Data Integration, Bias and Fairness

Information Retrieval, Extraction & Integration

Course 2022/23

Víctor Morcuende Castell

Guillermo Nájera Lavid

TABLE OF CONTENTS

1. Introduction

During the last years, data has become more and more important for many different contexts. It has become an utterly important agent in many fields such as machine learning, big data, artificial intelligence, and many others. This new context to which society is moving towards requires data extraction and understanding to be more precise, efficient, and especially reliable. It is because of this that a process such as data integration has become a key agent in the value chain of these activities. Data integration is of vital importance to the manipulation of data, since it is the process of unifying different data sources into a unified, harmonized, and standardized central database, for the purpose of combining them and perform studies, machine learning, and many other activities which require of clean and high-quality data to be performed effectively.

2. Dataset Selection & Description

The two datasets selected for this assignment are the following ones:

First, the **US Adult Census dataset**: also known as the Census Income dataset, is sourced from the 1994 Census database. It contains individual-level demographic and income information. This dataset is commonly used for tasks such as income prediction and classification, as well as exploring various socioeconomic factors. The key attributes in this dataset include:

- **age**: Age of the individual in years.
- **workclass**: The category of employment, such as Private, Self-employed, Government, etc.
- **fnlwgt**: Final weight, representing the number of people the census believes the entry represents.
- **education**: The highest level of education achieved by the individual.
- **education.num**: The highest level of education achieved by the individual, represented as a continuous numeric variable.
- **marital.status**: The individual’s marital status, such as Married, Divorced, Never-married, etc.
- **occupation**: The individual’s job or profession.
- **relationship**: The individual’s role in the family, such as Husband, Wife, etc.
- **race**: The individual’s race, such as White, Black, Asian, etc.
- **sex**: The individual’s gender, either Male or Female.
- **capital.gain**: The amount of capital gain (income from investments) for the individual.
- **capital.loss**: The amount of capital loss (losses from investments) for the individual.
- **hours.per.week**: The number of hours the individual works per week.
- **native.country**: The individual’s country of origin.
- **income**: The individual’s income, classified as either "<=50K" or ">50K" per year.

Second, the **US Census Demographic Data dataset** (being precise, the “acs2017_county_data.csv”): it contains aggregated demographic information from the 2017 American Community Survey (ACS) 5-Year Estimates at the county level. The ACS is an annual survey conducted by the US Census Bureau to collect detailed socioeconomic and demographic data. This dataset provides a snapshot of various aspects of the US population, such as income, education, employment, and housing characteristics. The key attributes in the dataset include:

- **CountyId**: A unique identifier for each county, based on the Federal Information Processing Standards (FIPS) code.
- **State**: The US state the county is located in.
- **County**: The name of the county.
- **TotalPop**: The total population of the county.
- **Men**: The number of men in the county.
- **Women**: The number of women in the county.
- **Hispanic**: The percentage of the population identifying as Hispanic or Latino.
- **White**: The percentage of the population identifying as White.
- **Black**: The percentage of the population identifying as Black or African American.
- **Native**: The percentage of the population identifying as American Indian or Alaska Native.
- **Asian**: The percentage of the population identifying as Asian.
- **Pacific**: The percentage of the population identifying as Native Hawaiian or Other Pacific Islander.
- **VotingAgeCitizen**: The number of US citizens of voting age (18 years or older).
- **Income**: The median household income in the county.
- **IncomeErr**: The margin of error for the median household income.
- **IncomePerCap**: The per capita income in the county.
- **IncomePerCapErr**: The margin of error for the per capita income.
- **Poverty**: The percentage of the population living below the poverty line.
- **ChildPoverty**: The percentage of children living below the poverty line.
- **Professional**: The percentage of the population employed in professional, scientific, management, administrative, and waste management services.
- **Service**: The percentage of the population employed in service occupations, such as healthcare support, protective service, etc.
- **Office**: The percentage of the population employed in office and administrative support occupations.
- **Construction**: The percentage of the population employed in construction, extraction, and maintenance occupations.
- **Production**: The percentage of the population employed in production, transportation, and material moving occupations.
- **Drive**: The percentage of workers who commute to work by car, truck, or van.
- **Carpool**: The percentage of workers who carpool to work.
- **Transit**: The percentage of workers who commute to work using public transportation, such as buses or trains.
- **Walk**: The percentage of workers who walk to work.

