support threshold impacts data distribution. A higher value of % K increases diversity, but it can also lead to greater divergence. A higher support threshold improves learning in larger subgroups but limits Error correction in minority groups.

3. Random Acquisition adds samples uniformly, while Targeted Acquisition focuses on problematic subgroups. However, the latter can worsen the situation if the added samples do not help in Error correction.

4. The main challenge is to maintain a balance between False Positives and False Negatives. If not balanced properly, addressing one type of Error may exacerbate the other.

5. Difficulty in Finding Optimal Parameters for Error Mitigation: For False Positives and False Negatives, it was easier to find parameters that improve bias mitigation, with good results across various settings. In contrast, optimizing for total Error was more complex, as small adjustments to parameters (%K, pruning, min support) had a significant impact on all metrics.

6. Accuracy and F1-score primarily reflect overall model performance and do not provide insights into subgroup-level behavior. These metrics may not reveal important issues within specific subgroups, making them less reliable for evaluating fairness or bias mitigation in detail.

In conclusion optimizing total Error requires more fine-tuned attention and a more challenging parameter search.

5.2.4 COMPAS Experimental Settings and Results

The preprocessed COMPAS Dataset consists of a total of 18,293 instances. These are divided as follows: 40% (7,317 instances) is allocated to the train set, while the remaining 60% is evenly split among the test set (3,659), validation set (3,658), and held-out set (3,659), with 20% each.

Table 5.17: Number of subgroups, number of problematic subgroups, one of the most divergent one with different minimum support and different the metrics for DT model.

Metric	Minimum Support	# Subgroups	#Divergent Subgroups	Most Divergent Subgroup
FP	5%	1149	54	{sex = Male, Violent Risk Level = Medium, Violent Recidivism Risk = 7}
	10%	349	16	{Violent Risk Level = Medium, Risk Level = High}
	15%	154	10	{Risk Level=High, race=African-American }
FN	5%	1149	416	{Prior Offenses = [0-5], Violent Risk Level = Low, Juvenile offenses = 0}
	10%	349	223	{Prior Offenses = [0-5], Violent Risk Level = Low, Juvenile Offenses = 0)}
	15%	154	109	{Prior Offenses = [0-5], Violent Risk Level = Low, Risk Level = Low}
ER	5%	1149	96	{Recidivism Risk = 10, sex = Male, Race = African-American}
	10%	349	23	{Prior Offenses = [6-10], Race = African-American, sex = Male}
	15%	154	10	{ sex = Male, Prior Offenses = [6-10]}

The Table 5.17 presents the most divergent subgroups concerning the False Positive, False Negative, and Error Metrics, along with their respective divergence values for the Decision Tree model.

The analysis considers different minimum support values, ranging from 5% to 15% of the validation set.

In the experiments conducted using the COMPAS dataset, the pruning parameter was set to 0, meaning no pruning was applied. This choice is justified by the relatively small number of instances in the dataset compared to the one used previously. As a result, it is expected that the number of identified subgroups will already be limited, and applying pruning could further reduce them to a point where they may not be sufficient for bias mitigation.

This assumption is confirmed by the results shown in the table: as the minimum support increases, the number of problematic subgroups decreases significantly. It is important to note that the total number of subgroups for a given support value remains the same across all three metrics since it depends solely on the support threshold. However, the number of problematic subgroups varies for each metric, as divergence is influenced not only by support but also by the specific metric used for evaluation.

Moreover, such a table provides a detailed overview of the problematic subgroups across different metrics. From this perspective, it becomes clearer what interpretable subgroups represent: since the subgroups are not encoded numerically, they are easily understandable. This makes it reasonable to assume, for example, that the subgroup {Recidivism Risk = 10, Sex = Male, Prior Offenses = [6-10]} is more likely to be subject to False Positives, just as the subgroup {Prior Offenses = [0-5], Violent Risk = Low, Risk Level = Low} is more likely to be subject to False Negatives.

In other words, the *interpretability* of these subgroups allows us to directly observe the characteristics of individuals most affected by classification Errors. Unlike purely numerical representations, categorical and structured attributes highlight meaningful patterns, making it easier to identify biases and disparities in the model's predictions.

After analyzing the subgroups experimentally, the following pages will provide details of the mitigation process for the COMPAS dataset.

5.2.4.1 COMPAS False Positive Mitigation

As a preliminary analysis of the solution's impact on False Positives por COMPAS dataset, Figure 5.4 is examined.

This image illustrates the impact of bias mitigation on False Positives using SMOTE-NC, applied to problematic subgroups identified with DivExplorer for a Decision Tree model. The x-axis represents the probability that synthetic points belong to class 0, ranging from 0.5 to 1. This aligns with the idea that, in mitigating False Positives, it is necessary to increase the number of synthetic data points representing problematic instances labeled as 0. The y-axis indicates the number of False Positives.

