Report - Exploring The Impact Of Data Smells On Fairness Problems in ML Solutions

Annamaria Basile
Mat:0522501844
a.basile40@studenti.unisa.it

Mario Cicalese
Mat:0522501799
m.cicalese21@studenti.unisa.it

Paolo Carmine Valletta
Mat:0522501828
p.valletta2@studenti.unisa.it

ABSTRACT

The swift expansion of *Artificial Intelligence* (AI) applications has sparked significant transformations across various sectors, prompting both practitioners and the academic community to address the task of developing products that alleviate global issues concerning equality, inclusivity, and fairness. Researchers have taken notable strides to mitigate inequalities and promote fairness in AI systems. In decision-making, fairness is the absence of bias or preference towards any individual or group based on their inherent or acquired traits. Quality assurance techniques are essential for managing AI-enabled systems that aim to identify and handle quality issues. These quality issues can severely impact various aspects of these systems. Through the different phases of the AI software development lifecycle, developers must consider several types of quality issues and guides to avoid suboptimal algorithms and low-quality datasets (i.e., Data Smells).

Additionally, our project aims to address two research questions which refer to data quality issues in the datasets (data smells) and fairness problems in ML solutions. Specifically, in the first RQ, we intend to explore whether the presence of quality issues (i.e., Data Smells) directly impacts unfairness in AI decision-making or vice versa. Understanding this relationship can offer insights into building more equitable and reliable AI systems.The second research question explores how the performance of bias mitigation algorithms might be affected by the presence of data smells in input datasets, which are commonly employed in literature to mitigate unfairness in machine learning solutions. Specifically, we intend to explore whether the presence of quality issues in the datasets directly impacts the performances of these algorithms. Addressing these two research questions can offer insights into which data engineering techniques minimize the level of unfairness in trained ML solutions. **This is the link to the GitHub repository: https://github.com/MarioCicalese/Fairness_datasmell_SE4AI.git**

1 CONTEXT OF THE PROJECT

Artificial Intelligence (AI) is increasingly prevalent, aiding individuals and companies in decision-making and automating human-like tasks. AI-intensive systems, equipped with advanced algorithms, are now deployed across various domains, showcasing remarkable efficiency and accuracy.

However, the development of artificial intelligence-enabled systems differs from other types of software because the effectiveness of the program in solving specific tasks **heavily relies on the data** and observations used to train the models. There can be various types of issues regarding data quality, which can arise from various reasons, such as data entry errors, inadequate data cleaning, or biases in the data. All these issues represent merely a subset of the *data smells* encountered in the field. In detail, data smells are defined as: *context-independent, data value-based indications of latent data quality issues caused by poor practices that may lead to problems in the future* [7]. Data smells are usually caused by the violation of recommended best practices in data management, software, or data engineering (e.g., data cleaning). As a result, poor-quality data can lead to abnormal behavior (e.g., bias).

Most AI systems and algorithms are data-driven and require data upon which to be trained. In the case of poor-quality training data that contains biases, the algorithms trained on them will learn these biases and reflect them in their predictions. The outcomes of these biased algorithms can then be fed into real-world systems and affect users' decisions [8]. Like people, algorithms are vulnerable to biases that render their decisions *unfair*. In the context of decision-making, *unfairness* is the *absence of any prejudice or favoritism toward an individual or group based on their inherent or acquired characteristics*[8]. In the literature, one of the methods used to mitigate unfairness in ML solutions is by employing *bias mitigation algorithms*[2]. Designed to identify and mitigate biases present in training data or algorithms themselves, thereby promoting fairness and reducing discrimination in AI systems. There are three types of algorithms:

- ***Pre-processing:*** This approach recognizes, tends to alter the sample distributions of protected variables or, more generally, perform specific transformations on the data to remove discrimination from the training data [4].
- ***In-processing:*** This approach often incorporates one or more fairness metrics into the model optimization functions in a bid to converge towards a model parameterization that maximizes performance and fairness [4].
- ***Post-processing:*** This approach tends to apply transformations to model output to improve prediction fairness [4].

If the algorithm is allowed to modify the training data, then pre-processing can be used. If it is allowed to change the learning procedure for a machine learning model, then in-processing can be used. If the algorithm can only treat the learned model as a black box without any ability to modify the training data or learning algorithm, then only post-processing can be used.

2 GOALS OF THE PROJECT

It has been shown by Recupito et al.[9] that the correlation between data smells and data quality is notably impactful, exhibiting a pronounced and substantial effect, especially in highly diffused data smell instances. Based on this insight, the first goal of this project is to study whether exists a relationship between anti-patterns in datasets (data smells) and fairness problem in machine learning solutions. To achieve this goal, we plan to perform a benchmark study

by employing state-of-the-art datasets known to create possible unfairness in trained ML solutions. After detecting the data smells in these datasets through the *Data Smell Detection* (DSD) tool [7] and after refactoring them, we will analyze the fairness of models trained on such datasets employing fairness metrics available on *AI Fairness 360* [3] (AIF-360):

- **SPD (Statistical Parity Difference):** calculates the disparity in favorable rates between the privileged and unprivileged groups.
- **AOD (Average Odds Difference):** captures the average discrepancy in false-positive rates and true-positive rates between the privileged and unprivileged groups.
- **EOD (Equal Opportunity Difference):** assesses the disparity in true-positive rates between the privileged and unprivileged groups.

We will analyze how the presence of data smells affects the level of fairness in our ML solutions. This goal will be decomposed into several sub-steps, namely: (1) Select state-of-the-art datasets known to induce unfairness in trained ML models without performing data engineering techniques (e.g., data cleaning). (2) Train shallow machine learning models on these raw datasets (with data smells) and, compute and analyze fairness metrics applied to both the model's prediction and the datasets. (3) Detect data smells in datasets employing the DSD tool. (4) Refactor data smells from the dataset. (5) Retrain the shallow ML models on tidy datasets (without data smells) and, compute and analyze fairness metrics applied to both the model's prediction and the datasets (as in Step 2). (6) Analyze and compare the results obtained from the fairness metrics with the number of data smells.

Thus, the second goal pertains to bias mitigation algorithms. Specifically, our goal is to study, whether there is a relationship between data smells and the performances of the bias mitigation Algorithms. To achieve this goal, we plan to perform a second benchmark study by employing the same datasets used for the first goal. The technique we plan to perform is also very similar to the previous one but with a few additional steps, specifically: after computing the fairness metrics in step (2), we will apply one bias mitigation algorithm for every technical approach (pre, in, and post-processing) and finally, we will iteratively recompute the fairness metrics (for each algorithm) to observe how the level of fairness issues change throughout the iteration. This added substep will be performed even after computing the fairness metrics in step (5). We will then analyze and compare the performances obtained from bias mitigation algorithms performed on raw and tidy datasets to understand how the presence of data smells affects the performances of these algorithms in our ML solutions.

As award criteria, listed in descending priority order, we would like to:

- Analyze CO2 emissions and energy consumption using CodeCarbon;
- Use ML-ops tool such as MLFlow to keep track of the experimented solutions;
- Experiment with deep learning solutions.

2.1 Research Questions

To formalize the main goal of our project, we applied the Goal-Question-Metric[1]. In detail, we defined the goal of our project:

OUR GOAL
Analyze: The relationship between data smell with fairness and bias mitigation algorithms
For the purpose of: understanding whether data smells influence the level of fairness and the performances of bias mitigation algorithms
From the point of view: of data engineering aiming to address fairness issues while ensuring quality
In the context of: ML-enabled System

Based on our goal, we have defined two research questions. The first research question pertains to data smells and fairness. Specifically, datasets may contain data quality issues (data smells), which, without a data pre-processing step will be fed into models. On the other hand, models could make unfair predictions if the training data aren't properly pre-processed. Thus, our goal is to investigate whether the presence of data smells in the dataset used for training the models, can impact the level of fairness problems. This step is critical because wether the unfairness are influenced by the presence of data smells in the input data, applying robust data engineering techniques (e.g., data cleaning) before the training, could serve as a solution to mitigate etichal bias. Specifically, the first research question is:

RQ$_1$. *Does the presence of data smells impact the fairness of machine learning models?*

As discussed previously, bias mitigation algorithms are employed in literature to address ethical bias issues within machine-learning solutions. However, it remains unclear whether the performances of these algorithms might be influenced by the presence of data smells in the input datasets. Hence, we plan to perform these algorithms on both datasets, with and without data smells in order to analyze the results and discern under which conditions these algorithms demonstrate greater efficiency. This is crucial, because, data engineering aims to reduce unfairness in ML solutions, and, comprehending which condition bias mitigation algorithms minimize unfairness is pivotal. Thus, we have defined the following research question:

RQ$_2$. *Does the presence of data smells impact on the performances of bias mitigation algorithms?*

Having established the research questions, the subsequent section will delve into the methodology we will employ to address these questions.

