Journal of Information Analytics, 1(1), 12-22

Journal of Information Analytics

Submit 05.12.2024
Revision 16.01.2025
Accept 18.02.2025
Publish 12.03.2025

Research Article

Open Access

Machine Learning-Based Forecasting of Minimum Wage in Turkey: A Case Study (2005-2024)

Nesrin Altınyüzük Gezer[1], Yunus Eroğlu[2], Suleyman Mete[2]

[1]Graduate School of Natural & Applied Sciences, Gaziantep University, Gaziantep, Türkiye

[2]Department of Industrial Engineering, Gaziantep University, 27100 Gaziantep, Türkiye

altinyuzuknesrinn@hotmail.com, eroglu@gantep.edu.tr, smete@gantep.edu.tr

Abstract

The minimum wage represents the lowest legally permissible level of compensation that employers are obligated to pay their employees. Determining the minimum wage involves a complex interplay of economic, social, and political factors, making it a crucial indicator for labor market policies. This study examines Turkey's minimum wage trends over the period 2005-2024, leveraging historical data to uncover the key determinants influencing wage adjustments. The aim is to predict the 2025 minimum wage by identifying the parameters that have significant effects on the determination of the minimum wage. Historical data was analyzed using machine learning and a forecasting analysis was performed for the 2025 minimum wage. In this context, studies were carried out using machine learning algorithms such as AdaBoost, Neural network, Linear regression. Orange 3 program was used with the data for machine learning and prediction processes. This research contributes to the literature by demonstrating the applicability of machine learning in economic forecasting, providing valuable insights for policymakers, employers, and labor market stakeholders.

Keywords Minimum wage, machine learning, forecasting, prediction models

Citation: Altınyüzük Gezer, N., Eroğlu, Y., & Mete, S. (2025). Machine learning-based forecasting of minimum wage in Turkey: A case study (2005-2024). *Journal of Information Analytics*, 1(1), 12-22.

Corresponding Author: Suleyman Mete smete@gantep.edu.tr

1. Introduction

Wage can be defined as the monetary level that employees earn because of physical or mental work to sustain their lives and meet their needs and the needs of their dependents (Çınar and Öz, 2018). A minimum wage is the lowest remuneration that employers can legally pay their employees. The purpose of minimum wages is to protect workers against unduly low pay and to help ensure a fair and equitable share of fruits of progress to all, and a minimum wage to all who are employed and in need of such protection (ILO, 2015).

This study aims to estimate the minimum wage in Turkey for the year 2025 by analyzing critical economic indicators that influence its determination. Various parameters play an active role in shaping the minimum wage, reflecting the intricate dynamics of the country's economy and labor markets. Specifically, the parameters identified in this study include the Wholesale Price Index (WPI), the Producer Price Index (PPI), Turkey's unemployment rate, and exchange rates for major currencies such as the USD and EUR. These indicators have been selected based on their historical correlation with wage adjustments and their capacity to reflect broader economic trends.

To ensure a robust analysis, the study incorporates a comprehensive dataset covering the period from 2005 to 2024. This dataset includes the historical values of WPI, PPI, unemployment rates, exchange rates for the USD and EUR, and the corresponding minimum wage figures during this timeframe. By analyzing the interplay between these variables, the study seeks to uncover patterns and relationships that can inform the projection of the minimum wage for 2025. This approach provides a holistic perspective, taking into account both domestic economic pressures, such as unemployment, and external factors, like currency fluctuations, which significantly impact Turkey's economic environment.

The findings of this study offer valuable insights for a wide range of stakeholders. Policymakers and labor organizations can leverage this data to make informed decisions during wage-setting negotiations, ensuring that the minimum wage reflects both the needs of workers and the realities of the economic environment. Moreover, these insights serve as a crucial resource for businesses and organizations, enabling them to prepare accurate and strategic budget forecasts for the coming year. By aligning financial planning with anticipated changes in labor costs, companies can better navigate economic uncertainties and maintain operational efficiency. This comprehensive analysis underscores the importance of integrating economic trends into decision-making processes, benefiting all parties involved in Turkey's labor market dynamics.

