eman ta zabal zazu

Universidad del País Vasco
Euskal Herriko Unibertsitatea

MASTER IN ECONOMICS
eap
Empirical Applications and Policies

The gender pay gap in Spain: A machine learning approach

Master thesis

Author: Ander Sanchez Maudo
Supervisors: Alaitz Artabe Echevarria, Ainhoa Vega Bayo

Master in Economics: Empirical Applications and Policies
University of the Basque Country (UPV/EHU)
July 21st 2023

Abstract

This thesis investigates the gender pay gap in Spain using Machine Learning (ML) techniques to provide insights and predictive models to understand and address this persistent problem. This study uses a large dataset that includes various labor market factors, such as education, jobs, and industry, to train ML models to predict and explain the gender wage gap. Using advanced methodology, the research aims to identify the key drivers of the wage gap and highlight potential areas where gender-based wage discrimination may exist. The results of this analysis indicate that the gender pay gap is between 14% and 16%. This indicates that according to the estimated models, men are paid between 14 and 16% more in Spain. On the other hand, it is not possible to establish the variables that can explain this gender gap, since most of it is unexplained.

Keywords: unadjusted gender gap, adjusted gender pay gap, machine learning, decomposition.

Table of contents

Index of tables

Index of figures

Introduction

The last century was undoubtedly the turning point in the process of women's emancipation, due to their massive participation in the labor market. Although this process was a milestone, gender inequalities still persist. In this sense, studies on the labor market with a gender perspective usually explore different intertwined gender gaps, such as women's participation or their income. Following Romer's (1990) line of argument, these disparities prevent the exploration of the full potential of women workers and, consequently, could jeopardize further development. Another issue related to gender equality in the labor market is the segregation of men and women in different sectors or activities, with the female ones being usually lower paid and less prestigious. This disparity represents not only a social injustice, but also an obstacle to economic development and equity in contemporary societies.

The wage gap between men and women is an issue of great relevance in the Spanish workplace. Although a significant reduction has been observed since 2007, the report conducted by the Ministry of Labor and Social Economy of the Government of Spain in 2022 reveals that the wage gap still stood at 9.4% in 2020 (Spanish Ministry of Labor, 2022). This income disparity between men and women remains a challenge that requires constant attention from the authorities and society in general. The wage gap is not the only aspect in which gender discrimination in the labor market manifests itself. The same study brings out that woman tend to face greater difficulties in terms of long-term unemployment and lower labor market participation compared to men. These findings highlight the need to comprehensively address gender inequalities in the professional sphere. Among the factors that have been identified as possible causes of the gender pay gap in Spain we can identify (European Commission, 2014): occupational segregation, in which women tend to be more concentrated in lower-paying sectors and occupations; lack of representation of women in managerial and decision-making positions; motherhood, which can lead to career interruptions and limit opportunities for advancement; and gender stereotypes embedded in society that influence the perception of women's job performance.

The analysis of this gender pay gap in Spain is something that has already been done in numerous studies, considering not only the monetary factor of the gap but also exploring different areas of discrimination. Although in our case the analysis will be purely on wage differences, mentioning the literature on this issue can be enriching.

Female employment rates are still quite low in the 21st century (Amuedo-Dorantes & De la Rica, 2006) although they have increased in the last decade (de la Rica et al., 2019), which has been linked to the participation gap and the wage gap. The lack of adequate family reconciliation policies leads many women to leave the labor market unless they have a significant economic advantage to work. However, younger women show a smaller participation gap, which supports the idea that many female workers retire after becoming mothers. In the field of labor research, a large number of studies have examined the differences between men and women in the Spanish and Southern European labor market (Kleven et al., 2018). Moreover, less educated women tend to drop out of the labor market after giving birth due to the lack of comparative advantages, leading to lower labor force participation. To better understand the penalization of children in the wage gap, the relationship between gender and age has been explored, but the results do not provide a plausible explanation as to why the gender gap persists in younger age groups. In addition to family responsibilities, other factors such as gender bias in hiring and aversion to competition in women have been observed (Kleven et al., 2018).

Education also plays an important role in wage disparities between men and women. Women with more education have a wage gap that is less explained by observable factors compared to those with less education. Moreover, in the Spanish labor market, the wage gap seems to increase with the position in the income distribution of women with higher education, which has been called the "ceiling pattern" (Ciminelli et al., 2021). However, this effect is not observed for less educated women, where the gap is larger at the bottom of the distribution, referred to as the "floor pattern". In addition, some factors driving the gender gap are family responsibilities, career interruption after motherhood and cultural gender biases. Conversely, women with higher education face the challenge of reconciling their family and work responsibilities, which can lead to the so-called "glass ceiling trap." These results are consistent with other research indicating that the gender pay gap is explained to a lesser extent by women with higher levels of education.

The academic literature has presented diverse perspectives on the floor and ceiling patterns in the gender gap. Some studies claim that the ceiling pattern prevails in European Union countries (Ciminelli et al., 2021), while others suggest that there is no clear correlation between educational attainment and gender disparities in Southern Europe. However, these approaches are not mutually exclusive, and it is crucial to keep in mind that the gender gap is influenced by multiple factors and may vary depending on the interaction with other variables (Arulampalam et al., 2007).

Regarding gender segregation in certain labor sectors, it has been observed that wage differentials are smaller in male-dominated industries with a lower percentage of female workers (de la Rica et al., 2019). This could suggest that women entering highly competitive sectors tend to be highly skilled and carefully selected professionals. The phenomenon of gender discrimination in the workplace has systemic and complex roots. It is normal to assume that if salaries were more dependent on the performance of each individual, this wage gap would be reduced, since both men and women would obtain similar performance. The reality is different in Spain, as shown in the study by de la Rica et al. (2010). In this country, the unadjusted pay gap is much larger in performance-based components compared to other forms of remuneration. The researchers attribute this disparity to monopolistic tendencies in the employment market, which allow employers to pay wages below the real value of the work performed. This is largely due to the domestic responsibilities that fall on women, which limits their bargaining power in the labor market (de la Rica and Llorens, 2008). In this sense, pay based solely on performance would not solve or eliminate the gender gap.

Interestingly, gender differences in total pay are strongly influenced by the gap in wage supplements, while the gap in base pay is significantly smaller (Amuedo-Dorantes & de la Rica, 2008). This suggests that wage components that are more regulated by the legal framework tend to be less "discriminatory" than those that depend more on individual job performance.

In summary, gender discrimination in the workplace is a complex phenomenon, rooted in systemic and cultural factors. The gender pay gap persists for a variety of reasons, including monopolistic tendencies, women's domestic responsibilities and the influence of wage components (Gneezy et al., 2003). It is essential to continue research and design policies that address these disparities and promote greater gender equality in the employment market.

Thus, as we can see, analyzing and addressing this labor market gap that still persists between men and women is a great challenge and offers numerous avenues of research. In this case, the analysis will focus on the wage gap between both groups for the Spanish case. For this purpose, the different literature and methodology used in previous studies has been studied.

Although there are numerous studies on the gender pay gap, most use traditional econometric techniques (Grimshaw & Rubery, 2002). Empirical research on this topic using ML techniques is more recent but is gaining ground (Bonoccolto-Topfer & Briel, 2022). On the other hand, it is also possible to perform the relevant decomposition of the gender wage gap to know the part of this wage gap that is given by the variables selected for the relevant analysis and the part that is not explained (and could be attributed to gender discrimination). The analysis of this gap can be performed on an adjusted or unadjusted basis. Several studies that perform the adjusted method with techniques such as Lasso or Ridge for countries such as Switzerland and Germany, find that there are substantial differences between using these types of techniques and not using them (Bonoccolto-Topfer & Briel, 2022). In this case, the ML results show a U-shaped gender wage gap across the distribution, with particularly high gaps at the top and bottom of the distribution. Furthermore, Homolka (2022) believes that the use of this regularization analysis realizes interactions between the variables chosen for analysis that perform significantly better and provide more reliable results.

In terms of decomposition, Blau and Kahn (2017) provide a summary of what is known about gender wage disparities. In numerous investigations, Oaxaca-Blinder decompositions or improved versions of this technique are used to split observed gender wage differences into an explained and an unexplained half. The unexplained part is linked to gender disparities in the coefficients, while the explained part results from variations in the control variables. Thus, the unexplained part represents pay disparities between people with the same observable traits, which is why some authors claim that it reflects discrimination (Goldin, 2014). Despite different model specifications, most empirical research yields a sizable gender pay gap residual. It should also be noted that the unexplained part of the gap may or may not be considered pure discrimination, as many papers claim that it does not necessarily measure gender wage discrimination, as this wage may be determined by differences in individual choices (such as educational or career choice). In addition, it is possible that there are factors not captured in the analysis that may influence the observed differences between men and women (Strittmatter & Wunsch, 2021).

Once we have studied our motivation to carry out this project, the contributions to be made have a methodological character to be considered, since the main ways of examining the wage difference between men and women will be the use of ML techniques such as Ridge, Lasso and Elastic Net, something that is not done in a recurrent way for the Spanish case. Subsequently, the Oaxaca-Blinder decomposition will be performed and it will be found out which part of this wage gap can be attributed to the difference explained by the covariates offered by the database to be used and which part remains unexplained. As for the research questions, they can be divided into two groups, those that respond to issues of wage differentials between men and women and those that respond to possible questions that can be asked about the methodology. As for those related to the wage gap between men and women, study to what extent the gender wage gap is significant according to the Wage Structure Survey data for 2018 would be our main objective. Once this wage gap is discovered, the next step would be to examine to what extent this gender wage gap can be explained by the covariates used in the analysis and what part of the gap remains unexplained. As well as suggesting to which factors this difference can be attributed. As for the methodology, an analysis and comparison of each of the models used will be made in order to know which of them is the most adequate depending on a series of criteria and metrics determined later on. On the other hand, finding out if these data are suitable or not for this type of analysis is also a question to be answered.

Therefore, our main objective in this project is to provide evidence on the existence of the gender pay gap and to see its significance, as well as to use different statistical and ML methods to deduce which part of the gender pay gap can be explained by the covariates used and which part is not possible to explain by our analysis. For this purpose, we have used the Wage Structure Survey corresponding to the year 2018.

The structure of the paper is as follows. Section 2 will present the data obtained from the Wage Structure Survey of different European countries together with the main descriptive statistics. Section 4 will discuss the different results obtained through the statistical analysis of ML methods such as, Lasso, Ridge and Elastic Net, described in detail in Section 3. Finally, Section 5 contains the conclusions and policy recommendations, as well as recommendations for future research.

Data

In this section we will provide a description of the Structure of Earnings Survey database and present descriptive results for different characteristics of the individuals in Spain surveyed in the process. These characteristics can be divided into both human capital characteristics (such as age, gender, education level and more) and employment market characteristics (type of contract, sector in which they operate). On the other hand, other important data are salary data. Wages are given on an hourly, monthly and annual basis. It is also segmented, so that a differentiation can be made between the base salary and the supplementary salary.

