DOI: https://doi.org/10.34123/jurnalasks.v17i1.803

Available online at https://jurnal.stis.ac.id/

Jurnal Aplikasi Statistika & Komputasi Statistik
Vol. 17 No. 1, June 2025
e- ISSN: 2615-1367, p-ISSN:2086-4132

Estimation of Gross Regional Domestic Product per Capita at the Sub-District Level in Bali, NTB, and NTT Provinces Using Machine Learning Approaches and Geospatial Data

I Made Satria Ambara Putra[1], Rindang Bangun Prasetyo[2*], Candra Adi Wiguna[3]

[1]BPS-Statistics West Kotawaringin Regency, West Kotawaringin, Indonesia, [2]Politeknik Statistika STIS, Jakarta, Indonesia, [3]Nanyang Technological University, Singapore
**Corresponding Author: E-mail address:* rindang@stis.ac.id

ARTICLE INFO

Article history:
Received 8 September, 2024
Revised 5 December, 2024
Accepted 4 February, 2025
Published 24 February, 2025

Keywords:
Big Data Geospatial; Gross Regional Domestic Product Per Capita; Machine Learning; Neural Network; Williamson Index

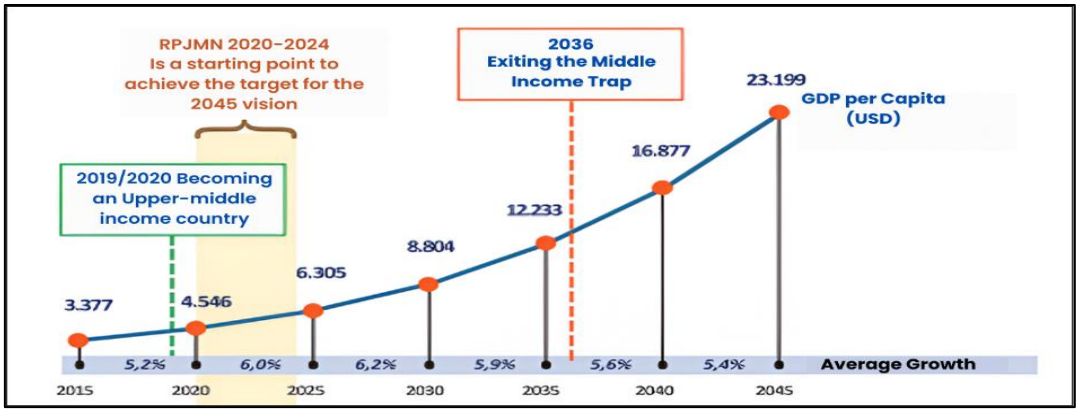
Abstract

Introduction/Main Objectives: This study aims to estimate Gross Regional Domestic Product (GRDP) per capita at the sub-district level. **Background Problems:** Currently, GRDP per capita is calculated only at the district level by BPS. **Novelty:** This study estimates GRDP per capita at the sub-district level using a model developed at the district level, applying machine learning and linear regression methods. **Research Methods:** The model was constructed using geospatial data sourced from satellite imagery, OpenStreetMap, (Village Potential Statistics) PODES, directories of large mining companies, and directories of the manufacturing industry at the district level. Linear regression and machine learning methods, including neural networks, random forest regression, and support vector regression, were used to develop the model. The research focuses on three provinces: Bali, West Nusa Tenggara (NTB), and East Nusa Tenggara (NTT). **Findings/Results:** The best-performing model was support vector regression, with MAE and MAPE evaluations of 10.33 million and 26.11%, respectively. The results indicate that sub-districts with high GRDP per capita are typically urban areas that serve as economic hubs. The Williamson Index results show that districts in the eastern region have higher inequality levels compared to those in the western region.

1. Introduction

The Indonesian government has incorporated the Sustainable Development Goals (SDGs) into its development strategy, as outlined in the National Medium-Term Development Plan (RPJMN), starting from the 2015-2019 period and continuing into the current 2020-2024 period. The RPJMN functions as a strategic framework for the government's initiatives to meet the objectives set forth in Indonesia's Vision 2045, which aims to establish a more Advanced Indonesia. Figure 1 shows the target Gross Domestic Product (GDP) per capita as part of the vision for an Advanced Indonesia. GDP is an economic indicator used to assess the performance of economic development in a country or region. It can be calculated from various metrics that evaluate a country's financial performance [1]. Indonesia's position from 2019 to 2020 has already reached upper-middle-income status, meaning it is above countries with middle-income per capita but still below those with high-income per capita. Indonesia is projected to escape the Middle-Income Trap by 2036. One of the key strategies to achieve this goal is to reduce regional disparities, ensuring that economic growth in each region aligns with national growth. To

monitor these conditions and track regional disparities, it is essential to use appropriate indicators. The Gross Regional Domestic Product (GRDP) per capita is one such indicator that can illustrate disparities between regions [2].



Source: Bappenas

Figure 1. Target of GDP per capita growth towards an advanced Indonesia

The Gross Regional Domestic Product (GRDP) is typically calculated at higher administrative levels, such as districts or provinces, and is generally not computed at the sub-district level [3]. This is primarily due to the limited scale and availability of economic data at the sub-district level, where economic activities tend to be smaller and more restricted compared to districts or provinces. The necessary economic data, such as industrial output, agricultural production, trade, and services, may either be unavailable or insufficiently detailed at the sub-district level, leading to challenges in data collection. Additionally, many economic activities in sub-districts are often integrated with surrounding areas, making it difficult to accurately isolate and assess the specific economic contributions of individual sub-districts. For these reasons, GRDP is more frequently calculated at higher levels, such as districts or provinces, where the data is more comprehensive and suitable for evaluating regional economic performance [4], [5].

GRDP at the sub-district level represents added value that can be directly observed by the community [6]. By utilizing GRDP data at the sub-district level, the success of economic development in these areas can be evaluated both directly and indirectly [6]. Regional disparities can be measured using the Williamson Index, which relies on GRDP per capita as a reference metric [7], [8]. Currently, the Williamson Index is typically calculated at the provincial or national level. However, with the availability of GRDP per capita data at the sub-district level, it would become possible to calculate the Williamson Index at the district level. Therefore, estimating GRDP per capita at the sub-district level is crucial for accurately assessing economic disparities and development.

The approach to calculating District GRDP can be done using two methods: the direct method and the indirect method. The direct method involves using primary data from each district to compute GRDP, though in practice, such data is often difficult to obtain. In contrast, the indirect method estimates gross value added by allocating or predicting it using indicators that closely correlate with the respective economic activities. This method leverages proxy indicators to approximate economic output when direct data is unavailable [3], [6]. One of the indirect methods that can be applied to produce GRDP estimates at the sub-district level is through prediction. Various estimation methods, including linear regression, can be used to estimate these values [9]. For instance, Pasaribu et al. [10] utilized a linear regression model to predict GRDP in Jakarta. Similarly, Agu et al. [11] predicted GDP using macroeconomic indicators by applying four regression techniques: Principal Component Regression (PCR), Ridge Regression (RR), Lasso Regression (LR), and Ordinary Least Squares (OLS). These methods help in estimating economic output when direct data is unavailable, providing a robust approach to forecasting regional economic performance.

