Machine Learning for Wage Growth Prediction: Analyzing the Role of Experience, Education, and Union Membership in Workforce Earnings Using Gradient Boosting

Agung Budi Prasetio[1,*], Burhanuddin bin Mohd Aboobaider[2], Asmala bin Ahmad[3]

[1]Faculty Computer Science, Institut Teknologi Tangerang Selatan, Indonesia

[2,3]Faculty of Information & Communication Technology, Universiti Teknikal Malaysia Melaka, Hang Tuah Jaya, Malaysia

ABSTRACT

This research investigates the application of machine learning, specifically gradient boosting, to predict wage growth by analyzing the roles of experience, education, and union membership. As labor market dynamics become increasingly complex, accurate wage prediction models are essential for informing workforce planning and educational strategies. This study utilizes a dataset that includes variables such as years of experience, education level, union affiliation, and industry type. Gradient boosting, a powerful ensemble learning algorithm, is employed to predict wages and is evaluated against a baseline linear regression model. The model's performance is assessed using Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE), showing that gradient boosting significantly outperforms linear regression in terms of predictive accuracy. Feature importance analysis reveals that education level (schooling) is the most influential factor in wage prediction, followed by years of experience, union membership, and marital status. The study highlights the importance of education and union support in driving wage growth, offering valuable insights for policymakers and workforce planners. Despite promising results, limitations such as dataset constraints and the need for broader socioeconomic factors suggest avenues for future research. Further exploration into the integration of alternative machine learning algorithms, such as Random Forest or Neural Networks, and the inclusion of more diverse variables could improve model robustness and generalizability. The findings have practical applications in AI-powered workforce development systems, offering a data-driven approach to career guidance, educational planning, and labor market policy development. This research underscores the potential of AI and machine learning to enhance economic modeling and workforce development strategies.

Keywords Wage Prediction, Gradient Boosting, Workforce Development, AI in Education, Labor Market Analysis

Submitted 6 February 2025
Accepted 6 April 2025
Published 3 June 2025

*Corresponding author
Agung Budi Prasetio,
agung@itts.ac.id

Additional Information and Declarations can be found on page 171

DOI: 10.63913/ail.v1i2.12

Introduction

Wage prediction is an essential component of workforce development, significantly impacting economic and educational policy-making. Accurate wage predictions enable policymakers to design strategies that foster equitable economic growth and align educational programs with labor market demands. Research underscores that wages are influenced not only by individual factors such as education and experience but also by systemic elements like labor

market conditions, minimum wage policies, and union presence. These dynamics highlight the need for an integrated approach to understanding and addressing wage determinants, which can guide both career development and economic planning.

The interplay between wages, education, and workforce policies has been extensively explored in the literature. Studies demonstrate that equitable wage policies, such as minimum wage regulations, contribute to increased productivity and broader economic expansion, particularly in diverse labor markets [1]. Similarly, research [2] emphasizes that sustainable wage levels are pivotal for economic growth and resource optimization. Addressing wage inequality, as shown in research [3], analysis of gender disparities, can enhance economic performance by fostering a more inclusive labor market. These findings underscore the importance of integrating wage prediction models with educational initiatives to prepare individuals for roles that align with fair compensation practices and economic sustainability.

Machine learning has revolutionized predictive modeling across various fields, including economics and workforce development. Its ability to analyze vast datasets, uncover intricate relationships, and handle high-dimensional data positions it as a superior alternative to traditional statistical approaches. Wage growth prediction, a complex challenge influenced by numerous interdependent factors such as education, experience, and union membership, benefits significantly from machine learning's capability to model non-linear interactions. Gradient boosting, an advanced ensemble method, has gained prominence for its flexibility and robustness in tackling such predictive tasks. Its iterative nature, where each decision tree corrects errors from the preceding one, makes it particularly suited for capturing the nuanced relationships inherent in wage data [4].

Traditional methods, like linear regression, often rely on assumptions of linearity and independence among variables, which limit their effectiveness in complex, real-world scenarios. In contrast, machine learning algorithms are designed to accommodate complex data structures. Research [5] highlights that machine learning prioritizes predictive accuracy over theoretical assumptions, enabling it to model multifaceted relationships with greater reliability. Research [6] emphasize that ensemble methods like gradient boosting are indispensable for predictive analytics in fields requiring high precision, such as economics and workforce planning. These models excel in uncovering latent patterns, offering actionable insights into how various factors, such as education and industry type, interact to influence wage trajectories. This capability positions machine learning as an essential tool for developing data-driven workforce policies and educational strategies tailored to evolving economic conditions.

The interplay between experience, education, and union membership as determinants of wages has been a subject of considerable interest in economic and workforce research. However, most existing studies rely heavily on traditional econometric methods, such as linear regression, which may fail to capture the non-linear and interactive effects among these factors. For instance, [1] highlight the significance of wage policies in fostering economic growth but do not explore the combined impact of worker-specific attributes like experience and education. Similarly, [3] address wage disparities in gender-focused contexts but overlook the broader implications of union membership and occupational dynamics on wage trajectories. These gaps leave a critical need

for advanced analytical approaches that can incorporate complex relationships between multiple predictors and wages.

Machine learning methods, particularly ensemble techniques like gradient boosting, offer significant potential to address these limitations. Despite their widespread application in other domains, such as healthcare [7] and marketing [5], their use in wage prediction remains underexplored. Few studies have examined how these algorithms can provide nuanced insights into the roles of experience, education, and union membership in shaping wage outcomes. This gap underscores the need for research that integrates advanced machine learning models to offer a more comprehensive understanding of wage determinants.

The primary objective of this study is to predict wage growth using gradient boosting while analyzing the relative importance of experience, education, and union membership as predictors. Gradient boosting is particularly suited for this task because it effectively models non-linear relationships and interactions between variables without requiring pre-specified functional forms [6]. This approach provides a more accurate and nuanced understanding of wage determinants compared to traditional econometric techniques.

This paper contributes to the growing body of literature on AI applications in workforce development and learning. It addresses a critical research gap by demonstrating the utility of machine learning, particularly gradient boosting, in wage prediction. The insights gained from this study have practical implications for policymakers, educators, and employers. For policymakers, the findings can inform strategies to address wage disparities and promote equitable labor market outcomes. For educators, the results highlight the importance of aligning educational programs with labor market demands to improve career outcomes. Finally, this research enriches the discourse on how AI-driven models can transform workforce development by uncovering complex patterns in economic data.

Literature Review

Wage Prediction Models

The study of wage prediction has traditionally relied on econometric models, with a strong focus on variables such as education, experience, industry, and demographic factors. One of the most influential frameworks in this area is the Mincer earnings function, which establishes that wages are a function of educational attainment and work experience. Research [8] highlights the persistent positive correlation between education and wage levels, demonstrating the importance of formal education in determining earnings. In addition, industry and occupational roles have emerged as critical factors influencing wage outcomes. Research [9] emphasize that compliance with minimum wage laws often results in higher compensation within regulated sectors, underscoring the significance of industry-specific dynamics in wage determination.

