Information Retrieval, Extraction and Integration

Data Integration, Bias and Fairness Assignment



| Group 8 Members |
| --- |
| Julio Nevado Delgado |
| David Cabornero Pascual |
| Wenqi Jiang |

# Scenario

Mental health has raised polemics nowadays. Especially in Spain (and generally in Europe), there is a debate about how working conditions, job instability and other political issues may affect people’s mental health.

In the technological sector, there are several features that make this kind of job stressful: tight deadlines, people working overtime to reach the goals, repetitive tasks while programming… In our assignment, we are going to focus on the mental health state of this sector and its relationship with some economic and health parameters.

In particular, there are several hypotheses we could try to demonstrate or reject in this assignment:

* The unemployment rate in a country could lead to more competency between coworkers, ending up with more stressful and less stable jobs.
* The inflation rate leads to an increase in prices, causing many families to decide to work longer hours for less pay. It could finally affect the mental health of the population in a country.
* How different factors, not only related to the subject itself but also related to the country he lives in (employment rate and inflation) affect the chances that an employee seeks for psychological treatment?

With all this, our **complex query** is: *Does the mental health quality of technological workers in a certain country depend on economical indicators such as the unemployment and inflation rate, and also on personal factors such as age or working conditions?*

# Datasets

Three datasets have been utilized in order to solve the previous complex query:

* Mental Health in Tech Survey [1]. This survey provides data about 1259 tech workers with personal data about sex, age and country, and several answers to mental health-related questions, such as interferences in their work because of any mental issue, whether they sought help or received treatment. This survey was carried out from 2014 to 2016.
* The employment rate in OECD countries [2]. This dataset includes the unemployment rate of every OECD country in the years we are interested in.
* The annual inflation in OECD countries [2]. This dataset includes the annual inflation rate in every OECD country. Since inflation generally reflects the purchasing power of the people in a country, this dataset could help us to conclude if the less purchasing power, the more mental health issues in a country. Again, there is data only about the years we are interested in.

# Conflicts

Our goal when integrating the previous datasets is being able to look at data from a unified perspective. To achieve that, we need a common schema for our final dataset.

In order to fulfill this, we make use of tools such as Python and OpenRefine, which allows us to surf over data in a quite interactive and manageable way.

However, the first we need is to make clear what we would like to obtain from the combination of the four previous datasets. Remember, we would like to compare employees' mental health.

## 3.1 Datasets

### 3.1.1 Mental health in tech survey dataset

One of the main issues can be found in the attribute “Gender” where many of the different possible values seem to refer to the same thing (“Cis Male”, “cis male” and “cis man”, or “Female” and “female”). Additionally, for the same variable some trash seems to be recorded in values such as “A little about you”. Additionally, options out of the classical “male” and “female” are so specific that it is difficult to gain information on them. A good option would be to combine the different options into only three:

* Male
* Female
* Others (including no binaries, LGTBIQ+ collective…).

Other problems are typical within data science problems. We refer to the missing values and outliers problem. Out of the 16 attributes of the dataset, the following ones contain one of them:

* Age (negative ages and age 99999999999)
* State (515 NA, not concerning us since we are not going to make use of this attribute)
* Self employed (18 NA)
* Work\_interfere (264 NA)
* Comments (1095 NA)

Additionally, some columns are going to be useless for our case and will have to be removed. These attributes are:

* Timestamp: refers to the time where each data item was recorded and therefore does not affect the well-being of the subject.
* State: It is missing for every subject who does not belong to the US. Additionally, datasets on inflation and employment rate contain information only at a country level. There is a mismatch on the scope of the data in this case.
* Comments: about 75% of data items do not include comments. We do not target natural language analysis so we will get rid of them.

### 3.1.2 Employment Rate and Inflation Rate datasets

Employment rate and Inflation rate datasets come from the same source, so they are quite similar. That is why they will be analyzed together.

Both datasets contain information for several years. We should take only the column belonging to 2015 since it belongs to the range of years we are studying.

