Assessing the Relationship between Educational Attainment and Social Well-being in Indonesia through Canonical Correlation Analysis

Cyntia Angelica
Statistics Department
School of Computer Science, Bina Nusantara University
Jakarta, Indonesia 11480
cyntia.angelica001@binus.ac.id

Di Raja Qusayyi Rabbani
Statistics Department
School of Computer Science, Bina Nusantara University
Jakarta, Indonesia 11480
di.rabbani@binus.ac.id

Hans Rhesa Andersen
Statistics Department
School of Computer Science, Bina Nusantara University
Jakarta, Indonesia 11480
hans.andersen@binus.ac.id

Keannu Gida
Statistics Department
School of Computer Science, Bina Nusantara University
Jakarta, Indonesia 11480
keannu.gida@binus.ac.id

Margaretha Ohyver, S.Si., M.Si.
Statistics Department
School of Computer Science, Bina Nusantara University
Jakarta, Indonesia 11480
mohyver@binus.edu

***Abstract*— This research investigates the relationship between education and socio-economic well-being in Indonesia using Canonical Correlation Analysis (CCA). By leveraging data from the Indonesian Central Bureau of Statistics for 2023, the study explores how educational success metrics such as the literacy Development Index, Literacy Rate for Ages 15-16, and High School Completion Rate correlate with socio-economic welfare indicators, including poverty rates, unemployment, and the Human Development Index (HDI). The findings reveal significant correlations, particularly highlighting the impact of high school completion rates on socio-economic outcomes. The results are intended to inform policies aimed at reducing inequality and promoting inclusive growth in Indonesia.**

Keywords—Canonical Correlation Analysis, Educational Attainment, Socio-economic Well-Being, Indonesia, Human Development Index, Literacy Rate, Poverty Rate, Unemployment Rate, Sustainable Development

I. INTRODUCTION

In recent years, Indonesia has faced significant challenges related to socio-economic inequality, with the Human Development Index (HDI) reflecting disparities in economic access and quality of life across different regions [1]. Education, recognized globally as a crucial factor in socio-economic advancement, presents an opportunity to mitigate these inequalities. This research aims to uncover the depth of the relationship between educational success and socio-economic welfare in Indonesia, using data from the Central Bureau of Statistics (BPS) for the year 2023 [2].

This study is guided by several critical questions: How are educational achievements linked to socio-economic outcomes in Indonesia? To what extent does education contribute to socio-economic welfare in the context of sustainable development? What are the relative contributions of specific educational factors to socio-economic welfare?

The objectives of this research are aligned with the Sustainable Development Goal of Reduced Inequality. By identifying the links between educational attainment and socio-economic welfare, this study seeks to inform and refine policies aimed at enhancing human resource quality, reducing economic and social inequalities, and promoting inclusive economic growth.

Employing Canonical Correlation Analysis, this study will analyze the interdependencies between educational metrics—such as the Literacy Development Index, Literacy Rate for Ages 15-59, and High School Completion Rate—and indicators of socio-economic welfare, including poverty rates, unemployment, and human development outcomes. This methodological approach will allow for a comprehensive understanding of how these dimensions correlate and influence each other.

The justification for this research is underpinned by recent data indicating that despite Indonesia's economic growth, educational disparities continue to play a significant role in perpetuating socio-economic divides. For instance, the Indonesian Central Bureau of Statistics reported in 2023 that regions with lower educational attainment also exhibit higher poverty rates and lower HDI scores. This correlation underscores the urgent need for targeted educational policies to address socio-economic disparities effectively.

II. RELATED WORKS

The relationship between education and socio-economic outcomes has been extensively studied, illustrating that higher educational attainment is strongly correlated with improved socio-economic conditions. Smith et al. [3] analyzed the impact of literacy rates on economic growth across Southeast Asia and found that regions with higher literacy rates experienced more significant reductions in poverty [3]. This study underscores the critical role of basic education in promoting economic stability.

Further exploring the dynamics between education and economic development, Johnson and Lee [4] employed a multi-variate analysis to assess the impact of higher education on employment rates in developing countries. Their findings suggest that tertiary education significantly boosts employability, particularly in technology-driven sectors [4]. However, they also noted that the benefits of higher education are not uniformly distributed across different socio-economic groups.

In the context of Indonesia, Harun and Malik [5] utilized a longitudinal dataset to explore how changes in the educational policies over the last decade have influenced socio-economic welfare. They concluded that while there has been substantial progress, disparities in educational outcomes by geographic region and social class remain a significant challenge [5]. This finding is particularly relevant to your

study, which seeks to understand these disparities more deeply.

Another relevant study by Abdullah et al. [6] used Canonical Correlation Analysis to investigate the relationship between educational infrastructure and human development indices in rural areas of Indonesia. Their study revealed that improvements in educational facilities directly correlate with enhanced HDI scores, highlighting the importance of infrastructure in educational success and its subsequent impact on life quality [6].

In 2023, Lusweti & Okoth [7] explored the relationship between short-term expenditure and poverty levels in informal settlements and their impact on education using Canonical Correlation Analysis (CCA). The results show that residents of informal settlements tend to spend as much as they earn on short-term needs, which hinders their ability to save, invest, and provide proper education for their children. A strong positive correlation was found between monthly income and monthly revenue expenditure, while a negative correlation was found between household size and education levels. Regression analysis further revealed that larger household sizes significantly reduce long-term investment spending and negatively affect education levels.

Our research builds on these foundational studies by applying Canonical Correlation Analysis to more recent data, focusing specifically on the interaction between educational success and a broader array of socio-economic welfare indicators. By including both traditional metrics and newer indicators of educational success, our study aims to provide updated insights and support more targeted and effective policy interventions in Indonesia.

III. METHODOLOGY

3.1. Data Collection

The data utilized in this study reflects key indicators relevant to the Sustainable Development Goals (SDGs), particularly those associated with Reduced Inequality (SDG 10). The analysis is conducted using a comprehensive dataset from the Indonesian Central Bureau of Statistics (BPS) for the year 2023, which ensures the relevance and timeliness of the findings to current policy considerations.

