Published by: Institute of Computer Science (IOCS)

Jurnal Teknik Informatika C.I.T Medicom

Journal homepage: www.medikom.iocspublisher.org

National Poverty Severity and Depth Index Analysis using PySpark Linear Regression Method

Febro Herdyanto[1]

[1)]Teknik Informatika, Pelita Bangsa University, West Java, Indonesia

Article Info

Article history

Received : diisi oleh editor
Revised : diisi oleh editor
Accepted : diisi oleh editor

Kata Kunci:

Linear Regression;
National Poverty;
PySpark Analysis;
Severity and Depth Index

Abstract

This research presents a thorough examination of poverty indices in Indonesia, employing PySpark for a comprehensive regional and sub-regional analysis. The primary objective is to unveil nuanced patterns and determinants influencing poverty, providing evidence-based insights for targeted policy interventions. The initial data preprocessing involves meticulous steps, including handling missing values and casting columns to ensure data quality. The exploration phase centers on Jawa Barat, where predictions are visually depicted through bar charts for both "Index Kedalaman Kemiskinan" and "Index Keparahan Kemiskinan." Subsequently, the analysis extends to the sub-regional level (kabupaten/kota), where distinct linear regression models are trained and visualized using line charts. The emphasis is on highlighting the top 5 regions displaying substantial reductions in poverty indices. The research findings contribute valuable insights into regional and sub-regional poverty variations, supporting evidence-based decision-making for effective poverty alleviation strategies in Indonesia. By leveraging machine learning techniques, this study demonstrates the potential for predictive models to offer actionable insights, enriching the ongoing discourse on poverty alleviation and providing a robust foundation for evidence-based policymaking.

Corresponding Author:

Febro Herdyanto,
Teknik Informatika
Pelita Bangsa Univeristy
Jl. Inspeksi Kalimalang Tegal Danas, Bekasi, West Java, Indonesia
febroherdyanto@mhs.pelitabangsa.ac.id

This is an open access article under the CC BY-NC license.

1. Introduction

Poverty remains a persistent and complex challenge in Indonesia, necessitating rigorous attention and comprehensive efforts for alleviation[1]. Despite ongoing initiatives and programs aimed at poverty reduction, the multifaceted nature of this issue demands in-depth analysis. Contributing factors include uneven economic growth, low education levels, regional disparities, and the effectiveness of public policies[2].

Unequal economic growth exacerbates economic disparities, leading to income inequality. Insufficient education impedes individuals' access to quality employment, while regional inequalities contribute to varying poverty levels across different areas. Assessing the impact of public policies on economic development is crucial in the context of poverty reduction. Effective public policies can significantly contribute to addressing poverty-related issues[1]. Evaluating the effectiveness of implemented policies becomes a crucial step in identifying weaknesses and assessing their impact on poverty rates in Indonesia[3].

This research utilizes data from the Badan Pusat Statistik (BPS) Indonesia, spanning from 2015 to 2023. The primary focus is on the Depth and Severity Indices of Poverty, considered a robust foundation for understanding the dynamics and characteristics of poverty in Indonesia over recent years[4].

By employing Linear Regression and PySpark methods, this study aims to provide new insights into the relationships between economic growth, education levels, regional disparities, public policies, and poverty rates[5]. A nuanced understanding of these factors is expected to lead to the formulation of targeted and effective policies for addressing the complex issue of poverty in Indonesia[6].

2. Research Method

The research methodology employed in this study is structured to comprehensively address key issues contributing to poverty in Indonesia. The analysis begins with a thorough examination of uneven economic growth, focusing on data collected from various regions and sectors between 2015 and 2023[7]. Statistical tools, particularly Linear Regression, are applied to assess the contributions of both informal and formal sectors to economic growth, providing insights into their implications for income distribution and poverty[8].