3 METHODOLOGICAL STEPS

In this section, we will outline the methodological steps we will apply to achieve our goal. Specifically, we have divided the following steps into two distinct *pipelines* to specify precisely all the steps we will apply to address each research question. The majority of steps

are in common between the two pipelines, however, we will write them only once.

3.1 Steps to address RQ1

(1) **Datasets selection**
Firstly, we will use of state-of-the-art datasets known to induce unfairness in trained ML models. More specifically we are going to use the three most unfair datasets in literature (in fairness research papers), namely:
- ***COMPAS:*** is a popular commercial algorithm used by judges and parole officers for scoring criminal defendant's likelihood of reoffending.[6]
- ***Adult:*** The adult dataset was created as a resource to benchmark the performance of machine learning algorithms on socially relevant data.[6]
- ***German Credit:*** the data summarizes applicants' financial situation, credit history, and personal situation, including housing and number of liable people.[6]

The following steps will focus on one dataset at a time, using them sequentially.

(2) **Training shallow ML models**
We will train at least two shallow machine learning models (excluding deep learning for now) on one dataset, such as Random Forest, Linear Regression, Naive Bayes, etc. The trained models will make predictions based on the input data (test set). These predictions will be used in the next step to compute the fairness metrics which will indicate the level of unfairness of each model.

(3) **Compute Fairness metrics**
Afterward the models' training, we will compute Fairness metrics based on the models' predictions. We will consider three metrics that have been widely adopted in the software fairness literature (as defined in [5]): **SPD, AOD, EOD**. The implementation that we will employ is available on the AI Fairness 360 site.
This step aims to calculate the fairness metrics on models trained on the raw dataset (with data smells), which will be used later in result analysis for comparison with the same metrics calculated on the tidy dataset. This comparative analysis will provide insights into the direct impact of data quality issues on the unfairness of machine learning models.

(4) **Detection of data smells**
At this point, we begin working on the data smells. The first step is to identify the data smells present in the dataset under analysis. We will detect data smells in the dataset using *Rule-based Detection tool* defined by Foidl et al. [7]. The tool focuses on rule-based smell detection and it can detect smells such as Long Data Value, Casing, Integer as String, Floating Point Number as String, or Integer as Floating Point Number. In detail, we will upload the dataset to the DSD tool, which will then generate a categorized list of detected data smells. This list will include the types of data smells identified, along with the total count of elements, the count of faulty elements for each type, and the values of those faulty elements. This step aims to identify potential issues or anomalies in the data that could impact the performance and fairness of machine learning models.

(5) **Refactoring of data smells**
At this stage, we can go ahead and refactor the identified data smells. Specifically, we will manually refactor the data smells considering the ones identified by the DSD tool. The refactoring will be carried out one type at a time, involving the modification of the detected faulty elements. The removal of data smells will produce the tidy dataset, which will be used as input to re-perform the previously described steps. This step is crucial because the dataset without data quality issues (tidy dataset) will be used to re-train (Step 2) the same previous models. This will allow us to make predictions and compute the fairness metrics (Step 3) without the presence of data smells. At this point, we will have the fairness metrics computed on both datasets (with and without data smells), which will be used in the result analysis to establish if the presence of data smells has impacted the level of unfairness in our ML models.

(6) **Iterate the process**
At this point we have completed the pipeline for one dataset, subsequently, we will apply the entire pipeline (all the steps defined in Subsection 3.1) to each of the remaining datasets described earlier. This approach allows us to compile a comprehensive table with the results for each dataset by the end of the process.

(7) **Result Analysis**
In the end, we will analyze the results obtained from running the pipeline on the three datasets. The focus of this step is to investigate whether the presence of data smells has impacted the unfairness of the models. By examining the fairness metrics across the datasets, we aim to identify if and how the fairness metrics changed based on the presence of data quality issues in the input datasets. Specifically, we will compare the fairness metrics computed on raw and tidy datasets, and we will investigate whether the fairness metrics improve without the presence of data smells. This analysis is crucial to understanding the impact of data quality issues on the reliability and equity of AI systems and to address the RQ1. Specifically, we will initially examine the results obtained from a single dataset, comparing the outcomes from both the raw and tidy datasets. Afterward, the comparison will be extended to encompass all datasets, aiming to identify any consistent trends across the three benchmarks.

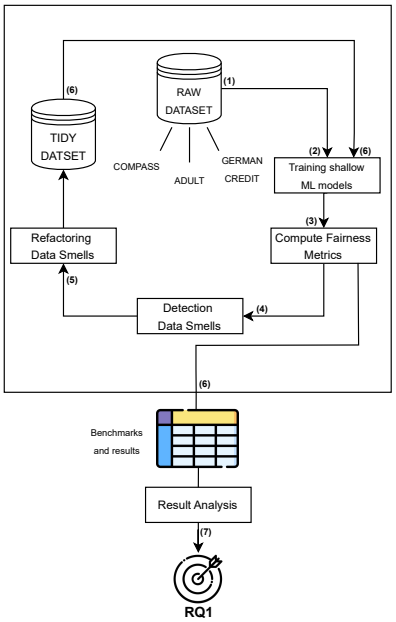

Figure 1: A summary of iterative pipeline

Figure 1 illustrates the iterative pipeline we will execute on the three datasets to achieve our goal and answer the RQ1.

3.2 Steps to address RQ2

This subsection outlines the methodological steps (second iterative pipeline) that we plan to perform to address the RQ2. This pipeline includes the same steps performed for the first iterative pipeline employed to address the RQ1. Thus, the first pipeline is a subset of the second. This pipeline includes an additional crucial step that involves the use of bias mitigation algorithms to address fairness. Indeed, this pipeline not only aims to compute the level of fairness in our models (like the first one) but also endeavors to mitigate it by performing bias mitigation algorithms. As for the first pipeline, these algorithms will be applied to both raw and tidy datasets to make comparisons on their outputs and to investigate whether the presence of data smells impacts the performances of the bias mitigation algorithms.

(1) **Dataset selection** and **Training Shallow ML models**
For the initial two phases, we follow the same procedures outlined for addressing RQ1 (items 1 and 2 in Section 3.1). Dataset selection involves all three datasets under consideration, while the models trained after step (2) in the first pipeline will be used to execute the bias mitigation algorithms. Consequently, since we already have fairness metrics outputs for each model, we proceed to compute the bias mitigation algorithms.

(2) **Perform bias mitigation algorithms on raw dataset**
To mitigate fairness in our raw models (models trained on raw datasets), we will perform a bias mitigation algorithm for every of three state-of-art approaches: **pre-, in, and, post-processing**. We will employ bias mitigation algorithms available on AI Fairness 360 (AIF-360).

(3) **Compute Fairness metrics**
After each bias mitigation algorithm is executed, we will compute the same fairness metrics described in step 3, which are based on the models. This step aims to compute fairness metrics on models trained using the raw dataset after executing the bias mitigation algorithms, which will be later used in result analysis to compare them with the same metrics calculated on the tidy dataset (without data smell).

(4) **Detection and Refactoring of data smells**
After executing bias mitigation algorithms on raw models, we aim to perform these algorithms even on the models trained on tidy dataset. However, we already have the tidy dataset yielded from Step 4 and 5 of the first pipeline (Section 3.1), So we don't have to redo these steps. Specifically, in Step 5, we have also re-trained the models on the yielded dataset, so we have already the tidy models. Thus, we can proceed with performing bias mitigation algorithms on these models to mitigate fairness.

(5) **Perform bias mitigation algorithms on tidy dataset**
At this stage, we can perform the bias mitigation algorithms on the models trained on the tidy dataset. The execution is the same as the previous one done on the models trained on the raw dataset. Likewise, there are no changes in the computation of fairness metrics. After each execution of bias mitigation algorithms, we will compute fairness metrics to examine how the fairness has been altered. This step aims to obtain the performances of bias mitigation algorithms on models trained on the tidy dataset which will be used in the result analysis step to investigate whether the presence of data smells has impacted the performances of the algorithms. The result of this step will be used to address the RQ2.