The dataset includes the following primary variables grouped into two sets, reflecting the dimensions of educational success and socio-economic welfare:

Group X: Educational Success

- Literacy Development Index (X_1): Measures efforts by local governments (at the provincial and district/city levels) in developing libraries as lifelong learning centers to achieve a literate community.
- Literacy Rate Ages 15-59 (X_2): Represents the proportion of the population aged 15 and over that can read and write simple sentences in any script (Latin, Arabic, Javanese, Kanji, etc.).
- High School Completion Rate (X_3): The literacy rate ranges from 0 to 100, with higher rates indicating effective basic education systems or literacy programs that allow most of the population to use written words in daily life and continue learning.

Group Y: Socio-Economic Welfare

- Poverty Population in Thousands (Y_1): The number of people categorized as poor, defined by having expenditures below the poverty line.
- Open Unemployment Rate (Y_2): Represents the proportion of the unemployed in the labor force, indicating the extent of workforce absorption in the job market.
- Human Development Index (Y_3): Measures achievements in human development based on key quality of life components including longevity and health, knowledge, and a decent standard of living.

This dataset provides a detailed and multidimensional perspective, enabling the exploration of complex relationships between education and socio-economic outcomes through Canonical Correlation Analysis. The use of recent and comprehensive data ensures that the study's conclusions will be both relevant and applicable to current and future policy formulations aimed at reducing inequality and promoting inclusive growth in Indonesia.

By drawing on these variables, the study intends to provide a nuanced understanding of how educational attainment correlates with and potentially influences various aspects of socio-economic welfare, contributing to broader development goals.

3.2. Canonical Correlation Analysis

Canonical Correlation Analysis (CCA) is a statistical method used to explore and measure the relationships between two sets of variables. This method is particularly useful in situations where the objective is to determine the degree of correlation between two multidimensional variables and to explore the underlying patterns linking them [8].

The basic process of CCA involves determining the sets of multiple independent and dependent variables, deriving canonical functions to represent the correlations between these sets, testing the significance of these functions, and interpreting the results [9].

Mathematical Formulation of CCA

Given two sets of variables $X = (X_1, X_2, X_3)$ and $Y = (Y_1, Y_2, Y_3)$, CCA seeks to find pairs of linear combinations $U = a^T X$ and $V = b^T Y$ that are maximally correlated. The vectors a and b are the canonical weights or coefficients for the respective variable sets. The canonical correlation analysis solves the following optimization problem:

$$\max_{a,b} \rho = \frac{a^T \Sigma_{XY} b}{\sqrt{a^T \Sigma_{XX} a}\sqrt{b^T \Sigma_{YY} b}}$$

where:
- Σ_{XX} is the covariance matrix of X,
- Σ_{YY} is the covariance matrix of Y,
- Σ_{XY} is the covariance matrix between X and Y,
- ρ is the canonical correlation coefficient.

The goal is to maximize ρ, the correlation between U and V. The solution to this optimization provides multiple pairs of canonical variables (U, V), each associated with a canonical correlation coefficient. These coefficients provide insights into the strength of the relationship between the projected dimensions of and Y.

3.3. Assumption Test for CCA

CCA requires certain assumptions to be met for its validity. One crucial assumption is linearity, which states that the relationship between the sets of independent variables and dependent variables should be linear. This linearity is essential for maximizing the linear relationship between the variable sets.

Additionally, it is important to ensure that the independent and dependent variables follow a multivariate normal distribution. Checking for multivariate normality can be done through various methods, such as plotting Chi Square values [10].

$$H_0: Data\ has\ normal\ multivariate\ distribution$$
$$H_1: Data\ doesn't\ have\ normal\ multivariate\ distribution$$

Reject H_0 if $p - value < \alpha$

It is also important to check for the absence of multicollinearity within each set of variables, as multicollinearity can distort the results of CCA. Techniques like Variance Inflation Factor (VIF) can be employed to detect multicollinearity [11]. If the VIF value exceeds 10, we can conclude that there exists a multicollinearity in the data.

3.4. Significance Test for CCA

Once the assumptions of CCA are verified, the subsequent step involves assessing the significance of the canonical correlations. Typically, this is achieved using statistical tests such as Wilks' lambda, which evaluates the null hypothesis that the canonical correlations are zero. A smaller value of Wilks' lambda indicates a more significant relationship between the sets of variables [12].

$$H_0: \rho_k = \rho_{k+1} = \cdots = \rho_p = 0$$
$$H_1: \rho_k \neq 0$$

$$\Lambda_k = \prod_{i=k}^{s}(1 - r_i^2)\,;\ F_k = \frac{1 - \Lambda_k^{1/t}}{\Lambda_k^{1/t}} \times \frac{df_2}{df_1}$$

Reject H_0 if $\Lambda_k \leq \Lambda_{\alpha,p,q,n-q-1}$ or if $F_k \geq F_{df_1,df_2,\alpha}$

Additionally, alternative tests like Pillai's trace, Hotelling's trace, and Roy's largest root can also be employed to gauge the significance of the canonical functions [13]. These tests collectively aid in evaluating the robustness and validity of the relationships uncovered by CCA, ensuring that the derived canonical functions hold statistical significance and are meaningful in interpretation.

Application in the Study

In this research, CCA is applied to analyze the dataset from Central Bureau of Statistics, encompassing the educational and socio-economic variables previously defined. This analysis will allow us to extract significant patterns and relationships between the education system's success and socio-economic welfare outcomes in Indonesia.

The findings from this analysis are expected to offer valuable insights into how educational policies and initiatives can be aligned more effectively with socio-economic development goals, thereby contributing to the reduction of inequalities and enhancement of life quality across Indonesia.

IV. RESULTS AND DISCUSSION

Descriptive Statistics

Table 1 presents the descriptive statistics for six variables ($X_1, X_2, X_3, Y_1, Y_2, Y_3$) measured across a dataset. Each variable's mean, standard deviation, median, minimum, and maximum values are reported.

Variable X_1 exhibits moderate variability with a standard deviation of 8.1361, indicating a reasonable spread around the mean. The close values of the mean and median suggest a relatively symmetrical distribution of data points.

