# Data Quality and Data Ethics: Towards a Trade-off Evaluation

Fabio Azzalini[1], Cinzia Cappiello[1], Chiara Criscuolo[1], Camilla Sancricca[1] and Letizia Tanca[1]

[1]Politecnico di Milano, Milan, Italy

**Abstract**

In the last decades, one of the main drivers for organizational success has been data-driven decision-making: strategic decisions are based on data analysis and interpretation. In this scenario, relying on dependable results becomes imperative. Therefore we must ensure that input data have good quality and the algorithms on which the analysis is based are fair: in general, Data Quality (DQ) and Data Ethics (DE) should be guaranteed.

However, maximizing DQ and DE simultaneously is non-trivial, since DQ improvement techniques can negatively affect DE and vice versa. Discovering which relationships exist between DQ and DE and thoroughly analyzing it is therefore of paramount importance. The goal of this paper is to study whether, in a given context, there is a trade-off between DQ and DE: specifically, we consider the Completeness dimension of DQ, and the Fairness dimension of DE. The results of our experiments, based on two real-world well-known datasets, provided details about this trade-off and allowed us to draw some guidelines.

**Keywords**

Data Quality, Data Ethics, Fairness

## 1. Introduction

In the last decades, data-driven culture spread in several domains. The availability of large amounts of data and algorithms has made our lives more efficient and easier, and strategic decisions are made based on data analysis and interpretation; therefore, relying on dependable results becomes imperative. We need to be sure that the data sources have good quality and the algorithms on which the analysis is based are fair and do not introduce bias in the decision process.

In fact, on the one hand, the performance of Machine Learning (ML) algorithms may be, for example, seriously affected by the poor quality of the training data [1]: *inaccurate, incomplete,* and *inconsistent* data may produce poor analysis results. Therefore, in addition to the well-known storage and processing problems related to data collection, addressing *Data Quality* (DQ) has become a fundamental issue [2, 3]. The most used DQ dimensions are Accuracy, Completeness, Consistency, and Timeliness [2]: *Accuracy* is the extent to which data are correct, reliable and certified; *Completeness* is the degree to which a given data collection includes the data describing the corresponding set of real-world objects; *Consistency* is the satisfaction of semantic rules defined over a set of data items; and *Timeliness* expresses how current the data are for the task at hand.

On the other hand, when Data Science is used to build decision-making tools that impact people's lives, the problem of *Data Ethics* (DE) becomes critically important. Even the most accurate application for collecting data might be affected by ethical issues since also high-quality data might lead to unfair outcomes. In [4], the authors note that, for Data Science to be reliable, DQ should also include some ethical dimensions because, in many critical fields, data can be considered of good quality *only if compliant with high ethical standards.* The authors propose to include the most common ethical requirements among the dimensions of quality, grouped in an Ethics Cluster: *Fairness*, defined as *the lack of bias*, since an algorithmic bias might result from training a system with biased data; *Transparency*, the possibility to control the knowledge extraction process to verify the reasons of the results; *Diversity*, the degree to which different kinds of objects are represented in a dataset; and finally, *Data Protection* that concerns the ways to protect data, algorithms, and models from unauthorized access.

It is already well known that there may be contrasting objectives also among the dimensions of DE, for instance, between Transparency and Data Protection. In the same way, the relationship between the DQ dimensions [2], and the ethical ones is complex. For example, commonly used DQ improvement techniques –e.g., imputing missing values using the mean value– might modify the overall distribution of values in the dataset, leading to a reduc-

*Joint Workshops at 49th International Conference on Very Large Data Bases (VLDBW'23) — the 12th International Workshop on Quality in Databases (QDB'23), August 28 - September 1, 2023, Vancouver, Canada*

fabio.azzalini@polimi.it (F. Azzalini); cinzia.cappiello@polimi.it (C. Cappiello); chiara.criscuolo@polimi.it (C. Criscuolo); camilla.sancricca@polimi.it (C. Sancricca); letizia.tanca@polimi.it (L. Tanca)

0000-0003-0631-2120 (F. Azzalini); 0000-0001-6062-5174 (C. Cappiello); 0000-0002-1345-2482 (C. Criscuolo); 0000-0002-3820-7870 (C. Sancricca); 0000-0003-2607-3171 (L. Tanca)

tion of Fairness; on the other hand, some Bias Mitigation techniques modify real data values to remove unfairness, thus lowering Accuracy, which is a fundamental dimension of DQ. However, there are also contexts in which the user does not care about Fairness, like in the analysis of sensors data or in forecasting raw-material prices. In these cases, we do not have protected attributes (e.g., sex, race, ethnicity, etc.) and not even proxy ones (e.g., education, zip code, etc.). Moreover, in some applications, differences in treatment and outcomes among different groups are justified and explained: for example, disproportional recruitment rates for males and females might be explained by the fact that more males have higher education [5], thus not always Fairness is an issue.

