Understanding Gender Inequality in Indonesia: An AI Approach to Evaluating Socio-Economic Factors for Sustainable Development

Hani Brilianti Rochmanto[1]*, Harun Al Azies[2]
[1]Department of Statistics, Faculty of Science and Technology, Universitas PGRI Adi Buana, Surabaya 60234, Indonesia
[2]Study Program in Informatics Engineering, Faculty of Computer Science, Universitas Dian Nuswantoro, 50131, Semarang, Indonesia
* Corresponding Author: Hani Brilianti Rochmanto. Email: rochmantohani@unipasby.ac.id

Abstract

Gender inequality in Indonesia remains a significant challenge, especially in achieving gender parity across various development indicators. This study evaluates the impact of different socioeconomic and demographic variables on the Gender Inequality Index (GII) across all regencies and cities in Indonesia in 2023, focusing on aligning with the fifth Sustainable Development Goal (SDG) of Gender Equality. The variables considered include Women's Income Contribution, Women in Professional Occupations, Women's Human Development Index, Women's Life Expectancy, Women's Labor Force Participation Rate, Women's Expected Years of Schooling, and Adjusted Per Capita Expenditure for Women. Three machine learning models, Support Vector Regression (SVR), AdaBoost Regressor, and Random Forest Regressor are applied to determine the relationship between these factors and the GII. Random Forest Regressor demonstrated the best performance with a Mean Squared Error (MSE) of 0.0109. The feature importance analysis reveals that the Women's Human Development Index has the highest impact at 43.92%, followed by Women's Life Expectancy at 23.25%, and Adjusted Per Capita Expenditure for Women at 10.27%. These findings highlight the pivotal role of human development and economic factors in shaping gender inequality in Indonesia, providing valuable insights for formulating targeted policies aimed at reducing gender disparities and promoting equitable development across the country.

Keywords: Gender Inequality; Artificial Intelligence; Socio-Economic Factors; Machine Learning; Random Forest

1. Introduction

Gender inequality remains a pressing issue globally, influencing various aspects of socio-economic development. In many developing countries, gender disparities in education, employment, and income distribution hinder progress toward achieving sustainable development (1). Gender inequality is not only a matter of social justice but also an economic concern, as studies have shown that reducing gender gaps can significantly boost economic growth and improve overall societal well-being (2). One of the critical international frameworks addressing this issue is the Sustainable Development Goals (SDGs), specifically Goal 5, which advocates for gender equality and the empowerment of all women and girls (3). Monitoring and analyzing gender inequality trends within specific regions provide crucial insights into the effectiveness of policies aimed at closing gender gaps.

In Indonesia, gender inequality persists despite notable advances in various social indicators. Although the country has shown improvements in women's access to education and health, significant gaps remain in labor force participation, income equality, and political representation. According to the World Economic Forum (2022), Indonesia ranks low in the Global Gender Gap Index, indicating a need for more targeted policies and interventions (4). Understanding the socio-economic factors contributing to

this inequality, particularly at the local level, is essential for formulating effective strategies for gender equality that align with Indonesia's broader development goals.

Gender Inequality Index (GII) serves as a comprehensive measure to assess disparities between men and women in terms of reproductive health, empowerment, and labor market participation (5). Given the complexity and interrelated nature of the variables influencing gender inequality, this research employs machine learning models to analyze the data and identify the most significant factors contributing to the GII. The use of advanced modeling techniques, such as machine learning, allows for more accurate and nuanced predictions of GII values, providing policymakers with valuable insights into the areas that need attention to reduce gender disparities. By leveraging modern machine learning techniques, this research aims to develop predictive models that can identify the most significant variables contributing to gender inequality. These models will help in understanding the regional disparities and offer insights into which socio-economic factors should be prioritized to promote gender equality.

