VILNIUS UNIVERSITY

FACULTY OF MATHEMATICS AND INFORMATICS

DATA SCIENCE STUDY PROGRAMME

Master's thesis

The Analysis of dynamics of women's and men's wages

Moterų ir vyrų darbo užmokesčio dinamikos analizė

Amanda Vilkončiūtė

Supervisor : Prof. Habil. Dr. Vydas Čekanavičius

Vilnius
2026

Summary

The gender pay gap remains a persistent issue in the modern labor markets, including Lithuania. The main purpose of this study is to analyse the gender pay gap in Lithuania, based on the most recent statistical wage survey data and to apply regression and predictive machine learning methods. The analysis consists of three main parts. The first part, which was dedicated to understanding the gender pay gap, revealed that the multivariate Regression explains about 48.6 percent of wage variation, meaning that more than half of the variation remains unexplained, while Oaxaca–Blinder Decomposition shows that only 19.1 percent of the gender wage gap can be explained by characteristics. In addition, Quantile Regression confirms that gender pay gap exists across the entire wage distribution and increases at higher wage levels. In the second part we predicted induvial hourly wages, using machine learning models. Among the evaluated methods, XGBoost achieved the best predictive performance, with a coefficient of determination of $R^2 = 0.97$. The last part of analysis was dedicated to predicting average occupational wages, for women and men separately, based on occupational structural characteristics. Best performing model for men was Random Forest with $R^2 = 0.93$, while for women was XGBoost $R^2 = 0.89$. By integrating causal regression with modern machine learning models, this study demonstrates the complementary value of these approaches in analyzing and predicting gender wage inequality.

Keywords: Multivariate Regression, Oaxaca–Blinder Decomposition, Quantile Regression, Random Forest, XGBoost, Support Vector Regression (SVR), Gender pay gap.

Santrauka

Lyčių darbo užmokesčio skirtumas išlieka nuolatine problema šiuolaikinėse darbo rinkose, įskaitant ir Lietuvą. Pagrindinis šio tyrimo tikslas - remiantis naujausiais statistiniais darbo užmokesčio tyrimų duomenimis, išanalizuoti lyčių darbo užmokesčio skirtumą Lietuvoje ir pritaikyti regresinius bei prognozuojamuosius mašininio mokymosi metodus. Analizę sudaro trys pagrindinės dalys. Pirmoji dalis, skirta lyčių darbo užmokesčio skirtumui suprasti. Rezultatai atskleidė, kad daugiamatė regresija paaiškina apie $48{,}6$ proc. darbo užmokesčio svyravimų, o tai reiškia, kad daugiau nei pusė svyravimų lieka nepaaiškinta, o Oachakos-Blinderio dekompozicija rodo, kad tik $19{,}1$ proc. lyčių darbo užmokesčio skirtumo galima paaiškinti charakteristikomis. Be to, kvantilinė regresija patvirtina, kad lyčių darbo užmokesčio skirtumas egzistuoja visame darbo užmokesčio pasiskirstyme ir didėja esant aukštesniems darbo užmokesčio lygiams. Antroje dalyje, kurioje naudojome mašininio mokymosi modelius, prognozavome individualius valandinius darbo užmokesčius. Iš vertintų metodų geriausią prognozavimo našumą pasiekė „XGBoost“, kurio determinacijos koeficientas buvo $R^2 = 0{,}97$. Paskutinė analizės dalis buvo skirta prognozuoti vidutinį darbo užmokestį moterims ir vyrams atskirai, profesijų lygyje, remiantis profesinėmis struktūrinėmis charakteristikomis. Geriausiai veikiantis modelis, prognozuojant vidutinius atlygimus vyrams buvo „Random Forest“ modelis, kurio $R^2 = 0{,}93$, o moterims – „XGBoost“ $R^2 = 0{,}89$. Integruodamas priežastinę regresiją su šiuolaikiniais mašininio mokymosi modeliais, šis tyrimas parodo šių metodų papildomąją vertę analizuojant ir prognozuojant lyčių darbo užmokesčio nelygybę.

Raktiniai žodžiai: Daugiamatė regresija, Oachakos-Blinderio dekompozicija, Kvantilinė regresija, Atsitiktinių miškų analizė, XGBoost, Atramos vektoriaus regresija (SVR), Lyčių darbo užmokesčio skirtumas.

List of Figures

List of Tables

Contents

Introduction

The gender pay gap is the difference between the incomes of women and men, reflecting social, economic and political factors. Although this gap has decreased over the past decades, this problem remains relevant in the modern labor market in many countries around the world. In 2023, the average gross hourly wage of women in Europe was 12% lower than that of men [6]. The largest gap is recorded in Latvia – as much as 19%. Due to the large gender pay gap, international institutions such as UNSECO, the United Nations, Eurostat ([7],[21],[12]) aim to reduce this gender pay gap and ensure equal pay for work of equal value in all countries. When monitoring progress towards Sustainable Development Goal 5, the European Union has made significant progress in most areas over the past five years, but the pay gap is still noticeable in almost all European Union and other countries. This problem is also relevant in Lithuania.

Although the Lithuanian labor market is characterized by a high number of working women and women acquire higher education more often than men, in the same fields of work, women earn as much as 11.5% less than men, according to statistics provided by Eurostat in 2023 [6]. The gap is recorded not only on a national scale, but also in specific professions. Although the gender pay gap is widely analyzed in international studies, this problem has not been studied in sufficient detail in Lithuania. Some previous studies have found econometric models, aggregated indicators, but they cannot understand the importance of actions. To not only to better understand the distribution of women's and men's salaries, but also to analyze which variables have the greatest impact on the salary gap, predict individual hourly wage and evaluate how many professions' average salaries can be predicted based on their structural characteristics for women and men separately, regression and machine learning models are used.

In this study, a Multivariate Regression, Quantile Regression and Oaxaca–Blinder Decomposition will be applied to describe what part of the salary is explained by personal and company characteristics of employees, including categorical and non-categorical variables. Machinie learning models, such as Random Forest, XGBoost and Support Vector Regression are used to aim not only to evaluate the importance of variables, but also, to predict individual hourly wages and professions' average salaries for women and men separately. For this we will use information from the publicly available database of the Lithuanian State Data Agency - Statistical Survey of the Structure of Wage, which is stored on the salaries of Lithuanian employees in October 2014, 2018 and 2022. To achieve an objective analysis, the data will be narrowed down to the level of professions. Only representatives of those professions will be selected in which there are a sufficient number of women and men, and their distribution is similar. By examining the distribution of data and using the model, the aim is to examine in detail the gender wage gap in various sections, by profession, education, work experience and even the size of companies.

The main goal of this work is to analyze the dynamics of women's and men's wages in Lithuania, based on the most relevant statistical wage survey data and to apply regression and predictive machine learning methods. To achieve this goal, several main tasks are set - to perform data preparation, analyze salary distributions according to different variables and apply forecasting models, assessing the

importance of variables and changes in the period 2014-2022. The analysis follows a structured analytical framework composed of several main different, but complementary parts. The analysis begins with a descriptive examination of wage patterns , complemented by a two-way ANOVA to evaluate the relationships between gender and occupation. The explanatory stage employs Multivariate Regression, Oaxaca–Blinder Decomposition, and Quantile Regression to determine how much of the wage gap can be explained by observable employee and firm characteristics across the wage distribution. Finally, the study applies predictive machine learning models—Random Forest, XGBoost, and Support Vector Regression—to evaluate the accuracy of individual hourly wage predictions and to forecast male and female wages separately at the occupational level.

1 Literature review

Modern studies that analyze the wages of men and women and their differences mostly use traditional regression and machine learning models. Traditional models, such as multiple linear regression, remain the main tools that explain the wages of women and men and their differences, using hypothesis testing. However, modern studies are increasingly using machine learning models not only to clarify the relationships between variables or explain the wage gap between women and men, but also to improve the accuracy of predictions. Although all researchers agree that there is a significant difference between the wages of women and men in different countries of the world, they choose different methods, data processing and model applications, which leads to different approaches to the gap and the explainability of the wage gap. Some use models to explain the wages of women and men, others develop predictive models to predict wages or the gender wage gap.
For example, A. Aiftincăi (2025) [1] and P. V. Patwa (2024) [14] investigate how machine learning models can be applied to wage forecasting, but do not analyze the gender wage gap. A. Aiftincăi (2025) [1] used macroeconomic data collected in Romania for the period 1991-2024, including factors such as inflation and consumer price indices. The best-performing model was Random Forest with $RMSE : 97.60 LEI$, which when converted to euros is $RMSE : 19.18$. Using the two best-performing models, Random Forest and XGBoost, A. Aiftincăi (2025) [1] predicted the overall average wage for the years 2025-2027. Unlike A. Aiftincăi (2025) [1], P. V. Patwa (2024) [14] analyzes individual employee salaries and focuses on feature engineering using salary information collected by a single US organization between 2013 and 2020. Using several more complex machine learning models, P. V. Patwa (2024) [14] achieved better results. The best-performing model is the Gradient Boosting model with $R^2 = 0.9698$. Although both studies are not directly focused on gender inequality, they demonstrate the usefulness of machine learning in modeling wage dynamics.
Others are dedicating their work to understanding what determines salaries in the IT sector, focusing on the gender pay gap, and developing a predictive model. M. Brandwijk (2021) [4] using the 2020 European IT salary survey database and applying Multiple Linear Regression and Random Forest Regression models, identified the most important variables - seniority, experience and age, and the gender pay gap becomes more pronounced with greater seniority. Also, M. Brandwijk (2021) [4] not only observed the distributions of salaries for women and men, but also predicted them. Although XGBoost showed a very low error on the training data ($RMSE = 1.703$), its performance on the testing part deteriorated significantly ($RMSE = 28.834$) due to overfitting. Multiple Linear Regression was the best-performing model with $RMSE = 24.459$ on the testing set. A completely different use of the XGBoost model for predicting the salaries of women and men in the IT sector was presented by S. A. Waisbrot et al. (2023) [22]. In his work, he proposed a new gender pay gap decomposition method based on machine learning and applied it to data on IT specialists working in Argentina. S. A. Waisbrot et al. (2023) [22] used contractual gender inversion in his work to calculate the gender pay gap - the model simulates how much the same person could earn if they were of the other sex, and all other variables remained the same. This would be how S. A. Waisbrot et al. (2023) [22] determined how much of the salary gap is due to pure discrimination. The results

obtained showed that the difference between men's and women's salaries is 20%, of which 7.7% can be explained only by direct discrimination, and 12.3% - by other factors, such as the total number of years of work experience, level of education and number of employees.

Some studies, in addition to sector-specific analyses, use large-scale datasets to assess the gender pay gap. For example, A. Strittmatter et al. (2021) [17] assesses the gender pay gap and analyze how methodological decisions affect the size of the unexplained portion of salaries, using a large dataset of 1.7 million employees information about wages, collected in Switzerland. Results revealed that Blinder-Oaxaca estimates of the unexplained gender wage gap are reduced 39% when gender comparability and a more flexible specification of the wage equation are used. Using stricter comparability criteria, more flexible salary functions, evidence-based methods including machine learning-based variable selection, and unexplained gender pay gap is important for job observation, in some cases by as much as 50%.Meanwhile, J. E. Olson et al. (2023) [13], using national data from Chile collected in the period 1990–2017, move from predictive modeling to explanatory modeling, applying causal machine learning to assess the impact of gender on wages. Similarly, S. A. Waisbrot et al. (2023) [22], J. E. Olson et al. (2023) [13] use modern machine learning models to estimate how much salary is reduced by being a woman but conduct a study on workers in all sectors. Of all the models used, the best result was shown by Polynomial Support Vector Regression, which achieves the lowest error ($MAE = €11{,}697$) and one of the lowest RMSE values (€$24{,}720$) for annual salary. Similarly, M. Töpfer et al. (2023) [11] demonstrated that large microdata sets combined with machine learning methods can improve gender pay gap analysis in the Spain, where wage prediction models are used to identify the most important wage drivers and indirectly infer the size of the gender pay gap across different groups of workers. In this work, ML models were applied to predict individual salaries in a large data set, and the predicted results were used to compare observed wage differences between men and women. Results revealed that the gender pay gap is between 14% and 16%.

J. M. Zawia et al. (2025) [24] propose to predict the percentage wage gap between women and men using machine learning models instead of wages. The authors calculated wage gaps by calculating the median wage of women and the median wage of men. SHAP analysis revealed that occupation is the most important factor, with median wages of men and sectors contributing to the largest differences. Random Forest was the best-performing forecasting model, achieving even $R^2 = 0.994$ with $RMSE = 2.964$ for hourly wage. Meanwhile, C. Tealdi et al. [18], like M. Töpfer et al. (2023) [11], uses machine learning models to estimate the gender pay gap, rather than to directly predict it. C. Tealdi et al. (2023) [18] apply advanced Double Machine Learning and regression techniques, showing that methodological decisions and the combination of ML models can significantly change the estimated unexplained size of the gender pay gap, in some cases reducing it by about 50% compared to traditional linear regressions.

The literature also shows a methodologically more diverse range of studies that analyze gender pay inequality. K. Albæk et al. (2014) [2] and S. Trivedi et al. (2025) [20] use Quantile Regression. The results of K. Albæk et al. (2014) [2] show that the gender pay gap in Denmark ranges from 10% in the low quantiles to more than 20% in the high quantiles, thus revealing greater inequality among high-earners. Similar trends were found by S. Trivedi et al. (2025) [20], emphasizing that the gap varies

from 8% (low quantiles) to $25-30\%$ (high quantiles), with the largest discrepancies being caused by occupational segregation and educational differences. A. Webster et al (2020) [23] confirmed in their study that the wage gap in lower occupations is more than 17%, using the Propensity-scope matching method. In contrast to A. Webster et al (2020)) [23], D. M. Blei et al (2024) [3] using foundation-model type transformers estimated that the wage decreases by $10-15\%$ if the person is a woman, and in highly skilled occupations the percentage is even higher. Another approach was presented by M. Töpfer et al. (2022) [19]. Töpfer et al. (2022) [19] used post-double-LASSO and high-dimensional regression in their work. It turned out that including a larger set of variables reduces the unexplained part of the wage gap by about 3-5 percentage points compared to the traditional Oaxaca–Blinder Decomposition.

Thus, modern research is becoming increasingly methodologically diverse and based on advanced analytical methods. Traditional regression and decomposition models remain important for explaining the gender pay gap. These models are complemented by machine learning models, quantile analysis and neural networks, which allow not only to predicting salaries or their salary gap, but also to analyze the impact of gender on different subgroups of employees. Research results confirm that women receive lower salaries than men in various countries around the world. This gap is determined by the sector of work, experience, age and profession. Most researchers use reliable machine learning models, such as Random Forest, XGBoost or SVR in their work, but adapt them for different purposes. Some predict total salary, others separate and predict male and female salaries separately, and still others predict the percentage difference between male and female salaries.

2 Methodology

The main objective of this study is to analyze the wages of women and men in Lithuania and evaluate predictive machine learning models for predicting individual hourly wages and wages of women and men at the occupational level. The analysis follows a structured analytical framework composed of several main different, but complementary parts related to the analysis of the gender pay gap in Lithuania, in 2022. Each part is dedicated to a different research objective.

In the first part of analysis, preliminary data analysis was conducted to describe wage patterns and differences between men's and women's salaries across gender, professions, education and other features. Additionally, Two-way ANOVA was applied. Unlike standard one-way ANOVA, which compares means across a single factor, the two-way ANOVA considers two factors: gender and profession. This approach allows us to estimate possible interactions between factors.

Second part is dedicated to explanatory analysis and aims to determine which part of the salary can be explained by the observed individual characteristics of employees and employees' companies. To achieve this goal, the Multivariate Regression model and the Oaxaca–Blinder Decomposition were used, and in order to better understand how salaries are distributed, the Quantile Regression method was applied.

