p-ISSN 2338 – 3216
e-ISSN 2528 – 1070

JURNAL STATISTIKA
UNIVERSITAS MUHAMMADIYAH SEMARANG
12(1) 2024: 34-47

THE PERFORMANCE ANALYSIS OF THE BEST MACHINE LEARNING MODEL FOR SULFUR DIOXIDE IN DKI JAKARTA

Panji Kuswanaji[1*], Rizky Addrian Aliyafi[2], Agung Hari Saputra[3]
[1 2 3] Undergraduate Applied Program of Meteorology, State College of Meteorology Climatology and Geophysics, Indonesia

***e-mail:** *pkuswanaji@gmail.com*

Article Info:
Received: January 17, 2024
Accepted: May 30, 2024
Available Online: July 29, 2024

Keywords:
Air Quality; Machine Learning; Model

Abstract: A good clean air is one of crucial things for humans health. A place with good and clean air can prevent humans from various kinds of respiratory diseases. One of the factors that can influence the cleanliness of the air in an area is the composition of Sulfur Dioxide (SO2). This study aims to examine sulfur dioxide (SO2) levels in Jakarta spanning eleven years, with the goal of determining the most accurate predictive model for SO2 concentrations., which is critical for public health and environmental management. The study incorporates quantitative methods, machine learning techniques, and statistical analysis. From this research there are three best models that has top performance, these are huber, exponential smoothing, and naive forecaster. The result shows that naive model has the best performance with MASE of 0.3864, RMSSE of 0.3098, MAE of 2.8857, RMSE of 3.7735, MAPE of 0.0593, and SMAPE of 0.0623.

1. INTRODUCTION

A good clean air is one of crucial things for humans health. A place with good and clean air can prevent humans from various kinds of respiratory diseases. One of the factors that can influence the cleanliness of the air in an area is the level of contaminants or pollutants in an area [1]. Sulfur dioxide (SO2) is an air pollutant that causes coughing and shortness of breath. The main sources of SO2 in the air come from the combustion process (coal or diesel), the metallurgical industry, and the sulfuric acid industry [2]. SO2 gas is also the cause of acid rain and chemical photo fog which disrupts human life [3].

SO2 is a pollutant that generally comes from smoke or air waste that appears as a result of burning fuel [4]. Besides that, SO2 is also possible to come from smoke from large industries that have massive combustion processes and natural processes like volcanic eruption [4]. The smoke from the eruption contains SO2 and SO4 which can affect air quality and can even influence climate change [4].

In the realm of environmental monitoring and management, accurately forecasting air quality parameters is crucial for safeguarding public health and ensuring sustainable urban development. Sulfur dioxide (SO2), a significant pollutant with potential adverse effects on respiratory health and the environment, requires meticulous analysis, particularly in the dynamic urban landscape of DKI Jakarta [5]. Here, industrialization and urbanization are on

[DOI: 10.14710/JSUNIMUS.12.1.2024.34-47]

the rise, making it imperative to comprehend and predict SO2 levels for effective pollution control and mitigation strategies.

The focus in this study is to find the best performance machine learning model to predict SO2 levels in DKI Jakarta. In order to get that, this study compares various machine learning models such as naive forecaster [6], exponential smoothing [7], huber [8], extreme gradient boosting [9], and other PyCaret available models. The reason these models are used is because these models can forecast data based on previous data and can be trained to predict the pattern of future data. The best performance model in this study is the model which has the least error value (MASE, RMSSE, MAE, RMSE, MAPE, and SMAPE).

2. LITERATURE REVIEW

2.1. Air Quality

Air quality refers to atmospheric conditions that include the composition of gases, particles and other chemicals in the air that can impact ecosystems and climate [10]. Air quality also has a big crucial role for the health quality in humans. This is because air quality can affect their respiratory, which is one of the most crucial system in humans body [11]. Some of the basic principles that shape our understanding of air quality include knowledge of air components, sources of air pollution, and the consequences of their impact on human health and the environment.

Air components consists of various components, such as nitrogen (N2), oxygen (O2), argon (Ar), and other gases. Vapor water (H2O) is also a variable part of the air composition that changes according to weather conditions. Apart from that, there are also air pollutants such as sulfur dioxide (SO2), nitrogen dioxide (NO2), ozone (O3), carbon monoxide (CO), and solid particles (PM10 and PM2.5) which can affect air quality [12].

Sources of air pollution comes from two main sources: natural and anthropogenic sources. Natural sources include volcanic eruptions, wind dust, and natural bacteria. Meanwhile, anthropogenic sources are related to human activities, such as vehicle emissions, industry, burning of fossil fuels and domestic waste. Understanding these two sources is important for developing air pollution control strategies [13].

Air pollution has a significant impact on human health. Gases such as SO2 and NO2 can cause respiratory tract irritation, while PM can enter the lungs and cause respiratory problems [14]. Some air pollutants have also been linked to long-term illnesses such as heart disease and chronic respiratory disorders. These health risks are the basis for setting air quality standards by regulatory agencies.

