JUSIKOM PRIMA (Jurnal Sistem Informasi dan Ilmu Komputer Prima)
Vol. 8 No. 1, August 2024 E-ISSN:2580-2879

Analysis And Prediction Of Global Population Using Random Forest Regression

Jepri Banjarnahor[1], Catherine Jeta Jones[2], Esthin Mitra Gulo[3], Angelia C.S Sianturi[4]*
Universitas Prima Indonesia Medan, Indonesia
jepribanjarnahor@unprimdn.ac.id

ABSTRACT

This research evaluates the performance of the random forest regression algorithm in predicting global population growth from time series data. The findings indicate that population growth predictions remain stable, with an annual increase of less than 1%. Model analysis using evaluation metrics such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and model scores demonstrates high quality, with average values below 0.5. These results imply that the model can deliver optimal and consistent outcomes. The model shows potential for accurate predictions when tested on datasets. Further analysis reveals a population increase of 0.88% in 2024, equating to an addition of approximately 70,206,291 people, and a rise of 0.91% in 2025, adding about 73,524,552 people.

Keywords: Analysis, Prediction, ,Random Forest Regression, world population, growth population.

INTRODUCTION

Population growth is a phenomenon that significantly impacts various aspects of human life. As the number of individuals within a population increases, it affects economic development, resource allocation, social structures, and environmental sustainability. Understanding and managing population growth is critical for ensuring that a society can meet the needs of its citizens and maintain a balance with the natural environment. This task, however, is complex and requires a structured approach informed by a deep understanding of the factors driving population growth and the methodologies available to analyze it [1].

Effectively managing population growth involves forecasting future trends, understanding the variables that contribute to these trends, and making informed decisions based on this knowledge. Previous research has explored population growth using a variety of methodologies, each offering unique insights into the phenomenon. Traditional demographic methods, such as analyzing birth and death rates, migration patterns, and fertility rates, have long been used to study population growth. However, as data collection and computational technologies have advanced, so too have the methods used to analyze population growth [2].

In recent years, researchers have increasingly turned to machine learning and advanced statistical models to study population dynamics. These methods allow for the analysis of large and complex datasets, enabling researchers to identify patterns and trends that might not be apparent using traditional methods. For example, Alter Lasarudin and Rubiyanto Maku (2022) used a neural network (NN) model to predict population growth trends. Their study was based on data obtained through interviews and observations, and they utilized time series techniques to model population changes over time. Neural networks are particularly useful in this context because they can learn from complex, non-linear data and adapt to changing patterns, making them well-suited for predicting population trends [3].

Neural networks, as used by Lasarudin and Maku, are a form of artificial intelligence that mimics the way the human brain processes information. These models consist of layers of interconnected nodes, or neurons, that process input data and generate predictions. One of the key advantages of neural networks is their ability to model complex relationships between variables, which is essential in population studies where multiple factors—such as economic conditions, migration patterns, and health care access—interact in ways that are not easily captured by traditional linear models.

In contrast, Hoiriyah Hoiriyah and Herry Adi Chandra (2023) employed multiple linear regression to forecast the national population growth rate. Multiple linear regression is a statistical technique that models the relationship between a dependent variable (in this case, population growth) and two or more independent variables (such as income levels, education, and health care access). This method assumes that there is a linear relationship between the independent variables and the dependent variable, which simplifies the analysis but may not fully capture the complexity of population dynamics.

Despite its limitations, multiple linear regression remains a widely used tool in demographic studies due to its simplicity and interpretability. It allows researchers to identify which factors have the most significant impact on population growth and to quantify the strength of these relationships. Hoiriyah and Chandra's study highlighted the importance of socio-economic factors in shaping national population trends and provided valuable insights for policymakers seeking to manage population growth effectively[4].

The primary distinction of the present study lies in its focus on global population growth analysis and prediction using the random forest regression algorithm. Random forest is an ensemble learning method that operates by constructing multiple decision trees during training and outputting the mode of the classes (for classification) or the mean prediction (for

regression). This method is particularly advantageous for analyzing large and complex datasets with numerous variables, making it an ideal tool for studying global population growth.

The random forest algorithm offers several key advantages over traditional methods like neural networks and multiple linear regression. Firstly, random forests are robust to overfitting, which is a common issue in predictive modeling. Overfitting occurs when a model performs well on training data but poorly on unseen data, often because the model has learned noise rather than the underlying pattern. By averaging the results of multiple decision trees, random forests reduce variance and improve the generalizability of the model, making it more reliable when applied to new data.

Secondly, random forests are capable of handling both numerical and categorical data, which is essential for a comprehensive analysis of population growth. The ability to model interactions between variables and assess the importance of each variable in the prediction process makes random forest a powerful tool for understanding the drivers of global population growth. By using this algorithm, the study aims to provide deeper and more accurate insights into the factors that contribute to population growth on a global scale.