Then, the focus shifts from the explainability of salary to predictive machine learning models. We aim to evaluate how well the models are able to predict individual hourly wages based on employee and company characteristics and to predict the salaries of women and men separately, at the occupational level. To address this goal, predictive machine learning models were applied - Random Forest, XGBoost and Support Vector Regression.

2.1 Regression-based analysis of the gender pay gap

This part of the study was designed to estimate the parts of the wage gap that can be explained by employee or company characteristics and those that remain unexplained even after controlling for the same characteristics and can be attributed to gender discrimination within the same profession or labor market biases. To estimate these parts, a Multivariate Regression model and Oaxaca–Blinder Decomposition were used. To better understand the structure of the wage distribution, a third method was used, Quantile Regression. The latter was applied to analyze the gender wage gap at different income levels (10th, 25th, 50th, 75th and 90th quantiles).

2.1.1 Two-Way ANOVA

Two-way analysis of variance (ANOVA) is a statistical method used to study how two categorical independent variables effect one continuous dependent variable, Kim, H. Y. (2014) [10]. As a preliminary step, a two-way analysis of variance (two-way ANOVA) was conducted to examine whether average hourly wages differ systematically by gender and occupation, and whether the gender wage gap varies across occupational groups. In this study, gross hourly wage was treated as the dependent

variable and gender and occupation as categorical independent variables.
The statistical model is specified as:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}, \tag{1}$$

where Y_{ijk} denotes the hourly wage of individual k belonging to gender group i and occupation j, μ is the overall mean wage, α_i represents the gender effect, β_j represents the occupation effect, $(\alpha\beta)_{ij}$ captures the interaction between gender and occupation, and ε_{ijk} is the error term.
The following null hypotheses were tested:

- H_{0A}: There is no difference in mean hourly wages between genders.
- H_{0B}: There is no difference in mean hourly wages across professions.
- H_{0AB}: There is no interaction effect between gender and profession on hourly wages.

To ensure the validity of the two-way ANOVA, the key assumptions were evaluated. First, the independence of observations was ensured by the study design. Second, the normality of residuals was assessed using graphical methods, including residual histograms and normal Q–Q plots. Third, the assumption of homogeneity of variances across groups was examined using Levene's test.

2.1.2 Multivariate Regression

Multiple regression is a statistical analysis method used to predict the value of a dependent variable based on the values of two or more independent variables, Sheposh, R. (2025) [16]. It allows us to understand complex relationships and make predictions by analyzing how multiple factors interact to influence an outcome.
In this study Multivariate Regression was applied to estimate the conditional gender pay gap after controlling for observable individual and job-related characteristics. In this case, the dependent variable is the logarithm of gross hourly earnings. The logarithmic transformation allows the regression coefficients to be interpreted as approximate percentage differences and reduces the effects of heteroscedasticity. The main explanatory variable is gender (0 for women and 1 for men). The coefficient on this variable reflects the average conditional wage gap between men and women, holding all other characteristics constant. Age is included as a structured categorical variable, dividing individuals into five age categories (20–24, 25–34, 35–44, 45–54 and $55+$). Work experience is measured by length of service. Educational attainment, firm size and occupation are included as categorical variables to account for differences in human capital, firm characteristics and occupational wage structures. Employment intensity is captured by the part-time rate, distinguishing between full-time and part-time work.A bonus indicator is also included.
The regression model is defined as follows:

$$\ln(wage_i) = \beta_0 + \beta_1 Male_i + X_i'\gamma + \varepsilon_i, \tag{2}$$

where ln$(wage_i)$ denotes the logarithm of hourly wage, $Male_i$ is a gender indicator variable, X_i is a vector of control variables, γ is the corresponding vector of coefficients, and ε_i is the error term. In our case, the vector of control variables X_i includes individual and job-related characteristics as well as occupation fixed effects. Specifically, X_i consists of age group indicators, tenure, education level, firm size, occupation dummies, working time status (full-time versus part-time), and bonuses. Tenure and bonuses are treated as continuous (interval) variables, while age is introduced in grouped form as an ordered categorical variable. Gender, education, firm size, occupation, and working time status are modeled as categorical variables.

After applying the model, the model fit was evaluated using the coefficient of determination (R^2), the root mean squared error (RMSE) and the F-statistic.

To ensure the validity of the Multivariate Regression results, the key assumptions were systematically evaluated. First, the normality of the residuals was assessed. Given the large sample size, formal normality tests such as the Shapiro-Wilk test were not applied. Instead, the normality of the residuals was examined using graphical methods such as histogram plotting and Q-Q plots, and numerical measures of skewness and kurtosis. Second, heteroscedasticity was tested using the Breusch-Pagan test. If $p < 0.001$, then we reject the null hypothesis of homoscedasticity and assume that heteroscedasticity exists. Third, multicollinearity of the explanatory variables was examined using the variance inflation factor (VIF). A low VIF value ($VIF < 4$) indicates that the explanatory variables do not have a strong linear relationship. Finally, outliers were assessed using Cook's distance. The diagnosis allows the detection of potentially influential cases without automatically excluding observations from the analysis.

2.1.3 Quantile Regression

Quantile Regression is a statistical method that models the relationship between predictor variables and different quantiles of a response variable's conditional distribution. In this case, Quantile Regression is used to examine how the gender wage gap varies across the hourly wage distribution and allows us to estimate conditional wage differences at different points in the wage distribution. The set of variables are the same we used in the Multivariate Regression - the dependent variable is the logarithm of gross hourly wage. This set of variables includes age group, work experience, education level, firm size, occupation, part-time employment status and the presence of performance-related bonuses.

Formally, the Quantile Regression model is specified as:

$$Q_\tau(\ln(wage_i) \mid X_i) = \beta_{0,\tau} + \beta_{1,\tau} Male_i + X_i' \gamma_\tau, \tag{3}$$

where $Q_\tau(\ln(wage_i) \mid X_i)$ denotes the τ-th conditional quantile of the logarithm of hourly wage, $Male_i$ is a gender indicator variable, X_i is a vector of control variables, and β_τ and γ_τ are quantile-specific coefficients.

Quantile Regression models were estimated at the 10th, 25th, 50th, 75th and 90th quantiles of the wage distribution. For each quantile, the coefficient associated with the gender indicator reflects the conditional gender pay gap at the corresponding income level. Statistical inference is based on

heteroskedasticity-robust standard errors.
Unlike Multivariate Regression, Quantile Regression does not require strong distributional assumptions. The key assumptions of Quantile Regression are that the conditional quantile of the dependent variable must be linear function of the explanatory variables and that the observations must be independently drawn. In this study, independence of observations is ensured by the survey design.

2.1.4 Oaxaca–Blinder Decomposition

The Blinder-Oaxaca Decomposition is a statistical method that decomposes differences in mean outcomes across two groups into a part that is due to group differences in the levels of explanatory variables and a part that is due to differential magnitudes of regression coefficients, Marek, H. (2022) [8]. This method allows us to separate the wage gap into an explained part based on individual characteristics of workers and their firms, and an unexplained part that may reflect discrimination in the labor market. As in the case of the regression models, the dependent variable is the logarithm of gross hourly earnings, allowing the decomposition results to be interpreted proportionally. The set of explanatory variables consists of age group, work experience, educational level, company size, occupational group, part-time employment status and the presence of performance-related bonuses. Gender is used as a grouping variable, distinguishing between men and women. All categorical variables are included as factor variables and the decomposition is performed using a bivariate specification.
The twofold Oaxaca–Blinder Decomposition expresses the total wage gap as:

$$\Delta = \underbrace{(X_m - X_f)\hat{\beta}_f}_{\text{Explained}} + \underbrace{X_f(\hat{\beta}_m - \hat{\beta}_f)}_{\text{Unexplained}}, \tag{4}$$

where X_m and X_f denote the average characteristics of men and women, respectively, and $\hat{\beta}_m$ and $\hat{\beta}_f$ are the estimated coefficients from group-specific wage regressions.
To assess the statistical uncertainty of the decomposition results, standard errors are obtained using bootstrap resampling with 200 replications.

2.2 Prediction of Individual Hourly Wages

The second part of the study is aimed at moving from salary-explanation to individual salary prediction for women and men based on the characteristics of employees and their companies. This part uses machine learning models such as Random Forest, XGBoost and Support Vector Regression. When applying the models, the prepared dataset was divided into training and testing parts. 70% was used for model training, and the remaining 30% was used for testing. Model performance was evaluated using Root Mean Squared Error (RMSE), Mean Absolute Percentage Error (MAPE) and the coefficient of determination (R^2).

2.2.1 Random Forest

Random forests are a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest, Breiman, L. (2001) [5]. In this study, Random Forest is used to predict individual hourly wages for women and men. In this study, Random Forest is used to predict individual hourly wages. As in previous models, the dependent variable is the logarithm of gross hourly wages, and the set of explanatory variables consists of information about the worker and his/her workplace.
Formally, the Random Forest predictor is defined as the average of predictions from B regression trees:

$$\hat{y}_i = \frac{1}{B} \sum_{b=1}^{B} T_b(x_i), \tag{5}$$

where $\hat{y}_i$ denotes the predicted logarithm of hourly wage for individual i, $T_b(\cdot)$ represents the prediction of the b-th regression tree, and x_i is a vector of explanatory variables.
The dataset was randomly divided into training and testing subsets, with 70% of observations used for model training and 30% reserved for testing. The Random Forest model was estimated using 500 trees, while the number of candidate variables was fixed at six. Model performance was evaluated using the root mean squared error (RMSE), the coefficient of determination (R^2) and the mean absolute percentage error (MAPE). Predicted log wages are transformed back into euros per hour to facilitate interpretation of prediction accuracy in monetary terms.
For visual evaluation of the model, a graph was drawn, which depicts the distribution of actual and predicted values. In the graph the 45-degree reference line represents perfect predictive accuracy, where predicted wages are equal to observed wages. Deviations from this line indicate prediction errors, with points above the line reflecting overestimation and points below the line reflecting underestimation of wages. Additionally, a feature importance plot was drawn, showing the most significant variables.

2.2.2 Extreme Gradient Boosting

Extreme Gradient Boosting (XGBoost) is a distributed, open-source machine learning library that uses gradient boosted decision trees, a supervised learning boosting algorithm that makes use of gradient descent. It is known for its speed, efficiency and ability to scale well with large datasets, IBM (2025) [9]. In this study, XGBoost is applied to predict individual hourly wages using a database prepared, in 2022. The dependent variable and the set of explanatory variables are the same as in the previous models.
Formally, the XGBoost model represents the predicted outcome as an additive function of K regression trees:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), \quad f_k \in \mathcal{F}, \tag{6}$$

where $\hat{y}_i$ denotes the predicted logarithm of hourly wage for individual i, x_i is a vector of explanatory variables, and $\mathcal{F}$ denotes the space of regression trees.

Model training is performed by minimising a regularised objective function:

$$\mathcal{L} = \sum_{i=1}^{N} \ell(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k), \tag{7}$$

where $\ell(\cdot)$ denotes the squared error loss function and $\Omega(f_k)$ is a regularisation term that penalises model complexity and reduces overfitting.
Categorical variables are transformed using single coding and entered into the model as binary indicators. The dataset was randomly divided into training (70%) and testing (30%) subsets. Model performance was evaluated using the root mean squared error (RMSE), the coefficient of determination (R^2) and the mean absolute percentage error (MAPE). Predicted log wages are transformed back into euros per hour to facilitate interpretation of prediction accuracy in monetary terms.
After evaluating the model's prediction accuracy, the distribution of actual and predicted values plot was drawn, which allows us to visually assess the model's ability to predict values. This graph shows at what salary level the model is able to predict values most accurately. To understand which variables are most important for predicting individual wages, a SHAP summary plot was drawn. SHAP is an interpretation system that assigns a contribution value to each explanatory variable for a given forecast. These values indicate how much each attribute increases or decreases the predicted wage.

2.2.3 Support Vector Regression

Support Vector Machines (SVR) are widely used in machine learning for classification problems, but they can also be applied to regression problems through Support Vector Regression (SVR), Sethi, A (2020) [15]. Unlike tree-based ensemble methods, SVR constructs predictions based on a subset of training observations known as support vectors.In this study, SVR is applied to predict individual hourly wage, using the dependent variable and the set of explanatory variable, which are the same like in other predictive models. In this study categorical variables were transformed using one-hot encoding and incorporated into the model as binary indicators. The resulting feature matrix was used as input for the SVR algorithm.
Formally, Support Vector Regression estimates the following function:

$$f(x) = \langle w, \phi(x) \rangle + b, \tag{8}$$

where $\langle w, \phi(x) \rangle$ denotes the scalar product in the feature space, $\phi(x)$ is a nonlinear mapping of the input variables into a high-dimensional feature space, w is a vector of weights, and b is the bias term. The SVR model is estimated by solving the following optimisation problem:

$$\min_{w,b,\xi_i,\xi_i^*} \frac{1}{2}\langle w, w \rangle + C \sum_{i=1}^{N} (\xi_i + \xi_i^*), \tag{9}$$

subject to:

$$y_i - f(x_i) \leq \varepsilon + \xi_i, \tag{10}$$

$$f(x_i) - y_i \leq \varepsilon + \xi_i^*, \tag{11}$$

$$\xi_i, \xi_i^* \geq 0, \tag{12}$$

where ε defines the width of the ε-insensitive loss function, C is a regularisation parameter controlling the trade-off between model complexity and prediction errors, and ξ_i, ξ_i^* are slack variables allowing deviations outside the ε-tube.

A radial basis function (RBF) kernel is employed to capture non-linear relationships between predictors and wages. The kernel parameter is set as the inverse of the number of predictors, while model hyperparameters are selected to balance bias and variance.

Again, the dataset was randomly divided into training (70%) and testing (30%) subsets. Model performance was evaluated using the root mean squared error (RMSE), the coefficient of determination (R^2) and the mean absolute percentage error (MAPE). Predicted log wages are transformed back into euros per hour to facilitate interpretation of prediction accuracy in monetary terms.

As with other machine learning models used in this study, the SVR model is used exclusively for prediction purposes. Model outputs and variable relationships are not interpreted as causal effects.

As with other machine learning models, after using the Support Vector Machines (SVR) model, the distribution of actual and predicted values plot and feature importance graph were drawn.

2.3 Predicting Gender Average Wages at the Occupational Level

The last part of the study is designed to evaluate how machine learning models able to predict the salaries of women and men separately, at the occupational level. In other words, we want to estimate how average occupational characteristics allows us to predict average wages, and whether these relationships differ by gender. For this part predictive machine learning models, such as Random Forest, XGBoost and Support Vector Regression (SVR) were applied. These models were also found in the previous part - predicting individual hourly wages, so a more detailed description of the models will not be repeated. All models were developed separately for men and women to compare the accuracy of the models and the importance of the variables between genders.

The study was conducted using a prepared database. Before using the predictive models, the data were aggregated at the occupation level to reduce noise and ensure statistical reliability. Aggregation was performed separately for men and women, thus forming two independent data samples. The following averages were calculated for each occupation: gross salary, age of employees, length of service, number of overtime hours, working time share, number of paid hours, amount of bonuses. These two samples were divided into training and testing parts, with 70% allocated to training the models and the remaining 30% to testing.

The predictive quality of the models was assessed using three standard regression metrics: RMSE (Root Mean Squared Error), R^2 (coefficient of determination) and MAPE (Mean Absolute Percentage Error). These metrics allow us to assess the accuracy of models and compare different methods with

each other. In order to determine the most significant aggregated variables, an analysis of the importance of variables was performed. In Random Forest models, importance is assessed using the percentage increase in mean squared error criterion (IncMSE), which shows how much the model's prediction accuracy deteriorates when information on a specific variable is changed/corrupted. A higher IncMSE value indicates a higher value of the variable in predicting wages. XGBoost models use a Gain indicator that reflects the contribution of each variable to the accuracy of the model. Additionally, SHAP (SHapley Additive exPlanations) analysis was applied to XGBoost models, allowing evaluate not only the importance of variables, but also the direction and strength of their impact on predicted wages. SVR models were subjected to variable standardization to ensure that all features were comparable. In addition, hyperparameter tuning was performed on the male sample using 5-fold cross-validation to test whether the optimized model parameters improved the prediction accuracy. As in the case of Random Forest, a graph of the importance of variables was drawn.