The data used in this project come from the Structure of Earning Survey (SES) carried out by Eurostat on a four-year basis, the first year being 2002 and the last year with available data being 2018. In this last year in Spain, the number of workers is 218,966 in which there are more than 25,000 different companies of different characteristics. Within these analyzed individuals 56% are women and 44% are men (Table 1). Data on the overall employed population in the last quarter of 2018 offers quite similar data. The percentage of men in the labor market was 54.45% while that of women was 45.45% (INE[1], 2023). Therefore, the sample in the Structure of Earnings Survey is quite representative of the Spanish labor market population.

Table 1. Number of observations in SES

	Individuals	*%*	*Companies*
Female	95,372	43.56%	
Male	123,594	56.44%	
Total	218,966		25,679

Source: Own calculations based on Structure of Earnings Survey for Spain (2018)

The Structure of Earnings Survey (SES) is a 4-yearly survey that provides EU-wide harmonized structural data on gross earnings, hours paid and annual days of paid holiday leave, which are collected under Council Regulation (EC) of the European Union[2] (EU). This large data collection represents a reach data source for policy-making and research purposes. It also provides detailed and comparable information on relationships between the level of hourly, monthly and annual remuneration, personal characteristics of employees (sex, age, occupation, length of service, highest educational level attained, etc.) and their employer (economic activity, size and economic control of the enterprise) (Eurostat, 2023).

The Structure of Earnings Survey offers precise and comprehensive information on earnings and the elements that make up earnings, as well as on level of education. But it has some restrictions. First, there is a problem because we do not have information about people who are not working, which affects our observations and can create a bias in the results. In addition, our research only focuses on work and does not consider other aspects of employees' lives, which limits our full understanding of their personal situation outside of work. This kind of information may be crucial for analyzing gender inequalities in the workplace. In addition, no information is obtained about the years of experience of each worker questioned.

Taking these factors into account, and although this lack of certain data can damage the analysis, there are other variables that make it possible to carry out an accurate and results-rich study. The fact that we have detailed data on hourly, monthly, base and supplementary salaries makes the data to be processed more attractive for the relevant analysis. This, together with the variables referring to the personal characteristics of each worker and those of each firm make the database a more than valid source of information for this Master Thesis.

2.1 Descriptive statistics

In this section we will present the descriptive statistics of the variables that will be used as covariates to explain the gender pay gap. These variables are divided into two segments, the first one dealing with the variables that are related to the worker and the second one with those that have to do with the characteristics of the companies surveyed[3].

2.1.1 Worker level variables

The worker level variables included the SES that we will focus on in this study of the gender pay gap are age[4], education level, occupation skill level and type of contract.

[1] INE stands for "Instituto Nacional de Estadistica" which offers and elaborates public statistics on the Spanish state.

[2] For more information on the legislation of the factors and statistical concepts and definitions see Eurostat.eu, last checked on 19th July 2020.

[3] For further information see **Categorical variable list and definitions** in Appendix.

[4] As the database does not contain information about the experience of each individual, age serves as a proxy, although in many cases the age of the individual and his or her years of experience are not related.

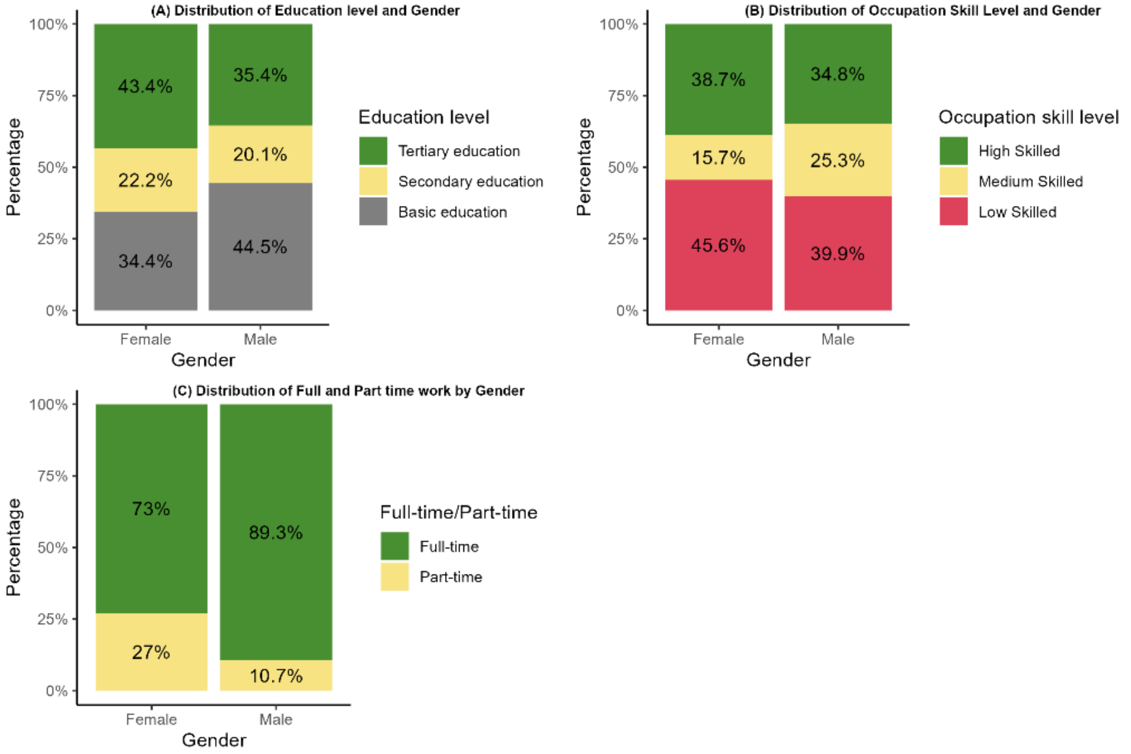
As mentioned above, almost 44% of the database is made up of women, but it is also important to know what percentage of women and men exists in each of the **age** groups analyzed (Figure A 1 in Appendix). This graph gives us a more exhaustive view of the distribution of individuals based on different age ranges, covering ages 14-19 in the first group, 20-29 in the second, 30-39 in the third, 40-49 in the fourth, 50-59 in the fifth and over 60 in the last group. Except for the first segment, in all other segments the proportion does not fall below sixty to forty percent, with the segment between 20 and 29 years of age being the one with the highest proportion of women (45%). As age increases, the number of women in the survey, and therefore working, decreases to 41% in proportion to the total number of individuals. Throughout the descriptive analysis, age will be an important factor to segment the data and to obtain a clearer analysis.

In terms of **education**, in the period and country analyzed, women obtain higher levels of tertiary education than men, with a difference of 8 percentage points. In secondary education the values are similar for both genders, with the female gender being slightly higher, and as for individuals with basic education, almost half of the men in the sample have only basic education, while in women these characteristics are fulfilled in only one third of the sample (Figure 1.A). INE data (2023) show the same proportion of men's educational level, although it varies somewhat in the case of women. The proportion of women with secondary education is the same, but the percentage of women with tertiary education decreases by 4% less than the data provided by Eurostat.

The high educational level of women mentioned above is not reflected in the **occupation skill** level, the percentage of workers in high skilled jobs is equal between the two groups, while in those jobs with low skill requirements, women occupy these jobs to a greater extent than men (Figure 1.B).

Focusing on the **type of contract**, it can be seen that women are more likely to use part-time contracts, with 27% of women represented in the database using part-time contracts compared to almost 11% of men (Figure 1.C). Furthermore, when the analysis is broken down into different age groups, it can be seen that the largest differences are found between 30 and 60 years of age, where 25% of women in this age group use part-time contracts compared to 8% of men (Figure A 2 in Appendix).

Figure 1. Worker Level variables



Source: Own calculations based on Structure of Earnings Survey for Spain (2018)

2.1.2 Firm level variables

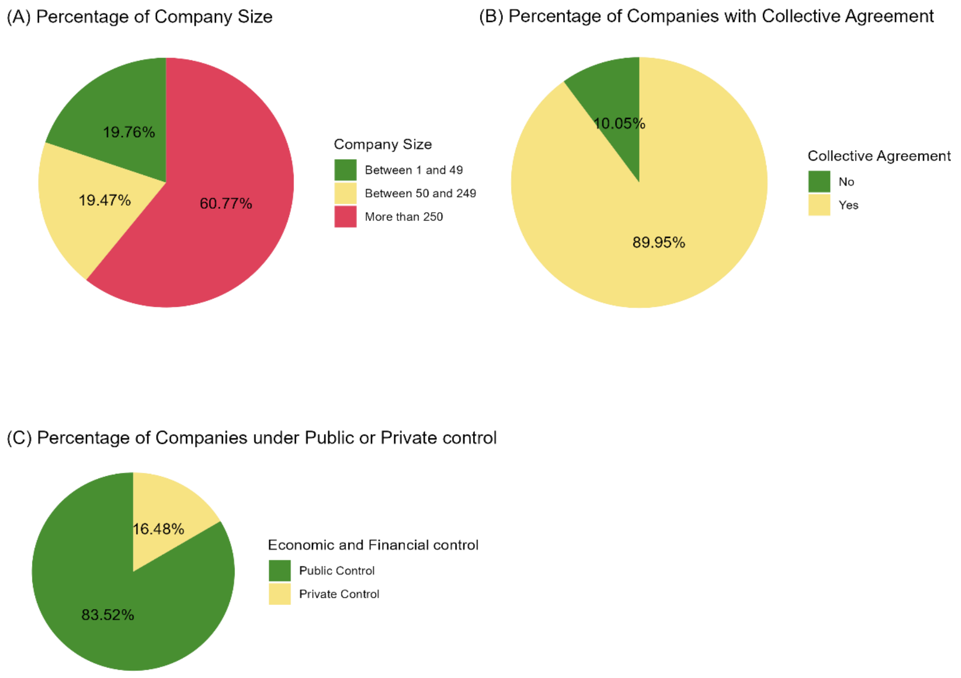
These variables are the size of the company, whether or not it has a collective pay agreement, its economic and financial control and its National Classification of Economic Activities (NACE) code. The NACE is a statistical classification and groups companies/self-employed persons according to the activities they carry out, with the objective of being able to group economic activities according to a defined criterion and the same category. Each NACE consists of a 5-digit code, each of which represents a more specific level of activity (Eurostat , 2008). The first two digits refer to the generic activity of each company and the following digits specify that activity. For the sake of simplicity, in this case only the first two digits of each activity will be taken into account[5].

In regard to the **size** of the company, we can say that the database is mostly composed of large companies (with a number of employees over 250), which make up more than 60% of the sample, while small and medium-sized companies only account for 19% of the sample (Figure 2.A). This data offered by the Structure of Earnings Survey is far from the reality in Spain. However, in Spain only 0.14% of companies were considered large (more than 250 employees) in 2018 (DIRCE[6], 2023), which is not reflected in the data being analyzed.

Regarding the percentage of companies that have a **collective pay agreement**, regardless of the type of agreement, the majority of companies meet the requirement (90%), with only a small number (10%) representing those companies that do not (Figure 2.B).

