Convergencia
ISSN: 1405-1435
ISSN: 2448-5799
Universidad Autónoma del Estado de México, Facultad de Ciencias Políticas y Administración

Edelsztein, Valeria Carolina; Waisbrot, Sebastián Ariel
Breaking down the Gender Pay Gap through a machine learning model
Convergencia, vol. 30, e20656, 2023
Universidad Autónoma del Estado de México, Facultad de Ciencias Políticas y Administración

DOI: https://doi.org/10.29101/crcs.v30i0.20656

Available in: https://www.redalyc.org/articulo.oa?id=10574559005

CONVERGENCIA
REVISTA DE CIENCIAS SOCIALES

https://doi.org/10.29101/crcs.v30i0.20656

Breaking down the Gender Pay Gap through a machine learning model

Descomponiendo la brecha salarial a través de un modelo de aprendizaje automático

Valeria Carolina Edelsztein* **https://orcid.org/0000-0001-6739-1825**

Universidad de Buenos Aires, Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Argentina, valecaroedel@yahoo.com

Sebastián Ariel Waisbrot **https://orcid.org/0000-0002-8056-4150**

Observatorio de Derecho Informático Argentino (O.D.I.A.), Argentina, seppo0010@gmail.com.

*Corresponding author: Valeria Carolina Edelsztein valecaroedel@yahoo.com

Abstract: Being able to decompose the gender pay gap (GPG) and determine the contribution of each component is important to design appropriate policies to reduce it. With the aim of providing a new tool to achieve this, in this paper, we propose a decomposition approach based on a machine learning model. The tool was implemented on a population of 5 742 Argentinean IT-related workers to obtain the value of the adjusted and unadjusted GPG in a four-phase process: sample characterization, development of a wage predictor, calculation of adjusted GPG, and analysis of the explained component of GPG. According to our analysis, there is a GPG of 20%, 7,7% of which can be explained exclusively by direct discrimination while 12,3% can be ascribed to other factors, such as total years of experience, educational level, and number of people in charge.

Key words: gender wage gap, labor market discrimination, machine learning, women's labor-force participation, wage disparities.

Received: 20/01/2023

Accepted: 17/05/2023

Published: 28/06/2023

Resumen: Poder descomponer la brecha salarial (GPG) y determinar la contribución de cada componente es importante para diseñar políticas adecuadas para reducirla. Con el objetivo de aportar una nueva herramienta para lograrlo, en este trabajo proponemos un enfoque de descomposición basado en un modelo de aprendizaje automático. La herramienta se implementó en una población de 5 742 trabajadores argentinos relacionados con la informática para obtener el valor de la GPG ajustada y no ajustada en un proceso de cuatro fases: caracterización de la muestra, desarrollo de un predictor de salarios, cálculo de GPG ajustada y análisis del componente explicado de la GPG. Según nuestro

análisis, existe una GPG de 20%, de la cual 7,7% puede explicarse exclusivamente por discriminación directa, mientras que 12,3% puede atribuirse a otros factores, como años totales de experiencia, nivel educativo y número de personas a cargo.

Palabras clave: brecha de género salarial, discriminación del mercado laboral, aprendizaje automático, participación de las mujeres en la fuerza laboral, disparidad salarial.

Introduction

For the past few decades, the interest in reducing the gender pay gap (GPG), that is to diminish the imbalances in wages between men and women, has increased in labor markets (Bishu and Alkadry, 2016: 65; Blau and Kahn, 2017: 789). However, despite the efforts made, gender parity in pay has proved hard to achieve and it persists worldwide, to a greater or lesser extent (Goldin, 2014: 1091; Blau and Kahn, 2017: 789). According to the World Economic Forum (2020), at current rates of change, the global gender gap will close in more than 250 years.

The Global Gender Gap Index (2020) shows that there has been no great progress towards closing the gap, and not even one out of the 153 countries reported have yet achieved gender parity in salaries. The International Labour Organization (ILO) (2018) use average hourly wages based on data for 73 countries estimates that the global GPG stands at around 16%, though using a factor-weighted approach the global estimate rises to 19%. Similar results emerge from the Global Gender Gap Report (World Economic Forum, 2020). According to the Organisation for Economic Co-operation and Development (OECD) data, the differential in men's and women's median income is about 13.5% and improving, while for non-OECD countries is around 15% and worsening. However, these figures vary greatly by country. For example, in 2021, within the European Union where the salary gap lies around 16%, the highest GPG was recorded in Latvia (22.3%) and in Estonia (21.7%), the lowest in Luxembourg (1.3%) and Romania (3.0%) (European Commission, 2021). GPGs are notorious even in countries with equal pay legislation. In the United States, the gap has remained around 17 to 20% for at least fifteen years despite the Equal Pay Act of 1963 (Fontenot *et al.*, 2018; U.S. Bureau of Labor Statistics, 2020), in the UK, with its Equal Pay Act of 1970, and France, which legislated in 1972, the gaps are nearly 17% and 12% respectively, and in Australia it remains around 12% (OECD, 2021).

Predictably, numbers do not get better when moving to Latin American and Caribbean countries. In 2018, this region fell in the middle of the overall global GPG, behind Western Europe, North America, Eastern Europe, and Central Asian countries (International Labour Organization, 2018). On average, women in the region work 25 hours more per month than the average man (United Nations, 2020), and half of them work for no pay or profit at all. Argentina, Brazil, Chile, Colombia, Mexico, Peru, and Venezuela prohibit gender pay discrimination and most countries embrace the ILO's notion of "work of equal value." Nevertheless, in Argentina women earn on average 29% less than men (D'Alessandro *et al.*, 2020). This difference is observed in all occupational categories and the gap becomes greater when analyzing the hierarchical positions.

Breaking down the gap

Accurately measuring the GPG is relevant to assess how far we are from equality since it is a symbol of women's position in the workforce in comparison to men. However, the unadjusted (raw) GPG is a complex indicator. Although it provides an overall picture of the difference between men and women salaries, it could mask the fact that this difference can be attributed not only to direct discrimination through 'unequal pay for equal work', but also related to many other factors equally or more powerful determinants of male-female earnings differentials, including social and historical factors, such as the concentration of one gender in certain activities ('segregation') ,and the ease of access to higher paid hierarchical positions ('glass ceiling') (Karamessini and Ioakimoglou, 2007: 31). Therefore, being able to decompose the GPG and to determine the contribution of its different components is important to design appropriate policies for reducing it. In this regard, some efforts have been made, typically through regression models that are based on Mincer-type wage equations, Oaxaca-Blinder decompositions or further developments of this method, or the Wellington approach to estimate the GPG along time, among other proposals (Karamessini and Ioakimoglou, 2007: 31; Chernozhukov *et al.*, 2013: 2205; Blau and Kahn, 2017: 789; Töpfer, 2017; Amado *et al.*, 2018: 357; European Commission. Statistical Office of the European Union, 2018).

Traditionally, economists' approaches to understand the GPG have rested on two sets of economic theories: *human capital model* and *models of labor market discrimination* (Grybaitė, 2006: 85; Ospino *et al.*, 2010: 237).

The human capital model is usually used to analyze the so-called *explained part* of the GPG (adjusted GPG), the one that can be attributed to differences in qualifications. The basic idea is that every person has some form of human capital, i.e., the abilities and skills acquired through education, training and working experience. According to this model, if women have less human capital than men, they should rightfully receive lower wages (Mincer and Polachek, 1974: S76; Polachek, 1981: 60; Grybaitė, 2006: 85). On average, making this assumption is valid and there are numerous reasons to explain it. For instance, women tend to have lesser labor market experience because they work part-time and intermittently due to the traditional division of labor by gender, maternity and the hourly dedication to housework and childcare. Moreover, all of these factors result in fewer incentives to invest in education and training (Becker, 1985: S33; Grybaitė, 2006: 85). It is important to point out, as several studies have already done, that human capital factors are based on broad assumption and does not take into account the fact that women and men cannot be studied as individuals independently from material and social frameworks since all decisions are made in a normative context where there are pre-established ideas about what women and men ought to do (Grybaitė, 2006: 85).