3 Data

The study, designed to analyze the salaries of men and women, used the publicly available database of the Lithuanian State Data Agency - Statistical survey of wage structure. The database consists of information on the hourly wage of Lithuanian employees, distinguishing employees' gender, education, work experience, age group, field of work and other categorical and non-categorical values. Data were collected every 4 years - in 2014, 2018 and 2022, in October, when conducting a statistical survey of the wage structure. The main attention will be paid to the data for 2022 to examine the most relevant information. The data for 2022 consists of 46561 rows (33 columns), of which 48% are women and 52% are men. In the database, professions were coded, therefore an additional database was used - the Lithuanian Occupations Classification, prepared based on the international system ISCO-08 (International Standard Classification of Occupations).

The data was prepared for further analysis. First, missing values were filled in and unnecessary columns with metadata were removed. Then, the different data structures were unified - column names were arranged. Also, interval age data was converted to numerical values. The main quantitative variables, such as age, length of service, overtime, hourly wage, and working time share, were transformed to numerical type and logically impossible values were removed. To better understand trends, we selected profession where are at least 100 men and 100 women, both genders accounted for at least 30 percent of all employees in that profession. In this way, professions dominated by one gender were eliminated. The combined dataset with 2014, 2018 and 2022 includes 47604 observations, of which 17152 are from 2022, this data will be used in applying regression and machine learning models.

Preliminary data analysis was conducted five selected professions that had the largest gender pay gap in the 2022 data: sales, marketing and development managers; professional services managers, medical doctors, database and network professionals, assemblers. These occupations were selected to visually reveal the structure of the gender gap and its trends, where the gap has the largest effect and is most pronounced in the labor market. The top 5 occupation subset is not intended to draw statistical conclusions generalizing to the entire labor market. Meanwhile, all causal and predictive models were trained using the full set of occupations.

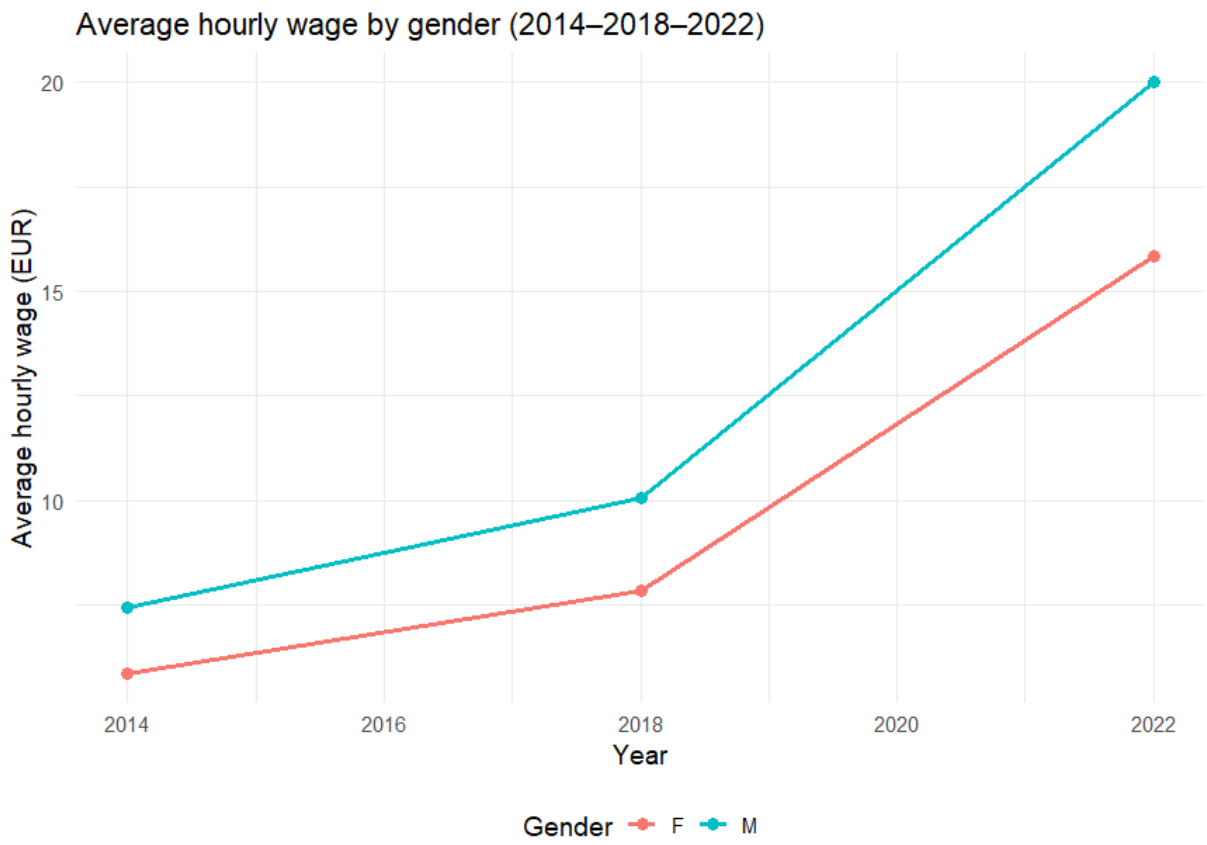
4 Results

This section presents the results of the study. First, the results of preliminary data analysis are presented together with Two-way ANOVA. Then, the results of a more depth study are presented, including the explanatory part, which consists of Multivariate Regression, Quantile Regression and Oaxaca–Blinder Decomposition models performances. The following are applications of machine learning models (Random Forest, XGBoost and the Support Vector Regression (SVR)) to predict individual wages and predict the salaries of women and men separately, at the occupational level. In addition graphical analysis will be shown to evaluate the ability of the models.

4.1 Preliminary data analysis

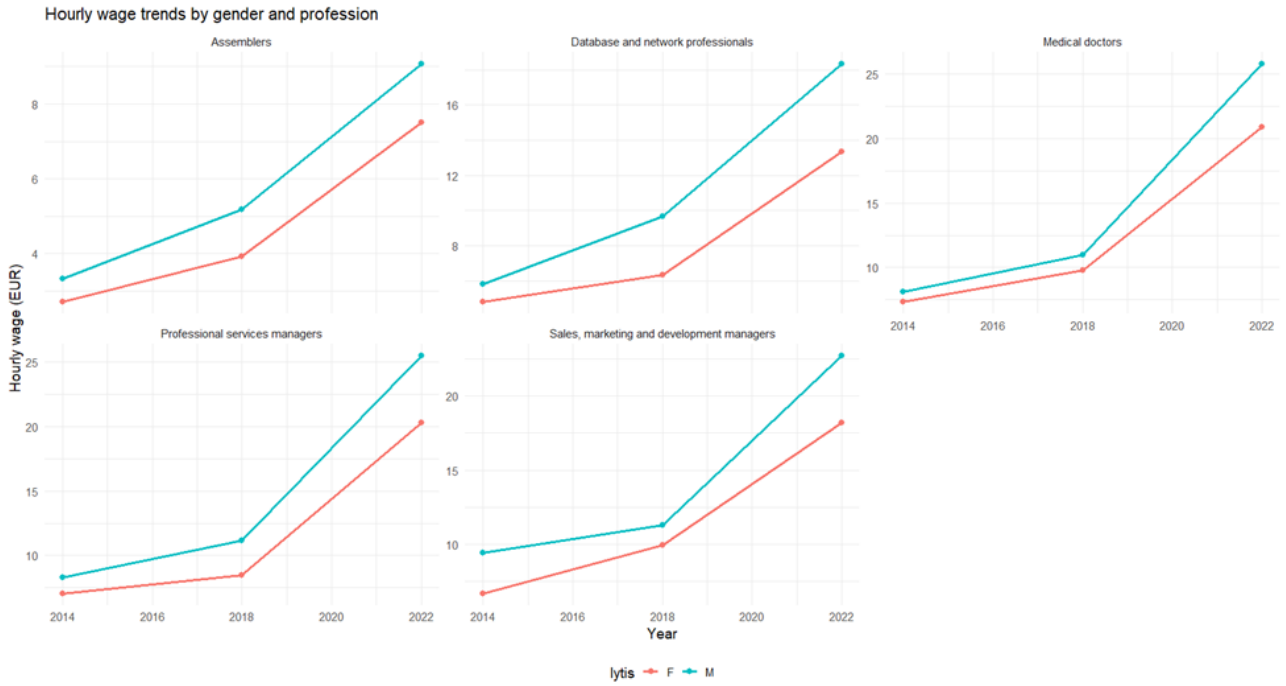
4.1.1 Descriptive statistics

The research sample reflects five occupational groups that are characterized by a similar ratio of women to men within the group and a significant difference in average hourly earnings between the genders. 1 figure. shows the variation in average hourly wages by gender over the years. In 2022, the average hourly wage for men in these five professions was €20.00 and for women €15.88, 23.2% lower than the hourly wage for men. Although wages have been growing since 2014, the gender pay gap has remained similar, at 23.7% in 2014 and 24.7% in 2018.



1 figure. *Average hourly wage trends by gender (2014-2022)*

Analyzing wage differences, it was noted that the largest differences in average hourly pay between the sexes are among representatives with higher education. A particularly pronounced gap is recorded between database and network professionals and sales, marketing and development managers. In 2022, male database and network professionals earned as much as 31.4% more than women, which is 4.97 euros per hour, and sales, marketing and development managers 22%.

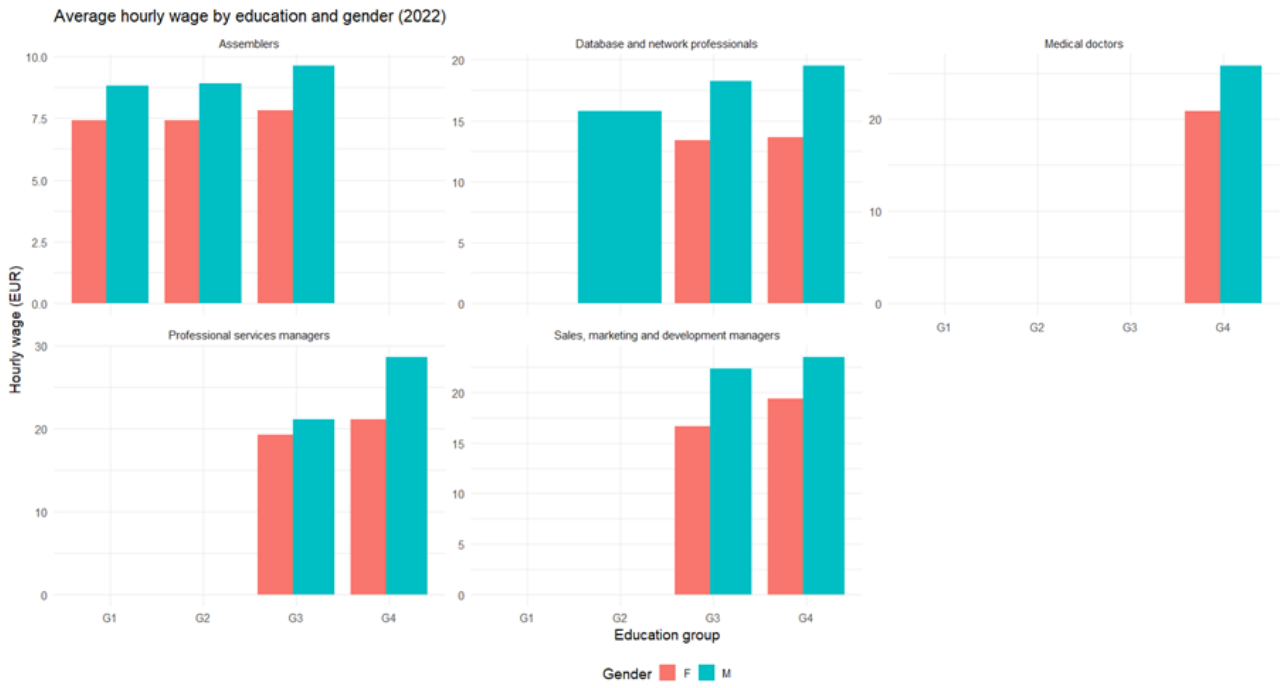


2 figure. *Average hourly wage trends by gender (2014-2022)*

The biggest change in gender pay gap in average hourly earnings was in the medical field, where the gap increased from 10.3% (2014) to 21.00% (2022). The consistent growth of the gap is recorded in professional services managers, where the gap increased from 15.7% to 22.7%.

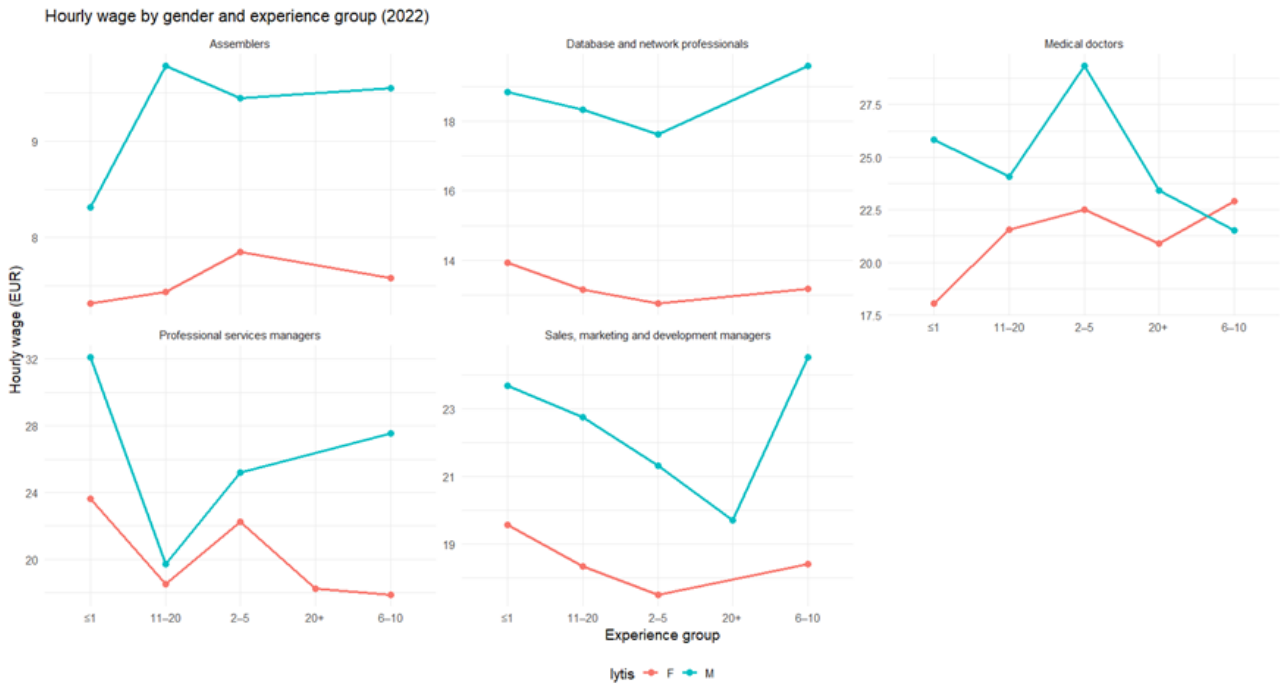
Initial results showed that there is a gender gap in average hourly earnings. While earnings increase with higher levels of education for both women and men, however, women usually earn less than men, even with higher education.

For example, in 2022, women with a higher education (university higher education - bachelor's/master's/PhD level) earn 13.2% less than men with a college education and even 12.1% less than men with a secondary education. A similar situation is observed in other areas of work, with women's average hourly wage below 8 euros, regardless of their education, while men's wages range from 8.82 to 9.62 per hour for assembler. Only one of the five occupational groups in which women with higher education earned roughly the same hourly wage as men with college-level education was professional services managers. However, their wage remained well below that of highly educated men 21.1 euros compared to 28.6 euros.



