Data Quality and Fairness: Rivals or Friends?

(Discussion Paper)

Fabio Azzalini[1], Cinzia Cappiello[1], Chiara Criscuolo[1], Sergio Cuzzucoli[1], Alessandro Dangelo[1], Camilla Sancricca[1] and Letizia Tanca[1]

[1]Politecnico di Milano - Dipartimento di Elettronica, Informazione e Bioingegneria

Abstract

In the last decade, data-driven decision-making is considered one of the main drivers for organizational success. Within this approach, decisions are based on insights and patterns identified through data analysis. In this scenario, input data must be reliable to guarantee the accuracy of the results: they should be correct and complete but also unbiased, i.e., both Data Quality (DQ) and Fairness should be guaranteed. However, maximizing DQ and Fairness simultaneously is not trivial, since data quality improvement techniques can negatively affect Fairness and vice versa. Understanding and thoroughly analyzing this relationship between DQ and Fairness is therefore paramount, and is this paper's goal. The results of our experiments, based on a well-known biased dataset (the Adult Census Income) provided details about this trade-off and allowed us to draw some guidelines.

Keywords
Data Quality, Fairness

1. Introduction

In the last decades, the possibility to use large amounts of data to extract information, and gain deeper knowledge in several domains, has caused the spread of a data-driven culture, making data collection and management extremely important. In fact, in this scenario, strategic decisions are made on the basis of data analysis and interpretation, and relying on dependable results becomes imperative. 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 decrease the accuracy of the 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].

Another key property for effective data-driven decision-making applications is the ethical level of the data. In fact, 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 it also conforms to high ethical standards*. Therefore the authors propose to include the most common ethical requirements among the dimensions of quality, grouped in an Ethics Cluster:

- *Fairness* is defined as *the lack of bias*, since an algorithmic bias might result from training a system with biased data.
- *Transparency* is the possibility to control the knowledge extraction process to verify the reasons of the results.
- *Diversity* is the degree to which different kinds of objects are represented in a dataset.
- Finally, *Data Protection* concerns the ways to protect data, algorithms and models from unauthorized access.

Looking at this list, it is immediate to see that there may be contrasting objectives also among the dimensions of Ethics, for instance as regards an obvious conflict between Transparency and Data Protection. In the same way, the relationship between many well-known DQ dimensions [2] and the ethical ones is complex. In fact, maximizing both aspects simultaneously is not trivial, since data quality techniques can negatively affect Fairness and vice versa. For example, commonly used DQ improvement techniques, e.g., imputing missing values using the mean value, can modify the overall distribution of values in the dataset and might lead to a reduction 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.

With this work, we want to investigate this trade-off, especially focusing on the *Completeness* and *Accuracy* dimensions of DQ, and on the *Fairness* dimension of ethics. To this aim, we have designed experiments that take as input a dataset and perform an assessment of DQ and Fairness before and after the application of some operations that should improve them. After the description of the experiments we conclude the paper with some takeaway messages for the researchers that look for the best strategy for applying changes in data to improve Fairness or DQ according to their needs.

The rest of the paper is organized as follows: Section 2 summarizes related work, while Section 3 introduces preliminary concepts of both areas of DQ and Fairness and describes the method we used to analyze the relationship between DQ and Fairness; 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 Fairness are in a very preliminary phase. In this section we will first present seminal works in Fairness and then introduce two first attempts at studying this important relationship. We do not focus on DQ systems since in this paper we will resort to well-known and established DQ techniques.

In the literature, one of the most notable solutions is *AI Fairness 360* [5], an open-source framework aiming to measure and enforce Fairness. Its aim is 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), in-processing (i.e., procedures that ensure that, during the learning phase, the algorithm does not pick up

the bias present in the data) and post-processing techniques (i.e., procedures that correct the algorithm's decisions with the objective of making them fair). The user can choose between four pre-processing techniques and five statistical measures to solve bias in the dataset.