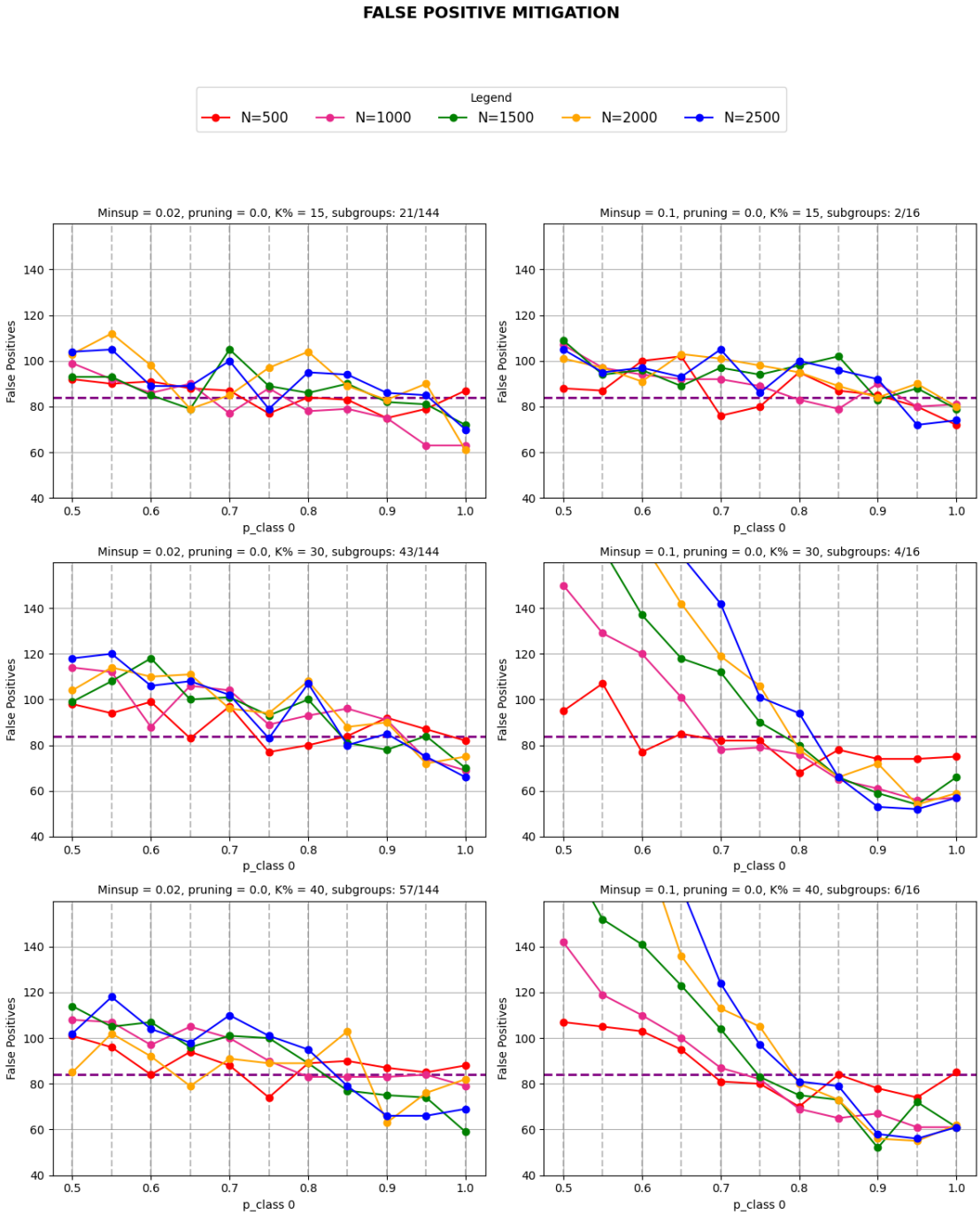
To perform the mitigation, synthetic samples (ranging from 0.5K to 2.5K) are injected into the training set to balance the data distribution and reduce False Positives. The experiments are conducted by varying the number of problematic subgroups used to identify problematic instances in the validation set. This approach allows for analyzing how different subgroup selections influence the effectiveness of the mitigation strategy.

On the left side of the figure, a lower minimum support threshold is applied, enabling the identification of a larger number of problematic subgroups. Conversely, on the right side, a higher support threshold results in fewer detected subgroups. Moving from top to bottom in the figure, the number of subgroups selected for data augmentation increases. From the analysis of the figure, we observe key trends in how the number of False Positives varies depending on the number of problematic subgroups and the minimum support threshold used in DivExplorer.

As the number of selected subgroups increases (moving from top to bottom in the figure), the mitigation effect generally improves, leading to a greater reduction in False Positives, especially for higher $p_{\text{class }0}$ values. This suggests that generating synthetic data from a larger set of problematic subgroups helps balance the dataset more effectively. However, when too many subgroups are included, the impact may stabilize or even decrease due to excessive data variability.

Comparing the left and right columns, a lower minimum support threshold (left) allows for identifying more subgroups, but many of them are less representative, leading to a less effective reduction in false positives. In contrast, with a higher minimum support threshold (right), fewer but more representative subgroups are selected, resulting in a more significant reduction in false positives.

Figure 5.4: False Positives trend generated with SMOTE-NC (0.5K-2.5K) as $p_{\text{class }0}$ varies for a Decision Tree. On the left, min_sup = 2%; on the right, min_sup = 10%; for both pruning parameter = 0%. Each row compares results for the same percentage of problematic subgroups used in mitigation.



In almost all cases, increasing $p_{\text{class 0}}$ reduces false positives, confirming that boosting the presence of label-0 instances through synthetic data helps mitigate bias. This effect is particularly evident when fewer subgroups are selected (right column), where the reduction is sharper and more effective. The Table 5.18 compares the three bias mitigation strategies:

Table 5.18: Comparison of Results for Targeted Data Acquisition, Random Data Acquisition and SMOTE-NC Data Generation Approaches for Decision Tree Model, metric: False Positive, MinimumSupport:10%.
Note: For each % K (15, 30, 40), the best results for each metric within the same sample size are marked with underline, while the overall best results are marked in **bold**.

% K	# Sub-groups	# Sam-ples	Approach	Accuracy	F1-Score	# FP	Δ_{avg}	Δ_{max}	Δ_{10}	Δ_{20}
-	-	-	Original	0.919	0.275	84	0.012	0.039	0.029	0.024
15	2	295	Random Acquisition	0.915	0.244	92	0.014	0.052	0.034	0.030
			Targeted Acquisition	0.920	0.263	76	0.012	0.052	0.029	0.021
			Generation $p = 1$	0.920	**0.284**	82	0.012	0.046	0.034	0.027
		500	Random Acquisition	0.918	0.246	81	0.014	0.052	0.034	0.030
			Generation $p = 1$	**0.921**	0.261	**72**	**0.010**	**0.036**	**0.025**	**0.020**
30	4	421	Random Acquisition	0.917	0.243	86	0.013	0.057	0.036	0.031
			Targeted Acquisition	0.922	0.260	66	0.010	0.042	0.025	0.018
			Generation $p = 1$	0.922	**0.270**	71	0.011	0.045	0.030	0.022
		1000	Random Acquisition	0.918	0.264	87	0.013	0.057	0.036	0.031
			Generation $p = 1$	**0.925**	0.262	**57**	**0.008**	**0.031**	**0.022**	**0.017**
40	6	559	Random Acquisition	0.918	0.246	82	0.012	0.043	0.033	0.028
			Targeted Acquisition	**0.923**	0.263	62	0.009	0.041	0.029	0.022
			Generation $p = 0.8$	0.920	0.263	75	0.011	0.045	0.033	0.026
		1000	Random Acquisition	0.918	**0.264**	87	0.012	0.043	0.033	0.028
			Generation $p = 1$	0.922	0.241	**61**	**0.009**	**0.023**	**0.015**	**0.013**

Targeted Data Acquisition, Random Data Acquisition, and SMOTE-NC Data Generation, in reducing False Positives and divergence metrics.

With a minimum support of 10% in the validation set, Targeted Acquisition is the most effective when the number of added samples is low. Selecting problematic instances from the holdout set with label 0 provides a direct reduction in False Positives. However, as the number of generated samples increases, Data Generation (SMOTE-NC) becomes the best approach, achieving the most significant reduction in False Positives and divergence metrics. This suggests that synthetic data generation enhances distribution homogeneity and minimizes gaps between problematic subgroups and the overall population.