Subsequently, the research delves into the impact of low education levels on poverty. Education-related data from the Badan Pusat Statistik (BPS) is scrutinized, emphasizing the evaluation of educational programs and regional accessibility. The analysis aims to discern the correlation between education levels and employment opportunities, thereby shedding light on the role of education in poverty alleviation[9].

Regional inequality, a critical concern in poverty management, is then addressed. Poverty distribution data across different regions is collected and analyzed to evaluate the extent to which regional disparities contribute to overall poverty rates[10]. This analysis assists in identifying specific regions experiencing high levels of poverty, forming the basis for the design of targeted policies tailored to local conditions[11].

Furthermore, the study assesses the effectiveness of public policies as a key factor in poverty reduction[12]. Data on implemented public policies, welfare programs, economic incentives, and educational aid are gathered and subjected to rigorous evaluation. The focus is on measuring the impact and sustainability of these policies, providing valuable insights into their role in addressing poverty-related challenges[13].

In terms of data collection, the study relies on comprehensive data from the Badan Pusat Statistik (BPS) for the period 2015-2023. The main variables of interest are the Depth and Severity Indices of Poverty, providing a nuanced understanding of the multifaceted nature of poverty in Indonesia[14].

The methodology involves integrating economic and social datasets, utilizing the strengths of PySpark for efficient data processing[15]. The use of Linear Regression as the primary analytical tool facilitates the identification of relationships between key variables, such as economic growth, education levels, and regional disparities, and the Depth and Severity Indices of Poverty[16].

The research also incorporates a predictive element, employing model testing and experimentation using Linear Regression. The dataset is divided into training and testing sets to evaluate the model's predictive capabilities. Model evaluation metrics such as Mean Squared Error (MSE) or R-squared are utilized, and adjustments are made to enhance the model's predictive performance based on the evaluation results[17].

In conclusion, the proposed research methodology combines quantitative analysis, statistical tools, and predictive modeling to offer a comprehensive understanding of the complex dynamics of poverty in Indonesia[18]. The structured approach ensures the reliability and relevance of the findings, contributing valuable insights for policymakers and stakeholders involved in poverty alleviation efforts[19].

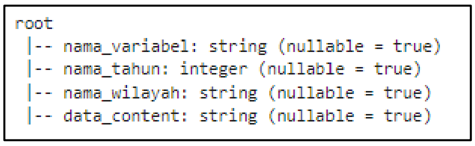
3. Result and Discussion

Prior to delving into the intricate details of Data Preprocessing and Exploration, the initial phase of our analysis involved the careful curation and loading of the poverty indices dataset[20]. This foundational step ensured a comprehensive understanding of the dataset's structure and laid the groundwork for subsequent insightful examinations

a. Data Preprocessing and Exploration
The initial phase of the analysis involved the preprocessing of the dataset using PySpark. The dataset, containing information on poverty indices, was loaded into a PySpark DataFrame[21]. The schema of the dataset was examined to understand the structure and types of variables. To ensure data quality, missing values (null or NaN) in the 'data_content' column were identified and subsequently removed from the dataset. The 'nama_tahun' and 'data_content' columns were cast to float types for further analysis.

```
df.printSchema()
```

Figure 1. Code for print schema of dataset.

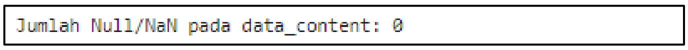


```
root
 |-- nama_variabel: string (nullable = true)
 |-- nama_tahun: integer (nullable = true)
 |-- nama_wilayah: string (nullable = true)
 |-- data_content: string (nullable = true)
```

Figure 2. Output schema of dataset.

```
null_count = df.filter(col("data_content").isNull() | isnan("data_content")).count()
print(f"Jumlah Null/NaN pada data_content: {null_count}")
```

Figure 3. Code to display the number of null/NaN values in the data_content column.



```
Jumlah Null/NaN pada data_content: 0
```