- **OtherTransp**: The percentage of workers who use other means of transportation to commute to work, such as bicycles or motorcycles.
- **WorkAtHome**: The percentage of workers who work from home.
- **MeanCommute**: The average time (in minutes) it takes for workers to commute to work.
- **Employed**: The number of employed individuals in the county.
- **PrivateWork**: The percentage of workers employed in the private sector.
- **PublicWork**: The percentage of workers employed in the public sector (government).
- **SelfEmployed**: The percentage of workers who are self-employed.
- **FamilyWork**: The percentage of workers involved in unpaid family work.
- **Unemployment**: The unemployment rate in the county.

3. Detected conflicts among the datasets

In this section, we will explain the main conflicts and challenges, both at data and schema levels, that we have detected while trying to integrate the datasets.

Data-level conflicts

Timeframe

The US Adult Census dataset (onwards Dataset 1) is extracted from the 1994 Census, whereas the US Census Demographic Data dataset (onwards Dataset 2) contains data from the 2017 American Community Survey 1-Year Estimates. The temporal gap between the two datasets might result in discrepancies and inconsistencies, as societal changes and demographic shifts have occurred over the years.

Data value

Dataset 1 has a numerical (but not continuous, since there are gaps in ages) age attribute for each individual, while Dataset 2 presents age as grouped ranges or percentages for each census county. Moreover, both datasets do not consider ages below 16. Also, there is a conflict with the "adult" value, as individuals with 18 years or more are represented as adult in Dataset 2, while in Dataset 1, 17 years old individuals are also considered adults. Therefore, to make the datasets compatible, it is necessary to standardize the age representation. This could involve creating age bins for Dataset 1 to match the age ranges in Dataset 2 or converting the age percentages in Dataset 2 to approximate individual-level ages, if feasible.

Missing values

'workclass' and 'occupation' from Dataset 1 contain missing values.

Data Precision

‘education’ and ‘occupation’ variables lack data precision, as there is no standardize manner of representing the data of each individual from these columns correctly.

Demographic representation

The categories used to represent race and ethnicity differ between the datasets, which leads to inconsistencies when integrating the data. For example, data is quantified (by counting the percentage of population) for Dataset 2, while there is no assessment for this issue in Dataset 1. To address this issue, a mapping of race and ethnicity categories between the two datasets could be created, and the categories should be standardized accordingly.

Data units

In Dataset 1, income is a binary classification (‘<=50K’ or ‘>50K’), while in Dataset 2, it is a continuous numerical value (median household income). To reconcile these units, the income in Dataset 2 could be discretized by choosing a threshold (e.g., $50,000) or by creating income bins in both datasets.

Granularity

Dataset 1 provides data at an individual level, while Dataset 2 offers aggregated data at a census county level. This difference in granularity can create challenges when attempting to merge or compare data across the datasets. One approach to addressing this issue is by aggregating individual-level data in Dataset 1 to match the census tract level of Dataset 2. Alternatively, it might be possible to disaggregate the census tract data in Dataset 2 to the individual level, although this would likely require additional data sources or assumptions.

Schema-level conflicts

Attribute differences

Both datasets have overlapping attributes, such as age, gender, race, income, and education. However, they also contain unique attributes that are not present in the other dataset. For instance, Dataset 1 includes attributes like ‘workclass’, ‘marital.status’, ‘occupation’, and ‘relationship’. In contrast, Dataset 2 focuses on additional demographic variables, such as ‘employment’ and housing characteristics. To integrate these datasets, it would be necessary to select a common set of attributes for analysis or expand the scope of the study to accommodate the unique attributes.

Aggregation Level

Dataset 1 is individual based, while Dataset 2 is aggregated at the county level. This means that combining the datasets would require either aggregating Dataset 1 or disaggregating Dataset 2.

Employment information

Dataset 1 has specific attributes for ‘workclass’, ‘occupation’, and ‘hours.per.week’, while Dataset 2 only has information about the type of work (‘Professional’, ‘Service’, ‘Office’, ‘Construction’, etc.) and employment status (‘PrivateWork’, ‘PublicWork’, ‘SelfEmployed’, etc.).