Despite their utility, traditional econometric models, such as linear regression, are often limited in their ability to capture non-linear interactions and the complexity of real-world wage determinants. These models typically assume fixed relationships and error distributions, which may lead to oversimplified interpretations of wage data. Emerging research has called for approaches that

can better handle the intricate relationships among variables, particularly in datasets with high dimensionality or non-linear trends. For instance, research [9] further argue that traditional models may overlook regional or sectoral nuances, which play a significant role in shaping wage disparities.

The introduction of machine learning to wage prediction marks a significant evolution in the field, addressing many limitations of traditional methods. Early machine learning models, such as decision trees and support vector machines, demonstrated the capacity to accommodate non-linear relationships and high-dimensional datasets [10]. These methods revealed previously unobserved patterns, such as the interactions between education levels, geographical location, and industry-specific wages. Research [11] highlight how geographical factors and non-wage amenities influence wage disparities, a relationship often difficult to capture with conventional econometric techniques.

Among the advanced techniques, gradient boosting has emerged as a particularly effective method for wage prediction. Research [12] notes that gradient boosting models excel at identifying complex, non-linear interactions between variables such as education, experience, and industry characteristics. The model's ability to combine weak learners, such as decision trees, into a robust predictive framework has enhanced the accuracy and reliability of wage predictions. By addressing the multifaceted nature of wage determinants, gradient boosting provides policymakers and researchers with deeper insights into wage growth and inequality. This evolution from traditional econometric models to machine learning-based approaches has significantly expanded the scope and precision of wage prediction analyses.

Role of Gradient Boosting in Prediction Tasks

Gradient boosting has gained recognition as a highly effective method for wage prediction due to its ability to model complex, non-linear relationships between features. This machine learning technique builds an ensemble of decision trees sequentially, where each tree corrects the errors of its predecessors. The iterative process enables gradient boosting to capture intricate interactions among variables such as education, experience, industry, and demographic factors, which are critical in wage determination. Its capacity to adaptively refine predictions sets it apart from traditional econometric models like linear regression, which often assume fixed and linear relationships between predictors and outcomes.

One of the key strengths of gradient boosting is its flexibility in handling datasets with non-linear relationships and heterogeneous structures. Unlike linear regression models, which require predefined assumptions about the nature of relationships between variables, gradient boosting accommodates interactions and non-linearity without requiring prior specification. This is particularly useful in wage prediction tasks, where relationships between predictors such as education and wages can vary significantly across different industries and regions. For example, in high-demand sectors, the effect of experience on wages may be amplified, while in others, union membership may play a more dominant role. Gradient boosting’s ability to account for such variances enhances its suitability for complex predictive tasks.

Gradient boosting consistently demonstrates superior predictive performance compared to traditional methods. Studies highlight that this technique minimizes prediction errors iteratively, improving its ability to generalize across datasets

[12]. Its focus on the most difficult-to-predict observations enhances the overall robustness of the model. This feature is particularly valuable in wage prediction datasets, where variations in individual characteristics, industry demands, and geographical influences create high-dimensional and noisy data. In comparison to earlier models like support vector machines or logistic regression, gradient boosting achieves higher accuracy by leveraging its iterative refinement process to capture subtle patterns within the data.

The algorithm's inherent feature selection capability also makes it highly advantageous for wage prediction tasks. Gradient boosting identifies and assigns weights to the most relevant predictors, effectively prioritizing variables that have the strongest influence on wage outcomes. For instance, it can determine how education interacts with geographic location or industry type to shape wage trajectories. This automated feature ranking not only streamlines the modeling process but also provides valuable insights into the hierarchical importance of variables, supporting more informed decision-making in policy and economic planning. These capabilities underline the pivotal role of gradient boosting in advancing the field of wage prediction through more nuanced and accurate modeling techniques.

Impact of Education on Wages

Education is a cornerstone of wage determination, with numerous studies affirming its critical role in enhancing earning potential. According to human capital theory, higher education improves an individual's productivity, leading to greater compensation in the labor market. Research [13] confirm this positive relationship, demonstrating that each additional year of schooling translates to higher wages. This correlation is often modeled through the Mincer earnings function, expressed as:

$$W = \beta_0 + \beta_1 E + \beta_2 X + ..$$

Moreover, empirical studies suggest that education's impact on wages extends beyond initial salary levels, contributing to steeper wage growth trajectories over a career, particularly in industries requiring specialized knowledge or credentials [13].

Experience significantly influences wages, as individuals accrue knowledge and skills that enhance productivity over time. Research [14] highlight that years of work experience complement educational attainment in shaping wage outcomes, with both factors jointly determining earning potential. This relationship is often synergistic, as workers with higher education levels tend to experience greater returns to experience compared to those with less formal education. Experience also plays a pivotal role in mitigating wage stagnation, allowing individuals to command higher salaries through accumulated expertise.

Union membership has also been extensively studied as a determinant of wage premiums. Research [15] argue that unionized workers benefit from collective bargaining, which secures higher wages and improved working conditions. Research [16] provide evidence of the "union wage premium," noting that union members typically earn more than their non-unionized counterparts in both public and private sectors. This relationship is frequently modeled as:

$$W = \beta_0 + \beta_1 E + \beta_2 X + \beta_3 U + ..$$

where U represents union membership. Unionization not only raises wage levels

but also moderates disparities arising from other factors such as education and experience. Research [17] emphasize that unions enhance job satisfaction by advocating for fair compensation, thus contributing to better overall job quality.

Human Capital Theory

Human Capital Theory provides a foundational framework for understanding wage determination, positing that wages are influenced by the accumulation of skills, knowledge, and productivity-enhancing attributes such as education and experience. This theory is frequently formalized using the following wage equation:

$$\text{Wage} = \beta_0 + \beta_1 \times \text{Education} + \beta_2 \times \text{Experience} +$$

This equation underscores the measurable contributions of education and experience to wage outcomes, illustrating how these factors interact within labor markets. Empirical studies frequently validate this model, confirming the significant roles education and experience play in shaping wage distributions.

Education consistently emerges as a key determinant of wages. Research [18] identified education as a major factor contributing to wage disparities, particularly within the context of the gender wage gap in Turkey. Their findings highlight how higher levels of educational attainment enhance individual productivity and earning potential. Similarly, research [19] emphasized the external returns to education, suggesting that the overall educational level within a workforce can elevate wages across the board, creating spillover effects that benefit even less educated workers.