Although *the human capital model* plays an important role in explaining the GPG, it does not account for the total gap. The remaining portion of the GPG, that is the *unexplained part* of the gap, is generally presumed to be due to *labor market discrimination* and refers to difference in salaries for workers that have the same abilities, experience and training. Therefore, it can be defined as direct discrimination since it accounts for 'unequal pay for equal work' (Grybaitė, 2006: 85). There are several models aimed at trying to understand this portion of the gap, such as the *statistical discrimination model* proposed by Edmund Phelps (1972: 659), other statistical models (Becker, 1971; Bergmann, 1974: 103; Aigner and Cain, 1977: 175; Lundberg and Startz, 1983: 340), or the *overcrowding model* developed by Barbara Bergmann (1974: 103). However, none of these models, while helpful in understanding some of the reasons behind the *unexplained part* of the GPG, manage to fully encompass it or propose ways to resolve inequity.

Both contributions to the GPG have varied over time. For instance, in the USA, by 2010, conventional human capital variables (education and labor-market experience) that were an important part of the GPG decades before, have decreased in importance probably due to the reversal of the gender difference in education and the substantial reduction in the experience gap (Blau and Kahn, 2017: 789). However, the persistence of an unexplained GPG suggests that labor-market discrimination continues to contribute. In 2014, the average adjusted GPG for the EU was 11.5% with a variation from 2.5% in Belgium to 24.2% in Lithuania. Notably, in many EU countries, the adjusted GPG is higher than the unadjusted figure. This means that women are expected to earn *more* than men due to better, on average, characteristics in the labour market (European Commission, 2018).

Since the unexplained part reflects differences in salaries of subjects with supposedly identical characteristics aside from gender, it could be claimed to reflect direct discrimination (Goldin, 2014: 1091).

To estimate the unexplained (adjusted) GPG, that is the pay penalty of being female is not an easy task since it is necessary to control for all relevant factors that are simultaneously correlated with salaries and gender (Fortin *et al.*, 2011), such as experience, educational level, abilities, position (Goldin, 2006: 1; Mandel and Semyonov, 2014: 1597; Blau and Kahn, 2017: 789; Töpfer and Brieland, 2022).

Machine Learning, a novel approach

As Qin and Chiang (2019: 465) point out, over the last 20 years there has been a revolution in statistical science given the possibility of 'extracting important patterns and trends, and understanding "what the data says"', so-called "learning from data" thanks to big data analysis and machine learning (ML).

Machine learning (ML) is a branch of artificial intelligence (AI) in which algorithms —that is sequences of statistical processing steps— are trained to find patterns in massive amounts of data in order to make decisions, inferences, and predictions. The resulting trained algorithm is the ML model.

There are four basic steps for building a ML model:

(1) *Select and prepare a training data set*. Representative to solve the problem in question. In some cases, the training data is *labeled* data, that is 'tagged' to call out features and classifications the model will need to

identify. Other data is *unlabeled*. Data is usually divided into subsets for training and cross-evaluation (also known as "leave one out method").

(2) *Choose an architecture, an algorithm to run on the training data set*. The type of algorithm depends on the type (labeled or unlabeled) and the amount of data in the training data set and on the type of problem to be solved. Common types of ML algorithms for using with labeled data include linear and logistic regression, and decision trees while algorithms for use with unlabeled data include clustering algorithms and association algorithms.

(3) *Training the algorithm to create the model*.

(4) *Using and improving the model*. The final step is to use the model with new data and, in the best case, to improve its accuracy and effectiveness over time.

The whole process is an iterative one.

In this regard, ML methods could offer a new approach to estimate the adjusted GPG (Karimian *et al.*, 2019; Bonaccolto-Töpfer and Briel, 2022). In the literature, relevant variables to control for the estimation of the adjusted GPG are typically chosen based on economic reasoning. However, there is a limited understanding of the functional form, which includes identifying relevant interactions and polynomials (Bonaccolto-Töpfer and Briel, 2022). Additionally, certain character skills may have a nonlinear impact on wages. Estimating the adjusted GPG becomes particularly complicated when there is a lot of heterogeneous data since numerous factors may contribute to pay differences between genders, and their relevance may vary depending on the wage level being considered. ML methods provide a more systematic approach to avoid arbitrary selection of variables (Bonaccolto-Töpfer and Briel, 2022).

Being able to perform a GPG decomposition controlling simultaneously for several factors, even in heterogeneous samples could help clarify the different effect each of the variables have to diminish this gap, and globally understand how each of the components vary over time.

Undoubtedly, the most effective approach for assessing the level of discrimination would be to compare an individual's salary with what they would earn in the exact same conditions if they were of the opposite gender (Alatrista-Salas *et al.*, 2017). However, this method is unrealistic since it is impossible to observe someone's characteristics as the opposite gender. Nevertheless, by utilizing ML, we can simulate this scenario to some extent and that is what we attempted to do in this research.

In this article, we will show how through a ML model and based on specific information (gender, age, years of experience, number of employees at their company, position, etc.), provided by a population of 5,742 IT-related workers through an anonymous online survey conducted by a community called *Sysarmy* we can infer with certain degree of precision, this person's salary. Based on this model, we will propose a decomposition approach for the GPG to find out the value for the unexplained GPG considering size sample disparity and the factors determinants of male-female earnings differentials in the explained part of the GPG.

Methodology

This investigation consisted of four phases: (I) characterization of the sample, (II) creating the salary predictor, (III) adjusting the GPG, (IV) and analyzing the explained part of the GPG. Subsequently, we will develop the methodology deployed in each of them and in the following section. We will share our results.

Phase I. Characterization of the sample

Our original idea was to generate a salary predictor for IT-related workers. To do that, a ML model was developed from real salary data provided by an open and anonymous survey on IT-related workers. The answers were from a self-selected population, and therefore did not represent the entire IT population of Argentina, but it allowed us to analyze trends.

- *Period.* The recollection was made between the months of December 2019 and January 2020 in Argentina.
- *Features.* Workers were asked about their gender identity, years of experience, workplace, number of employees in their companies, level, and area of studies, among other factors. Full datasets and features are of public access (Sysarmy, 2020). In Table 1,[1] we have listed the features that were considered for this study along with a brief description or the elective options for each of them.
- *Income data.* In Argentina (a country with high rates of informal labor), the salary is typically paid monthly and is the value

1 All tables and figures are at the end of this article (Editor's Note).

commonly used for economic estimators (for example, poverty and indigence indices are estimated based on monthly family income, subsidies for electricity and gas are also granted based on these incomes). Therefore, we considered for the analysis only those participants who have a full-time monthly salary (40 hours per week) and their gross monthly income. This, of course, implies a selection bias but minimizes the error of including individuals, who work as freelancers and also standardizes the number of worked hours regardless of gender.

The salary values mentioned in the article correspond to Argentine pesos. Between December 2019 and 2020 —the period during which the information was collected—, on average, in Argentina, one US dollar was equivalent to 63 Argentine pesos. The inflation indexes for these months were 3.7% and 2.3%, respectively. This could bring some variability to the analysis because we did not have the date of each record to normalize the values if there were any salary adjustments.