The linear regression method is one of the simplest and most commonly used techniques to model the relationship between continuous variables [12]. However, its primary limitation is that it can only capture linear associations between the dependent and independent variables. To address this limitation, machine learning techniques can be utilized. Machine learning algorithms can identify and model complex, non-linear relationships, provide more accurate predictions, offer flexibility in model selection and development, address overfitting through regularization techniques, and efficiently process large datasets, making them highly suitable for more advanced statistical modeling and prediction tasks [13].

Several researchers have applied machine learning (ML) techniques to predict GRDP. Muchisha et al. [14] developed and evaluated the performance of six widely used ML algorithms, including Random Forest, LASSO, Ridge, Elastic Net, Neural Networks, and Support Vector Machines, to forecast Indonesia's quarterly GDP growth in real-time. Similarly, Richardson et al. [15] demonstrated that ML algorithms have the potential to significantly outperform simple autoregressive benchmarks and dynamic factor models when predicting New Zealand's GDP. Sa'adah and Wibowo [16] applied two deep learning techniques, LSTM and RNN, to model GDP fluctuations, even during the COVID-19 pandemic. Experimental results from Lai [17] showed that neural networks are reliable in predicting GDP and have practical applications. Meanwhile, Sukono et al. [18] predicted Gross Regional Domestic Product (GRDP) using a genetic algorithm approach based on the Cobb-Douglas model, comparing the results with those obtained from the ordinary least squares (OLS) method. The study conducted by Puttanapong et al. [13] in estimating Provincial GDP in Thailand utilized three methods: neural networks, random forests, and support vector machines. These methods were chosen because neural networks are highly effective in capturing complex and non-linear relationships within data, random forests are known for their ability to produce accurate models by leveraging multiple decision trees, thereby reducing the risk of overfitting, and SVMs perform well with high-dimensional data, making them suitable for analyzing complex datasets [13].

The primary challenge in estimating Gross Regional Domestic Product (GRDP) at the sub-district level lies in the issue of data completeness. Not all economic sectors have data available at such a granular level, necessitating the use of additional data that specifically covers the sub-district level [6]. Furthermore, the data sources traditionally used for GRDP calculations are often costly and time-intensive to collect. Consequently, there is a need for alternative methods and data sources that are both cost-effective and time-efficient while providing comprehensive coverage at the sub-district level.

Recent studies have increasingly explored the use of alternative data sources, such as geospatial data derived from satellite imagery through remote sensing and OpenStreetMap (OSM). These data sources offer the potential to supplement the limited data currently available for calculating GRDP per capita at more granular levels. Remote sensing refers to the use of sensors to detect electromagnetic radiation, which can be processed to generate interpretable images of the earth's surface, thereby yielding valuable information [19]. One significant advantage of remote sensing is its ability to provide extensive spatial coverage, even at very fine scales, coupled with the availability of large volumes of data [19]. Similarly, OpenStreetMap data can accurately represent regional characteristics down to micro levels, and its open-source nature ensures it is freely accessible [20].

Previous research by Fasial and Shakera [21] investigated the use of remote sensing techniques to predict Gross Domestic Product (GDP) through built-up index analysis across nine major cities in Canada. Similarly, Puttanapong et al. [13] demonstrated that the application of geospatial data and machine learning, specifically utilizing the Random Forest algorithm, achieved a prediction accuracy of 97.7% for Provincial Gross Domestic Product. In addition to the Random Forest method, their study also employed two other machine learning algorithms, namely Neural Networks and Support Vector Machines [13]. These findings indicate that the application of machine learning algorithms can significantly enhance the accuracy and comprehensiveness of Provincial GDP predictions, while also facilitating a more detailed integration with geospatial data. Furthermore, Putri et al. applied machine learning techniques to estimate poverty at a granular level, building a model based on sub-district data to estimate poverty at a 1.5 km grid scale [22]. Therefore, this study aims to estimate the GRDP per capita at the sub-district level in the provinces of Bali, NTB, and NTT using linear regression, neural networks, random forest regression, and support vector regression. The models are constructed at the district level to estimate the lower-level sub-districts.

2. Material and Methods

2.1. Gross Regional Domestic Product Per Capita

Gross Regional Domestic Product (GRDP) data, both at current prices and constant prices, is one of the key indicators for understanding the economic condition of a region or area within a specific period [23]. In general, GRDP represents the total value added generated by all business units in a region or can be interpreted as the total value of final goods and services produced by all economic units [23]. GRDP at current prices reflects the value added of goods and services calculated based on the prevailing prices in the given year. Meanwhile, GRDP at constant prices represents the value added of goods and services calculated using fixed prices from a specific base year as a reference [23]. Badan Pusat Statistik is an Indonesian government agency responsible for collecting GRDP data.

GRDP per capita is the result of dividing the value added generated by all economic activities by the total population [24]. The size of the population, whether large or small, will affect the GRDP per capita, while the GRDP itself is highly dependent on the potential of natural resources and the production factors available in the region [24]. GRDP per capita at current prices reflects the value of GRDP divided by the total population or per individual [24].

2.2. Geospatial Data

Geospatial data is information about the location, size, and characteristics of objects, both natural and man-made, that are located above or below the Earth's surface [25]. Geospatial information is processed geospatial data that can be utilized as a tool in formulating policies, making decisions, and carrying out activities related to terrestrial space. Geospatial data and information also become very important and beneficial in supporting the development process across various sectors of life [25]. There are five sources of geospatial data: public participatory geographic information systems (PPGIS), participatory geographic information systems (PGIS), volunteered geographic information (VGI), open data, and big data [26]. In this research, geospatial data from satellite imagery sourced from big data and geospatial data from OpenStreetMap sourced from VGI are used.

2.3. Yeo-Johnson Transformation

Yeo-Johnson is an extension of the Box-Cox transformation. Box-Cox Transformation is used to transform positive data to approximate a normal distribution. This transformation can only be applied to positive data. Yeo-Johnson Transformation is a generalization of Box-Cox that can be applied to both positive and negative data. This makes Yeo-Johnson more flexible in handling various types of data [27]. The data transformations are defined by the following equations [27].

$$\psi(x,\lambda)\begin{cases}\dfrac{(x+1)^{\lambda}-1}{\lambda} & x \geq 0;\ \lambda \neq 0\\ \log(x+1) & x \geq 0;\ \lambda = 0\\ \dfrac{(-x+1)^{2-\lambda}-1}{2} & x < 0;\ \lambda \neq 2\\ -\log(-x+1) & x < 0;\ \lambda = 2\end{cases} \quad (1)$$

Where x is the original value and λ is the transformation parameter.

2.4. Pearson Correlation

Pearson correlation is used to determine the strength of the linear relationship between two variables and to identify the direction of the relationship that occurs [28]. The interpretation of the correlation coefficients is presented in Table 1 below.