Another significant aspect of this study is the use of data from sources previously unexamined, specifically Kaggle.com. Kaggle is an online platform that hosts a wide range of datasets for machine learning and data science projects. The data available on Kaggle is often crowdsourced and curated, providing a rich resource for researchers. By incorporating data from Kaggle, this study aims to explore new dimensions of population growth that have not been previously considered. This novel approach allows for the inclusion of diverse variables and datasets, enhancing the robustness and relevance of the study's findings[5].

The insights gained from this research have important implications for policy-making. By understanding the factors that drive global population growth, governments and international organizations can develop more effective strategies for managing population changes. This includes planning for infrastructure, healthcare, education, and resource allocation—all of which are critical for sustainable development. Moreover, the study's findings can inform international efforts to address global challenges such as climate change, food security, and migration, which are closely linked to population dynamics.

In conclusion, the proposed approach in this study is expected to offer deeper and more relevant insights into the dynamics of global population growth, supporting more effective planning and decision-making in the context of sustainable development. By leveraging the power of random forest regression and novel data sources, this research contributes to a more nuanced

understanding of population growth, providing a valuable tool for policymakers and researchers alike[6].

LITERATURE REVIEW

This The analysis and prediction of global population growth is a complex and multifaceted endeavor that requires sophisticated methodologies capable of handling the intricacies of the data involved. This study is classified as a descriptive-analytical research effort, focusing on the use of the Random Forest Regression algorithm to analyze and predict global population growth trends. The primary objective of this research is to enhance the accuracy of population growth predictions, identify key factors influencing these trends, and assess the implications of these predictions for sustainable development policies.

The Importance of Accurate Population Growth Predictions

Accurate predictions of population growth are crucial for effective planning and policy-making. Governments and international organizations rely on population forecasts to plan for future demands on resources, infrastructure, healthcare, education, and employment. Inaccurate predictions can lead to resource shortages, inadequate infrastructure, and social unrest. Therefore, the development and application of advanced predictive models are essential for understanding and managing population dynamics on a global scale.

Traditional methods of predicting population growth, such as linear regression and time series analysis, have been widely used in demographic studies. However, these methods often fall short when dealing with the complexities of global population growth, which involves numerous variables that interact in non-linear ways. The Random Forest Regression algorithm, an ensemble learning method that builds multiple decision trees and merges them to improve prediction accuracy, is particularly well-suited for this task. Its ability to handle large, complex datasets with numerous variables makes it an ideal tool for studying global population growth.

The Random Forest Regression Algorithm

The Random Forest Regression algorithm has gained popularity in recent years due to its robustness and versatility. It is particularly effective in situations where the relationship between variables is non-linear or where the data contains a mixture of numerical and categorical variables. This algorithm works by constructing multiple decision trees during the training phase and outputting the average of the predictions from these trees. This ensemble

approach helps to reduce overfitting, a common issue in machine learning where a model performs well on training data but poorly on new, unseen data.

One of the key advantages of Random Forest Regression is its ability to provide insights into the importance of different variables in the prediction process. By analyzing the contribution of each variable to the overall prediction, researchers can identify which factors are most influential in driving population growth. This is particularly valuable in the context of global population studies, where numerous socio-economic, environmental, and demographic factors interact to influence growth trends.

Methodological Approaches In Population Growth Studies

Previous studies on population growth have employed a variety of methodological approaches. For instance, Alter Lasarudin and Rubiyanto Maku (2022) utilized a neural network model to predict population growth trends based on historical data obtained through interviews and observations. Their approach, which involved the use of time series techniques, allowed for the modeling of complex, non-linear relationships between variables. However, while neural networks are powerful tools for prediction, they can be prone to overfitting and often require large amounts of data to perform well.

Similarly, Hoiriyah Hoiriyah and Herry Adi Chandra (2023) used multiple linear regression to forecast national population growth rates. Multiple linear regression is a simpler and more interpretable method than neural networks, but it assumes a linear relationship between variables, which may not always be the case in population studies. The limitations of these methods highlight the need for more advanced techniques, such as Random Forest Regression, that can handle the complexities of global population growth data.

In this study, the research methodology involves several exploratory steps designed to measure and enhance the prediction accuracy of the Random Forest Regression algorithm. These steps include testing the algorithm with historical population data, developing a predictive model, and refining the model to improve its accuracy. The ultimate goal is to create a model that can provide more realistic projections of global population growth, offering a stronger foundation for stakeholders involved in global population planning.

Key Factors Influencing Population Growth

One of the primary objectives of this research is to identify the main factors affecting global population growth. Population growth is influenced by a wide range of factors, including birth rates, death rates, migration patterns, economic conditions, healthcare access, education levels, and environmental factors. By using the Random Forest Regression algorithm, this study aims to provide a more holistic understanding of how these variables contribute to population growth. The Random Forest Regression algorithm is particularly well-suited for this type of analysis because it can model complex interactions between variables and assess the importance of each variable in the prediction process. This allows researchers to identify not only the direct effects of individual variables on population growth but also the indirect effects that occur through interactions with other variables. This comprehensive analysis is expected to yield valuable insights into the drivers of global population growth and inform more effective policy-making.

The Research Process

The research process in this study involves four key stages: preprocessing, feature extraction, model prediction, and model validation.