(6) **Iterate the process**
As the first pipeline, the previous steps described are applied on a single dataset. We will iterate this pipeline on the three datasets to have more benchmarks and results that are useful for the result analysis.

(7) **Result Analysis**
At this stage, we have computed metrics for both the raw dataset (with and without bias) and the tidy dataset (with and without data smells). Ultimately, we will analyze the results obtained from running the entire pipeline on the three datasets. The focus of this step is to investigate whether the presence of data smells impacts the performances of the bias mitigation algorithms (RQ2). By examining the fairness metrics across the datasets, we aim to identify if and how the performance of the algorithms has changed based on the presence of data quality issues in the input datasets.

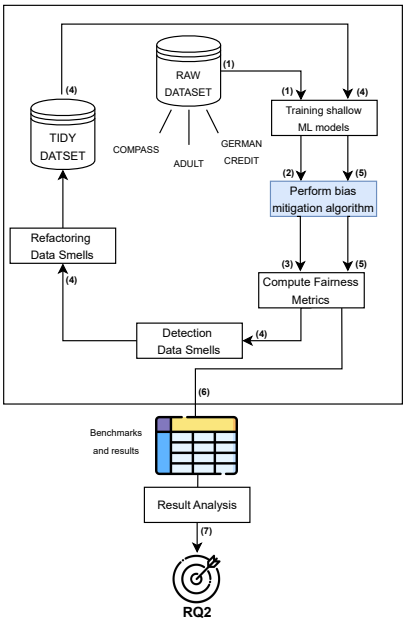

Figure 2: A summary of iterative pipeline

Figure 2 illustrates the iterative pipeline we will execute on the three datasets to achieve our goal and address the RQ2. Compared to the pipeline of the first research question, the additional step highlighted in blue involves the execution of bias mitigation algorithms.

4 METHODS AND RESULTS FOR RQ1

4.1 COMPAS

This dataset was created for an independent audit of racial biases in the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) risk assessment tool developed by *Northpointe*, which estimates the likelihood of a defendant becoming a recidivist.

The instances in this dataset are associated with two target variables (is_recid and is_violent_recid; we will consider is_recid only), indicating whether defendants were booked in jail for a criminal offense (potentially violent) that occurred after their COMPAS screening but within two years.

This dataset is used in fairness literature for various tasks, such as *fair classification, fairness evaluation, fair risk assessment, fair task assignment, fair representation learning,* and more.

Compas is known in literature because it has two **Sensitive features:**

- **sex**, privileged: **Female**, unprivileged: **Male**;
- **race**, privileged: **Caucasian**, unprivileged: **Not Caucasian**;

4.1.1 Data pre-processing. We performed minimal data preprocessing to maintain the dataset in its raw form as much as possible, only eliminating columns that were deemed irrelevant. Specifically, we removed columns that were unnecessary for our purposes. Additionally, some columns indicate data recorded if the individual was a recidivist, but we are interested in whether or not an individual is a recidivist. Thus, we have to delete them to **avoid data leakage**. This step was essential to ensure that our model training was based on relevant and high-quality data.

Lastly, rows with null values have been removed, and we have converted columns that represent dates (string values) to integer values. To prepare the remaining data for model training, we focused on encoding the categorical features. The categorical columns in the dataset included: **sex, age_cat, race, c_charge_degree, c_charge_desc**.
We utilized the *pd.get_dummies* function to convert these categorical variables into a one-hot encoded format. This transformation created binary columns for each category, which allowed us to effectively incorporate these features into our machine learning model. By using one-hot encoding, we ensured that the model could interpret the categorical data without imposing any ordinal relationship between the categories.

4.1.2 Models Training. During this phase, following the data preprocessing, we split the dataset into training and test sets using an 80/20 split. We used three of the most well-known and commonly used classification ML models: **Decision Tree**, **Support Vector Machine (SVM)**, and **Random Forest**. For each of these models, we followed a training and evaluation process to analyze their performance and compare the results.

- For the Decision Tree model, we initialized the classifier with a random state parameter to ensure the reproducibility of the results.
- For the SVM model, we used a pipeline that includes a StandardScaler to normalize the data and an SVM classifier with a linear kernel.
- For the Random Forest model, we configured the classifier with 100 decision trees and the Gini criterion for splitting.

The performance metrics for all three models were calculated using the same evaluation function. In this way, we ensured a fair comparison between the three classification models, using the same training and test data sets, as well as the same performance metrics function to evaluate the accuracy, recall, and F1 score of each model.

4.1.3 Compute Fairness Metrics. Once the model has been trained, we computed the SPD, AOD, and EOD fairness metrics using the model's predictions. Specifically, we have computed the Fairness metrics considering the following two sensitive features:

- **Sex**: privileged: **Female**, unprivileged: **Male**;
- **Race**: privileged: **Caucasian**, unprivileged: **Not Caucasian**;

and considering **value 0 as the favorable label (is not recidivist)** and **value 1 as the unfavorable label (is recidivist)**.
Table 1 shows the values yielded from the fairness metrics considering sex as sensitive feature:

Model	SPD	AOD	EOD
Decision Tree	-0.035	-0.018	-0.024
Support Vector Machine (SVM)	-0.121	-0.132	-0.061
Random Forest	-0.161	-0.174	-0.089

Table 1: Compas Fairness metrics values for sex feature

Precisely, the significance and interpretation of the results are:

- **SPD**: A value less than 0 indicates that males have a lower probability of receiving a favorable outcome (non-recidivist prediction) than females;
- **AOD**: A value less than 0 indicates that males have lower True Positive Rates (TPR) and higher False Positive Rates (FPR) compared to females;
- **EOD**: a value less than 0 indicates that males have a lower True Positive Rate (TPR) than females, meaning males have a lower probability of being correctly predicted as non-recidivist than females.

In conclusion, all the fairness metrics computed suggest that **the models are discriminatory towards males.**
Table 2 shows the values yielded from the fairness metrics considering race as sensitive feature:

Model	SPD	AOD	EOD
Decision Tree	-0.088	-0.09	-0.042
Support Vector Machine (SVM)	-0.127	-0.127	-0.075
Random Forest	-0.164	-0.169	-0.093

Table 2: Compas Fairness metrics values for race feature

In detail, the significance and interpretation of the results are:

- **SPD**: a value less than 0 indicates that Not Caucasians have a lower probability of receiving a non-recidivist prediction than Caucasians;
- **AOD**: a value less than 0 indicates that Not Caucasians have lower True Positive Rates (TPR) and higher False Positive Rates (FPR) compared to Caucasians;
- **EOD**: a value less than 0 indicates that Not Caucasians have a lower True Positive Rate (TPR) than Caucasians, meaning Not Caucasians have a lower probability of being correctly predicted as non-recidivist than Caucasians.

In conclusion, the computed fairness metrics suggest that **the models are discriminatory towards Not Caucasians.**

4.1.4 Detection of Data Smells. In this section, we delve into the process of detecting data smells within the dataset. With DSD tool we aim to systematically identify various types of data smells present in the dataset, through this tool, we can generate a categorized list of detected data smells. We have uploaded the COMPAS dataset to the DSD Tool, and it yields the following data smells shown in Table 3:

We found 5 types of data smells in the Compas dataset:

- ***Duplicated Value smell***: occurs when data values are syntactically equal across several data instances;
- ***Casing smells***: occurs when in data value s there are inconsistencies in the use of uppercase and lowercase letters;
- ***Integer As Floating Point Number Smell***: occurs when integer data values are incorrectly stored as floating point numbers;
- ***Missing Value Smell***: occurs when there are null or missing values.
- ***Extreme Value Smell***: occurs when data instances have extreme data values relating to other data instances;

Feature	Data Smell Type	Faulty Element Count
age	Duplicated Value Smell	11753
	Extreme Value Smell	41
age_cat	Duplicated Value Smell	11757
c_charge_degree	Duplicated Value Smell	11757
c_charge_desc	Casing Smell	1871
	Duplicated Value Smell	10776
	Missing Value Smell	749
c_days_from_compas	Integer As Floating Smell	11015
	Missing Value Smell	742
	Extreme Value Smell	184
c_jail_in	Missing Value Smell	1180
c_jail_out	Missing Value Smell	1180
	Duplicate Value Smell	109
c_offense_data	Missing Value Smell	2600
	Duplicated Value Smell	8886
days_b_screening_arrest	Missing Value Smell	1180
	Integer As Floating Smell	10577
	Suspect Sign Smell	474
	Extreme Value Smell	246
decile_score	Duplicated Value Smell	11757
	Suspect Sign Smell	15
is_recid	Duplicated Value Smell	11757
	Suspect Sign Smell	719
juv_fel_count	Duplicated Value Smell	11754
	Extreme Value Smell	143
juv_misd_count	Duplicated Value Smell	11756
	Extreme Value Smell	171
juv_other_count	Duplicated Value Smell	11753
	Extreme Value Smell	207
priors_count	Duplicated Value Smell	11753
	Extreme Value Smell	273
race	Duplicated Value Smell	11757
sex	Duplicated Value Smell	11757

Table 3: Data smells found in Compas dataset

- ***Suspect signal Smell***: occurs when data instances have values that are unusual, unexpected, or not logically consistent within the context of the dataset.