This research aims to study if, in a given context, a trade-off between Data Quality and Data Ethics exists and, in this case, give guidelines to the user according to that specific context. In this paper, we focus on the *Completeness* dimension of DQ, and on the *Fairness* dimension of DE. To this aim, we have designed experiments that take a dataset as input and perform an assessment of these dimensions before and after applying some operations that should improve them. The rest of the paper is organized as follows: Section 2 summarizes related work, while Section 3 introduces preliminary concepts of both areas of Data Quality and Data Ethics and describes the method we used to analyze the relationship between Completeness of DQ and the Fairness dimension of DE; Section 4 presents the experiments we conducted on a real-world dataset, and Section 5 concludes the paper.

## 2. Related Work

Research studies on the relationship between DQ and DE are in a very preliminary phase. In this section, we will first present seminal works on Fairness and then introduce two first attempts at studying its important relationship with Completeness. We do not focus on DQ systems since, in this paper, we will resort to well-known and established DQ definitions and techniques [2].

In the literature, one of the most notable solutions aiming to measure and enforce Fairness is *AI Fairness 360* [6], an open-source framework. It aims to mitigate data bias, quantified using different statistical measures, by exploiting pre-processing (i.e., procedures that, before the application of a prediction algorithm, make sure that the learning data are fair) techniques and statistical measures to solve bias in the dataset. Similarly, *Fairlearn* [7], another pre-processing, open-source, community-driven project, aims to help data scientists improve Fairness of their ML systems by means of statistical Fairness metrics. Both papers focus on techniques that manipulate the data to make them fairer; however, they do not consistently consider the impact that their techniques have on DQ.

A system that considers also DQ is described in the paper by Abraham et al. [8], who proposed *FairLOF*, a Fairness-aware outlier-detection framework. This work starts from the fact that underrepresented groups, although relevant in the dataset, could be identified as outliers, and specifically, on calibrating the so-called *local outlier factor*, by means of which a fairer outlier detection is possible. Though this system actually focuses on a specific problem, it can be considered a starting point for studying the relationship between DQ and DE. A similar system has been presented by Biswas et al. [9], whose goal is to investigate the impact of data preparation pipelines on algorithmic Fairness, focusing on deep-learning techniques. The authors conduct a detailed evaluation of several Fairness metrics applied to different deep-learning applications and discover that many data preparation actions can introduce bias in the data and, consequently, in the final prediction. However, they do not employ any Fairness improvement technique inside their pipelines, considering only how DQ techniques impact Fairness, and not vice versa.

Guha et al. [10] conducted a study to investigate whether errors, e.g., missing values, outliers, and label noise, can be related to demographic characteristics. Moreover, they investigate if automated data cleaning actions could impact Fairness. In their study, they discovered that tuples related to disadvantaged groups were more affected by the presence of missing values; instead, the number of mislabeled data was lower in the disadvantaged groups w.r.t the privileged ones. Moreover, they proved that, in general, the probability that automated data cleaning contributes to worsening Fairness is higher w.r.t. improving it. Finally, there is a work on the specific relationship between Fairness and missing values [11]. We discuss our diverse settings in Section 4.2.3.

## 3. Experiment Design

This section presents the method we used to investigate the relationship between DE w.r.t. *Fairness*, and the DQ, w.r.t. the *Completeness*. Figure 1 schematizes the typical Data Science pipeline used to derive knowledge from data. The pipeline begins with the *Acquisition and Extraction* step: the information relevant to the data-science task is collected. The second step of the pipeline aims to solve the Data Quality issues: *DQ Improvement and Annotation* procedures are used to "sanitize" the data sources in such a way as to make them complete, correct and consistent. In the third phase, if needed, *Data Integration* provides a unified view of the data sources acquired in the first phase. Finally, in the last two steps, the predictive models are learned (*Analysis and Modeling*), and data and results are visualized (*Visualization and Evaluation*). We position our solution between the first and second step of the

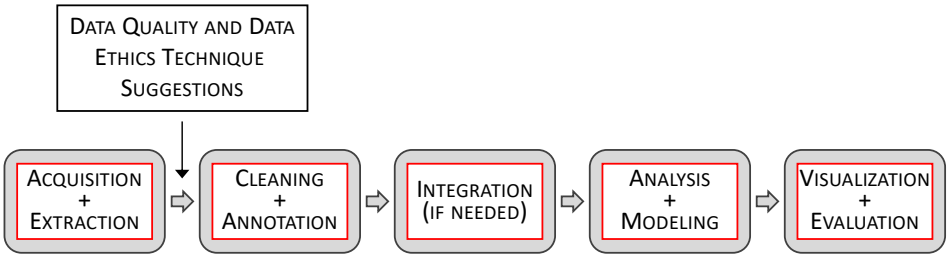


**Figure 1:** Data Science Pipeline

Data Science pipeline. Before describing the work, we introduce some preliminary theoretical concepts related to various DQ and DE aspects.