Preprocessing: This stage involves cleaning and preparing the data for analysis. This includes handling missing data, normalizing the data, and transforming categorical variables into numerical forms that can be used in the model.

Feature Extraction: In this stage, the relevant features (or variables) that are expected to influence population growth are identified and extracted from the data. Feature extraction is crucial for improving the accuracy of the model, as it ensures that the model focuses on the most important variables.

Model Prediction: During this stage, the Random Forest Regression algorithm is applied to the data to make predictions about future population growth. The algorithm builds multiple decision trees based on the training data and combines their predictions to generate a final forecast.

Model Validation and Evaluation: The final stage involves validating the model's predictions by comparing them to actual population growth data. This is done using a variety of statistical metrics to assess the accuracy and reliability of the model. The model is also evaluated to ensure that it is not overfitting the data and that it generalizes well to new, unseen data.

The results of this study are expected to provide a more accurate and comprehensive understanding of global population growth trends. By identifying the key factors driving these

trends and improving the accuracy of population growth predictions, this research will contribute to more effective planning and policy-making in the context of sustainable development.

In conclusion, the use of the Random Forest Regression algorithm in this study represents a significant advancement in the field of population growth analysis. By leveraging the strengths of this algorithm, this research aims to provide deeper insights into the dynamics of global population growth, supporting the development of more effective and sustainable policies for managing population change on a global scale..

Evaluative steps for algorithm accuracy include testing historical data and developing a predictive model that can provide more realistic projections. This aims to offer a stronger foundation for stakeholders involved in global population growth planning. Additionally, this study aims to identify the main factors affecting global population growth. In-depth analysis using the Random Forest Regression algorithm is expected to provide a more holistic understanding of the contribution of each variable to the population growth phenomenon. The research process in prediction and analysis using the Random Forest method includes four stages: preprocessing, feature extraction, model prediction, model validation, and overall evaluation [7].

METHOD

To ensure the research progresses smoothly and is completed on time, a set of research procedures has been established. The research procedures are as follows:.

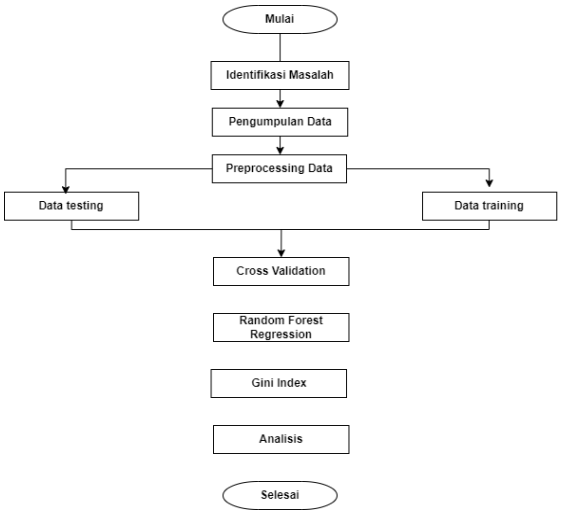

Figure 1. Research methodology flowchart

Problem Identification and Data Collection

The first step in this research involves identifying the key variables necessary for predicting global population growth. These variables are selected based on their potential impact on population dynamics and are drawn from a wide range of demographic, socio-economic, and environmental factors. The selection of these variables is crucial, as it determines the scope of the study and the relevance of the findings. The variables may include birth rates, death rates, migration patterns, economic indicators (like GDP per capita), educational attainment levels, and environmental factors such as urbanization rates and climate conditions.

Data collection is carried out using reliable global databases such as the World Bank, United Nations, and Kaggle.com. These databases provide extensive and up-to-date information on various indicators that influence population growth. The data collected spans from 2020 to 2023, offering a relevant and contemporary basis for the study. The collected data serves as the foundation for all subsequent stages of the research, from preprocessing to model validation and analysis.

Data Preprocessing

Data preprocessing is a critical step in preparing raw data for further analysis. The goal is to clean and transform the data into a format that can be effectively processed by machine learning algorithms, ensuring the quality and accuracy of the results. This step involves several key processes:

1. Data Cleaning: This involves identifying and addressing any inconsistencies or errors in the data, such as missing values, outliers, or duplicates. Missing data can be particularly problematic, and various techniques such as mean imputation, median imputation, or regression-based imputation may be used to fill in gaps. Outliers may be removed or transformed to prevent them from skewing the results.
2. Data Transformation: Transformation includes normalizing or standardizing the data to ensure that all variables operate on the same scale. This is especially important in machine learning, where differences in scale can lead to biased predictions. Normalization typically involves rescaling the data to a range of 0 to 1, while standardization adjusts the data to have a mean of 0 and a standard deviation of 1.
3. Encoding Categorical Variables: If the dataset includes categorical variables, these need to be converted into numerical formats to be used in machine learning models. Common

techniques include one-hot encoding, which creates binary columns for each category, and label encoding, which assigns a unique number to each category.