Similarly, *Fairlearn* [6], another pre-processing, open-source, community-driven project, aims to help data scientists improve the Fairness of their ML systems by means of statistical Fairness metrics. These works focus on techniques that manipulate the data, seeking to make them more fair. They do not consistently consider the impact that their techniques have on both the DQ and Fairness dimensions.

A preliminary system that considers both DQ and Fairness is the paper by Abraham et al. [7] who proposed *FairLOF*, a Fairness-aware outlier detection framework. This work starts from the fact that underrepresented groups could be identified as outliers, although they are relevant in the dataset. Specifically, it focuses on calibrating the so-called *local outlier factor*, a local outlier detection method by means of which a fairer outlier detection is possible. Though this system actually focuses on a specific DQ problem, it can be considered as a starting point for studying the relationship between DQ and Fairness.

A similar system has been presented by Biswas et al. [8]. The authors' 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, thus considering only how data quality techniques impact Fairness, and not vice versa.

3. Experiment Design

This section presents the method we used to investigate the relationship between DQ and Fairness. Before describing the work, we introduce some preliminary theoretical concepts related to various Data Quality and Data Ethics aspects.

Data Quality Data Quality (DQ) is defined as "fitness for use," i.e., the ability of a data collection to meet the user requirements [9]. Data Quality is a multi-dimensional concept: a DQ model is composed of *DQ dimensions* representing the different aspects to consider (i.e., errors, duplicates, format errors, typos, or missing values). As already mentioned, our work focuses on the Accuracy and Completeness dimensions:

- *Accuracy* is 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 [2]. It is associated with syntactic and semantic issues that might create a discrepancy between the value stored in the dataset and the correct value.
- *Completeness* characterizes the extent to which a dataset represents the corresponding real-world [2]. 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.

Fairness   Fairness is one of the most important dimensions of *Data Ethics*. The most used definition of Fairness is: "it is the absence of any prejudice or favoritism toward an individual or a group based on their inherent or acquired characteristics" [10, p.100].

Fairness is based on the idea 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 [11]. A protected group is a set of individuals identified by the same value of a protected attribute (e.g.: females, young people, Hispanic people).

There is no unique definition of Fairness, but many facets exist, and a model is considered fair if it satisfies some or all these definitions. The most used technique to identify unfairness in datasets is to train a classification algorithm to predict the binary value of the target class and then use Fairness metrics to understand whether the prediction of this model encompasses discrimination: if the metrics results show unfairness, we can conclude that also the original 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., being a woman determines that the salary is lower than 50k$/year, and being a man determines that the salary is higher than 50k$/year). Informally:

- *Disparate Impact* is the probability to get a positive outcome regardless of whether the person is in the protected group [12];
- *Predictive Parity* evaluates if both protected and unprotected groups have equal probability that a group member with positive predictive value belongs to the negative class [11];
- *False Positive Ratio*: evaluates if the probability of having a false positive prediction is the same for all protected groups [11].

The Method Adopted   This section presents the two pipelines we defined to execute the experiments and study the relationship between DQ and Fairness. In the first one, we inject errors in the dataset, causing data quality issues, then apply DQ improvement techniques and measure their impact on Fairness. In the second pipeline, since the Adult Census Income dataset[1] already contains bias w.r.t. the income of US citizens we do not need to inject further bias to perform the experiments, therefore, we already start from a biased dataset, apply bias mitigation techniques and measure their impact on DQ. By means of these results, we evaluate the relationship between Fairness and DQ.

The dataset used for the experiments had the following characteristics: *(i)* it contained bias, so the classification algorithm would be affected by unfairness, *(ii)* it had been pre-processed so that the classification algorithm could be executed in a correct manner. For example, missing values had to be dealt with, and encoding and normalization operations were performed.

The last operation to be performed before entering the two pipelines was applying a first classification algorithm in order to compute the Fairness level of the dataset. As for the DQ measure, we already knew that it was 100%. We now describe the two pipelines.