3 figure. *Average hourly wage by education and gender (2022)*

The ANOVA test confirmed that there are statistically significant differences in wages based on education level, gender and profession (all $p < 0.05$). Education has a particularly strong effect, with F-value of 174.01 ($p < 0.01$) , that mean higher education on wages is not the same for men and women, confirming that earning differ across education groups. Welch Two Sample T-test revealed that database and network professionals highly educated and experienced women earn less than less educated and less experienced men ($p < 0.001$). Looking at how the average hourly wage changes for men and women separately, we notice that in only one of the five professions do female medical workers earn 0.39 euros more than men when their work experience is 6-10 years. From 11 years of experience, men's salaries become higher.



4 figure. *Average hourly wage by education and gender (2022)*

Although women's average hourly wages usually are lower, regardless of their level of education and work experience companies are more likely to hire women with higher education. As many as 55.2% of women working in companies have higher education, compared to 51.8% of men. The ANOVA test showed that hourly wage is statistically determined by gender, education, company size, working time, work experience, age and profession, while the number of overtime hours has no significant effect ($p > 0.05$).

4.1.2 Two-way ANOVA

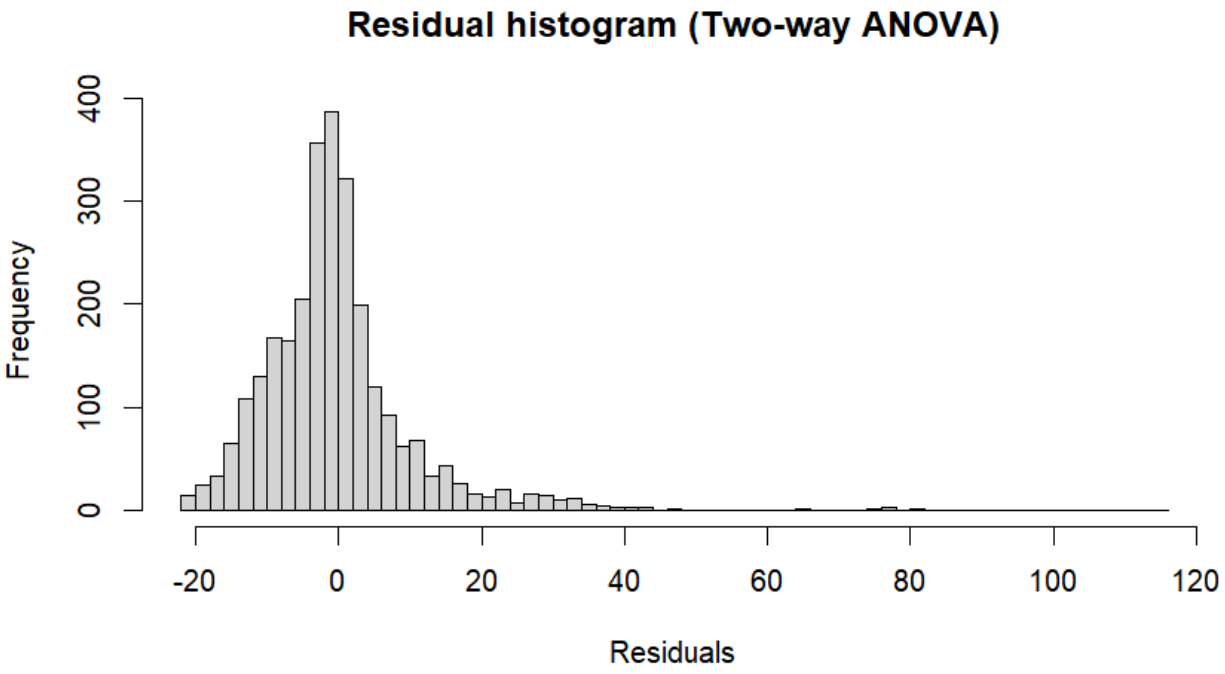
The preliminary analysis and graphical evidence suggest wage differences across gender and occupational groups. To evaluate whether these observed differences are statistically significant and to examine whether the gender wage gap varies across professions, a two-way analysis of variance (ANOVA) was conducted using 2022 data. The dependent variable was gross hourly wage, while gender and occupation were treated as categorical independent variables.

The results, presented in 1 table. indicate statistically significant main effects of both gender and occupation on hourly wages. The main effect of gender is statistically significant with $F = 483.99$ and $p < 0.001$, meaning that average hourly wages differ between women and men across occupations. However, the corresponding effect size is only $\eta^2 = 0.019$, it means that gender alone explains only a small share of total wage variation. Meanwhile, main effect of occupation suggest stronger effect on wages and the corresponding effect size with $F = 316.50$ ($p < 0.001$) and $\eta^2 = 0.274$. Although, the interaction between gender and occupation is also statistically significant 8.425 ($p < 0.001$), the corresponding effect size with is very small, only $\eta^2 = 0.007$. These results provide evidence to reject the null hypotheses of equal mean wages across gender, equal mean wages across occupations, and no interaction between gender and occupation.

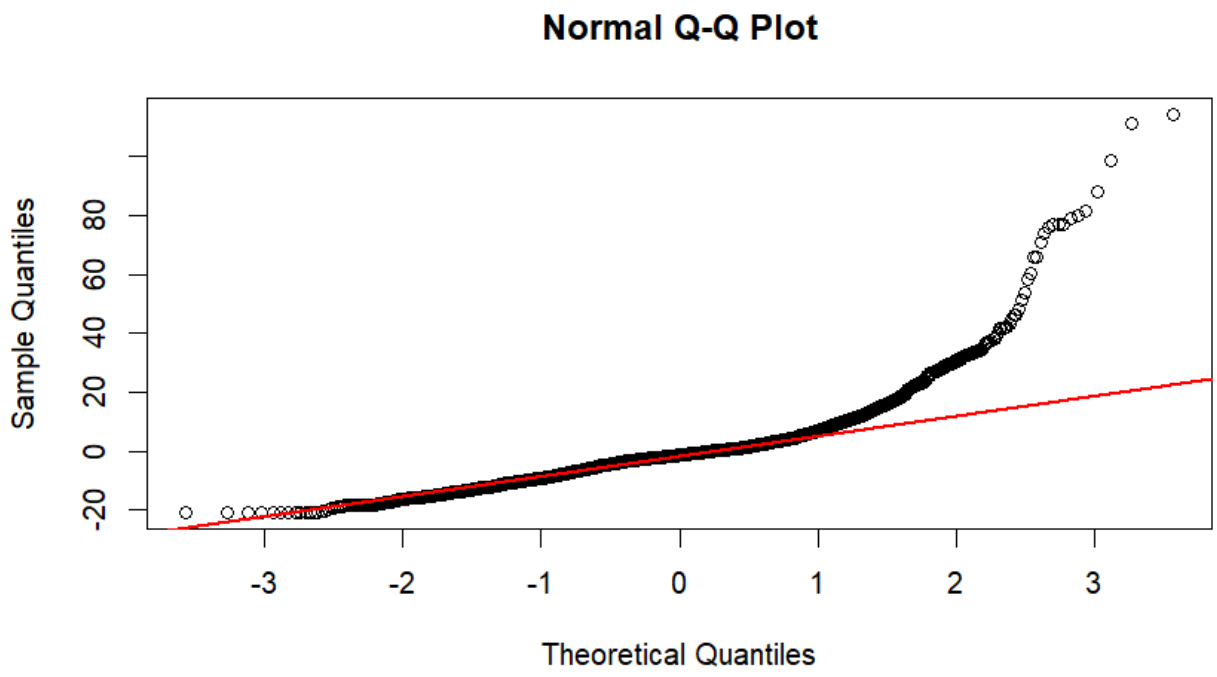
Source of variation	df	Sum of Squares	Mean Square	F-value	η^2
Gender	1	25 198	25 198	483.995^{***}	0.019
Occupation	22	362 516	16 478	316.500^{***}	0.274
Gender $\times$ Occupation	22	9 650	439	8.425^{***}	0.007
Residuals	17 785	925 946	52	–	0.700

1 table. *Two-way ANOVA results for hourly wages by gender and occupation (2022)*

Testing assumption, we denoted, what residuals histogram (5 figure.) shows that residuals are approximately symmetrically distributed around zero and the distribution is approximately bell-shaped, while in Q–Q plot (6 figure.) significant deviations are observed in the upper tail, where points systematically rise above the reference line. Given the large sample size, these deviations are not considered problematic, and the normality assumption is deemed satisfied.



5 figure. *Residual histogram (Two-way ANOVA)*



6 figure. *Q-Q plot (Two-way ANOVA)*

Levene's test for homogeneity of variances indicates statistically significant heterogeneity across groups ($p < 0.001$). To evaluate the robustness, a heteroskedasticity-robust ANOVA using HC3-adjusted standard errors was applied. The robust results confirm that the main effects of gender and occupation, as well as their interaction, remain statistically significant.

To better understand the significant interaction effect, post-hoc comparisons were applied using estimated marginal means. These comparisons confirms that statistically significant gender wage differences are present in many occupations, especially among high-skilled professions such as medical doctors, finance professionals, database and network professionals, and professional services managers. Moreover, in several other professions—such as business services agents, university and higher education teachers, and certain elementary occupations—the differences in average hourly wages between women and men are not statistically significant. These findings underscore the heterogeneous nature of the gender wage gap.

4.2 Regression-based analysis of the gender pay gap

4.2.1 Multivariate Regression

The Multivariate Regression model, obtained from the wage equation in Equation 2 revealed that the coefficient of the gender indicator is positive and highly statistically significant. The results show that women with the same individual and company characteristics earn 11.5% less than men, in 2022 data. Most variables are strongly related to wages ($p < 0.001$). Higher levels of education (G3 and G4) are associated with higher wages. Working in larger companies and bonuses are positively associated with hourly wages. Part-time work is associated with significantly lower wages

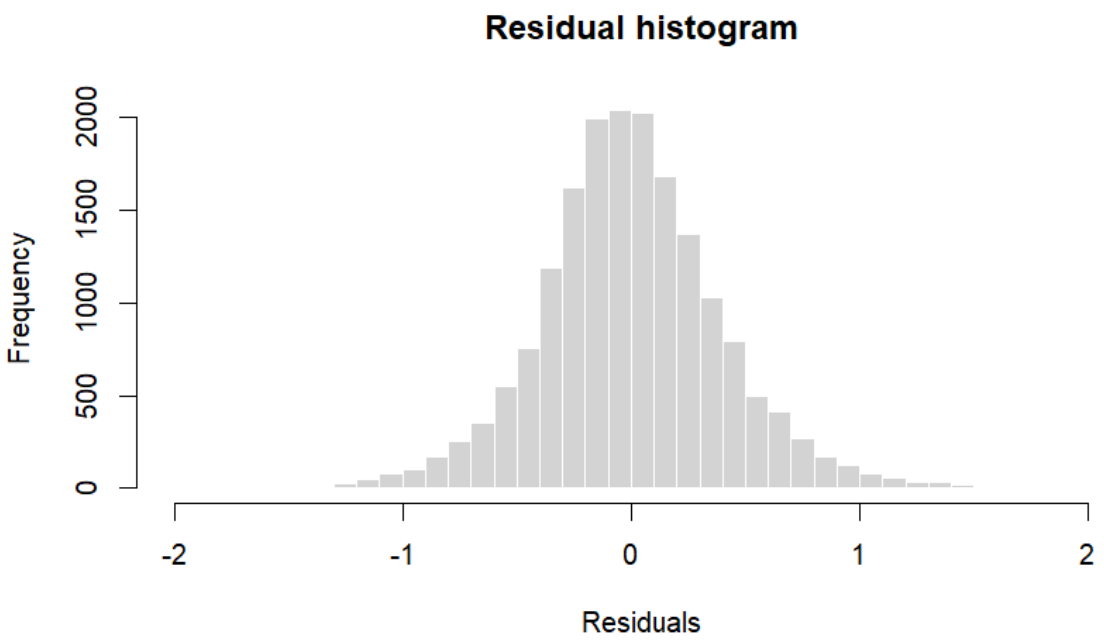
($\hat{\beta} = -0.624$, $p < 0.001$), which corresponds to an approximate $\exp(-0.624) - 1 \approx -46.4\%$ wage gap compared to full-time work. Full regression results for this specification are reported in 2 table.. The model fit statistics are presented at the bottom of the table. The model does not have high explanatory power with $N = 17831$ and the $R^2 = 0.4862$. The overall model is statistically significant, as indicated by the F statistic.

Variable	Estimate	Std. Error	t-value	p-value
Intercept	2.056	0.018	111.57	< 0.001
Gender (lytis_num)	0.115	0.007	17.08	< 0.001
Age group (linear)	-0.093	0.007	-12.75	< 0.001
Age group (quadratic)	-0.077	0.006	-12.36	< 0.001
Age group (cubic)	0.022	0.006	3.50	< 0.001
Tenure (stazas)	0.0019	0.0005	3.84	< 0.001
Education G2	0.024	0.014	1.73	0.083
Education G3	0.139	0.016	8.84	< 0.001
Education G4	0.237	0.016	14.40	< 0.001
Firm size 50–249	0.219	0.009	25.02	< 0.001
Firm size (all)	0.031	0.017	1.88	0.061
Firm size >250	0.233	0.008	28.87	< 0.001
Assemblers	-0.308	0.019	-15.95	< 0.001
Business services agents	-0.211	0.027	-7.89	< 0.001
Business services managers	0.240	0.016	15.10	< 0.001
Client information workers	-0.212	0.022	-9.82	< 0.001
Creative and performing artists	-0.272	0.024	-11.18	< 0.001
Database and network professionals	0.179	0.020	9.10	< 0.001
Cleaners and helpers	-0.575	0.016	-36.88	< 0.001
Finance professionals	0.082	0.013	6.21	< 0.001
Food processing workers	-0.270	0.022	-12.55	< 0.001
Legal professionals	0.232	0.024	9.68	< 0.001
Manufacturing labourers	-0.359	0.018	-19.70	< 0.001
Transport clerks	-0.176	0.018	-9.52	< 0.001
Medical doctors	0.414	0.018	22.58	< 0.001
Other elementary workers	-0.560	0.022	-25.79	< 0.001
Professional services managers	0.440	0.024	18.57	< 0.001
Regulatory associate professionals	-0.079	0.020	-4.01	< 0.001
Sales and purchasing agents	-0.118	0.016	-7.21	< 0.001
Sales and marketing managers	0.333	0.018	18.07	< 0.001
Sales and PR professionals	0.061	0.013	4.82	< 0.001
Shop salespersons	-0.438	0.017	-25.31	< 0.001
University teachers	-0.012	0.023	-0.54	0.592
Wood trades workers	-0.277	0.021	-12.98	< 0.001
Part-time	-0.624	0.086	-7.28	< 0.001
Bonuses	0.00002	7.5×10^{-7}	26.17	< 0.001
R^2		0.486		
F-statistic		479.7		
Observations		17 831		

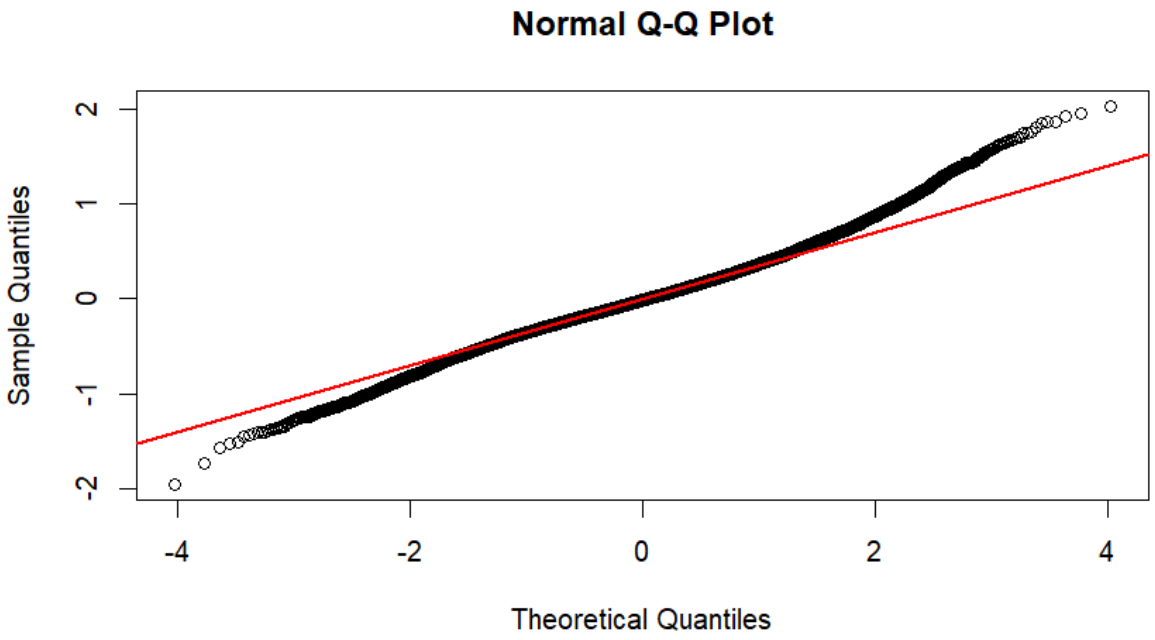
2 table. *Linear regression results (dependent variable: log wage)*

To ensure the reliability of the estimated coefficients, a series of assumptions tests were performed. Residual normality was assessed using graphical methods and numerical measures of skewness and kurtosis. As illustrated in Figures 7 figure. and 8 figure., the residual distribution is approximately symmetric and centered around zero with minor deviations in the tails, which are not considered problematic given the large sample size.The residual distribution exhibits low skewness (0.24) and moderate excess kurtosis (1.19), indicating slightly heavier tails than the normal distribu-

tion. Given the large sample size, these deviations are considered insignificant and the normality assumption is considered satisfied.