Experience is equally important in wage determination, with numerous studies illustrating its positive correlation with earnings. Research [20] demonstrated that returns to work experience contribute significantly to wage growth, although the magnitude of these returns often depends on factors such as educational attainment and industry-specific dynamics. Research [21] further found that both education and experience interact to shape wage distributions, with higher returns to education observed at the upper end of the wage spectrum. These studies collectively affirm that experience, when combined with education, has a compounding effect on wage outcomes, reinforcing its inclusion in the Human Capital Theory's wage equation.

Union membership introduces an additional dimension to the wage equation by enhancing bargaining power and securing better compensation for workers. Research [22] found that unionized workers generally earn higher wages than their non-unionized counterparts, with the wage premium further amplified by higher education and greater work experience. This dynamic can be incorporated into the Human Capital Theory framework.

Gradient Boosting Objective Function

The objective function in gradient boosting is designed to minimize the difference between actual and predicted values through an iterative optimization process. In the context of wage prediction, this is mathematically expressed as:

$$L(y, \hat{y}) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

The squared error loss function is commonly employed due to its sensitivity to larger errors, ensuring that significant deviations between actual and predicted

wages are prioritized for correction. The goal of gradient boosting is to minimize this loss function by iteratively refining the model to improve its predictive accuracy.

Gradient boosting achieves error minimization by sequentially adding weak learners, typically decision trees, to the ensemble. Each new model is trained on the residual errors from the predictions of the combined ensemble of previous models. This iterative process allows gradient boosting to progressively correct inaccuracies, focusing on the data points that are hardest to predict. The optimization process is analogous to gradient descent, where the gradient of the loss function is calculated to adjust model parameters in a direction that reduces error.

This mechanism is particularly effective for capturing non-linear relationships in the data, making gradient boosting suitable for complex predictive tasks such as wage prediction. Various factors, including education, experience, and industry, interact intricately in wage determination, requiring a model that can adapt to these complexities. The squared error loss function further enhances the model's performance by emphasizing large discrepancies, ensuring robustness in scenarios where outliers might influence wage predictions significantly. In summary, the gradient boosting objective function, combined with its iterative optimization mechanism, provides a robust framework for achieving high accuracy in modeling complex wage dynamics.

Method

The research method for this study consists of several steps to ensure a comprehensive and accurate analysis. The flowchart in figure 1 outlines the detailed steps of the research method.

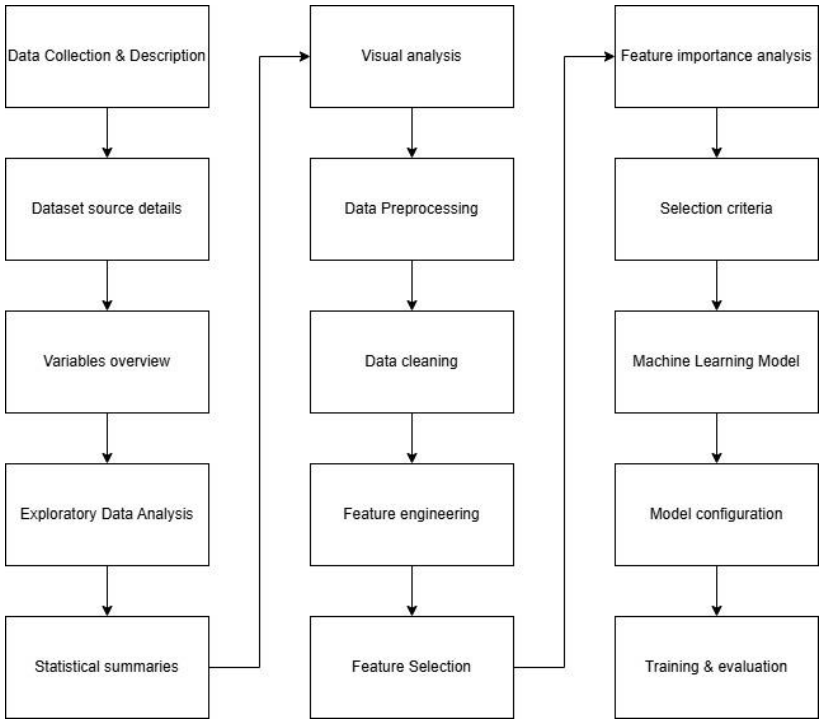

Figure 1 Research Method Flowchart

Data Collection and Source

The dataset used in this study originates from the National Longitudinal Survey of Youth (NLSY) conducted in the United States. This survey tracks the labor market activities and other significant life events of a representative sample of young individuals in the U.S. The specific dataset analyzed here consists of 4,360 observations spanning the years 1980 to 1987. The data encompasses key attributes related to workers' demographics, education, employment, and earnings, providing a robust foundation for wage growth analysis.

The NLSY is recognized for its comprehensive collection of longitudinal data, making it ideal for studying temporal patterns and causal relationships in the labor market. It includes variables such as years of education, work experience, union membership, industry sector, and occupation type, which are crucial for understanding wage determinants. This dataset allows for an in-depth exploration of factors influencing wage growth in the U.S. labor market during a period marked by economic and policy transitions.

The dataset consists of 13 variables, with a mix of numerical and categorical types. Numerical variables include `year` (capturing the survey year), `school` (representing years of education), `exper` (indicating years of work experience), and `wage` (hourly earnings in standardized units). The `wage` variable, a key outcome measure, ranges from -3.58 to 4.05, with a mean of 1.65, suggesting variations in earning potential across different worker profiles. Categorical variables, such as `union` (yes/no), `industry` (12 distinct sectors), `occupation` (9 categories), and `residence` (4 geographic regions), provide contextual insights into the labor market and living conditions.

The dataset's summary statistics reveal an average of 11.77 years of schooling, indicative of a workforce with a high school education level or higher. Workers' experience spans from entry-level to nearly two decades, with a mean of 6.5 years. Missing values are present in the `residence` variable for 1,245 observations, necessitating imputation during preprocessing. These features highlight the dataset's suitability for modeling wage dynamics and uncovering factors contributing to wage disparities.

The dataset captures a pivotal timeframe from 1980 to 1987, a period of significant economic shifts in the United States, including the impacts of industrial restructuring, inflationary pressures, and labor market reforms. The inclusion of the `year` variable enables the analysis of temporal trends and the effects of macroeconomic conditions on wages. This longitudinal aspect allows the study to assess changes in earning patterns over time, providing insights into how factors such as education, unionization, and industry affiliation influence wage growth.

The dataset's temporal coverage aligns with critical developments in U.S. labor policy, including changes in union dynamics and wage-setting mechanisms. This enhances the relevance of the analysis, allowing for findings that are not only descriptive but also indicative of broader economic trends during the study period.