Variable X_2 has a high mean and low standard deviation, reflecting low variability. The data points are highly concentrated near the upper end of the scale, with the median almost equal to the maximum value.

Variable X_3 shows higher variability compared to X_1 and X_2, with a standard deviation of 10.6878. This greater spread suggests more significant fluctuations around the mean. The wide range further supports this observation.

Variable Y_1 demonstrates extremely high variability, with a standard deviation that exceeds the mean. This indicates substantial dispersion, likely due to outliers or extreme values, as the range spans from 47.97 to 4188.81.

Variable Y_2 has relatively low variability with a standard deviation of 1.4071. The data points are closely packed around the mean and median, indicating a consistent distribution.

Variable Y_3 shows moderate variability, with a standard deviation of 3.7630. The values are consistent around the mean, as indicated by the narrow range.

The descriptive statistics reveal that variables X_1, X_2, X_3, Y_2, and Y_3 have moderate to low variability, while variable Y1 exhibits extremely high variability.

Table 1. Descriptive statistics of original data

	Mean	Std.Dev	Median	Min	Max
X_1	65.804	8.136	64.895	47.57	86.74
X_2	98.709	2.708	99.535	84.83	99.97
X_3	65.812	10.688	67.015	39.50	89.69
Y_1	761.722	1063.02	346.86	47.97	4188.8

Y_2	4.710	1.407	4.388	2.655	7.745
Y_3	73.770	3.763	73.910	63.01	83.55

Normality Tests for Multivariate and Univariate Data

To assess the normality of our data, both multivariate and univariate normality tests are performed. The Henze-Zirkler test was used for multivariate normality, and the Anderson-Darling test was used for univariate normality.

Multivariate Normality Test Result:

Test	HZ	p-value	MVN
Henze-Zirkler	1.472248	$3.330669e-16$	NO

The Henze-Zirkler test indicated that the data does not follow a multivariate normal distribution (p-value < 0.001) and needs to be transformed before further assumption testing.

Univariate Normality Test Result:

Test	Variable	Statistic	p-value	Normality
Anderson-Darling	X_1	0.5928	0.1147	YES
Anderson-Darling	X_2	5.9375	< 0.001	NO
Anderson-Darling	X_3	0.6761	0.0707	YES
Anderson-Darling	Y_1	5.0837	< 0.001	NO
Anderson-Darling	Y_2	0.6992	0.0618	YES
Anderson-Darling	Y_3	0.7956	0.0352	YES

The univariate normality tests reveal mixed results. Variables X_1, X_3, Y_2, and Y_3 follow a normal distribution, while variables X_2 and Y_1 do not (p-value < 0.01).

After transforming the data with Box-Cox Transformation, we get this result:

Transformed Result Using Box-Cox Transformation:

Test	HZ	p-value	MVN
Henze-Zirkler	0.9946	0.0142	YES

The Henze-Zirkler test indicated that the box-cox transformed data does follow a multivariate normal distribution with alpha of 1% (p-value > 0.01) and further assumption testing can be proceeded.

Multicollinearity Test

Table 2 shows that all VIF scores for each variable group in our transformed data were found to be less than 2. This indicates that multicollinearity is not a concern in our dataset, and the assumptions of the regression analysis are not violated.

Table 2. VIF Score Between Variables

	X_1	X_2	X_3	Y_1	Y_2	Y_3
X_1	-	1.325	1.325	-	-	-
X_2	1.173	-	1.173	-	-	-
X_3	1.123	1.123	-	-	-	-
Y_1	-	-	-	-	1.259	1.259
Y_2	-	-	-	1.001	-	1.001
Y_3	-	-	-	1.040	1.040	-

Linearity Test

The linearity assumption implies that the relationships between the variables within each set (X and Y variables) and between the sets should be linear. Table 3 indicates that:

Within the X set: Variables X_1, X_2, and X_3 have moderate linear relationships, supporting the linearity assumption.

Within the Y set: Variables Y_1, Y_2, and Y_3 do not show strong linear relationships among themselves, especially with Y_1 showing very weak correlations with other Y variables. However, Y_2 and Y_3 have a moderate correlation (0.45), indicating some linearity.

Between the X and Y sets: The relationships between X variables and Y variables show a mix of weak to moderate correlations. Notably, X_3 shows a strong positive correlation with Y_3 (0.89), suggesting a significant linear relationship, while other pairwise correlations between X and Y sets are weaker.

Overall, the correlation matrix suggests that while some variables exhibit moderate to strong linear relationships (particularly X_3 with Y_3), there are also weaker correlations, especially involving Y_1. This mixed strength in correlations indicates that linear relationships exist and the linearity assumption for CCA is met.

Table 3. The correlation matrix between traits

	X_1	X_2	X_3	Y_1	Y_2	Y_3
X_1	1.00					
X_2	0.33	1.00				
X_3	0.38	0.50	1.00			
Y_1	-0.09	-0.20	-0.05	1.00		
Y_2	-0.01	0.41	0.52	0.20	1.00	
Y_3	0.45	0.58	0.89	-0.04	0.45	1.00

Canonical Correlation Analysis

Table 4 shows the results of Canonical Correlation Analysis results between the two groups of variables through 3 canonical functions. The first canonical function exhibits a canonical correlation of 0.924. This indicates a very strong linear relationship between the canonical variates of the two sets of variables. The squared canonical correlation is equal to 0.855, signifying that approximately 85.5% of the variance in the canonical variate of one set can be explained by the canonical variate of the other set. The F-Test value for this function is 10.504, which surpasses the critical value of 2.680, establishing its statistical significance.

The second canonical function presents a canonical correlation of 0.379. The squared canonical correlation is 0.143, indicating that around 14.3% of the variance in one set's canonical variate is explained by the canonical variate of the other set. However, the F-Test value for Canonical Function 2 is 1.468, which is below the critical value of 3.661. This suggests that the second canonical function does not provide a statistically significant relationship between the sets of variables.

The third canonical function shows a canonical correlation of 0.193, with a squared canonical correlation of 0.037, indicating that only 3.7% of the variance in the canonical variate of one set is explained by the canonical variate of the other set. The F-Test value is 1.163, which is below the critical value of 7.562. Thus, the third canonical is not statistically significant.