Table 1. Interpretation of correlation coefficient

Range of Values	Level of Intensity
0.80 – 1.000	Very Strong
0,60 – 0,799	Strong
0,40 – 0,699	Strong Enough
0,20 – 0,399	Weak
0,00 – 0,199	Very Weak

2.5. Mutual Information

Mutual Information has emerged in recent years as an important measure for feature selection criteria, particularly in machine learning [29]. By using Mutual Information, we can uncover unexpected relationships between the measured variables. So far, the Pearson correlation coefficient has been the most popular, but this measure is not sufficient because many phenomena are non-linear in nature [30]. A high Mutual Information value indicates that the association between two variables is stronger [31].

2.6. Variance Inflation Factor

Multicollinearity is the occurrence of a high correlation among factors in a model. Multicollinearity can produce biased results when researchers attempt to determine how effectively each factor can be used to predict or understand the response variable in a statistical model [32]. Overall, multicollinearity can result in wider confidence intervals and less reliable probability values for predictors [32]. This means that the findings from the model with multicollinearity may not be reliable. There are several techniques used to detect multicollinearity, and this study employs the Variance Inflation Factor (VIF). VIF is used to measure the extent to which the variance of the estimated regression coefficients increases when the independent variables are correlated with each other.

2.7. Linear Regression

Linear regression is the simplest and certainly the most common method for measuring the relationship between continuous variables, and this method is easiest to understand through practical examples. Regression models are often used for practical decision-making as well as for more theoretical or scientific investigations [12].

2.8. Machine Learning

Machine learning is a field that lies at the intersection of computer science, statistics, and various other disciplines, focusing on the automatic improvement over time, as well as inference and decision-making under uncertainty [33]. Machine learning requires valid data as learning material (during the training process) before being used in testing to produce optimal output [34]. The ability of networked and cellular computing systems to collect and transfer large amounts of data has grown rapidly, a phenomenon often referred to as "big data." Researchers who gather such data often turn to machine learning to seek solutions in obtaining useful insights, making predictions, and facilitating decision-making from these data sets [33].

2.9. Evaluation Metrics

The evaluation of the optimal model selection is conducted by considering the level of accuracy. The accuracy level is calculated by taking into account the residual values produced by the model that has been built. This research uses several evaluation metrics, namely Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE). The formulas for these three metrics are as follows [35].

$$RMSE = \sqrt{\frac{\sum_{t=1}^{n}(F_t - A_t)^2}{n}} \tag{1}$$

$$MAE = \frac{1}{n}\sum_{t=1}^{n}|F_t - A_t| \tag{2}$$

$$MAPE = \frac{100}{n}\sum_{t=1}^{n}\frac{|F_t - A_t|}{A_t} \tag{3}$$

In this context, F_t represents the forecasted value for time period t, A_t denotes the actual observed value for the same period, and n indicates the total number of time periods. A lower MAPE value signifies a higher accuracy of the forecasting model.

2.10. Williamson Index

The Williamson Index is a tool for measuring the development of a region by comparing it with more advanced areas. In general, the Williamson Index assesses the level of development disparity in a region [8]. The Williamson index describes the relationship between regional disparity and economic levels using economic data from both developed and developing regions [8]. The basis for calculating the Williamson Index involves GDP per capita and the population size in each region. Statistically, the formulation of the Williamson Index can be expressed with the following formula.

$$IW = \frac{\sqrt{\sum_i^n (y_i - \hat{y})^2 \frac{p_i}{p}}}{\hat{y}} \quad (4)$$

With IW being the Williamson Index, y_i represents the GDP per capita at the district/city level, $\hat{y}$ represents the GDP per capita of the province, p_i represents the population of each district/city, p is the total population of the province, and n is the number of districts in the province. This study will obtain the GDP per capita at the sub-district level. Therefore, the IW calculation can be conducted at the district level. When the Williamson Index value moves further away from 0, it indicates that income inequality between regions in that area is increasing. Conversely, if the Williamson Index value approaches 0, it shows that income inequality between regions in that area is decreasing [8].

2.11. Analysis Method

This research uses six data sources. First, the data source comes from satellite imagery. Satellite data from remote sensing used in this research includes Night Time Light (NTL) from the NOAA-VIIRS satellite, Normalized Difference Vegetation Index (NDVI), Normalized Difference Water Index (NDWI), Normalized Difference Built-Up Index (NDBI) from the Sentinel-2 satellite, Land Surface Temperature from the MODIS satellite, and Carbon Monoxide (CO), Nitrogen Dioxide (NO2), and Sulfur Dioxide (SO2) from the Sentinel-5P satellite. This data collection is conducted using a cloud-based platform, namely Google Earth Engine. Secondly, the data source used comes from OpenStreetMap (OSM). The collection of OSM data is done through Overpass Turbo. Overpass Turbo is a web-based mining tool for running queries to access OpenStreetMap data and displaying the results on an interactive map. Tourism data is obtained by entering the query or keyword "tourism," while to get road length data, the keywords "highway=primary," "highway=secondary," and "highway=tertiary" are used.

Thirdly, the data sources used come from the village potential data of 2021 collected by Badan Pusat Statistik (BPS). The researcher proposed data collection through the Statistical Service Information System (Silastik). The data collected includes the number of households using electricity, the number of educational facilities, the number of health facilities, the number of micro and small industries, the number of financial services, the number of villages with kiosks selling agricultural production tools, the number of food and beverage accommodation providers, the number of places of worship, the number of villages with internet access for online gaming cafes and other facilities, the number of village-owned enterprises, and the number of villages with waste banks.

Fourth, the data sources used are from the directory of large mining companies in 2021 and the directory of the manufacturing industry in 2021. Fifth, the data sources used are from the official websites of the BPS of each province to obtain GRDP per capita data at the district administrative level for the year 2021. GRDP per capita is the response variable in this study. The sixth is the population data taken from regional publications in numbers, this data is used for the calculation of the Williamson Index. All data collected is based on the year 2021.

In this study, the model is built at the district level. The resulting model is used to estimate GRDP per capita at the sub-district level. The model development uses data from 9 provinces, namely the Special Region of Yogyakarta, DKI Jakarta, Banten, East Java, Central Java, West Java, West Nusa Tenggara, East Nusa Tenggara, and Bali. However, the predictions were only made in three provinces: Bali, NTB, and NTT. The tools used in this research are Google Earth Engine, Overpass Turbo, QGIS, and Google Colab for data collection, data preprocessing, model development, and mapping. In addition, Google Drive and spreadsheets are used for data storage. The general steps are outlined in the following flowchart.

Data collection and preprocessing is the first stage of this research. Through preprocessing, the data is cleaned, processed, and transformed to be ready for further analysis. Preprocessing steps such as data cleaning help ensure that the data used for training and testing models or algorithms is of high quality and relevant. The collected data consists of data at the district and sub-district administrative levels. The district administrative data is used for model development purposes, while the sub-district administrative data is used for estimating per capita GRDP at the sub-district level. The district administrative data includes 160 districts and cities from 9 provinces, namely Bali Province, West Nusa Tenggara, East Nusa Tenggara, Special Region of Yogyakarta, Jakarta Special Capital Region, Banten, East Java, Central Java, and West Java. Meanwhile, the sub-district administrative data consists of 389 sub-districts from 3 provinces, namely Bali Province, West Nusa Tenggara, and East Nusa Tenggara.