2. Literature Review

There is limited academic research that directly focuses on minimum wage estimation using machine learning. However, there are some studies that apply machine learning methods to salary and wage estimation.

Ghei and Lee (2020) discuss a study using machine learning methods to predict annual wages. The study compares the performance of basic Mincer regression on Current Population Survey (CPS) data with advanced machine learning algorithms such as XGBoost, LightGBM and deep learning. The best results are obtained with the Stacking method, which is a combination of different models. For forecasts limited to a small set of variables, the machine learning methods provided only a modest improvement

compared to Mincer regression, but greatly outperformed forecasts with an expanded set of variables. In particular, Stacking provided significant improvements in the RMSE-log, RMSE-level and absolute deviation measures, combining the strengths of different algorithms to achieve higher prediction accuracy.

Çınar and Öz (2018) applying the human capital model to wage estimation for the service sector in Turkey, it analyzes the effect of individuals' characteristics such as education and experience on wage levels. In the study, a survey was applied to 2000 people working in Bursa. The study examined the contribution of years of education, work experience and other demographic factors to wage differences based on Jacob Mincer's wage equation. The data were collected from individuals working in the service sector and analyzed with econometric methods. The findings showed that education level and work experience have a significant and positive effect on wages, but these effects offer a diminishing return after a certain level. The study emphasizes that service sector wage policies can be improved with strategies aimed at developing human capital.

Cazcarra (2024) conducted a study analyzing the impact of increasing the minimum wage on income inequality between 2001-2021 in Spain. Using national census data provided by the Spanish Tax Administration, he showed that increasing the minimum wage reduced income inequality. It was found that increasing the minimum wage did not lead to inflation or unemployment, but rather increased net employment, kept prices under control, and increased company profit margins. Analysis using Multivariate Linear Regression, Random Forest Regressor, and Time Series Regression Model machine learning models revealed that increasing the minimum wage increased the country's wealth, increased employment and company profits, and was an effective method for wealth redistribution. For example, the multivariate linear regression model explained 99.3% of the Gini index, using p-values and coefficients to determine the effects of independent variables on the Gini index. The random forest regression model optimized the mean square error (MSE) to 0.0039 and determined the significance of the variables. These findings show that the minimum wage increase is effective in reducing income inequality and increasing economic welfare

3. Material and Method

In the application part of the study, the last 20 years of dataset was used. In this context, WPI (Wholesale Price Index), PPI (Producer Price Index), Turkey's unemployment rate and exchange rates data between 2005-2024 were collected from TÜİK. Table 1 shows the last 20 years of data on the parameter values affecting the minimum wage.

Table 1. Last 20 years wage dataset

Year	Minimum Wage (TRY)	Year+1	WPI (Wholesale Price Index)	PPI (Producer Price Index)	Unemployment Rate (%)	USD/TRY Average Exchange Rate	EUR/TRY Average Exchange Rate
2005	350,15	380,46	7,72	2,66	10,6	1,3405	1,67
2006	380,46	419,15	9,65	11,58	10,2	1,4297	1,798
2007	419,15	503,26	8,39	5,94	10,3	1,3003	1,7773
2008	503,26	546,48	10,06	8,11	11	1,2976	1,8969
2009	546,48	599,12	6,53	5,93	14	1,5457	2,1508
2010	599,12	658,95	6,4	8,87	11,9	1,499	1,9886
2011	658,95	739,8	10,45	13,33	9,8	1,6708	2,3244
2012	739,8	803,68	6,16	2,45	9,2	1,7922	2,3041
2013	803,68	891,03	7,4	6,97	9,7	1,9033	2,529
2014	891,03	1000,54	8,17	6,36	9,9	2,1865	2,9042
2015	1000,54	1.300,99	8,81	5,71	10,3	2,7191	3,0187
2016	1.300,99	1.404,06	8,53	9,94	10,9	3,0181	3,3375
2017	1.404,06	1.603,12	11,92	15,47	10,9	3,6445	4,1159
2018	1.603,12	2.020,90	20,3	33,64	11	4,8301	5,6789
2019	2.020,90	2.324,71	11,84	7,36	13,7	5,6712	6,3481
2020	2.324,71	2.825,90	14,6	25,15	13,2	7,0034	8,014
2021	2.825,90	5.500,35	36,08	79,89	13,4	8,8557	10,4408
2022	5.500,35	11.402,32	64,27	97,72	10,4	16,5512	17,3642
2023	11.402,32	17.002,12	64,77	44,22	9,4	23,7482	25,6852
2024	17.002,12	?	44,38	28,52	9	32,7825	35,4779