- *Data Quality Improvement Pipeline:* The input dataset is free of DQ problems. For this reason, we had to inject errors in order to evaluate the impact of DQ improvement

---

[1]https://archive.ics.uci.edu/ml/datasets/adult

techniques. By injecting in a uniform manner a different percentage of DQ errors (from 90% to 0%, with a decreasing step of 10%) related to a specific DQ dimension, the *Error Injection* phase generated ten instances of the original dataset at different levels of quality. For example, for the Completeness dimension, a certain number of values are replaced with null values. The obtained ten dirty datasets were the input of the *Data Quality Improvement* phase, in which a DQ improvement technique was applied. In the case of the Completeness dimension, an imputation technique was selected. The ten clean datasets obtained as output were analyzed in the second *Evaluation* phase, in order to check the impact of the DQ improvement on the Fairness measures. This procedure was repeated for a number of different imputation methods.

- *Fairness Improvement Pipeline:* As regards Fairness, we did not have an error injection phase since the considered database was already biased. The improvement phase (i.e., *Bias Mitigation* phase) consisted of applying a bias mitigation technique to remove unfairness. The repaired dataset, output of this phase, was analyzed in the second *Evaluation* phase in order to assess the impact of the bias mitigation on the DQ level. This phase was repeated for all the selected bias mitigation techniques. Since some of these techniques act by directly substituting the data values with other (faked) values, they also allow for partial bias repairs. For example, Correlation Remover [6], fully described in the next section, modifies the actual values in order to minimize the correlation between the feature attributes and the sensitive ones. If possible, ten repaired datasets (with a level of repair from 10% to 100%) were the output of this phase.

The last step was the analysis of the results of the two second *Evaluation* phases.

4. Experiments

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

4.1. Experimental setup

DQ Improvement phase   In this paper, we consider two types of Data Imputation techniques: Density-based, i.e., missing values are imputed for each feature with the same distribution of the non-empty values, and Rare-based, i.e., the less frequent value is imputed.

Bias Mitigation phase   Two bias mitigation techniques are proposed in order to remove the unfairness from data. The former one, Correlation Remover [6], removes the negative correlation between the protected attribute and the classification label by modifying the non-protected attributes that are in turn correlated to the protected one: mathematically speaking, it poses a minimization problem of 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 latter, Optimized Preprocessing [5], solves an optimization problem with the objective of minimizing the difference between the modified distribution and the original one; specifically, it aims to reduce discrimination by mapping different feature

attributes and classification labels of the individuals inside a dataset while keeping the protected attributes unchanged, to limit the dependency of the prediction on the sensitive attributes.

Evaluation Metrics   To evaluate the DQ level of the dataset in the *Evaluation* phase (see Section 3), the Accuracy dimension has been selected. To this aim, the distance between the original and the final dataset has been computed. Thus, the number of values N_{match} that match between the original and the final dataset was extracted and the Accuracy dimension has been measured as follows:

$$Accuracy = \frac{N_{match}}{N_{tot}} \tag{1}$$

where N_{tot} is the total number of cells.

The three metrics selected to evaluate Fairness (see Section 3) were applied by using the following formulas:

$$Disparate\ Impact\ (DIR) = \frac{P(\hat{Y} = 1 | G = discr)}{P(\hat{Y} = 1 | G = priv)} \tag{2}$$

$$Predictive\ Parity\ (PPR) = \frac{P(Y = 0 | \hat{Y} = 1, G = discr)}{P(Y = 0 | \hat{Y} = 1, G = priv)} \tag{3}$$

$$False\ Positive\ Ratio\ (FPR) = \frac{P(\hat{Y} = 1 | Y = 0, G = discr)}{P(\hat{Y} = 1 | Y = 0, G = priv)} \tag{4}$$

where:

- G: protected attribute that has two values *discr* (=discriminated), *priv* (=privileged);
- X: additional data regarding the individual;
- Y: actual classification result, two values (or labels) 0 or 1;
- $\hat{Y}$: the algorithm predicted decision for the individual, two values of the outcome 0 or 1.