Machine learning approaches have been increasingly applied in social science research to address complex issues, such as gender inequality, due to their ability to handle large datasets and uncover intricate patterns (6,7). Support Vector Regression (SVR) has been shown to perform well in predicting socio-economic indicators due to its robustness in handling non-linear relationships (8). Moreover, ensemble learning methods like AdaBoost and Random Forest have proven effective in improving prediction accuracy by combining multiple weak learners into a strong predictive model (9). These methods provide a significant advantage in analyzing multi-dimensional data. Recent studies have used these machine learning methods to demonstrated their superior performance compared to classical regression techniques. For example, Random Forest and AdaBoost models applied to predict key outputs of non-traditional machining (NTM) processes. The study finds that traditional Linear Regression performs poorly due to the non-linear nature of NTM processes, while Random Forest and AdaBoost significantly improve predictive accuracy (9). Similarly, SVR has been applied for modeling the life satisfaction in older people, with results showing that SVR achieved the best performance in predicting life satisfaction (10). These approaches allow for more accurate forecasting and understanding of the key drivers behind socio-economic disparities, which are essential for effective policy-making.

In the context of gender inequality, applying machine learning can provide a more nuanced analysis of how different factors such as income contribution, professional participation, and education influence gender disparities. By analyzing the socio-economic factors through these advanced techniques, this study seeks to offer new insights into which areas of intervention can most effectively reduce gender inequality in Indonesia. This study aims to address two key research questions: How can machine learning models be applied to analyze and predict the Gender Inequality Index (GII) in Indonesia, and which socioeconomic and demographic factors have the most significant impact on the GII? These questions are crucial for understanding the relationship between gender inequality and the various factors that influence it, ultimately guiding the development of targeted interventions. The objective of this study is to analyze the socio-economic factors influencing the Gender Inequality Index (GII) across all districts and cities in Indonesia for the year 2023, using machine learning approaches—specifically Support Vector Regression (SVR), AdaBoost Regressor, and Random Forest Regressor. By identifying the most significant factors contributing to gender inequality, this research aims to provide a foundation for targeted policy-making that can address regional disparities in gender inequality.

2. Method

This study applies a quantitative research design to examine the influence of socioeconomic and demographic factors on gender inequality across Indonesia's regencies and cities. Specifically, machine learning techniques are employed to model and predict the Gender Inequality Index (GII) based on key

indicators reflecting women's socioeconomic status and development. Before conducting the analysis, the dataset was preprocessed, and the following steps were applied. The data was split into a training set (80%) and a testing set (20%) to facilitate model evaluation. Next, data standardization was performed using a StandardScaler to ensure all variables were on a consistent scale. Three machine learning models were employed in this study, namely Support Vector Regression (SVR), AdaBoost Regressor, and Random Forest Regressor. For model optimization, hyperparameter tuning of the Random Forest Regressor was conducted using GridSearchCV to identify the best-performing combination of parameters. The performance of these models was evaluated using Mean Squared Error (MSE) as the primary metric, and the model with the lowest MSE was selected as the best-performing model. Finally, feature importance analysis was carried out on the trained Random Forest model to determine the most significant variables influencing the Gender Inequality Index (GII). The entire process of this study is visualized in Figure 1, which provides a comprehensive overview of the methodology employed, including data preparation, model selection, and evaluation steps.