Figure 4. Output of display the number of null/NaN values in the data_content column.

```
columns_to_cast = ["data_content", "nama_tahun"]
for col_name in columns_to_cast:
df = df.withColumn(col_name, df[col_name].cast("float"))
```

Figure 5. Convert data types of data_content, nama_tahun, and other variables to float.

b. Region-specific Analysis
The analysis focused on a specific region, in this case, "JAWA BARAT." The dataset was filtered to include only relevant entries for this region. The number of rows after filtering was verified to ensure sufficient data for model training. Subsequently, a Vector Assembler was employed to combine the 'nama_tahun' column into a feature vector. A Linear Regression model was then trained to predict the 'data_content' values.

```
# Ensure there is training data
if jumlah_baris_setelah_filtering > 0:

    # Use VectorAssembler to combine columns into a feature vector
    assembler = VectorAssembler(inputCols=["nama_tahun"], outputCol="features")
    df_assembled = assembler.transform(df_wilayah)
    # Linear Regression Model
    lr = LinearRegression(featuresCol="features", labelCol="data_content")
    model = lr.fit(df_assembled)

    # Predictions
    predictions = model.transform(df_assembled)

    # Fetch data for plotting
    actual_data = df_wilayah.select("nama_tahun", "data_content").toPandas()
    predicted_data = predictions.select("nama_tahun", "prediction").toPandas()

    # Plotting using a bar chart
    plt.figure(figsize=(12, 6))

    # Chart for Poverty Depth Index
    plt.subplot(1, 2, 1)
    plt.bar(actual_data["nama_tahun"], actual_data["data_content"], label="Actual")
    plt.bar(predicted_data["nama_tahun"], predicted_data["prediction"], label="Predicted", alpha=0.7)
    plt.title(f"Index Kedalaman Kemiskinan - {wilayah}")
    plt.xlabel("Tahun")
    plt.ylabel("Index Value")
    plt.legend()

    # Chart for Poverty Severity Index
    plt.subplot(1, 2, 2)
    plt.bar(actual_data["nama_tahun"], actual_data["data_content"], label="Actual")
    plt.bar(predicted_data["nama_tahun"], predicted_data["prediction"], label="Predicted", alpha=0.7)
    plt.title(f"Index Keparahan Kemiskinan - {wilayah}")
    plt.xlabel("Tahun")
    plt.ylabel("Index Value")
    plt.legend()
    plt.tight_layout()
    plt.show()

else:
    print(f"Tidak ada data yang cukup untuk melatih model untuk {wilayah}.")
```

Figure 6. Data Filtering, Feature Vector Creation, and Model Training for Region-specific

c. Model Evaluation

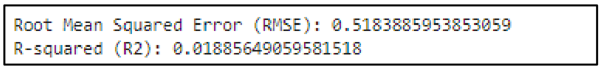
The trained model's performance was evaluated using standard metrics, including Root Mean Squared Error (RMSE) and R-squared (R2)[22]. These metrics provide insights into the accuracy and goodness of fit of the regression mode[23]l.

```
# Model Evaluation
evaluation = model.evaluate(df_assembled)

# Fetch evaluation metrics
rmse = evaluation.rootMeanSquaredError
r2 = evaluation.r2

# Display evaluation metrics
print(f"Root Mean Squared Error (RMSE): {rmse}")
print(f"R-squared (R2): {r2}")
```

Figure 7. Code to check Model Evaluation



```
Root Mean Squared Error (RMSE): 0.5183885953853059
R-squared (R2): 0.01885649059581518
```

Figure 8. Output of Model Evaluation

d. Visualization of Predictions

The predictions of the model were visualized using bar charts for both the "Index Kedalaman Kemiskinan" and "Index Keparahan Kemiskinan." The actual and predicted values were plotted side by side, providing a clear comparison over the years. This visualization aids in understanding the model's ability to capture the underlying patterns in the data.