Data preparation

First, we tried to detect and eliminate anomalies —that is, extreme values that did not make sense for salaries, number of employees, age, etc.— that were introduced maybe intentionally due to typing errors or misunderstanding of the questions. For instance, someone answered that their company had over a billion employees, clearly an outlier. Many other workers entered that their salaries were $1, perhaps unemployed people who still wanted to participate in the survey. A person indicated that he had been employed by the same company for 2000 years; it was probably someone who misinterpreted the statement and understood that he was being asked since what year he had been working for his current company.

From a total of 5,982 responses, we finally considered 5,766 in the present analysis, that is, 96% of the total. We were, then, faced with two types of data: numerical (e.g., years of experience and salary) and categorical, that is non-numerical (e.g., gender and province), which had to be transformed into numerical variables.

Numerical data. Despite the fact that it was possible to operate directly with these values, it is important to take into account that they do not necessarily follow a linear progression.

Regarding experience, for example, we might think that it does not have the same effect on salary to go from having no experience to having 1 year, that to go from having 10 to 11 years. That is to say, the more years of experience a person has, the less impact in the salary will have the addition of a new one. A good way to adjust this behavior is by using a logarithmic function. We could also apply logarithm to the salary values themselves because it is not the same to earn $1,000 more for someone who earns $10,000 than for a person who receives $200,000. It is important to note that monotonic transformations of the inputs do not affect the result in a tree model, like the one we used.

Categorical data. To transform non-numerical into numerical data, we created a matrix assigning binary values to each of the categorical characteristics.

Argentina is made up of 24 jurisdictions: 23 provinces and the Autonomous City of Buenos Aires (CABA). Since the cost of living in each jurisdiction is quite uneven, salaries tend to be too. Although people from different provinces responded to the survey, some of the districts had very few answers and that made generalization difficult. So, we decided to group them into larger blocks based on the average salary for each province and cultural similarities that we expected to cause their indicators to behave similarly. Based on this analysis, we divided the provinces as follows:

- Northwest: Catamarca, Jujuy, La Rioja, Salta, Santiago del Estero, Tucumán.
- Northeast: Chaco, Corrientes, Entre Ríos, Formosa, Misiones.
- Cuyo: Mendoza, San Juan, San Luis.
- Pampean Plain: La Pampa, Santa Fe, Córdoba, Buenos Aires.
- Patagonia: Chubut, Neuquén, Río Negro, Santa Cruz, Tierra del Fuego.
- AMBA: CABA and part of Buenos Aires.

Therefore, for geographical data, we used one column for each region and assign each person a binary value for each column. It is important to note that, when grouping, information is lost on provinces that had public technology development policies such as Tierra del Fuego and San Luis.

This approach had the advantage of accounting for features that are not mutually exclusive, such as programming languages: multiple options could be selected in the question and that information could be reflected in

the matrix. Consequently, a man from CABA, who uses Java and JavaScript in his work, will be represented by the values shown in Table 2.

Sample characterization

In order to adequately describe the original sample, we chose certain features of the Sysarmy form that we considered most relevant (some of which would later be used to interpret the explained part of the GPG, see Phase IV), and we analyzed the wage distribution according to these features: age, career, level of education attained, institution where the person studied, specific occupation within the company, number of employees in the company, number of dependents (workers they coordinate), years of experience, and geographical distribution.

Phase II. Creating the salary predictor

Then, we moved towards the construction of the model itself: from the actual data collected we wanted to build a model that would be able to estimate wages from new inputs. The construction of a model requires two steps, (1) *training* and (2) *prediction*. In the training stage, the model received data with labels that represent the 'correct' result, in this case, the salary. In the prediction step, the 'correct' result was unknown and the model had to predict it. It would be very easy to make a model that memorizes the 'correct' answers and gives them as outputs whenever it is asked, but despite being able to perfectly predict these values, this model would have little predictive capacity in the face of unknown data that do not fit exactly those from whom it learned.

So, what we needed was to train the model with a set of known data and evaluate another set of data, also known, but that the model had never 'seen'. This would allow us to know how good the model was. Since we had relatively little data, we chose a technique called *cross validation* (*leave one out method*), which consists of dividing the data into several groups and training the model many times. At each of those times, a different group is excluded from the training and used for the prediction. In this way, if we cross-validate with five groups, we are going to train five different models (models A1 to A5), each with four fifths of the data, and we are going to ask each model to infer the data, that is predict the salary, of the remaining group. Thus, once the process is finished, we will have a prediction for each value, reached by the model that did not 'see' that data when training, and

we could estimate how close to the 'correct' result the predictions were through a coefficient of determination (R^2). For example, if one person earns $100 and the model estimates $110 and for another that earns $200 it predicts $180, then R^2 is 0.9 because it has a 10% error in each estimate. The best possible coefficient is 1, equivalent to say that all the predictions were correct.

Using XGBoost[2], a model based on decision trees that tends to work best in ML. We obtained R^2 for models A1-A5. The code is publicly accessible in Github (Waisbrot, 2022).

Then, it was time to obtain the salary predictor (model A). We created it by training on the whole dataset and using the architecture that had been validated in the previous step (with the five groups A1-A5) (Waisbrot, 2020).

Phase III. Adjusting the GPG

When sharing the salary predictor with the public through social networks to test how well it worked, some users noticed that changing gender from "male" to "female" but keeping all other variables exactly the same often led to a decrease in salary. We decided to explore this situation.

From here onwards, we worked only with the data of those individuals who had identified themselves as male or female (5,742 responses) and deliberately excluded the category "others" because, unfortunately, there were very few data (24 responses).

First, we calculated the unadjusted GPG from the actual salaries reported by the IT workers in the survey through comparing the median values for the salary of men and women. Then, we tried to estimate whether there were differences exclusively by gender, that is the adjusted GPG.

To do that, we constructed the same structure as for model A and trained it with the same data but reversing the gender for each entry yielding five models (B1-B5) and ten expected outcomes: each person had an outcome with a model that accounts for their gender and one outcome

2 We applied logarithms for the analysis of certain features because we were not sure what type of architecture we were going to use to create the models. Based on previous work, we estimated that three characteristics were good predictors of a person's salary: gender, the province where he/she works and how many years of experience he/she has. So, a first simple model we built took into account only these characteristics and ran a linear regression. Then, we decided to use XGBoost and in that case, applying logarithms is indistinct because it does not use functions but splits.

with the model that include their gender reversed. Thus, we learned how much the model predicted it should pay, keeping all the variables the same (education, experience, etc.) except gender.

We also considered the difference in the sample size: we had six times more responses from men than from women. Accordingly, we created another model that compensated for this skew (model C).

Access to the code for the construction of each model is public in Github (Waisbrot, 2020).

Phase IV. Analyzing the explained part of the GPG

To analyze the *explained part* of the GPG we used model A —as it considered the whole dataset— to establish which of the features the prediction model considered relevant, that is which were the more important salary predictors. With this information, we looked at what the variation in salary by gender was for each of these features in the original data set (characterization of the sample in Phase I). It should be noted that it is not enough just to look at the distribution of the data. We must also consider the number of responses in each group. If there were a large pay gap in a group where there are few people, its impact on the total gap would not be significant. Therefore, both variables were analyzed in each case through different methods of plotting numeric data.

Results and Discussion

From the data collected in the survey, we obtained 820 responses from women with a median salary value of $62,050 and 4,922 from men, with a median salary value of $77,000. Therefore, an unadjusted GPG of 20% was obtained in favor of men. In other words, women were paid $0,80 for each $1 paid to men.