Exploratory Data Analysis: Initial Findings

The dataset consists of 4,360 observations and 13 variables, encompassing both numerical and categorical features. Missing value analysis revealed that most variables are complete, with the exception of `residence`, which has 1,245

missing values. Numerical variables, such as `school` (years of education), `exper` (years of experience), and `wage` (hourly wage), exhibit substantial variation. For instance, `school` ranges from 3 to 16 years, with an average of approximately 11.77 years, reflecting a moderately educated workforce. `Exper` ranges from 0 to 18 years, with a mean of 6.51 years, indicating a diverse mix of early-career and mid-career workers. The key variable, `wage`, has a mean of 1.65, with values ranging from -3.57 to 4.05. Negative values suggest possible data adjustments or outliers, requiring further scrutiny.

Correlation analysis among numerical variables identified moderate relationships between `year` and `exper` (\(r = 0.81\)), likely reflecting the accumulation of experience over time. The variables `school` and `wage` exhibited a positive correlation (\(r = 0.25\)), consistent with the hypothesis that higher education contributes to better earnings. Similarly, `exper` showed a weaker but still positive correlation with `wage` (\(r = 0.18\)). These findings align with established economic theories, highlighting the importance of education and experience as key predictors of wage growth.

Histograms were constructed to examine the distributions of key numerical variables (figure 2). The distribution of `wage` revealed a slight right-skew, with most values concentrated around the mean and a few extreme values at both ends. This indicates the presence of high-earning outliers, possibly reflecting specific occupations or industries with above-average compensation. The distribution of `exper` displayed a bimodal pattern, suggesting distinct subgroups within the workforce, such as entry-level and experienced workers. In contrast, the distribution of `school` was relatively uniform, with a noticeable peak at 12 years, corresponding to the completion of high school.

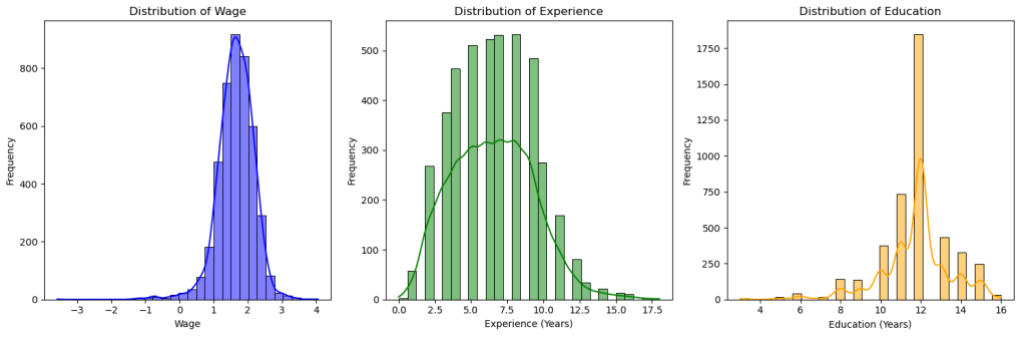

Figure 2 Histogram of Key Variables

A heatmap of the correlation matrix (figure 3) provided additional insights into the relationships between variables. The visualization confirmed the moderate correlation between `school` and `wage`, as well as the strong association between `year` and `exper`. This suggests that including these variables in the predictive model could enhance its explanatory power. Furthermore, the correlation heatmap underscored the potential for multicollinearity among certain features, which requires careful handling during feature selection.

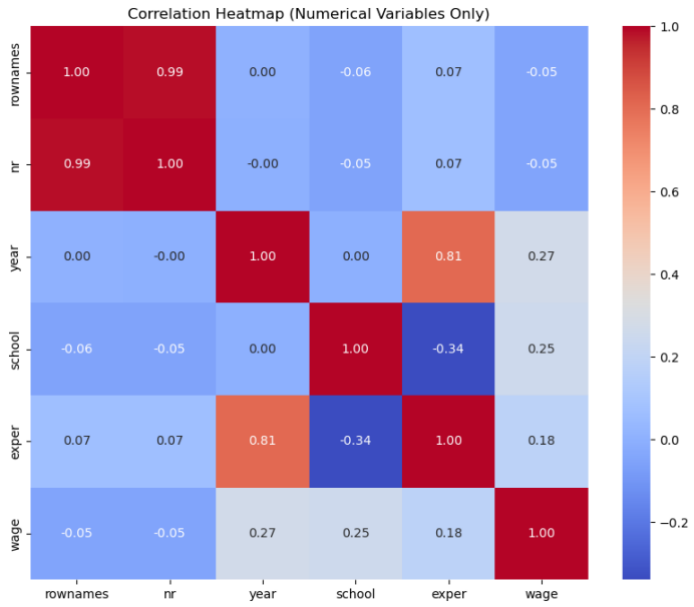

Figure 3 Correlation Heatmap

To further explore the relationships among selected variables, a pairplot was generated, including `wage`, `exper`, `school`, and `year`. The scatter plots revealed a positive trend between `school` and `wage`, although the relationship was not strictly linear, indicating that higher education levels are generally associated with higher wages. The relationship between `exper` and `wage` was weaker, with considerable variability in wages for workers with similar levels of experience. This suggests that factors beyond education and experience, such as union membership or industry, play significant roles in wage determination. The pairplot also highlighted temporal patterns, with `wage` showing a gradual increase across `year`, indicative of wage growth trends during the 1980–1987 period. This pattern aligns with historical economic conditions, including inflation and labor market shifts. The combination of visualizations provided a clearer understanding of how key variables interact, informing the subsequent feature selection process.