## 3.1. Preliminaries

*Data Quality* (DQ) is defined as "fitness for use," i.e., the ability of a data collection to meet the user requirements [12]. Data Quality is a multi-dimensional concept: a DQ model is composed of *DQ dimensions* representing the different aspects to be considered (i.e., errors, duplicates, format errors, typos, or missing values). The experiments concentrate on the Completeness DQ dimension. *Completeness* characterizes the extent to which a dataset represents the corresponding real-world. For instance, in a relational database, Completeness is strictly related to the presence of null values. A simple way to assess the Completeness of a table is to calculate the ratio between the number of non-null values and the number of cells in the table. It is important to specify that we also use the *Accuracy* dimension to evaluate the resulting data correctness. Accuracy is, in fact, defined as the closeness between a data value $v$ and a data value $v'$, considered as the correct representation of the real-life phenomenon that the value $v$ aims to represent. It is associated with syntactic and semantic issues that might create a discrepancy between the value stored in the dataset and the correct value. How each of these two dimensions is used will be explained in the description of the method.

*Fairness* whose most used definition is: "the absence of any prejudice or favoritism toward an individual or a group based on their inherent or acquired characteristics" [13, p.100], is one of the most important dimensions of *Data Ethics* (DE). Fairness is based on the concept of *protected or sensitive attribute*. A protected attribute is a characteristic for which non-discrimination should be established, such as religion, race, sex, and so on [14]. A protected group is a set of individuals identified by having the same value of a protected attribute (e.g.: females, young people, Hispanic people). There is no unique metric of Fairness, but many facets exist, and a model is considered fair if it satisfies some or all these metrics. The most used technique to identify unfairness in datasets is to train a classification algorithm to predict the binary value of the target class that can be a positive outcome like obtaining a loan or having a high income, or a negative outcome like not obtaining a loan or having a low income; and then use Fairness metrics to understand whether the prediction of this model encompasses discrimination for the protected group: if the metrics results show discrimination, we can conclude that also the dataset contains unfair behaviors since the model learned the bias from it. Specifically, we measure the importance of protected attributes in determining the result of the model. The following statistical metrics study how specific values of the protected attributes impact the result of the prediction algorithm (e.g., women are very frequently associated with salaries lower than 50k$/year, while men earn more than 50k$/year). Informally: *Disparate Impact Ratio* is the probability to get a positive outcome regardless of whether the person is in the protected group [15]; *Predictive Parity Ratio* evaluates if both protected and unprotected groups have equal probability that a group member with positive predictive value belongs to the negative class [14]; *False Positive Ratio*: evaluates if the probability of having a false positive prediction is the same for all protected groups [14].

## 3.2. A Method to analyze the DQ and DE tradeoff

This section presents the two pipelines we defined to execute the experiments. In the first one, which can be applied to datasets affected by ethical issues and ethics-compliant datasets, we injected errors in the input dataset, causing data quality issues, and then applied DQ improvements techniques, measuring their impact on DE. In the second pipeline, we applied DE improvement techniques to a dataset affected by ethical problems and measured their impact on DQ. Through these results, we studied the trade-off between DQ and DE. In our experiments, we considered the trade-off between the *Completeness* DQ dimension and the *Fairness* DE dimension, while the *Accuracy* DQ dimension is used to evaluate the final DQ