Using XGBoost we obtained an average R^2 of 0.5175 for models A1-A5 (in other words, we could explain more than half of the salary with our models), and an average R^2 of 0.6244 for model B.

Model B predicted a median salary for women of $74,243 and of $80,492 for men, that is an adjusted GPG of 7,76%. Furthermore, 12.3% of the gap can be attributed to other factors besides gender (*explained* part).

The salary distribution for males (orange) and females (blue) was plotted according to the actual data collected and from the results obtained

with model B by means of violin plots. This method allows to visualize, both summary statistics and the probability density of the data at different values: for each group the medians, the maximum frequency of each distribution (in both graphics the orange distribution for male salaries shows a peak at higher values), and the areas can be compared (the bigger the difference of orange and blue surfaces, the bigger the gap) (Figure 1).

If instead of calculating the difference in medians, we calculate the average of the point-to-point difference between the predicted wage for each individual considering their gender as female or male and normalizing it by the male wage, then, the difference turns out to be 6.92%.

It is interesting to note that model B seemed to have learned that greater the person's experience, greater the difference between men and women must be (Figure 2). This could be interpreted as the well-known combination of 'sticky floors' and 'glass ceiling' (Ciminelli *et al.*, 2021).

Accounting for the sample size disparity

There were approximately six times more data from men than from women (4922 responses vs. 820 responses, respectively). This disparity brought with it a problem: the model was more unfair to women, i.e., it was less penalized when making a mistake with women. To compensate for this asymmetry, we built model C.

Model C architecture was similar to that of model B (gender reversal) but in order to train it, we "cloned" each woman five times so that each of them was worth six. By giving more "weight" to the data collected from women, the model became fairer, equally penalizing errors for men and women. Thus, the adjusted GPG decreased from 6.92% to 5.77%. Why do we say that the model is fairer? Because if we use it to estimate the salary a woman should receive in a company, it will show that, instead of paying her 6.92% less, we should pay her "only" 5.77% less than men. There is still a gap due to gender alone, but now it is smaller.

Salary predictors: the explained part of the GPG

To estimate salaries, the prediction model A takes as main characteristics years of experience, degree, number of employees in the company, profession, age, level of education attained, college they attended, whether they had finished their degree, and number of people that person has in charge (Figure 3).

Since these characteristics seem to be the most important contributors to a person's salary, we decided to analyze them for the original sample data to better understand the *explained part* of the GPG.

Features that significantly contribute to de GPG

- *Experience*. The main predictor of salary was the experience of the person: the more experienced, the higher the pay. However, as experience increases, the gap between men and women widens and there are fewer women with 10 or more years of experience (Figure 4). Given these, the *experience* contribution to the *explained* GPG is significant.
- *Degree*. In Figure 5, we see that the gap favors men in almost all careers. The most feminized are graphic design and bachelor's degree in administration, where the gap is less significant.
- *Level of education attained*. For each level of study, we see that the income distribution of men is equal to or higher than that of women as shown in Figure 6. The only exception is in the "Secondary in progress" category. However, again, the limited number of data (there are only three women in that category) makes it difficult to draw a conclusion in this regard and, moreover, its influence on the total gap is very low. A relevant detail is that even though, proportionally, more women have completed university or higher education, the distribution of salaries at all levels favors men.
- *Age*. According to the salaries distribution by age, the GPG is small in all groups except the last one (older than 39), where the highest proportion of members is found (Figure 7). The number of employed women decreases significantly above the age of 30. This result reflects known trends linked to childbearing and shows the robustness of the sample, despite being self-selected.

Features that do not significantly contribute to de GPG

- *Number of employees in the company*. Larger companies, with 2,000 employees or more, are the largest contributors to the wage gap but there is no preference in terms of company size by gender.
- *Number of dependents*. As for the number of dependents, the vast majority of people have no dependents and the differences in other

groups do not seem significant. Therefore, we could think that the contribution to the unadjusted GPG is not significant.

- *Profession.* In terms of profession, most of respondents were developers and, in this group the gender pay gap is not evident. However, the well-known "segregation" emerges: QA and UX are the most feminized occupations in contrast with SysAdmin and DevOps.

Conclusions

In this paper we proposed a decomposition approach based on a machine learning model to find out the value of the adjusted and unadjusted GPG among a population of 5742 Argentinean IT-related workers.

From our analysis, based on the current data, there is a GPG of 20%, of which 7.7% can be explained exclusively by direct discrimination (adjusted GPG) while 12.3% can be attributed to other factors, such as total years of experience, degree, level of education attained and age.

We also found evidence of glass ceiling, sticky floor and segregation phenomena and inferred that the influence of age on GPG has a direct correlate with motherhood.

Our proposal has certain limitations, of course: it is possible that the model does not sufficiently fit the data, that there are variables that were not taken into account (because there was no control over the features that were incorporated in the Sysarmy form) and, above all, that there are selection biases given that the sample was not chosen by statistical methods but was self-selected. This self-selection could be the cause of the observed differences. Moreover, one problem of using this model is the fact that the wage distribution is skewed. This, in the future, could be improved by changing the architecture.

In any case, our results are consistent with those in the literature (Blau and Kahn, 2017: 789; European Commission, 2018) and complement results obtained by other research groups (Töpfer and Brieland, 2022). Unlike classic models, ML models allow to work with heterogeneous samples and to juggle a large number of interactions all at once, thus providing new insights to the GPG analysis. It poses a helpful tool for an impending problem that must be tackled from all possible approaches with a main objective: to design appropriate policies for reducing and, eventually, closing the gap.

References

Aigner, Dennis y Cain, Glen (1977), "Statistical Theories of Discrimination in Labor Markets", en *ILR Review*, núm. 30, vol. 2, Estados Unidos: Sage.

Alatrista-Salas, Hugo *et al.* (2017), "Measuring the gender discrimination: A machine learning approach", en *2017 IEEE Latin American Conference on Computational Intelligence (LA-CCI). 2017 IEEE Latin American Conference on Computational Intelligence (LA-CCI)*. Doi: https://doi.org/10.1109/LA-CCI.2017.8285682 Disponible en: https://ieeexplore.ieee.org/document/8285682 [20 de enero de 2023].

Amado, Carla *et al.* (2018), "Measuring and decomposing the gender pay gap: A new frontier approach", en *European Journal of Operational Research*, núm. 1, Holanda: Elsevier.

Becker, Gary (1971), *The Economics of Discrimination*, Estados Unidos: University of Chicago Press.

Becker, Gary (1985), "Human Capital, Effort, and the Sexual Division of Labor", en *Journal of Labor Economics*, núm. 1, Estados Unidos: University of Chicago Press.

Bergmann, Barbara (1974), "Occupational Segregation, Wages and Profits When Employers Discriminate by Race or Sex", en *Eastern Economic Journal*, núm. 2, Alemania: Springer.

Bishu, Sebawit y Alkadry, Mohamad (2016), "A Systematic Review of the Gender Pay Gap and Factors That Predict It", en *Administration and Society*, núm. 49, vol. 1, Estados Unidos: Sage.

Blau, Francine y Kahn, Lawrence (2017), "The Gender Wage Gap: Extent, Trends, and Explanations", en *Journal of Economic Literature*, núm. 55, vol. 3, Estados Unidos: American Economic Association.

Bonaccolto-Töpfer, Marina y Briel, Stephanie (2022), "The gender pay gap revisited: Does machine learning offer new insights?", en *Labour Economics*, 78. Doi: https://doi.org/10.1016/j.labeco.2022.102223 Disponible en: https://www.sciencedirect.com/science/article/abs/pii/S0927537122001130?via%3Dihub [20 de enero de 2023].