4.1.5 Refactoring of Data Smells. In this step of the pipeline, we begin the refactoring process aimed at addressing the data smells identified in the preceding phase. We have performed the following steps to address the data smells of table 3:

- Firstly, we have handled the Missing Value Smell: for `c_jail_in` and `c_jail_out`, we have employed a data imputation technique. However, the statistical operator can only be applied to numeric values; in this case, we have dates. So, rather than considering the dates, we considered how many days the defendant spent in jail (`c_jail_out - c_jail_in`). By doing this, we have an integer feature where we can apply the mean operator to replace the missing values. After the creation of the new feature, the attributes `c_jail_in` and `c_jail_out` were removed;
- The `c_charge_desc` attribute is defined in natural language, so it is difficult to make a deductive imputation. Thus, we have dropped it along with the missing values. We also dropped `c_offense_date`, `c_case_number`, `c_days_from_compas`, and `days_b_screening_arrest`.

Since some of them have 2000+ missing values and could potentially be irrelevant to whether a defendant has exhibited recidivism, we decided to drop these columns.
- To address the casing smell in the `c_charge_desc` attribute, we have mapped its values to lower case.
- The suspect value in the `is_recid` attribute is due to a -1 value in one row; we simply removed this row.
- To handle the Extreme Value Smell, we have applied the *MinMax normalization* technique. However, even after applying normalization, some attributes still had outliers. Thus, we removed the rows that had outlier values. Due to these eliminations, the attributes: `juv_misd_count`, `juv_fel_count`, and `juv_other_count` only contained the value 0 in all rows, making them non-discriminative. Therefore, they are useless for training the model, so we deleted them.

Once we have employed these steps, all the data smells of Table 3 have been removed. Thus, now we have the tidy Compas dataset that will be used in the next step to recompute the Fairness metrics on the dataset without data smells.

4.1.6 Compute Fairness Metrics on Tidy Compas. Once we have addressed all the data smells in the Compas dataset, we can recompute the fairness metrics to investigate if the metrics' outputs have been influenced by the presence of data smells. We have computed the fairness metrics using the same parameters as in Section 4.1.3. In Table 4, the new values obtained from the metrics on the tidy dataset considering the sex-sensitive feature are presented. The value in parentheses indicates the difference compared to the corresponding metric calculated on the raw Compas dataset previously.

Model	New SPD	New AOD	New EOD
Decision Tree	-0.043 (+23%)	-0.044 (+144%)	-0.015 (-38%)
SVM	-0.007 (-94%)	0.002 (-102%)	-0.011 (-82%)
Random Forest	-0.086 (-47%)	-0.093 (-47%)	-0.056 (-37%)

Table 4: Tidy Compas Fairness metrics values for sex feature

As can be observed from Table 4, by addressing the data smells, the metrics values improved in 7 out of 9 cases, with an average change of -31%.
In Table 5, the new values obtained from the metrics on the tidy dataset considering the race-sensitive feature are presented.

Model	New SPD	New AOD	New EOD
Decision Tree	-0.062 (-30%)	-0.06 (-33%)	-0.041 (-2%)
SVM	-0.004 (-97%)	-0.004 (-97%)	-0.001 (-99%)
Random Forest	-0.072 (-56%)	-0.079 (-53%)	-0.043 (-54%)

Table 5: Tidy Compas Fairness metrics values for race feature

As can be observed from Table 5, by addressing the data smells, the metrics values improved in 9 out of 9 cases, with an average change of -58%.

4.2 ADULT

The Adult dataset was created as a resource to benchmark the performance of machine learning algorithms on socially relevant data. Each instance of the dataset represents a person who participated in the March 1994 US Current Population Survey.

The data includes demographic and socio-economic features such as profession, education, age, sex, race, and financial condition.

The target variable for the prediction task in this dataset is a binary indicator of whether a respondent's income is above $50,000. This dataset, like COMPAS, is used in fairness literature for various tasks, with the most important being *fair classification*.

Adult is known in literature because it has two **Sensitive features:**

- **sex**, privileged: **Male**, unprivileged: **Female**;
- **race**, privileged: **White**, unprivileged: **Not White**;

4.2.1 Data pre-processing. We did not perform any data preprocessing on the Adult dataset, as all the columns were important for the classification task. To prepare data for model training we address two steps:

- we focused on encoding the categorical features of the dataset, that are: ***sex, race, workclass, education, marital-status, occupation, relationship, native-country***;
- We have encoded the *income* column, which originally contained string values indicating income levels (*<=50k* or *>50k*), into a binary column. Here, 0 represents income less than or equal to $50,000, while 1 represents income greater than $50,000.

4.2.2 Models Training. In this phase of the pipeline, we have carried out the identical steps employed for training the same models on the COMPAS dataset, as outlined in Subsection 4.1.2.

4.2.3 Compute Fairness Metrics. We have computed the SPD, AOD, and EOD Fairness metrics using the models' predictions. Specifically, we have computed the Fairness metrics considering the two sensitive features defined previously:

- **sex**, privileged: **Male**, unprivileged: **Female**;
- **race**, privileged: **White**, unprivileged: **Not White**;

and considering **value 1 as the favorable label (income > 50k)** and **0 as the unfavorable label (income <= 50k)**.
Table 6 shows the values obtained from the fairness metrics considering sex as a sensitive feature:

Model	SPD	AOD	EOD
Decision Tree	-0.195	-0.106	-0.11
Support Vector Machine (SVM)	-0.162	-0.073	-0.08
Random Forest	-0.181	-0.091	-0.104

Table 6: Adult Fairness metrics values for sex feature

Also in this case, the fairness metrics yielded negative values, meaning that the privileged group (Male) has a higher probability of receiving a favorable outcome (income > 50k) compared to the unprivileged group (Female). Thus, the metrics' outputs suggest that **the models are discriminatory towards Females**.

Table 7 shows the values obtained from the fairness metrics considering race as a sensitive feature:

Model	SPD	AOD	EOD
Decision Tree	-0.085	-0.046	-0.057
Support Vector Machine (SVM)	-0.062	-0.019	-0.024
Random Forest	-0.091	-0.044	-0.051

Table 7: Adult Fairness metrics values for race feature

The fairness metrics yielded negative values, meaning that the privileged group (White) has a higher probability of receiving an income > 50k prediction compared to the unprivileged group (Not White). Thus, the metrics' outputs suggest that **the models are discriminatory towards Not White people**.

4.2.4 Detection of Data Smells. In this phase, we performed the same operations as outlined in Section 4.1.4. We have uploaded the Adult dataset to the DSD Tool, and it yields the following data smells shown in Table 8:

Feature	Data Smell Type	Faulty Element Count
age	Integer As String Smell	48842
	Duplicated Value Smell	48841
capital-gain	Integer As Floating Smell	48842
	Missing Value Smell	1
	Extreme Value Smell	331
capital-loss	Integer As Floating Smell	48842
	Missing Value Smell	1
	Extreme Value Smell	2216
education	Duplicated Value Smell	48842
	Missing Value Smell	1
education-num	Integer As Floating Smell	48842
	Missing Value Smell	1
	Extreme Value Smell	330
fnlwgt	Integer As Floating Smell	48842
	Missing Value Smell	1
	Extreme Value Smell	506
hours-for-week	Integer As Floating Smell	48842
	Missing Value Smell	1
	Extreme Value Smell	681
income	Duplicated Value Smell	48842
	Missing Value Smell	1
marital-status	Duplicated Value Smell	48842
	Missing Value Smell	1
native-country	Duplicated Value Smell	48841
	Missing Value Smell	1
occupation	Duplicated Value Smell	48842
	Missing Value Smell	1
race	Duplicated Value Smell	48842
	Missing Value Smell	1
relationship	Duplicated Value Smell	48842
	Missing Value Smell	1
sex	Duplicated Value Smell	48842
	Missing Value Smell	1
workclass	Duplicated Value Smell	48842
	Missing Value Smell	1

Table 8: Data smells found in Adult dataset

We found 5 types of data smells in the Adult dataset: four already found previously: (**Integer As Floating Point Number Smell, Duplicated Value Smell, Missing Value smell, and Extreme Value Smell**) and one new: **Integer As String Smell** that occurs when numeric values, specifically integers, are stored as strings (text).