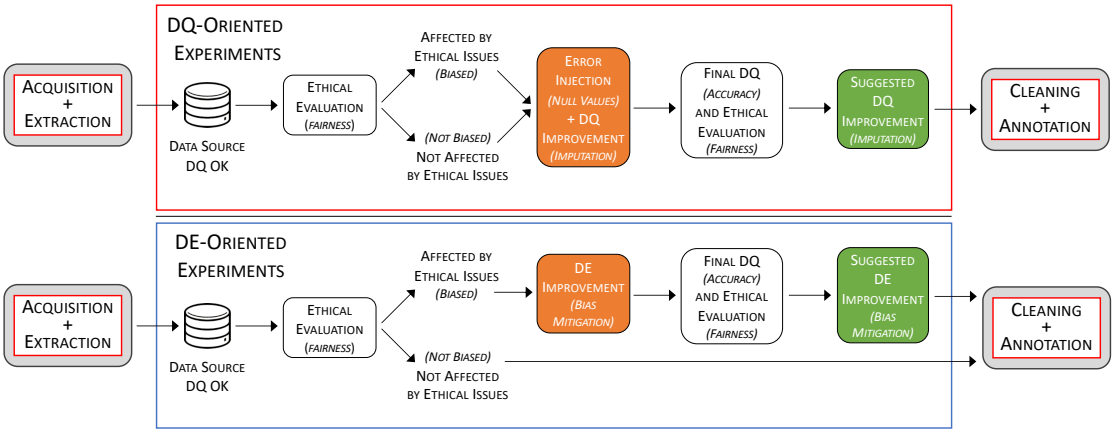


**Figure 2:** The method adopted

level in both pipelines. We used the Adult Census Income dataset[1] and the German Credit dataset[2] and considered 'sex' as the protected attribute. Since the Adult Census Income dataset already contains bias w.r.t. the income of US citizens, injecting further bias to perform the experiments was not necessary, therefore, we used it in both pipelines. The German Credit dataset, instead, is not affected by bias – thus, we could not apply Bias Mitigation techniques, and we tested it only in the first pipeline. The first operation, performed in both pipelines, is the *Ethical Evaluation*, in our case based on a classification algorithm that computes the Fairness level of the dataset. For the DQ level, we already knew that it was 100% for both datasets. We now describe the two pipelines shown in Figure 2.

*DQ-Oriented Experiments.* The input dataset was free of DQ problems. For this reason, we had to inject errors in order to evaluate the impact of the DQ improvement techniques. In our case, to affect *Completeness*, we replaced existing values with null values. By injecting a different percentage of uniformly distributed DQ errors[3] (from 90% to 0%, with a decreasing step of 10%) the *Error Injection* phase generates ten instances of the original dataset, at different levels of quality. These 'dirty' versions are the input of the *DQ Improvement* phase, in which a DQ improvement technique is applied. In our case, an Imputation technique was selected. The ten repaired datasets obtained as output were analyzed in the *Final Evaluation* phase, to check the impact of the DQ improvement on the Fairness and Accuracy measures, used to evaluate respectively the lack of bias and the data correctness. This procedure was repeated for different Imputation methods. The pipeline output is the *Suggested DQ Improvement* step in which we suggest the best DQ improvement technique based on Accuracy and Fairness results. The final users can choose the Imputation technique with the minimum impact on Fairness according to their preferred trade-off.

*DE-Oriented Experiments.* Also in this case the input dataset was free of DQ problems. As regards Fairness, we did not have an error-injection phase since, this time, the considered dataset (Adult Census Income) was already biased. The *DE Improvement* phase consisted of applying a Bias Mitigation Technique to remove unfairness. Also here, the repaired dataset was analyzed in the *Final Evaluation* phase, both Fairness and Accuracy are measured, repeating this phase for all the selected Bias Mitigation techniques. Some of these techniques, since they act by directly replacing the data values with other (fake) values, also allow controlling the amount of bias repair executed. For example, Correlation Remover [7], fully described in the next section, modifies the actual values to minimize the correlation between the feature attributes and the sensitive ones. The output of the pipeline is the *Suggested DE Improvement* step in which we propose the best DE improvement technique based on both DQ and DE evaluation results. The final users can choose the Bias Mitigation technique having the minimum impact on Accuracy according to their preferred trade-off.

## 4. Experiments

In this section, we first introduce the experimental setup and then describe the results, both from the DQ and the DE perspectives.

[1] https://archive.ics.uci.edu/ml/datasets/adult
[2] https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data)
[3] Related to a specific DQ dimension

## 4.1. Experimental setup

*DQ Improvement phase.* In this paper, we consider three Data Imputation techniques: Density-based, where missing values are imputed for each feature with the same distribution of the non-empty values; Mode Imputation, where the most frequent value is imputed; and Rare-based, where the less frequent value is imputed.