For a fixed , Data Generation consistently achieves better results with more generated samples. The reduction in Δ_{avg} and Δ_{max} indicates improved subgroup alignment, while Δ_{10} and Δ_{20} confirm that even the most problematic subgroups benefit from mitigation. These metrics measure the absolute variation in average, maximum, top 10, and top 20 divergences on the test set before and after mitigation. Accuracy and F1-score do not necessarily follow this trend, as they are general evaluation metrics rather than subgroup-specific ones.

At lower sample sizes, Targeted Acquisition is more effective at reducing False Positives due to its focused selection of problematic instances. With a higher number of generated samples, Data Generation using SMOTE-NC is the superior strategy, consistently reducing divergence metrics and False Positives more effectively than the other approaches.

The choice of bias mitigation strategy depends on the number of available samples. With fewer added samples, Targeted Acquisition is preferable for its immediate reduction of False Positives. If more synthetic samples can be generated, Data Generation proves to be the most effective long-term strategy, ensuring greater divergence reduction and better overall data distribution balance.

Since the proposed solution is model-agnostic, the method was tested on different models to evaluate its effectiveness in reducing False Positives. In this analysis, K-Nearest Neighbors (KNN) was used instead of Logistic Regression, as Logistic Regression produced zero False Positives, making a meaningful comparison impossible. The results in Table 5.19 show that the

Generation Acquisition strategy with SMOTE-NC improves False Positive mitigation and reduces subgroup divergence as the number of generated samples increases.

Table 5.19: Comparison of results for different models, metric: False Positive. For GB K:30%, MinimumSupport: 2%, and Pruning: 0%. For KNN K:15%, MinimumSupport: 2%, and Pruning:0%. For RF K:20%, MinimumSupport: 2%, and Pruning:0%. *Note:* For each model type (GB, KNN, RF), the best results for each metric are marked in **bold**. If nothing is in bold, then the metric is worse or equal than the initial one.

Model	**# Samples**	**Approach**	**Accuracy**	**F1-Score**	**# FP**	Δ_{avg}	Δ_{max}	Δ_{10}	Δ_{20}
GB	-	Original	0.927	0.032	4	0.001	0.043	0.037	0.032
	500	Random Acquisition	0.926	0.032	5	0.001	0.043	0.037	0.032
		Generation $p = 0.95$	**0.928**	0.043	2	0.001	0.037	0.034	0.023
	1500	Random Acquisition	**0.928**	**0.054**	2	0.001	0.043	0.037	0.032
		Generation $p = 0.95$	**0.928**	0.033	**1**	**0.000**	**0.022**	**0.019**	**0.018**
KNN	-	Original	0.925	0.179	26	0.002	0.131	0.108	0.100
	500	Random Acquisition	**0.929**	0.196	15	0.002	0.108	0.100	0.095
		Generation $p = 0.95$	0.926	0.182	22	0.002	**0.070**	**0.066**	0.063
	1500	Random Acquisition	0.927	**0.234**	26	0.002	0.108	0.100	0.095
		Generation $p = 0.6$	0.927	0.183	**20**	0.002	0.071	**0.066**	**0.061**
RF	-	Original	0.934	0.359	28	0.004	0.014	0.007	0.003
	500	Random Acquisition	0.934	0.365	28	0.005	0.016	0.007	0.003
		Generation $p = 0.95$	**0.936**	0.361	22	**0.003**	0.011	0.005	**0.002**
	1500	Random Acquisition	0.934	**0.370**	29	0.005	0.016	0.007	0.003
		Generation $p = 1$	**0.936**	0.350	**20**	**0.003**	**0.009**	**0.004**	**0.002**

In general, Generation Acquisition strategies lead to a reduction in the total number of False Positives and a decrease in divergence metrics (Δ_{max}, Δ_{avg}, Δ_{10}, and Δ_{20}), indicating greater fairness across subgroups.

For Gradient Boosting (GB), the Generation Acquisition strategy with p=0.95 and 1500 samples yields the best results, reducing the number of

False Positives from 4 to 1, lowering Δ_{max} from 0.043 to 0.022, and slightly improving Accuracy (0.928). In the case of KNN, Random Acquisition with 500 samples improves Accuracy (0.929) and F1-Score (0.196) while reducing False Positives from 26 to 15. Similarly, Generation Acquisition with $p = 0.95$ lowers Δ_{max} from 0.131 to 0.070, but does not achieve noticeable improvements in other metrics. For Random Forest (RF), Generation Acquisition with $p = 1$ and 1500 samples delivers the best results, reducing False Positives from 28 to 20 and lowering Δ_{max} from 0.014 to 0.009. Accuracy also improves slightly (0.936) although the F1-Score experiences a minor decrease (from 0.359 to 0.350).

For all models, increasing the number of synthetic samples tends to enhance bias mitigation. In particular, using a higher number of generated samples results in the most significant reductions in False Positives and subgroup disparities. Accuracy and F1-Score, being overall performance metrics, vary depending on the chosen strategy but do not always directly reflect improvements in fairness.

These results confirm that increasing the number of synthetic samples with SMOTE-NC is an effective strategy for reducing False Positives and subgroup discrepancies, improving overall fairness without excessively compromising the model's general performance.

5.2.4.2 COMPAS FP Mitigation Main Outcomes

The analysis demonstrates that bias mitigation using SMOTE-NC effectively reduces False Positives, particularly when synthetic points are predominantly assigned to class 0. Though its impact varies based on factors such as synthetic sample size, subgroup selection, and parameter settings.

Key observations:

1. The Generation strategy, proposed in this study, proves to be the most effective approach when a sufficiently large number of synthetic samples are injected into the training set.

2. Targeted Acquisition outperforms other methods when the number of additional samples is limited, highlighting its suitability for early-stage data augmentation.

3. Higher thresholds detect fewer subgroups, but more representative, while a lower minimum support thresholds identify more subgroups, but less representative. Here selecting too many subgroups can introduce noise, limiting effectiveness. A low support threshold is more effective, as it captures more representative problematic subgroups enhances mitigation effectiveness.

4. While Accuracy and F1-score provide a general performance overview, they do not fully capture fairness improvements at the subgroup level. A slight decrease in F1-score in some cases aligns with a stronger reduction in divergence, emphasizing the need for fairness-aware evaluation metrics.

5. The best overall performance is achieved using the Generation Acquisition strategy when a sufficiently large number of samples are included. This confirms that bias mitigation strategies must be carefully tailored based on the dataset characteristics and the balance between fairness and model Accuracy.

In conclusion, SMOTE-NC is effective in mitigating False Positives, but achieving fairness requires careful parameter tuning. Targeted Acquisition works best for fewer samples, while Data Generation with SMOTE-NC ensures long-term improvements in fairness and subgroup balance.

5.2.4.3 COMPAS False Negative Mitigation

Again, as a preliminary analysis of the solution's impact on False Negatives for COMPAS dataset, Figure 5.5 is examined.