4. Data Splitting: The dataset is divided into two subsets: a training set and a testing set. The training set, which typically comprises 70-80% of the data, is used to train the machine learning model, while the testing set, which makes up the remaining 20-30%, is used to evaluate the model's performance. This split ensures that the model's accuracy is assessed on data it has not seen during training, providing a realistic measure of its predictive capabilities.

Data Training and Testing

After preprocessing, the next step is to train the machine learning model using the processed data. In this study, the Random Forest Regression algorithm is used due to its ability to handle complex, high-dimensional data and its robustness against overfitting.

The training process involves feeding the model with the training data, allowing it to learn the relationships and patterns within the data. The model constructs multiple decision trees during training, with each tree built from a different subset of the data. These trees collectively form a "forest" that generates predictions by averaging the output of individual trees, thereby improving accuracy and reducing variance.

Once the model has been trained, it is tested on the unseen testing data. This step evaluates the model's ability to generalize its predictions to new data. The accuracy of the model's predictions is measured using various performance metrics, such as Mean Squared Error (MSE), which quantifies the average squared difference between the predicted and actual values.

The formula for Mean Squared Error (MSE) is:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Where:

- \(n \) is the number of data points,
- \(y_i \) is the actual value,
- \(\hat{y}_i \) is the predicted value.

A lower MSE indicates a model that more accurately predicts population growth.

Cross-Validation

To further ensure the reliability of the model, cross-validation is performed. Cross-validation is a technique used to assess how well a model generalizes to an independent dataset. It involves dividing the data into several subsets or "folds." The model is trained on all but one fold and tested on the remaining fold. This process is repeated multiple times, with each fold serving as the test set once. The results from each iteration are then averaged to provide a more robust estimate of the model's performance.

Cross-validation reduces the risk of overfitting, where a model performs well on the training data but poorly on unseen data. It also ensures that the model's performance is not dependent on the specific split of the data into training and testing sets, providing a more reliable measure of its predictive power**.**

Random Forest and Gini Index

After training and validating the model, the Random Forest algorithm is employed to generate predictions. A key feature of Random Forest is its use of the Gini index to evaluate the purity of nodes within the decision trees. The Gini index measures the likelihood of incorrectly classifying a randomly chosen element if it were randomly labeled according to the distribution of labels in the dataset.

The Gini index formula is:

$$\text{Gini} = 1 - \sum_{i=1}^{n} p_i^2$$

Where:

- \(p_i \) is the proportion of samples belonging to class \(i \) at a particular node.

A node with a lower Gini index indicates a more homogeneous or "pure" node, meaning that most of the data points in that node belong to a single class. The algorithm uses the Gini index to select features that result in the greatest decrease in impurity, thereby optimizing the decision tree structure for more accurate predictions.

Analysis

Once the predictions are generated, the final step involves analyzing the results. This analysis focuses on evaluating the model's performance, identifying patterns in the data, and detecting any anomalies in the predictions. The analysis also involves comparing the predicted population growth rates with actual historical data to assess the model's accuracy and reliability.

This evaluation process is essential for drawing relevant conclusions and understanding the implications of the findings. The insights gained from this analysis can inform policymakers and researchers about the key drivers of global population growth, helping them to develop more effective strategies for managing population changes in the context of sustainable development.

In conclusion, this methodical approach—spanning from problem identification and data collection to model evaluation and analysis—ensures that the research provides accurate and meaningful predictions of global population growth. By leveraging the strengths of the Random Forest Regression algorithm and the Gini index, this study offers valuable insights into the factors driving population change and contributes to the broader field of demographic research and policy-making.

RESULT

Data Analysis

In this study, an analysis of population growth data based on historical datasets of world population growth obtained from the Kaggle.com source was carried out. The purpose of this research is to analyze and predict global population growth using the Random Forest Regression algorithm method. The dataset used includes world population information by country from the period 2020 to 2023, which is then analyzed using the Python programming language. The data variables used in this dataset after going through the data selection process consist of several attributes, such as rank, country name, continent, population in 2023, population in 2022, population in 2020, area, population density, population growth rate and percentage of country population to world population.

Data Preprocessing

Data preprocessing is a set of methods applied to clean, transform, and prepare initial data so that it can be used effectively by machine learning algorithms [9]. At this stage the author cleans

the data so that the data used in the algorithm model can get maximum results and further processing is more structured [10]. In this research, the data preprocessing stage carried out has three stages as follows.

Data selection

Data selection is an important stage in the research process where we select the most relevant and significant subset of data for further analysis [11]. The purpose of data selection is to reduce data complexity and improve the quality of the model to be developed. Below are some of the steps performed as a world population data selection process.