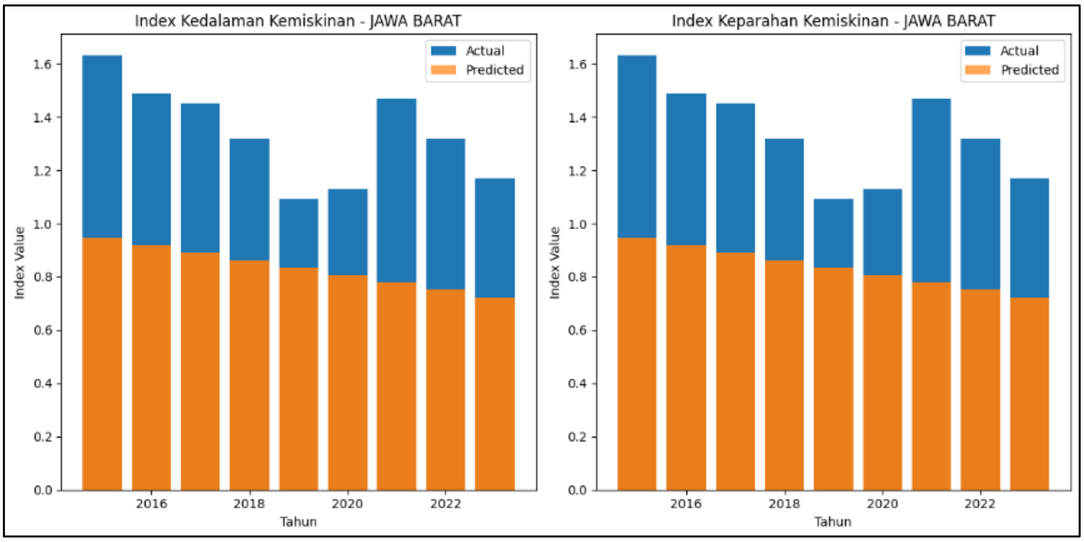


Figure 9. Visualization of Region-specific analysis

e. Regional Disparities

To address regional disparities, the analysis extended to focus on provinces. Separate models were trained for both "Index Kedalaman Kemiskinan" and "Index Keparahan Kemiskinan" at the provincial level. Line charts were created using Plotly Express, showcasing the actual values over time and overlaying them with predicted values.