4.2.5 Refactoring of Data smells. In this phase, after identifying the data smells in the dataset, we initiate the refactoring process to resolve them. We have performed the following steps to address the data smells in the table 8:

- Firstly, we addressed the null values in the dataset. Upon inspection, we discovered a row where all columns contained null values. Consequently, we began by dropping this particular entry from the dataset.
- The `age` attribute had the issue where its numerical data was improperly stored as a string, exhibiting the *Integer as String* smell. We resolved this by using the `pd.to_numeric` function from *pandas* to convert the column into numerical values.
- The attributes `age`, `capital-gain`, `capital-loss`, `education-num`, `fnlwgt`, and `hours-per-week` exhibited the *Integer as Floating Point* smell. To resolve this, we converted all these columns to the `int64` data type.
- To handle the Extreme Value Smell, we removed the rows containing outlier values for some attributes. Despite these eliminations, the attribute `hours-per-week` still has outliers. Additionally, the columns `capital-gain` and `capital-loss` contained only the value 0 in all rows, making them non-discriminative. Consequently, we deleted these columns as they were useless for training the model.

Once we have employed these steps, all the data smells, except the column *hours-per-week* of Table 8 have been removed. Thus, now we have the tidy Adult dataset that will be used in the next step to recompute the Fairness metrics on the dataset without data smells.

4.2.6 Compute Fairness Metrics on Tidy Adult. After addressing the data smells, we proceeded to retraining and recalculate the fairness and performance metrics on the tidy dataset devoid of any data smells to investigate if the metrics' outputs have been influenced by the presence of data smells. We have computed the fairness metrics using the same parameters as in Section 4.2.3.

In Table 9, the new values obtained from the metrics on the tidy dataset considering the *sex* sensitive feature are presented. The value in parentheses indicates the difference compared to the corresponding metric calculated on the raw Adult dataset previously.

Model	New SPD	New AOD	New EOD
Decision Tree	-0.155 (-20%)	-0.088 (-17%)	-0.074 (-33%)
SVM	-0.122 (-25%)	0.066 (-10%)	-0.064 (-20%)
Random Forest	-0.015 (-92%)	-0.075 (-18%)	-0.066 (-37%)

Table 9: Tidy Adult Fairness metrics values for sex feature

As can be observed from Table 9, by addressing the data smells, the metrics values improved in 9 out of 9 cases, with an average change of -30%.

In Table 10, the new values obtained from the metrics on the tidy dataset considering the *race* sensitive feature are presented.

Model	New SPD	New AOD	New EOD
Decision Tree	-0.081(-5%)	-0.057(+24%)	-0.068(+19%)
SVM	-0.047(-24%)	-0.001(-95%)	0.017(-170%)
Random Forest	-0.084(-8%)	-0.052(+18%)	-0.061(+20%)

Table 10: Tidy Adult Fairness metrics values for race feature

As can be observed from Table 10, by addressing the data smells, the metrics values improved in 5 out of 9 cases, with an average change of -25%.

4.3 GERMAN CREDIT

The German Credit dataset was created to study the problem of automated credit decisions at a regional Bank in southern Germany. Instances represent loan applicants from 1973 to 1975, who were deemed credit-worthy and were granted a loan, bringing about a natural selection bias. A binary variable encoding whether each loan recipient punctually paid every installment is the target of a classification task.

The target variable in this dataset is a binary indicator representing an individual's credit risk for the prediction task.

This dataset, like COMPAS and Adult, is used in fairness literature for various tasks, with the most important being fair classification.

German Credit is known in literature because it has two **Sensitive features**:

- **sex**, privileged **Male**, unprivileged: **Female**;
- **age**, privileged **Old**, unprivileged: **Young**;

4.3.1 Data Pre-processing. The German credit dataset doesn't have any NULL or missing values. Furthermore, all the columns were deemed important for the classification task, so none of them were removed. However, to prepare the data for model training, we addressed three steps:

- The main issue with this dataset is that categorical attributes are encoded to represent discrete values. For example, in the sex column, the encoded values *A91, A93*, and *A94* represent Males, while *A92*, and *A95* represent Females. However, we mapped only the sex column to the actual values to compute the Fairness metrics correctly;
- The dataset contains the column *age in years* which represents ages as integer values. However, to compute the Fairness metrics, we need the values *Old* and *Young*. Thus, we created a new column named *age* where if the integer age value is less than or equal to 25, it is mapped as *Young*, and for integer age values greater than 25, it is mapped as *Old*;
- Lastly, we encoded the categorical features using the one-hot encoding technique.

4.3.2 Models Training. In this phase of the pipeline, we carried out the same steps employed for the previous two datasets and trained the same models, as outlined in Subsection 4.1.2.

4.3.3 Compute Fairness Metrics. We have computed the SPD, AOD, and EOD Fairness metrics using the models' predictions. Specifically, we have computed the Fairness metrics considering the two sensitive features defined previously:

- **sex**, privileged **Male**, unprivileged: **Female**;
- **age**, privileged **Old**, unprivileged: **Young**;

Table 11 shows the values obtained from the fairness metrics considering sex as a sensitive feature:

Model	SPD	AOD	EOD
Decision Tree	-0.012	-0.005	-0.015
Support Vector Machine (SVM)	-0.021	-0.051	0.035
Random Forest	0.023	0.0	0.066

Table 11: German Credit Fairness metrics values for sex feature

In this case, the Fairness metrics yielded an unexpected result compared to what we had anticipated. Not all the output values are negative, indicating that Males are privileged over Females. Instead, there are more positive values, indicating that Females are privileged over Males, and even a value of 0, which suggests that there is no discrimination.
However, all the values are close to 0. Since all the values are approximately zero, they can be considered negligible, indicating that **there is no significant discrimination present.** Table 12 shows the values obtained from the fairness metrics considering sex as a sensitive feature:

Model	SPD	AOD	EOD
Decision Tree	-0.089	-0.055	-0.127
Support Vector Machine (SVM)	-0.379	-0.447	-0.242
Random Forest	-0.276	-0.143	-0.343

Table 12: German Credit Fairness metrics values for age feature

The fairness metrics yielded negative values, indicating that the privileged group (Old) has a higher probability of being categorized as having good credit compared to the unprivileged group (Young). These results reveal significant disparities between privileged and unprivileged groups, with particularly unfavorable outcomes for the unprivileged group. Therefore, the metrics' outputs suggest that **the models are discriminatory towards young people.**

4.3.4 Detection of data smells. In this phase, we performed the same operations as outlined in Section 4.1.4. We have uploaded the German Credit dataset to the DSD Tool, and it yields the following data smells shown in Table 13:

The only data smell identified is the *Duplicated Value Smell*. Unfortunately, all the features affected by the *Duplicated Value Smell* are categorical features. These smells will be resolved once we apply the *one-hot encoding* technique to these features. However, for training the raw part of this dataset, we have already applied this technique. **This means that the raw dataset used for training the models and the tidy dataset (raw dataset without the duplicated value smell) are the same**. Consequently, **the German**

Feature	Data Smell Type	Faulty Element Count
Housing	Duplicated Value Smell	1000
Job	Duplicated Value Smell	1000
Property	Duplicated Value Smell	1000
Purpose	Duplicated Value Smell	1000
Target	Duplicated Value Smell	1000
Telephone	Duplicated Value Smell	1000
age	Duplicated Value Smell	1000
sex	Duplicated Value Smell	1000

Table 13: Data smells found in German Credit dataset

credit dataset cannot be used for this study because it does not have enough data smells.
However, we have tested other datasets known to be unfair, and after some tests, **we have selected the *Bank Marketing Dataset* to replace the German Credit Dataset for the entire project.**

4.4 BANK MARKETING

The Bank Marketing Dataset was produced to support a study of success factors in telemarketing of long-term deposits within a Portuguese bank, with data collected over the period 2008-2010. The classification target is a binary variable indicating client subscription to a term deposit.
It is known in literature because it has one **Sensitive Feature**: **age**.