Air pollution also has an impact on the environment. Acid rain, caused by SO2 and NO2 emissions, can damage soil and water ecosystems [15]. Ozone in the atmosphere can affect plant growth and cause damage to vegetation. Understanding these impacts is important to maintain biodiversity and ecosystem balance.

2.2. Machine Learning

Machine Learning is an analytical tool used to predict future values based on historical information and relevant factors [16]. Machine Learning plays a pivotal role across diverse fields, aiding decision-makers in anticipating future trends, making informed choices, and planning strategies [17]. This literature review provides a comprehensive overview of machine

[DOI: 10.14710/JSUNIMUS.12.1.2024.34-47]

learning, delving into their methodologies, applications, and the evolving landscape of predictive analytics.

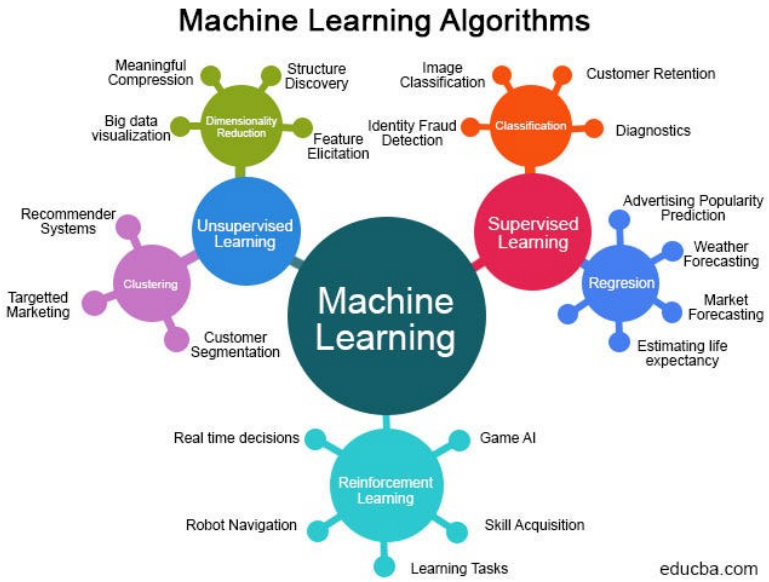

Fig. 1 Machine Learning Algorithms

Figure 1 is the machine learning algorithms. The use of machine learning methods, such as non-linear regression, support vector machines, and neural networks, has become increasingly common in the development of forecasting models [18]. The advantage of machine learning lies in its ability to handle complex and non-linear patterns in data, increasing the accuracy of predictions [19].

Through the application of this basic theory, machine learning developers can select and design models that suit the desired data characteristics and forecasting objectives [20]. Integrating diverse approaches and in-depth understanding of the dynamics that influence machine learning data is the key to success in using machine learning in various application fields.

2.3. Model

Model is a crucial aspect of decision-making across various industries, including finance, economics, supply chain management, and weather prediction [21]. Over the years, researchers and practitioners have developed numerous forecasting models to improve the accuracy of predictions. This literature review provides an overview of the key models and discusses their strengths, weaknesses, and recent developments.

In the realm of forecasting models, there's a continuous evolution as new techniques and hybrid approaches emerge. Time series models like Autoregressive Integrated Moving Average (ARIMA) and Exponential Smoothing (ETS) remain fundamental, but machine learning and deep learning models are gaining prominence, especially in handling complex and high-dimensional data [22]. Hybrid models that amalgamate the strengths of multiple approaches are growing in popularity. The choice of a model hinges on the specific problem, available data, and the desired level of accuracy and interpretability [23]. As technology and data availability continue to advance, the forecasting field will likely witness further

[DOI: 10.14710/JSUNIMUS.12.1.2024.34-47]

innovations and refinements, ultimately enhancing the accuracy and reliability of future predictions.

3. METHODOLOGY

This research uses a quantitative descriptive method where the data in the research can be directly analyzed statistically to draw conclusions .The analysis in this research uses a Python-based library program designed to simplify the machine learning model development process by providing a simple and automated interface for common tasks such as data exploration, feature selection, model selection, and parameter tuning. In this research PyCaret is used for a program in Python that can perform an Automated Model Selection. This can be used by using the compare_models() function.

3.1. Data Processing

The first data processing that needs to be done after downloading the datasets from Open Data Jakarta website (https://data.jakarta.go.id/) is to create a new excel file containing one parameter of the ISPU data you want to research and the time. The dataset is then combined into one sheet with row ('Date', 'SO2'). 'Date' is a variable that indicates time, namely daily and 'so2' is the value or index of SO2 levels. This processing aims to enable PyCaret to read the dataset. Table 1 is an example of the dataset.

Table 1. Example of Sulfur Dioxide Dataset for PyCaret

Date	SO2
01/01/2010	2
02/01/2010	3
03/01/2010	4

3.2. Initial Setup

After processing the downloaded data, the next step is to make an initial setup for configuring the parameter of the model. Configuring parameters as the default rules to be used in experiments is done with the setup function, which will serve as the foundation for data processing. This ensures that the data to be processed can produce accurate model results. Table 2 is the parameter initial setup of the model configuration.