When analyzing the gender pay gap, it can be interpreted that companies with public **economic-financial control** will close the existing gender gap, which is why the analysis is also carried out on the basis of this control (Figure 2.C). The majority of the companies in the database are publicly controlled, i.e., 83%. While the percentage of privately controlled companies drops to 17%.

Figure 2. Company characteristics


(A) Percentage of Company Size

(B) Percentage of Companies with Collective Agreement

(C) Percentage of Companies under Public or Private control

Source: Own calculations based on Structure of Earnings Survey for Spain (2018)

[5] To see the list of each of the different codes to be taken into account, see Table A 2 in Appendix.

[6] DIRCE stands for "Directorio central de empresas" in Spanish. It is a unique information system that brings together all Spanish companies and their local units located in the national territory.

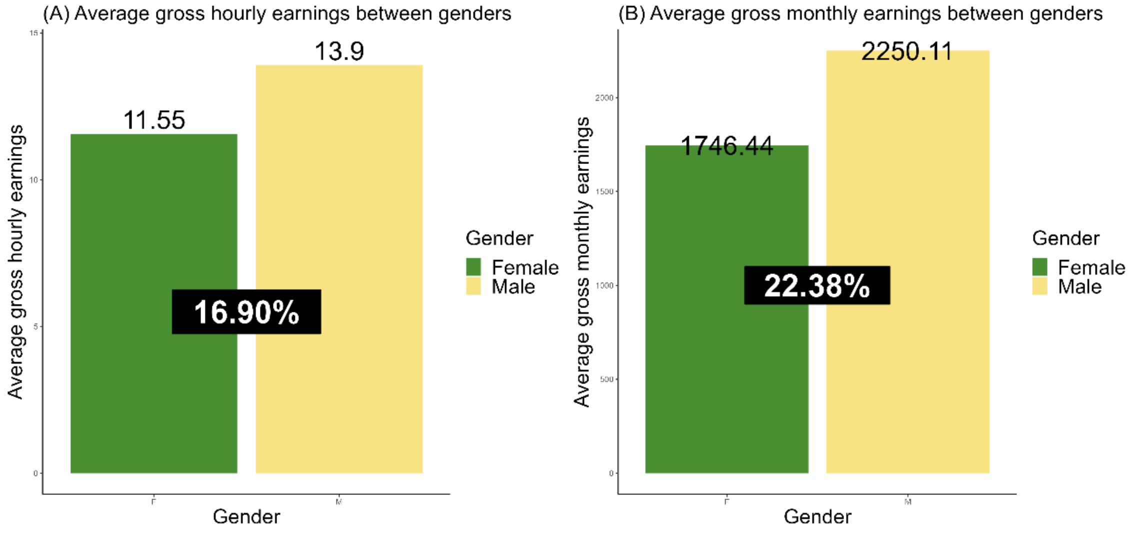
The NACE code serves to give us more information about the enterprises and in which business field each one operates. The most represented code is manufacturing, with almost a quarter of the companies in the database belonging to this group. Other groups also stand out, although they do not reach 10%, such as health and social work, transporting and storage and other business activities. Those with the lowest number of companies are mining and quarrying and other services activities, with values of 0.78% and 1.02%, respectively (Table A 2 in Appendix). Furthermore, if we separate the percentage of workers by gender in each group, we can see which sectors contain a higher number of female workers and therefore which are more or less feminized or masculinized. Labor-intensive sectors, such as construction or mining and quarrying, each have a higher percentage of men than 80% (Table A 2 in Appendix). On the contrary, the sectors with a higher proportion of women are those related to service and care, such as health and social work and education.

2.2 Unadjusted gender pay gap

Having outlined all the personal and company characteristics to be considered in the analysis, this section will describe the unadjusted wage gap. The unadjusted gender pay gap is a descriptive statistic that measures the average wage difference between men and women. Comparing the values of the averages means that the extreme values have a great influence on the result of the analysis and therefore, this type of analysis should be interpreted in a non-literal way. Its measurement covers both possible discrimination between men and women through ‘unequal pay for equal work’ and the differences in the average characteristics of male and female employees (Leythienne & Ronkowski, 2018).

Comparing the monthly salary of each gender, we can observe that the average monthly salary among men exceeds 2,250 euros, while women do not reach 1,750 euros. On average, women earn 500 euros less than men, or 22.38% less than men's salaries (Figure 3.B). The differences in hourly wages decrease considerably, reducing the wage gap to almost 17% (Figure 3.A).

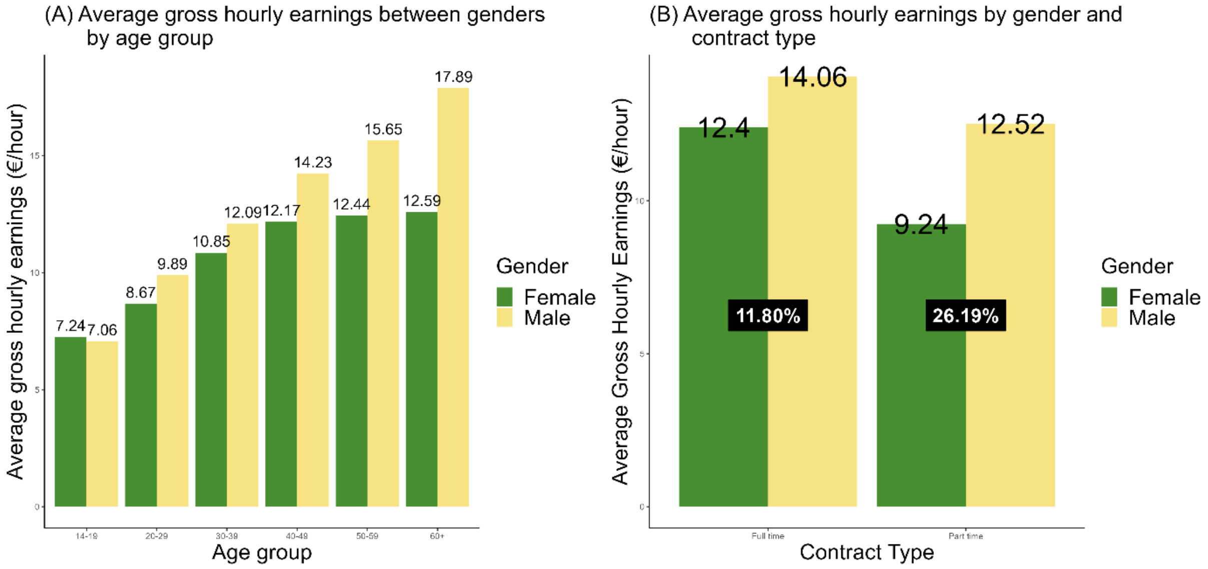
Figure 3. Average gross hourly and monthly earnings



Source: Own calculations based on Structure of Earnings Survey for Spain (2018)

Segregating these data into age groups or contract types could give us different perspectives and a richer look at interpreting the data. As can be seen in Figure 4 A, women earn less at all stages of their career (except for those under 20 years of age), and from the age of 40 onwards their wage growth stagnates at around 12 euros/hour while that of men increases considerably. In the age group between 20 and 29, the wage gap in percentage terms is 12.34%; when we look at the age group over 60, this percentage rises to 29.63%, i.e., the wage gap increases by more than 240%. It should be noted that this difference may be an effect of the existing and evident discrimination between women and men in the twentieth century, it is clear that this discrimination and wage gap has decreased in recent years and it could be that in the future this difference between the elderly will be reduced because of the generational replacement yet to be completed. In terms of the contractual difference, for those who work full-time, the difference in average salary is 11.80% (Figure 4.A), while for part-time workers it rises to 26%. This, together with women's greater use of part-time contracts (Figure 1.C), creates a greater inequality between the two groups of analysis.

Figure 4. Average gross hourly earnings by age groups and contract type



(A) Average gross hourly earnings between genders by age group

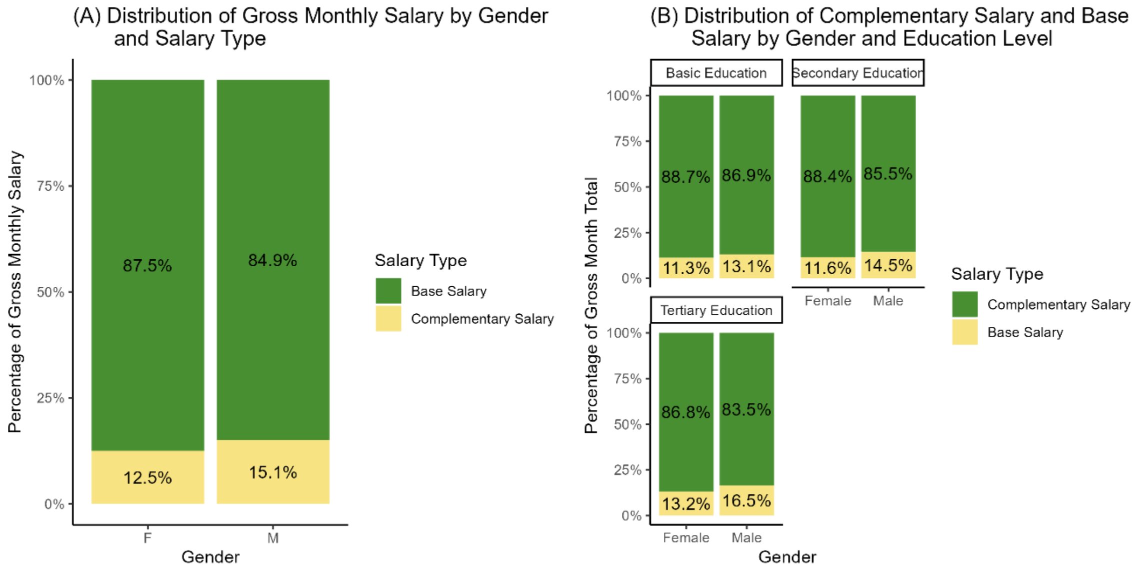
(B) Average gross hourly earnings by gender and contract type

Source: Own calculations based on Structure of Earnings Survey for Spain (2018)

So far, the data presented here are based on the total wage received by workers, either in monthly or hourly terms. For a more exhaustive review of the data, this total wage has been separated into two types: the base wage and the supplementary wage. Employee remuneration refers to all forms of payment or retribution directed to employees and derived from their activities with the company (Dessler, 2009). This remuneration is divided into two parts, and the sum of these parts is the total remuneration of the employee. Base salary refers to the fixed amount that an employee receives regularly, either in the form of a monthly salary or in the form of hourly pay. Supplementary salary are the programs intended for those employees with a high level of performance or sufficient to merit an extra salary that is distinguished from the base salary itself. In the case of our database, the complementary salary is made up of the so-called annual bonuses and allowances not paid at each paid period, annual payments in kind (both prorated in monthly units), monthly earnings related to overtime and monthly special payments for shift work.