4.4.1 Data Pre-processing. The Bank Marketing dataset does not have any NULL or missing values, and none of the columns have been removed. To prepare the data for model training, we have addressed three steps:

- Mapped the target feature with value 1 if its original value was 'yes', and with value 0 if its original value was 'no';
- Mapped the age feature as a categorical feature, rather than a numeric feature, with three possible values: age < 25, 25 <= age < 60, and age >= 60, because we will need these three age intervals for computing the Fairness metrics later;
- Lastly, we have encoded the categorical features using the *one-hot encoding* technique.

4.4.2 Models Training. In this phase of the pipeline, we carried out the same steps employed for the previous two datasets and trained the same models, as outlined in Subsection 4.1.2.

4.4.3 Compute Fairness Metrics. We have computed the SPD, AOD, and EOD Fairness metrics using the models' predictions. Specifically, we have computed the Fairness metrics considering the sensitive feature defined previously:

- **age**, privileged **25 <= age < 60**, unprivileged: **age < 25 or age >= 60**

Table 14 shows the values obtained from the Fairness metrics considering age as a sensitive feature:

Model	**SPD**	**AOD**	**EOD**
Decision Tree	0.316	0.064	0.142
Support Vector Machine (SVM)	0.359	0.088	0.166
Random Forest	0.362	0.064	0.162

Table 14: Bank Marketing Fairness metrics values for age feature

All the Fairness metrics have yielded positive values, meaning that the unprivileged group (age < 25 or age >= 60) has a higher probability of being predicted with the favorable label (*deposit = yes*) compared to the privileged group (25 <= age < 60). Therefore, the metrics' outputs suggest that **the models are discriminatory towards people who are 25 to 59 years old**.

4.4.4 Detection of data smells. Also in this phase, we performed the same operations as outlined in Section 4.1.4. We have uploaded the Bank Marketing dataset on the DSD Tool, and it yields the following data smells shown in Table 15:

Feature	Data Smell Type	Faulty Element Count
age	Extreme Value Smell	132
	Duplicated Value Smell	11160
balance	Suspect Sign Smell	688
	Extreme Value Smell	173
	Duplicated Value Smell	9341
campaign	Extreme Value Smell	210
	Duplicated Value Smell	11156
contact	Duplicated Value Smell	11162
day	Duplicated Value Smell	11162
default	Duplicated Value Smell	11162
deposit	Duplicated Value Smell	11162
duration	Extreme Value Smell	201
	Duplicated Value Smell	10832
education	Duplicated Value Smell	11162
housing	Duplicated Value Smell	11162
job	Duplicated Value Smell	11162
loan	Duplicated Value Smell	11162
marital	Duplicated Value Smell	11162
month	Duplicated Value Smell	11162
pdays	Extreme Value Smell	176
	Duplicated Value Smell	11028
poutcome	Duplicated Value Smell	11162
previous	Extreme Value Smell	220
	Duplicated Value Smell	11150

Table 15: Data smells found in Bank marketing dataset

We found 3 types of data smells in the bank dataset that were already identified previously: **Extreme Value Smell**, **Duplicated Value Smell** and **Suspect Sign Smell**.

4.4.5 Refactoring of Data smells. In this phase, after identifying the data smells, we began the refactoring process to resolve them. We have performed the following steps to address the data smell in the table 15:

- Initially, we examined the attributes *pdays* and *previous*, noting that most of the values in these attributes were 0. Since these attributes did not provide meaningful information for training the model, we decided to remove them from our analysis.
- To handle the Extreme Value Smell, we used the MinMax normalization technique. However, despite the application of this technique, some attributes continued to have outliers. Consequently, to eliminate them, we removed the rows that contained outlier values.
- Regarding the management of the data smell suspect sign, the elimination of outliers led to the removal of this data smell.

Once we have employed these steps, all the data smells, have been removed. Now, we have the tidy Bank dataset that will be used in the next step to recompute the Fairness metrics on the dataset without the data smells.

4.4.6 Compute Fairness Metrics on Tidy Bank. After correcting the data smells, we re-examined the equity and performance metrics using the dataset cleaned of these issues. This allowed us to assess whether the results of the metrics had been skewed by the presence of the data smells. Fairness metrics were calculated using the same parameters as in section 4.4.3. In Table 16, there are the new values obtained from the metrics on the tidy dataset, considering the sensitive age attribute.

Model	New SPD	New AOD	New EOD
Decision Tree	0.44 (+39%)	0.152 (+138%)	0.302 (+112%)
SVM	0.554 (+54%)	0.287 (+226%)	0.324 (+95%)
Random Forest	0.568 (+57%)	0.28 (+338%)	0.33 (+104%)

Table 16: Bank Marketing Fairness metrics values for age feature

As can be observed from Table 16, by addressing the data smells, all the metrics values have increased, moving further away from zero, with an average change of +129%. We believe that this significant increase is due to the removal of outliers. In fact, to remove all the outliers, approximately 40% of the rows in the dataset were deleted.

4.5 Result analysis

In this subsection, we are going to analyze the Fairness metrics obtained from raw and tidy datasets. This analysis is going to help us understand the extent to which these metrics are influenced by the presence of data smells within our datasets.

We have considered three datasets: Compas, Adult, and Bank Marketing, excluding German Credit for the reasons specified earlier.

After removing the data smells, all the metrics underwent positive and negative changes. Most of them showed improvement, indicating that the data smells impacted the calculation of fairness metrics. Specifically, In the COMPAS and Adult datasets, on average, the metrics have improved, getting closer to 0. For Bank Marketing Dataset, all the metrics have worsened.

Answer to RQ$_1$. *To summarize the results, after we addressed the data smells from the datasets, all the Fairness metrics outputs changed. Thus, The presence of data smells has influenced each Fairness metric calculated, either positively or negatively.*

5 METHODS AND RESULTS FOR RQ2

After addressing RQ1 and discovering that data smells have influenced the presence of Fairness in our models, our aim now is to investigate whether data smells could also influence the performance of bias mitigation algorithms. As defined previously, to address RQ2, we are going to apply pre-, in-, and post-processing algorithms to our models trained on both raw and tidy datasets. We will then analyze and compare the fairness metrics obtained after employing these algorithms. We are going to use these three bias mitigation algorithms available on the AIF-360 website:

- **Reweighing:** is a pre-processing technique that Weights the examples in each (group, label) combination differently to ensure fairness before classification:
- **Exponentiated Gradient Reduction:** is an in-processing technique that reduces fair classification to a sequence of cost-sensitive classification problems, returning a randomized classifier with the lowest empirical error subject to fair classification constraint;
- **Equalized odds:** is a post-processing technique that solves a linear program to find probabilities with which to change output labels to optimize equalized odds.

These algorithms will be applied to models used to address the RQ1, thus, we will not train new models.

5.1 COMPAS

In this step, we are going to apply the three algorithms on models that have been trained on raw Compas Dataset (Section 4.1.2) and tidy Compas Dataset (Section 4.1.5).