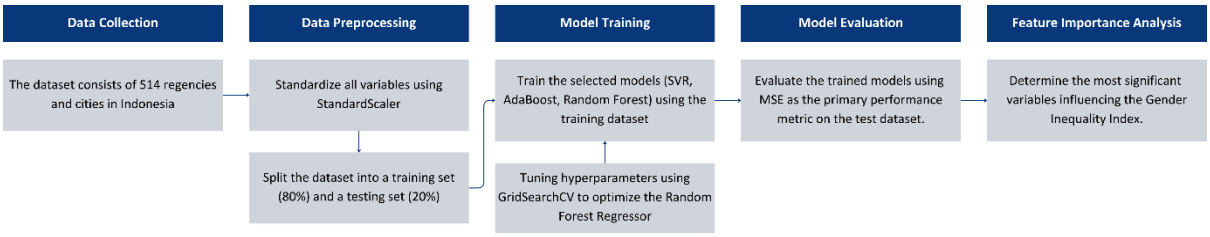

Figure 1. Machine Learning Workflow for Analyzing Gender Inequality in Indonesia

2.1. Data and Variables

This study utilized secondary data sourced from Statistics Indonesia (BPS). The primary dataset includes variables related to gender inequality and socio-economic factors for 514 regencies and cities in Indonesia for the year 2023. The key dependent variable is the Gender Inequality Index (GII), which measures disparities in reproductive health, empowerment, and labor market participation. The independent variables in this study include several socio-economic factors: Women's Income Contribution (X1), Women in Professional Occupations (X2), Women's Human Development Index (X3), Women's Life Expectancy (X4), Women's Labor Force Participation Rate (X5), Women's Expected Years of Schooling (X6), and Adjusted Per Capita Expenditure for Women (X7). These variables were chosen for their relevance to Sustainable Development Goal (SDG) 5, which emphasizes the achievement of gender equality and the empowerment of women and girls.

2.2. Data Preprocessing

In this study, the dataset was first standardized using the StandardScaler to ensure that all features were on the same scale. StandardScaler standardizes the features by removing the mean and scaling them to unit variance, resulting in a distribution with a mean of 0 and a standard deviation of 1 (11). The scaling is performed according to the following formula:

$$Z_{scaled} = \frac{X - \mu}{\sigma}$$

where μ is the mean and σ is the standard deviation of the feature. This step is crucial as it prevents features with larger ranges from dominating the learning process, ensuring that all features contribute equally to the model (12).

After standardization, the data was split into 80% training data and 20% testing data. The training set was used to train the machine learning models, while the testing set was reserved for evaluating the models' performance and ensuring their generalizability to unseen data.

2.3. Analytical Method

Three machine learning models were employed to analyze the relationship between the socio-economic variables and the Gender Inequality Index, namely Support Vector Regression (SVR), AdaBoost Regressor, and Random Forest Regressor. The models were selected based on their ability to handle complex interactions among variables.

1. Support Vector Regression (SVR)
   Support Vector Regression (SVR) is an extension of the support vector machine (SVM) technique, which is widely applied for both linear and non-linear regression tasks (13). SVR operates by finding a hyperplane in a high-dimensional feature space that best fits the data points while maintaining a margin of tolerance defined by a parameter, epsilon (ε), around the predicted function (14). The general form of the hyperplane equation is given by:

   $$y = wX + b$$

   where w represents the weights and b is the intercept at $X = 0$. The tolerance margin defined by ε, controls how much deviation from the hyperplane is allowed.

2. AdaBoost Regressor
   The AdaBoost Regressor (ABR) is an ensemble learning method that improves predictive performance by combining multiple weak learners into a stronger model. Originally developed for classification tasks, AdaBoost has been extended to regression problems where it sequentially builds a series of decision trees (15). Initially, all data points have equal weights. After training the first weak learner, errors are calculated, and higher weights are given to misclassified points. Subsequent learners are trained on the adjusted dataset, focusing on reducing errors in challenging instances. The final model aggregates the predictions of all weak learners, assigning greater weight to those that perform better (16).

3. Random Forest Regressor
   The Random Forest Regressor (RFR) is an ensemble learning method that combines multiple decision trees to improve the accuracy and robustness of predictions. It works by constructing a large number of decision trees during training, where each tree is built from a random subset of the data. Each tree makes an independent prediction, and the final output of the Random Forest model is the average (for regression) of all individual tree predictions (17). By considering random subsets of both data points and features, Random Forest prevents over-reliance on specific variables, thus improving model performance and reducing variance (9).

2.4. Performance Evaluation

The performance of the machine learning models employed in this study was evaluated using the Mean Squared Error (MSE) as the primary metric. MSE is a common metric used to assess the accuracy of a regression model by measuring the average squared differences between the actual and predicted values. These differences, known as "errors," are squared to ensure all values are positive and to give greater weight to larger errors. A lower MSE indicates that the predictions are closer to the actual values, suggesting a better model fit (18). The MSE is calculated as follows:

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

Where Y_i represents the observed values, $\hat{Y}_i$ denotes the predicted value, and n is the total number of observations.