7 figure. *Residual histogram*



8 figure. *Normal Q-Q Plot*

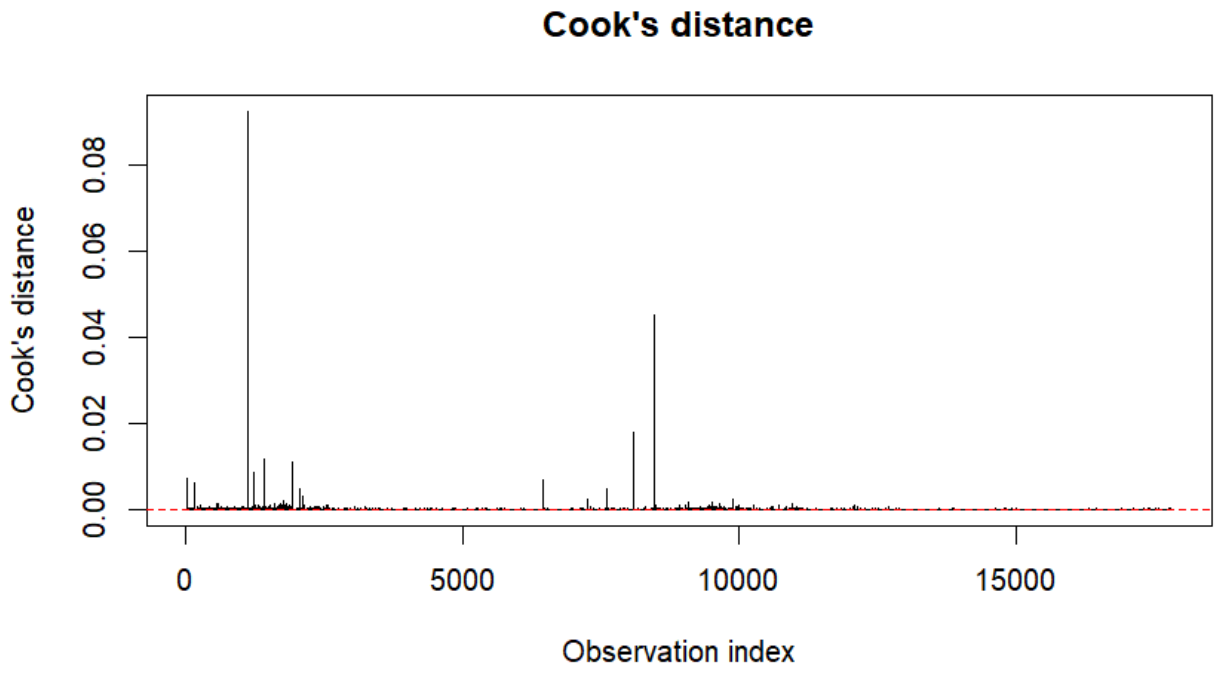
The Breusch–Pagan test indicates the presence of heteroskedasticity with $p < 0.05$. Therefore, heteroskedasticity-robust standard errors are employed to ensure valid statistical inference. While coefficient estimates remain unchanged, the statistical significance of some variables changes slightly.

The lower education category (issilavinimasG2) becomes statistically significant. 3 table. depicts Multicollinearity diagnostics based on the variance inflation factor reveal no evidence of strong linear dependence among the explanatory variables.

Variable	GVIF	Df	$\text{GVIF}^{1/(2\cdot Df)}$
Gender (lytis_num)	1.164	1	1.079
Age group	1.652	3	1.087
Tenure (stazas)	1.334	1	1.155
Education (issilavinimas)	2.524	3	1.167
Firm size (im_dydis)	1.272	3	1.041
Occupation (Pavadinimas)	3.838	22	1.031
Part-time	1.005	1	1.003
Bonuses (premijos)	1.044	1	1.022

3 table. *Multicollinearity diagnostics (GVIF)*

Finally, 9 figure. presents Cook's distance values for individual observations. The majority of observations exhibit very low Cook's distance. Although, some observations exceed $4/N$ threshold due to the large sample size, no single observation has significant influence on the regression results.



9 figure. *Cook's distance*

Overall, while not all regression assumptions are fully satisfied, the diagnostic analysis suggests that the identified deviations do not undermine the validity of the main results.

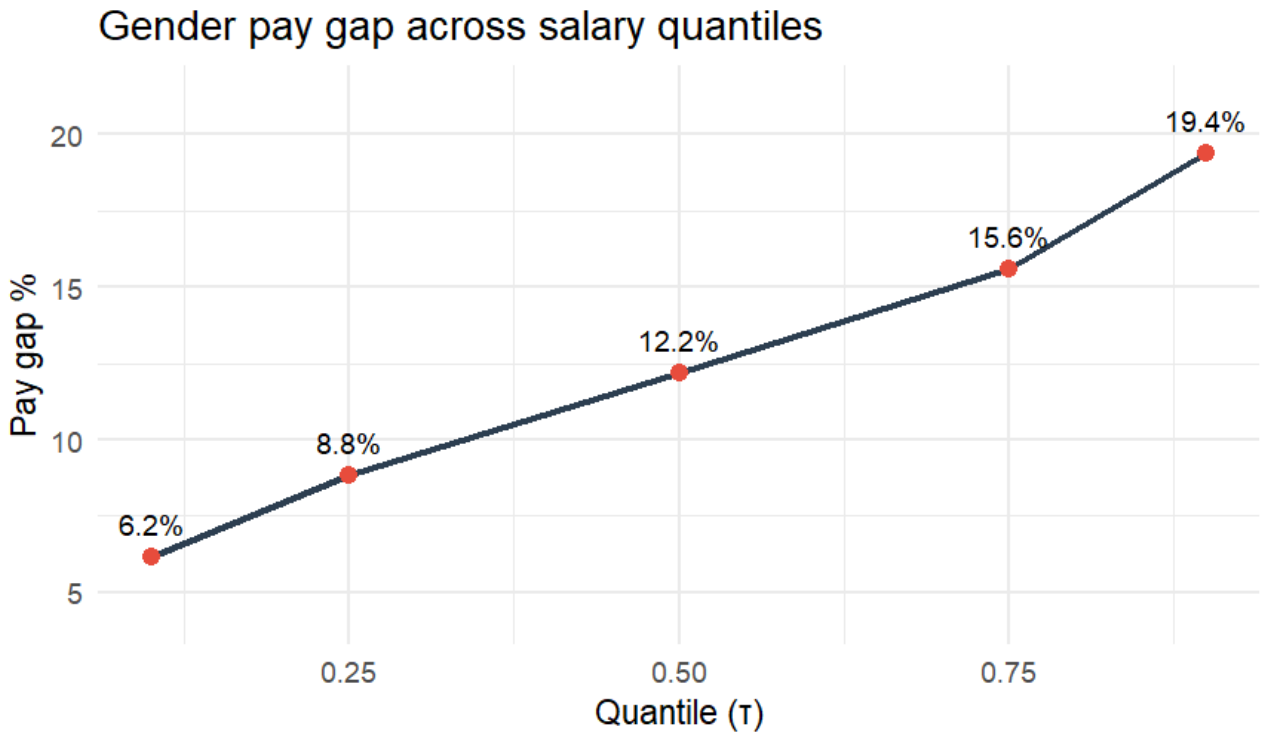
4.2.2 Quantile Regression

Moving to the second regression model, obtained from the wage equation in Equation 3, the Quantile Regression framework allows the gender wage gap to be examined across different points of the wage distribution. 4 table. presents the results of the Quantile Regression. Here we see that the estimated coefficient of the gender indicator (men = 1, women = 0) is positive and statistically significant in all quantiles considered ($\tau = 0.10, 0.25, 0.50, 0.75$ and 0.90). This shows that, depending on the observed characteristics, such as age, length of service, education, company size, occupation, working hours and bonuses, men systematically earn higher wages than women across the entire wage distribution. Also, the estimated gender effect size increases with wage quantile. At the 10th percentile, the estimated coefficient is 0.06, which corresponds to an implied gender wage gap of approximately 6.2%. The gap increases to 8.8% at the 25th percentile and reaches 12.2% at the median of the wage distribution. At the upper end of the distribution, the estimated gap increases further, reaching 15.6% at the 75th percentile and 19.4% at the 90th percentile, as it is shown in 10 figure.

Quantile (τ)	Coefficient	Std. Error	t-statistic	Implied gap (%)
0.10	0.060	0.006	9.73	6.18
0.25	0.084	0.008	10.57	8.76
0.50	0.115	0.007	16.13	12.18
0.75	0.145	0.008	19.07	15.63
0.90	0.177	0.010	17.90	19.36

4 table. *Gender pay gap estimates across wage quantiles (Quantile Regression)*

Note: The dependent variable is log wages. The implied percentage gap is calculated as $100 \times (e^{\beta} - 1)$.



10 figure. *Gender pay gap across salary quantiles*

The results reveal that in the higher wage quantiles, gender inequality is not uniform across income levels. The gender pay gap increases as wages increase. These results confirm previous results obtained in the preliminary data analysis section and multivariate regression.

4.2.3 Oaxaca–Blinder Decomposition

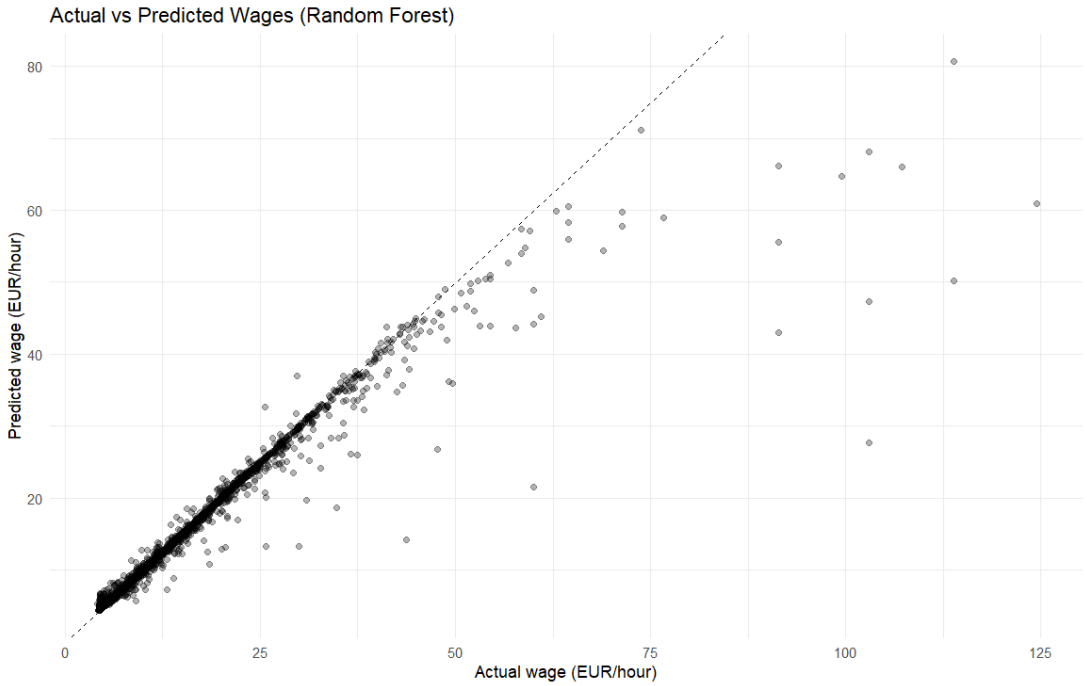
Although the regressions analysis identifies a significant gender wage gap, it does not provide a direct decomposition of the gap into explained and unexplained components. Therefore, the Oaxaca–Blinder Decomposition is employed to quantify the extent to which differences in observable characteristics account for the wage gap.
Oaxaca–Blinder Decomposition results revealed that the average log hourly wage of women is 2.279, compared to 2.435 for men, resulting in a total wage gap of -0.156 log points, which is 14.5%. Approximately 19.1% of this difference can be explained by differences in observable characteristics between men and women, such as age, education, company size, work experience, and bonuses. The remaining 80.9% of the difference is unexplained and can be attributed to discrimination or unanticipated factors such as labor productivity.

Overall, the regression-based analysis confirms that statistically significant gender pay gap exists in Lithuania, in 2022. Multivariate Regression results indicate that, after controlling for observable characteristics, women earn on average 11.5% less than men. Quantile Regression reveals that the gender pay gap increases across the wage distribution. Oaxaca–Blinder Decomposition further shows that only 19.1% of the observed wage gap can be explained by differences in characteristics, while the remaining part (80.9%) remains unexplained and may reflect structural or discriminatory factors.