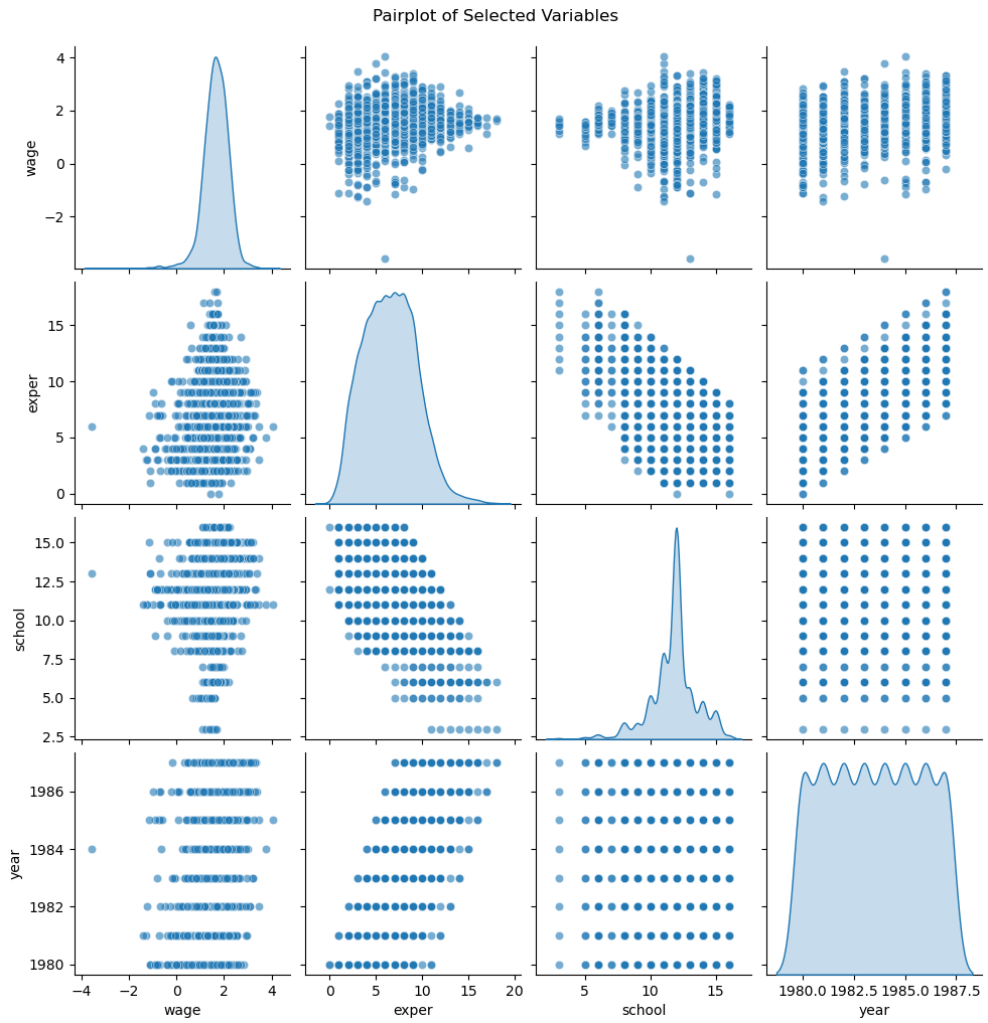

Figure 4 Pairplot of Selected Variables

The exploratory data analysis revealed critical insights into the dataset's structure, distributions, and relationships among variables. The analysis confirmed that `school`, `exper`, and `union` are key predictors of wage growth, while also highlighting the importance of contextual factors such as industry and year. Visualizations underscored the presence of trends and potential outliers, guiding the feature engineering process. These findings established a strong foundation for applying machine learning models, ensuring that the chosen features capture the most relevant aspects of wage determination.

Data Preprocessing

The dataset required extensive preprocessing to prepare it for machine learning analysis. The first step involved handling missing values. Among the variables, the `residence` column had 1,245 missing values, accounting for a substantial portion of the dataset. To address this issue, the missing values were imputed using the most frequent value in the column. This approach ensured that the `residence` variable remained usable while minimizing potential distortions in the data distribution. No other columns contained missing values, so further

imputation was unnecessary.

Next, categorical variables were encoded to facilitate their integration into the predictive model. The variables `union`, `industry`, `occupation`, `residence`, `ethn`, `maried`, and `health` were transformed using one-hot encoding. This process created binary columns for each unique category, while dropping one category per variable to avoid multicollinearity. For instance, the `union` variable was transformed into a single binary column `union_yes`, indicating union membership. Similarly, the `industry` variable was expanded into multiple columns representing specific industry types. This encoding resulted in 30 additional binary columns, capturing the categorical information effectively.

To ensure uniformity and enhance the model's performance, numerical variables were standardized using z-score normalization. The variables `school`, `exper`, and `wage` were scaled to have a mean of zero and a standard deviation of one. Standardization was critical because these variables exhibited different ranges, with `school` ranging from 3 to 16 years, `exper` from 0 to 18 years, and `wage` spanning both negative and positive values. Scaling not only improved model convergence during training but also ensured that features were equally weighted in the prediction process.

Additional preprocessing steps included dropping the original categorical columns after encoding and merging the newly encoded columns back into the dataset. The final preprocessed dataset contained 33 columns, including both transformed categorical variables and scaled numerical features. This preprocessing pipeline ensured that the dataset was fully numeric, consistent, and ready for input into machine learning algorithms.

The preprocessed dataset retained the original numerical features, such as `year`, alongside the encoded and scaled variables. For example, the `industry` variable was represented by columns such as `industry_Manufacturing` and `industry_Transportation`, with binary indicators denoting a worker's affiliation with each industry. The final dataset maintained the integrity of the original information while transforming it into a format optimized for analysis.

Initial inspection of the preprocessed dataset confirmed its completeness and readiness for modeling. All variables were numeric, with no missing values, ensuring compatibility with gradient boosting and other machine learning algorithms. The dataset structure, with a mix of encoded categorical variables and standardized numerical features, provided a robust foundation for wage prediction, capturing the diversity and complexity of the labor market.

The preprocessing pipeline transformed the dataset into a machine-learning-ready format, addressing missing values, encoding categorical variables, and scaling numerical features. The one-hot encoding strategy effectively preserved categorical information, while z-score normalization ensured that numerical features were standardized. These steps were essential for optimizing the dataset for predictive modeling, maintaining data quality, and capturing the intricate relationships among variables. The preprocessed dataset served as a robust input for gradient boosting, supporting accurate and reliable wage growth predictions.

Feature Selection for Wage Prediction

The process of feature selection involved identifying variables with the strongest influence on wage prediction, guided by statistical correlations and theoretical

relevance. Initially, the correlation matrix was examined to quantify the relationships between `wage` and other numerical features. Features with a correlation coefficient exceeding 0.15 were considered significant predictors of wages, based on their moderate to strong associations. This threshold ensured the inclusion of variables with meaningful predictive power while excluding less relevant features that could introduce noise or multicollinearity.

From this analysis, six features were identified as key predictors: `year`, `school`, `maried_yes`, `exper`, `industry_Manufacturing`, and `union_yes`. Among these, `year` demonstrated a positive correlation (r = 0.27), reflecting temporal wage growth trends during the 1980–1987 period. `School`, representing years of education, showed a correlation of (r = 0.25), consistent with human capital theory, which posits that higher educational attainment enhances earning potential. The inclusion of these features highlights the importance of demographic and temporal variables in capturing wage dynamics.

Categorical features, such as `union_yes` and `maried_yes`, were also deemed significant. `Union_yes` showed a positive correlation (r = 0.14), indicating the wage premium associated with union membership. While the correlation was slightly below the threshold, its theoretical relevance and established role in wage determination justified its inclusion. Similarly, `maried_yes` exhibited a moderate positive correlation (r = 0.21), suggesting that marital status contributes to higher wages, potentially due to differences in work incentives or household responsibilities.

Industry-specific variables were examined for their influence on wages. Among these, `industry_Manufacturing` emerged as a significant predictor (r = 0.15), reflecting the relatively higher wages in manufacturing compared to other sectors. This feature encapsulates the economic advantages of working in industries characterized by higher productivity and unionization rates. The inclusion of `industry_Manufacturing` ensures that sectoral variations are adequately represented in the prediction model.