*Bias Mitigation phase.* Three Bias Mitigation techniques are proposed to remove the unfairness from data. The first one, Correlation Remover [7], removes the negative correlation between the protected attribute and the classification label by modifying the non-protected attributes that are in turn correlated with the protected one: mathematically speaking, it poses a minimization problem on the correlation between the feature attributes and the sensitive ones by centering the sensitive values, training a linear regressor on the non-sensitive ones and reporting the residual. The second one is Learning Fair Representation [6], which maps each data tuple (corresponding to an individual) to a 'prototype', an artificial representation of the data containing the same protected attribute but with modified values for the other features, to remove the correlation between the protected attributes and the target ones. To do so, this method uses a neural network with the objective of retaining as much information as possible. The last one, Optimized Preprocessing [6], solves an optimization problem with the objective of minimizing the difference between the modified distribution and the original one; specifically, it aims to reduce the discrimination by mapping different feature attributes to the classification labels of the individuals inside the dataset, while keeping the protected attributes unchanged, to limit the dependency of the prediction on the sensitive attributes. In all three cases, the techniques involve only the numerical features.

*Evaluation Metrics.* To evaluate the DQ level of the dataset, during the *Evaluation* phase, the *Accuracy* metric has been selected. To this aim, the distance between the original and the final dataset has been computed. Thus, we extracted the number $N_{match}$ of values that correspond to each other in the original and the final dataset, and measured the Accuracy as $\frac{N_{match}}{N_{tot}}$ where $N_{tot}$ is the total number of cells.

Since there is no standard system for measuring Fairness, we used two different systems. For the *DQ-Oriented Experiments*, we measured Fairness by means of a set of already defined formulas. Instead, for the *DE-Oriented Experiments*, we computed the Fairness metrics offered by the *Fairlearn* [7] mitigation tool. The two results are comparable since there is a very small delta between the two. For the *DQ-Oriented Experiments*, the three metrics, taken from [14, 15], selected to evaluate Fairness (see Section 3) are expressed as: *Disparate Impact Ratio* (DIR) $\frac{P(\hat{Y}=1|G=discr)}{P(\hat{Y}=1|G=priv)}$; *Predictive Parity Ratio* (PPR): $\frac{P(Y=0|\hat{Y}=1,G=discr)}{P(Y=0|\hat{Y}=1,G=priv)}$; *False Positive Ratio* (FPR) $\frac{P(\hat{Y}=1|Y=0,G=discr)}{P(\hat{Y}=1|Y=0,G=priv)}$; where $G$ is a protected attribute that has two values *discr* (=discriminated), *priv* (=privileged); $Y$ is the actual classification result, two values (or labels) 0 or 1; and $\hat{Y}$ is the algorithm-predicted decision for the individual, two values of the outcome 0 (negative outcome) or 1 (positive outcome). The ideal value for all three metrics is 1, which means both groups are treated equally. If the value is between 0 and $1-t$, the discriminated group is treated unfairly, whereas if the value is greater or equal to $1+t$, the privileged group is treated unfairly. Parameter $t$ is a threshold value that must be set by an expert. In our experiment we set the $t$ parameter equal to 0.2.

*Dataset and classification algorithm.* As explained in Section 3, we considered two datasets. The first one is the Adult Census Income dataset, typically used to predict whether the income of an individual exceeds 50k$ per year. It comprises 48842 tuples, described by 15 attributes, including the target class. This dataset contains more than one protected attribute ('race', 'sex', and 'native country'), but our study considered only the attribute 'sex'. The second one is the German Credit dataset, which collects information on individuals that are classified based on the fact that they are deemed good or bad payers when asking for a loan. It comprises 1000 tuples, consisting of 20 attributes, including the target class. The sensitive attribute is 'personal-status-sex', i.e., the marital status, from which the protected attribute 'sex' can be derived. Differently from the previous one, this dataset is not affected by bias with respect to 'sex'. Finally, we used as *classification algorithm* the Decision Tree Classifier offered by the *scikit-learn* Python library.

## 4.2. Result evaluation

This section presents the main results we obtained. In Figure 3, the $x$-axis represents the Completeness level; instead, in Figure 4, the $x$-axis shows the degree of Bias Mitigation. In both figures, the $y$-axis represents the level of the evaluated metrics.

### 4.2.1. DQ-Oriented Experiments

The plots shown in Figure 3 focus on the *DQ-Oriented Experiments* in which the *Accuracy* and *Fairness* results are compared for the three Imputation techniques explained in Section 4.1.