Chernozhukov, Víctor *et al.* (2013), "Inference on Counterfactual Distributions", en *Econometrica*, núm. 81, vol. 6, Estados Unidos: Wiley.

Ciminelli, Gabriele *et al.* (2021), "Sticky floors or glass ceilings? The role of human capital, working time flexibility and discrimination in the gender wage gap", en *OECD Economics Department Working Papers*, núm. 1668, Francia: Organisation for Economic Co-operation and Development (OECD).

D'Alessandro, Mercedes *et al.* (2020), "Los cuidados, un sector económico estratégico. Medición del aporte del Trabajo doméstico y de cuidados no remunerado al Producto Interno Bruto". Disponible en: https://observatorio.senadoer.gob.ar/materiales/material/los-cuidados-un-sector-economico-estrategico-medicion-del-aporte-del-trabajo-domestico-y-de-cuidados-no-remunerado-al-producto-interno-bruto/ [20 de enero de 2023].

European Commission (2021), *The gender pay gap situation in the EU*, *European Commission - European Commission*. Disponible en: https://ec.europa.eu/info/policies/justice-and-fundamental-rights/gender-equality/equal-pay/gender-pay-gap-situation-eu_en [20 de enero de 2023].

European Commission. Statistical Office of the European Union (2018), *A decomposition of the unadjusted gender pay gap using structure of earnings survey data: 2018 edition.*

Doi: 10.2785/796328. Disponible en: https://op.europa.eu/en/publication-detail/-/publication/fb389f61-6f7c-11e8-9483-01aa75ed71a1/language-en [20 de enero de 2023].

Fontenot, Kayla *et al.* (2018), "Income and Poverty in the United States: 2017", en *United States, Census Bureau*. Disponible en: https://www.census.gov/library/publications/2018/demo/p60-263.html [20 de enero de 2023].

Fortin, Nichole *et al.* (2011), "Chapter 1 - Decomposition Methods in Economics", en Ashenfelter, Orley y Card, David [eds.], *Handbook of Labor Economics*, Holanda: Elsevier.

Goldin, Claudia (2006), "The Quiet Revolution That Transformed Women's Employment, Education, and Family", en *American Economic Review*, vol. 96, núm. 2, Estados Unidos: American Economic Association.

Goldin, Claudia (2014), "A Grand Gender Convergence: Its Last Chapter", en *The American Economic Review*, vol. 104, núm. 4, Estados Unidos: American Economic Association.

Grybaitė, Virginija (2006), "Analysis of theoretical approaches to gender pay gap", en *Journal of Business Economics and Management*, vol. 7, núm. 2, Reino Unido: Taylor & Francis.

International Labour Organization (2018), "Global Wage Report 2018/19: What lies behind gender pay gaps". Disponible en: http://www.ilo.org/global/publications/books/WCMS_650553/lang--en/index.htm [20 de enero de 2023].

Karamessini, Maria y Ioakimoglou, Elias (2007), "Wage determination and the gender pay gap: A feminist political economy analysis and decomposition", en *Feminist Economics*, vol. 13, núm. 1, Reino Unido: Taylor & Francis.

Karimian, Hamid *et al.* (2019), "A Machine Learning Framework to Identify Employees at Risk of Wage Inequality: U.S. Department of Transportation Case Study", en *The American Economic Review*, vol. 73, núm. 3, Estados Unidos: American Economic Association.

Mandel, Hadas y Semyonov, Moshe (2014), "Gender pay gap and employment sector: sources of earnings disparities in the United States, 1970-2010", en *Demography*, vol. 51, núm. 5, Alemania: Springer.

Mincer, Jacob y Polachek, Solomon (1974), "Family Investments in Human Capital: Earnings of Women", en *Journal of Political Economy*, vol. 82, núm. 2, Estados Unidos: The University of Chicago Press.

Organisation for Economic Co-operation and Development (OECD) (2021), "Earnings and wages - Gender wage gap - OECD Data, OECD data - Gender wage gap". Disponible en: http://data.oecd.org/earnwage/gender-wage-gap.htm [20 de enero de 2023].

Ospino, Carlos *et al.* (2010), "Oaxaca-Blinder wage decomposition: Methods, critiques and applications. A literature review", en *Revista de Economía del Caribe*, núm. 5, Colombia: Editorial Universidad del Norte.

Phelps, Edmund (1972), "The Statistical Theory of Racism and Sexism", en *The American Economic Review*, vol. 62, núm. 4, Estados Unidos: American Economic Association.

Polachek, Solomon (1981), "Occupational Self-Selection: A Human Capital Approach to Sex Differences in Occupational Structure", en *The Review of Economics and Statistics*, vol. 63, núm. 1, Estados Unidos: MIT Press.

Qin, Joe y Chiang, Leo (2019), "Advances and opportunities in machine learning for process data analytics", en *Computers & Chemical Engineering*, núm. 126. Doi:

https://doi.org/10.1016/j.compchemeng.2019.04.003 Disponible en: https://www.sciencedirect.com/science/article/abs/pii/S0098135419302248?via%3Dihub [20 de enero de 2023].

Sysarmy (2020), "Resultados de la Encuesta de sueldos 2020.1, sysarmy". Disponible en: https://sysarmy.com/blog/posts/resultados-de-la-encuesta-de-sueldos-2020-1/ [20 de enero de 2023].

Töpfer, Marina (2017), "Detailed RIF decomposition with selection: The gender pay gap in Italy", en *Hohenheim Discussion Papers in Business, Economics and Social Sciences*, vol. 26, Alemania: University of Hohenheim.

Töpfer, Marina y Briel, Stephanie (2022), "The gender pay gap revisited: Does machine learning offer new insights?", en *Labour Economics*, núm. 78, Holanda: Elsevier.

United Nations (2020), "The World's Women 2020. Trends and Statistics". Disponible en: https://worlds-women-2020-data-undesa.hub.arcgis.com [20 de enero de 2023].

U.S. Bureau of Labor Statistics (2020), "Current Population Survey (CPS) Tables". Disponible en: https://www.bls.gov/cps/tables.htm [20 de enero de 2023].

Waisbrot, Sebastián (2020), "Estimación Sueldo 2020.1". Disponible en: https://seppo0010.github.io/sysarmy-sueldos-2020.1 / [20 de enero de 2023].

Waisbrot, Sebastián (2022), "Seppo0010/sysarmy-sueldos-2020.1". Disponible en:https://github.com/seppo0010/sysarmy-sueldos-2020.1/blob/d9a7c959a033429f669c3a98ef6c278bec192f23/notebook/Brecha%20de%20g%C3%A9nero.ipynb [20 de enero de 2023].

World Economic Forum (2020), "Global Gender Gap Report 2020, World Economic Forum". Disponible en: https://www.weforum.org/reports/gender-gap-2020-report-100-years-pay-equality/ [20 de enero de 2023].

Annex

Table 1

Features considered for the construction of the machine learning model

Feature	Options / Description
Age	Numerical data
I identify as...	Man, women, others.
Where are you working	Any of the 24 provinces into which the country is divided.
Years of experience	Numerical data.
Level of studies achieved	Elementary school, high school, tertiary education, college, postgraduate, PhD.
Status	Refers to the status of the level of education attained. Ongoing, incomplete and completed.
Career	Different IT-related careers such as electronic engineering, system analyst, computing, graphic design, among others.
College	The most relevant universities and colleges were IT-related careers are taught such as the University of Buenos Aires (UBA), the National University of Córdoba (UNC), the National University of La Plata (UNLP), the National Technological University (UTN), among others.
Did you do any specialization course?	On my own, provided by an employer, no.
Number of employees at your current company	Numerical data.
Main activity	Services / Software or Digital Consultant, Software base product, other.
How many employees do you manage?	Numerical data.
Do you contribute to open source?	Yes or no question.
Do you code as a hobby?	Yes or no question.