The aim is to make a minimum wage estimate for 2025 by teaching the last 20-year values with machine learning. Machine learning techniques were employed for the analysis process. Initially, the dataset was loaded, preprocessed, and subjected to exploratory data analysis to understand its structure and key features. The data was then split into training and testing subsets, ensuring proper validation. Various machine learning algorithms, specific models such as Linear Regression, Gradient Boosting, Neural Network, Adaboost, Stochastic Gradient Descent were trained on the training subset using cross-validation to optimize hyperparameters and minimize overfitting.

After training, the performance of each model was evaluated using appropriate metrics, such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), R-squared (R^2) etc. Based on these metrics, the model yielding the best results was selected. Finally, the chosen model was used to generate predictions on unseen data, and the results were analyzed to interpret the model's behavior and assess its real-world applicability.

Estimating the minimum wage with data from previous years is a time series analysis. In time series analysis, it is a common approach to predict future values using past data. In order to predict future values of economic variables, such as minimum wage forecasts, historical data needs to be analyzed. In

this context, machine learning models are usually trained and tested using the 1-year shift method. While this method allows the model to learn based on historical data, it also allows to evaluate how the model performs not only with training data but also with new data that it will actually encounter in the testing phase. In other words, testing by shifting by one year provides a more realistic measure of the model's ability to predict the future and reduces the risk of overfitting. This strategy helps the model learn the changing dynamics over time, so that it can make more accurate predictions when faced with future data.

In this study, the minimum wage data in the dataset has been shifted by 1 period and a column named "Year+1" has been added to calculate the 2025 minimum wage as target.

After the data was exported to Orange with the "File" function, it was displayed as a table with the "Data table" function and the data was checked. Then, with the "Select Columns" function, unnecessary data was ignored, and the target data was determined as follows. In Figure 1, it is seen that the data to be estimated is Year+1.