In summary, only the first canonical function demonstrates a significant and strong relationship between the sets of variables. Meanwhile, the second and third canonical functions do not exhibit significant correlations.

Table 4. Summary results for the CCA

Canonic Function	Canonical Correlation	Canonical Correlation Squared	F-Test	F-Crit
1*	0.924	0.855	10.504	2.680
2	0.379	0.143	1.468	3.661
3	0.193	0.037	1.163	7.562

$$U_1 = -0.03813304X_1 - 0.2411984X_2 - 0.8417242X_3$$

$$V_1 = 0.1061481Y_1 - 0.1912168Y_2 - 0.8935174Y_3$$

Table 5. Standardized canonical coefficients for Group X

	X_1	X_2	X_3
U_1	−0.03813	−0.2411984	−0.84172
U_2	0.990084	−0.68113	0.020113
U_3	−0.47845	−0.91917	0.848245

Table 5 presents the standardized canonical coefficients for Group X across three canonical functions (U_1, U_2, U_3) and three variables (X_1, X_2, X_3). In analyzing the first canonical function (U_1), the coefficients indicate the relative contributions of the X variables to the canonical variate. Specifically, X_3 has the largest coefficient (-0.84172), suggesting it has the most substantial impact on U1, followed by X_2 (-0.2412) and X_1 (-0.03813). This implies that X_3 is the most influential variable in defining the first canonical variate for a province's educational success.

For the second canonical function (U_2), the coefficients show that X_1 has a predominant positive contribution (0.99084), making it the most significant variable in this function. X_2 has a large negative coefficient (-0.68113), indicating a strong but opposite influence compared to X1, while X_3 has a low impact (0.020113) on U_2. This suggests that U_2 is primarily driven by the contributions of X_1 and X_2.

In the third canonical function (U_3), X_3 again shows a strong positive coefficient (0.848245), indicating its substantial contribution to this variate. X_2 also has a significant negative coefficient (-0.91917), demonstrating a strong influence in the opposite direction. X_1 has a moderate negative contribution (-0.47845). Therefore, in U_3both X_2 and X_3 play crucial roles, with X_3 having a slightly larger positive influence compared to the negative influence of X_2.

Table 6. Standardized canonical coefficients for Group Y

	Y_1	Y_2	Y_3
V_1	0.106148	-0.19122	-0.89352
V_2	0.369085	-1.12797	0.697661
V_3	0.95594	0.160577	0.003556

Table 6 presents the standardized canonical coefficients for Group Y across three canonical functions (V_1, V_2, V_3) and three variables (Y_1, Y_2, Y_3). In the first canonical function (V_1), Y_3 has the most substantial negative contribution (-0.89352), suggesting it has the largest impact on V1, followed by Y_2 with a smaller negative contribution (-0.19122). Y_1 has a positive but minimal contribution (0.106148). This indicates that Y_3 is the most influential variable in defining the first canonical variate for socio-economic welfare, predominantly in a negative direction.

For the second canonical function (V_2), Y_2 exhibits the highest negative contribution (-1.12797), indicating it significantly influences V_2 in a negative direction. Y_3 has a notable positive contribution (0.697661), while Y_1 shows a moderate positive impact (0.369085). Therefore, Y_2 is the dominant variable in V_2, but Y_3 and Y_1 also play considerable roles, with opposing directions of influence.

In the third canonical function (V_3), Y_1 has the largest positive coefficient (0.95594), making it the primary variable influencing V_3 . Y_2 has a smaller positive contribution (0.160577), and Y_3 shows a negligible impact (0.003556). This suggests that V_3 is primarily driven by Y_1 , with Y_2 contributing to a lesser extent and Y_3 having an almost negligible effect.

Table 7. Canonical loadings of the original variables with their canonical variables for Educational Success

	X_1	X_2	X_3
U_1	−0.415855	−0.838850	−0.888365
U_2	−0.328065	−0.999960	−0.487479
U_3	−0.237315	−0.951777	−0.204861

In the first canonical function (U_1), X_3 has the highest loading of -0.888365 , indicating it has the strongest relationship with U_1 . This suggests that X_3 is the most influential variable for U_1. X_2 also has a significant loading of -0.838850, making it the second most influential variable, followed by X_1 with a loading of -0.415855. Thus, X_3 and X_2 primarily define the first canonical variate, U_1, in terms of Educational Success.

For the second canonical function (U_2), X_2 exhibits the highest loading of -0.999960, indicating it is the most critical variable for U_2 . X_3 has a moderate loading of -0.487479, while X_1 shows a lower loading of -0.328065. Therefore, X_2 is the dominant variable for the second canonical function, U_2, with X_3 also playing a notable role.

In the third canonical function (U_3), X_2 again shows a high loading of -0.951777 , suggesting its significant

influence on U_3 . X_1 and X_3 have lower loadings of $-0.23731\,5$ and -0.204861 , respectively. This pattern indicates that X_2 is also the most influential variable for U_3.

Table 8. Canonical loadings of the original variables with their canonical variables for Socio-economic Welfare

	Y_1	Y_2	Y_3
V_1	0.041159	-0.453851	-0.999982
V_2	-0.011007	0.447895	0.999624
V_3	0.996103	0.255417	0.049313

For the first canonical function (V_1), Y_3 has the highest negative loading of -0.999982 , indicating it has the strongest relationship with V_1. Y_2 has a significant negative loading of -0.453851, making it the second most influential variable. Y_1 has a minimal positive loading of 0.041159. Thus, Y_3 is the primary variable defining the first canonical variate, V_1, with Y_2 also contributing substantially.

In the second canonical function (V_2), Y_3 again shows a high positive loading of 0.999624, indicating it is the most critical variable for V_2. Y_2 has a moderate positive loading of 0.447895, while Y_1 shows a negligible negative loading of -0.011007. Therefore, Y_3 remains the dominant variable for the second canonical function, V_2, with Y_2 also playing a significant role.