4.3 Prediction of Individual Hourly Wages

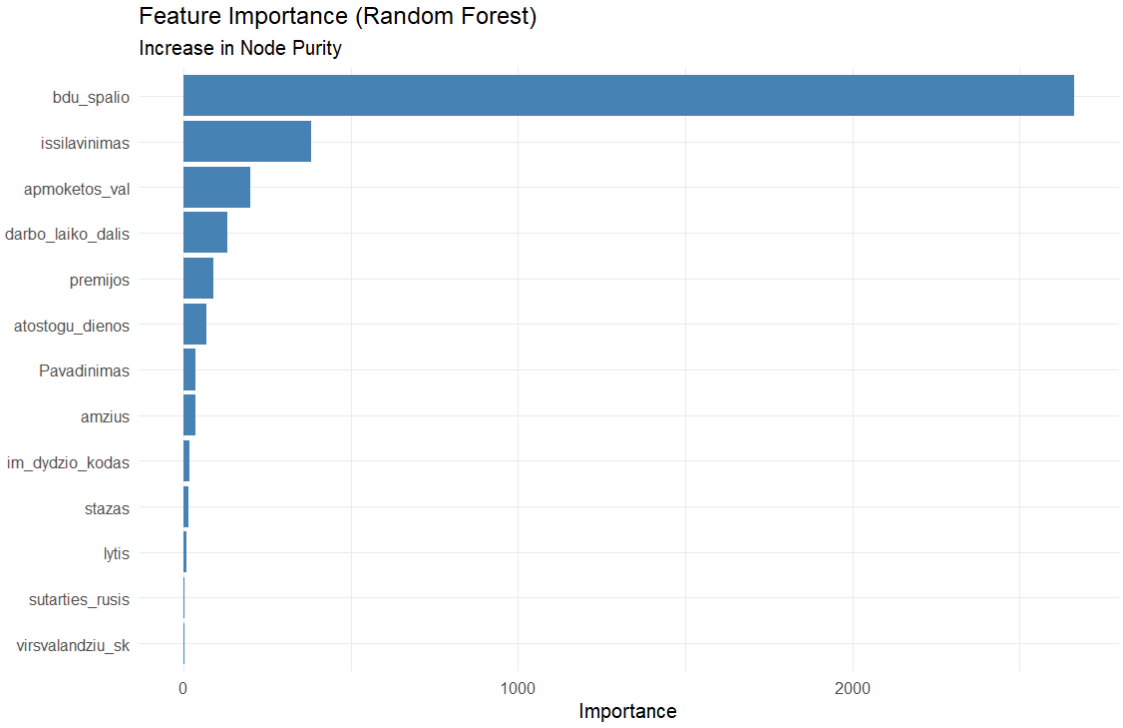
4.3.1 Random Forest

The results of the Random Forest model in predicting individual wages did not show ideal accuracy in the 2022 data set. In the testing sample, the absolute forecast error (RMSE) reached 2.605, and the Model's coefficient of determination is $R^2 = 0.93$, which means that the model explains 93% of the actual hourly wage dispersion. Additionally, the mean absolute percentage error (MAPE) was calculated, which amounted to 1.99%, which indicates an acceptable level of forecasting accuracy.
11 figure. shows a graph of the relationship between actual and predicted hourly wages obtained using the Random Forest model. Most observations are close to the 45-degree line, indicating a good fit and accuracy between actual and predicted values. From the graph, we can see that the prediction accuracy is particularly accurate at low and medium wages. The model has more difficulty predicting the observed values at high wages.



11 figure. *Actual vs Predicted Wages (Random Forest)*

12 figure. depicts the importance of variables in predicting wages. The most influential predictor is $bduspalio$ - gross wages in October. Employee and job-related characteristics, such as education ("$issilavinimas$"), paid hours ("$apmoketosval$"), and full-time position ("$darbolaikodalis$"), also play an important role in predicting wages. In contrast, gender is not very important, indicating that wage differences are primarily determined by structural and job-related characteristics, rather than gender alone.



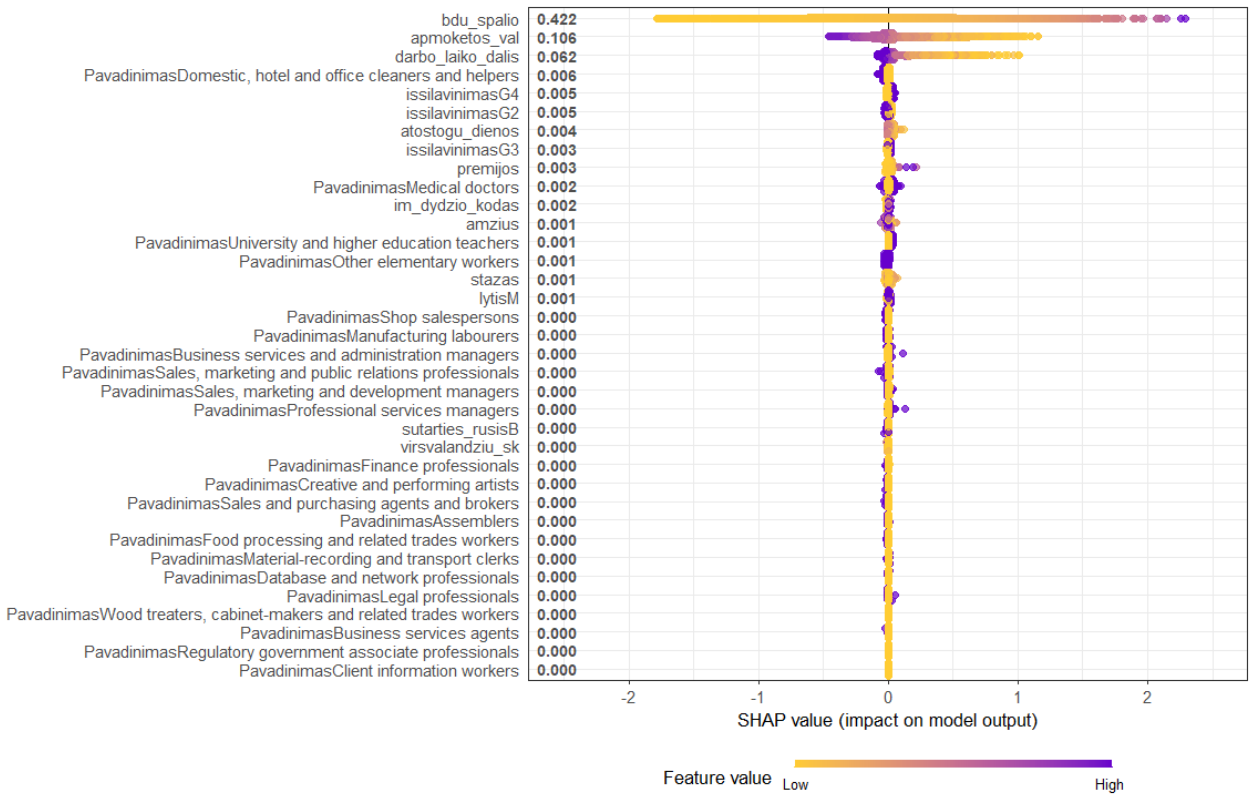
12 figure. *Feature Importance (Random Forest)*

4.3.2 XGBoost

Meanwhile, the XGBoost model offers even better accuracy in predicting individual wages. This model achieved a Model's coefficient of determination of $R^2 = 0.978$, explaining almost 98 percents of the actual hourly wage dispersion. The RMSE reaches 1.431, and the mean absolute percentage error (MAPE) does not even reach 1% (0.98%), indicating good model accuracy in the 2022 dataset. Not only do the calculated estimates show good accuracy of the XGBoost model, but the distribution of actual and predicted values plot also confirms this, represented in 13 figure.. The best accuracy is recorded among lower wages, with deviations recorded at higher wages, as in the case of Random Forest.

13 figure. *Actual vs Predicted Wages (XGBoost)*

The SHAP analysis results for XGBoost presented in 14 figure. show that the most significant variable is $bduspalio$ - gross wages in October. From the graph, we see that ($bduspalio$) makes the largest contribution (0.422), significantly exceeding the influence of other variables. Variables related to working time, such as paid hours (0.106) and working time share (0.062), are of further importance. In contrast, education and demographic characteristics, including gender, show relatively low average absolute values of SHAP, indicating a limited direct contribution to wage predictions, considering work-related factors.



14 figure. *SHAP analysis*

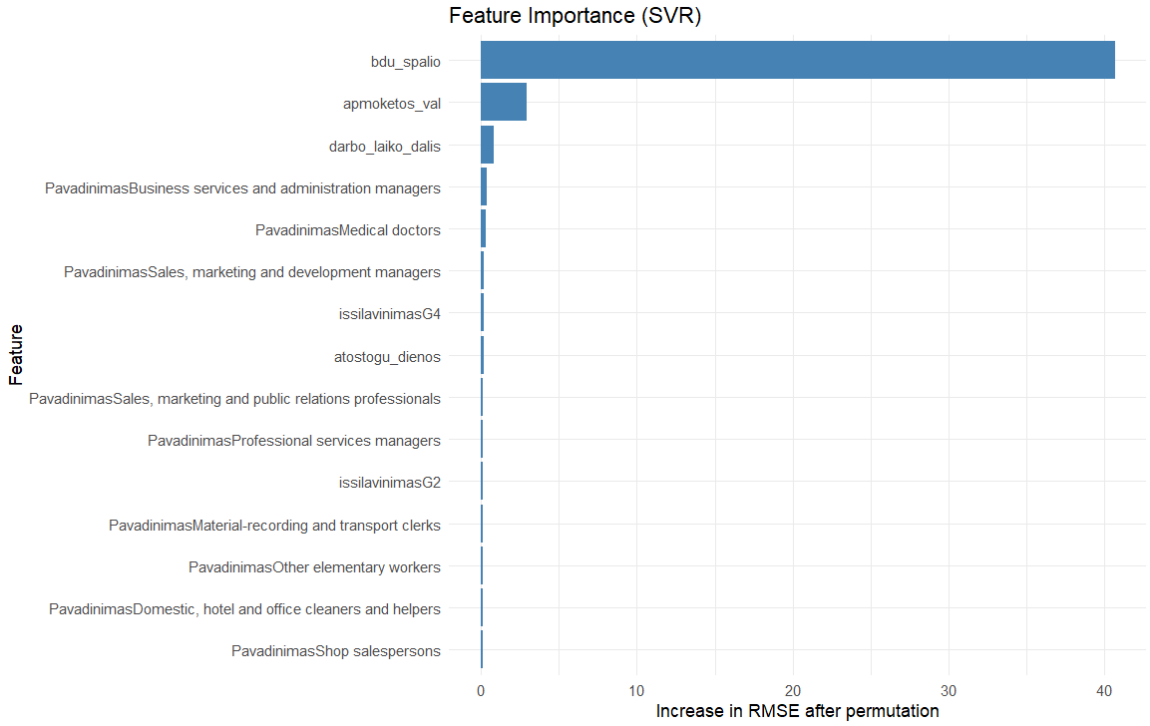
4.3.3 Support Vector Regression (SVR)

Support Vector Regression (SVR) performed worse than Random Forest and XGBoost in predicting individual wages. The model achieved a Model's coefficient of determination of $R^2 = 0.812$ and a poor RMSE of 4.032. The mean absolute percentage error (MAPE) exceeded 5% (5.786%), indicating poor but acceptable accuracy.

Overall, the 15 figure. shows a reasonable fit between predicted and observed wages, especially at low and middle wages, where most observations are close to the 45-degree reference line. This suggests that the SVR was able to capture the overall wage changes in the data. However, compared to other machine learning models used previously, SVR has a significantly higher prediction error, especially at higher wages.

15 figure. *Actual vs Predicted Wages (SVR)*

The feature importance plot (16 figure.) confirms that the most important variable in predicting individual wages is $bduoktobris$ - gross wages in October. Paid hours and part-time work (full-time part) are also very important in predicting values. Notably, gender does not emerge as an important predictor in the SVR model. This finding is consistent with the results obtained from Random Forest and XGBoost models.



16 figure. *Feature Importance (SVR)*

The performance of three machine learning models: Random Forest, XGBoost and Support Vector Regression in predicting individual hourly wages based on employee and firm characteris-

tics are represented in 5 table. Among the evaluated models, XGBoost achieved the best overall performance, exhibiting the lowest prediction error and the highest coefficient of determination ($R^2 = 0.97$). The Random Forest model also performed well, although its predictive accuracy was slightly lower compared to XGBoost. In contrast, the Support Vector Regression model showed weaker performance, characterized by higher prediction errors and a lower explanatory power.

Model	RMSE	R^2	MAPE (%)
XGBoost	1.43	0.97	0.98
Random Forest	2.60	0.93	2.0
SVR	4.03	0.81	5.8

5 table. *Performance of three machine learning models—Random Forest, XGBoost and Support Vector Regression—in predicting individual hourly wages*

4.4 Predicting Gender Average Wages at the Occupational Level

4.4.1 Random Forest

The Random Forest model was first machine learning model which was applied to predict average wages at the occupational level separately for men and women. Since the model predicts salaries for men and women separately, the performance of the model was also evaluated separately for each gender.

6 table. illustrates the predictive performance of the Random Forest models estimated separately for men and women. The results indicate a high predictive accuracy for both genders. For men, the Random Forest model achieves an $R^2 = 0.93$ and quite high the absolute forecast error $RMSE = 3.966$. However, the mean absolute percentage error (MAPE) of 16.34% indicates low, but acceptable accuracy. For women, the predictive performance is slightly lower, with an $R^2 = 0.885$, but still remains high and achieving the absolute forecast error $RMSE = 2.031$. Similarly, for men, the MAPE result exceeds 10%, indicating that small discrepancies exist, but forecasts are generally reliable.

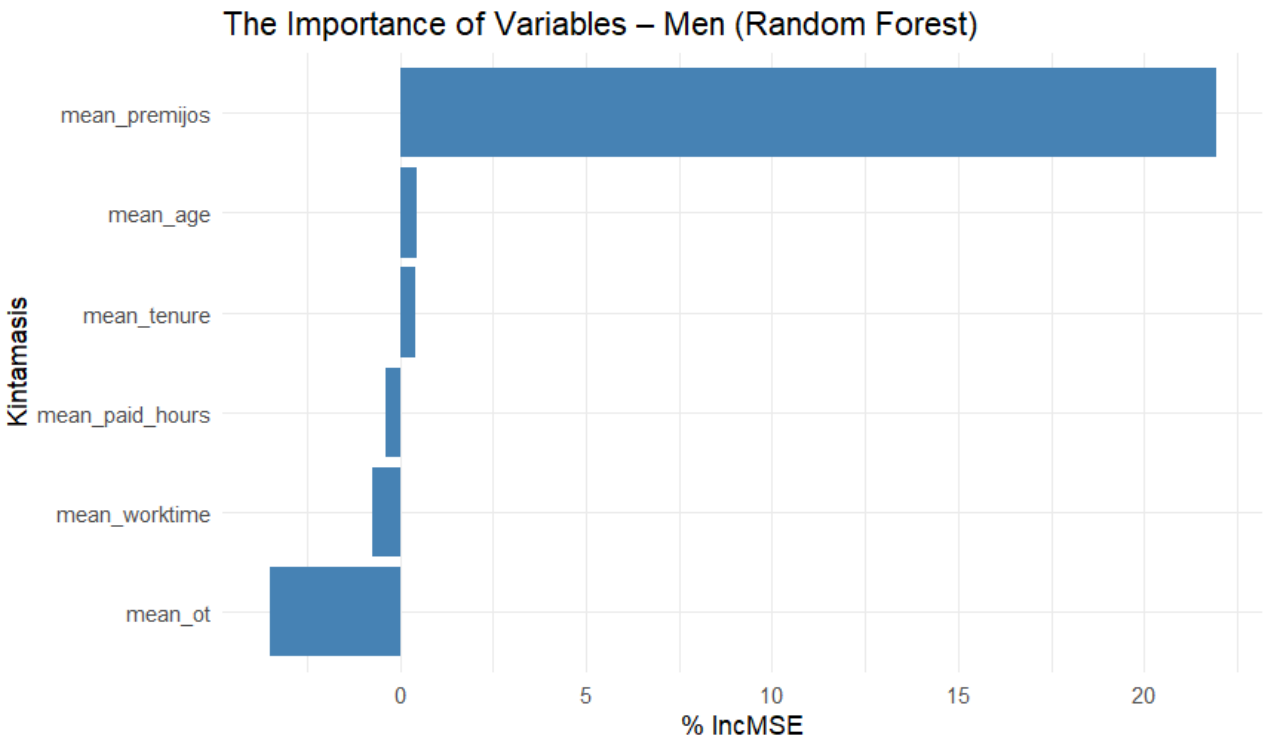
Gender	RMSE	R^2	MAPE (%)
Men	3.966	0.93	16.34
Women	2.031	0.885	11.336