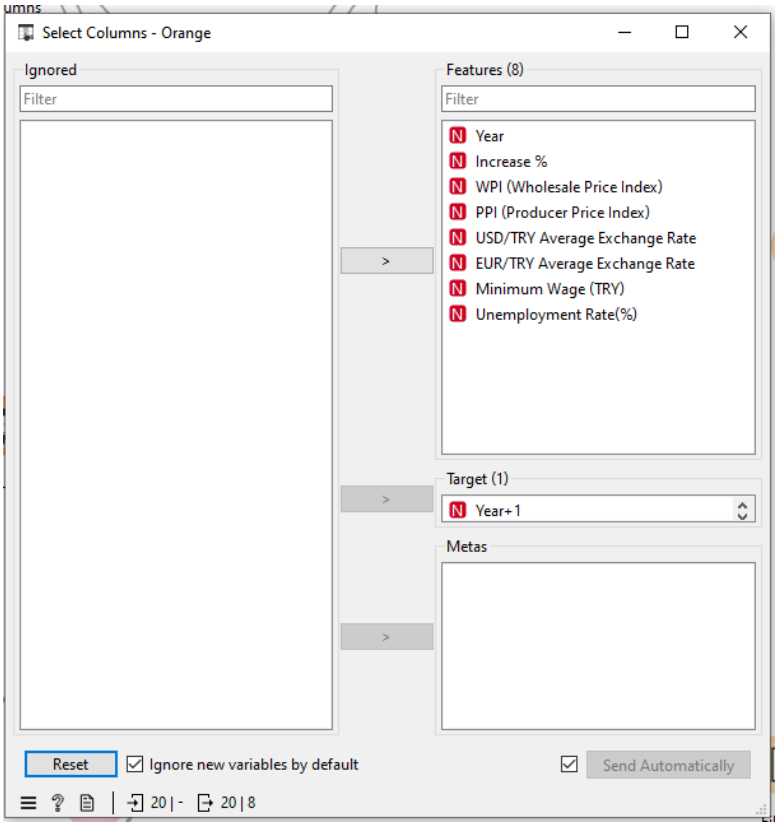

Figure 1. Determining target data

Five machine learning techniques were used in the analysis with Orange. These are Neural Network, Adaboost, Linear Regression, Stochastic Gradient Descent, Gradient Boosting learning methods.

3.1 Machine Learning Algorithms

a. Neural Network

Artificial Neural Network (ANN) has been a hot topic in artificial intelligence since the 1980s. It abstracts the human brain neural network from the perspective of information processing, establishes a simple model and composes different networks according to different connections (Dong and Hu, 1997). Trying to simulate brain neural network processing, memory information in the way of information processing.

In engineering and academia are often directly referred to as neural network or neural network. Neural network is a computing model, by a large number of nodes (or neurons) connected to each other (Jenkins and Tanguay, 1995). Each node represents a specific output function, called the activation function. The connection between every two nodes represents a weight for the signal passing through the connection, which is called the weight, which is equivalent to the memory of the artificial neural network (Bnlsabi, 1993). The output of the network will vary depending on how the network is connected, the weight value, and the incentive function. However, the network itself is usually an approximation to some kind of algorithm or function in nature, or it may be an expression of a logic strategy (Luo et al., 1998).

b. Adaboost

The AdaBoost algorithm corrects the misclassifications made by weak classifiers, and it is less susceptible to overfitting than most learning algorithms. Recognition performances of the AdaBoost-based classifiers are generally encouraging (Freund and Schapire, 1997).

c. Linear Regression

Linear regression is the building block for many modern modeling tools. In particular, when the sample size is small or the signal is relatively weak, linear regression often provides a satisfactory approximation to the underlying regression function (James et al., 2023).

d. Stochastic Gradient Descent (SGD)

Used to provide optimized learning processes on large datasets, this method provides a stochastic approach to minimize the cost function.

e. Gradient Boosting

Based on the principle of boosting tree-based models, this method aims to gradually reduce errors by succeeding in adding weak learners. With these 5 machine learning techniques in Orange, the dataset was first trained and then predicted. Figure 2 shows the process flow in Orange.

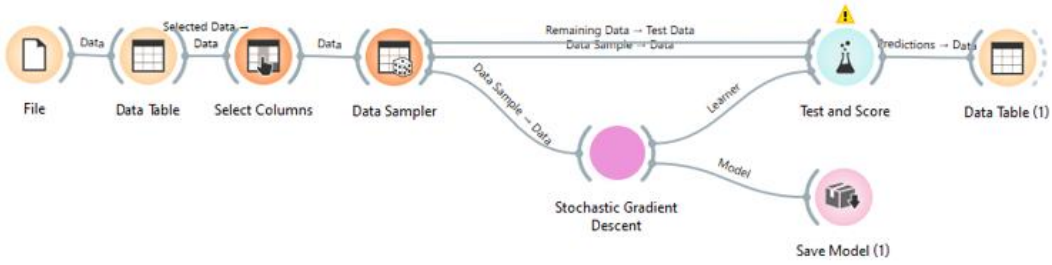

Figure 2. Process flow in Orange

The data analysis process began with loading data into Orange. In the first step, the general structure of the data was examined by providing data representation. Then, the target value (the output to be predicted) and the inputs (the variables used by the model for prediction) were determined. In order to make a more accurate assessment of the data, the training and test data were separated into separate groups by applying the sample separation (train-test split) method. In this step, the 'Test & Score' function used for testing and model validation was activated. Along with the sample separation process, the testing process was carried out by matching the model with the training data. The results were

observed, and the performance of the model was evaluated. Figure 3 shows the flow of loading the model into Orange to make predictions.