Demographic representation

Dataset 1 represents race using a single ‘race’ attribute with possible values such as ‘White’, ‘Black’, ‘Asian-Pac-Islander’, ‘Amer-Indian-Eskimo’, and ‘Other’. In contrast, Dataset 2 has separate attributes for each racial group (‘Hispanic’, ‘White’, ‘Black’, ‘Native’, ‘Asian’, ‘Pacific’).

Gender discrepancies

Dataset 1 represents gender as ‘Male’ or ‘Female’, while Dataset 2 has separated ‘Men’ and ‘Women’ columns with population counts. To align these representations, the gender distribution in Dataset 2 could be calculated or the gender data in Dataset 1 could be reorganized.

Geographic discrepancies

Dataset 2 has geographic information (census county, state, and county) that is not present in Dataset 1. However, Dataset 1 contains information about the individual’s country of origin, while in Dataset 2 this information does not appear.

4. Bias and Fairness Analysis

In this section, we will discuss the methods used to identify, analyze, and mitigate the bias and fairness of our datasets. To do so, we will use the tool Aequitas, which is a bias and fairness audit toolkit that helps analyzing the performance of machine learning models across different demographic groups.

However, before applying Aequitas, we realized that first we had to preprocess the datasets and create a machine learning model, since Aequitas needs a fixed structure to be able to work properly on our data. Therefore, the main steps that we followed to be able to use Aequitas were:

1. Preprocess the datasets: we had to modify the variables which had missing values as well as those that contained irrelevant or nonsense values. As a result, ‘workclass’, ‘occupation’ from Dataset 1 and ‘ChildPoverty’ and ‘Income’ from Dataset 2 were modified. After that, we applied a Label Encoder for the purpose of transforming the categorical variables to numeric ones, having a complete numeric dataset which could be fed to a machine learning model.

2. Machine learning model: once the datasets were on point, we created a Logistic Regression model to be able to extract the ‘score’ and ‘label_value’ variables, which are basically the target and predicted variables of the model, that are also needed for Aequitas to work. At this point, we decided to use the ‘income’ variable from both datasets as the target variable, since it was the most logical approach to follow. Consequently, for the Dataset 1 we had to convert ‘<=50K’ to 0 and ‘>50K’ to 1, and then apply the same procedure to the continuous values for the Dataset 2, as Aequitas required that the ‘score’ and ‘label_value’ variables were binary.

Finally, after we performed the procedure mentioned, we began the bias and fairness analysis. To achieve this part, we had to follow a set of steps for each dataset:

1. Choose a set of attributes: the attributes where we are going to look for bias or fairness. For Dataset 1, we chose ‘age’, ‘race’ and ‘sex’ variables. For Dataset 2, we selected the ‘Men’, ‘Women’, ‘Hispanic’, ‘White’, ‘Black’, ‘Native’, ‘Asian’ and ‘Pacific’ variables.

2. Identify a reference group: to assess the direction of bias and to be used as baseline to calculate relative disparities in this audit. We decided to go with the ‘Automatically select group with the lowest bias metric for every attribute’, since we thought it was the one with the most logical sense.

3. Select Fairness Metrics: we decided to select all (Equal Parity, Proportional Parity, False Positive Rate Parity, False Discovery Rate Parity, False Negative Rate Parity and False Omission Rate Parity).

4. Fairness Threshold: we opted to go with the default setting.

Dataset 1 Analysis

As we said previously, we are going to focus on the ‘age’, ‘race’ and ‘sex’ variables.

Summary of the Audit Results:

- Equal Parity: Failed
- Proportional Parity: Failed
- False Positive Rate Parity: Failed
- False Discovery Rate Parity: Failed

- False Negative Rate Parity: Passed
- False Omission Rate Parity: Failed

Equal Parity (Failed)

Equal Parity aims to ensure that all protected groups have equal representation in the selected set. The audit found that the age groups 37-48, 28-37, and 48-90 had disparities of 4.91, 2.82, and 5.30 times the representation of the reference group (17-28), respectively. This unequal distribution could be due to biases in data collection, data preprocessing, or inherent imbalances in the population. The lack of equal representation in the dataset indicates that some groups might be over or underrepresented, leading to skewed results of certain demographic groups.