*Biased dataset.* The three plots at the top of Figure 3 show the results for the Adult dataset. In general, the Mode and the Density-based Imputations reach higher Accuracy with respect to the Rare-based one, since the

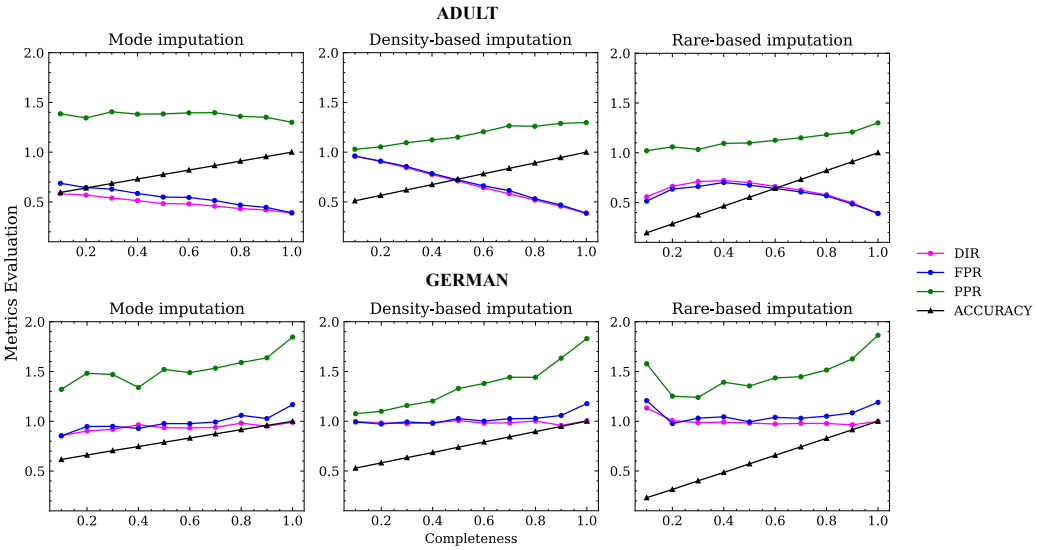


**Figure 3:** DQ-Oriented Experiments: effects of Data Imputation

latter modifies the original distribution of values more than the others. From the Fairness point of view, we can observe that the Predictive Parity Ratio (PPR) metric can assume values greater than $1+t$, i.e., 1.2. This means that the privileged class (men) is treated unfairly for that specific Fairness aspect; i.e., the probability of belonging to class 0 (low income) for a man that instead was predicted to class 1 (high income) is lower than the probability of belonging to class 0 for a woman predicted to class 1. On the contrary, False Positive Ratio (FPR) always takes opposite values with respect to PPR. These two metrics are symmetrical since they represent opposite Fairness aspects: FPR evaluates whether the probability of predicting class 1 is the same both for men and women belonging to class 0.

As we can notice, in this specific experiment, the Mode Imputation introduces minimal changes to the Fairness metrics since imputing the most frequent value does not affect the distribution of the original ones.

Instead, the Density-based Imputation is much better: in fact, as the percentage of injected errors increases, Fairness increases for all three metrics. This is related to a vast majority of the class 0 in the dataset; since the Imputation follows the value distribution, it means that those labels (class 0) have a higher probability of being assigned to men (who are over-represented). In this way, the dataset will be balanced. We can conclude that the application of this Imputation method improves Fairness.

Finally, when applying the Rare-based Imputation, when Completeness varies between 100% and 40%, the Fairness increases; for Completeness values below 40%, Fairness decreases very quickly. In this specific case, this happens because, by imputing the less frequent values, the dataset will be more balanced in favor of the protected class. As the percentage of injected errors grows, the rare values become too many, unbalancing the dataset again.

*Unbiased dataset.* The three plots at the bottom of Figure 3 show the results for the German dataset. Since the two datasets have a similar distribution, after the application of the Imputation techniques, the Accuracy takes similar values as in the previous case.

Since the dataset is already fair, FPR and DIR metrics assume values around 1, while the PPR is almost 2. After applying the Imputation techniques, FPR and DIR are not affected, while the value of PPR is closer to 1 (i.e., the probability of belonging to class 0 (bad credit) for a man predicted to class 1 (good credit) is lower than the probability of belonging to class 0 for a woman predicted to class 1), therefore the PPR has improved with respect to its initial value. In this case, the Imputation techniques balanced the PPR, improving it as much as they modify the original distribution of the values. In fact, Rare-based Imputation, which modifies the original distribution more, introduces unbalance, causing further deterioration of Fairness over the 60% injected errors.

From these results, we can notice a trade-off between Accuracy and Fairness; from the *DQ-Oriented Experiments* we see that this trade-off can be more or less emphasized depending on the DQ improvement technique applied.