The difference in terms of the type of wages received between genders is very low (Figure 5.A), with men's use of supplementary wages being slightly higher than women's, 15.1% versus 12.5%. This difference in the supplementary salary is translated into the base salary, where women have a higher percentage. Due to this small difference, all the analyzes of the unadjusted gender pay gap in terms of base and supplementary salary are similar, in terms of age range, there is practically no peculiarity, in all range’s men get a slightly higher percentage of the use of supplementary salary (Figure A 2in Appendix). The same happens when we make the distinction on the basis of educational level, where the difference remains stable at no more than 3% (Figure 5.B).

Figure 5. Base and Complementary salary related distributions



Source: Own calculations based on Structure of Earnings Survey for Spain (2018)

Methodology

In this section we explain the methodology to carry out the analysis of the adjusted gender pay gap. In the first part, the Ordinary Least Square (OLS) model is used to estimate this gap, and in the second part, a series of ML techniques such as Ridge, Lasso and Elastic Net are used. Lastly, the results are decomposed using the Blinder-Oaxaca method in order to deduce which part of the gap can be explained by the variables selected in the models and which part remains unexplained.

3.1 Adjusted gender gap: Ordinary Least Squares

In this section we will explore more comprehensively the differences in hourly rates, the data being in logarithms. The reason for this analysis is the fact that by running a linear model we can control for differences in gender means and the results are not inflated by extreme values.

The explanatory variables are divided into two categories, those that bring together the different characteristics of the worker and those that bring together the characteristics of the firm in which the worker works. In the first group we find personal characteristics such as age, educational level, etc., and in the second, firm size, NACE code, etc. (Appendix Table A 1).

Once the relevant variables have been chosen, the following model is defined:

Equation 1. OLS Model specification

$$log\, y_i = \alpha + \beta_1 Female_1 + \beta'_2 w'_i + \beta'_3 c'_i + \varepsilon_i$$

Where y_i represents the dependent salary variable (salary per hour), α the constant term, $Female$ is the dummy variable that takes the value 1 if the subject is female and 0 if the subject is male. β_1, β_2 and β_3 are parameters to be estimated. In fact, β_1 is the parameter that will measure the gender pay gap. The vector of covariates of workers' characteristics is represented by w', c' represents the vector of covariates of company characteristics and ε_i the error term.

In turn, we have to bear in mind that the use of the OLS method has a number of limitations (ClockBackward, 2009). An OLS regression is not suitable for dealing with outliers, the model performs poorly with excessively large or small values. Moreover, the assumption that the OLS model is suitable for linear models leads to problems in model estimation, as in reality almost no model is linear. Additionally, the method is not only compromised with extremely low or high values, but also suffers when the variables to be included are numerous.

3.2 Adjusted gender gap: Regularization

Regularization is a much-debated topic in ML methodology and Bayesian statistics. In reality, there is not a big difference between the term ML and the statistical models that are often used in econometrics or statistics, the difference between these two is their main objective and how the researcher deals with the processes involved (Watts, 2014; Yarkoni & Westfall, 2017). When using these types of processes, the main goal is to generate a model that explains the research question adequately with the available data. Thanks to this, ML helps to make more accurate predictions, as the model is able to fit the data provided (Breiman et al., 2001).

Since the main goal of ML is to predict the data accurately, all its strategies are aimed at avoiding over-fitting/under-fitting of the data. The most commonly used strategy is to divide the data into training data and test data. The training data usually consists of a random sample of about 70% of the initial data and the test data consists of the remaining 30% of the initial data. This helps us to build a model on the training set as many times as necessary until the most appropriate model for the data is obtained.

On this basis, this project addresses three different types of regularization approaches, such as Ridge, Lasso and Elastic Net. (Cimentada, 2020).

The **Ridge** is nothing more than a linear regression with some adaptation. In a linear regression the objective is to minimize the sum of the squared residuals; that is, for each fitted regression line, the predicted value is compared, subtracted from the actual value and squared and then summed over all the fitted regression lines. This sum is called the Residual Sum of Squares (RSS) and the linear model chooses the option with the lowest RSS for efficiency and accuracy. The Ridge regression takes the RSS and adds the so-called shrinkage penalty, forcing each β_j coefficient to zero by squaring it:

Equation 2. Ridge Model specification

$$Ridge\,regularization = \; RSS + \; \lambda \sum_{k=1}^{n} \beta_j^2$$

The main objective of this shrinkage is to force each coefficient to be as small as possible without compromising the RSS. The lambda term (λ) can be interpretated as the weight, the greater the weight, the greater the weight given to the shrinkage term of the equation. If this were 0, the penalty term would be zero, reducing the regression to the simple RSS term. The advantage of the ridge regression is that it trades off excellent training data fitting for better generalization. In other words, we enhance bias for reduced variance (making our predictions more reliable) by forcing the coefficients to be smaller. For better generalization on fresh data, very large coefficients are penalized, which is the overall idea underlying ridge regression. Keeping that intuition in mind, it's crucial to remember that the ridge regression's predictors must be normalized. How come this is the case? A predictor's coefficient may be penalized more than those of other predictors because of the scale of the predictor (James et al., 2013).

The **Lasso** regularization is very similar to the Ridge, except that a number of modifications are made to the penalty term. Instead of squaring the coefficients in the penalty term, the lasso regularization takes the absolute value of the coefficient:

Equation 3. Lasso Model specification

$$Lasso\,regularization = \; RSS + \alpha \sum_{k=1}^{n} |\beta_j|$$

The big difference between Lasso and Ridge is that the former can force the coefficients to be exactly zero, while the latter could only approximate them to zero. This means that Lasso selects variables with large coefficients without sacrificing the model's RSS. Ridge regression has the drawback that whereas the training error almost always decreases as the number of variables rises, the test error does not. Since it is possible to take the coefficients until they obtain a value of zero, the Lasso makes a selection of the variables with the higher coefficients and eliminates those which do not have a strong relationship which usually makes it a more appropriate model for interpretation because it removes redundant variables while ridge can be useful if you want to keep a number of variables in the model, despite them being weak predictors.

The **Elastic Net** is the simple sum of the two regularization methods seen above, the Ridge and the Lasso:

Equation 4. Elastic Net Model specification

$$Elastic\ Net\ regularization = RSS + \underbrace{\lambda \sum_{k=1}^{n} \beta_j^2}_{\text{Ridge Penalty Term}} + \underbrace{\alpha \sum_{k=1}^{n} |\beta_j|}_{\text{Lasso Penalty Term}}$$

Boehmke and Greenwell (2020) state that although Lasso models perform feature selection, when two strongly correlated features are pushed towards zero, one may be pushed fully to zero while the other remains in the model. Furthermore, the process of one being in and one being out is not very systematic. In contrast, the Ridge regression penalty is a little more effective in systematically handling correlated features together. Consequently, the advantage of the Elastic Net penalty is that it enables effective regularization via the Ridge penalty with the feature selection characteristics of the Lasso penalty.

3.3 Blinder-Oaxaca Decomposition

The Blinder-Oaxaca decomposition seeks to distinguish between the proportion of the variance in mean results between two groups that can be attributed to group differences in explanatory variable levels and the proportion that can be attributed to group variations in regression coefficient magnitude (Blinder, 1973; Oaxaca, 1973). In order to explain how this method is implemented in our database and processes, we will follow an article by Hlavac (2022) in which the processes are briefly explained[7]. According to the latter author, this decomposition can be carried out in a threefold or twofold manner. In this case the twofold method will be used.

Bearing in mind that we have two groups in our sample, the first being Male and the second Female, the mean outcome difference to be explained ($Diff.$) is simply the difference of the mean outcomes for observations in both groups (Jann, 2008), denoted as $E(Y_M)$ and $E(Y_F)$, respectively:

Equation 5. Difference of the mean outcomes between female and male

$$Diff = E(Y_F) - E(Y_M)$$

In determining the components of this decomposition, the estimation of the unknown vector of non-discriminatory coefficients is necessary. Oaxaca (1973) states that there is reason to believe that discrimination is directed towards one of the groups under examination. In the case of discrimination against women, i.e., positive discrimination against men, the decomposition can be summarized as follows:

Equation 6. Decomposition Model specification

$$Diff = \underbrace{(\bar{X}_F - \bar{X}_M)'\hat{\beta}_F}_{\text{Explained}} + \underbrace{(\hat{\beta}_F^{intercept} - \hat{\beta}_M^{intercept}) + \bar{X}'_F(\hat{\beta}_F - \hat{\beta}_M)}_{\text{Unexplained}}$$

where F identifies as women and M as men. Analogously, $\hat{\beta}_M$ and $\hat{\beta}_F$ are the male and female estimated coefficient vectors, respectively, and $\bar{X}_M$ and $\bar{X}_F$ are the mean values vector of each variable for the given groups. To obtain the value of the coefficients and means, each model was

[7] For a more comprehensive account of the process, see Jann (2008).

estimated separately. First, the means and coefficients were calculated for women and then for men. Thus, the data for all models are obtained separately for women and men. Thanks to this, it has been possible to carry out the specifications expressed in Equation 6. We can also see that these mean outcomes are divided into two parts, the part that can be explained by cross-group differences in the explanatory variables, and the part that remains unexplained by these differences. Often, we tend to think that this unexplained part refers to the existing discrimination of the two groups analyzed, in this case gender, but in reality, this can also result from the influence of the unobserved variables.

Each of these sections will explore different groups of covariates already described above[8], these covariates will be grouped into two distinct groups. The first group called worker level characteristics, will cover the data concerning the worker (Section 3.2.1) and the second will be company characteristics (Section 3.2.2). Thus, the explained and unexplained parts can be separated as follows[9], where w stands for worker characteristics and c for company characteristics:

Equation 7. Decomposition Explained part

$$Explained = (\bar{X}_F^W - \bar{X}_M^W)'\beta_F^W + \left(\bar{X}_F^C - \bar{X}_M^C\right)'\beta_F^C$$

Equation 8. Decomposition Unexplained part

$$Unexplained = \left(\hat{\beta}_F^{intercept} - \hat{\beta}_M^{intercept}\right) + \bar{X}_F^W(\beta_F^W - \beta_M^W) + \bar{X}_F^C\left(\beta_F^C - \beta_M^C\right)$$

Thus, the main objective is to analyze the explained and unexplained part separately. By doing this, one can find out which groups of covariates have a higher weight in each of the groups.

The method used (Blinder-Oaxaca Decomposition) contains a number of drawbacks in the analysis. The results of the difference between genders may vary depending on the choice of the non-discriminatory wage structure, in this case, women are chosen as the reference for the decomposition (Oaxaca & Ransom, 1994; Neumark, 1988). On the other hand, when using categorical variables in the unexplained part, the result is not invariant to the choice of the excluded category in the gender pay gap. To solve this problem, we used the so-called deviation contrast transformation (Gardeazabal & Ugidos, 2004) which tries to impose a zero-sum constraint on the coefficients of the individual categories and expressing the effects as deviations from the overall mean.

Results

The results obtained are presented in two different sections. In the first section we examine the results and error coefficients of each of the four different models selected and already described, and then make a comparison of them. In the second section we proceed to decompose the sizable gender pay gap, in order to detect which part can be explained by the models and which part remains unexplained.