Dataset and classification algorithm   The Adult Census Income dataset is obtained as an extraction of the 1994 U.S. Census database. It is typically used to predict whether the income of an individual exceeds 50k$ per year. It comprises 48842 tuples, described by 15 attributes, including the targeted class. There are some sensitive attributes, such as 'race', 'sex', and 'native country'. In particular, for the experiments shown in this paper, we compute the *Evaluation* considering the *sex* as a protected attribute (see Section 3). Finally, the Decision Tree Classifier offered by the *scikit-learn*[2] Python library was used as *classification algorithm*.

4.2. Result evaluation

This section presents the main results we obtained. In Figure 1, the *x*-axis represents the Completeness level; instead, in Figure 2, the *x*-axis shows the degree of bias mitigation. In both Figures, the *y*-axis represents the level of the evaluated metrics.
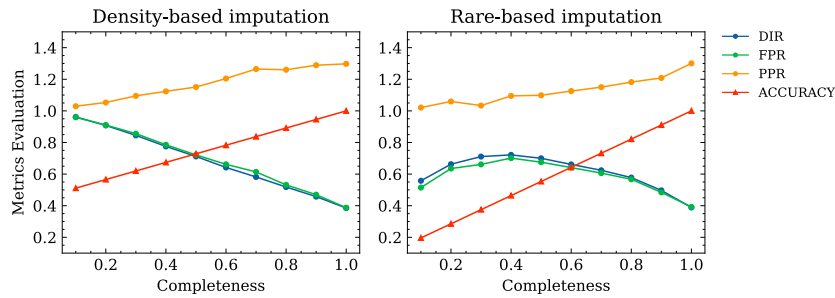
Figure 1: Data Quality perspective: effects of Data Imputation on the Adult dataset

Data Quality Perspective The plots shown in Figure 1 focus on the *DQ perspective* in which the *Evaluation* results are compared for the two imputation techniques explained above (see Section 4.1). As we can notice, the Density-based Imputation is widely better: the Accuracy measure does not reach less than 50% for the Accuracy assessment, and the Fairness measures increase as the percentage of injected errors increases. This is related to a vast majority of the target class that has a value lower than 50k\$/year in the dataset; since the imputation follows the value distribution, it means that those labels 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. Instead, applying the Rare-based Imputation, we have a big deterioration of Accuracy, and from 100% to 40% of Completeness, the Fairness increases; instead, 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 classes. As the percentage of injected errors increases, the rare values become too many, unbalancing the dataset. We can also observe that the Predictive Parity (PPR) metric can assume values greater than 1. This means that the privileged class (men) is discriminated for that specific Fairness aspect; False Positive Ratio (FPR) always takes opposite values with respect to PPR. The two metrics are symmetrical since they represent opposite Fairness aspects. From these results, we notice that a trade-off between DQ and Fairness is present and that, from the *DQ perspective*, this trade-off can be more or less emphasized depending on the DQ improvement technique applied.

Fairness Perspective The plots shown in Figure 2 focus on the *Fairness perspective*. The *Evaluation* results are compared for two different bias mitigation techniques: Correlation Remover offered by Fairlearn [6] and Optimized preprocessing in AI Fairness 360 tool [5] (see Section 4.1). In applying Correlation Remover, the Fairness metrics (DIR, FPR and PPR) slightly increase with an important loss in Accuracy (from 1.0 to 0.6) for a partial bias mitigation between 0 and 1. This happens because the removal of correlation strongly modifies the data, greatly decreasing the Accuracy. On the other hand, Optimized Preprocessing increases one metric over three (FPR), and the Accuracy remains unchanged before and after the mitigation process. This happens because data modification is at a minimum, not affecting the Accuracy. From the