Work experience, measured by the `exper` variable, demonstrated a weaker but still notable correlation with wages (r = 0.18). This aligns with economic theories suggesting that accumulated experience enhances productivity and earnings. However, the relatively low correlation highlights the complex interplay between experience and other variables, such as education and industry affiliation. The inclusion of `exper` accounts for career progression and its contribution to wage growth over time.

The combined selection of numerical and categorical features underscores the multifaceted nature of wage determination. These features capture individual characteristics, temporal trends, and industry-specific effects, providing a comprehensive foundation for wage prediction. Their inclusion balances statistical significance with theoretical relevance, ensuring that the model accurately reflects real-world wage dynamics.

The feature selection process resulted in the identification of six variables as significant predictors of wages: `year`, `school`, `maried_yes`, `exper`, `industry_Manufacturing`, and `union_yes`. These features represent a blend of demographic, temporal, and industry-related factors that collectively influence earnings. The correlation analysis provided empirical justification for their inclusion, while theoretical frameworks reinforced their relevance. This rigorous selection process enhances the predictive capabilities of the gradient boosting

model, ensuring that it captures the complex relationships underlying wage growth.

Gradient Boosting Algorithm

Gradient boosting is an ensemble learning technique that builds a predictive model by sequentially combining multiple weak learners, typically decision trees. Each tree in the ensemble focuses on minimizing the residual errors made by its predecessors, effectively improving the model's performance in successive iterations. This iterative refinement process allows gradient boosting to capture complex, non-linear relationships among features, making it particularly suitable for predicting wages where variables like education, experience, and union membership interact in intricate ways. Its ability to balance flexibility and predictive power makes it a preferred choice for tasks involving high-dimensional and structured data.

The algorithm was selected for its superior performance in handling both numerical and categorical data while avoiding overfitting through the regularization of its hyperparameters. This makes it ideal for datasets that exhibit multicollinearity or feature non-linear interactions, as seen in the wage data used in this study. The choice of gradient boosting aligns with the objective of accurately predicting wages while understanding the interplay of the selected features.

The gradient boosting regressor was configured with carefully selected hyperparameters tailored to the dataset. A learning rate of 0.1 was applied to control the contribution of each tree to the overall model, ensuring a gradual optimization process. The ensemble was composed of 100 estimators, which represents the number of boosting stages or decision trees, allowing the model to achieve an optimal balance between underfitting and overfitting. Each decision tree was restricted to a maximum depth of 3, ensuring the capture of meaningful patterns without introducing excessive complexity.

The model’s configuration also included a random state of 42 to ensure the reproducibility of results and consistency in the splitting of training and testing datasets. These hyperparameter choices were guided by common practices in machine learning and the need to maintain interpretability and efficiency in the model. This setup ensured that the gradient boosting model was appropriately tailored to the structure and size of the dataset.

The dataset was divided into training and testing subsets, with 80% of the data allocated for training and 20% for testing. The training data was used to fit the gradient boosting model, allowing it to learn the relationships between the selected features and the target variable, wage. During this process, the model iteratively built decision trees, each focusing on correcting the residual errors of the previous trees. The testing data provided an independent set of observations to evaluate the model's generalization ability.

The features selected for training included `year`, `school`, `maried_yes`, `exper`, `industry_Manufacturing`, and `union_yes`. These variables were chosen based on their theoretical relevance and statistical significance, ensuring that the model captured key factors influencing wage prediction. The model’s training process incorporated these features to create a comprehensive representation of wage determinants.

The model’s performance was assessed using error metrics suitable for

regression tasks, specifically Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). MAE measures the average absolute differences between predicted and actual values, providing a straightforward understanding of prediction accuracy. RMSE, which emphasizes larger errors due to its quadratic nature, was used to evaluate the overall robustness of the model. These metrics were selected to ensure a balanced assessment of the model's accuracy and its ability to handle variations in the data.

This methodology provided a systematic approach to applying gradient boosting for wage prediction, ensuring the algorithm was configured and evaluated in alignment with best practices in machine learning. The emphasis on appropriate feature selection, robust training, and meaningful error metrics underscores the rigor of this methodological framework.

Result and Discussion

Comparison of Feature Importance

Feature importance analysis revealed notable differences between the Gradient Boosting and Linear Regression models. For Gradient Boosting, the most influential feature was "school," with an importance score of 0.388886, followed by "year" (0.229752) and "union_yes" (0.111090). In contrast, the Linear Regression model's coefficients highlighted "school" (0.308709) as the most impactful feature, followed by "union_yes" (0.293160) and "maried_yes" (0.229581). While both models emphasized the significance of education and union membership, Gradient Boosting captured more nuanced relationships by incorporating additional non-linear effects from "year" and "exper."

The importance of "school" across both models underscores the critical role of education in wage determination, aligning with human capital theory. Similarly, "union_yes" and "maried_yes" consistently emerged as significant predictors, reflecting the benefits of union membership and marital status on earnings. However, the Gradient Boosting model's ability to identify complex interactions, particularly with "year" and "industry_Manufacturing," highlights its advantage over linear approaches.

Model Performance

The predictive accuracy of the Gradient Boosting model and Linear Regression model was evaluated using RMSE and MAE. The Gradient Boosting model achieved a Train RMSE of 0.8639 and a Test RMSE of 0.8805, alongside a Train MAE of 0.6189 and a Test MAE of 0.6265. In comparison, the Linear Regression model demonstrated slightly higher Train RMSE and Test RMSE values of 0.8983 and 0.8863, respectively. Similarly, the Linear Regression model's Train MAE and Test MAE values were 0.6510 and 0.6404, indicating marginally reduced accuracy compared to the Gradient Boosting model. These results suggest that Gradient Boosting provides a more robust predictive performance for wage growth prediction. The Gradient Boosting model's ability to model non-linear relationships contributed to its improved performance over the Linear Regression model, which assumes a strictly linear relationship between features and the target variable. The consistent performance across training and testing datasets indicates that the Gradient Boosting model generalizes well, capturing the intricate interactions between the features that influence wage dynamics.

Discussion of Model Comparison

The comparison between Gradient Boosting and Linear Regression underscores the importance of selecting appropriate modeling techniques for wage prediction. Gradient Boosting's ensemble nature and capability to handle non-linear relationships offer a significant edge, particularly in datasets with complex feature interactions. Although Linear Regression provides a simpler and interpretable baseline, its assumptions limit its ability to capture subtle patterns within the data. The findings indicate that while both models identify key predictors of wages, Gradient Boosting achieves superior predictive accuracy and a more comprehensive understanding of feature importance. These results suggest that advanced machine learning algorithms like Gradient Boosting are better suited for wage growth analysis, offering actionable insights for policy makers and economic planners.