Additionally, for the employment rate dataset, we lack yearly information, so we are forced to average data from the four quarters of a year to get the data we seek.

## 3.2 Inter-dataset problems

### 3.2.1 Employment Rate and Inflation datasets

As said, in the case of the employment rate and inflation datasets both of them are quite similar since both come from the same source, OECD stats. However, there is a problem regarding the units of the attributes of interest. For the dataset 2 (employment rate), this information is in terms of percentages while for the case of the dataset on inflation the information is in terms of parts per unit.

Many algorithms are quite sensitive on the units of the attributes so it would be very interesting if we expressed that information using a common scale.

### 3.2.1 Problem with countries included

We can see in the following table that not for every country every piece of data exists. Countries in yellow remark those cases where we have data for every attribute.

Only for 26 countries full information is available out of the 59 countries for which we have any kind of information. This means we will consider employment and inflation information for only those subjects belonging to one of those 26 countries.

Another problem, only in the case of China, is the case of different names. Namely, in the Mental Health dataset, China is written as “China” but in the case of the inflation rate dataset China is referred to as “China (People’s Republic of)”.

Finally, another disparity is with decimal symbols. For Age, the decimal mark is a period, while for inflation and employment the symbol is a comma.

|  | **Mental health** | **Employment rate** | **Inflation rate** |
| --- | --- | --- | --- |
| **Argentina** |  |  | x |
| **Australia** | x | x | x |
| **Austria** | x | x | x |
| **Bahamas** | x |  |  |
| **Belgium** | x | x | x |
| **Bosnia and Herzegovina** | x |  |  |
| **Brazil** | x |  | x |
| **Bulgaria** | x |  |  |
| **Canada** | x | x | x |
| **Chile** |  | x | x |
| **China** | x |  | x (China (People’s Republic of)) |
| **Colombia** | x | x | x |
| **Costa Rica** | x | x |  |
| **Croatia** | x |  |  |
| **Czech Republic** | x | x | x |
| **Denmark** | x | x | x |
| **Estonia** |  | x | x |
| **Finland** | x | x | x |
| **France** |  | x | x |
| **Georgia** | x |  |  |
| **Germany** | x | x | x |
| **Greece** | x | x | x |
| **Hungary** | x | x | x |
| **Iceland** |  | x | x |
| **India** | x |  | x |
| **Indonesia** |  |  | x |
| **Ireland** | x | x | x |
| **Israel** | x | x | x |
| **Italy** | x | x | x |
| **Japan** | x | x | x |
| **Korea** |  | x | x |
| **Latvia** | x | x |  |
| **Lithuania** |  | x |  |
| **Luxembourg** |  | x | x |
| **Mexico** | x |  | x |
| **Moldova** | x |  |  |
| **Netherlands** | x |  | x |
| **New Zealand** | x |  | x |
| **Nigeria** | x |  |  |
| **Norway** | x | x | x |
| **Philippines** | x |  |  |
| **Poland** | x | x | x |
| **Portugal** | x | x | x |
| **Romania** | x |  |  |
| **Russia** | x | x | x |
| **Saudi Arabia** |  |  | x |
| **Singapore** | x |  |  |
| **Slovak Republic** |  | x | x |
| **Slovenia** | x | x | x |
| **South Africa** | x | x | x |
| **Spain** | x | x | x |
| **Sweden** | x | x | x |
| **Switzerland** | x | x | x |
| **Thailand** | x |  |  |
| **Turkey** |  | x | x |
| **United Kingdom** | x | x | x |
| **United States** | x | x | x |
| **Uruguay** | x |  |  |
| **Zimbabwe** | x |  |  |