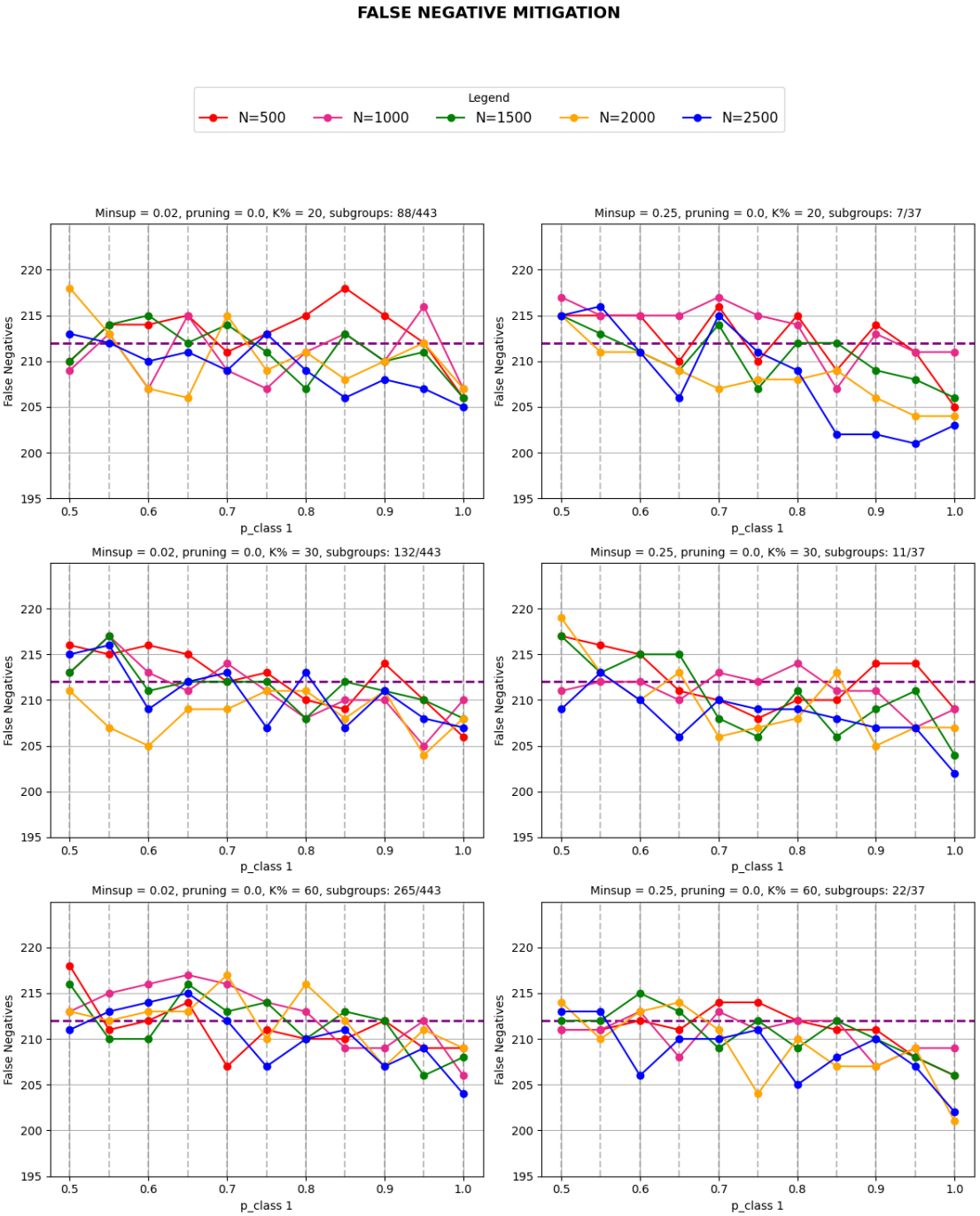
This image illustrates the impact of bias mitigation on False Negatives using SMOTE-NC, applied to problematic subgroups identified with DivExplorer for a Decision Tree model. The x-axis represents the probability that synthetic points belong to class 1, ranging from 0.5 to 1. This aligns with the idea that, in mitigating False Negatives, it is necessary to increase the number of synthetic data points representing problematic instances labeled as 1. The y-axis indicates the number of False Positives.

To perform the mitigation, synthetic samples (ranging from 0.5K to 2.5K) are injected into the training set to balance the data distribution and reduce False Positives. The experiments are conducted by varying the number of problematic subgroups used to identify problematic instances in the validation set. This approach allows for analyzing how different subgroup selections influence the effectiveness of the mitigation strategy.

On the left side of the figure, a lower minimum support threshold is applied, enabling the identification of a larger number of problematic subgroups. Conversely, on the right side, a higher support threshold results in fewer detected subgroups. Moving from top to bottom in the figure, the number of subgroups selected for data augmentation increases. From the analysis of the figure, we observe key trends in how the number of False Negatives varies depending on the number of problematic subgroups and the minimum support threshold used in DivExplorer.

As before, a higher minimum support threshold (right column) proves more effective. Since the mitigation strategy targets False Negatives, increasing the support threshold ensures that the identified subgroups contain more class-1 instances, leading to a more reliable synthetic data generation process. This results in a greater reduction of False Negatives compared to a lower support threshold (left column), where subgroups may contain fewer meaningful class-1 instances.

Figure 5.5: False Negatives trend generated with SMOTE-NC (0.5K-2.5K) as $p_{\text{class 1}}$ varies for a Decision Tree. On the left, min_sup = 2%; on the right, min_sup = 25%; for both pruning parameter = 0%. Each row compares results for the same percentage of problematic subgroups used in mitigation.



As the number of selected subgroups increases (moving from top to bottom), the effect of mitigation is generally more pronounced. A larger set of subgroups provides more diverse instances for data augmentation, helping to balance the dataset and reduce False Negatives more effectively. However, the impact stabilizes or fluctuates when too many subgroups are included, suggesting that excessive subgroup diversity may limit the benefits of synthetic data generation.

Overall, SMOTE-NC is most effective in mitigating False Negatives when applied to subgroups with a sufficiently high support threshold, ensuring that synthetic data is generated from well-represented class-1 instances. A balance must be maintained, as selecting too many or too few subgroups can affect the overall effectiveness of the mitigation strategy.

Table 5.20 not only highlights the reduction of False Negatives through different bias mitigation strategies but also provides a detailed comparison of divergence metrics. While Accuracy and F1-score offer a global measure of model performance, they do not provide insights into subgroup disparities. The table allows for a deeper evaluation of how each approach impacts both overall classification and fairness across different groups.