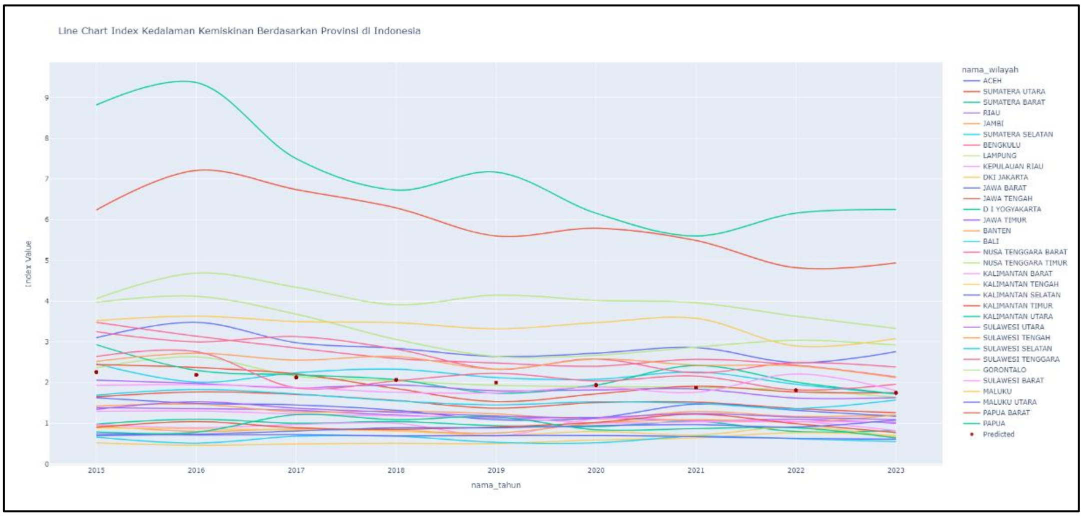


Figure 10. Provincial Analysis with Predictive Modeling - Index of Poverty Depth

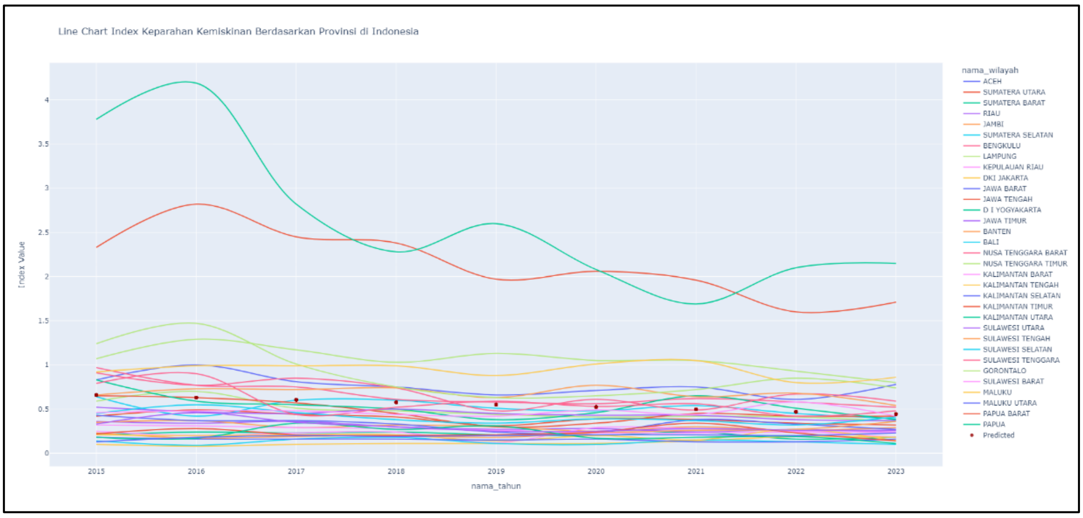


Figure 11. Provincial Analysis with Predictive Modeling - Index of Poverty Severity

f. Top 5 Regions with Significant Reduction
An additional analysis highlighted the top 5 regions with the most significant reduction in poverty indices. Line charts were generated for both depth and severity indices, displaying the actual and predicted values. This detailed analysis of specific regions provides valuable insights for policymakers and stakeholders

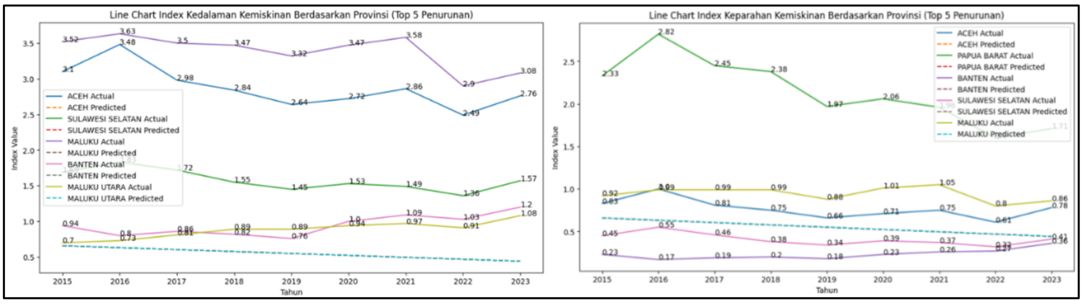


Figure 12. Top 5 Regions Analysis - Index of Poverty Depth and Severity

g. Sub-Regional Analysis
The analysis delved further into sub-regional data, specifically focusing on kabupaten/kota (districts/cities). Linear Regression models were trained separately for both "Index Kedalaman Kemiskinan" and "Index Keparahan Kemiskinan" at this level. Visualizations, including line charts, illustrated the predicted and actual values, emphasizing the top 5 regions with substantial reductions.