Identification Data

In the obtained dataset, each variable relevant to the research is selected, and variable names are modified to be more easily understood. The changes are illustrated in the image below.

	rank	cca3	country	continent	2023 population	2022 population	2020 population	2015 population	2010 population	2000 population	1990 population	1980 population	1970 population	area (km²)	density (km²)	growth rate	world percentage
0	1	IND	India	Asia	1428627663	1417173173	1396387127	1322866505	1240613620	1059633675	870452165	696828385	557501301	3287590.00	481	0.81%	17.85%
1	2	CHN	China	Asia	1425671352	1425887337	1424929781	1393715448	1348191368	1264099069	1153704252	982372466	822534450	9706961.00	151	-0.02%	17.81%
2	3	USA	United States	North America	338996563	338289857	335942003	324607776	311182845	282398554	248083732	223140018	200328340	9372610.00	37	0.50%	4.25%
3	4	IDN	Indonesia	Asia	277534122	275501339	271857970	259091970	244016173	214072421	182159874	148177096	115228394	1904569.00	148	0.74%	3.47%
4	5	PAK	Pakistan	Asia	240485658	235824862	227196741	210969298	194454498	154369924	115414069	80624057	59290872	881912.00	312	1.98%	3.00%
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
229	230	MSR	Montserrat	North America	4386	4390	4500	5059	4938	5138	10805	11452	11402	102.00	43	-0.09%	0.00%
230	231	FLK	Falkland Islands	South America	3791	3780	3747	3408	3187	3080	2332	2240	2274	12173.00	0	0.29%	0.00%
231	232	NIU	Niue	Oceania	1935	1934	1942	1847	1812	2074	2533	3637	5185	261.00	7	0.05%	0.00%
232	233	TKL	Tokelau	Oceania	1893	1871	1827	1454	1367	1666	1669	1647	1714	12.00	189	1.18%	0.00%
233	234	VAT	Vatican City	Europe	518	510	520	564	596	651	700	733	752	0.44	1177	1.57%	0.00%

Figure 2. Dataset before selection

	rank	negara	benua	populasi_2023	populasi_2022	populasi_2020	luas_area	kepadatan	pertumbuhan	persentase_populasi_dunia
0	1	India	Asia	1428627663	1417173173	1396387127	3287590.0	481	0.81%	17.85%
1	2	China	Asia	1425671352	1425887337	1424929781	9706961.0	151	-0.02%	17.81%
2	3	United States	North America	339996563	338289857	335942003	9372610.0	37	0.50%	4.25%
3	4	Indonesia	Asia	277534122	275501339	271857970	1904569.0	148	0.74%	3.47%
4	5	Pakistan	Asia	240485658	235824862	227196741	881912.0	312	1.98%	3.00%
5	6	Nigeria	Africa	223804632	218541212	208327405	923768.0	246	2.41%	2.80%
6	7	Brazil	South America	216422446	215313498	213196304	8515767.0	26	0.52%	2.70%
7	8	Bangladesh	Asia	172954319	171186372	167420951	147570.0	1329	1.03%	2.16%
8	9	Russia	Europe	144444359	144713314	145617329	17098242.0	9	-0.19%	1.80%
9	10	Mexico	North America	128455567	127504125	125998302	1964375.0	66	0.75%	1.60%
10	11	Ethiopia	Africa	126527060	123379924	117190911	1104300.0	112	2.55%	1.58%
11	12	Japan	Asia	123294513	123951692	125244761	377930.0	338	-0.53%	1.54%
12	13	Philippines	Asia	117337368	115559009	112190977	342353.0	394	1.54%	1.47%
13	14	Egypt	Africa	112716598	110990103	107465134	1002450.0	113	1.56%	1.41%
14	15	DR Congo	Africa	102262808	99010212	92853164	2344858.0	45	3.29%	1.28%

Figure 3. Dataset after selection

Transformation Data

The data transformation process is the stage where data is modified and unified into a suitable format for the application of machine learning [12]. At this stage, the author checks the data types in the dataset. For more details, see the image below.

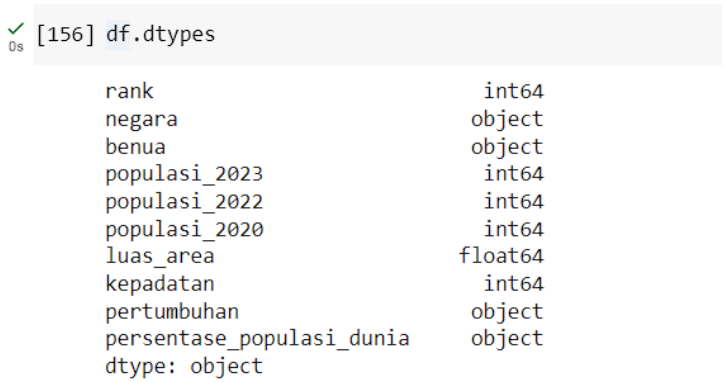

Figure 4. Type of data variable

After checking the dataset variables, it was found that almost every variable in the dataset has an integer data type, which can be directly applied to the Random Forest Regression algorithm. Additionally, the continent variable needs to be transformed so it can be effectively used by the algorithm.

Exploratory Data Analysist

At this stage, the researcher conducts data exploration to analyze anomalies in the dataset to maximize the results of the research. Exploratory Data Analysis (EDA) is a process used to analyze a data set to summarize its main characteristics and understand the dataset's condition [13]. For more details, see the figure below.