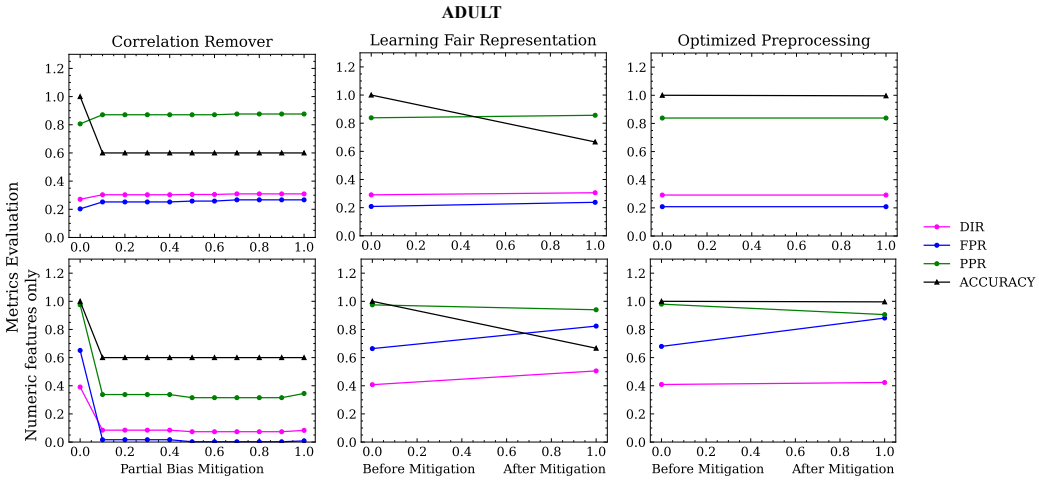


**Figure 4:** DE-Oriented Experiments: effects of Bias Mitigation

#### 4.2.2. DE-Oriented Experiments

The plots shown in Figure 4 focus on the *DE-Oriented Experiments*. We compared the *Accuracy* and *Fairness* results for the Bias Mitigation techniques explained in Section 4.1. The results of the experiments conducted on the entire dataset are represented at the top of Figure 4. The Bias Mitigation techniques we used focus only on numerical attributes, thus, the results shown at the bottom of Figure 4 show the same experiments based only on the numerical features. We now present our results by analyzing one Bias Mitigation technique at a time.

*Correlation Remover.* When applying Correlation Remover for a partial Bias Mitigation between 0 and 1, the Fairness metrics (DIR, FPR, and PPR) slightly improve, but with an important loss in Accuracy (from 1.0 to 0.6). This happens because the removal of correlation strongly modifies the data, greatly affecting Accuracy. Considering the case in which only the numerical features are involved, the Fairness metrics are negatively affected. This represents a case of over-correction. By modifying the entire dataset, data are too far from the original ones, and the results are no longer reliable.

*Learning Fair Representation.* Applying Learning Fair Representation, we have the same loss in Accuracy as for Correlation Remover, since it modifies the numerical features in order to remove correlations. However, this technique also aims to minimize information loss, thus, does not cause such a radical modification as the previous method. Therefore, Fairness improvement is minimal considering the full dataset, while considering only the numerical features, two metrics over three improve (DIR and FPR).

*Optimized Preprocessing.* Using Optimized Preprocessing, the Accuracy remains unchanged before and after the mitigation process. This happens because there is no data modification, but only weights are given to the numerical features in order to reduce the correlation between the protected attribute and the prediction. However, applying this technique to the full dataset is not sufficient to improve Fairness because the categorical features still affect the prediction. Moreover, applying this technique considering only the numerical features improves one Fairness metric (FPR) over three.

In the *DE-Oriented Experiments* we detected a trade-off between Accuracy and Fairness, and this relationship can be more or less strong depending on the Bias Mitigation technique that is applied.

#### 4.2.3. A brief comparison

We can now summarize the differences between our work and the approach of [11] presented in Section 2: in [11] the authors studied only the Completeness dimension of DQ, while we also evaluate the results using Accuracy; the Fairness metric studied in [11] is only one, while we studied two more metrics; in [11] the initial dataset used for the experiments is an unclean one, while we control the process by applying error injection to a previously cleaned dataset; finally, in [11] the Imputation techniques used are only *Mode* and *Mean*, while we also apply *Rare* and *Density-based* Imputation techniques.