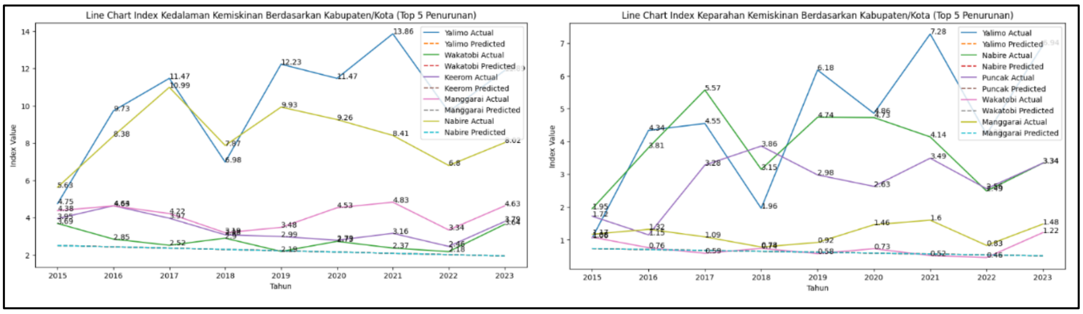


Figure 13. Top 5 Sub-Regions Analysis - Index of Poverty Depth and Severity

h. Insights and Implications

The results from this comprehensive analysis offer insights into the factors influencing poverty indices in Indonesia. Regional and sub-regional variations have been explored, providing a nuanced understanding of the poverty landscape. The predictive models demonstrate the potential for leveraging machine learning techniques to gain actionable insights for targeted policy interventions. These findings contribute to the ongoing discourse on poverty alleviation strategies in Indonesia and provide a foundation for evidence-based decision-making.

4. Conclusion

This research sheds light on the intricate dynamics of poverty in Indonesia, offering a comprehensive understanding of contributing factors and potential solutions. Unequal economic growth, low education levels, regional disparities, and policy effectiveness were identified as crucial elements influencing poverty rates[24].

The analysis, spanning from 2015 to 2023, utilized data from the Badan Pusat Statistik (BPS), focusing on the Depth and Severity Indices of Poverty. Employing Linear Regression and PySpark methods, the study unraveled insights into the relationships between economic growth, education, regional variations, and public policies with poverty rates.

The research methodology, incorporating statistical tools and predictive modeling, facilitated a nuanced exploration of economic disparities and educational impacts on poverty. It further highlighted regional inequalities and critically assessed the effectiveness of public policies.

Key findings included the exacerbation of income inequality due to uneven economic growth, the role of education in shaping employment opportunities, and the need for targeted policies addressing regional disparities. The predictive models, especially at regional and sub-regional levels, provided actionable insights for policymakers[25].

Model evaluations using metrics like Root Mean Squared Error (RMSE) and R-squared demonstrated the reliability of the Linear Regression approach. Visualizations, including bar and line charts, enhanced the interpretation of predictions and regional disparities, emphasizing the top 5 regions with substantial poverty reductions[26].

In essence, the research contributes to the ongoing discourse on poverty alleviation in Indonesia, offering evidence-based insights for policymakers and stakeholders. The findings underscore the importance of tailored policies that consider regional nuances and the multifaceted nature of poverty, paving the way for effective strategies in the country's pursuit of sustainable development and poverty reduction.

References

[1] T. Agus Triono, R. Candra Sangaji, D. Program, and F. Bisnis dan Ekonomi, "Faktor Mempengaruhi Tingkat Kemiskinan di Indonesia: Studi Literatur Laporan Data Kemiskinan BPS Tahun 2022," *Journal of Society Bridge*, vol. 1, no. 1, pp. 59–67, Jan. 2023, doi: 10.59012/JSB.V1I1.5.