Variable Importance Analysis

The Gradient Boosting model highlighted several variables as critical predictors of wage growth, as evidenced by their feature importance scores. Among these, education (measured by the 'school' variable) emerged as the most influential predictor, contributing approximately 38.89% to the model's predictive accuracy. This finding underscores the pivotal role of education in determining wages, aligning with the human capital theory, which posits that higher education levels directly enhance an individual's earning potential. The 'year' variable, accounting for 22.98% of the importance, also played a significant role, reflecting temporal factors such as inflation, changes in economic conditions, and evolving labor market dynamics.

Other notable variables included union membership ('union_yes') and marital status ('maried_yes'), with contributions of 11.11% and 9.85%, respectively. These variables highlight the impact of collective bargaining and social factors on wage determination. Workers affiliated with unions typically benefit from better wages due to negotiated contracts, while marital status may serve as a proxy for stability or employer perceptions of worker reliability. 'Industry_Manufacturing' (9.10%) and experience ('exper') (8.08%) were also influential, emphasizing the importance of sector-specific wage standards and the accumulation of skills over time in shaping earnings.

Linear regression provided additional insights into variable significance through its coefficient values. Consistent with the Gradient Boosting results, education and union membership exhibited strong coefficients, further confirming their impact on wage outcomes. However, the linear model's inability to capture non-linear relationships limited its explanatory power compared to Gradient Boosting. For instance, while experience was important in both models, the Gradient Boosting model's ability to consider interaction effects made its predictions more robust.

Feature importance was visualized using bar charts (figure 5), which vividly depicted the relative contributions of each variable to wage prediction. The prominence of education as the leading factor reinforces the need for policies promoting access to higher education and skills training. Similarly, the moderate influence of industry and union membership suggests the need for sector-specific interventions and strengthening of collective bargaining mechanisms to reduce wage disparities. Overall, the analysis confirmed that wage growth is a multifaceted phenomenon influenced by both individual attributes, such as

education and experience, and contextual factors, such as industry and union presence.

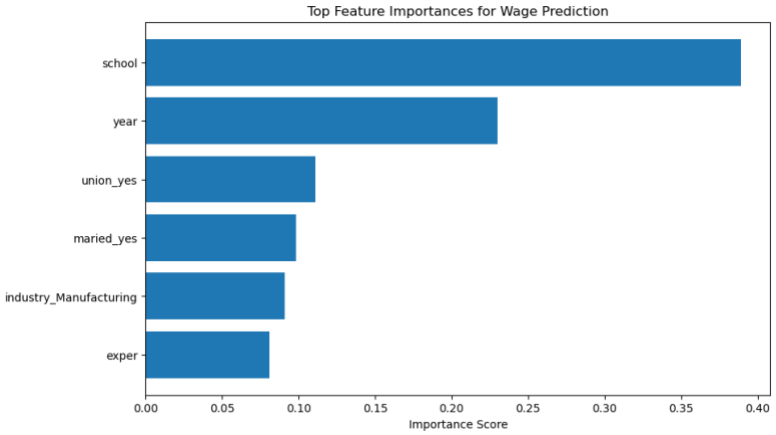

Figure 5 Feature Importance Bar

Interpretation of Results

The analysis revealed clear patterns in the relationship between key variables and wage outcomes. Experience demonstrated a positive correlation with wages, consistent with expectations that accumulated skills and tenure enhance productivity and earning potential. Similarly, education emerged as a critical determinant, with higher levels of schooling strongly linked to increased wages. Union membership also contributed significantly, reflecting the role of collective bargaining in securing better pay for workers across various industries. However, the magnitude of these effects varied depending on industry and other contextual factors, underscoring the complexity of wage dynamics. Unexpected findings included cases where higher education did not correspond to proportionately higher wages in certain industries, such as manufacturing. This anomaly may be attributed to industry-specific wage caps or an oversupply of educated workers in these sectors, leading to diminished returns on education. Additionally, the relatively modest impact of experience compared to education suggested that, while tenure is valuable, its influence on wages is often mediated by formal qualifications and industry standards.

Implications for Workforce Development

These findings carry important implications for workforce development and policy-making. The pronounced impact of education underscores the need for accessible and high-quality educational programs to enhance workforce competitiveness. Investments in vocational training and higher education can equip workers with the skills needed to meet evolving labor market demands. Additionally, targeted initiatives to address disparities in wage returns across industries could improve equity and incentivize workers to pursue education in high-demand fields. The significance of union membership highlights the importance of fostering strong labor organizations and collective bargaining mechanisms. Policymakers and industry leaders can leverage these insights to promote fair wage practices and reduce income inequality. Encouraging unionization in underrepresented sectors could provide workers with a stronger voice and improve wage outcomes, particularly in industries where education

and experience alone are insufficient to secure equitable pay. These findings suggest a need for integrated workforce strategies that balance educational access, skills development, and industry-specific interventions. By addressing both individual and systemic factors influencing wage growth, policymakers can create a more inclusive and equitable labor market that fosters long-term economic growth and worker prosperity.

Conclusion

This study demonstrated the significant influence of experience, education, and union membership on wage prediction. Among the variables examined, education emerged as the most critical factor, affirming its role as a key determinant of earning potential. Higher levels of education consistently correlated with increased wages, underscoring the importance of investing in educational attainment. Experience also positively impacted wages, though its effect was more nuanced, often mediated by industry-specific conditions and formal qualifications. Union membership played a substantial role, particularly in industries where collective bargaining secured higher wages and better employment conditions. These findings highlighted the multifaceted nature of wage determination, shaped by both individual attributes and contextual factors.

The research contributed to understanding wage determinants by integrating advanced machine learning techniques with traditional economic variables. Gradient Boosting provided a robust framework for analyzing non-linear relationships and interactions among predictors, offering enhanced predictive accuracy compared to baseline models like linear regression. This approach bridged gaps in traditional econometric methods, enabling a more comprehensive analysis of wage dynamics. Policymakers, educators, and workforce planners can leverage these insights to design targeted interventions, such as improving access to education, supporting unionization efforts, and addressing industry-specific wage disparities. The study emphasized the utility of AI-driven approaches in enriching economic modeling and decision-making processes.

Despite its strengths, the study faced several limitations. The dataset, while robust, focused on a specific timeframe and demographic, potentially limiting the generalizability of the findings to broader or more diverse populations. Additionally, the study considered a finite set of variables, leaving room to explore other socioeconomic factors, such as geographic mobility, digital skills, and workplace conditions. Future research could expand on this work by incorporating alternative machine learning algorithms, such as Random Forest or Neural Networks, to validate and refine the findings. Exploring cross-national datasets or longitudinal data could also provide deeper insights into the global and temporal dynamics of wage growth.