Feature	Options / Description
Do you have on-call schedules?	Active, passive, no.
Salary in dollars	Yes or no question whether the participant receives his or her salary in dollars (They are mostly pay in pesos argentinos).
I work as...	Different IT-related jobs such as Architect, Data Analyst, Consultant, Developer, HelpDesk, Project Manager, QA / Tester, SysAdmin, among others.
What OS do you use on your work computer?	GNU/linux, macOS, Windows.
And in your phone?	Android, iOS, Windows, I do not own a phone / it is not a smartphone
Sexual orientation	Bisexual or queer, heterosexual, homosexual, other.
What tech events did you attend last year?	Pyconar, nodeconfar, meetups, others.
Tech you use (yes/no)	Answer with yes or no to each tech among many such as bsd, amazon web services, cobol, azure, docker, firebase, heroku, java, javascript, kubernetes, linux, matlab, python, ruby, etc.
Another benefits (yes/no)	Answer with yes or no to each of many possible benefits such as: cellphone or internet plan, language classes, discounts, gym, university or postgraduate payments, parking, extended parental leave, etc.

Table 2

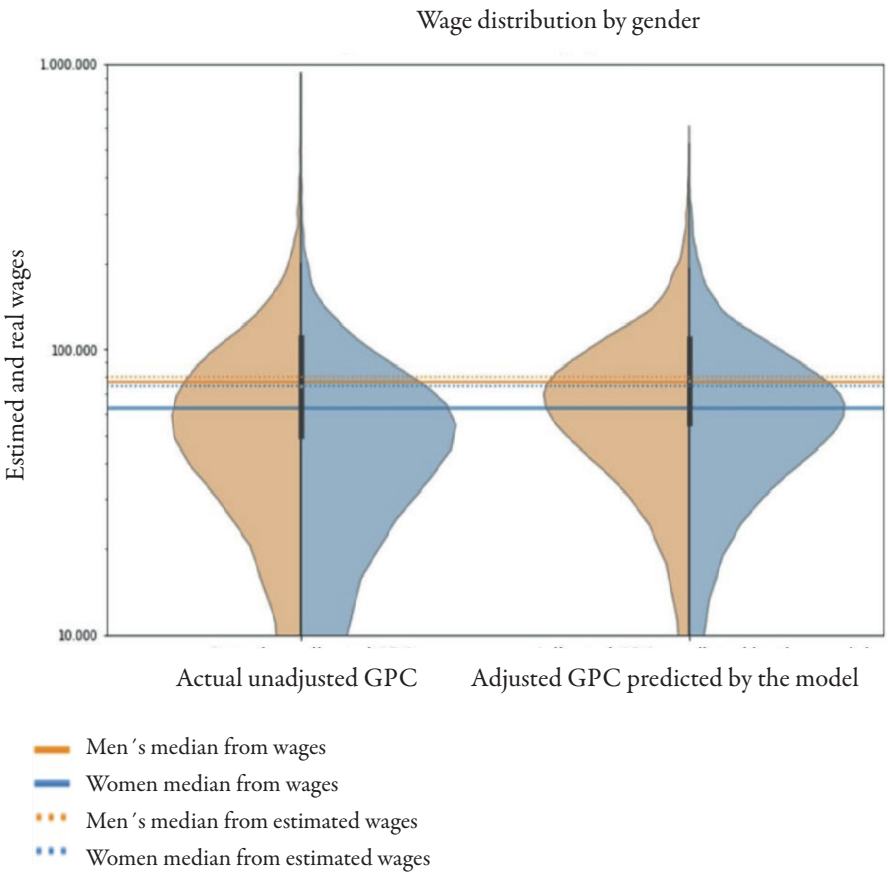
Matrix example for the transformation of categorical to numeric data of a man from CABA, who uses Java and JavaScript

Column	Value
I identified as = man	1
I identified as = women	0
Region = AMBA	1
Region = Cuyo	0
...	...
Programming languages = Java	1
Programming languages = JavaScript	1
Programming languages = rust	0
...	...

Source: Authors' own elaboration based on data collected by Sysarmy.

Figure 1

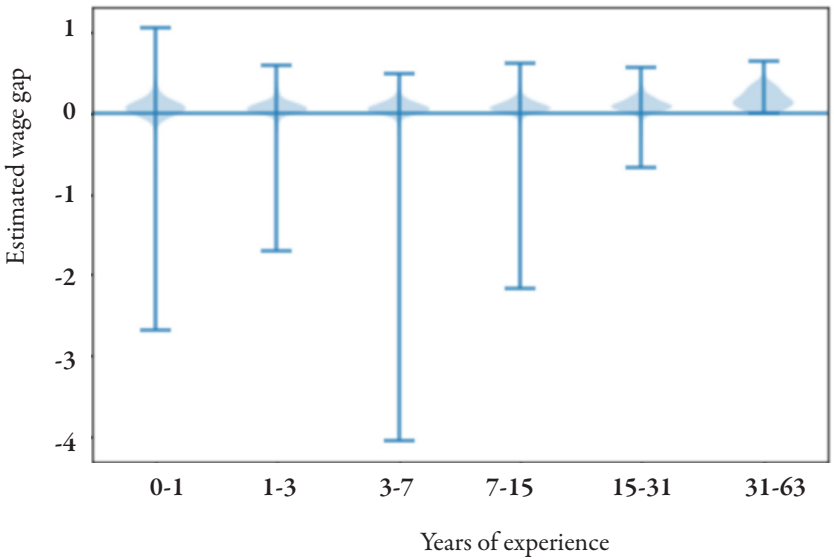
Wage distribution by gender showing the actual salaries for men and women according to the survey data and the salaries predicted by model B



Source: Authors' own elaboration based on data collected by Sysarmy.

Figure 2

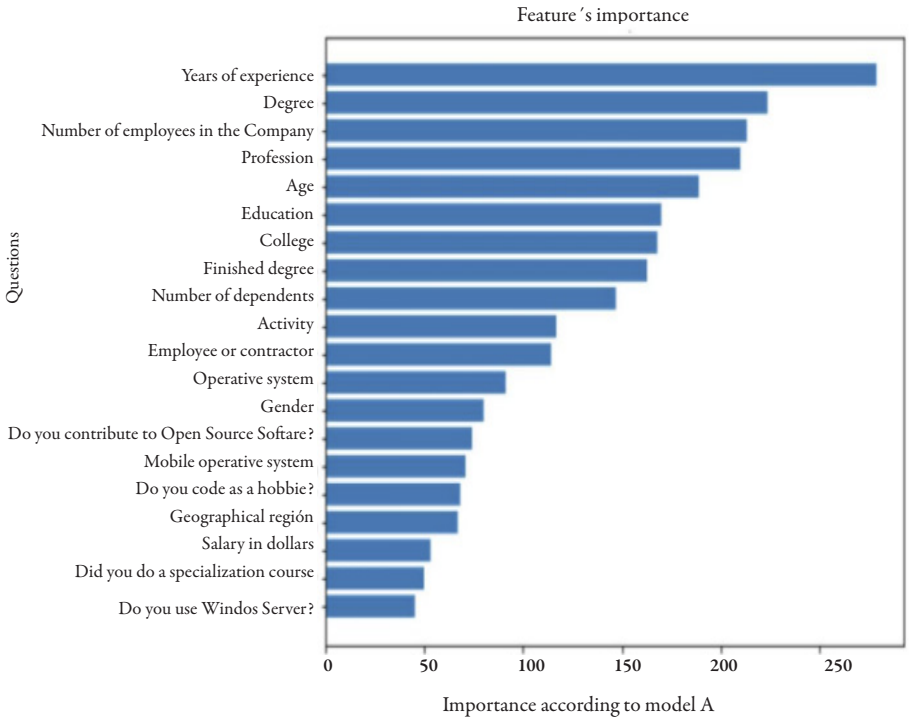
GPG distribution according to years of experience



Source: Authors' own elaboration based on data collected by Sysarmy.