4.1 Model estimations and comparison

Having estimated all the models mentioned in the previous section, the results of these models are shown in Table 2. The model that offers the highest value of wage inequality is the OLS, followed by the Lasso and Elastic Net (which almost coincide in their values) and the lowest value of inequality is the Ridge model. The OLS result is -16.32%, indicating that women earn 16.32% less than men in terms of hourly wages, in the Lasso and Elastic Net this value drops by almost 1 percentage points to 15.17%. The largest drop, as already mentioned, is in the Ridge model, which estimates that women receive a wage 14.48% lower than men. In short, between the models with the highest and lowest percentages there is a difference of almost 2 percentage points.

[8] See 3.2.1 and 3.2.2 for further description.

[9] It should be noted that in ML methods the models themselves perform interactions between the different variables, this group will also be generated for those ML models.

Furthermore, in the OLS model all variables obtain a p-value of less than 0.05[10], which means that all variables are statistically significant at a significance level of 0.05 (5%).

In addition, additional metrics such as the Root Mean Square Error (RMSE), the R^2, the Standard Error and the penalty terms ($\lambda\ and\ \alpha$) have been calculated for the cases of regularization models. These metrics not only provide additional information on the performance and quality of the models, but also allow to evaluate and compare the accuracy in relation to the calculation of the gender wage gap. Thus, not only do we obtain data on the gender gap itself, but we also obtain information about the methods used, and thus be able to determine which of them works more efficiently and gives us more accurate results.

Table 2. Different model estimation results

	OLS	*RIDGE*	*LASSO*	*ELASTIC NET*
ADJUSTED GENDER GAP ($\hat{\beta}_1$)	-16.3234%	-14.4654%	-15.16023%	-15.15952%
STANDARD ERROR	0.07363599	0.006523244	0.00731325	0.007303904
RMSE	0.4026416	0.3867647	0.3856518	0.3856534
R^2	0.4232422	0.4638775	0.4669584	0.4669542
PENALTY TERM (λ)		0.02450519		0.0000490104
PENALTY TERM (α)			0.000024505	0.0000245052

Source: Own calculations based on Structure of Earnings Survey for Spain (2018)

The RMSE is a measure of the difference between model predictions and actual values, the value of RMSE for OLS is 0.4 while for all regularization models it is 0.385; in definitive, the difference between the two seems not to be significant. When the RMSE value is 1, it implies that, on average, the model predictions differ by about 1 unit of the target variable compared to the actual values. On the other hand, an RMSE value of 0 indicates that there is no average difference between the model predictions and the actual values (Chai & Draxler, 2014). Since these results are so similar (0.4 and 0.385), we can infer that all four models perform similarly in terms of prediction accuracy. Being close values, it indicates that both models are producing similar predictions and have comparable performance. However, it is important to note that RMSE alone does not provide a complete assessment of model quality. It is advisable to consider other metrics, such as R^2, the standard error and the penalty term, and to perform a thorough comparison of the models based on their performance and generalizability.

Unlike the RMSE, in the coefficient of determination (R^2) the closer it is to 1, the better the fit of the model to the data (Hahn, 1973). An R^2 value of 0.42 indicates that the OLS model explains approximately 42% of the variability in wages, while an R^2 value of 0.46 indicates that the models with regularization explain approximately 46% of the variability. Therefore, the models with regularization perform slightly better in terms of capturing variability in wages and fitting the observed data.

On the other hand, the standard error of each model has also been calculated. The standard error is a measure of the precision of the sample mean (Altman & Bland, 2005). The lower the value of the standard error coefficient, the higher the precision of the estimate and, therefore, the more reliable the estimated coefficient. The differences between the four models are small, especially in those of ML. The OLS method offers a higher number compared to the ML methods. Therefore, it could be argued that the Ridge is the model with the highest estimation accuracy in terms of standard error, as it offers the lowest value of all.

[10] See Table A 3 in Appendix.

Finally, penalty terms are used in regularization models to control the complexity of the model and avoid overfitting. In the case of the OLS model, there is no penalty term. In contrast, regularization models, such as Ridge, Lasso and Elastic Net, introduce a penalty term. This penalty term helps to restrict the magnitude of the model coefficients, which results in lower model complexity and better generalization to new data. For the Ridge model, the penalty term is 0.0245. For the Lasso model, the penalty term is 0.0000246. For the elastic net model, which combines Ridge and Lasso, the penalty term for λ is 0.0000491, while for α it is 0.0000245052. These values reflect the magnitude of the penalty applied to the coefficients of each model. In the case of the Elastic Net, this includes a combination of the penalties from the Ridge and Lasso models. The specific values obtained for the penalty terms depend on the data and the model fit. These penalty terms influence how the final set of model coefficients is estimated and selected, balancing the trade-off between fit to the data and model complexity.

In conclusion, and having studied each of the values offered by the models, we can arrive at a series of statements. The first model (OLS) is the model that offers the highest value for the difference in wages between the two groups (16.3234%). It performs reasonably well with an acceptable RMSE and R^2 but being the only one that does not belong to the branch of regularization models analyzed here it does not get a penalty term, and this makes it more prone to overfitting and less flexible to handle data with high dimensionality or multicollinearity. Furthermore, the value of the standard error is the highest of all models, even though it is considerably low (0.07363). The Ridge model shows an improvement in performance compared to the OLS model, with a slightly better RMSE and R^2 as well as standard error. The difference in wages drops to 14.4654%. In addition, the penalty term of 0.0245 indicates that the magnitude of the model coefficients has been restricted to control for complexity and improve generalization. The Lasso model also performs similarly to the Ridge model, with almost identical RMSE, standard error and R^2. The value of the difference lies between the OLS and the Ridge, with a value of 15.16023%. However, the penalty term of 0.0000246 indicates that the Lasso model may have applied a stronger penalty to the coefficients, suggesting a greater reduction in the magnitude of the coefficients and possibly a greater ability to select features. The Elastic Net model shows similar performance to the Ridge and Lasso models, although the greatest similarity is with Lasso, as the adjusted gender pay gap is -15.15952%. The RMSE, standard error and R^2 are also identical to those obtained in Lasso. The penalty term of 0.0000491 indicates that a combination of the penalties from the Ridge and Lasso models has been applied to control complexity and balance feature selection and reduction in the magnitude of the coefficients (Breiman, 2001). All these reasons argue that the three ML models can provide comparable results in terms of predictive accuracy and ability to explain wage variability.

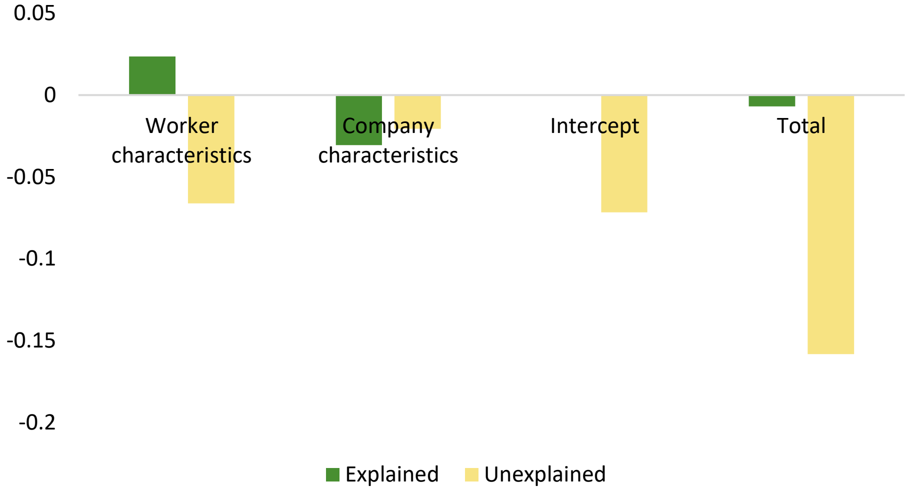
4.2 Decomposition

The decomposition has been performed on each of the models, the first to be tested being the **OLS model**. In this model we have obtained a value of wage inequality of a total of 16.32% in favor of men; that is, the latter earn 16.32% more per hour than women according to the specified model.

The explained part represents a small part of the total wage gap (Figure 6), with a coefficient value of 0.0069 (4.24% as a percentage of total inequality[11]). When analyzing the human capital variables, it is observed that they contribute negatively (-14.31%) to the wage gap. This indicates that, on average, women have a slightly higher level of human capital (education, work experience, skills, etc.) than men in the group studied, and even so, they obtain lower values in terms of wages per hour worked. On the other hand, company characteristics contribute positively (18.52%) to the explained wage difference. This could mean that men are employed in firms with characteristics that favor higher wages compared to women, either because of the sector in which they operate or because of the size of the firm.

[11] For values in % see Appendix Table A 5.

Figure 6. OLS model decomposition



Source: Own calculations based on Structure of Earnings Survey for Spain (2018)

The unexplained part of the wage difference between men and women is high, indicating that there is a significant proportion of the total difference that cannot be attributed to the human capital and firm characteristics variables included in the model. Importantly, the difference in the human capital component is notable (40.05%), while the difference in the firm characteristics' component is lower (12.43%). This difference in the results point to that those disparities in human capital between men and women may play a more significant role in the unexplained wage gap compared to firm characteristics. The unexplained human capital component (40.05%) indicates that the relevant model cannot explain the existing differences in salary through the chosen human capital variables. This may be because women and men do not have such different characteristics in this facet, yet they have a lower hourly salary.

On the other hand, the unexplained firm characteristics component (12.43%) suggests that differences in the specific characteristics of the firms in which men and women work have a minimal influence on the unexplained wage gap. This implies that, although firm characteristics may have some effect, their impact on the wage gap is generally quite low compared to other variables.

The intercept of the unexplained component with a percentage of 43.31% indicates the presence of an unexplained portion of the gender wage gap that cannot be attributed to either human capital or the firm characteristics included in the analysis. This intercept represents the portion of the wage gap that cannot be explained by the variables considered in the model. It may be due to a combination of unobserved factors, such as bias and discrimination in the employment market, gender stereotypes, structural or cultural barriers, or any other factor contributing to the gender wage gap.

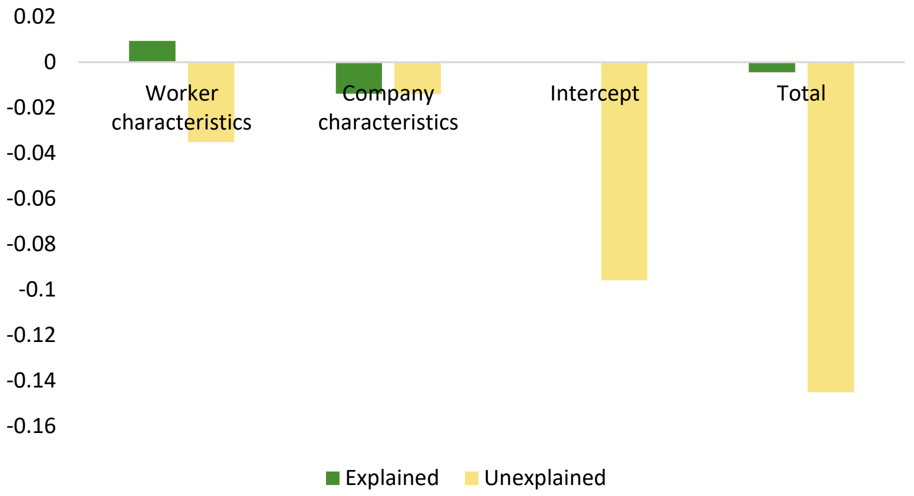
As for the **Ridge Model**, the explained component represents 2.98% of the inequality (Appendix) or in other words has a coefficient of -0.004459[12] (Table A 4). Within this component, worker characteristics contribute negatively (-6.29%), suggesting that, on average, women have a slightly higher level of human capital (as in the previous model) than men in the group studied, but still obtain lower values in terms of wages per hour worked. On the other hand, firm characteristics contribute positively (9.27%) to the explained wage gap.