5.1.1 pre-processing algorithm. Firstly, we have applied the Reweighing algorithm considering the sex feature. The original Fairness values are in Table 1 and Table 4. In Table 17 there are the new values obtained from the Fairness metrics on the raw and tidy dataset considering the sex feature after the Reweighing algorithm has been applied:

Raw Model	SPD	AOD	EOD
Decision Tree	-0.03	-0.013	-0.009
SVM	-0.079	-0.087	-0.027
Random Forest	-0.135	-0.141	-0.071
Tidy Model	**SPD**	**AOD**	**EOD**
Decision Tree	-0.018	-0.03	0.017
SVM	-0.01	0.002	-0.014
Random Forest	-0.08	-0.085	-0.045

Table 17: Compas Fairness metrics on sex feature after Reweighting alghoritm

The algorithm has performed well on the raw decision tree, with values close to 0, even better than the tidy decision tree. However, for the other raw models, the metrics do not improve significantly compared to the original ones.
Then, we have done the same on the race feature. The original Fairness values for race feature are in Table 2 and Table 5. In Table 18 there are the new values obtained from the Fairness metrics on the raw and tidy dataset considering the race feature after the Reweighing algorithm has been applied:

Raw Model	**SPD**	**AOD**	**EOD**
Decision Tree	-0.113	-0.115	-0.052
SVM	-0.117	-0.115	-0.069
Random Forest	-0.155	-0.153	-0.092
Tidy Model	**SPD**	**AOD**	**EOD**
Decision Tree	-0.054	-0.054	-0.034
SVM	-0.006	-0.005	-0.002
Random Forest	-0.073	-0.079	-0.041

Table 18: Compas Fairness metrics on race feature after Reweighting alghoritm

5.1.2 in-processing algorithm. Afterwards, we applied the Exponentiated Gradient Reduction algorithm to the sex feature. The original Fairness values are in Table 1 and Table 4. In Table 19 there are the new values obtained from the Fairness metrics on the raw and tidy dataset considering the sex feature after the Exponentiated Gradient Reduction algorithm has been applied:

Raw Model	**SPD**	**AOD**	**EOD**
Decision Tree	0.024	0.016	0.066
SVM	-0.016	-0.011	0.027
Random Forest	-0.022	-0.011	0.012
Tidy Model	**SPD**	**AOD**	**EOD**
Decision Tree	-0.005	-0.026	0.038
SVM	-0.034	-0.015	-0.022
Random Forest	-0.034	-0.015	-0.022

Table 19: Compas Fairness metrics on sex feature after Exponentiated GR alghoritm

Then, we have done the same on the race feature. The original Fairness values for race feature are in Table 2 and Table 5. In Table 20 there are the new values obtained from the Fairness metrics on the raw and tidy dataset considering the race feature after the Exponentiated Gradient Reduction algorithm has been applied:

Raw Model	**SPD**	**AOD**	**EOD**
Decision Tree	-0.02	-0.011	0.0
SVM	-0.031	0.001	-0.026
Random Forest	-0.028	0.003	-0.019
Tidy Model	**SPD**	**AOD**	**EOD**
Decision Tree	0.031	0.06	0.012
SVM	-0.044	-0.024	-0.034
Random Forest	-0.038	-0.015	-0.036

Table 20: Compas Fairness metrics on race feature after Exponentiated GR alghoritm

5.1.3 post-processing algorithm. Lastly, we have applied the Equalized odds algorithm considering the sex feature. The original Fairness values are in Table 1 and Table 4. In Table 21 there are the new values obtained from the Fairness metrics on the raw and tidy dataset considering the sex feature after the Equalized odds algorithm has been applied:

Raw Model	**SPD**	**AOD**	**EOD**
Decision Tree	-0.015	0.002	-0.004
SVM	-0.01	-0.005	0.003
Random Forest	-0.01	-0.001	-0.001
Tidy Model	**SPD**	**AOD**	**EOD**
Decision Tree	-0.01	0.0	0.0
SVM	0.0	0.007	-0.001
Random Forest	-0.007	0.002	-0.002

Table 21: Compas Fairness metrics on sex feature after Equalized odds alghoritm

Then, we have done the same on the race feature. The original Fairness values for race feature are in Table 2 and Table 5. In Table 22 there are the new values obtained from the Fairness metrics on the raw and tidy dataset considering the race feature after the Equalized odds algorithm has been applied:

Raw Model	**SPD**	**AOD**	**EOD**
Decision Tree	-0.01	0.001	-0.003
SVM	-0.014	-0.001	-0.004
Random Forest	-0.014	0.002	0.004
Tidy Model	**SPD**	**AOD**	**EOD**
Decision Tree	-0.009	0.004	-0.009
SVM	-0.001	0.001	-0.001
Random Forest	-0.006	-0.001	0.002

Table 22: Compas Fairness metrics on race feature after Equalized odds alghoritm

5.2 ADULT

In this step of the pipeline, we used both the raw Adult dataset with data smells and the tidy Adult dataset without data smells. The raw dataset was pre-processed as described in Subsection 4.2.1,

while the tidy dataset was prepared by applying the steps outlined in Subsection 4.2.5. Subsequently, we applied all three previously described techniques: pre-processing, in-processing, and post-processing. This comprehensive approach ensures a robust evaluation of the dataset's performance and fairness, both with and without the presence of data smells.

5.2.1 pre-processing algorithm. In the pre-processing stage, we applied Reweighing algorithm to mitigate bias in the input data before training the models. We focused on the sex feature. The original Fairness values are in Table 6 and Table 9. In Table 23 there are the new values obtained from the Fairness metrics on the raw and tidy dataset considering the sex feature after the Reweighing algorithm has been applied:

Raw Model	**SPD**	**AOD**	**EOD**
Decision Tree	-0.189	-0.086	-0.068
SVM	-0.046	0.121	0.234
Random Forest	-0.184	-0.089	0.097
Tidy Model	**SPD**	**AOD**	**EOD**
Decision Tree	-0.17	-0.111	-0.11
SVM	0.06	0.268	-0.484
Random Forest	-0.156	-0.092	-0.096

Table 23: Adult Fairness metrics on sex feature after Reweighting alghoritm

Then, we have done the same on the race feature. The original Fairness values for race feature are in Table 7 and Table 10. In Table 24 there are the new values obtained from the Fairness metrics on the raw and tidy dataset considering the race feature after the Reweighing algorithm has been applied:

Raw Model	**SPD**	**AOD**	**EOD**
Decision Tree	-0.06	-0.007	0.002
SVM	-0.026	0.038	0.063
Random Forest	-0.088	-0.037	-0.038
Tidy Model	**SPD**	**AOD**	**EOD**
Decision Tree	-0.079	-0.049	-0.053
SVM	-0.043	0.008	0.033
Random Forest	-0.08	-0.045	-0.049

Table 24: Adult Fairness metrics on race feature after Reweighting alghoritm

5.2.2 in-processing algorithm. The in-processing stage involved modifying the learning algorithms to account for fairness during the model training process. We applied the Exponentiated Gradient Reduction algorithm to the sex feature, while training the models. The original Fairness values are in Table 6 and Table 9. In Table 25 there are the new values obtained from the Fairness metrics on the raw and tidy dataset considering the sex feature after the Exponentiated Gradient Reduction algorithm has been applied:

Raw Model	**SPD**	**AOD**	**EOD**
Decision Tree	-0.007	0.044	0.035
SVM	-0.018	0.139	0.271
Random Forest	0.013	0.033	-0.026
Tidy Model	**SPD**	**AOD**	**EOD**
Decision Tree	-0.011	0.054	0.022
SVM	-0.008	0.148	0.262
Random Forest	-0.013	0.025	-0.056

Table 25: Adult Fairness metrics on sex feature after Exponentiated GR alghoritm

Then, we have done the same on the race feature. The original Fairness values for race feature are in Table 7 and Table 10. In Table 26 there are the new values obtained from the Fairness metrics on the raw and tidy dataset considering the race feature after the Exponentiated Gradient Reduction algorithm has been applied:

Raw Model	**SPD**	**AOD**	**EOD**
Decision Tree	-0.008	-0.013	-0.063
SVM	-0.019	0.03	0.047
Random Forest	0.011	0.001	-0.061
Tidy Model	**SPD**	**AOD**	**EOD**
Decision Tree	-0.023	0.01	-0.007
SVM	0.01	0.068	0.093
Random Forest	0.025	0.047	0.004

Table 26: Adult Fairness metrics on race feature after Exponentiated GR alghoritm

5.2.3 post-processing algorithm. Lastly, Post-processing techniques were applied to the trained models to adjust their predictions and enhance fairness. In this stage, we have applied the Equalized odds algorithm considering the sex feature. The original Fairness values are in Table 6 and Table 9. In Table 27 there are the new values obtained from the Fairness metrics on the raw and tidy dataset considering the sex feature after the Equalized odds algorithm has been applied:

Raw Model	**SPD**	**AOD**	**EOD**
Decision Tree	-0.082	0.009	0.019
SVM	-0.086	0.006	0.013
Random Forest	-0.09	0.0	0.0
Tidy Model	**SPD**	**AOD**	**EOD**
Decision Tree	-0.055	0.011	0.024
SVM	-0.053	-0.002	0.005
Random Forest	-0.065	0.007	-0.017

Table 27: Adult Fairness metrics on sex feature after Equalized odds alghoritm

Then, we have done the same on the race feature. The original Fairness values for race feature are in Table 7 and Table 10. In Table 28 there are the new values obtained from the Fairness metrics

on the raw and tidy dataset considering the race feature after the Equalized odds algorithm has been applied:

Raw Model	SPD	AOD	EOD
Decision Tree	-0.043	0.0	-0.001
SVM	-0.045	0.002	0.002
Random Forest	0.048	0.004	0.011
Tidy Model	**SPD**	**AOD**	**EOD**
Decision Tree	-0.029	0.011	0.024
SVM	-0.028	-0.001	-0.006
Random Forest	-0.033	0.003	0.005

Table 28: Adult Fairness metrics on race feature after Equalized odds alghoritm

5.3 BANK MARKETING

In this step of the pipeline, we used both the raw and tidy versions of the Bank dataset. The raw dataset was pre-processed as per subsection 4.4.1, while the tidy dataset followed the steps in subsection 4.4.5. We then applied the three techniques: pre-processing, in-processing, and post-processing.