Proportional Parity (Failed)

Proportional Parity tries to guarantee that all protected groups are represented in the selected set proportionally to their share of the population. For age, the audit found disparities of 6.17, 3.23, and 5.30 times for the age groups 48-90, 28-37, and 37-48, respectively, compared to the reference group (17-28). This suggests that the dataset does not reflect the true proportions of the population, which could be due to sampling bias, data preprocessing issues, or other data collection problems. As the dataset does not accurately represent the population, the model's predictions may be biased towards certain demographic groups, leading to unfair treatment.

False Positive Rate Parity (Failed)

False Positive Rate Parity aims to ensure that all protected groups have the same false positive rates as the reference group. The audit found disparities of 3.36, 4.48, and 2.24 times for the age groups 37-48, 48-90, and 28-37, respectively, compared to the reference group (17-28). This indicates that different age groups have varying false positive rates, which might be due to algorithmic bias, data preprocessing issues, or other factors affecting the model's predictions. The unequal distribution of false positive rates might result in some demographic groups experiencing a higher likelihood of being falsely identified, leading to unfair treatment.

False Discovery Rate Parity (Failed)

False Discovery Rate Parity aims to ensure that all protected groups have equally proportional false positives within the selected set compared to the reference group. For age, the audit found disparities of 2.38 and 1.31 times for the age groups 17-28 and 28-37, respectively, compared to the reference group (37-48). This suggests that different age groups have varying false discovery rates, which might be due to the algorithm's inability to accurately identify true positives across demographic groups. This lack of equal distribution in the false discovery rates indicates that the model might be less reliable in detecting true positives for certain demographic groups. Consequently, the audit failed this metric.

False Negative Rate Parity (Passed)

False Negative Rate Parity aims to ensure that all protected groups have the same false negative rates as the reference group. Based on the fairness threshold used, all groups passed the audit for this metric. This indicates that the false negative rates are similar across different demographic groups, suggesting that the model is more consistent in terms of false negative rates. The equal distribution of false negative rates is a positive sign in terms of fairness.

False Omission Rate Parity (Failed)

False Omission Rate Parity aims to ensure that all protected groups have equally proportional false negatives within the non-selected set compared to the reference group. The audit found disparities of 7.64, 8.66, and 5.79 times for the age groups 48-90, 37-48, and 28-37, respectively, compared to the reference group (17-28). This indicates that different age groups have varying false omission rates, which might be attributed to the model's inability to correctly identify true negatives for specific demographic groups. The unequal distribution of false omission rates indicates potential biases and unfairness in the dataset or the algorithm, which may lead to an unfair treatment of the data.

In conclusion, the Aequitas Bias Report reveals significant fairness concerns in the audited dataset, with only False Negative Rate Parity passing the audit. The disparities in representation and error rates across demographic groups suggest that there may be inherent biases in the dataset, the data collection preprocess or process, or the algorithm used to generate predictions.

Dataset 2 Analysis

As we said previously, we are going to focus on the 'Men', 'Women', 'Hispanic', 'White', 'Black', 'Native', 'Asian' and 'Pacific' variables.

Summary of the Audit Results:

- Equal Parity: Failed
- Proportional Parity: Failed
- False Positive Rate Parity: Failed
- False Discovery Rate Parity: Failed
- False Negative Rate Parity: Failed
- False Omission Rate Parity: Failed

Equal Parity (Failed)

This metric failed the audit, which indicates that protected groups in the dataset are not equally represented in the selected set. In other words, certain groups are being disproportionately selected or not selected, causing an imbalance in representation. This could be a result of biased data collection, underrepresentation, or other systemic factors that have led to an unequal representation of the groups.

Proportional Parity (Failed)

The audit failed for all groups except for the ‘Pacific’ variable, meaning that the selection of the groups is not proportional to their population percentages. The failure of this test may be attributed to the algorithm not considering the proportional population percentages of the protected groups during the selection process. This could be due to the algorithm’s inability to capture the importance of the protected attribute or because it is prioritizing other factors that lead to disproportional selection.

False Positive Rate Parity (Failed)

This metric failed the audit for all groups as well, suggesting that the false positive rates for the protected groups are not equal. This test may have failed because the algorithm is more likely to incorrectly predict positive outcomes for certain protected groups compared to the reference group. This could be due to biases in the training data, improper feature selection, or algorithmic bias that disproportionately affects certain groups.