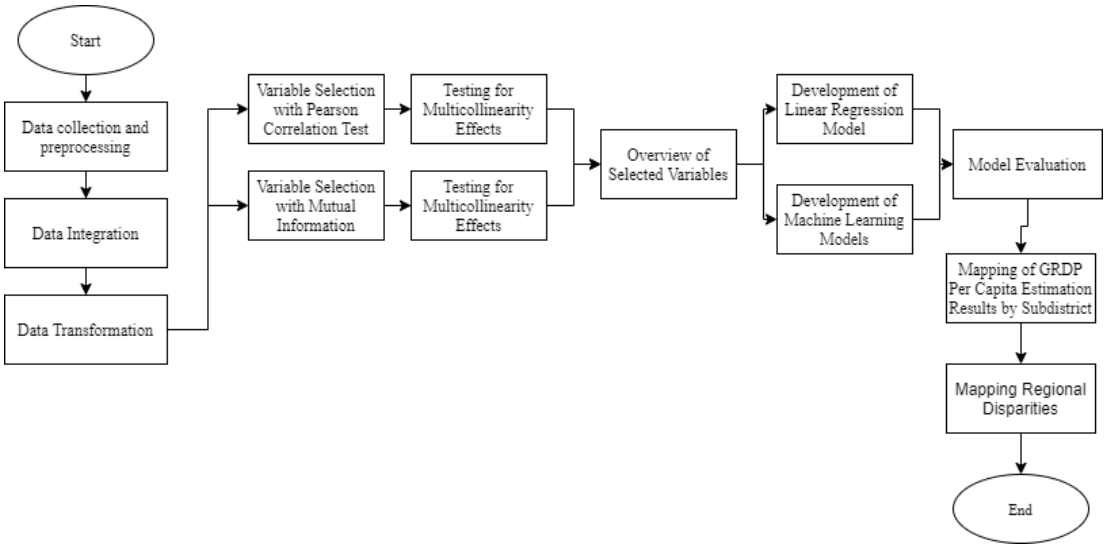


Figure 2. Flowchart of research stages

The expected outcome of the data integration steps is a dataset that encompasses all the necessary data, such as satellite imagery, OpenStreetMap, village potential data, mining, and manufacturing industries. After collecting data and preprocessing it, data from various sources was obtained at the administrative levels of the sub-district and district. At this stage, the data is combined based on its administrative level. Data at the sub-district administrative level is needed to estimate the per capita GRDP at the sub-district level, while data at the district administrative level is necessary for model development. The next step is data transformation. The purpose of data transformation is to improve data quality or to prepare the data for analysis or model development. The transformation used is the Yeo-Johnson transformation. In the Python programming language, the "sklearn.preprocessing" library provides the "PowerTransformer" function for calculating the Yeo-Johnson transformation.

Variable selection means choosing among many variables which ones will be included in the model, that is, selecting the appropriate variables from a complete list by eliminating those that are irrelevant or redundant. Practicality is one of the reasons why variables must be chosen. According to the principle of parsimony, a simple model with fewer variables is preferred over a complex model with many variables. In this study, variable selection for the development of the linear regression model was conducted using Pearson correlation analysis, and variables were chosen based on moderate to strong relationships. After that, the selected variables underwent a multicollinearity test. Variable selection for the development of the machine learning model used mutual information analysis, and 10 variables with the highest values were chosen. Following that, the selected variables underwent a multicollinearity test. Variables that indicate the presence of multicollinearity will be eliminated in the model development. The purpose of this stage is to improve the model's performance.

After the district data set is completed, variable selection is carried out, and the model is built at the district level. The model for estimating GDP per capita is built by applying linear regression and machine learning methods. The machine learning algorithms used are neural network (NN), random forest regression (RFR), and support vector regression (SVR). The data set is divided into 80% training data (128 data points) and 20% testing data (32 data points). The model is built using 80% of the data set. In this research, the GridSearchCV approach is applied to build and select the best model. The evaluation uses the K-Fold Cross Validation method, which is the default evaluation method when using GridSearchCV. Next, evaluate the model using RMSE, MAE, and MAPE values to obtain the best model for estimating per capita GDP at the sub-district level. The model is selected based on the smallest RMSE, MAE, and MAPE values.

3. Result and Discussion

3.1. Candidate Variable Data Exploration

It can be identified in Figure 3 that the relationship between GRDP per capita and the predictor variables has a non-linear pattern, while it can also be identified that the patterns and relationships among other variables also exhibit a non-linear pattern. The variables used are still in different units. Therefore, a transformation is needed to improve data quality. The transformation used in this research

is the Yeo-Johnson transformation because it can handle both positive and negative values, as well as manage variables that have different units, such as the variables used in this study. This transformation can produce the best evaluation compared to other transformations. After that, a Pearson correlation test and multicollinearity test were conducted for the development of the linear regression model, while variable selection was done using mutual information and a multicollinearity test for the development of the machine learning model. This is done so that the model can provide accurate and precise predictions.