Figure 3

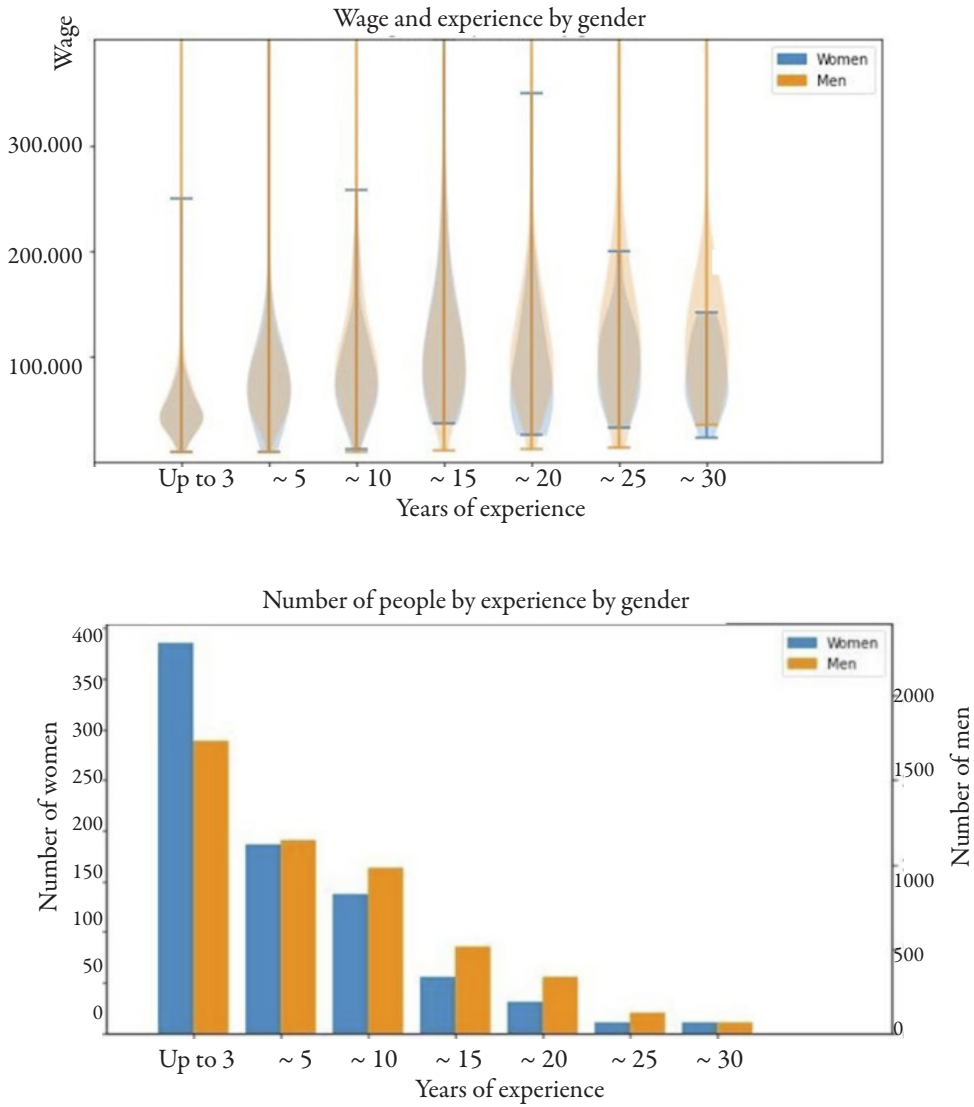
Main characteristics that model A considers when estimating the salary



Source: Authors' own elaboration based on data collected by Sysarmy.

Figure 4

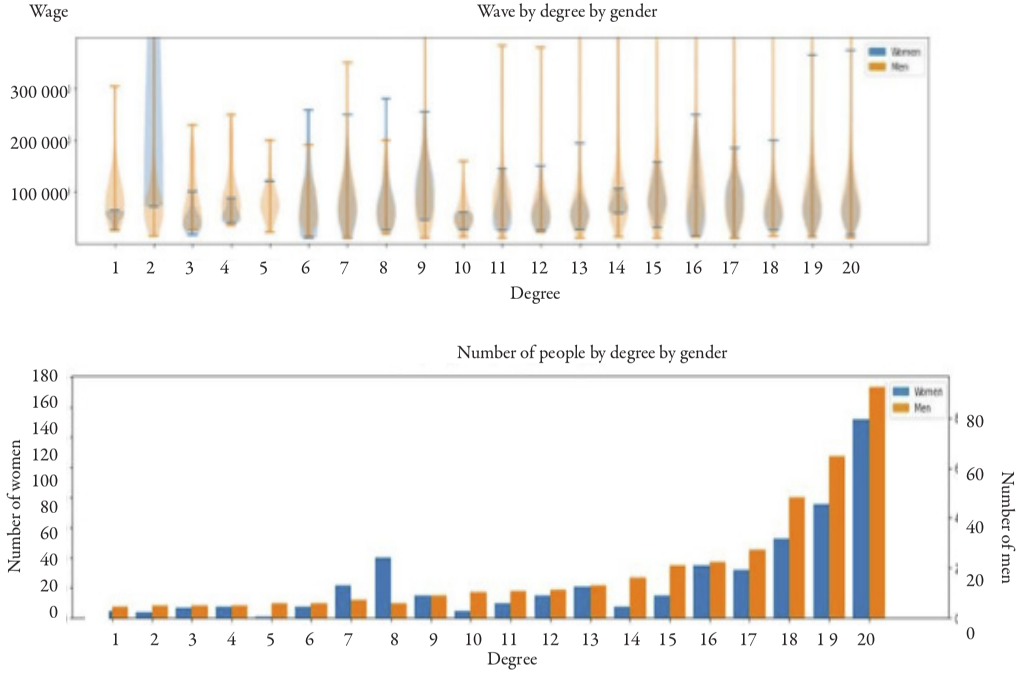
Salaries by years of experience and experience distribution of the sample (women, blue; men, orange)



Source: Authors' own elaboration based on data collected by Sysarmy.