When the number of added samples is low, generally, Random Data Acquisition proves to be the most effective in reducing False Negatives.
However, as the number of generated samples increases, SMOTE-NC becomes the most effective strategy in reducing both False Negatives and divergence metrics. This indicates that synthetic data generation helps create a more homogeneous distribution, reducing gaps between problematic subgroups and the overall population.

In this specific case, Random Acquisition performances often yield better local results because there are very few problematic instances of class 1. As a result, SMOTE-NC struggles to generate effective synthetic samples, limiting its ability to reduce False Negatives as efficiently as in other scenarios. This highlights the strong influence of data imbalance on the performance of synthetic data generation techniques.

Table 5.20: Comparison of Results for Targeted Data Acquisition, Random Data Acquisition and SMOTE-NC Data Generation Approaches for Decision Tree Model, metric: False Negative, MinimumSupport:25%.
Note: For each % K (20, 30, 60), the best results for each metric within the same sample size are marked with underline, while the overall best results are marked in **bold**.

% K	# Sub-groups	# Sam-ples	Approach	Accuracy	F1-Score	# FN	Δ_{avg}	Δ_{max}	Δ_{10}	Δ_{20}
-	-	-	Original	0.919	0.275	212	0.124	0.209	0.197	0.179
20	7	413	Random Acquisition	0.916	0.246	218	0.108	0.187	0.174	0.157
			Targeted Acquisition	**0.922**	**0.282**	212	0.124	0.209	0.197	0.179
			Generation $p = 1$	0.913	0.273	**208**	0.121	0.224	0.206	0.185
		1000	Random Acquisition	0.918	0.264	214	0.108	0.187	0.174	0.157
			Generation $p = 0.75$	0.912	0.248	215	**0.102**	**0.173**	**0.165**	**0.148**
30	11	542	Random Acquisition	0.916	0.243	219	0.105	0.183	0.170	0.153
			Targeted Acquisition	**0.922**	0.277	213	0.124	0.205	0.197	0.181
			Generation $p = 1$	0.907	0.261	208	0.127	0.224	0.210	0.190
		1500	Random Acquisition	0.920	**0.297**	215	0.105	0.183	0.170	0.153
			Generation $p = 0.6$	0.909	0.241	**206**	**0.103**	**0.173**	**0.165**	**0.148**
60	22	1105	Random Acquisition	0.917	0.260	215	0.108	0.198	0.182	0.161
			Targeted Acquisition	0.925	0.272	217	0.115	0.190	0.183	0.167
			Generation $p = 0.95$	0.902	0.231	214	0.107	**0.177**	**0.169**	**0.155**
		2500	Random Acquisition	0.923	0.291	**210**	0.108	0.198	0.182	0.161
			Generation $p = 0.9$	0.899	0.240	**210**	**0.106**	0.192	0.179	0.160

Again, since the proposed solution is model-agnostic, the method is tested on different models to evaluate its effectiveness in reducing False Negatives.

The results in Table 5.21 show that the Generation Acquisition strategy generally improves this metric and reduces subgroup divergences as the number of generated samples increases.

Table 5.21: Comparison of results for different models, metric: False Negative.
For GB K:45%, MinimumSupport: 25%, and Pruning: 0%. For KNN K:20%, MinimumSupport: 25%, and Pruning:0%. For RF K:30%, MinimumSupport: 25%, and Pruning:0%.
Note: For each model type (GB, KNN, RF), the best results for each metric are marked in **bold**. If nothing is in bold, then the metric is worse or equal than the initial one.

Model	**# Samples**	**Approach**	**Accuracy**	**F1-Score**	**# FN**	Δ_{avg}	Δ_{max}	Δ_{10}	Δ_{20}
GB	-	Original	0.927	0.032	175	0.011	0.017	0.015	0.011
	500	Random Acquisition	0.926	0.032	173	0.018	0.028	0.022	0.018
		Generation $p = 0.75$	0.927	0.032	174	0.011	0.017	0.015	0.011
	1500	Random Acquisition	**0.928**	**0.054**	173	0.018	0.028	0.022	0.018
		Generation $p = 0.5$	0.918	0.020	**172**	**0.008**	**0.011**	**0.010**	**0.008**
KNN	-	Original	0.925	0.179	158	0.065	0.112	0.112	0.103
	500	Random Acquisition	**0.929**	0.196	157	0.084	0.152	0.152	0.137
		Generation $p = 0.95$	0.928	**0.215**	**154**	0.078	0.135	0.135	0.122
	3000	Random Acquisition	0.926	0.196	156	0.084	0.152	0.152	0.137
		Generation $p = 0.85$	0.922	0.167	159	**0.064**	**0.107**	**0.107**	**0.103**
RF	-	Original	0.934	0.359	133	0.143	0.253	0.249	0.231
	500	Random Acquisition	0.934	0.365	132	0.144	0.253	0.249	0.231
		Generation $p = 0.85$	0.932	0.357	132	0.143	0.258	0.249	0.229
	3000	Random Acquisition	**0.936**	**0.379**	131	0.144	0.253	0.249	0.231
		Generation $p = 0.95$	0.918	0.321	**130**	0.143	0.264	**0.242**	**0.223**

For Gradient Boosting (GB), the generation strategy with 1500 additional samples ($p = 0.5$) achieves the most significant reduction in False Negatives and subgroup divergence metrics. The Random Acquisition approach does not appear to offer substantial improvements in these aspects.

For K-Nearest Neighbors (KNN), the generation strategy with $p = 0.95$ results in the lowest number of False Negatives while also improving the F1-score. However, with a higher number of generated samples ($p = 0.85$, 3000 samples), the divergence metrics decrease further, suggesting a trade-off between overall performance and model fairness improvement.

For Random Forest (RF), the generation strategy with $p = 0.95$ achieves the lowest number of False Negatives but does not significantly reduce subgroup divergence. An important factor to consider is that SMOTE-NC generates new samples based on existing data, but in this case, the number of minority class instances was limited, reducing the effectiveness of data generation and the potential improvement in fairness metrics.

In summary, the results confirm that increasing the number of synthetic samples can help reduce False Negatives and subgroup divergences, but the effectiveness of this strategy depends on the original class distribution and the model used.

5.2.4.4 COMPAS FN Mitigation Main Outcomes

The analysis demonstrates that bias mitigation using SMOTE-NC effectively reduces False Negatives, particularly when synthetic points are predominantly assigned to class 1. However, its impact varies based on factors such as synthetic sample size, subgroup selection, and parameter settings. Key Observations

1. The Generation strategy, proposed in this study, proves to be the most effective approach when a sufficiently large number of synthetic samples are injected into the training set.