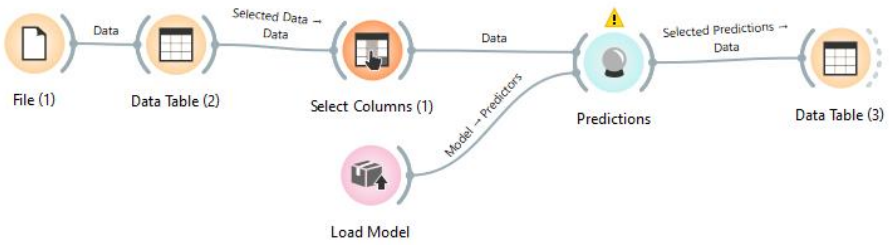

Figure 3. Loading the model with Orange

The mean squared error (MSE) or mean squared deviation (MSD) of an estimator (of a procedure for estimating an unobserved quantity) measures the average of the squares of the errors—that is, the average squared difference between the estimated values and the true value.

If vector n of predictions is generated from a sample of n data points on all variables, and Y is the vector of observed values of the variable being predicted, with Ŷ being the predicted (e.g. as from a least-squares fit), then the within-sample MSE of the predictor is computed as (Wikipedia, 2024a):

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2$$

RMSE (Root Mean Squared Error): The root mean square deviation (RMSD) or root mean square error (RMSE) is either one of two closely related and frequently used measures of the differences between true or predicted values on the one hand and observed values or an estimator on the other.

The RMSD of an estimator $\hat{\theta}$ with respect to an estimated parameter θ is defined as the square root of the mean squared error (Wikipedia, 2024b):

$$\mathrm{RMSD}(\hat{\theta}) = \sqrt{\mathrm{MSE}(\hat{\theta})} = \sqrt{\mathrm{E}((\hat{\theta} - \theta)^2)}.$$

MAE (Mean Absolute Error): Measures the average absolute difference between predicted and actual values. It is more robust to outliers than MSE (Eroğlu, 2024).

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n}|y_i - x_i|}{n} = \frac{\sum_{i=1}^{n}|e_i|}{n}.$$

It is thus $|e_i| = |y_i - x_i|$ and arithmetic average of the absolute errors , where y_i is the prediction and x_i the true value (Wikipedia, 2024c).

MAPE (Mean Absolute Percentage Error): Shows the percentage difference between predicted and actual values, useful for understanding relative errors (Eroğlu, 2024).

$$\mathrm{MAPE} = 100\frac{1}{n}\sum_{t=1}^{n}\left|\frac{A_t - F_t}{A_t}\right|$$

Where A_t is the actual value and F_t is the forecast value. Their difference is divided by the actual value A_t. The absolute value of this ratio is summed for every forecasted point in time and divided by the number of fitted points n (Wikipedia, 2024d).

R^2 (R-Square): Represents the proportion of variance in the target variable explained by the model. An R^2 value closer to 1 indicates a better fit (Eroğlu, 2024).

R^2 formulation is as below where SS_{res} is residual sum of squares and SS_{tot} total sum of squares; (Wikipedia, 2024e).

$$R^2 = 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}}$$

4. Results and Discussion

The results obtained in the study were compared using five different metrics (MSE, RMSE, MAE, MAPE and R^2) to evaluate the performance of each model. These metrics comprehensively measure the accuracy, error rate and overall performance of each model. Table 2 shows the performance results of the machine learning models.

Table 2. Results of machine learning models

Model	MSE	RMSE	MAE	MAPE	R^2
Neural Network	173.595.602	416.648	261.018	0.111	0.992
Stochastic Gradient Descent	66.205.537	257.304	204.619	0.222	0.997
Linear Regression	473.863.523	217.684	164.023	0.178	0.998
AdaBoost	4.265.772	65.313	37.237	0.076	1
Gradient Boosting	0.020	0.142	0.106	0.000	1

MSE (Mean Squared Error) and RMSE (Root Mean Squared Error) values indicate the distance of the model's predictions from the actual data. Lower values in these metrics indicate that the model makes more accurate predictions.