[2] J. Suprijati and S. R. Damayanti, "PENGENTASAN KEMISKINAN KOTA DAN DESA 31 PROPINSI DI INDONESIA MELALUI PERTUMBUHAN EKONOMI YANG DIBENTUK DARI PMDN DAN PEKERJA," 2022.

[3] J. Perbendaharaan *et al.*, "INDONESIAN TREASURY REVIEW PENGARUH DANA DESA TERHADAP KEMISKINAN: STUDI TINGKAT KABUPATEN/KOTA DI INDONESIA."

[4] "Badan Pusat Statistik Indonesia." Accessed: Jan. 08, 2024. [Online]. Available: https://www.bps.go.id/id

[5] R. K. Mishra, "PySpark MLlib and Linear Regression," *PySpark Recipes*, pp. 235–259, 2018, doi: 10.1007/978-1-4842-3141-8_9.

[6] H. S. Kudale, M. V Phadnis, P. J. Chittar, K. P. Zarkar, and B. K. Bodhke, "A REVIEW OF DATA ANALYSIS AND VISUALIZATION OF OLYMPICS USING PYSPARK AND DASH-PLOTLY," 2093. [Online]. Available: www.irjmets.com

[7] T. Agus Triono, R. Candra Sangaji, D. Program, and F. Bisnis dan Ekonomi, "Faktor Mempengaruhi Tingkat Kemiskinan di Indonesia: Studi Literatur Laporan Data Kemiskinan BPS Tahun 2022." [Online]. Available: https://www.bk3s.org/ojs/index.php/jsb

[8] A. Karim, "Perbandingan Prediksi Kemiskinan di Indonesia Menggunakan Support Vector Machine (SVM) dengan Regresi Linear," *Jurnal Sains Matematika dan Statistika*, vol. 6, no. 1, 2020.

[9] A. Mufida, S. Retno Faridatussalam, and J. A. Yani Tromol Pos, "ANALISIS TINGKAT KEMISKINAN MASYARAKAT PROVINSI BANTEN TAHUN 2015-2019," *Jurnal Pendidikan Sejarah dan Riset Sosial Humaniora*, vol. 2, no. 2, pp. 34–45, 2021, Accessed: Jan. 08, 2024. [Online]. Available: https://ejournal.penerbitjurnal.com/index.php/humaniora/article/view/58

[10] D. Sari, "Poverty Mapping And Poverty Analysis In Indonesia", doi: 10.21082/jae.v28n1.2010.95-111.

[11] S. A. Mahfuza, Z. Azmi, and G. Syahputra, "Data Mining Untuk Mengestimasi Angka Kemiskinan Di Sumatera Utara Menggunakan Metode Regresi Linier Berganda," *Jurnal CyberTech*, vol. 4, no. 6, 2021, [Online]. Available: https://ojs.trigunadharma.ac.id/

[12] A. Nikola Putra, H. Fricylya Br Tobing, O. Sanityasa Rahajeng, and R. Julaeni Yuhan, "The Indonesian Journal of Social Studies Penerapan Path Analysis terhadap Faktor-Faktor yang Mempengaruhi IPM dan Kemiskinan di Indonesia Tahun 2019," vol. 3, no. 1, pp. 37–45, 2020, Accessed: Jan. 08, 2024. [Online]. Available: https://journal.unesa.ac.id/index.php/jpips/index

[13] S. Riaman, S. Supian, and A. Talib Bon, "Poverty Level Analysis in Indonesia Using the Stochastic Restricted Maximum Likelihood Approach Method".

[14] Riaman, S. Supian, Sukono, and A. T. Bon, "Poverty Level Analysis in Indonesia Using the Stochastic Restricted Maximum Likelihood Approach Method," *Proceedings of the International Conference on Industrial Engineering and Operations Management*, pp. 4077–4086, 2021, doi: 10.46254/AN11.20210732.