3. Results and Discussion

3.1. Result

Before conducting the analysis, the independent variables were normalized using StandardScaler. This transformation is essential for models like Support Vector Regression (SVR) and AdaBoost Regressor, which are sensitive to the scale of the input data. Then, the dataset was split into training (80%) and testing (20%) sets, ensuring that the machine learning models could be trained on a large portion of the data while maintaining an unbiased testing set for evaluation. The correlation heatmap in Figure 3 illustrates the relationships between the dependent variable Y and seven independent variables.

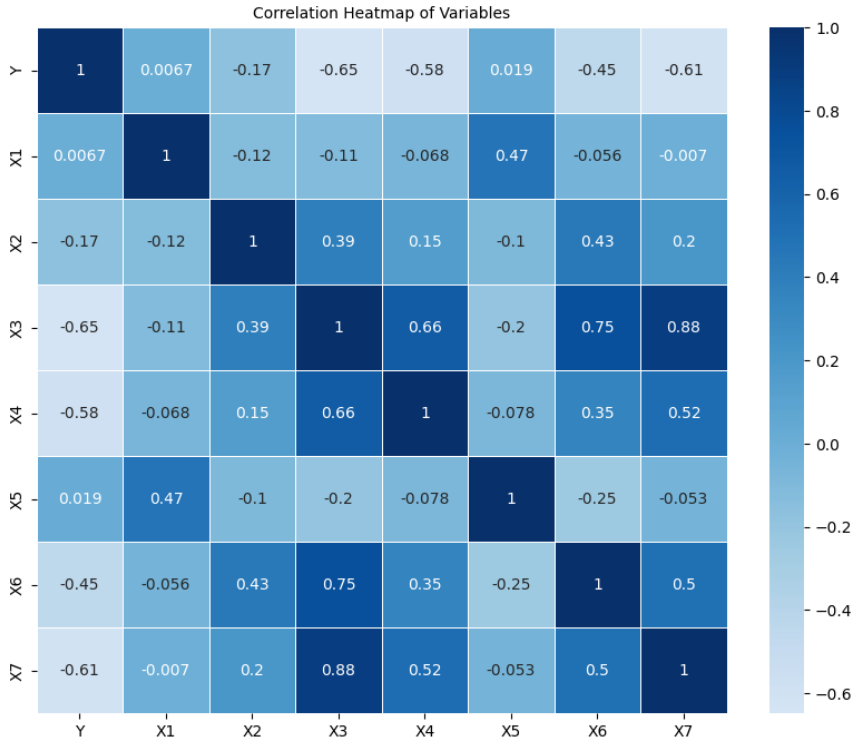

Figure 3. Correlation Heatmap of Factors Affecting Gender Inequality in Indonesia

Based on the heatmap, strong negative correlations are observed between GII and Women's Human Development Index (X3), Women's Life Expectancy (X4), and Adjusted Per Capita Expenditure for Women (X7), indicating that higher human development, life expectancy, and economic well-being among women are linked to lower gender inequality. Women's Expected Years of Schooling (X6) and Women in Professional Occupations (X2) shows a moderate negative correlation with GII, while Women's Income Contribution (X1) and Female Labor Force Participation (X5) have minimum impact.

In response to the first research question, three machine learning models—Support Vector Regression, AdaBoost Regressor, and Random Forest Regressor—were applied to predict the Gender Inequality Index (GII) across Indonesia's districts and cities. The performance of these models was

evaluated using Mean Squared Error (MSE). As shown in Table 1, the Tuned Random Forest Regressor achieved the lowest MSE of 0.0109, making it the best-performing model compared to SVR and AdaBoost Regressor, with MSE values of 0.0133 and 0.0119, respectively. These results demonstrate that the Random Forest Regressor outperformed the other models in capturing the complex relationships between the socio-economic variables and GII.