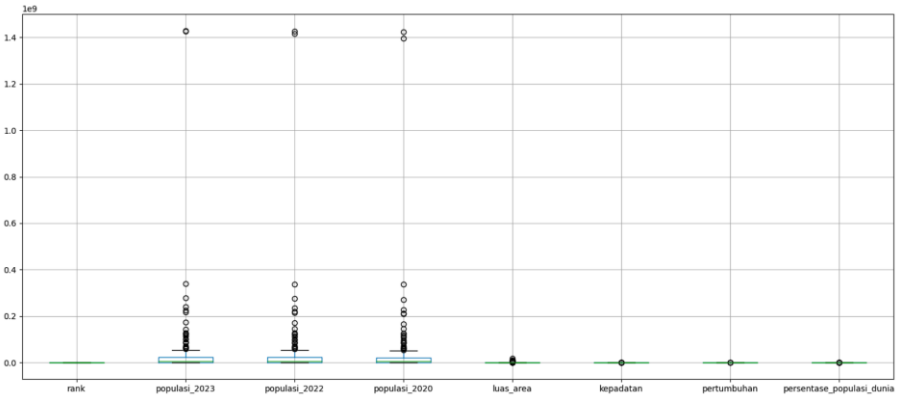

Figure 5. Outliers

Based on the analysis performed on the outlier checks, no significant anomalies were found. The data distribution showed relatively normal and homogeneous characteristics, with no strong linear relationships or striking anomalies among the observed variables.

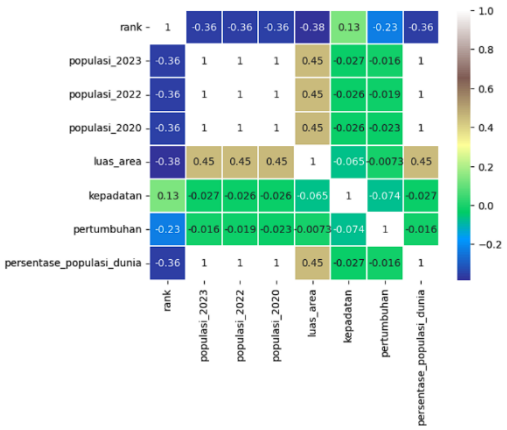

Figure 6. Heatmap

	count	mean	std	min	25%	50%	75%	max
populasi_2023	234.000000	34374424.743590	137386405.597263	518.000000	422598.250000	5643895.000000	23245367.250000	1428627663.000000
populasi_2022	234.000000	34074414.713675	136766424.804728	510.000000	419738.500000	5559944.500000	22476504.750000	1425887337.000000
populasi_2020	234.000000	33501070.952991	135589876.924439	520.000000	415284.500000	5493074.500000	21447979.500000	1424929781.000000
luas_area	234.000000	581449.983590	1761840.665609	0.440000	2650.000000	81199.500000	430425.750000	17098242.000000
kepadatan	234.000000	451.282051	1979.398922	0.000000	39.500000	97.500000	242.750000	21403.000000
rank	234.000000	117.500000	67.694165	1.000000	59.250000	117.500000	175.750000	234.000000
pertumbuhan	234.000000	0.973675	1.234990	-7.450000	0.232500	0.820000	1.685000	4.980000
persentase_populasi_dunia	234.000000	0.429444	1.716421	0.000000	0.010000	0.070000	0.290000	17.850000

Figure 7. Data value checking

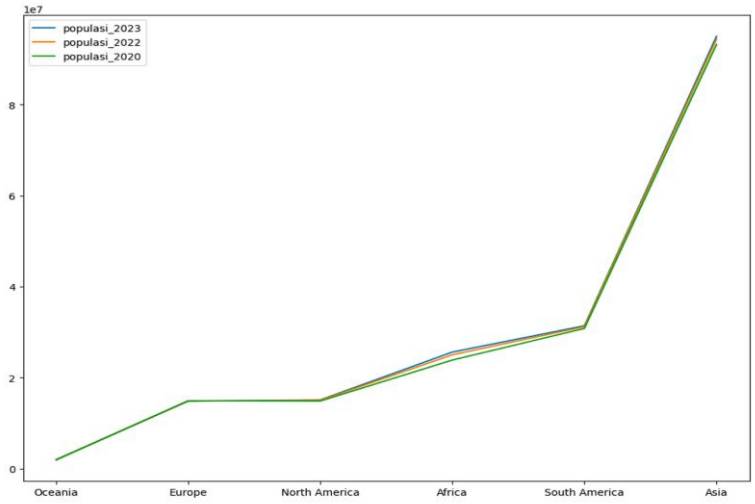

Figure 8. Population per year by Continent

Based on the analysis conducted, correlation examination, and data distribution assessment, no significant anomalies were found. The data distribution exhibits relatively normal and homogeneous characteristics, without strong linear relationships or striking anomalies among the observed variables.

Dataset Splitting

After completing the data preprocessing stage, the prepared data is then utilized for the machine learning model training process. In this process, the model learns to recognize patterns and correlations within the processed data, enabling it to produce accurate predictions [14].