---
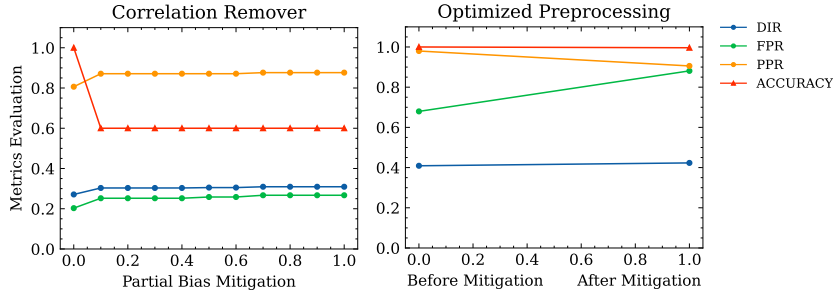
Figure 2: Fairness perspective: effects of Bias Mitigation on the Adult dataset

Fairness perspective, a trade-off between DQ and Fairness is present, and this trade-off can be more or less evident depending on the bias mitigation technique applied.

Takeaway message   From our experiments, we can notice that in some particular cases, the bias mitigation technique that less affects the DQ 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 DQ and Fairness do not exist. Therefore, as a takeaway message, we can affirm that the best DQ improvement/bias mitigation techniques to apply strictly depends on the analysis goal. If a user is more interested in preserving Fairness aspects, s/he 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, the user will apply a subset of DQ improvement tasks that could deteriorate Fairness. It is worth noting that situations may also exist in which the DQ and the Fairness aspects are not in conflict; however, this is strictly context-dependent and could be rare to observe.

5. Conclusions and Future Work

In this work, we analyzed the relationship between Data Quality (DQ) and Fairness. In particular, through a series of experiments, we demonstrated that between DQ and Fairness, a *trade-off* is present. In fact, the experiments showed us that the application of Fairness improvement operations could lead to a deterioration of the DQ and vice-versa. Analyzing the experiments more in detail, we can also state that the amount of DQ deterioration after Fairness improvements depends on the bias mitigation technique, as well as the deterioration of Fairness can depend on the selected DQ improvement technique. Future work will focus on the definition of clear guidelines to recommend the best choice of DQ improvement/bias mitigation techniques to be applied depending on the scope of the analysis. Moreover, we could enrich the gathered knowledge with more datasets, DQ dimensions, Fairness metrics and bias mitigation techniques [13, 14].

Acknowledgments

References

[1] A. Jain, H. Patel, L. Nagalapatti, N. Gupta, S. Mehta, S. Guttula, S. Mujumdar, S. Afzal, R. Sharma Mittal, V. Munigala, 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).

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

[5] R. K. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilović, et al., Ai fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias, IBM Journal of Research and Development 63 (2019) 4–1.

[6] S. Bird, M. Dudík, R. Edgar, B. Horn, R. Lutz, V. Milan, M. Sameki, H. Wallach, K. Walker, Fairlearn: A toolkit for assessing and improving fairness in ai, Microsoft, Tech. Rep. MSR-TR-2020-32 (2020).

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

[8] 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.

[9] R. Y. Wang, D. M. Strong, Beyond accuracy: What data quality means to data consumers, Journal of management information systems 12 (1996) 5–33.

[10] N. A. Saxena, K. Huang, E. DeFilippis, G. Radanovic, D. C. Parkes, Y. Liu, How do fairness definitions fare? testing public attitudes towards three algorithmic definitions of fairness in loan allocations, Artif. Intell. 283 (2020) 103238.

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

[12] 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.

[13] F. Azzalini, C. Criscuolo, L. Tanca, E-fair-db: functional dependencies to discover data bias and enhance data equity, ACM Journal of Data and Information Quality 14 (2022) 1–26.

[14] F. Azzalini, C. Criscuolo, L. Tanca, Fair-db: A system to discover unfairness in datasets, in: 2022 IEEE 38th International Conference on Data Engineering (ICDE), IEEE, 2022, pp. 3494–3497.