6 table. *Random Forest model performance by gender*

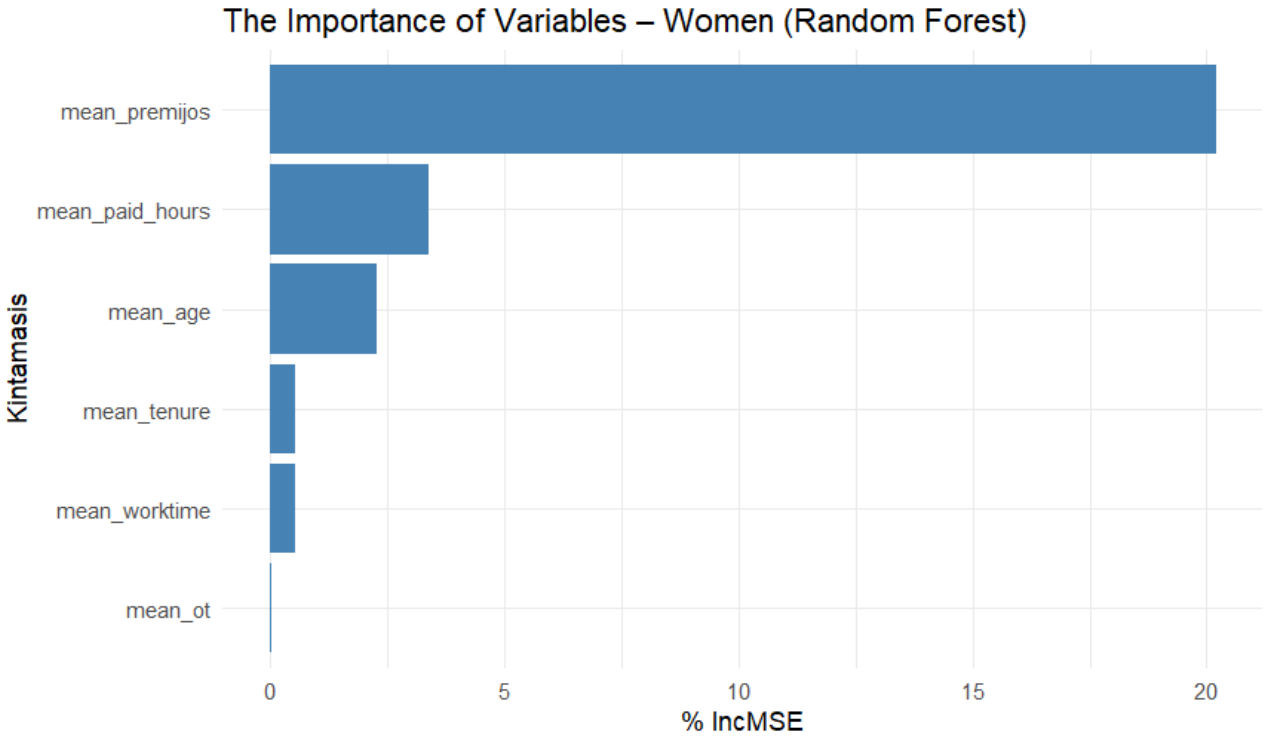
In 17 figure. and 18 figure. presents the importance of variables in Random Forest models, constructed separately for men and women, in predicting average wages at the occupational level. In both models, mean bonuses are most significant variable with the largest impact on wage predictions. This suggests that at the occupational level, the bonus structure is the main factor explaining differences in mean wages for both male and female groups. However, the importance of secondary variables differs by gender. In the male model, the influence of the remaining variables is very limited, and the importance of some variables (for example, overtime – mean ot) even acquires negative

values.

Additional Random Forest analysis at the occupation level showed that average wages in men's occupations are more predicted by structural characteristics than those in women's. This suggests that the structure of men's wages across occupations is more systematic, while the structure of women's wages is more heterogeneous.



17 figure. *The Importance of Variables – Men (Random Forest)*



18 figure. *The Importance of Variables – Women (Random Forest)*

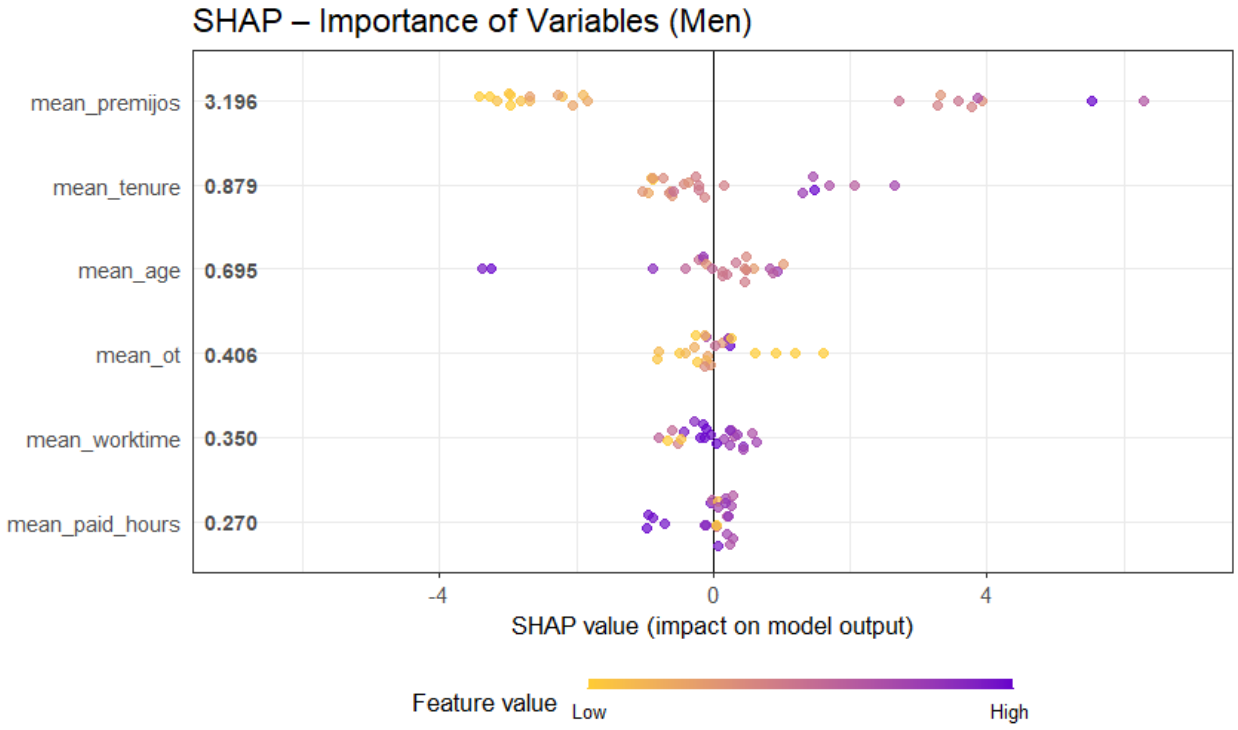
4.4.2 XGBoost

The XGBoost model achieved better model performance in predicting average salaries at the occupation level. The results provided in 7 table. shows that the XGBoost model has high prediction accuracy in both samples. The male model has $R^2 = 0.91$ and RMSE of 3.61, indicating that the model explains most of the variation in occupational-level wages. The coefficient of determination is slightly lower in the female model ($R^2 = 0.89$), but the prediction errors are significantly lower ($RMSE = 1.65$; $MAPE = 10.6\%$).

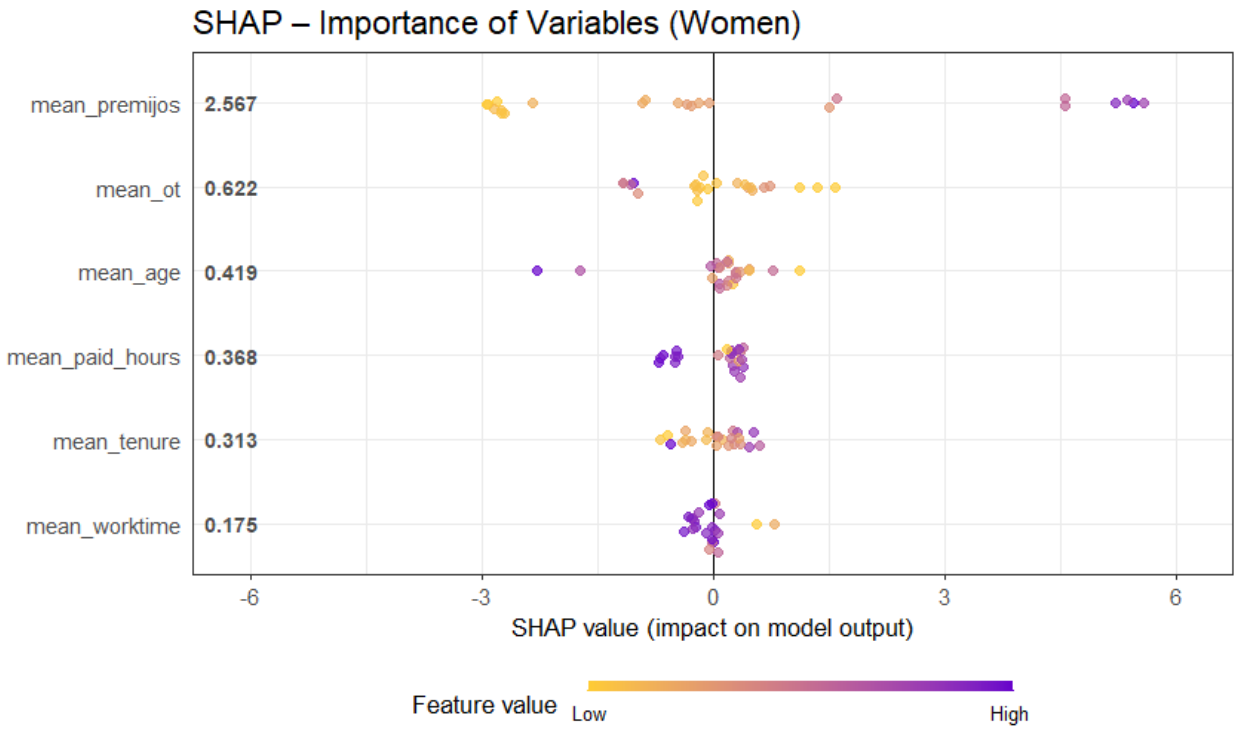
Gender	RMSE	R^2	MAPE (%)
Men	3.610	0.91	16.623
Women	1.648	0.89	10.614

7 table. *Random Forest model performance by gender*

SHAP values show not only the importance of variables, but also the direction and heterogeneity of the relationship, i.e. how different values of the variable increase or decrease the predicted salary. As in the case of the Random Forest model, for women and men, the main predictor is average bonuses, with average absolute values of SHAP of 2.567 and 3.20. For men, the average bonus is followed by length of service (0.88) and age (0.70), while for women, work intensity, measured by overtime (0.62) and paid hours (0.37), plays a relatively larger role than length of service (0.31) or age (0.42).These results imply that wage differentiation in male-dominated occupations is more strongly driven by bonuses and experience, whereas in female-dominated occupations it is more linked to actual working-time characteristics.



19 figure. *SHAP – Importance of Variables (Men)*



20 figure. *SHAP – Importance of Variables (Women)*

4.4.3 Support Vector Regression (SVR)

Support Vector Regression with a radial basis function (RBF) kernel was applied to predict average occupational wages using six aggregated explanatory variables. Model performance was eval-

uated using five-fold cross-validation.

As reported in Table 8 table., for men the optimal SVR specification is obtained at $C = 1$, achieving the lowest root mean squared error ($RMSE = 4.35$) and an associated coefficient of determination of $R^2 = 0.525$. meanwhile, for the female, Support Vector Regression model performance improves as the cost parameter C increases, indicating that low values of C lead to underfitting. The optimal specification is obtained at $C = 4$ with $\sigma = 0.181$, achieving the lowest $RMSE = 2.55$ and an associated coefficient of determination of approximately $R^2 = 0.68$.

Cost parameter (C)	RMSE	R^2	MAE
0.25	4.899	0.520	3.910
0.50	4.572	0.542	3.786
1.00	**4.347**	**0.525**	**3.599**
2.00	4.501	0.489	3.681
4.00	4.666	0.434	3.771
8.00	4.990	0.366	4.112
16.00	5.315	0.330	4.436
32.00	5.446	0.345	4.627
64.00	5.541	0.340	4.775
128.00	5.541	0.340	4.775

8 table. *Support Vector Regression (RBF kernel) tuning results for men*

Note: Five-fold cross-validation was used to select the optimal model based on the minimum RMSE. The kernel parameter was held constant at $\sigma = 0.181$.

Cost parameter (C)	RMSE	R^2	MAE
0.25	3.612	0.624	3.042
0.50	3.193	0.661	2.601
1.00	2.568	0.684	2.083
2.00	2.583	0.680	2.055
4.00	**2.545**	**0.680**	**1.957**
8.00	2.545	0.680	1.957
16.00	2.545	0.680	1.957
32.00	2.545	0.680	1.957
64.00	2.545	0.680	1.957
128.00	2.545	0.680	1.957

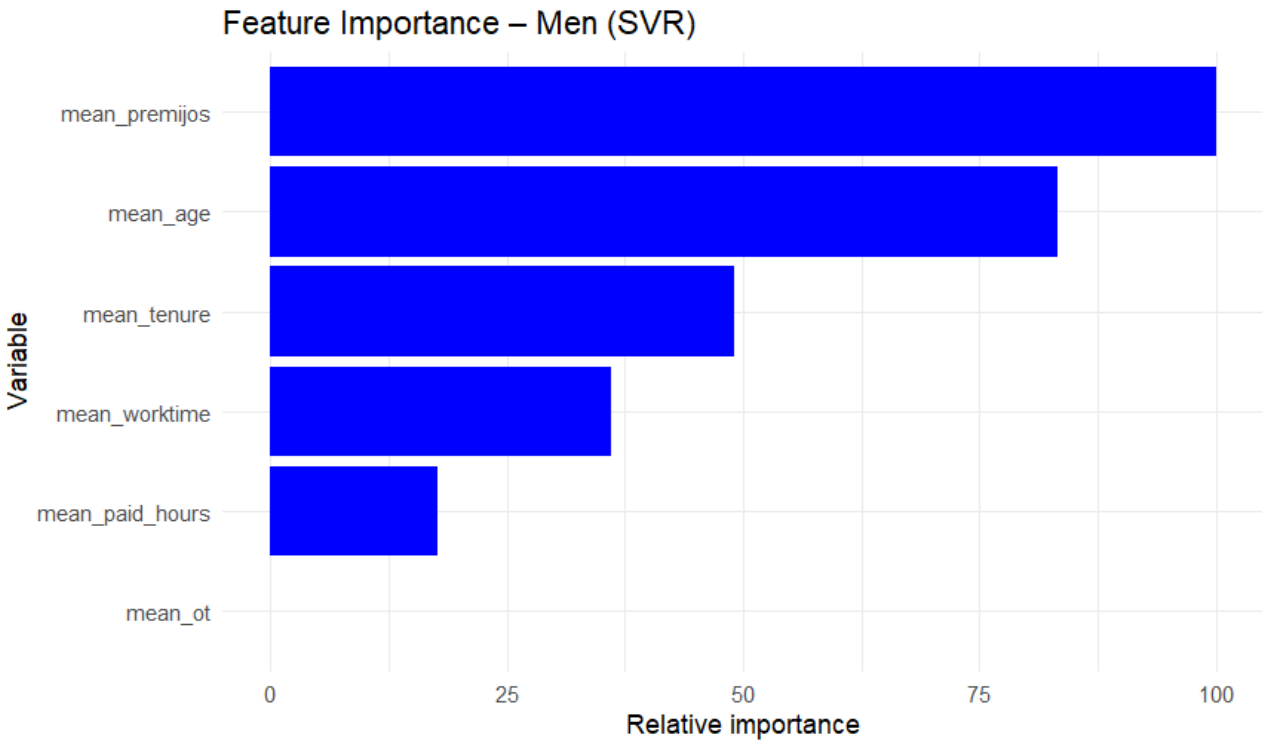
9 table. *Support Vector Regression (RBF kernel) tuning results for women*

Note: Five-fold cross-validation was used to select the optimal model based on the minimum RMSE. The kernel parameter was held constant at $\sigma = 0.181$.