Figure 5

Salaries by degree and degree distribution of the sample (women, blue; men, orange). Since the number of women who responded to the survey was much lower than the number of men, to establish a visual comparison, we had tonormalize. For this reason, the y-axes have different scales for men and women in the lower graph

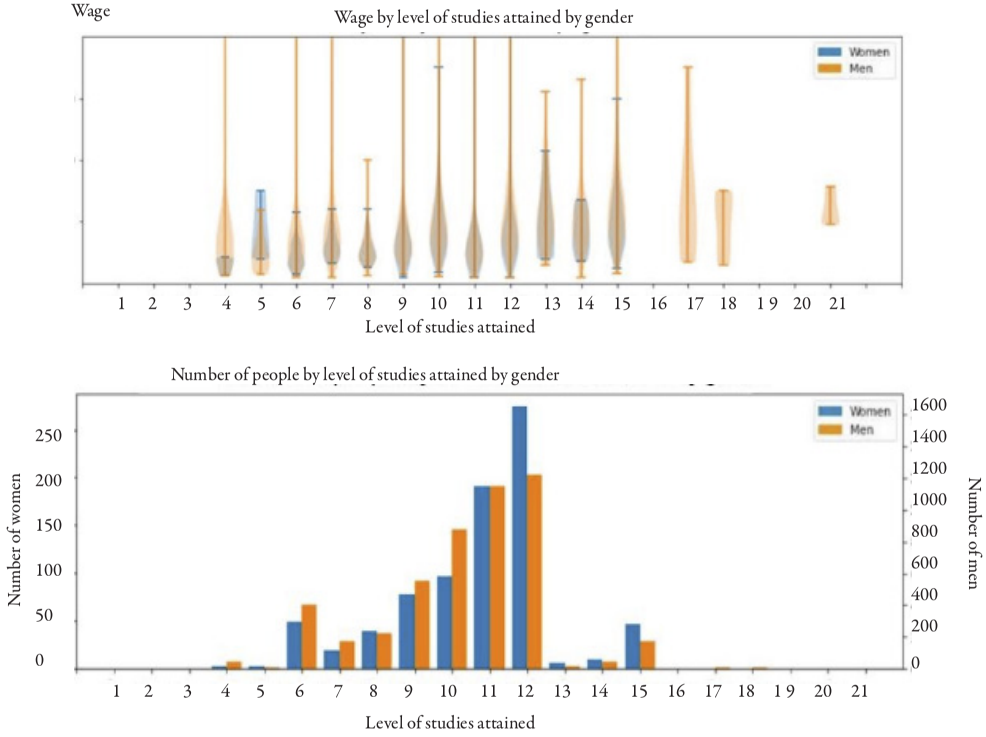


Source: Authors' own elaboration based on data collected by Sysarmy.

Reference

1. Computer Analyst
2. University Programmer Analyst
3. Telecommunication Engineering
4. Higher Technician in Systems Analysis
5. Computer Systems Technician
6. Programmer Analyst
7. Bachelor´s Degree in Administration
8. Graphic Design
9. Bachelor´s Degree in Systems Analysis
10. Computer Science Technician
11. Computer Engineering
12. Advanced Programming Technician
13. Programming Technician
14. Electronic Engineering
15. Bachelor´s Degree in Computer Science
16. Bachelor´s Degree in Computer Science
17. Degree in Information Systems
18. Systems Analyst
19. Computer Engineering
20. Information Systems Engineering

Figure 6
Salaries by level of education attained and level of education distribution of the sample (women, blue; men, orange). It should be noted that categories 17, 18 and 21 involve a very small number of people and, for this reason, the bars are not observed in the lower graph

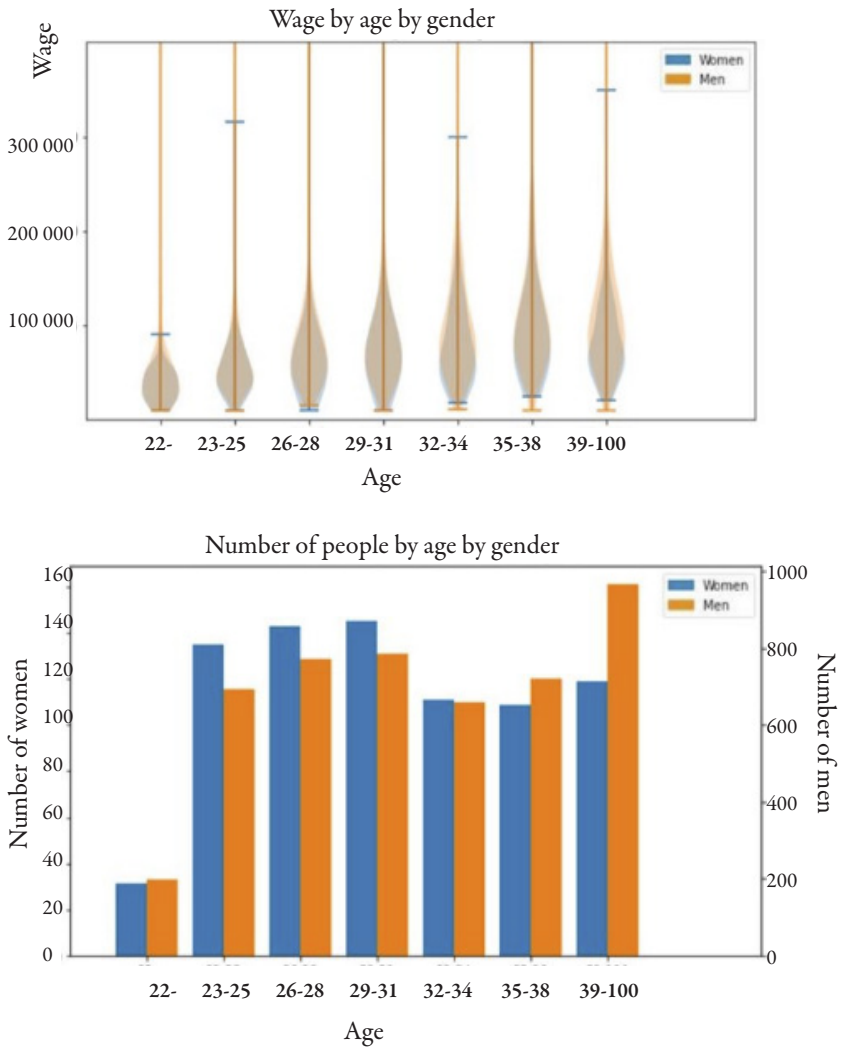


Reference
1. Elementary incomplete
2. Elementary in progress
3. Elementary complete
4. Highschool in progress
5. Highschool in progress
6. Highschool complete
7. Tertiary incomplete
8. Tertiary in progress
9. Tertiary complete
10. College degree incomplete
11. College degree in progress
12. College degree complete
13. Masters degree incomplete
14. Masters degree in progress
15. Masters degree complete
16. PhD incomplete
17. PhD in progress
18. PhD complete
19. Posdoctorate incomplete
20. Posdoctorate in progress
21. Posdoctorate complete

Source: Authors' own elaboration based on data collected by Sysarmy.

Figure 7

Salaries by age and age distribution of the sample (women, blue; men, orange)



Source: Authors' own elaboration based on data collected by Sysarmy.

Valeria Edelsztein. PhD in Chemistry. Centro de Formación e Investigación en Enseñanza de las Ciencias (CEFIEC-UBA), Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Argentina. Research lines: gender perspectives in science, history and philosophy of science. Recent publications: 1) Edelsztein, Valeria and Vázquez, Cecilia (2021), "Checkable nutrition: a scientific literacy experience for students", in *International Journal of Science Education,* vol. 43, no. 5. Doi: https://doi.org/10.1080/09500693.2021.1884315. 2) Lagares, F., Edelsztein, V., Parisi, G., and Rieznik, A. A. (2022), "The effect of handedness on mental arithmetic: A longitudinal large-scale investigation through smart mobile devices", in *Journal of Applied Research in Memory and Cognition*. Doi: https://doi.org/10.1037/mac0000047. 3) Edelsztein, Valeria, Tarzi, Olga and Galagovsky, Lydia (2020), "Chemical senses: a context-based approach to chemistry teaching for lower secondary school students", in *Chemistry Teacher International,* vol. 2, no. 2. Doi: https://doi.org/10.1515/cti-2019-0003

Sebastián Waisbrot. Programmer. Facultad de Derecho, Universidad de Buenos Aires. Member of the Observatorio de Derecho Informático Argentino (O.D.I.A.), Argentina. Research lines: programming.