## 3.3 Desired Schema

## 

1. Age: Age of the employee
2. Gender: Gender of the employee
3. Country: Which country?
4. self\_employed: Are you self-employed?
5. family\_history: Do you have a family history of mental illness?
6. treatment: Have you sought treatment for a mental health condition?
7. work\_interfere: If you have a mental health condition, do you feel that it interferes with your work?
8. no\_employees: How many employees does your company or organization have?
9. remote\_work: Do you work remotely (outside of an office) at least 50% of the time?
10. tech\_company: Is your employer primarily a tech company/organization?
11. benefits: Does your employer provide mental health benefits?
12. care\_options: Do you know the options for mental health care your employer provides?
13. wellness\_program: Has your employer ever discussed mental health as part of an employee wellness program?
14. seek\_help: Does your employer provide resources to learn more about mental health issues and how to seek help?
15. anonymity: Is your anonymity protected if you choose to take advantage of mental health or substance abuse treatment resources?
16. leave: How easy is it for you to take medical leave for a mental health condition?
17. mental\_health\_consequence: Do you think that discussing a mental health issue with your employer would have negative consequences?
18. phys\_health\_consequence: Do you think that discussing a physical health issue with your employer would have negative consequences?
19. coworkers: Would you be willing to discuss a mental health issue with your coworkers?
20. supervisor: Would you be willing to discuss a mental health issue with your direct supervisor(s)?
21. mental\_health\_interview: Would you bring up a mental health issue with a potential employer in an interview?
22. phys\_health\_interview: Would you bring up a physical health issue with a potential employer in an interview?
23. mental\_vs\_physical: Do you feel that your employer takes mental health as seriously as physical health?
24. obs\_consequence: Have you heard of or observed negative consequences for coworkers with mental health conditions in your workplace?
25. Employment: Employment rate in this country
26. Inflation: Inflation rate in this country

# Bias and Fairness

## 4.1 Bias Detection: AIF360

AI Fairness 360 is an extensible open source toolkit can help we examine, report, and mitigate discrimination and bias in machine learning models throughout the AI application lifecycle. In our case, we wanted to know if there would be gender bias in the results when searching for the impact of economic circumstances on the mental health of technology workers in a specific country. So how do we quantify such bias? There are a number of different methods, and we used a metric known as the “Disparate Impact Ratio”:

The differential impact ratio is simply the ratio of positive results in the unprivileged group (in this case, females) divided by the ratio of positive outcomes in the privileged group (males). According to the AIF360 tool, an acceptable lower bound is 0.8, which means that a disparate impact violation occurs when the unprivileged group receives a good outcome less than 80% of the time that the privileged group does. And similarly, we can assess how close the predictions are to equality of odds. Average Odds Error is a relaxed version of equality of odds, which returns the average of the absolute difference in FPR and TPR for the unprivileged and privileged groups: i.e.:

After calculated, our Disparate Impact Ratio is 1.2549 and Average Odds Error is 0.0652.

We used Reweighing in the AIF360 toolkit, which is a preprocessing technique that weights the examples in each (group, label) combination differently to ensure fairness before classification.

After Reweighing, we got 1.2362 as Disparate Impact Ratio, and 0.0357 as Average Odds Error.

So we can conclude from the calculated result data that the gender bias in our query results may not be as serious as believed.

## 4.2 Fairness Detection: Aequitas

The tool used in this case is Aequitas, an intuitive web tool which allows us to identify bias and fairness in our datasets.

Aequitas needs a very specific format of the input dataset. The following is needed:

· label\_value: the ground truth for the dataset. In our case, we chose treatment (if a subject decides to take a treatment or not).

· score: the label assigned by a classifier. We will use a naïve bayes classifier.

· Categorical attributes

· Numerical attributes: Aequitas will split them to work with categorical attributes.

We will choose a set of protected attributes, which are the attributes we are going to “audit” looking for bias or fairness. For each attribute, we will identify a privileged group (or reference group as stated in Aequitas) in order to assess the direction of bias, if it exists. In other words, the reference group will be like the benchmark, the ground, over which we are going to look for biases for other groups inside the same attribute.

For simplicity, we will choose the option “Majority groups” which takes as reference group the largest one for each attribute.

Additionally, Aequitas allows us to choose several metrics to compute. These metrics are:

· Equal parity: means that every possible class is equally represented. If sex was “male” or “female”, an equal parity of 0 would mean that 50% of examples are “male” and 50% of examples are “female”.