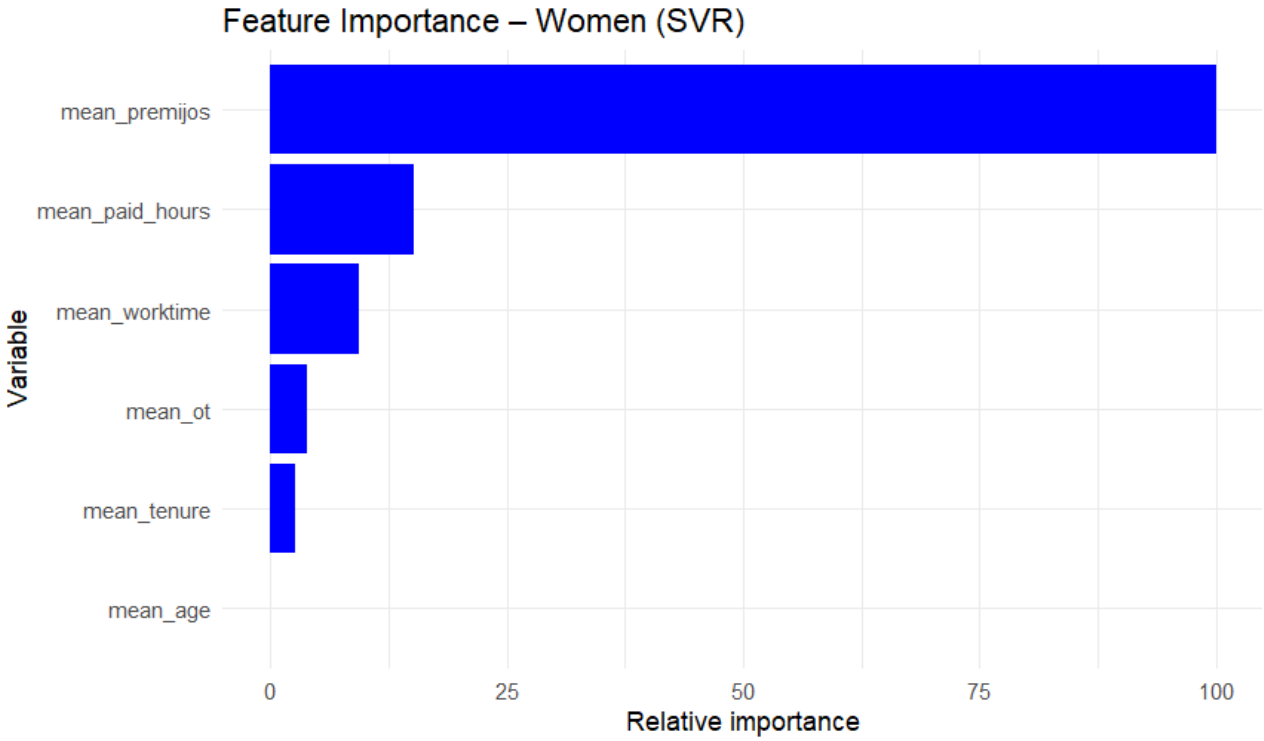
As with previous machine learning models, the SVR variable importance plots (21 figure. and 22 figure.) show that the most important variable in predicting average salaries for both men and women at the occupational level is average bonuses. For men, this is followed by average age and length of service. In contrast to men, variables related to working time, especially paid hours and

overtime, play a relatively greater role for women. This suggests that female occupational wages are more closely linked to actual labor input rather than experience.

Although the SVR model shows the worst performance, the results of the variable importance plots are consistent with the findings of the Random Forest and SHAP studies, supporting the conclusion that bonuses are the main determinant of wages for both genders, and that secondary factors differ significantly between men and women.



21 figure. *Feature Importance – Men (SVR)*



22 figure. *Feature Importance – Women (SVR)*

The performance of three predictive machine learning models: Random Forest, XGBoost and Support Vector Regression in predicting average hourly wages at the occupational level, separately for women and men, using aggregated occupational characteristics are represented in 10 table. For men, the Random Forest model achieved the best predictive accuracy ($R^2 = 0.93$) and the lowest prediction error among the evaluated methods. In contrast, the best performing model was XGBoost with $R^2 = 0.91$, for women. The Support Vector Regression model exhibited weaker predictive performance for both genders. The analysis of variable importance reveals that average education level, occupational structure, firm-related characteristics and bonus-related variables were significant in determining average wages for both women and men. Although, the set of important predictors is broadly similar across genders, their relative importance differs.

Model	Gender	RMSE	R^2	MAPE (%)
Random Forest	Men	3.966	0.930	16.34
	Women	2.031	0.885	11.34
XGBoost	Men	3.610	0.910	16.62
	Women	1.648	0.890	10.61
SVR	Men	4.347	0.525	3.59
	Women	2.545	0.680	1.96

10 table. *Performance of three machine learning models: Random Forest, XGBoost and Support Vector Regression—in predicting average hourly wages at the occupational level*

5 Conclusions

This study was designed to analyze the gender pay gap in Lithuania, based on the most relevant statistical wage survey data and to apply regression and predictive machine learning methods. The analysis consists of three main parts: to understand the gender pay gap, evaluate machine learning model performance to predicting individual hourly wages and average occupational wages based on occupational structural characteristics, for women and men separately. For this research we used the publicly available database of the Lithuanian "State Data Agency Statistical survey of wage structure". This database was collected in every four years-2014, 2018 and 2022 in October. Nevertheless, the main focus was on 2022 data, while it is the most relevant information about wages in Lithuania. By combining regression methods with modern machine learning techniques, the research aimed to both explain the gender pay gap and evaluated the predictive power of statistical and machine learning models at the individual and occupational levels.

The results of preliminary data revealed that the men's average hourly wage was 11.5% bigger than women's average hourly wage, in 2022. In several occupation the gender pay gap exceeded 20%, in 2022. For example, male database and network professionals earned as much as 31.4% more than women, which is 4.97 euros per hour, and sales, marketing and development managers 22%. Also, the results of two-way ANOVA confirmed that statistically significant wage differences between women and men across occupations exist. The significant relationship between gender and occupation reveals that the gender pay gap is heterogeneous and varies across professional groups, especially in high-skilled and managerial occupations.

In the first part of analysis, Multivariate regression analysis showed that, after controlling for observable characteristics such as education, work experience, firm size, occupation, working time and bonuses model can explains only 48.62% of wage variation, meaning that more than half of the variation remains unexplained. Additionally, when assessing the reliability of the multivariate Regression results, an analysis of the main assumptions of the model was performed. Although the residual histogram and Q-Q plot are approximately symmetrical and centered around zero, some deviations from normality were observed in the tails of the distribution. In the case of a large sample, such deviations are considered insignificant and do not have a significant impact on the coefficient estimates. The presence of heteroscedasticity was confirmed by the Breusch-Pagan test, and therefore heteroscedasticity-resistant standard errors were used. Also, Cook's distance analysis revealed that no observation had a disproportionately large impact on the model results. Quantile Regression, also, confirmed that the gender pay gap exists across the entire wage distribution and increases from lower to higher wage quantiles, reaching its highest values among top earners. At the 10th percentile, the estimated gap was 6 percent, while at the 90th percentile it increases to almost 18 percent. The Oaxaca–Blinder Decomposition provided additional insight into the structure of the wage gap. Only about one-fifth of the total wage difference could be explained by differences in characteristics, while the remaining majority remained unexplained.

In the second part of the study, which aims to evaluate the capabilities of machine learning models to predict individual hourly wages, three models were evaluated - Random Forest, XGBoost, and

Support Vector Regression. The best performing model was XGBoost with high $R^2 = 0.97$ value and low $RMSE = 1.43$. Random Forest also performs well with a little lower $R^2 = 0.93$ and bigger $RMSE = 2.60$, while Support Vector Regression shows the weakest predictive accuracy, $R^2 = 0.81$. The feature importance results showed that wage levels are primarily driven by structural job-related characteristics such as education, paid hours, firm size and bonuses, while gender itself played a relatively minor role in prediction accuracy, for all models.

The third part, in which main goal was to evaluate the accuracy of machine learning models to predict average wages for women and men separately at the occupational level, revealed that the best performing model for men was Random Forest with high $R^2 = 0.93$ value and for women XGBoost with $R^2 = 0.890$. However, $RMSE$ result for men was high 3.966. As in the second part, the support vector regression model appeared weakest for both men and women, $R^2 = 0.68$ for women and only $R^2 = 0.525$ for men. Differences in features importance across genders indicate that the determinants of wages may operate differently for women and men even within the same occupational structures. Men's wages are primarily driven by work intensity–related factors such as paid hours, overtime and bonuses, as well as firm size, whereas women's wages are more strongly associated with structural and human capital characteristics, including education level, tenure, age and working time arrangements.

Several limitations should be acknowledged. The study relies on the Statistical Survey of the Structure of Wages dataset, which does not include several potentially important determinants of wages. Factors such as career interruptions, parental leave history, negotiation behavior, job performance, informal responsibilities, or employer-specific pay-setting practices are not observed. Omitting these variables may contribute to a large unexplained component and limit the ability to fully separate discrimination from unobserved productivity-related factors. Also, data was collected at three discrete time points (2014, 2018 and 2022). Although these data allow for comparisons across years, they do not constitute a true panel dataset. As a result, the study cannot track individual wage trajectories over time.

Overall, although, this study has some limitations, but it confirms that a significant gender pay gap exists in Lithuania, despite the same characteristics of women and men. Machine learning models are able to predict individual hourly wages based on employee and company characteristics and predict average wages for women and men separately, at the occupation level. Regression models can only explain less than half of the gender pay gap, but provide useful insights.

References and sources

[1] A. Aiftincai. “An Exploratory Application of Machine Learning Algorithms in Estimating Net Salaries in Romania.” In: *Romanian Economic Journal* (2025). URL: https://rejournal.eu/sites/rejournal.versatech.ro/files/articole/2025-06-24/3779/aiftincai.pdf.

[2] K. Albæk, B. Thomsen. “Decomposing Wage Distributions on a Large Data Set – A Quantile Regression Analysis of the Gender Wage Gap.” In: *Working Paper* (2014). URL: https://pure.vive.dk/ws/files/232486/WP_06_2014.pdf.

[3] D. M. Blei, K. Vafa, S. Athey. “Estimating Wage Disparities Using Foundation Models.” In: (2025). URL: https://arxiv.org/abs/2409.09894.

[4] M. Brandwijk. “Analysing the gender pay gap in IT through salary prediction: a data driven approach.” Master’s thesis. Tilburg University, 2021. URL: https://arno.uvt.nl/show.cgi?fid=157118.

[5] L. Breiman. “Random Forests.” In: *Machine Learning* 45.1 (2001), pages 5–32. URL: https://link.springer.com/article/10.1023/A:1010933404324.

[6] Eurostat. *Gender pay gap statistics*. URL: https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Gender_pay_gap_statistics.

[7] Eurostat. *SDG 5 – Gender equality*. URL: https://ec.europa.eu/eurostat/statistics-explained/index.php?oldid=640502&title=SDG%5C_5%5C_-%5C_Gender%5C_equality.

[8] M. Hlavac. *oaxaca: Blinder-Oaxaca Decomposition in R*. 2022. URL: https://cran.r-project.org/web/packages/oaxaca/vignettes/oaxaca.pdf.

[9] IBM. *What is XGBoost?* IBM Think: Machine Learning Topic. 2025. URL: https://www.ibm.com/think/topics/xgboost.

[10] H. Y. Kim. “Statistical notes for clinical researchers: Two-way analysis of variance (ANOVA)—exploring possible interaction between factors.” In: (2014). URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC3978106/.

[11] A. S. Maudo. “The Gender Pay Gap in Spain: A Machine Learning Approach.” In: (2023). URL: https://addi.ehu.es/bitstream/handle/10810/62099/Ander_Sanchez_2023_TFM.pdf.

[12] U. Nations. *Gender Equality: Sustainable Development Goal 5*. URL: https://www.un.org/sustainabledevelopment/gender-equality/.

[13] J. E. Olson, K. Michell, W. Kristjanpoller. “Determining the gender wage gap through causal machine learning.” In: *Neural Computing and Applications* (2023). URL: https://link.springer.com/article/10.1007/s00521-023-08221-9.

[14] P. V. Patwa. “Unveiling Patterns in Employee Compensation: A Feature- Driven Analysis using Machine Learning Algorithms.” Master’s thesis. 2024. URL: https://esource.dbs.ie/server/api/core/bitstreams/4a40e069-c3e2-421b-a384-8abaa6a0de52/content.

[15] A. Sethi. *Support Vector Regression Tutorial for Machine Learning*. 2020. URL: https://www.analyticsvidhya.com/blog/2020/03/support-vector-regression-tutorial-for-machine-learning/.

[16] R. Sheposh. “Multiple regression.” In: *EBSCO Research Starters: Mathematics* (2025). URL: https://www.ebsco.com/research-starters/mathematics/multiple-regression.

[17] A. Strittmatter, C. Wunsch. “The Gender Pay Gap Revisited with Big Data: Do Methodological Choices Matter?” In: (2021). URL: https://docs.iza.org/dp14128.pdf.

[18] C. Tealdi, R. Forshaw, M. E. Schaffer, V. Iakovlev. “Using Machine Learning Methods to Estimate the Gender Wage Gap.” In: (2024). URL: https://pure.hw.ac.uk/ws/portalfiles/portal/91472316/Schaffer_TES2023.pdf.

[19] M. Töpfer, S. Briel. “The gender pay gap revisited: Does machine learning offer new insights?” In: *Labour Economics* (2020). URL: https://www.econstor.eu/bitstream/10419/213883/1/1689243643.pdf.

[20] S. Trivedi, S. Mishra. “A Quantile Regression Modelling Approach to Study Gender Wage Gap in India.” In: (2025). URL: https://csu.gov.cz/a-quantile-regression-modelling-approach-to-study-gender-wage-gap-in-india.

[21] UNESCO. *UNESCO Priority Gender Action Plan*. URL: https://amcow.ams3.cdn.digitaloceanspaces.com/resources/UNESCO%20Priority%20Gender%20Action%20Plan.pdf.

[22] S. A. Waisbrot, V. C. Edelsztein. “Breaking down the Gender Pay Gap through a Machine Learning Model.” In: *Economía: Teoría y Práctica* (2023). URL: https://www.scielo.org.mx/scielo.php?pid=S1405-14352023000100006&script=sci_arttext.

[23] A. Webster, F. Pastore, K. Meara. “The gender pay gap in the USA: a matching study.” In: *Journal of Population Economics* (2020). URL: https://link.springer.com/article/10.1007/s00148-019-00743-8.

[24] J. M. Zawia, L. S. Chin, M. A. Ismail. *Predictive Modelling of the Gender Pay Gap using Machine Learning*. Internal copy, downloaded. 2025.

Appendix 1. Use of artificial intelligence tools

The artificial intelligence tool ChatGPT (OpenAI, GPT-4.x version), which is freely available, was used in the preparation of this thesis. This tool was used for the following purposes:

- To generate initial ideas for the structure of the text.
- To generate a LaTeX tabular structure based on the provided data.
- To assist with LaTeX formatting, including table captions, labels, cross-references, and alignment.
- To improve language and text, including grammar and style improvements.
- To assist with program code formatting and syntactic corrections, without generating core algorithms or research logic.
- To upload code from Rstudio to Github.

All numerical results, analyses and interpretations were provided by the author. The AI tool did not generate the research results and was not used to draw conclusions. All AI generated input was reviewed and approved by the author.

Appendix 2. Code

The source code developed for this Master's thesis is publicly available at the following GitHub repository: `https://github.com/AmandaVilkonciute/Master-thesis`.