The unexplained component of the wage gap accounts for 97.02% of the total inequality. Importantly, the difference in the unexplained human capital component is notable (23.46%), while the difference in the firm characteristics' component is ow compared to the latter (9.41%).

[12] Note that when performing the relevant decomposition calculations, the total inequality may vary slightly from the gap calculated above.

This difference in the results suggests that disparities in human capital between men and women may play a more important role in the unexplained wage gap compared to firm characteristics. Finally, the coefficient of the intercept increases relative to that obtained in the OLS model (from -0.0715 to -0.09594), which may indicate that the Ridge model provides even greater evidence of inequality and discrimination against women, although such a statement would require further analysis.

Figure 7. Ridge model decomposition



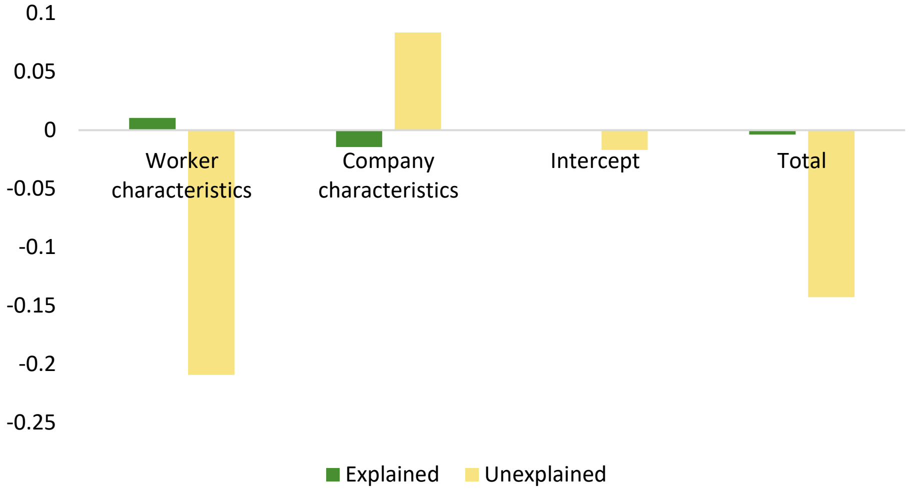
Source: Own calculations based on Structure of Earnings Survey for Spain (2018)

Unlike the OLS model, ML models perform interactions between the different covariates included in the model, so that all of them are related to all of them. These interactions capture the complex and nonlinear relationships that may exist between different worker and firm characteristics. These interactions have been included in the two groups, depending on whether the base variable of the interaction belonged to one group or the other.

In the analysis of the **Lasso Model** in Figure 8, significant changes are observed compared to the Ridge model in terms of the decomposition of the wage gap. The explained component of the Lasso model accounts for 2.57% of the total inequality. Within this component, worker characteristics contribute negatively to a greater extent than in the Ridge model (-7.18%), firm characteristics, however, contribute positively (9.75%) to the explained component.

On the other hand, the unexplained component of the wage gap in the Lasso model accounts for 97.43% of the total inequality. This is where the most important changes are observed. The intercept of the unexplained component has a positive value (11.53%), proposing the presence of a part of the wage gap that cannot be explained by the variables included in the Lasso model. This change in the sign of the intercept indicates that there is a portion of the wage inequality that is attributed to factors not observed or not considered in the analysis. The unexplained component of the workers' characteristics is very high (142.84%). This indicates that there is a significant portion of the wage gap that cannot be attributed to the workers' characteristics considered in the model.

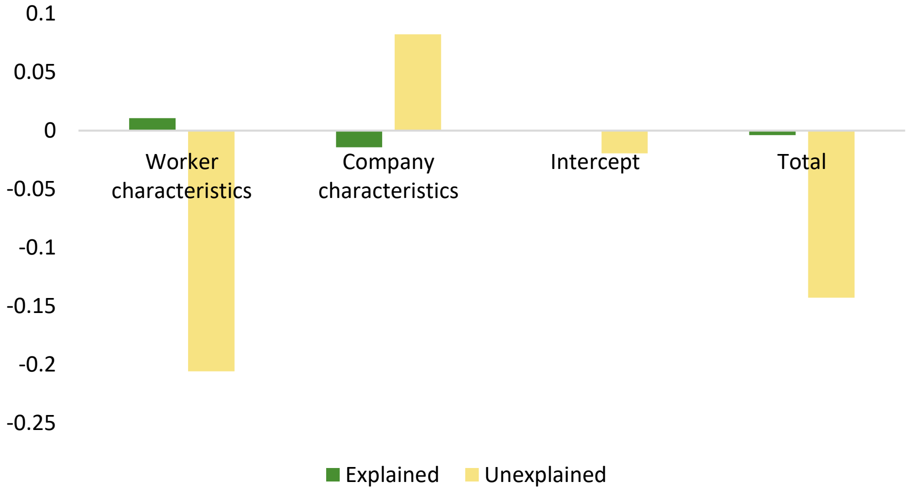
Figure 8. Lasso model decomposition



Source: Own calculations based on Structure of Earnings Survey for Spain (2018)

The **Elastic Net Model** presents very similar results to the Lasso model. The explained component of the Elastic Net model is 2.57%, indicating that this part of the wage inequality can be explained by the variables included in the model. As in the Lasso model, worker characteristics have a more pronounced negative contribution (-7.20%), suggesting that women possess a slightly higher level of human capital compared to men in the group studied. Firm characteristics also contribute positively with a similar value than in the Lasso model (9.77%) in the explained component, indicating that men are employed in firms with characteristics that favor higher wages compared to women.

Figure 9. Elastic Net model decomposition



Source: Own calculations based on Structure of Earnings Survey for Spain (2018)

As for the unexplained component of the wage gap, it accounts for 97.43% of the total inequality, which is consistent with the Lasso model. Similar changes are observed in comparison with the Ridge model and the Lasso model. The intercept of the unexplained component is positive (13.28%), indicating the presence of factors not observed or not considered in the model that contribute to the wage gap. The unexplained component of workers' characteristics is very high (140.37%), insinuating the presence of unobserved factors related to human capital that influence the wage gap. On the other hand, the unexplained component of firm characteristics is something to bear in mind, with a value of -56.22%, similar to the Lasso model.

After having explored and described the results obtained by all the models presented, it is easy to notice the lack of explanation offered by the data to somehow justify this sizable gender pay gap. Across the four models, we have consistently observed that the unexplained part of the wage gap is significantly higher than the explained part. This raises questions about the underlying factors generating this wage inequality and the difficulty in identifying and addressing them. One of the possible explanations for this low explained part is the limitation of the variables used in the models. Even if variables related to human capital and firm characteristics are included, it is possible that not all relevant differences between men and women in the workplace are fully captured. There are several dimensions of human capital and firm characteristics that can influence wages, such as specific work experience, networks, or organizational structure. If these important variables are omitted or not adequately captured, the model's ability to explain the wage gap is compromised.

In addition, models may fail to consider direct discrimination factors or hidden biases in pay decisions. Despite advances in gender equality, unconscious biases, gender stereotypes and discrimination still persist in the workplace. These factors may influence hiring, promotion and compensation decisions, but are not reflected in the model variables. Therefore, the unexplained portion of the wage gap may reflect precisely these forms of discrimination and bias that are not being addressed in the analysis. It is also important to consider the differences in employment opportunities between men and women. There are structural and social barriers that may limit women's access to certain higher-paying jobs or industries. These differences in opportunities may contribute to the unexplained part of the wage gap. If the model does not account for these differences in job opportunities, the observed wage inequality may not be adequately explained. Finally, cultural and normative factors play a key role in generating and perpetuating the gender wage gap. Entrenched societal gender roles and cultural expectations can influence occupational choices, wage decisions, and career development opportunities. These factors can be difficult to measure and capture in models, leading to a low explained share.

In conclusion, the high proportion of the unexplained part of the wage gap in the models studied indicates that there are complex and diverse factors that contribute to this gender wage inequality. The limitations of the variables used, the presence of direct discrimination or hidden biases, differences in job opportunities, and cultural and normative factors are just some of the possible explanations for this unexplained wage gap. Addressing this problem requires a comprehensive and multifaceted approach that considers the interaction of multiple factors and promotes changes at both the structural and cultural levels in society and in the workplace.

Conclusions and discussion

This thesis has addressed a topic of great relevance and interest in contemporary society: the gender wage gap in Spain. Throughout this study, an exhaustive analysis of the gender pay gap in the Spanish employment context has been carried out. Both traditional methods, such as the OLS model, and ML techniques have been used to guarantee the veracity and accuracy of the data obtained. The incorporation of ML techniques has considerably enriched our analysis, allowing us to obtain a deeper and more accurate understanding of the determinants of the gender wage gap in the Spanish context. Our main objective has been to deepen our knowledge of the causes and dimensions of the gender pay gap specifically in the Spanish context. Likewise, we have sought to identify whether the use of ML techniques could yield more favorable and revealing results compared to conventional methods previously applied in other studies. On the other hand, the use of the decomposition of the wage gap components has revealed both those variables that can partially explain this disparity, as well as those factors that remain unexplained, suggesting possible areas of discrimination or inequality not yet detected.

These approaches have opened up new possibilities for the study of this complex social issue and have provided valuable results that could guide future research and public policies aimed at addressing the gender wage gap more effectively and fairly in Spain.

The results obtained are robust and consistent across all models. It has been shown that the gender pay gap exists, with an average of approximately 15% in each of the cases. This means that, on average, women earn around 15% less than men in the data sample used for this study. These results are worrying and reflect a wage inequality that needs to be addressed at the societal and political level. An essential part of the analysis consisted of decomposing the gender wage gap into two components: what is explained and what is not explained by the available covariates. The results show that only a small fraction of the wage gap can be explained by the variables present in the database, while the vast majority (about 95% in all models) remains unexplained.

The fact that only a small percentage of the gender wage gap can be explained by the available covariates may be due to several reasons. First, it may suggest that important variables are missing from the analysis, such as cultural factors, gender bias in wage decision making or the influence of traditional gender roles on the distribution of job opportunities. On the other hand, it is also possible that there are unobservable or difficult to measure factors that contribute significantly to the gender pay gap. These could include direct or indirect discrimination, unconscious biases and other forms of gender inequality embedded in social and work structures.