MAE (Mean Absolute Error) and MAPE (Mean Absolute Percentage Error) measure error rates as absolute values. These metrics reflect how accurate the predictions are and how many errors the model makes.

R^2 (R-squared) is a coefficient that shows how well the model fits the data and the explanatory power of the independent variables on the dependent variable. A high R^2 value indicates that the model explains the data set well.

Gradient Boosting and AdaBoost Models

When the results are examined, Gradient Boosting and AdaBoost models stand out with their low error rates and high accuracy values. Both models have low MSE, RMSE, MAE and MAPE values and exhibit high R^2 values.

Gradient Boosting has modeled the data set flawlessly by showing excellent performance in all metrics. The fact that the MSE and RMSE values are almost zero reveals that the model's predictions are extremely close to the real data. The fact that the MAE and MAPE values are also close to zero shows that the model minimizes prediction errors. The R^2 value is also close to 1, indicating that this model explains the

data by 100%. These results indicate that the Gradient Boosting model is the model that produces the most accurate and reliable predictions for this data set.

AdaBoost is also a model that exhibits high performance in a similar way. The MSE and RMSE values are quite low, and the MAE and MAPE values are at a minimum level. This shows that AdaBoost has a high capacity to make accurate predictions, and that the model's performance is quite reliable. In addition, the R^2 value is close to 1, explaining the data with very high accuracy. Although AdaBoost shows slightly higher error rates compared to Gradient Boosting, it still stands out as a strong alternative.

Neural Network and Linear Regression Models

Neural Network and Linear Regression models stand out with higher error rates and lower accuracy levels.

The Neural Network model demonstrates a high performance in terms of R^2 value, explaining 99.2% of the data set. However, since the MSE, RMSE, MAE and MAPE values are high, it is seen that there are large errors in the model's predictions. This shows that although the Neural Network model learns the data quite well, its overall accuracy is not sufficient, and that improvement is needed. In particular, the MAPE value of 11.1% indicates that the model's predictions are significantly incorrect.

Linear Regression performed very poorly on this dataset. The high MSE, RMSE, MAE and MAPE values indicate that the linear regression model's estimates are far from the real data. In particular, the MSE value (473,863,523) is quite high, and the model's estimates contain a lot of errors. Although the R^2 value is high at 0.998, this model was generally inadequate due to its high error rates.

Stochastic Gradient Descent

The Stochastic Gradient Descent model provides much lower values in error metrics such as MSE and RMSE than the Linear Regression and Neural Network models, indicating that the accuracy of the predictions is high. This shows that SGD manages to approach the lowest error more quickly and efficiently when updating the model parameters.

5. Conclusion

As a result, Gradient Boosting achieved the most successful results among the models tested within the scope of this study. Gradient Boosting modeled the dataset very well with low error rates and high accuracy values and provided excellent prediction accuracy. AdaBoost also stood out as a very effective model and exhibited high performance in terms of prediction accuracy and error rates.

However, Neural Network and Linear Regression models attracted attention with their higher error rates and low accuracy levels and produced less effective results for this dataset. Although Neural Network has learned the dataset well with its high R^2 value, its error rates in predictions are quite high. Linear Regression can be considered as the least successful model with both low accuracy and high error rates.

The Stochastic Gradient Descent (SGD) model, when compared to other models, still has some error rates compared to advanced techniques such as Gradient Boosting and AdaBoost, but generally achieves better results compared to Linear Regression and Neural Network models. SGD's MSE, RMSE, MAE and

MAPE values indicate lower error rates than the performance of these models. Since the R^2 value shows that the model explains the data by 99.7%, it can be said that it has a very successful prediction capacity.