Cross Validation

After the model training process is completed, the cross-validation process is conducted to evaluate the model's performance more meticulously and reduce the risk of overfitting. Using this method, the outcome of cross-validation consists of a series of performance evaluation scores for the model in each fold. The formula used to calculate the performance evaluation in each fold is provided below. For detailed results of cross-validation, please refer to the figure below.

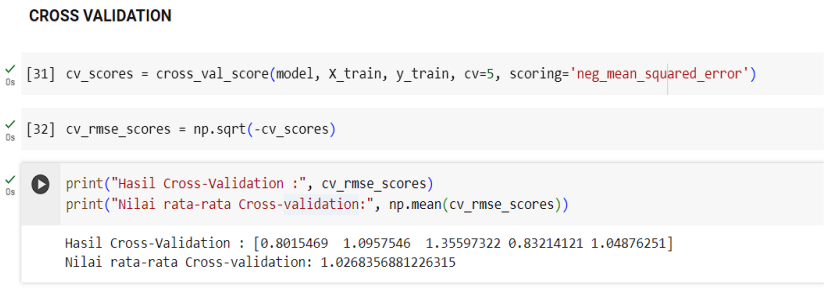

Figure 9. Cross Validation

The illustration above reveals the utilization of 5 folds, with each value in the cross-validation results denoting the model's performance evaluation on a specific fold. The observed values indicate variations in the model's performance across folds. However, the average model performance is computed as 1.0268356881226315. Consequently, the average cross-validation outcome offers an estimation of the model's overall performance by considering the performance variations across folds. This provides a broader insight into the model's predictive capability concerning unseen data.

Implementation Of Random Forest Regression

In evaluating the outcomes derived from this research, assessment will be conducted based on five criteria aligned with the grading scale depicted in the following table.

Very Good	> 0.5
Good	< 1.0
Adequate	< 1.5
Insufficient	1.5 to 2.0
Poor	> 2.0

Table 1. Assesment Criteria

After applying the Random Forest Regression algorithm, the accuracy of the model will be tested using Mean Square Error (MSE), Root Mean Square Error (RMSE), and Mean Absolute Error (MAE). The formulas are as follows:

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

Description :

Mean Square Error (MSE) is the average of the squared differences between the actual values (y_i) and the predicted values $(\hat{y}_i)$.

$$RMSE = \sqrt{MSE}$$

Description :

Root Mean Square Error (RMSE) is the square root of the Mean Square Error (MSE).

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |y_1 - \hat{y}_1|$$

Description :

Mean Absolute Error (MAE) is the average of the absolute differences between the actual values (y_i) and the predicted values $(\hat{y}_i)$.

Explanation of the formulas:

- y_i : Actual data (*true data*) for I =1,2,3,…,n
- $\hat{y}_i$: Predicted value for i = 1,2,3,…,n
- n : Number of data points.

And after applying the Random Forest Regression algorithm to the test data, which was previously trained, the following results were obtained.

Year	World Population	Growth Percentage	Summary of growth	Density
2024	8,045,311,447	0.88%	70,206,291	54
2025	8,118,835,999	0.91%	73,524,552	55

Table 2. Prediction Result

Mean Squared Error (MSE)	0.24420957741453037
Root Mean Squared Error (RMSE)	0.4941756544130137
Mean Absolute Error (MAE)	0.237832264957265
Score Model	0.3613816137857798

Table 3. Results score model

Based on the results obtained above, the population growth rate is predicted to increase by less than <1% in one year, indicating a relatively insignificant growth rate. Drawing conclusions from the achieved values for MSE, RMSE, MAE, and the model score, which are all less than 0.5, indicates excellent performance. This suggests that the model is capable of providing optimal results and can be expected to consistently deliver satisfactory outcomes when applied to different datasets in the future study.

CONCLUSION

The research findings demonstrate that the Random Forest Regression algorithm is highly effective in analyzing and predicting global population growth from time series data. The model achieved an evaluation score of 0.3613816137857798, which is notably below the 0.5 threshold. This low score indicates a high level of accuracy, signifying that the model's predictions are closely aligned with actual population trends.

The robustness of the Random Forest Regression algorithm in handling complex datasets, especially in a domain as intricate as global population dynamics, is evident from these results. The model's ability to consistently produce accurate predictions underscores its potential as a powerful tool for demographic forecasting. This high level of precision suggests that the algorithm is not only reliable but also adaptable to different datasets, making it a valuable asset for future research endeavors.

Moreover, the model’s performance in this study suggests it could be effectively applied to a wide range of demographic datasets, providing accurate predictions that could inform policy decisions and strategic planning. This potential for broad application further emphasizes the model's utility in the ongoing analysis and management of global population growth, contributing to more informed and sustainable development strategies.

REFERENCES

A. Lasarudin dan R. Maku, "PREDIKSI PERTUMBUHAN JUMLAH PENDUDUK MENGGUNAKAN ALGORITMA NEURAL NETWORK (NN)," 2021.

P. Kurniawan *dkk.*, "Prediksi Jumlah Penduduk Jakarta Selatan Menggunakan Metode Regresi Linear Berganda," *Jurnal Sistem dan Teknologi Informasi (JustIN)*, vol. 10, no. 4, hlm. 518, Des 2022.