The findings of this study have practical implications for developing AI-powered tools to guide workforce planning and educational initiatives. Predictive models based on the identified variables can inform career counseling platforms, enabling individuals to make data-driven decisions about education and employment pathways. Policymakers can utilize these insights to design workforce development programs that prioritize high-impact interventions, such as subsidizing education in high-demand fields or supporting labor unions in underrepresented industries. These applications demonstrate the potential of combining machine learning with socioeconomic research to address critical

challenges in labor market policy and workforce development.

Declarations

Author Contributions

Conceptualization: A.B.P.; Methodology: B.M.A.; Software: B.M.A.; Validation: A.A.; Formal Analysis: A.B.P.; Investigation: B.M.A.; Resources: A.A.; Data Curation: A.B.P.; Writing Original Draft Preparation: B.M.A.; Writing Review and Editing: A.B.P.; Visualization: A.B.P.; All authors have read and agreed to the published version of the manuscript.

Data Availability Statement

The data presented in this study are available on request from the corresponding author.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] L. Fanti and L. Gori, “On Economic Growth and Minimum Wages,” *J. Econ.*, vol. 103, no. 1, pp. 59–82, 2011, doi: 10.1007/s00712-011-0190-3.

[2] S. Marlida, “The Impact of Labour, Wage, and Human Development Index on Economic Growth,” *J. Ekon. Dan Bisnis Airlangga*, vol. 33, no. 2, pp. 188–199, 2023, doi: 10.20473/jeba.v33i22023.188-199.

[3] A. Ali and L. Jiang, “Examining the Relationship Between Inequalities in Gender Wage and Economic Growth in Pakistan,” *Pak. J. Gend. Stud.*, vol. 12, no. 1, pp. 39–52, 2016, doi: 10.46568/pjgs.v12i1.198.

[4] Z. Liu, “Review on the Influence of Machine Learning Methods and Data Science on the Economics,” *Appl. Comput. Eng.*, vol. 22, no. 1, pp. 137–141, 2023, doi: 10.54254/2755-2721/22/20231208.

[5] M. Y. Amare and S. Šimonová, “Global Challenges of Students Dropout: A Prediction Model Development Using Machine Learning Algorithms on Higher Education Datasets,” *SHS Web Conf.*, vol. 129, no. 12, p. 09001, 2021, doi: 10.1051/shsconf/202112909001.

[6] M. I. Jordan and T. M. Mitchell, “Machine Learning: Trends, Perspectives, and Prospects,” *Science*, vol. 349, no. 6245, pp. 255–260, 2015, doi: 10.1126/science.aaa8415.

[7] I. I. B. Toleva, “Machine Learning for Decision Making in Medicine and Healthcare,” *J. Pharm. Negat. Results*, vol. 14, no. 2, pp. 2508–2516, 2023, doi: 10.47750/pnr.2023.14.s02.295.

[8] M. Atolia, “Trade Liberalization and Rising Wage Inequality in Latin America: Reconciliation With HOS Theory,” *J. Int. Econ.*, vol. 71, no. 2, pp. 467–494, 2007,

doi: 10.1016/j.jinteco.2006.06.005.
[9] H. Bhorat, R. Kanbur, and B. Stanwix, “Partial Minimum Wage Compliance,” *Iza J. Labor Dev.*, vol. 4, no. 1, p. 18, 2015, doi: 10.1186/s40175-015-0039-1.
[10] D. Cengiz, A. Dubé, A. Lindner, and D. Zentler-Munro, “Seeing Beyond the Trees: Using Machine Learning to Estimate the Impact of Minimum Wages on Labor Market Outcomes,” *J. Labor Econ.*, vol. 40, no. S1, pp. S203–S247, 2022, doi: 10.1086/718497.
[11] S. Liu and Y.-C. Su, “The Geography of Jobs and the Gender Wage Gap,” *Wp*, vol. 2020, no. 2028, pp. 1-66, 2020, doi: 10.24149/wp2028.
[12] X. Liu, “Salary Grades Prediction Using Machine Learning,” *Appl. Comput. Eng.*, vol. 8, no. 1, pp. 248–255, 2023, doi: 10.54254/2755-2721/8/20230152.
[13] S. Kampelmann, F. Rycx, Y. Saks, and I. Tojerow, “Does Education Raise Productivity and Wages Equally? The Moderating Role of Age and Gender,” *Iza J. Labor Econ.*, vol. 7, no. 1, p. 1, 2018, doi: 10.1186/s40172-017-0061-4.
[14] H. I. Shahiri, Z. Osman, and P. Kihong, “Union Relevance in the Malaysian Labour Market,” *Asian-Pac. Econ. Lit.*, vol. 30, no. 2, pp. 45–56, 2016, doi: 10.1111/apel.12153.
[15] Z. Parolin and T. VanHeuvelen, “The Cumulative Advantage of a Unionized Career for Lifetime Earnings,” *Ilr Rev.*, vol. 76, no. 2, pp. 434–460, 2022, doi: 10.1177/00197939221129261.
[16] D. G. Blanchflower and A. Bryson, “The Wage Impact of Trade Unions in the UK Public and Private Sectors,” *Economica*, vol. 77, no. 305, pp. 92–109, 2009, doi: 10.1111/j.1468-0335.2008.00726.x.
[17] M. J. Maleka, C. Schultz, L. V. Hoek, L.-A. Paul-Dachapalli, and S. Ragadu, “Union Membership as a Moderator in the Relationship Between Living Wage, Job Satisfaction and Employee Engagement,” *Indian J. Labour Econ.*, vol. 64, no. 3, pp. 621–640, 2021, doi: 10.1007/s41027-021-00322-0.
[18] İ. İlkkaracan and R. Selim, “The Gender Wage Gap in the Turkish Labor Market,” *Labour*, vol. 21, no. 3, pp. 563–593, 2007, doi: 10.1111/j.1467-9914.2007.00378.x.
[19] S. Kirby and R. Riley, “The External Returns to Education: UK Evidence Using Repeated Cross-Sections,” *Labour Econ.*, vol. 15, no. 4, pp. 619–630, 2008, doi: 10.1016/j.labeco.2008.04.004.
[20] J. Bray, J. Hinde, and A. Aldridge, “Alcohol Use and the Wage Returns to Education and Work Experience,” *Health Econ.*, vol. 27, no. 2, pp. e87-e100, 2017, doi: 10.1002/hec.3565.
[21] P. Adhikari, K. Kc, and S. R. Bhatta, “Does Education and Experience Matter in the Distribution of Wages in Nepal? A Quantile Regression Approach,” *Econ. J. Dev. Issues*, vol. 28, no. 1, pp. 1–14, 2019, doi: 10.3126/ejdi.v28i1-2.33193.
[22] S. H. Yang and B. Y. Jeong, “Gender Differences in Wage, Social Support, and Job Satisfaction of Public Sector Employees,” *Sustainability*, vol. 12, no. 20, p. 8514, 2020, doi: 10.3390/su12208514.