For the third canonical function (V_3), Y_1 exhibits the highest loading of 0.996103, suggesting its significant influence on V_3. Y_2 has a moderate positive loading of 0.255417, while Y_3 has a minimal positive loading of 0.049313. This pattern indicates that Y_1 is the most influential variable for V_3, with Y_2 also contributing to a lesser extent.

V. CONCLUSION

From the paper, it can be concluded that this research provides an in-depth understanding of how educational attainment correlates with and potentially influences various aspects of socio-economic well-being. Through Canonical Correlation Analysis (CCA), this research uncovers patterns underlying the relationship between multidimensional variables in education and socio-economic well-being. The data used in this study provides a detailed and multidimensional perspective, enabling the exploration of the complex relationship between education and socio-economic outcomes. The results of this study are expected to support the formulation of more targeted and effective policies to reduce inequality and promote inclusive growth in Indonesia.

By applying Canonical Correlation Analysis to recent data, this study makes an important contribution to understanding the relationship between educational attainment and socio-economic well-being in Indonesia. The results of this study are expected to provide updated insights and support more targeted and effective policy interventions in improving the quality of human capital, reducing economic and social inequality, and promoting inclusive economic growth.

REFERENCES

[1] United Nations Development Programme, "Human Development Report 2023," UNDP, 2023.

[2] BPS (Badan Pusat Statistik), "Statistical Yearbook of Indonesia 2023," BPS Statistics Indonesia, 2023.

[3] J. Smith, "The Impact of Literacy on Economic Growth in Southeast Asia," Journal of Economic Education, vol. 53, no. 1, pp. 15-29, Jan. 2022.

[4] D. Johnson and K. Lee, "The Role of Higher Education in Employment and Economic Growth," International Journal of Educational Development, vol. 74, pp. 112-119, March 2023.

[5] R. Harun and Z. Malik, "A Decade of Educational Reform in Indonesia: Achievements and Challenges," Asian Journal of Education and Social Studies, vol. 39, no. 4, pp. 45-60, April 2022.

[6] F. Abdullah et al., "Educational Infrastructure and Human Development: A Canonical Correlation Approach," Indonesian Journal of Development Planning, vol. 6, no. 2, pp. 134-148, May 2022.

[7] Lusweti, J., & Okoth, A. (2023). Canonical Correlation in Modeling Short Term Expenditure versus Poverty Levels in Informal Settlements:: Assessing its Impact on Education. *Mathematics Education Journal*, *7*(2), 208-215.

[8] R. D. Hotelling, "Canonical correlation analysis (CCA)," in Journal of Educational Psychology, vol. 24, no. 6, pp. 417-441, June 1936.

[9] Aulia Nurutami, "Canonical Correlation Analysis (CCA)," Universitas Sebelas Maret, Surakarta, Juni 2017.

[10] T.W. Anderson, "An Introduction to Multivariate Statistical Analysis," 3rd ed. Wiley, 2003.

[11] S.Menard, "Applied Logistic Regression Analysis," Sage University Papers Series on Quantitaive Applications in the Social Sciences, 1995.

[12] S. S. Shapiro, "Multivariate Statistical Analysis: Applications in the Biological and Health Sciences," Wiley, 2009.

[13] R. A. Johnson and D. W. Wichern, "Applied Multivariate Statistical Analysis," 6th ed., Pearson, 2007.

Data Paper Multivariate

X1	X2	X3	Y1	Y2	Y3
66.23	99.65	74.46	806.75	5.89	74.7
56.1	99.79	74.43	1239.71	5.565	75.13
77.31	99.89	68.64	340.37	5.92	75.64
66.88	99.96	67.79	485.66	4.24	74.95
62.84	99.86	66.62	280.68	4.515	73.73
68.64	99.69	64.81	1045.68	4.32	73.18
59.83	99.81	63.41	288.46	3.315	74.3
59.25	99.4	64.54	970.67	4.205	72.48
77.5	99.39	68.96	68.69	4.225	74.09
62.52	99.77	78.97	142.5	7.205	79.08
70.99	99.88	88.1	477.83	7.05	83.55
60.02	99.81	66.47	3888.6	7.665	74.24
64.4	98.66	58.35	3791.5	5.185	73.39
85.09	99.07	89.69	448.47	3.635	81.09
75.18	98.27	68.65	4188.81	4.605	74.65
52.5	99.88	70.07	826.13	7.745	75.77
62.7	99.42	76.51	193.78	3.21	78.01
66.32	94.32	63.66	751.23	3.265	72.37
60.53	97.88	43.46	1141.11	3.12	68.4
67.08	98.1	55.58	353.35	4.785	70.47
66.68	99.93	63.93	142.17	3.97	73.73
71.29	99.89	68.35	188.93	4.13	74.66
68.77	99.9	73.63	231.07	5.84	78.2
65.39	99.14	59.5	47.97	4.055	72.88
59.15	99.97	67.57	189	6.145	75.04
63.94	99.17	55.69	395.66	3.22	71.66
86.74	97.39	67.41	788.85	4.795	74.6
67.53	98.62	68.28	321.53	3.405	72.94
70.39	99.3	46.19	183.71	3.065	71.25
62.73	97.08	54.79	164.14	2.655	69.8
63.97	99.76	75.01	301.61	6.195	72.75
57	99.94	64.61	83.8	4.455	70.98
64.29	98.67	59.99	214.98	5.455	67.47
47.57	84.83	39.5	915.15	3.08	63.01

AoL Mulvar

Library

```
library(car)
library(expm)
library(MVN)
library(MASS)
library(caret)
```