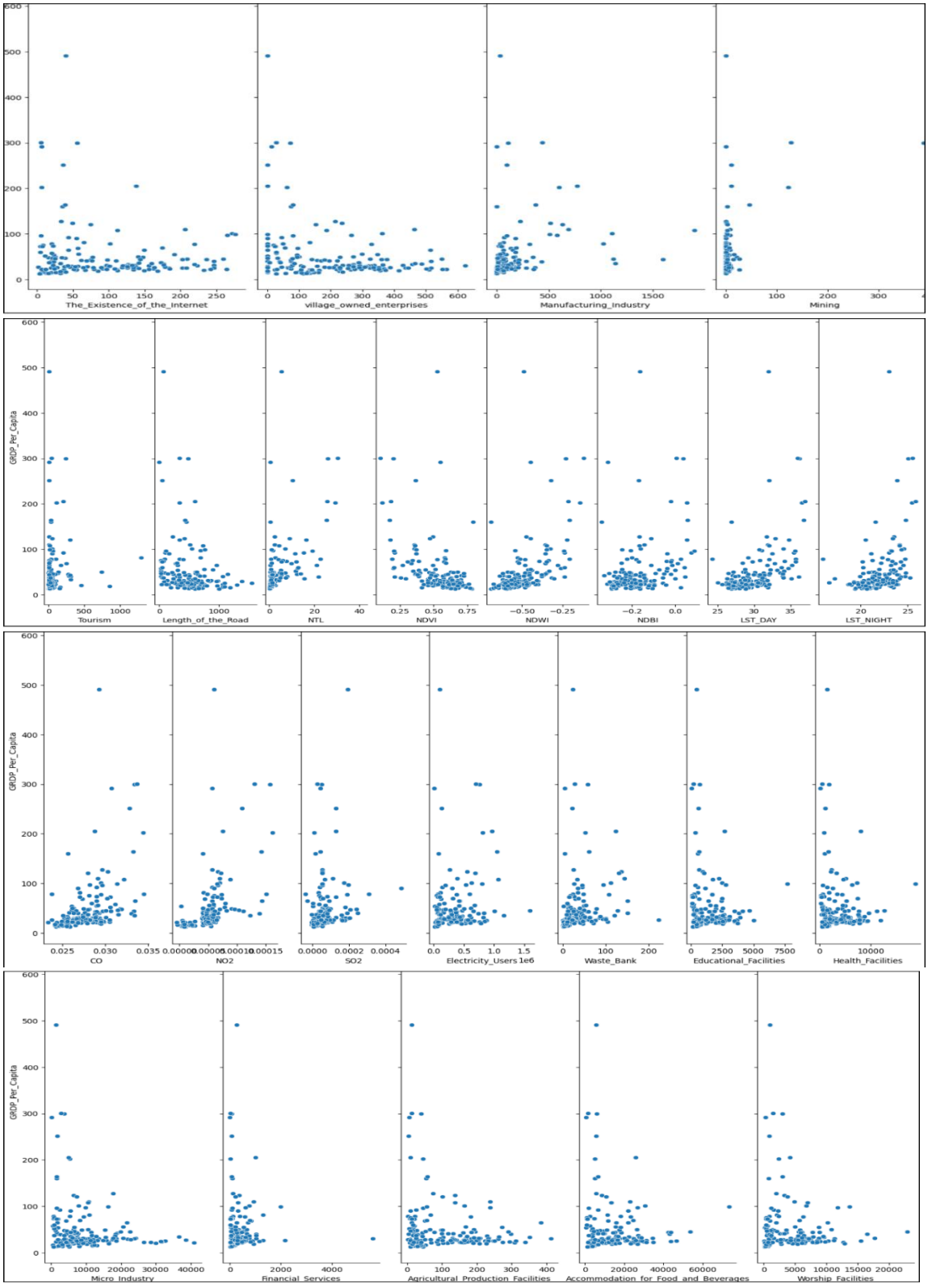


Figure 3. Pairplot

3.2. Variable Selection

In building a linear regression model, selection is carried out using the Pearson correlation test with a moderate to strong relationship. The results of the calculations are shown in Table 2, sorted from the highest value.

Table 2. Results of the pearson correlation calculation

Variable	Pearson	Degree of Closeness	Direction
Night Time Light (NTL)	0.67	Strong	Positive
Normalized Difference Water Index (NDWI)	0.66	Strong	Positive
Nitrogen Dioxide (NO_2)	0.61	Strong	Positive
Day Time Land Surface Temperature (LST Day)	0.57	Strong Enough	Positive
Carbon Monoxide (CO)	0.55	Strong Enough	Positive
Industri Manufaktur	0.48	Strong Enough	Positive
Night Time Land Surface Temperature (LST Night)	0.44	Strong Enough	Positive
Normalized Difference Built-Up Index (NDBI)	0.42	Strong Enough	Positive
Normalized Difference Vegetation Index (NDVI)	-0.61	Strong	Negative

Based on Table 2, it can be seen that the selected variables used in the development of the linear regression model are Night Time Light (NTL), Normalized Difference Water Index (NDWI), Nitrogen Dioxide (NO2), Day Time Land Surface Temperature (LST Day), Carbon Monoxide (CO), Manufacturing Industry, Night Time Land Surface Temperature (LST Night), Normalized Difference Built-Up Index (NDBI), and Normalized Difference Vegetation Index (NDVI). Next, a multicollinearity test was conducted using VIF. The results of the VIF calculations are displayed in Table 3.

Based on Table 3, it can be identified that the variables Normalized Difference Water Index (NDWI), Day Time Land Surface Temperature (LST Day), Normalized Difference Built-Up Index (NDBI), and Normalized Difference Vegetation Index (NDVI) have VIF values of 99.64, 20.48, 12.13, and 122.76, respectively, where a VIF value > 10 indicates that the estimated regression coefficients are weak due to multicollinearity. Variables with VIF values > 10 will be eliminated. Therefore, the selected variables for the development of the linear regression model are Night Time Light (NTL), Nitrogen Dioxide (NO2), Carbon Monoxide (CO), Manufacturing Industry, and Night Time Land Surface Temperature (LST Night).

Table 3. Results of VIF calculation after pearson correlation test

Variable	VIF
Night Time Light (NTL)	5.32
Normalized Difference Water Index (NDWI)	99.64
Nitrogen Dioxide (NO_2)	4.74
Day Time Land Surface Temperature (LST Day)	20.48
Carbon Monoxide (CO)	7.32
Manufacturing Industry	0.48
Night Time Land Surface Temperature (LST Night)	0.44
Normalized Difference Built-Up Index (NDBI)	0.42
Normalized Difference Vegetation Index (NDVI)	122.76

Before building the machine learning model, variable selection is carried out by examining the mutual information values. After that, 10 variables are selected based on the highest mutual information values. The results of the calculations can be seen in the following Table 4, which is sorted by the highest mutual information values.

Table 4. Results of mutual information calculation

Variable	Mutual Information Value
Night Time Light (NTL)	0.41
Nitrogen Dioxide (NO_2)	0.32
Normalized Difference Vegetation Index (NDVI)	0.30
Normalized Difference Water Index (NDWI)	0.29
Manufacturing Industry	0.27
Carbon Monoxide (CO)	0.26
The Number of Villages with the Presence of Kiosks Selling Supplies	0.22
Jumlah Fasilitas Kesehatan	0.21
Day Time Land Surface Temperature (LST Day)	0.21
Night Time Land Surface Temperature (LST Night)	0.19

It should be identified that the 10 variables with the highest mutual information values are Night Time Light (NTL), Nitrogen Dioxide (NO2), Normalized Difference Vegetation Index (NDVI), Normalized Difference Water Index (NDWI), Manufacturing Industry, Carbon Monoxide (CO), Number of Villages with Agricultural Production Supply Kiosks, Number of Health Facilities, Day Time Land Surface Temperature (LST Day), and Night Time Land Surface Temperature. (LST Night). Next, a multicollinearity test was conducted on the selected variables using the Variance Inflation Factor (VIF). The results of the VIF calculation are obtained in the following Table 5.

Table 5. Results of VIF calculation after analyzing autual information values

Variable	VIF
Night Time Light (NTL)	6.07
Nitrogen Dioxide (NO_2)	4.80
Normalized Difference Vegetation Index (NDVI)	95.08
Normalized Difference Water Index (NDWI)	90.93
Manufacturing Industry	3.87
Carbon Monoxide (CO)	2.61
The Number of Villages with the Presence of Kiosks Selling Supplies	2.61
Jumlah Fasilitas Kesehatan	3.04
Day Time Land Surface Temperature (LST Day)	14.32
Night Time Land Surface Temperature (LST Night)	5.25

Based on Table 5, it can be identified that the variables Normalized Difference Vegetation Index (NDVI), Normalized Difference Water Index (NDWI), and Day Time Land Surface Temperature (LST Day) have VIF values of 95.08, 90.93, and 14.32, respectively, where a VIF value > 10 indicates that the estimated regression coefficients are weak due to multicollinearity (Shrestha, 2020). Variables with VIF values > 10 will be eliminated. Therefore, the selected variables for the development of the machine learning model are Night Time Light (NTL), Nitrogen Dioxide (NO2), Manufacturing Industry, Carbon Monoxide (CO), the Number of Villages with Agricultural Production Supply Kiosks, the Number of Health Facilities, and Night Time Land Surface Temperature (LST Night).