Table 1. Performance Comparison of Machine Learning Models Based on Mean Squared Error

Model	MSE
Support Vector Regression	0.0133
AdaBoost Regressor	0.0119
Random Forest Regressor	0.0109

To address the second research question, a feature importance analysis was conducted using the Tuned Random Forest Regressor to determine the most influential factors affecting GII. The analysis revealed that the Women's Human Development Index (X3) had the greatest impact on the GII, contributing 43.92% to the model's predictions. Women's Life Expectancy (X4) was the second most significant factor, with a contribution of 23.25%, followed by Adjusted Per Capita Expenditure for Women (X7), which accounted for 10.27%. Other contributing factors, such as Women's Labor Force Participation (X5) and Women's Expected Years of Schooling (X6), also had significant but relatively smaller effects, reinforcing the multifaceted nature of gender inequality. These results are visualized in Figure 3, emphasizing the prominent role of socio-economic factors in determining GII levels across Indonesian districts and cities.

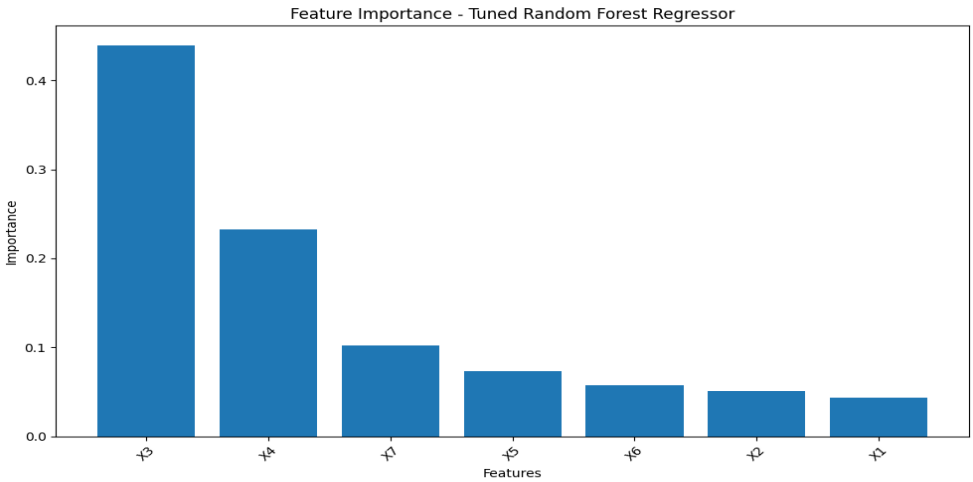

Figure 3. Dominant Factors Affecting Gender Inequality in Indonesia

3.2. Discussion

The results of this study highlight the effectiveness of the Tuned Random Forest Regressor in predicting the Gender Inequality Index (GII) across Indonesia's districts and cities, with a Mean Squared Error (MSE) of 0.0109, outperforming other models like SVR and AdaBoost Regressor. These findings align with previous studies that have demonstrated the superior performance of ensemble methods, particularly Random Forest, in handling complex interactions among variables (19). Compared to traditional regression methods, which often assume linear relationships, Random Forest is more flexible

due to its decision tree-based approach (9). This ability to model intricate dependencies among socio-economic variables explains why the Random Forest model performed better than SVR and AdaBoost in this study, reflecting findings in other contexts, such as medical waste prediction (20).

The feature importance analysis revealed that the Women's Human Development Index (X3) was the most significant predictor of GII, contributing 43.92% to the model. This finding is consistent with prior research that highlights the critical role of human development in reducing gender inequality (21). Aspects of human development, such as education, healthcare, and access to social services, are often seen as foundational in improving women's empowerment and participation in the labor market (22). This result underscores the importance of continuing efforts to enhance women's access to education and health services, as these are key drivers of gender equality.