Based on the methodology employed, the use of ML methods to analyze the gender pay gap has proven to be effective and accurate compared to traditional approaches. Linear regression models, such as OLS, while useful, may have limitations in capturing the complexity of the relationships between variables and the nonlinearity present in the data. In contrast, ML models, such as Ridge, Lasso and Elastic Net, have demonstrated better performance at all levels of error.

First, these ML models achieve greater efficiency by reducing overfitting and improving model generalization. In addition, ML models also offer higher accuracy in estimating the coefficients of the predictor variables. As for the comparison between the three ML models used (Ridge, Lasso and Elastic Net), it is observed that Lasso shows superior performance. The lower values of the root mean square error (RMSE) and the higher values of R^2 indicate that the Lasso model fits the data better and provides greater explanatory power for the wage gap between men and women in the data sample examined. It is different in the case of the standard error, where we find the lowest value for the Ridge model, although it is not very different from the others.

However, during the development of the study we encountered certain limitations that could have influenced our results and conclusions. One of them lies in the lack of sufficient variables that play a crucial role in wage determination. Despite using decomposition methods to identify explained and unexplained contributions to the wage gap, the absence of some relevant covariates could have affected the accuracy of our conclusions. In addition, data availability and quality also posed a challenge. The use of ML techniques, while enriching our analysis, also comes with its own limitations. Some ML models may be subject to biases and difficulties in interpreting the results, which could have affected the accuracy of our conclusions. Finally, it is important to keep in mind that the gender wage gap and its determinants may vary over time and depending on the evolution of labor and social policies. Since our research was based on data from a specific period, the results may not reflect future or past situations.

In order to address the gender pay gap comprehensively, further research represents a vital opportunity to delve deeper into this pervasive issue. The ultimate goal of this research may be to conduct a comparative analysis between different European countries or to undertake a more comprehensive examination of the factors contributing to the significant gender pay gap. Both avenues offer valuable insights that can inform policies and interventions to reduce the gap and promote gender equality in the workplace.

One possible direction for future research is to conduct a comparative analysis of the gender pay gap in several European countries. The availability of a standardized dataset, with information collected in a similar way in all European Union (EU) Member States, makes this approach particularly suitable. By taking advantage of this dataset, researchers can better understand how the gender pay gap varies from country to country and study the influence of different socioeconomic, cultural and political contexts. This type of research can be instrumental in identifying best practices and successful policies implemented in specific countries that have effectively reduced the significant gender pay gap. Comparative analysis can bring out the impact of particular legal frameworks, business practices and social attitudes on wage disparities. In addition, it can provide valuable information on the challenges faced by different countries in addressing the gender pay gap and help develop specific strategies to close the gap.

Another avenue of research is to conduct a more in-depth analysis of the gender wage gap in order to uncover the underlying factors contributing to wage disparities. While the existing dataset has provided valuable initial insights, a more granular approach is needed to comprehensively understand the significant gender pay gap. To achieve this, a multidimensional approach can be taken that incorporates a variety of factors known to influence the wage gap. For example, information on work experience, tasks conducted at the job, training and other relevant variables, perhaps related to IQ and/or personality, can be collected through tailored questionnaires. Such questionnaires would provide a wealth of detailed data on individual characteristics, making it possible to pinpoint the key drivers of the significant gender pay gap.

In conclusion, this study has shown that the gender pay gap is an undeniable reality in the data sample studied. The results of the regression models confirm the existence of this gender wage gap, and although only a small part can be explained by the available variables, most of it remains unexplained. These results highlight the need to address gender pay inequalities from multiple perspectives, including equal pay policies, promoting diversity and gender equity in the workplace, as well as promoting cultural change towards a more egalitarian and fair society for all. Only through concerted efforts can we achieve a society in which men and women are treated equally in terms of pay and job opportunities.

Bibliography

Altman, D., & Bland, M. (2005). Standard deviations and standard errors. *BJM*, 903.

Amuedo-Dorantes, C., & De la Rica, S. (2006). The Role of Segregation and Pay Structure on the Gende Wage Gap. Evidence from Matched Employer-Employee Data for Spain. *The B.E. Journal of Economic Analysis & Policy, 5 (1)*, 1-34.

Arulampalam, W., Booth, A., & Bryan, M. (2007). Is There a Glass Ceiling over Europe? Exploring the genderpay gap across the wages distribution. *Industrial* , 163-186.

Blau, F., & Kahn, L. (2017). The gender wage gap: Extent, trends, and explanations. *Journal Economics, 55 (3)*, pp. 789-865.

Blinder, A. (1973). Wage Discrimination: Reduced Form and Structural Estimates. *Journal of Human Resouces*, 8 (4), 436-455.

Boehmke, B., & Greenwell, B. M. (2020). *Hands-on Machine Learning with R*. Retrieved from GitHub Pages: https://bradleyboehmke.github.io/HOML/

Bonoccolto-Topfer, M., & Briel, S. (2022). The gender pay gap revisited: Does machine learning offer new insights? *Labour Economics, 78*.

Breiman, L. e. (2001). Statistical Modeling: The Two Cultures (with Comments and a Rejoinder by the Author). *Statistical Science, 16 (3)*, pp. 199-231.

Burawoy, M. (2005). For Public Sociology. *American Sociological Review, 70 (1)*, 4-28.

Chai, T., & Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)? Arguments against avoiding RMSE in the literature. *Geoscientific Model Development, 7*, pp. 1247-1250.

Cimentada, J. (2020, July 08). *Machine Learning for Social Scientists*. Retrieved from GitHub Pages: https://cimentadaj.github.io/ml_socsci/

Ciminelli, G., Schwellnus, C., & Stadler, B. (2021). Sticky floors or glass ceilings? The role of human capital, working time flexibility and discrimantion in the fender wage gap. *OECD Economics Department Working Papers*, 1668.

ClockBackward. (2009). *Ordinary Least Squares Linear Regression: Flaws, Problems and Pitfalls.* Retrieved from ClockBackward Essays.

de la Rica, S. D. (2010). Performance Pay and the Gender Wage gap: Evidence from Spain. *IZA*, No. 5032.

de la Rica, S. D., & Llorens, V. (2008). Ceilings or floors? Gender wage gaps by education in Spain. *Journal of Popular Economics*, 751-776.

de la Rica, S., Gorjón, L., & Vega-Bayo, A. (2019). *Brechas de Género en el Mercado en el mercado laboral de euskadi (2019).* ISEAK.

Dessler, G. (2009). *Administracion de recursos humanos* (Octava ed.). Ciudad de Mexico: Pearson.

European Comission. (2014). *How to combat the wage gap between men and women.* Luxembourg: European Union.

Eurostat . (2008). *NACE Rev. 2. Statistical classification of economic activities in the European Community.* Luxembourg: Eurostat Methodologies and Working papers.

Eurostat. (2023, June 10). *Eurostat*. Retrieved from ec.europa.eu: https://ec.europa.eu/eurostat/cache/metadata/en/earn_ses2018_esms.htm

Gardeazabal, J., & Ugidos, A. (2004). More on identification in detailed wage decompositions. *The Review of Economics and Statistics, 86 (4)*, pp. 1034-1036.

Gneezy, U., Niederle, M., & Rustichini, A. (2003). Performance in Competitive environments: Gender Differences. *The Quarterly Journal of Economics, 118 (3)*, 1049-1074.

Goldin, C. (2014). A grand gender convergence: Its last chapter. *American Economic Review, 104 (4)*, pp. 1091-1119.

González, M. J., Cortina, C., & Rodríguez, J. (2019). The Role of Gender Stereotypes in Hiring: A field Experiment. *European Sociological Review, 35 (2)*, 187-204.

Grimshaw, D., & Rubery, J. (2002). *The adjusted gender pay gap: a critical appraisal of standdard decomposition techniques.* Manchester: Manchester School of Management.

Hahn, G. J. (1973). The coefficient of determination exposed! *Chemical Technology, 3*, 609.

Hlavac, M. (2022). *oaxaca: Blinder-Oaxaca Decomposition in R.* Bratislava, Slovakia: Social Policy Institute.

Homolka, D. (2022). *Closing the Gender Pay Gap: Can Machine Learning Help ?* Nova School of Business and Economics.

James et al. (2013). *An Introduction to Statistical Learning* (Vol. 112). Springer.

Jann, B. (2008). The Blinder-Oaxaca Decomposition for Linear Regression Models. *Stata Journal*, 8 (4), 453-479.

Kleven, H., Landais, C., & Søgaard, J. (2018). Children and Gender Inequality: Evidence from Denmark. *National Bureau of Economics Research*.

Leythienne, D., & Ronkowski, P. (2018). *A decomposition of the unadjusted gender pay gap using structure of earnings survey data.* Eurostat.

Ministerio de trabajo y economia social del gobierno de españa. (2022). *La situacion de las mujeres en el mercado de trabajo 2022.*

Neumark, D. (1988). Employers' discriminatory behavior and the estimation of wage discrimination. *Journal of Human Resorces, 23 (3)*, pp. 279-295.

Oaxaca, R. (1973). Male-Female Wage Differentials in Urban Labor Markets. *International Economic Review*, 14 (3), 693-709.

Oaxaca, R., & Ransom, M. (1994). On discrimination and the decomposition of wage diferentials. *Journal of Economics, 61 (1)*, pp. 5-21.

Romer, P. (1990). Endogenous Technological Change. *Journal of Political Economy*, 98(5), 71-102.

Strittmatter, A., & Wunsch, C. (2021). The Gender Pay Gap revisited with Big Data: Do Methodological Choices Matter? *IZA*.

Watts, D. J. (2014). Common Sense and Sociological Explanations. *American Journal of Sociology, 120 (2),* pp. 313-51.

Yarkoni, T., & Westfall, J. (2017). Choosing Prediction over Explanation in Psychology: Lessons from Machine Learning. *Perspectives on Psychological Science, 12 (6),* pp. 1100-1122.

Appendix

Categorical variable list and definitions

Age:

- From 14 to 19.
- From 20-29.
- From 30-39.
- From 40-49.
- From 50-59.
- 60 or more.

Education Level:

- Group 1: Includes less than primary, primary and lower secondary.
- Group 2: Includes upper secondary and post-secondary (non-tertiary).
- Group 3[13]: Includes short-cycle tertiary, bachelor, master, doctoral or the equivalent to each of them.

Occupation Skill Level[14]:

- High skilled: For managers, professionals and technicians and associate professionals from the International Standard Classification of Occupations (ISCO-08).
- Medium skilled: For clerical support workers, service and sales workers and skilled agricultural, forestry and fishery workers.
- Low skilled: For craft and related trades workers, plant and machine operator and assemblers and elementary occupations.

Type of contract:

- Full-time
- Part-time

Firm size:

- Small firms: hiring up to 49 employees
- Midsize firms: between 50 and 250 employees
- Large firms: 250 employees or more

Collective Pay Agreement:

- Have a Collective Pay Agreement. This may include: an agreement at national or interconfederal agreement, an industry agreement, enterprise agreement, agreement applying only to workers in the local unit, agreement for individual industries in individual regions or any other type of agreement.
- Do not have a Collective Pay Agreement.