3.3. Development of GRDP Per Capita Estimation Model

In this research, the creation of linear regression and machine learning models aims to estimate the GDP per capita map based on the spatial characteristics of an area by combining data from satellite imagery, OpenStreetMap, village potential data, directories of large mining companies, and directories of the manufacturing industry that have undergone a previous variable selection stage. This model was built using linear regression and machine learning with three algorithms: neural network, random forest regression, and support vector regression. The GridSearchCV method with 5-fold cross-validation is used for the purpose of tuning parameters in machine learning models.

The first machine learning model built is a neural network. The second machine learning model built is Random Forest Regression. The third machine learning model built is Support Vector Regression. The model with the best combination of hyperparameters is chosen to map GRDP per capita. The specifications of the machine learning model used are presented in Table 6.

Table 6. Hyperparameter specifications

Method	Specifications	Value
Neural Network	Epochs	20
	Batch_Size	32
	Units	128
	Activation	Tanh
	Learning_Rate	0.001
Random Forest Regression	n_estimators	100
	Max_depth	None
	Min_samples_split	10
	Min_samples_leaf	2
Support Vector Regression	C	1
	Gamma	0.1
	kernel	rbf

The selected specifications for the neural network model are an epoch of 20 means the model is trained for 20 full iterations through the entire training dataset, allowing weight updates more than once to improve accuracy. Batch 32 means the training data is divided into groups of 32 samples, and weight updates are performed after each group is processed. The number of neurons in the hidden layer is 128. Activation using hyperbolic tangent (tanh) means the function is used as the activation function, mapping inputs to a range between -1 and 1. Learning rate 0.001 means the model updates its weights in small steps of 0.001 during training to optimize performance.

The second machine learning model built is Random Forest Regression. N_estimators set to 100 in the Random Forest Regression model means the model uses 100 decision trees to make predictions. The final result is the average of the predictions from all the trees. No max_depth means there is no maximum depth limit set for the decision trees in the model. Min samples split 10 means that each decision tree in the model will only split a node if there are at least 10 samples in it. Min_samples_leaf 2 means that each leaf in the decision tree must contain at least 2 samples.

The third machine learning model built is Support Vector Regression. C set to 1 in model represents the regularization value used to control the trade-off between a larger margin and errors in the training data. Gamma set to 0.1 in the Support Vector Regression (SVR) model represents a parameter that controls how far the influence of a single data sample extends to the model. RBF Kernel (Radial Basis Function) means the kernel function used in the model to measure the similarity between data points. To determine the better and selected method for predicting GRDP per capita at the sub-district level, a numerical evaluation was conducted using the three measures outlined in point 2.9.

Table 7. Model evaluation results

Method	RMSE (Million IDR)	MAE (Million IDR)	MAPE
Neural Network	15.26	10.43	28.77%
Random Forest Regression	15.33	10.37	28.63%
Support Vector Regression	15.97	10.33	26.11%
Linear Regression	20.42	13.66	38.60%

The four machine learning models and linear regression produced a MAPE of less than 50%, which means that according to the MAPE significance table, the forecasts are quite accurate. The machine learning model using the SVR algorithm yielded an MAE of 10.33 million IDR and a MAPE of 26.11%. The SVR model achieved the smallest MAE and MAPE evaluations among the other models. Therefore,

this study uses SVR to estimate GDP per capita at the administrative district level in the provinces of Bali, NTB, and NTT.

3.4. Mapping of GRDP Per Capita at the Sub-District Level

This research produces a map of GRDP per capita at the administrative district level. This map is expected to assist local governments in monitoring the welfare of their communities. The visualization of GRDP per capita at the sub-district level for the year 2021 in the provinces of Bali, NTB, and NTT, along with direct monitoring or verification in the field using Google Earth. The research locus map of Bali, West Nusa Tenggara, and East Nusa Tenggara provinces can be seen in Figure 4. The mapping was conducted using the natural breaks method and divided the GRDP per capita results into 5 classes. The mapping results are shown in Figure 5.

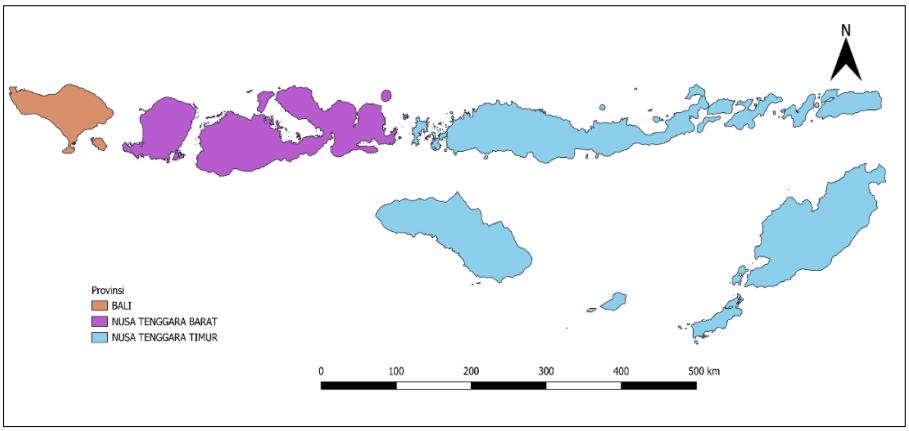


Figure 4. Map of Bali, NTB, And NTT Provinces

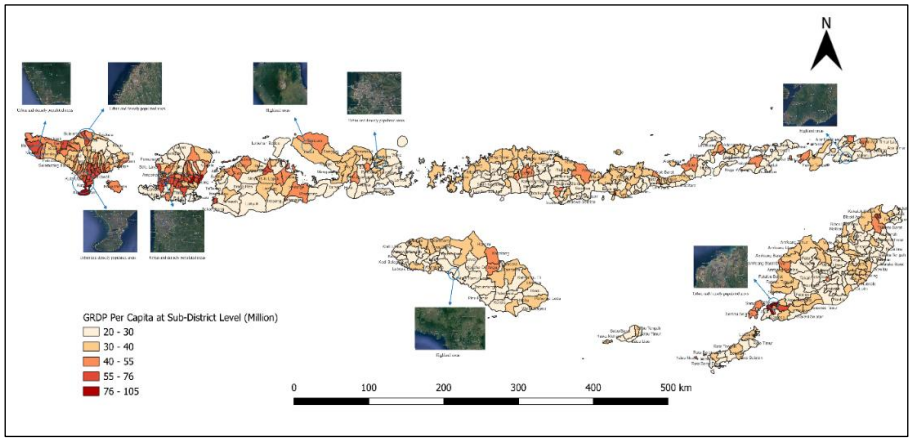


Figure 5. GRDP per capita for Districts in Bali, NTB, and NTT

It can be identified in the image that the districts in the western part have a higher GRDP per capita compared to the GRDP per capita in the eastern part, so the further east the district is, the lower the GRDP per capita value. Urban areas that serve as centers of economic activity have a high GRDP per capita, such as in the southern district of Bali Province, specifically Kuta District, which has a GRDP per capita of 90.31 million. The Cakranegara District in the city of Mataram, NTB Province, which is an urban area, also has a high GDP per capita of 55.05 million. The Kelapa Lima District in the city of Raja, NTT, which is an urban and densely populated area, has a high GRDP per capita of 105.29 million. In addition, districts that are dominated by highlands have a low GRDP per capita, such as Alor Barat Daya District, which has a GDP per capita value of 25.86 million.