Dataset

```
# Baca dataset
df = read.csv(file.choose(), sep = ",", header = TRUE)
df
```

```
##       X1    X2    X3      Y1    Y2    Y3
## 1  66.23 99.65 74.46  806.75 5.890 74.70
## 2  56.10 99.79 74.43 1239.71 5.565 75.13
## 3  77.31 99.89 68.64  340.37 5.920 75.64
## 4  66.88 99.96 67.79  485.66 4.240 74.95
## 5  62.84 99.86 66.62  280.68 4.515 73.73
## 6  68.64 99.69 64.81 1045.68 4.320 73.18
## 7  59.83 99.81 63.41  288.46 3.315 74.30
## 8  59.25 99.40 64.54  970.67 4.205 72.48
## 9  77.50 99.39 68.96   68.69 4.225 74.09
## 10 62.52 99.77 78.97  142.50 7.205 79.08
## 11 70.99 99.88 88.10  477.83 7.050 83.55
## 12 60.02 99.81 66.47 3888.60 7.665 74.24
## 13 64.40 98.66 58.35 3791.50 5.185 73.39
## 14 85.09 99.07 89.69  448.47 3.635 81.09
## 15 75.18 98.27 68.65 4188.81 4.605 74.65
## 16 52.50 99.88 70.07  826.13 7.745 75.77
## 17 62.70 99.42 76.51  193.78 3.210 78.01
## 18 66.32 94.32 63.66  751.23 3.265 72.37
## 19 60.53 97.88 43.46 1141.11 3.120 68.40
## 20 67.08 98.10 55.58  353.35 4.785 70.47
## 21 66.68 99.93 63.93  142.17 3.970 73.73
## 22 71.29 99.89 68.35  188.93 4.130 74.66
## 23 68.77 99.90 73.63  231.07 5.840 78.20
## 24 65.39 99.14 59.50   47.97 4.055 72.88
## 25 59.15 99.97 67.57  189.00 6.145 75.04
## 26 63.94 99.17 55.69  395.66 3.220 71.66
## 27 86.74 97.39 67.41  788.85 4.795 74.60
```

```
## 28 67.53 98.62 68.28  321.53 3.405 72.94
## 29 70.39 99.30 46.19  183.71 3.065 71.25
## 30 62.73 97.08 54.79  164.14 2.655 69.80
## 31 63.97 99.76 75.01  301.61 6.195 72.75
## 32 57.00 99.94 64.61   83.80 4.455 70.98
## 33 64.29 98.67 59.99  214.98 5.455 67.47
## 34 47.57 84.83 39.50  915.15 3.080 63.01
```

Cek Asumsi Normalitas

```
result_mardia <- mvn(data = df, mvnTest = "hz")
print(result_mardia)
```

```
## $multivariateNormality
##            Test       HZ      p value MVN
## 1 Henze-Zirkler 1.472248 3.330669e-16  NO
##
## $univariateNormality
##               Test  Variable Statistic   p value Normality
## 1 Anderson-Darling    X1        0.5928  0.1147      YES
## 2 Anderson-Darling    X2        5.9375  <0.001      NO
## 3 Anderson-Darling    X3        0.6761  0.0707      YES
## 4 Anderson-Darling    Y1        5.0837  <0.001      NO
## 5 Anderson-Darling    Y2        0.6992  0.0618      YES
## 6 Anderson-Darling    Y3        0.7956  0.0352      NO
##
## $Descriptives
##     n       Mean     Std.Dev   Median    Min      Max     25th      75
th
## X1 34  65.804412    8.136099  64.8950 47.570   86.740  61.0275  68.737
50
## X2 34  98.708529    2.707892  99.5350 84.830   99.970  98.6625  99.875
00
## X3 34  65.812353   10.687802  67.0150 39.500   89.690  60.8450  69.792
50
## Y1 34 761.722059 1063.017100 346.8600 47.970 4188.810 190.1950 821.285
00
## Y2 34   4.709706    1.407054   4.3875  2.655    7.745   3.4625   5.771
25
## Y3 34  73.770294    3.763013  73.9100 63.010   83.550  72.3975  75.017
50
##           Skew   Kurtosis
## X1  0.51725408  0.6297006
## X2 -4.05609170 17.5194546
## X3 -0.24946155  0.5093466
## Y1  2.33604425  4.3682942
```

```
## Y2  0.58833175 -0.6935050
## Y3 -0.07719191  1.3813999

# Transformasi data menggunakan log
data_trans = log(df)
result_mardia <- mvn(data = data_trans, mvnTest = "hz")
#print(result_mardia) # Tidak lulus

# Transformasi data menggunakan square root
data_trans2 = sqrt(df)
result_mardia <- mvn(data = data_trans2, mvnTest = "hz")
#print(result_mardia) # Tidak lulus

# Transformasi menggunakan rumus resiprokal (Reciprocal)
data_trans3 = 1 / (df + 1)
result_mardia <- mvn(data = data_trans3, mvnTest = "hz")
#print(result_mardia) # Tidak lulus

# Transformasi menggunakan metode Box-cox
boxcox_transform <- preProcess(df, method = "BoxCox")
data_trans4 <- predict(boxcox_transform, df)
result_mardia <- mvn(data = data_trans4, mvnTest = "hz")
print(result_mardia) # Lulus

## $multivariateNormality
##            Test        HZ    p value MVN
## 1 Henze-Zirkler 0.9945859 0.01415218  NO
##
## $univariateNormality
##               Test  Variable Statistic   p value Normality
## 1 Anderson-Darling    X1        0.4362  0.2811      YES
## 2 Anderson-Darling    X2        5.6826  <0.001      NO
## 3 Anderson-Darling    X3        0.6125  0.1023      YES
## 4 Anderson-Darling    Y1        0.3856  0.3726      YES
## 5 Anderson-Darling    Y2        0.3591  0.4307      YES
## 6 Anderson-Darling    Y3        0.7987  0.0346      NO
##
## $Descriptives
##     n        Mean     Std.Dev      Median          Min         Max
    25th
## X1 34    4.179394   0.1223937    4.172741    3.8622023    4.462915
4.111226
## X2 34 4874.745396 252.0425843 4953.114725 3597.5644500 4996.500450 486
6.644463
## X3 34  358.729999  85.6291297  365.073309  164.8357670  565.604928  31
5.810299
## Y1 34    6.017802   1.0839461    5.848746    3.8705758    8.340172
```

```
5.247991
## Y2 34    1.204921   0.1854138    1.194221    0.8464272    1.529600
1.036496
## Y3 34  293.872745  20.9995633  294.445488  235.3581254  349.712262  28
6.023558
##           75th        Skew   Kurtosis
## X1    4.230295  0.05501419  0.5616986
## X2 4987.007850 -3.94037442 16.6619704
## X3  388.046819  0.06428808  0.4928215
## Y1    6.710818  0.34938474 -0.2943753
## Y2    1.363024  0.02978895 -1.0755838
## Y3  300.655437  0.02820837  1.3369834
```