2. Random Acquisition outperforms other methods when the number of additional samples is limited.

3. Higher minimum support thresholds detect fewer subgroups, leading to a more reliable synthetic data generation process and sharper reductions in False Negatives. Conversely, lower thresholds capture a larger number of subgroups, offering gradual mitigation but introducing noise when too many subgroups are selected. In small datasets, a low support threshold is more effective as it identifies more problematic subgroups without overwhelming the original data distribution.

4. While Accuracy and F1-score provide a general performance overview, they do not fully capture fairness improvements at the subgroup level. A slight decrease in F1-score in some cases aligns with a stronger reduction in divergence, emphasizing the need for fairness-aware evaluation metrics.

5. The best overall performance is achieved using the Generation Acquisition strategy when a sufficiently large number of samples are included. This confirms that bias mitigation strategies must be carefully tailored based on the dataset characteristics and the balance between fairness and model accuracy.

For these reasons SMOTE-NC is effective in mitigating False Negatives, but achieving fairness requires careful parameter tuning. Targeted Acquisition works best when fewer samples are available, while Data Generation with SMOTE-NC ensures long-term improvements in fairness and subgroup balance.

5.2.4.5 COMPAS Error Mitigation

For the last time, as a preliminary analysis of the solution's impact on Errors for COMPAS dataset, Figure 5.6 is examined.

This image illustrates the impact of bias mitigation on Errors using SMOTE-NC, applied to problematic subgroups identified with DivExplorer for a Decision Tree model. The x-axis represents the probability that synthetic points belong to class 0; since this is a binary classification problem, the complementary value represents the probability of belong to class 1. The y-axis indicates the number of Errors.

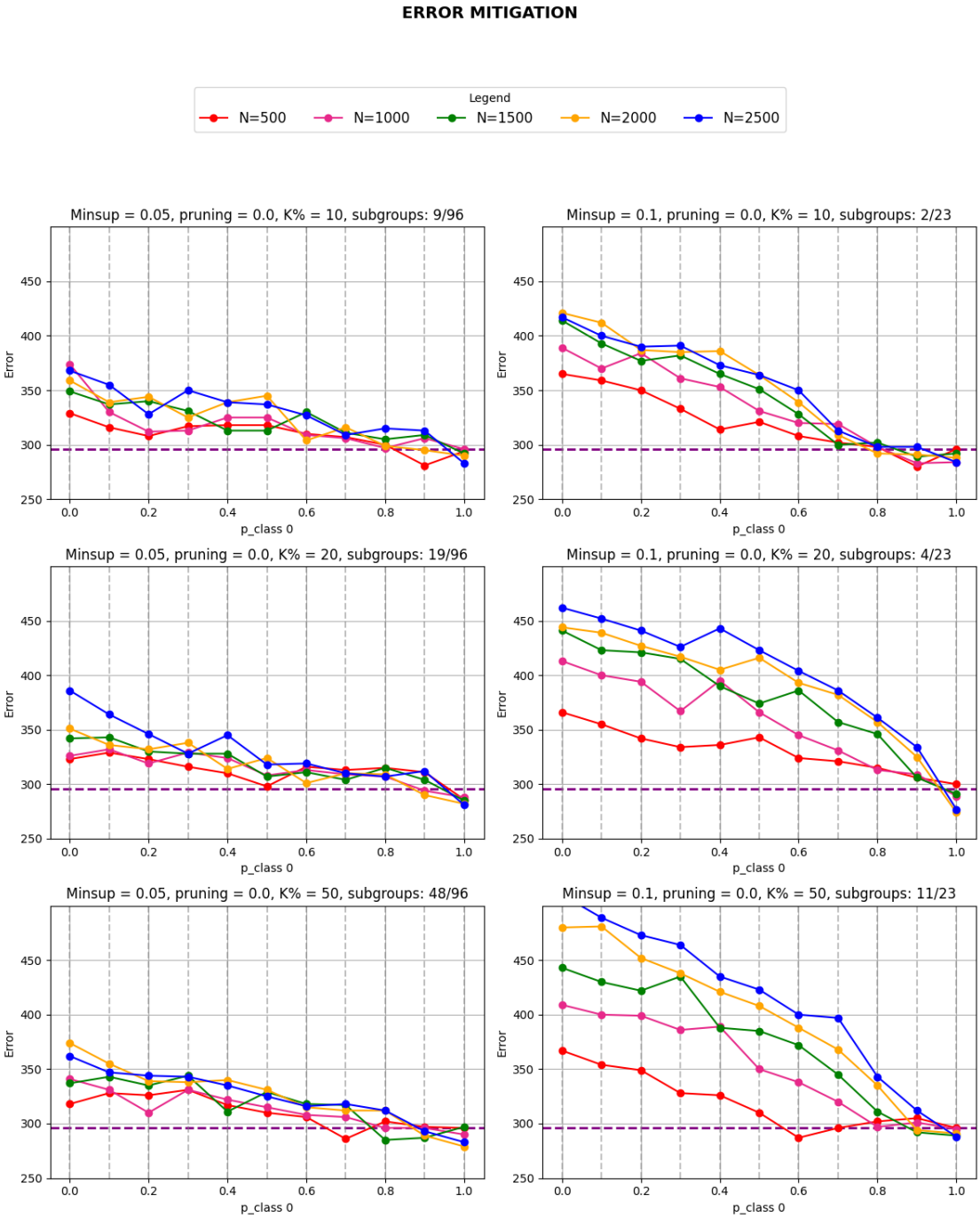
To perform the mitigation, synthetic samples (ranging from 0.5K to 2.5K) are injected into the training set to balance the data distribution and reduce the Errors. The experiments are conducted by varying the number of problematic subgroups used to identify problematic instances in the validation set. This approach allows for analyzing how different subgroup selections influence the effectiveness of the mitigation strategy.

On the left side of the figure, a lower minimum support threshold is applied, enabling the identification of a larger number of problematic subgroups. Conversely, on the right side, a higher support threshold results in fewer detected subgroups. Moving from top to bottom in the figure, the number of subgroups selected for data augmentation increases. The figure shows how Error mitigation is affected by the number of problematic subgroups and the minimum support threshold used in DivExplorer.

A higher minimum support threshold (right column) generally leads to a more stable Error reduction because the identified subgroups contain more representative instances, making SMOTE-NC more effective. In contrast, a lower support threshold (left column) detects more subgroups, but they may include less meaningful instances, leading to less consistent results.

As the number of selected subgroups increases (top to bottom), Error mitigation becomes more pronounced, particularly when a higher number of synthetic points is injected. However, when too many subgroups are selected, the effect stabilizes, indicating that excessive subgroup diversity may reduce the overall impact of synthetic data generation.

Figure 5.6: Error trend generated with SMOTE-NC (0.5K-2.5K) as $p_{\text{class }0}$ varies for a Decision Tree. On the left, min_sup = 5%; on the right, min_sup = 10%; for both pruning parameter = 0%. Each row compares results for the same percentage of problematic subgroups used in mitigation.