Overall, areas dominated by highlands are estimated to be lower than those in the surrounding urban areas. Regions with high GRDP per capita have high values of NTL, CO, and NO_2 as well. The region also has a significant number of manufacturing industries, a number of villages with kiosks selling agricultural production tools, and a large number of healthcare facilities. In addition, the region also has warmer nighttime temperatures compared to areas with low GRDP per capita.

3.5. Mapping Regional Inequality

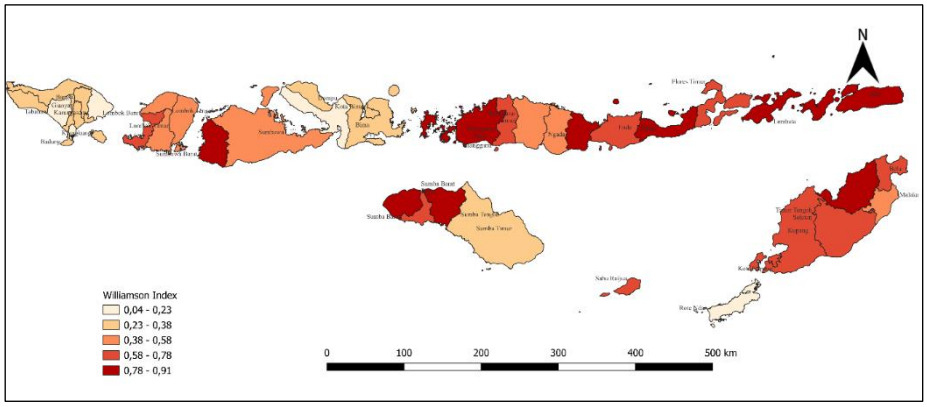


Figure 6. Distribution Map of the Williamson Index in Bali, NTB, and NTT

Figure 6 shows the mapping results of the Williamson Index by district in the provinces of Bali, West Nusa Tenggara (NTB), and East Nusa Tenggara (NTT). It can be seen that the eastern region has a higher disparity between areas compared to the western region. In Bali Province, Denpasar City has the lowest level of regional inequality within the area, with a Williamson Index value of 0.0382, while Klungkung Regency has the highest level of regional inequality with a Williamson Index value of 0.3571. In NTB Province, Mataram City has the lowest level of regional inequality within the area, with a Williamson Index of 0.1859, while West Sumbawa Regency has the highest level of inequality with a Williamson Index value of 0.8076. In NTT Province, Rote Ndao Regency has the lowest level of inequality with a Williamson Index value of 0.22, while West Manggarai Regency has the highest level of inequality with a Williamson Index value of 0.9106.

4. Conclusion

Overall, the estimation model for GRDP per capita at the sub-district level, using linear regression and machine learning, produced a MAPE of less than 50%, indicating that the model forecasts with reasonable accuracy. The support vector regression model was chosen to estimate GRDP per capita at the sub-district level because it resulted in the smallest MAE and MAPE, with values of 10.33 million and 26.11%, respectively. The estimation results indicate that high GRDP per capita is found in urban areas that serve as centers of economic activity, while low GRDP per capita is observed in highland areas with minimal economic activity. Regions with high GRDP per capita also exhibit high levels of NTL, CO, and NO2. These areas are characterized by a large number of manufacturing industries, numerous villages with kiosks selling agricultural production inputs, and a high concentration of healthcare facilities. Additionally, these regions experience warmer nighttime temperatures compared to areas with lower GRDP per capita. Analysis using the Williamson Index reveals that the eastern region has a higher level of inequality compared to the western region.

Ethics approval

Not required.

Acknowledgments

Not required.

Competing interests

All the authors declare that there are no conflicts of interest.

Funding

This study received no external funding.

Underlying data

This research uses secondary data obtained from Badan Pusat Statistik, Satellite Imagery, and OpenStreetMap.

Credit Authorship

I Made Satria Ambara Putra: Data Curation, Visualization, Writing-Original Draft Preparation. **Rindang Bangun Prasetyo:** Methodology, Investigation, Supervision. **Candra Adi Wiguna:** Validation, Writing-Reviewing and Editing.

References

[1] BPS, Gross Domestic Product of Indonesia by Expenditure 2019-2023. Jakarta: Badan Pusat Statistik, 2024.

[2] M. Nasution, "Ketimpangan Antar Wilayah & Hubungannya dengan Belanja Pemerintah: Studi di Indonesia [Regional Disparities & Their Relationship with Government Spending: A Study in Indonesia]," *J. budg.*, vol. 5, no. 2, pp. 84–102, Nov. 2020, doi: 10.22212/jbudget.v5i2.101.

[3] Dinas Komunikasi dan Informatika Kota Depok, *Indikator Ekonomi Kecamatan Kota Depok 2018 [Economic Indicators of Subdistricts in Depok City, 2018]*, 2019.

[4] R. Capello, "Regional Economics, 0 ed." *Routledge*, 2015. doi: 10.4324/9781315720074.

[5] N. M. Coe, P. F. Kelly, and H. W.-C. Yeung, *Economic geography: a contemporary introduction, Third edition*. Hoboken, NJ: Wiley-Blackwell, 2020.

[6] BPS Kab. Sleman, *Produk Domestik Regional Bruto Kecamatan di Kabupaten Sleman 2016 [Gross Regional Domestic Product of Subdistricts in Sleman Regency, 2016]*, Kabupaten Sleman: Sleman: Badan Pusat Statistik Kabupaten Sleman, 2016.

[7] Sjafrizal, *Ekonomi wilayah dan perkotaan, Cetakan ke-1 [Regional and Urban Economy, 1st Edition]*, Jakarta: PT RajaGrafindo Persada, 2012.

[8] BPS Provinsi Jawa Tengah, *Analisis Indeks Williamson Provinsi Jawa Tengah 2017-2021 [Analysis of the Williamson Index in Central Java Province, 2017-2021]*, Jawa Tengah: Semarang: Badan Pusat Statistik Jawa Tengah, 2021.

[9] S. Pratama, "Prediksi Harga Tanah Menggunakan Algoritma Linear Regression [Land Price Prediction Using the Linear Regression Algorithm]," *Technologia*, vol. 7, no. 2, Jun. 2016, doi: 10.31602/tji.v7i2.624.

[10] V. L. Delimah Pasaribu et al., "Forecast Analysis of Gross Regional Domestic Product based on the Linear Regression Algorithm Technique," *TEM Journal*, pp. 620–626, May 2021, doi: 10.18421/TEM102-17.