[13] Initially, group 3 was segregated into two different groups, one in which the years of tertiary education did not exceed 4 years and the other in which it did. In order to simplify it, these two groups have been merged.

[14] For the sake of simplicity, the group 0 called *Armed Forces Occupations* has been eliminated as the sample belonging to that group was considerably small.

Economic and Financial Control:

- Public control
- Private control

NACE code:

- Mining and quarrying.
- Manufacturing.
- Electricity, gas, steam and air conditioning supply and water collection, treatment and supply.
- Wholesale and retail trade.
- Hotels and Restaurants.
- Transporting and Storage.
- Financial and insurance activities
- Real Estate activities.
- Other business activities.
- Public Administration and Defense.
- Education.
- Health and social work.
- Other services activities.
- Arts, entertainment and recreation.

Table A 1. Regression model, explained and explanatory variables

	Variable name	*Categories*[15]
Dependent variables	Total Salary Per Hour/Month Base Salary Per Hour/Month Complementary Salary Per Hour/Month	N/A
Personal explanatory variables	Gender	Female **Male**
	Age[16]	**29 or below** 30-39 40-49 50 or above
	Education Level	**Primary and lower secondary** Upper secondary and post-secondary All tertiary education
	Occupation Level	**Low skilled** Medium skilled High skilled

[15] For more information about category specifications, see **Categorical variable list and definitions** in Appendix. The reference categories are emboldened.

[16] In contrast to the descriptive analysis, the age groups have been simplified here: the 14-19 age group has been merged with the 20-29 age group and the 60+ age group has been merged with the 50-59 age group.

	Contract Type	Full-time
		Part-time
Company explanatory variables	Company size	**Small (49 or below)**
		Medium (50-249)
		Large (250 or above)
	Collective Pay Agreement	**Yes**
		No
	Economic/Financial control	**Public control**
		Private control
	NACE code	**Manufacturing**
		For the rest see **Categorical variable list and definitions** in Appendix

Source: Own calculations based on Structure of Earnings Survey for Spain (2018)

Table A 2. NACE code related information

	N	%	Female	Male
Mining and quarrying	1,707	0.78%	7.81%	92.19%
Manufacturing	54,613	24.94%	28.18%	71.82%
Electricity, gas, steam and air conditioning supply	8,031	3.67%	17.41%	82.59%
Construction	12,557	5.73%	11.73%	88.27%
Wholesale and retail trade	17,956	8.20%	49.17%	50.83%
Hotels and Restaurants	8,361	3.82%	55.86%	44.14%
Transporting and Storage	20,126	9.19%	34.94%	65.06%
Financial and insurance activities	14,675	6.70%	47.71%	52.29%
Real Estate activities	10,242	4.68%	51.59%	48.41%
Other business activities	19,850	9.07%	55.74%	44.26%
Public Administration and Defense	10,051	4.59%	51.38%	48.62%
Education	7,962	3.64%	60.79%	39.21%
Health and social work	20,377	9.31%	73.12%	26.88%
Other services activities	2,241	1.02%	65.15%	34.85%
Arts, entertainment and recreation	10,217	4.67%	42.75%	57.25%
	218,966	100.00%	43.56%	56.44%

Source: Own calculations based on Structure of Earnings Survey for Spain (2018)

Table A 3. OLS regression results

		Coefficients	P-value
	(Intercept)	1.837111	< 2e-16 ***
Female	Female	-0.16323	< 2e-16 ***
Age	30-39	0.191376	< 2e-16 ***
	40-49	0.314221	< 2e-16 ***
	50 and more	0.421691	< 2e-16 ***
Educational level	Secondary Education	0.117793	< 2e-16 ***

	Tertiary education	0.326532	< 2e-16 ***
Company size	Medium size	0.137323	< 2e-16 ***
	Big size	0.189271	< 2e-16 ***
Occupational Skill Level	Medium skilled	0.01298	1.36e-07 ***
	High skilled	0.320246	< 2e-16 ***
Contract type	Full time contract	0.077496	< 2e-16 ***
Does the company have a collective agreement?	Yes	0.041102	< 2e-16 ***
Economic/Financial Control	Public control	0.159235	< 2e-16 ***
NACE code	Mining and quarrying	0.137433	< 2e-16 ***
	Electricity, gas, steam and air conditioning supply	0.030829	2.42e-10 ***
	Construction	-0.06034	< 2e-16 ***
	Wholesale and retail trade	-0.0674	< 2e-16 ***
	Hotels and Restaurants	-0.09357	< 2e-16 ***
	Transport and Storage	-0.06327	< 2e-16 ***
	Financial and insurance activities	0.052839	< 2e-16 ***
	Real Estate activities	-0.21289	< 2e-16 ***
	Other business activities	-0.20715	< 2e-16 ***
	Public Administration and Defense	-0.16713	< 2e-16 ***
	Education	-0.20919	< 2e-16 ***
	Health and social work	-0.20564	< 2e-16 ***
	Other services activities	-0.21357	< 2e-16 ***
	Arts, entertainment and recreation	-0.13405	< 2e-16 ***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4027 on 218938 degrees of freedom

Multiple R-squared: 0.4232, Adjusted R-squared: 0.4232

F-statistic: 5950 on 27 and 218938 DF, p-value: < 2.2e-16

Source: Own calculations based on Structure of Earnings Survey for Spain (2018)

Table A 4. Decomposition results in coefficients for all models

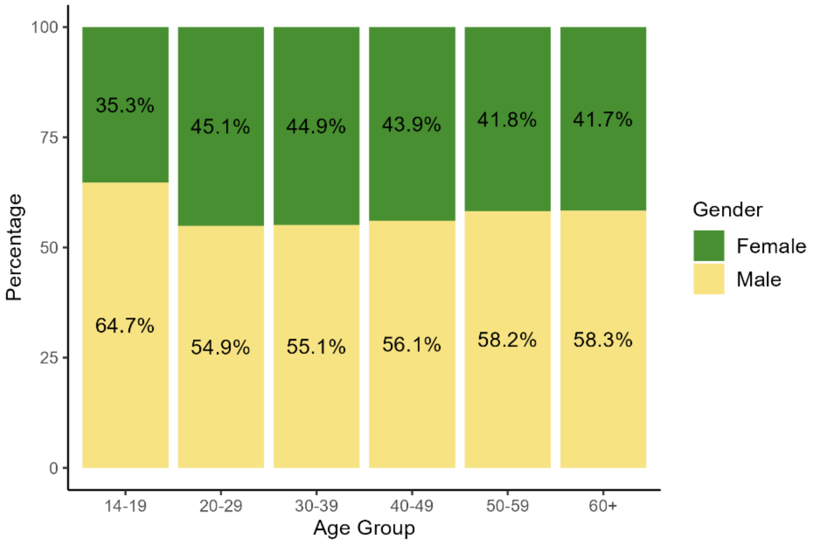
	OLS		RIDGE		LASSO		ELASTIC NET	
	Explained	Unexplained	Explained	Unexplained	Explained	Unexplained	Explained	Unexplained
Worker characteristics	0.02362812	-0.06614657	0.00940944	-0.03507768	0.01051867	-0.20925858	0.01055351	-0.20565474
Company characteristics	-0.03058495	-0.02053855	-0.01386903	-0.01406829	-0.01427747	0.08341124	-0.01431949	0.08237049
Intercept		-0.07152622		-0.09594604		-0.01689741		-0.01945675
Total	-0.00695683	-0.1582113	-0.00445959	-0.145092	-0.0037588	-0.1427447	-0.00376598	-0.142741
Total difference (Female - Male)	-0.165168128		-0.149551588		-0.146503503		-0.146506975	

Table A 5. Decomposition results in % for all models

	OLS		RIDGE		LASSO		ELASTIC NET	
	Explained	Unexplained	Explained	Unexplained	Explained	Unexplained	Explained	Unexplained
Worker characteristics	-14.31%	40.05%	-6.29%	23.46%	-7.18%	142.84%	-7.20%	140.37%
Company characteristics	18.52%	12.43%	9.27%	9.41%	9.75%	-56.93%	9.77%	-56.22%
Intercept	0.00%	43.31%	0.00%	64.16%	0.00%	11.53%	0.00%	13.28%
Total	4.21%	95.79%	2.98%	97.02%	2.57%	97.43%	2.57%	97.43%
Total difference (Female - Male)	-16.51681%		-14.95516%		-14.65035%		-14.65070%	

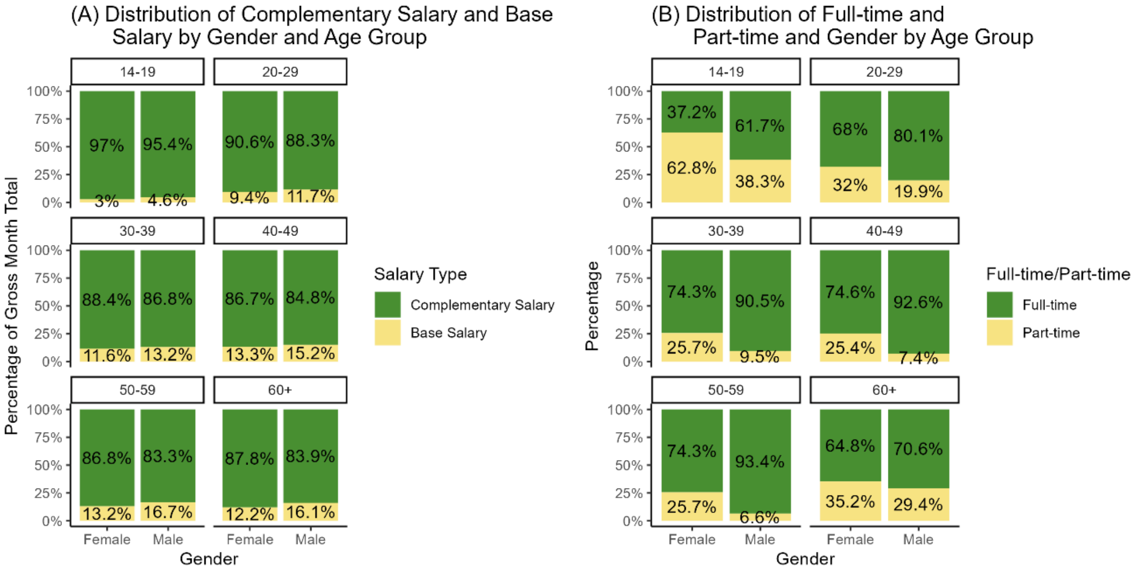
Source: 1. Own calculations based on Structure of Earnings Survey for Spain (2018)

Figure A 1.Distribution of Age and Gender



Source: Own calculations based on Structure of Earnings Survey for Spain (2018)

Figure A 2. Distribution of Salary and Contract type by Age groups


(A) Distribution of Complementary Salary and Base Salary by Gender and Age Group

(B) Distribution of Full-time and Part-time and Gender by Age Group

Source: Own calculations based on Structure of Earnings Survey for Spain (2018)