The second most influential factor was Women's Life Expectancy (X4), contributing 23.25% to GII predictions. This aligns with literature that emphasizes the relationship between women's health outcomes and gender equality (22). Longer life expectancy often reflects better access to healthcare and social services, which in turn supports women's greater participation in economic and political life. Health improvements lead to enhanced productivity and societal contribution, reinforcing the overall goal of gender equality. Adjusted Per Capita Expenditure for Women (X7) was also a significant factor, with a contribution of 10.27%. This highlights the economic dimensions of gender inequality, showing that financial independence and economic access are key to improving gender equality (23). Economic empowerment, particularly through increased earnings and economic participation, has been shown to reduce inequality and enhance women's agency in various contexts.

The findings suggest that women's human development and health outcomes are the most critical factors to address when targeting gender inequality in Indonesia. Policies aimed at improving access to education, healthcare, and social services for women are likely to have the greatest impact in reducing the GII. Moreover, economic policies that focus on enhancing women's financial independence, such as increasing their income contribution and employment in professional sectors, should be prioritized to close gender gaps. These insights offer a data-driven basis for policymakers to design targeted interventions that address the root causes of gender inequality. By using machine learning models like Random Forest, decision-makers can more accurately forecast where interventions will be most effective, making policy efforts more efficient and impactful.

4. Conclusion

This study employed machine learning models to predict the Gender Inequality Index (GII) across Indonesia’s districts and cities, identifying key socio-economic factors that influence gender inequality. Among the models used, the Tuned Random Forest Regressor exhibited the best performance, with the lowest Mean Squared Error (MSE) of 0.0109, outperforming other models such as Support Vector Regression and AdaBoost Regressor. This highlights the Random Forest model’s ability to capture the complex, non-linear relationships between variables related to gender inequality. The feature importance analysis identified that the Women's Human Development Index (X3) and Women's Life Expectancy (X4) were the most significant contributors to GII, reinforcing the critical role of education, healthcare, and overall quality of life in reducing gender disparities. Additionally, economic factors, particularly Adjusted Per Capita Expenditure for Women (X7), also played a significant role, emphasizing the need for economic empowerment to bridge the gender gap. The findings of this study provide valuable insights for policymakers seeking to address gender inequality in Indonesia. Interventions that focus on improving women’s human development, health, and economic participation are likely to yield the most significant reductions in gender disparities. This research not only contributes to the existing literature on gender

inequality but also demonstrates the utility of machine learning models, particularly Random Forest, in socio-economic policy analysis.

5. Conflict of Interest

The authors declare that there is no conflict of interest related to the writing or publication of this article.

6. Acknowledgement

The authors would like to express their gratitude to the Statistics Indonesia (BPS) for providing the data and resources necessary for this study. The authors also acknowledge the support from Department of Statistics Universitas PGRI Adi Buana Surabaya and Faculty of Computer Science at Universitas Dian Nuswantoro for providing the necessary resources, guidance, and support throughout this study.

7. References

[1] Enaifoghe A, Maseko TI. The challenges of gender inequality on sustainable development in leadership in Africa. EUREKA: Social and Humanities. 2023 Sep 30;(5):56–66.

[2] Ali, Audi, Marc, Chan B, Roussel, Yannick. The Impact of Gender Inequality and Environmental Degradation on Human Well-Being in The Case of Pakistan: A Time Series Analysis [Internet]. 2021 [cited 2024 Oct 25]. Available from: https://mpra.ub.uni-muenchen.de/106655/

[3] Birdthistle N, Hales R. Attaining the 2030 Sustainable Development Goal of Gender Equality. Birdthistle N, Hales R, editors. Emerald Publishing Limited; 2024.

[4] Hastuti D, Sudrajat. Gender Gap in Education and Employment in Asia: Indonesia and South Korea Compared. Proceedings of the International Conference of Social Science and Education (ICOSSED 2021). 2023;116–21.

[5] Braverman-Bronstein A, Ortigoza AF, Vidaña-Pérez D, Barrientos-Gutiérrez T, Baldovino-Chiquillo L, Bilal U, et al. Gender inequality, women's empowerment, and adolescent birth rates in 363 Latin American cities. Soc Sci Med. 2023 Jan;317:115566.