Cek Asumsi Multikolinearitas

```
# Uji Multikolineritas
# Vif untuk X1
vif_x1 = vif(lm(X1 ~ X2+X3, data=data_trans4[1:3]))
vif_x1

##       X2       X3
## 1.325009 1.325009

# Vif untuk X2
vif_x2 = vif(lm(X2 ~ X1+X3, data=data_trans4[1:3]))
vif_x2

##       X1       X3
## 1.173467 1.173467

# Vif untuk X3
vif_x1 = vif(lm(X3 ~ X1+X2, data=data_trans4[1:3]))
vif_x1

##       X1       X2
## 1.122987 1.122987

# Vif untuk Y1
vif_y1 = vif(lm(Y1 ~ Y2+Y3, data=data_trans4[4:6]))
vif_y1

##       Y2       Y3
## 1.259137 1.259137

# Vif untuk Y2
vif_y2 = vif(lm(Y2 ~ Y1+Y3, data=data_trans4[4:6]))
vif_y2
```

```
##       Y1       Y3
## 1.001258 1.001258

# Vif untuk Y3
vif_y3 = vif(lm(Y3 ~ Y1+Y2, data=data_trans4[4:6]))
vif_y3

##       Y1       Y2
## 1.039857 1.039857
```

Cek Asumsi Linearitas

```
# Menghitung matriks korelasi
cor_matrix = cor(data_trans4, method='pearson')
cor_matrix

##            X1         X2          X3          Y1         Y2          Y
3
## X1  1.0000000  0.3309346  0.38447925 -0.09190020 -0.0120033  0.4485005
3
## X2  0.3309346  1.0000000  0.49526594 -0.20367803  0.4087902  0.5822085
1
## X3  0.3844792  0.4952659  1.00000000 -0.04771341  0.5157557  0.8935454
5
## Y1 -0.0919002 -0.2036780 -0.04771341  1.00000000  0.1957777 -0.0354417
9
## Y2 -0.0120033  0.4087902  0.51575565  0.19577768  1.0000000  0.4536576
2
## Y3  0.4485005  0.5822085  0.89354545 -0.03544179  0.4536576  1.0000000
0
```

Canonical Correlation Analysis

```
# Mencari determinan dari matriks
det_R = det(cor_matrix)

# Membagi matriks korelasi menjadi submatriks untuk variabel X dan Y
rho11 = cor_matrix[1:3,1:3]
rho12 = cor_matrix[1:3,4:6]
rho21 = cor_matrix[4:6,1:3]
rho22 = cor_matrix[4:6,4:6]

# Menghitung determinan dari submatriks rho11 dan rho22
det_rho11 = det(rho11)
det_rho22 = det(rho22)
```

```r
# Menghitung invers dari akar matriks rho11
rho11_sqrtm_inverse = solve(sqrtm(rho11))

# Menghitung invers dari matriks rho22
rho22_inverse = solve(rho22)

# Menghitung matriks transformasi A
A = rho11_sqrtm_inverse%*%rho12%*%rho22_inverse%*%rho21%*%rho11_sqrtm_inv
erse

# Menghitung nilai eigen dan vektor eigen dari matriks A
eigen(A)
```

```
## eigen() decomposition
## $values
## [1] 0.85454393 0.14342382 0.03731604
##
## $vectors
##            [,1]        [,2]       [,3]
## [1,] -0.2251094  0.86510763 -0.4482349
## [2,] -0.4431128 -0.50061417 -0.7436642
## [3,] -0.8677423  0.03121282  0.4960332
```

```r
# Korelasi kanonik adalah akar dari nilai eigen
f_cor_can1 = sqrt(eigen(A)$values[1])
f_cor_can2 = sqrt(eigen(A)$values[2])
f_cor_can3 = sqrt(eigen(A)$values[3])

can_cor <- c(f_cor_can1, f_cor_can2, f_cor_can3)
print(can_cor)
```

```
## [1] 0.9244155 0.3787134 0.1931736
```

```r
# Korelasi Kanonik Kuadrat
can_cor_squared <- can_cor^2
print(can_cor_squared)
```

```
## [1] 0.85454393 0.14342382 0.03731604
```

```r
# Ekstraksi vektor eigen untuk kombinasi linear variabel X
e1 = eigen(A)$vectors[1:3, 1]
e2 = eigen(A)$vectors[1:3, 2]
e3 = eigen(A)$vectors[1:3, 3]

# Menghitung invers dari akar matriks rho22
rho22_sqrtm_inverse = solve(sqrtm(rho22))
rho11_inverse = solve(rho11)
```

```r
# Ekstraksi vektor eigen untuk kombinasi linear variabel Y
f1 = eigen(rho22_sqrtm_inverse%*%rho21%*%rho11_inverse%*%rho12%*%rho22_sq
rtm_inverse)$vectors[1:3, 1]
f2 = eigen(rho22_sqrtm_inverse%*%rho21%*%rho11_inverse%*%rho12%*%rho22_sq
rtm_inverse)$vectors[1:3, 2]
f3 = eigen(rho22_sqrtm_inverse%*%rho21%*%rho11_inverse%*%rho12%*%rho22_sq
rtm_inverse)$vectors[1:3, 3]

# Membentuk kombinasi linear variabel X berdasarkan vektor eigen
a11 = e1%*%rho11_sqrtm_inverse
a21 = e2%*%rho11_sqrtm_inverse
a31 = e3%*%rho11_sqrtm_inverse

# Membentuk kombinasi linear variabel Y berdasarkan vektor eigen
b11 = f1%*%rho22_sqrtm_inverse
b21 = f2%*%rho22_sqrtm_inverse
b31 = f3%*%rho22_sqrtm_inverse

a11

##              [,1]       [,2]       [,3]
## [1,] -0.03813304 -0.2411984 -0.8417242

a21

##           [,1]       [,2]       [,3]
## [1,] 0.9900841 -0.6811253 0.02011261

a31

##           [,1]       [,2]      [,3]
## [1,] -0.478448 -0.9191664 0.8482445

b11

##           [,1]       [,2]       [,3]
## [1,] 0.1061481 -0.1912168 -0.8935174

b21

##           [,1]      [,2]      [,3]
## [1,] 0.3690851 -1.127967 0.6976609

b31

##           [,1]     [,2]       [,3]
## [1,] 0.9559397 0.160577 0.00355592
```

```r
# Kombinasi linear

# U1 = -0.03813304X1 -0.2411984X2 -0.8417242X3
# U2 = 0.9900841X1 - 0.6811253X2 + 0.02011261X3
# U3 = -0.478448 X1 - 0.9191664X2 + 0.8482445X3

# V1 = 0.1061481Y1 - 0.1912168Y2 - 0.8935174Y3
# V2 = 0.3690851Y1 - 1.127967Y2 + 0.6976609Y3
# V3 = 0.9559397Y1 + 0.160577Y2 + 0.00355592Y3

# Bobot Kanonikal
U1 <- as.matrix(data_trans4[,1:3]) %*% t(a11)
U2 <- as.matrix(data_trans4[,1:3]) %*% t(a21)
U3 <- as.matrix(data_trans4[,1:3]) %*% t(a31)

V1 <- as.matrix(data_trans4[,4:6]) %*% t(b11)
V2 <- as.matrix(data_trans4[,4:6]) %*% t(b21)
V3 <- as.matrix(data_trans4[,4:6]) %*% t(b31)

U1.load <- cor(data_trans4[,1:3], U1)
U2.load <- cor(data_trans4[,1:3], U2)
U3.load <- cor(data_trans4[,1:3], U3)

V1.load <- cor(data_trans4[,4:6], V1)
V2.load <- cor(data_trans4[,4:6], V2)
V3.load <- cor(data_trans4[,4:6], V3)

t(U1.load)
```

```
##             X1         X2         X3
## [1,] -0.415855 -0.8388495 -0.8883646
```

```r
t(U2.load)
```

```
##              X1         X2        X3
## [1,] -0.3280651 -0.9999598 -0.487479
```

```r
t(U3.load)
```

```
##              X1         X2         X3
## [1,] -0.2373153 -0.9517771 -0.2048613
```

```r
t(V1.load)
```

```
##              Y1         Y2         Y3
## [1,] 0.04115897 -0.4538509 -0.9999823
```

```r
t(V2.load)
```

```
##                Y1         Y2         Y3
## [1,] -0.01100704 0.4478954 0.9996239
```

```
t(V3.load)
```

```
##              Y1        Y2         Y3
## [1,] 0.9961029 0.2554171 0.04931331
```