Although the previously described figure provides insight into how the number of Errors evolves when new data is added to the training set, it does not offer any information on divergence metrics. For this reason, Table 5.22 compares the three different mitigation methods as the number of inspected subgroups varies, while keeping the minimum support fixed at 10%.

Table 5.22: Comparison of Results for Targeted Data Acquisition, Random Data Acquisition and SMOTE-NC Data Generation Approaches for Decision Tree Model, metric: Errors, MinimumSupport:10%.
Note: For each % K (10, 20, 50), the best results for each metric within the same sample size are marked with underline, while the overall best results are marked in **bold**.If nothing is in bold, then the metric is worse than the initial one.

% K	# Sub-groups	# Sam-ples	Approach	Accuracy	F1-Score	# ER	Δ_{avg}	Δ_{max}	Δ_{10}	Δ_{20}
-	-	-	Original	0.919	0.275	296	0.023	0.076	0.062	0.054
10	2	383	Random Acquisition	0.917	0.248	304	0.024	0.081	0.070	0.063
			Targeted Acquisition	**0.922**	0.281	**287**	**0.022**	0.069	0.058	0.051
			Generation $p = 0.1$	0.920	0.262	293	0.024	0.072	0.063	0.052
		1500	Random Acquisition	0.920	**0.297**	293	0.024	0.081	0.070	0.063
			Generation $p = 1$	0.920	0.244	292	0.023	**0.068**	**0.055**	**0.048**
20	4	496	Random Acquisition	0.917	0.244	304	0.025	0.086	0.073	0.064
			Targeted Acquisition	0.923	**0.296**	280	0.021	**0.071**	**0.057**	**0.051**
			Generation $p = 1$	0.920	0.270	292	0.023	0.090	0.066	0.059
		2000	Random Acquisition	0.918	0.293	299	0.025	0.086	0.073	0.064
			Generation $p = 1$	**0.925**	0.267	**274**	**0.020**	0.086	0.068	0.059
50	11	790	Random Acquisition	0.920	0.269	294	0.023	0.086	0.071	0.061
			Targeted Acquisition	**0.927**	**0.310**	**267**	**0.019**	**0.064**	**0.053**	**0.047**
			Generation $p = 0.8$	0.921	0.287	288	0.021	0.102	0.076	0.065
		1000	Random Acquisition	0.918	0.264	301	0.023	0.086	0.071	0.061
			Generation $p = 1$	0.919	0.261	295	0.023	0.075	0.063	0.058

It is observed that, unlike in previous cases, here it is not possible to select labels ad hoc for Targeted Acquisition. As a result, the problematic instances from the holdout set that are added to the training set may sometimes worsen the metrics instead of improving them. However, in this case, even without the specific selection of labels, Targeted Acquisition generally performs better.

Moreover, as the number of added samples increases, the Generation Strategy in this case does not always prove to be the most effective, as already observed in previous cases. It is important to note that the total number of errors does not decrease dramatically. This is because injecting too many instances of class 0 or class 1 improves one of the two metrics (False Negatives or False Positives) while worsening the other, ultimately keeping the total error count high.

As for Accuracy and F1-Score, the same considerations as before apply: while they provide a general performance overview, they do not fully capture fairness improvements at the subgroup level.

For the last time, recall that since the proposed solution is model-agnostic,its impact is evaluated on overall Errors across different models. Table 5.23 shows that the Generation Acquisition strategy can improve False Negative reduction but may increase False Positives, leading to a relatively stable total Error count.

For Gradient Boosting (GB), the generation strategy with $p = 0.9$ and 2000 additional samples achieves the lowest number of Errors while also improving subgroup divergence metrics. However, the overall reduction in Errors is not drastic, as improvements in False Negatives are counterbalanced by potential increases in False Positives.

For K-Nearest Neighbors (KNN), the results show a trade-off: the Random Acquisition approach reduces the total number of Errors the most, while the Generation Acquisition strategy with p=0.5 achieves better subgroup fairness metrics. This suggests that although the generation strategy enhances fairness, it may not always lead to a significant reduction in total Errors.

Table 5.23: Comparison of results for different models, metric: Error.
For GB K:30%, MinimumSupport: 15%, and Pruning: 0%. For KNN K:40%, MinimumSupport: 15%, and Pruning:0%. For RF K:40%, MinimumSupport: 10%, and Pruning:0%.
Note: For each model type (GB, KNN, RF), the best results for each metric are marked in **bold**. If nothing is in bold, then the metric is worse or equal than the initial one.

Model	# Samples	Approach	Accuracy	F1-Score	# ER	Δ_{avg}	Δ_{max}	Δ_{10}	Δ_{20}
GB	-	Original	0.927	0.032	179	0.013	0.057	0.039	0.030
	500	Random Acquisition	0.926	0.032	180	0.013	0.060	0.040	0.030
		Generation $p = 1$	0.927	0.043	178	0.013	0.055	0.038	**0.029**
	2000	Random Acquisition	0.927	0.053	177	0.013	0.060	0.040	0.030
		Generation $p = 0.9$	**0.928**	**0.054**	**175**	**0.012**	**0.054**	**0.037**	**0.029**
KNN	-	Original	0.925	0.179	184	0.010	0.053	0.045	0.038
	500	Random Acquisition	**0.929**	0.196	**172**	0.010	0.051	0.041	0.029
		Generation $p = 0.4$	0.925	0.208	183	**0.009**	0.048	0.042	0.038
	5000	Random Acquisition	0.927	**0.234**	177	0.010	0.051	0.041	0.029
		Generation $p = 0.5$	0.928	0.200	176	0.010	**0.041**	**0.035**	**0.028**
RF	-	Original	0.934	0.359	161	0.009	0.054	0.045	0.040
	500	Random Acquisition	0.934	0.365	160	0.009	0.057	0.047	0.041
		Generation $p = 0.5$	**0.936**	**0.386**	**156**	0.008	**0.052**	**0.039**	0.034
	2500	Random Acquisition	0.935	0.368	158	0.009	0.057	0.047	0.041
		Generation $p = 0.8$	**0.936**	0.381	**156**	**0.007**	**0.052**	**0.039**	**0.033**

For Random Forest (RF), the generation strategy with $p = 0.8$ achieves the lowest number of Errors and subgroup divergence. However, as SMOTE-NC relies on available minority class data to generate new instances, the effectiveness of this strategy is limited when the original dataset has an imbalanced class distribution.

In general, increasing the number of generated samples tends to improve performance on subgroup-related metrics, leading to better fairness and reduced divergence across different groups.