In this context, it can be said that the Stochastic Gradient Descent model shows slightly lower accuracy than Gradient Boosting and AdaBoost but still offers a strong alternative. This model can be an effective option for those looking for a simpler and faster solution.

As a result, it can be said that the Gradient Boosting model is the most successful model on this dataset, offering excellent prediction accuracy with minimal errors. AdaBoost also stands out as an important alternative, demonstrating high performance and accuracy, although it shows slightly higher error rates compared to Gradient Boosting. Additionally, Stochastic Gradient Descent (SGD) performs well, delivering lower error rates than models like Linear Regression and Neural Networks, though it does not achieve the same level of accuracy as Gradient Boosting and AdaBoost. Therefore, Gradient Boosting and AdaBoost are the models to prefer for this dataset due to their superior performance, while Stochastic Gradient Descent provides a reliable alternative with a good balance between performance and computation efficiency.

When the estimates made with the model are evaluated, the Stochastic Gradient Descent (SGD) model estimated the minimum wage as 22170 TRY for 2025. The minimum wage announced for 2025 in Turkey was determined as 22104 TRY. Compared to the estimates of other models, the SGD model presented a result that is quite close to the actual value.

References

Bnlsabi, A. (1993). Some analytical solutions to the general approximation problem for feed forward neural networks. Neural Networks, 6, 991–996.

Cazcarra, M. L., (2024). Machine learning analysis of the impact of increasing the minimum wage on income inequality in Spain from 2001 to 2021," arXiv preprint arXiv:2402.02402, 2024. https://arxiv.org/abs/2402.02402

Çınar, M., & Öz, R., (2018). Hizmet sektörü için ücret tahmini: insan sermayesi modeli, International Conference on Economics and Administrative Sciences (IZCEAS), Bursa, Turkiye, Aralık.

Dong, J., & Hu, S. (1997). The progress and prospects of neural network research. Information and Control, 26(5), 360–368.

Eroğlu, Y. (2024). IMDb score estimation using movie dialogues: A text mining and machine learning hybrid approach, International Data Science and Statistics Congress, Ankara, Türkiye.

Freund, Y., & Schapire, R.E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting, *Journal of Computer and System Sciences*, 55 (1), 119–139. https://ieeexplore.ieee.org/document/4454220.

Ghei, D., & Lee, S.M., (2020). Annual wage prediction: machine learning competition, https://static1.squarespace.com/static/62014b6e32679565e6c69891/t/621e43aaace5152a7b76ff1a/1646150570815/2020_Winners_sangminlee-dhanajay-report-updated.pdf

International Labor Organization, "How to define a minimum wage?," https://www.ilo.org/global/topics/wages/minimum-wages/definition/WCMS_439072/lang--en/index.htm.

James, G., Witten, D., Hastie, T., Tibshirani, R., & Taylor J. (2023). Linear regression, in *An Introduction to Statistical Learning: With Applications in Python*, Cham: Springer International Publishing, 69–134.

Jenkins, B.K., & Tanguay, A.R. (1995). Handbook of neural computing and neural networks. Boston: MIT Press.

Luo, Z.H., Xie, Y., & Zhu, C. (1997). The study of convergence of CMAC learning process. Acta Automatic Sinica, 23(4), 455–461.

Wikipedia "Mean absolute percentage error," Wikipedia, Oct. 4, 2024 https://en.wikipedia.org/wiki/Mean_absolute_percentage_error

Wikipedia, "Coefficient of determination," *Wikipedia*, Dec. 17, 2024. https://en.wikipedia.org/wiki/Coefficient_of_determination.

Wikipedia, "Mean absolute error," Wikipedia, Dec. 21, 2024. https://en.wikipedia.org/wiki/Mean_absolute_error.

Wikipedia, "Mean squared error," Wikipedia, Oct. 22, 2024. https://en.wikipedia.org/wiki/Mean_squared_error.

Wikipedia, "Root means square deviation," Wikipedia, Oct. 15, 2024. https://en.wikipedia.org/wiki/Root_mean_square_deviation.