[6] Vlasceanu M, Amodio DM. Propagation of societal gender inequality by internet search algorithms. Proceedings of the National Academy of Sciences. 2022 Jul 19;119(29).

[7] Agrawal H. GENDER INEQUALITY IN ARTIFICIAL INTELLIGENCE: FEMINIST POLITICAL ECONOMIC APPROACH. International Journal Of Creative and Innovative Research In All Studies [Internet]. 2024 [cited 2024 Oct 25];7(3). Available from: http://www.ijciras.com/PublishedPaper/IJCIRAS1983.pdf

[8] Jassim MS, Coskuner G, Zontul M. Comparative performance analysis of support vector regression and artificial neural network for prediction of municipal solid waste generation. Waste Management & Research: The Journal for a Sustainable Circular Economy. 2022 Feb 4;40(2):195–204.

[9] Shanmugasundar G, Vanitha M, Čep R, Kumar V, Kalita K, Ramachandran M. A comparative study of linear, random forest and adaboost regressions for modeling non-traditional machining. Processes. 2021 Nov 1;9(11).

[10] Shen X, Yin F, Jiao C. Predictive Models of Life Satisfaction in Older People: A Machine Learning Approach. Int J Environ Res Public Health. 2023 Feb 1;20(3).

[11] Raju VNG, Lakshmi KP, Jain VM, Kalidindi A, Padma V. Study the Influence of Normalization/Transformation process on the Accuracy of Supervised Classification. In: 2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT). IEEE; 2020. p. 729–35.

[12] Shantal M, Othman Z, Bakar AA. A Novel Approach for Data Feature Weighting Using Correlation Coefficients and Min–Max Normalization. Symmetry (Basel). 2023 Dec 1;15(12).

[13] Montesinos López OA, Montesinos López A, Crossa J. Support Vector Machines and Support Vector Regression. Multivariate Statistical Machine Learning Methods for Genomic Prediction. Cham: Springer International Publishing; 2022.

[14] Parbat D, Chakraborty M. A python based support vector regression model for prediction of COVID19 cases in India. Chaos Solitons Fractals. 2020 Sep;138(109942).

[15] Khan AA, Chaudhari O, Chandra R. A review of ensemble learning and data augmentation models for class imbalanced problems: Combination, implementation and evaluation. Expert Syst Appl. 2024 Jun;244:122778.

[16] Gupta KK, Kalita K, Ghadai RK, Ramachandran M, Gao XZ. Machine Learning-Based Predictive Modelling of Biodiesel Production-A Comparative Perspective. Energies (Basel). 2021 Feb 1;14(4).

[17] Saleh MdA, Rasel HM. Performance evaluation of Machine Learning based regression models for rainfall forecasting [Internet]. 2024 Jan. Available from: https://www.researchsquare.com/article/rs-3856741/v1

[18] Khan M, Noor S. Performance Analysis of Regression-Machine Learning Algorithms for Predication of Runoff Time. Agrotechnology. 2019;08(01).

[19] Khosravi M, Arif S Bin, Ghaseminejad A, Tohidi H, Shabanian H. Performance Evaluation of Machine Learning Regressors for Estimating Real Estate House Prices [Internet]. 2022. Available from: https://www.preprints.org/manuscript/202209.0341/v1

[20] Erdebilli B, Devrim-İçtenbaş B. Ensemble Voting Regression Based on Machine Learning for Predicting Medical Waste: A Case from Turkey. Mathematics. 2022 Jul 1;10(14).

[21] Osakede UA, Aramide VO, Adesipo AE, Akunna LC. Correlates of human development in Africa: Evidence across gender and income group. Research in Globalization. 2023 Jun 1;6.

[22] Al Azies H. Gender Equality and Women's Participation in Indonesian Employment: An Empirical Study from A Spatial Perspective. In 2023. p. 358–68.

[23] Barnat N, MacFeely S, Peltola A. Comparing Global Gender Inequality Indices: How Well Do They Measure the Economic Dimension? Journal of Sustainability Research. 2019;1(2).