However, the impact on overall Errors remains balanced due to trade-offs between False Positives and False Negatives. The results confirm that the Generation Acquisition strategy can effectively reduce the number of Errors and improve subgroup divergence, but it does not necessarily lead to a sharp decrease in total Errors.

When False Negatives decrease, False Positives may increase, and conversely, when False Negatives rise, False Positives may decrease. This highlights the importance of carefully selecting the number of generated samples to optimize both overall performance and fairness.

5.2.4.6 COMPAS Error Mitigation Main Outcomes

The analysis demonstrates that bias mitigation using SMOTE-NC effectively impacts overall Error reduction, particularly in relation to the balance between False Positives and False Negatives. The effectiveness of this approach depends on several factors, including the number of synthetic samples generated, the subgroup selection criteria, and specific parameter configurations. Key observations:

1. The Generation Strategy, as proposed in this study, is not always the most effective also when a large number of synthetic samples are injected into the training set. However this strategy generally improves subgroup divergence metrics, reducing disparities between problematic groups and the overall population.

2. Targeted Acquisition often performs best even when the number of additional samples is low. Despite the inability to select labels ad hoc in this scenario, the method effectively mitigates the impact of problematic instances from the holdout set, leading to overall metric improvements.

3. A higher minimum support threshold detects fewer subgroups, leading to a more stable but sometimes limited reduction in Errors. Conversely, a lower minimum support threshold identifies a larger number of subgroups, which can result in more effective bias mitigation. However, excessive subgroup diversity may introduce noise, reducing the overall benefits of synthetic data generation.

4. The total number of Errors does not decrease dramatically because increasing synthetic instances of either class (0 or 1) improves one metric (False Negatives or False Positives) at the cost of the other. This balance suggests that careful tuning of synthetic data generation is required to optimize fairness without disproportionately affecting accuracy.

SMOTE-NC proves to be an effective tool for mitigating bias and improving fairness metrics, particularly when applied through the Generation Strategy. However, the impact on overall Error reduction is nuanced: reducing False Negatives often leads to an increase in False Positives, and vice versa. Therefore, selecting an appropriate number of synthetic samples is crucial to balancing fairness with model accuracy.

Chapter 6

Conclusions and Future Works

In this final chapter, the conclusions are presented in 6.1, while the potential future works are discussed in 6.2.

6.1 Conclusions

Machine Learning is now widely used in various decision-making domains, directly impacting people's lives. However, ML models can inherit and amplify biases present in the data, compromising the fairness of their predictions. For this reason, it is crucial to develop bias mitigation strategies that improve the reliability and transparency of these systems.

This thesis proposes a new model-agnostic bias mitigation method for tabular data, which automatically identifies problematic subgroups and generates new representative data to balance the training set. The data generation process is performed using SMOTE-NC, a pre-existing method traditionally used to balance imbalanced datasets. However, in this work, SMOTE-NC has been employed in an innovative way, specifically aimed at bias mitigation, thus expanding its application scope.

Experimental results have shown that the proposed approach can improve prediction fairness depending on the dataset characteristics, by adjusting the parameters described in the thesis. This process may lead to a reduction in overall accuracy, but it is important to note that global accuracy does not provide insight into the quality of predictions for problematic subgroups.

For example, increasing the number of synthetic instances with a negative class can reduce False Positives (FP) at the cost of increasing False Negatives (FN) and decreasing overall accuracy. However, if the goal is to minimize FP, the adopted strategy proves effective, highlighting the importance of a flexible approach that allows balancing the trade-off between fairness and accuracy based on the specific objectives of the problem.

Ultimately, this work contributes to the development of more balanced and representative datasets, enabling a training phase that better accounts for fairness and reduces the risk of bias in machine learning model decisions.

6.2 Future Works

This work opens several directions for future development. One possible extension of the proposed strategy is its adaptation to multi-class classification by modifying the probability with which the generated samples belong to a specific class. This would allow the method to be applied in more complex scenarios, expanding its applicability.

Moreover, the current strategy may not be optimal for certain metrics, such as the total number of Errors. Since reducing False Positives (FP) increases False Negatives (FN) and vice versa, the overall number of Errors may not decrease significantly. For this reason, it would be beneficial to extend the method so that it can mitigate bias across multiple metrics simultaneously, better balancing the effects on different types of Errors.

Bibliography

[1] Harini Suresh and John Guttag. A framework for understanding sources of harm throughout the machine learning life cycle. In *Proceedings of the 1st ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, EAAMO '21, New York, NY, USA, 2021. Association for Computing Machinery.

[2] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks, 2016. Accessed: 2024-11-09.

[3] Dave Gershgorn. Amazon reportedly scraps internal ai recruiting tool that was biased against women, 2018. Accessed: 2024-11-3.

[4] James Vincent. Uk a-level results algorithm was biased against disadvantaged students, report finds, 2020. Accessed: 2024-11-09.

[5] Eliana Pastor, Luca de Alfaro, and Elena Baralis. Looking for trouble: Analyzing classifier behavior via pattern divergence. In *Proceedings of the 2021 International Conference on Management of Data*, SIGMOD '21, page 1400–1412, New York, NY, USA, 2021. Association for Computing Machinery.

[6] Nima Shahbazi, Yin Lin, Abolfazl Asudeh, and H. V. Jagadish. Representation bias in data: A survey on identification and resolution techniques. *ACM Comput. Surv.*, 55(13s), July 2023.

[7] US Census Bureau. The asian and pacific islander population in the united states: May 2019, 2019. Accessed: 2024-12-24.

[8] Jiaxuan Li, Duc Minh Vo, and Hideki Nakayama. Partition-and-debias: Agnostic biases mitigation via a mixture of biases-specific experts, 2023.

[9] Manh Khoi Duong and Stefan Conrad. *Trusting Fair Data: Leveraging Quality in Fairness-Driven Data Removal echniques*, page 375–380. Springer Nature Switzerland, 2024.

[10] Emmanouil Panagiotou, Arjun Roy, and Eirini Ntoutsi. Synthetic tabular data generation for class imbalance and fairness: A comparative study, 2024.

[11] Alkis Koudounas, Eliana Pastor, Giuseppe Attanasio, Luca de Alfaro, and Elena Baralis. Prioritizing data acquisition for end-to-end speech model improvement. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7000–7004, 2024.

[12] Will Kurt. A gentle introduction to bayesian statistics, 2015. Accessed: January 7, 2025.

[13] Mauro Gasparini. *Modelli Probabilistici e Statistici*. Pearson, Milan, Italy, 2009.

[14] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research (JAIR)*, 16:321–357, 2002.

[15] Maria Antonietta Longo. Official code repository. [Online]. https://github.com/MariaAntoniettaL/Synthetic-Data-Generation-Bias-Mitigation-Subgroup-Based-for-Structured-Data.