· Proportional parity: means that the dataset reproduces the actual distribution of categories. If there existed a 60% of “male” people and 40% of “female” people, then a dataset with proportional parity would have a 60% of “male” examples and a 40% of “female” ones.

· False Positive Rate Parity: False Positive Rate Parity is referred to the False Positive Ratio being equal for every category of the audited attribute. In other words, we would have a fair FPRP if the FPR was equal for every group, in our example, “male” and “female” examples.

· False Discovery Rate Parity: the same as False Positive Rate Parity but for the case of False Discovery Rate.

· False Negative Rate Parity: the same as False Positive Rate Parity but for the case of False Negative Rate.

· False Omission Rate Parity: the same as False Positive Rate Parity but for False Omission Rate.

We will focus on analyzing equal parity since we would like to focus equally on subjects from any sex, age, and country.

The procedure to be followed is:

1. Discretize continuous attributes. The one of interest in this case is age, which has been split as:

a. Teenager: 0-20

b. Twenties: 20-40 (we joined twenties with thirties to obtain more subjects in that interval)

c. Forties: 40-50

d. Old: 50-

2. Train a multilayer perceptron trying to predict if a treatment is sought or not

3. Try to detect bias or fairness in protected attributes: age and sex.

### Age

Equal parity is a criterion failed. In other words, the integrated dataset lacks fairness and is biased with respect to the majority group. For the case of Age, the following is obtained:

Previous picture states that the fact that a subject decides to go through a psychological treatment is biased towards the twenties (people between 20 and 30 years old). We can see that, while disparity is very small for the groups twenties (20 - 30), old (50-), and teenagers (-20), it is huge, about 27%, with respect to subjects belonging to forties (40-50).

In other words, the distribution over the dependent variable (treatment) for the “forties” tier is 27% different compared to “twenties” tier. If for male subjects we had a 50% of “male” for treatment and a 50% of male for no treatment, this would mean that the classifier yielded, for instance, a 23% for treatment and a 77% for no treatment for the subjects that are “female”.

Since the fairness threshold is set to 80%, the audit fails since the outcome is out of the interval 80% - 125%.

### Sex

For the case of sex, the following results are obtained.

Similarly to the case of Age, fairness is not achieved for the case of Sex. Again, the difference between “male” and one of the other classes, “other” in this case is small enough to be inside the fairness range. However, comparing “male” with “female” the disparity is 36%, out of that margin of 20% due to the fact of using a fairness threshold of 80%.

### Future Choices based on Aequitas

We have seen that the dataset is highly biased for both attributes, Age and Sex. There exist several solutions in these cases, among which we would consider the following:

· Reweighting: basically, consist of considering these categories weighting higher those that are “discriminated” by the model.

· A simpler approach would be Reject Option Based Classification which, around the boundary, is going to favor unprivileged groups. In other words, it is as if the boundary was extended to make the domain of the unprivileged groups bigger.

· Doing nothing: Due to several factors, people from different ages and belonging to different genders may be more likely to start psychological therapy.

· Looking for additional datasets.

We can see that there exist several options. However, to choose the most appropriate one we should get to know better the data collection process and the nature of the data itself.

# 5. Difficulties

The development of the following work has conveyed the following issues:

* Understanding attributes
* Homogenize data types between different attributes.
* Finding an appropriate discretization for attributes.
* Obtaining the appropriate schema for Aequitas.
* Fully understanding Aequitas outcomes and concepts like protected attributes, privileged groups.
* Getting to know a framework such as AIF360 in little time.

References

[1] Kaggle.com. 2022. Mental Health in Tech Survey. [online] Available at: <https://www.kaggle.com/datasets/osmi/mental-health-in-tech-survey> [Accessed 28 April 2022].

[2] Stats.oecd.org. 2022. *OECD Statistics*. [online] Available at: <https://stats.oecd.org/> [Accessed 28 April 2022].

[3] AIF360 2022. [online] Available at: <https://aif360.mybluemix.net/> [Accessed 28 April 2022].