R. Kurnia Armanda, "Prediksi Pertumbuhan Penduduk Kecamatan Cimaragas Kabupaten Ciamis Dengan Metode Artificial Neural Network," vol. 3, no. 2, hlm. 170–178, 2023.

A. Pratama Yudha, R. Puji Cahyono, S. Informasi Akuntansi, dan T. Komputer, "Analisis Kepuasan Pengunjung Menggunakan Metode Random Forest Untuk Wisata Pantai pada Pesawaran."

N. Nur, F. Wajidi, S. Sulfayanti, dan W. Wildayani, "Implementasi Algoritma Random Forest Regression untuk Memprediksi Hasil Panen Padi di Desa Minanga," *Jurnal Komputer Terapan*, vol. 9, no. 1, hlm. 58–64, Jun 2023.

L. Britanthia, C. Tanujaya, B. Susanto, dan A. Saragih, "Perbandingan Metode Regresi Logistik dan Random Forest untuk Klasifikasi Fitur Mode Audio Spotify," *Indonesian Journal of Data and Science (IJODAS)*, vol. 1, no. 3, hlm. 68–78, 2020.

F. Al Farikhi *dkk.*, "Perbandingan Algoritma Classification and Regression Tree (CART) dan Random Forest (RF) untuk Klasifikasi Penggunaan Lahan pada Google Earth Engine Informasi artikel A B S T R A K Sejarah artikel."

J Banjarnahor, F Sinaga, DS Sitorus, WAA Sitanggang, Application of Decision Tree Method in ECG Signal Classification For Heart Disorder Detection- Sinkron: jurnal dan penelitian teknik informatika, 2024.

W Setiawan, J Banjarnahor, MF Shandika, M Radhi, ANALYSIS OF CLASSIFICATION OF LUNG CANCER USING THE DECISION TREE CLASSIFIER METHOD, - Jurnal Sistem Informasi dan Ilmu Komputer Prima …, 2023

J Banjarnahor, F Zai, J Sirait, DW Nainggolan, Comparison Analysis of C4. 5 Algorithm and KNN Algorithm for Predicting Data of Non-Active Students at Prima Indonesia University- Sinkron: jurnal dan penelitian teknik informatika, 2023

SH Sinaga, AAM Duha, J Banjarnahor, Analisis Prediksi Deteksi Stroke Dengan Pendekatan Eda Dan Perbandingan Algoritma Machine Learning - Jurnal Ilmiah Betrik, 2023

A Tanzil, RA Barasa, Y Laia, J Banjarnahor, Analysis of Method C5. 0 in Triggering Factors The Number of Covid-19 Increases or Decreases After Getting the Vaccine, - Jurnal Sistem Informasi dan Ilmu Komputer Prima …, 2023

K. Ciptady, M. Harahap, J. Jonvin, Y. Ndruru, dan I. Ibadurrahman, "Prediksi Kualitas Kopi Dengan Algoritma Random Forest Melalui Pendekatan Data Science," *Data Sciences Indonesia (DSI)*, vol. 2, no. 1, Sep 2022.

"ANALISIS PERBANDINGAN ALGORITMA C4.5 DAN ALGORITMA KNN UNTUK MEMPREDIKSI DATA MAHASISWA NON AKTIF DI UNIVERSITAS PRIMA INDONESIA."

A. Rizal, D. C. R. Novitasari, dan Moh. Hafiyusholeh, "Pengelompokan Karyawan Berdasarkan Kesalehan Menggunakan Perbandingan Fuzzy C-Means, K-Means, dan Probabilistic Distance Clustering," *Jurnal Fourier*, vol. 11, no. 2, hlm. 69–77, Okt 2022.

P. Rosyani, A. Suhendi, D. H. Apriyanti, dan A. A. Waskita, "Color Features Based Flower Image Segmentation Using K-Means and Fuzzy C-Means," *Building of Informatics, Technology and Science (BITS)*, vol. 3, no. 3, hlm. 253–259, Des 2021.

L. Wulandari dan B. Olga Yogantara, "Algorithm Analysis of K-Means and Fuzzy C-Means for Clustering Countries Based on Economy and Health," vol. 15, no. 2, hlm. 1979–276, 2022, doi: 10.30998/faktorexacta.v15i2.12106.

Fadellia Azzahra, N. Suarna, dan Y. Arie Wijaya, "Penerapan Algoritma Random Forest Dan Cross Validation Untuk Prediksi Data Stunting," *Kopertip : Jurnal Ilmiah Manajemen Informatika dan Komputer*, vol. 8, no. 1, hlm. 1–6, Feb 2024.

A. Fauzan dan D. Ahmad, "ANALISIS HASIL PREDIKSI MAGNITUDO GEMPA DI WILAYAH KOTA PADANG MENGGUNAKAN TEKNIK RANDOM FOREST," *Jurnal Lebesgue : Jurnal Ilmiah Pendidikan Matematika, Matematika dan Statistika*, vol. 4, no. 3, hlm. 1569–1576, Des 2023.