[15] A. Testas, "Multiple Linear Regression with Pandas, Scikit-Learn, and PySpark," *Distributed Machine Learning with PySpark*, pp. 53–74, 2023, doi: 10.1007/978-1-4842-9751-3_3.

[16] D. Maulud and A. M. Abdulazeez, "A Review on Linear Regression Comprehensive in Machine Learning," *Journal of Applied Science and Technology Trends*, vol. 1, no. 4, pp. 140–147, Dec. 2020, doi: 10.38094/jastt1457.

[17] O. : Murdiyana and D. Mulyana, "ANALISIS KEBIJAKAN PENGENTASAN KEMISKINAN DI INDONESIA," 2017. [Online]. Available: www.bps.go.id,

[18] S. G. Edoumiekumo, T. M. Karimo, and S. S. Tombofa, "Income Poverty in Nigeria: Incidence, Gap, Severity and Correlates," *American Journal of Humanities and Social Sciences*, vol. 2, no. 1, pp. 1–9, 2014, doi: 10.11634/232907811604499.

[19] F. Handoyo, A. Hidayatina, and P. Purwanto, "The Effect of Rural Development on Poverty Gap, Poverty Severity and Local Economic Growth in Indonesia," *Jurnal Bina Praja: Journal of Home Affairs Governance*, vol. 13, no. 3, pp. 369–381, Dec. 2021, doi: 10.21787/JBP.13.2021.369-381.

[20] "Review_of_Data_Preprocessing_Techniques".

[21] T. Asadollahi, S. Dadfarnia, A. M. H. Shabani, J. B. Ghasemi, and M. Sarkhosh, "QSAR Models for CXCR2 Receptor Antagonists Based on the Genetic Algorithm for Data Preprocessing Prior to Application of the PLS Linear Regression Method and Design of the New Compounds Using In Silico Virtual Screening," *Molecules 2011, Vol. 16, Pages 1928-1955*, vol. 16, no. 3, pp. 1928–1955, Feb. 2011, doi: 10.3390/MOLECULES16031928.

[22] D. Chicco, M. J. Warrens, and G. Jurman, "The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation," *PeerJ Comput Sci*, vol. 7, pp. 1–24, Jul. 2021, doi: 10.7717/PEERJ-CS.623/SUPP-1.

[23] H. Z. Abyaneh, "Evaluation of multivariate linear regression and artificial neural networks in prediction of water quality parameters," *J Environ Health Sci Eng*, vol. 12, no. 1, pp. 1–8, Jan. 2014, doi: 10.1186/2052-336X-12-40/FIGURES/7.

[24] J. S. Monang Tambun dan Rita Herawaty and F. Statistisi BPS Provinsi Sumatera Utara, "Pemodelan Faktor-Faktor yang Mempengaruhi Indeks Kedalaman Kemiskinan dan Indeks Keparahan Kemiskinan Kabupaten/Kota di Sumatera Utara Menggunakan Regresi Data Panel," *Publikauma : Jurnal Administrasi Publik Universitas Medan Area*, vol. 6, no. 1, pp. 100–110, Jun. 2018, doi: 10.31289/PUBLIKA.V6I1.1574.

[25] I. Ahmaddien and I. Ahmaddien, "Faktor determinan keparahan dan kedalaman kemiskinan jawa barat dengan regresi data panel," *FORUM EKONOMI: Jurnal Ekonomi, Manajemen dan Akuntansi*, vol. 21, no. 1, pp. 87–96, Mar. 2019, doi: 10.30872/jfor.v21i1.5225.

[26] N. Pokhriyal and D. C. Jacques, "Combining disparate data sources for improved poverty prediction and mapping," *Proc Natl Acad Sci U S A*, vol. 114, no. 46, pp. E9783–E9792, Nov. 2017, doi: 10.1073/PNAS.1700319114/SUPPL_FILE/PNAS.201700319SI.PDF.