5.3.1 pre-processing algorithm. Initially, in the pre-processing phase, the Reweighing algorithm was applied on both the raw dataset and the tidy dataset. The original Fairness values for age feature are in Table 14 and Table 16.In Table 29, we have the results obtained considering the age feature on both the raw dataset an the tidy dataset.

Raw Model	SPD	AOD	EOD
Decision Tree	0.257	-0.009	0.126
SVM	0.075	-0.207	-0.106
Random Forest	0.319	0.032	0.103
Tidy Model	**SPD**	**AOD**	**EOD**
Decision Tree	0.377	0.126	0.159
SVM	0.23	-0.013	-0.042
Random Forest	0.512	0.217	0.253

Table 29: Bank Marketing Fairness metrics on age feature after Reweighting alghoritm

5.3.2 in-processing algorithm. Secondly, in the in-processing step, we have applied the Exponentiated Gradient Reduction alghoritm to the raw dataset and to the tidy dataset. The original Fairness values are in Table 14 and Table 16. In Table 30, we have the results obtained considering the age feature on both the raw dataset and the tidy dataset.

Raw Model	SPD	AOD	EOD
Decision Tree	0.034	-0.251	-0.081
SVM	-0.045	-0.342	-0.197
Random Forest	-0.004	-0.345	-0.123
Tidy Model	**SPD**	**AOD**	**EOD**
Decision Tree	0.006	-0.284	-0.139
SVM	0.065	-0.244	-0.121
Random Forest	0.072	-0.238	-0.127

Table 30: Bank Marketing Fairness metrics on age feature after Exponentiated GR alghoritm

5.3.3 post-processing algorithm. In the last, in the post-processing step, we applied the Equalized odds alghoritm to the raw dataset and to the tidy dataset. The original Fairness values are in Table 14 and Table 16. In Table 31, we have the results obtained considering the age feature on both the raw dataset and the tidy dataset.

Raw Model	SPD	AOD	EOD
Decision Tree	0.164	-0.012	-0.001
SVM	0.169	-0.001	0.001
Random Forest	0.187	0.015	-0.016
Tidy Model	**SPD**	**AOD**	**EOD**
Decision Tree	0.097	0.001	0.001
SVM	0.182	-0.025	0.011
Random Forest	0.186	-0.029	0.032

Table 31: Bank Marketing Fairness metrics on age feature after Equalized odds alghoritm

5.4 Result Analysis

After analyzing all the tables, we found some interesting trends that helped us to investigate whether the presence of data smells could influence the performances of the algorithms that we have employed. The trends found are consistent across all three datasets. They are:

- the pre-processing Reweighting is the algorithm that performed the worst regardless of the type of model used (raw and tidy). After it was applied, the metrics improved only slightly compared to the original ones;
- On the other hand, The post-processing Equalized Odds algorithm is the algorithm that performed the best. In fact, in most cases, it returned values equal to or very close to 0, eliminating fairness issues within the models, regardless of whether it was applied to models trained on the raw dataset or the tidy dataset.

Thus, considering these trends and the results obtained from the 3 datasets that we have used during this project, the data smells don't influence the performance of bias mitigation algorithms. This means that the performance of the algorithms did not change depending on whether they were applied to the raw dataset or the tidy dataset. The Fairness metrics calculated after applying the algorithms always changed in the same proportion, regardless of

whether the algorithms were applied to the raw or tidy models. However, we have noticed that there are some types of algorithms that work very well (post-processing) and others that don't make much difference (pre-processing).

Answer to RQ$_2$. *To summarize the results, after employing bias mitigation algorithms on raw and tidy models, there is no evidence that the performance of these algorithms changed based on the presence of data smells. This is because the fairness metrics calculated after applying the algorithms always changed in the same proportion, regardless of whether the algorithms were applied to the raw or tidy datasets.*

6 CONCLUSION

In this study, we explored the impact of data smells on fairness issues in machine learning (ML) solutions. We aimed to address two primary research questions: whether the presence of data quality issues (data smells) affects the fairness of AI decision-making and how these data smells impact the performance of bias mitigation algorithms. The project involved a comprehensive analysis using state-of-the-art datasets known for inducing unfairness in ML models, such as the COMPAS, Adult, German Credit, and Bank Marketing datasets.

We employed various fairness metrics, in detail we used AOD, EOD, SPD. We also used data smell detection tools to evaluate and refactor the datasets. Through a detailed methodology that included pre-processing, in-processing, and post-processing bias mitigation techniques, we analyzed the performance of models trained on both raw and tidy datasets.

Our findings indicate that the presence of data smells significantly impacts the fairness metrics of ML models, demonstrating that addressing these data quality issues is crucial for developing fair AI systems (4.5).

Additionally, the performance of bias mitigation algorithms was not significantly influenced by data smells (5.4).

This study emphasizes the importance of understanding the relationship between data quality and fairness to develop more equitable and reliable AI systems.

7 FUTURE WORK

In the future, it would be worthwhile to use a larger number of datasets to verify the validity of RQ1 and RQ2. Specifically, it would be interesting to determine whether RQ1 still holds with a larger number of datasets and thus verify if the presence of data smells indeed affects (mostly negatively) the fairness of the model. Even more interesting would be to verify RQ2, as no evidence was found across three datasets to prove that the performance of bias mitigation algorithms was influenced by the presence or absence of data smells.

Additionally, it would be valuable to conduct a study on which type of algorithm performs better than others, given that in this project, the post-processing algorithm performed the best across all three datasets compared to other types.

REFERENCES

[1] V.R. Basili. 1992. *Software Modeling and Measurement: The Goal/question/metric Paradigm.* University of Maryland. https://books.google.it/books?id=GcpwAACAAJ

[2] Rachel KE Bellamy, Kuntal Dey, Michael Hind, Samuel C Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilović, et al. 2019. AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development* 63, 4/5 (2019), 4–1.

[3] Rachel K. E. Bellamy, Kuntal Dey, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, Seema Nagar, Karthikeyan Natesan Ramamurthy, John Richards, Diptikalyan Saha, Prasanna Sattigeri, Moninder Singh, Kush R. Varshney, and Yunfeng Zhang. 2018. AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias. (2018).

[4] Simon Caton and Christian Haas. 2024. Fairness in Machine Learning: A Survey. *ACM Comput. Surv.* 56, 7, Article 166 (apr 2024), 38 pages. https://doi.org/10.1145/3616865

[5] Zhenpeng Chen, Jie M Zhang, Federica Sarro, and Mark Harman. 2024. Fairness Improvement with Multiple Protected Attributes: How Far Are We?. In *Proceedings of the IEEE/ACM 46th International Conference on Software Engineering*. 1–13.

[6] Alessandro Fabris, Stefano Messina, Gianmaria Silvello, and Gian Antonio Susto. 2022. Algorithmic fairness datasets: the story so far. *Data Mining and Knowledge Discovery* 36, 6 (2022), 2074–2152.

[7] Harald Foidl, Michael Felderer, and Rudolf Ramler. 2022. Data smells: categories, causes and consequences, and detection of suspicious data in AI-based systems. (2022), 229–239.

[8] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)* 54, 6 (2021), 1–35.

[9] Gilberto Recupito, Raimondo Rapacciuolo, Dario Di Nucci, and Fabio Palomba. 2023. Unmasking Data Secrets: An Empirical Investigation into Data Smells and Their Impact on Data Quality. (2023).