## 5. Conclusions

*Takeaway message.* From our experiments, we have noticed that the application of Data Imputation techniques, in some particular cases, e.g., Density-based Im-

putation and Rare-based imputation on the Adult dataset, can contribute to improving Fairness. Moreover, in the experiments, starting from unbiased data, Fairness was not affected by the application of the Imputation techniques. In most cases, we noticed a trade-off: the Bias Mitigation technique that less affects the Accuracy, in general the Optimized Preprocessing technique, is not the one that improves Fairness the most, and vice versa; for these cases, we can deduce that techniques that succeed in preserving both Accuracy and Fairness do not exist. Therefore, as a takeaway message, we can affirm that ***the best Data Imputation/Bias Mitigation technique to apply strictly depends on the analysis goal***. If users are more interested in preserving Fairness aspects, they will concentrate on a subset of techniques at the cost of losing DQ; if the major interest is to optimize the improvement of the DQ, they will apply a subset of DQ improvement tasks that could affect Fairness. It is worth noting that situations may also exist in which Accuracy and Fairness are not in conflict; however, this is strictly context-dependent.

*Conclusions.* In this work, we analyzed the relationship between Data Quality (DQ) and Data Ethics (DE). Specifically, we focus on the Completeness dimension of DQ, and on the Fairness dimension of DE. Through a series of experiments, we demonstrated that between DQ and DE a *trade-off* is present. In fact, the experiments showed us that the application of Fairness improvement operations can lead to a deterioration of Accuracy, used to evaluate the DQ, and vice versa. Analyzing the experiments more in detail, we can also state that the amount of Accuracy deterioration after Fairness improvements depends on the Bias Mitigation technique, as well as the deterioration of Fairness can depend on the selected Imputation technique. Future work will focus on the definition of clear guidelines to recommend the best choice of DQ/DE improvement techniques to be applied depending on the scope of the analysis. Moreover, we could enrich the gathered knowledge with more datasets, DQ and DE dimensions, and Bias Mitigation techniques [16, 17].

## Acknowledgments

This research was supported by EU Horizon Framework grant agreement 101069543 (CS-AWARE-NEXT) and by project ICT4Dev, funded by AICS (Italian Agency for Development Cooperation).

## References

[1] A. Jain, et al., Overview and importance of data quality for machine learning tasks, in: Proceedings of the 26th ACM SIGKDD, 2020, pp. 3561–3562.

[2] C. Batini, M. Scannapieco, Data and Information Quality - Dimensions, Principles and Techniques, Data-Centric Systems and Applications, Springer, 2016.

[3] C. Sancricca, C. Cappiello, Supporting the design of data preparation pipelines (2022) 149–158.

[4] D. Firmani, L. Tanca, R. Torlone, Ethical dimensions for data quality, JDIQ 12 (2019) 1–5.

[5] F. Kamiran, I. Žliobaitė, Explainable and non-explainable discrimination in classification, Discrimination and Privacy in the Information Society: Data mining and profiling in large databases (2013) 155–170.

[6] R. K. Bellamy, et al., Ai fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias, IBM Journal of Research and Development 63 (2019) 4–1.

[7] S. Bird, et al., Fairlearn: A toolkit for assessing and improving fairness in ai, Microsoft, Tech. Rep. MSR-TR-2020-32 (2020).

[8] S. S. Abraham, Fairlof: fairness in outlier detection, Data Science and Engineering 6 (2021) 485–499.

[9] S. Biswas, H. Rajan, Fair preprocessing: towards understanding compositional fairness of data transformers in machine learning pipeline, in: Proceedings of the 29th ACM Joint Meeting on ESEC/FSE, 2021, pp. 981–993.

[10] S. Guha, F. A. Khan, J. Stoyanovich, S. Schelter, Automated data cleaning can hurt fairness in machine learning-based decision making, in: 2023 IEEE 39th International Conference on Data Engineering (ICDE), IEEE, 2023, pp. 3747–3754.

[11] F. Martínez-Plumed, C. Ferri, D. Nieves, J. Hernández-Orallo, Fairness and missing values, arXiv preprint arXiv:1905.12728 (2019).

[12] R. Y. Wang, D. M. Strong, Beyond accuracy: What data quality means to data consumers, JMIS 12 (1996) 5–33.

[13] N. A. Saxena, et al., How do fairness definitions fare? testing public attitudes towards three algorithmic definitions of fairness in loan allocations, Artif. Intell. 283 (2020) 103238.

[14] S. Verma, J. Rubin, Fairness definitions explained, in: Proceedings of the FairWare@ICSE, 2018, pp. 1–7.

[15] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, A. Galstyan, A survey on bias and fairness in machine learning, ACM Comput. Surv. 54 (2022) 115:1–115:35.

[16] F. Azzalini, C. Criscuolo, L. Tanca, E-fair-db: functional dependencies to discover data bias and enhance data equity, JDIQ 14 (2022) 1–26.

[17] F. Azzalini, C. Criscuolo, L. Tanca, Fair-db: A system to discover unfairness in datasets, in: ICDE, IEEE, 2022, pp. 3494–3497.