False Discovery Rate Parity (Failed)

The audit failure for this metric (except for the ‘Pacific’ variable) means that the proportion of false positives within the selected set is not equal among the protected groups when compared to the reference group. This could be caused by biases in the training data or the algorithm’s inability to generalize well for certain groups due to limited or unrepresentative data.

False Negative Rate Parity (Failed)

The audit failed for all groups except for the ‘Hispanic’, ‘White’, and ‘Pacific’ variables, suggesting that the false negative rates for the protected groups are not equal. This test may have failed because the algorithm is more likely to incorrectly predict negative outcomes for certain protected groups compared to the reference group. Similar to the False Positive Rate Parity, the reasons for this could include biases in the training data, improper feature selection, or algorithmic bias that disproportionately impacts certain groups.

False Omission Rate Parity (Failed)

The failure of this metric for all groups suggests that the false omission rates for the protected groups are not equal. The failure of this test might be related to the algorithm’s tendency to have a higher proportion of false negatives within the non-selected set for certain protected groups compared to the reference group, which could be due to biases in the data or the algorithm.

In summary, the Aequitas Bias Report indicates that the data has failed all parity tests, suggesting significant disparities among the protected groups. The results highlight the need for further investigation and potential adjustments to ensure fair representation and treatment for all groups.

Mitigating Bias

To address and try to mitigate the bias and fairness described above, it is essential to investigate the reasons behind the disparities and consider different strategies. Some approaches to take are:

- Improve data collection: ensure that the data being used to train the algorithm is representative of the population that is being target.

- Data pre-processing: apply techniques like re-weighting, re-sampling (over-sample under-represented groups or under-sample over-represented groups), or adversarial training to make the data more balanced and reduce the impact of biases on the algorithm’s performance.

- Data post-processing: adjust the decision-making thresholds for different groups to reduce disparities in the algorithm’s outputs. This may involve setting different thresholds for false positives or false negatives to achieve parity across groups.

- Algorithmic adjustments: explore alternative algorithms or fairness-aware machine learning models that can learn representations that do not discriminate against protected groups. Some methods include adversarial training, fairness constraints, or fairness regularizers.

- Regular audits and monitoring: continuously monitor the dataset and model performance to ensure fairness and adjust the mitigation strategies as necessary.

By implementing these strategies, it is possible to make the dataset more representative and fairer to all demographic groups, leading to more accurate, fair, and unbiased insights.

5. Conclusions

Doing this assignment, we have learnt the importance data integration has in the process of studying and working with data. Moreover, working with tools such as Aequitas has helped us understand the importance of studying both bias and fairness, which ultimately shows whether a set of data might be ready to be used in an information retrieval process.

However, it has not been without having to face multiple challenges that we have had to identify certain aspects of the different tools used that made it possible to work with them. Otherwise, it would certainly have been difficult to continue and progress with our work.

For instance, FairLens, one of the tools shown in class to perform bias analysis did not properly work in windows systems that did not have the C++ compilation modules from Visual Studio installed, something quite obscure and tedious to identify. Moreover, we were also unable to work with the FairLens library in macOS, given the fact that it is not compatible with the ARM distributions of this OS.

Nevertheless, and as has already been commented before, the tool we ultimately used was Aequitas, considering all the problems mentioned above regarding the FairLens library. Making use of Aequitas, we were able to perform all the desired tasks, but not before having to deal with the fact that the tool would not properly function in the case that we did not properly arrange the dataset in its desired schema, something that we did not identify obviously enough.

Regarding the datasets, we think it has been an adequate choice for the tasks that were to be performed. These two datasets, although represent almost the same information, did not follow the same data distribution, representation, and most importantly, did not follow the same schema, but instead radically different ones. We thought of this at the time of performing the selection and decided this was the right way to go, mainly due to the difficulties it would bring at the time of evaluating the different metrics, something that we thought would be beneficial for the learning process.

Thanks to this work, we were able to profoundly understand the foundations of data integration, the opportunities it brings when done correctly, and the immense consequences that are to be faced when it is poorly performed or is not even done at all. We think the importance of data integration will exponentially increase in the foreseeable future, which makes awareness for it to be crucial and more necessary than ever before, reasons why we thought of this assignment to be very beneficial, enlightening, and constructive.