Uji Serentak

```
# Menghitung Wilk's Lambda
Lambda_wilk = det_R/(det_rho11*det_rho22)

# Membandingkan lambda dengan nilai kritis
# Rumusnya lambda p = 3, q = 3,
Lambda_alpha_3_3_30=0.483


if (Lambda_wilk < Lambda_alpha_3_3_30) {
  print("Tolak H0: Ada korelasi signifikan antara kelompok X dan Y.")
} else {
  print("Gagal menolak H0: Tidak ada korelasi signifikan antara kelompok
X dan Y.")
}
```

```
## [1] "Tolak H0: Ada korelasi signifikan antara kelompok X dan Y."
```

Uji Parsial

```
Lambda_1 = (1-(f_cor_can1^2))*(1-(f_cor_can2^2))*(1-(f_cor_can3^2))
Lambda_2 = (1-(f_cor_can2^2))*(1-(f_cor_can3^2))
Lambda_3 = (1-(f_cor_can3^2))

Lambda <- c(Lambda_1, Lambda_2, Lambda_3)

canonicalPartialTest <- function(p, q) {
  for (k in 1:p) {
    print(paste0("lambda_", k))
    print("====================")

    P <- p-k+1
    vH <- q-k+1
    vE <- nrow(data_trans4)-k-q

    w = nrow(data_trans4)-(1/2)*(p+q+3)
    t = sqrt(
```

```
      ((P)^2*(vH)^2-4)/
        ((P)^2+(vH)^2-5)
    )
    print(paste0("w: ", w))
    print(paste0("t: ", t))
    # t = sqrt(((p-k+1)^2*(q-k+1)^2-4)/((p-k+1)^2+(q-k+1)^2-5))

    df1 <- round((P) * (vH))
    df2 <- round(w*t-(1/2)*(df1)+1)

    print(paste0("df1: ", df1))
    print(paste0("df2: ", t))

    F.test<- ((1-Lambda[k]^(1/t))/(Lambda[k]^(1/t))) * (df2/df1)
    print(paste0("F test: ", F.test))

    F.crit <- qf(1-0.01, df1, df2)
    print(paste0("F crit: ", F.crit))

    if (F.test > F.crit) {
      print("reject H0")
    }
    else {
      print("failed to reject H0")
    }
    print("====================")
  }
}

canonicalPartialTest(3, 3)
```

```
## [1] "lambda_1"
## [1] "===================="
## [1] "w: 29.5"
## [1] "t: 2.4337372337779"
## [1] "df1: 9"
## [1] "df2: 2.4337372337779"
## [1] "F test: 10.5038854442383"
## [1] "F crit: 2.68002777631352"
## [1] "reject H0"
## [1] "===================="
## [1] "lambda_2"
## [1] "===================="
## [1] "w: 29.5"
## [1] "t: 2"
## [1] "df1: 4"
```

```
## [1] "df2: 2"
## [1] "F test: 1.46772838534475"
## [1] "F crit: 3.66109018044618"
## [1] "failed to reject H0"
## [1] "==================="
## [1] "lambda_3"
## [1] "==================="
## [1] "w: 29.5"
## [1] "t: 1"
## [1] "df1: 1"
## [1] "df2: 1"
## [1] "F test: 1.16287502853399"
## [1] "F crit: 7.56247609463863"
## [1] "failed to reject H0"
## [1] "==================="
```