[11] S. C. Agu, F. U. Onu, U. K. Ezemagu, and D. Oden, "Predicting gross domestic product to macroeconomic indicators," *Intelligent Systems with Applications*, vol. 14, p. 200082, May 2022, doi: 10.1016/j.iswa.2022.200082.

[12] T. M. H. Hope, "Linear regression, in Machine Learning," *Elsevier*, 2020, pp. 67–81. doi: 10.1016/B978-0-12-815739-8.00004-3.

[13] N. Puttanapong, N. Prasertsoong, and W. Peechapat, "Predicting Provincial Gross Domestic Product Using Satellite Data and Machine Learning Methods: A Case Study of Thailand," *Asian Development Review*, vol. 40, no. 02, pp. 39–85, Sep. 2023, doi: 10.1142/S0116110523400024.

[14] N. D. Muchisha, N. Tamara, A. Andriansyah, and A. M. Soleh, "Nowcasting Indonesia's GDP Growth Using Machine Learning Algorithms," *IJSA*, vol. 5, no. 2, pp. 355–368, Jun. 2021, doi: 10.29244/ijsa.v5i2p355-368.

[15] A. Richardson, T. Van Florenstein Mulder, and T. Vehbi, "Nowcasting GDP using machine-learning algorithms: A real-time assessment," *International Journal of Forecasting*, vol. 37, no. 2, pp. 941–948, Apr. 2021, doi: 10.1016/j.ijforecast.2020.10.005.

[16] S. Sa'adah and M. S. Wibowo, "Prediction of Gross Domestic Product (GDP) in Indonesia Using Deep Learning Algorithm, in 2020 3rd International Seminar," *Research of Information Technology and Intelligent Systems (ISRITI)*, Yogyakarta, Indonesia: IEEE, Dec. 2020, pp. 32–36. doi: 10.1109/ISRITI51436.2020.9315519.

[17] H. Lai, "A comparative study of different neural networks in predicting gross domestic product," *Journal of Intelligent Systems*, vol. 31, no. 1, pp. 601–610, May 2022, doi: 10.1515/jisys-2022-0042.

[18] J. Saputra, B. Subartini, J. H. F. Purba, S. Supian, and Y. Hidayat, *An Application of Genetic Algorithm Approach and Cobb-Douglas Model for Predicting the Gross Regional Domestic Product by Expenditure-Based* in Indonesia, 2019.

[19] A. F. Syah, "Penginderaan Jauh dan Aplikasinya di Wilayah Pesisir dan Lautan [Remote Sensing and Its Applications in Coastal and Marine Areas]," *Jurnal Kelautan*, vol. 3, pp. 18–28, 2010.

[20] Z. Wang et al., "Exploring the Potential of OpenStreetMap Data in Regional Economic Development Evaluation Modeling," *Remote Sensing*, vol. 16, no. 2, p. 239, Jan. 2024, doi: 10.3390/rs16020239.

[21] K. Faisal and A. Shaker, "The Use of Remote Sensing Technique to Predict Gross Domestic Product (GDP): An Analysis of Built-Up Index and GDP in Nine Major Cities in Canada," *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, vol. XL–7, pp. 85–92, Sep. 2014, doi: 10.5194/isprsarchives-XL-7-85-2014.

[22] S. R. Putri, A. W. Wijayanto, and S. Pramana, "Multi-source satellite imagery and point of interest data for poverty mapping in East Java, Indonesia: Machine learning and deep learning approaches," *ScienceDirect*, Accessed: Jul. 04, 2024. [Online]. Available: https://www.sciencedirect.com/ science/article/abs/pii/S2352938522001975

[23] BPS Kab. Musi Rawas, *Produk Domestik Regional Bruto Kabupaten Musi Rawas Menurut Pengeluaran 2016-2020 [Gross Regional Domestic Product of Musi Rawas Regency by Expenditure, 2016-2020]*, Kabupaten Musi Rawas: Musi Rawas: Badan Pusat Statistik Kabupaten Musi Rawas, 2021.

[24] BPS Kab. Badung, *Produk Domestik Regional Bruto Kabupaten Badung Menurut Lapangan Usaha [Gross Regional Domestic Product of Badung Regency by Business Sector]*, Kabupaten Badung: Badung: Badan Pusat Statistik Kabupaten Badung, 2022.

[25] M. Urbac, A. Junaidi, M. Syukur, N. Nurhamidah, and R. Ferial, "Kajian Aspek Geospasial Untuk Percepatan Pembangunan dan Pemberdayaan Desa Binaan Kota Padang [Study of Geospatial Aspects for Accelerating Development and Empowering Fostered Villages in Padang City]," *JLBI*, vol. 12, no. 4, pp. 198–204, Dec. 2023, doi: 10.32315/jlbi.v12i4.83.

[26] D. Reynard, "Five Classes of Geospatial Data and The Barriers to Using Them," *Geography Compass*, vol. 12, no. 4, p. e12364, Apr. 2018, doi: 10.1111/gec3.12364.

[27] P. Pan, R. Li, and Y. Zhang, "Predicting punching shear in RC interior flat slabs with steel and FRP reinforcements using Box-Cox and Yeo-Johnson transformations," *Case Studies in Construction Materials*, vol. 19, p. e02409, Dec. 2023, doi: 10.1016/j.cscm.2023.e02409.

[28] B. T. Suryanto, A. A. Imron, and D. A. R. Prasetyo, "The Correlation between Students Vocabulary Mastery and Speaking Skill," *ijoeel*, vol. 3, no. 1, pp. 10–19, Jun. 2021, doi: 10.33650/ijoeel.v3i1, 2042.

[29] G. Zeng, "A Unified Definition of Mutual Information with Applications in Machine Learning," *Mathematical Problems in Engineering*, vol. 2015, pp. 1–12, 2015, doi: 10.1155/2015/201874.

[30] P. Laarne, M. A. Zaidan, and T. Nieminen, "ennemi: Non-linear correlation detection with mutual information," *SoftwareX*, vol. 14, p. 100686, Jun. 2021, doi: 10.1016/j.softx.2021.100686.

[31] J. Zhao, Y. Zhou, X. Zhang, and L. Chen, "Part mutual information for quantifying direct associations in networks," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 113, no. 18, pp. 5130–5135, May 2016, doi: 10.1073/pnas.1522586113.

[32] N. Shrestha, "Detecting Multicollinearity in Regression Analysis," *AJAMS*, vol. 8, no. 2, pp. 39–42, Jun. 2020, doi: 10.12691/ajams-8-2-1.

[33] M. I. Jordan and T. M. Mitchell, "Machine learning: Trends, perspectives, and prospects," *Science*, vol. 349, no. 6245, pp. 255–260, Jul. 2015, doi: 10.1126/science.aaa8415.

[34] I. Cholissodin, *Buku Ajar AI, Machine Learning & Deep Learning [AI, Machine Learning & Deep Learning textbook]*, Malang: Malang: Filkom Universitas Brawijaya, 2020.

[35] P.-C. Chang, Y.-W. Wang, and C.-H. Liu, "The development of a weighted evolving fuzzy neural network for PCB sales forecasting," *Expert Systems with Applications*, pp. 86–96, 2007.