Table 2. Initial Parameter Setup for Sulfur Dioxide Dataset

Description	Value	Description	Value
session_id	123	Fold Generator	SlidingWindowSplitter
Target	so2	Fold Number	5
Approach	Univariate	Enforce Prediction Interval	FALSE
Exogenous Variables	Present	Splits used forhyperparameters	all

[DOI: 10.14710/JSUNIMUS.12.1.2024.34-47]

3.2. Dataset Analysis

The next step after configuring the intial setup to run the model is to analyze the dataset model. This analysis can be done by plotting the clasical decomposition of the dataset, the diagnostics plots of the dataset, the difference plots of the dataset. The first dataset analysis is the classical decomposition of the dataset, this research uses seasonal period of 48. In this step, there are four plots that appears the actual graph of dataset, the seasonal graph, trend graph, and residual graph. The actual graph is a graph that contains the timeseries data. The sasonal graph is a graphical representation of data that displays patterns or trends that repeat at regular intervals over time [24]. This type of graph is particularly useful for analyzing and visualizing the seasonal variation in a set of data points, such as monthly sales figures, temperature fluctuations, or other periodic patterns. A trend graph, also known as a trend chart or time series plot, is a graphical representation of data points in a time-ordered sequence [25]. The primary purpose of a trend graph is to visualize the underlying trend or pattern in the data over time. A residual graph, in the context of statistical analysis or modeling, is a graphical representation of the residuals or errors derived from a regression analysis [26]. Residuals are the differences between the observed values and the predicted values from a regression model.From all the graphs that has been plotted, we can analyze the dataset. If the dataset still has oddity like outliers, the dataset needs to normalize first.

Besides the classical decomposition of the dataset, dataset analysis also can be done by analyzing the diagnostics plots and the difference plots of the dataset. Diagnostics plots, and difference plots are graphical tools used in statistical analysis and regression modeling to assess the assumptions and performance of a statistical model. These plots help analysts and researchers identify potential issues with their models, such as violations of assumptions, outliers, or patterns that may indicate the need for model refinement. In these plots, there are six graphs. These graphs are the actual graph, ACF graph, PACF graph, Q-Q plots graph, and periodogram graph. ACF graph (Autocorrelation Function graph) is a graphical representation of the autocorrelation of a time series [27]. Autocorrelation measures the correlation between a time series and its own lagged values. In other words, it quantifies the degree of similarity between a data point and the data points that precede or follow it in time. PACF graph (Partial Autocorrelation Function graph) is a graphical representation of the partial autocorrelation of a time series [27]. The partial autocorrelation at lag k measures the correlation between the series and its lagged values while controlling for the influence of the intermediate lags (1 to k-1). A Q-Q (Quantile-Quantile) plot is a graphical tool used to assess whether a set of data follows a specified theoretical distribution, often the normal distribution [28]. It compares the quantiles of the observed data against the quantiles of the expected distribution, typically a normal distribution. A periodogram is a graphical representation of the spectral density of a time series [29]. It is a tool commonly used in signal processing and time series analysis to identify periodic patterns or frequencies within a set of data. With all of these graphs, the dataset can be analyzed in order to check the quality of the dataset.

3.2. Model Comparison

Comparing the model in this research begins by using the compare_models() function in the PyCaret to find the Mean Absolute Scaled Error (MASE), Root Mean Square Scaled Error (RMSSE), Mean Absolute Error (MAE), Root Mean Square Scaled Error (RMSE), Mean Absolute Percentage Error (MAPE), and Symmetric Mean Absolute Percentage Error (SMAPE) of 26 machine learning models in PyCaret. The purpose of this comparison is to compare the predicted results of the 26 available models in PyCaret to find the best model. This

[DOI: 10.14710/JSUNIMUS.12.1.2024.34-47]

test was carried out using MASE, RMSSE, MAE, RMSE, MAPE, and SMAPE accuracy tests to see the accuracy of predictions of the models.

MAE serves as a metric for assessing machine learning models [30]. It is computed by taking the average of the absolute differences between the actual and predicted values within the model. In mathematical terms, MAE can be represented as formula 1.

$$MAE = \frac{\sum |yi - yi'|}{n} \quad (1)$$

MASE is an evaluation metric used to measure the accuracy of a forecasting model [30]. This metric is calculated by comparing the MAE (Mean Absolute Error) of the forecasting model in use with the MAE of a random walk forecasting model. A smaller MASE value indicates a more accurate forecasting model. In mathematical terms, it can be defined as formula 2.

$$MASE = \frac{(MAE\ forecasting\ model)}{(MAE\ random\ model)} \quad (2)$$

MAPE (Mean Absolute Percentage Error) is an evaluation metric used to measure the accuracy of a forecasting or prediction model in terms of a percentage [22]. MAPE is employed to gauge how closely a forecasting model approximates the actual values in percentage terms [17]. It quantifies the average percentage error of predictions relative to the actual values. In this context, the smaller the MAPE value, the better the performance of the model in forecasting future values. MAPE can be defined as formula 3.

$$MAPE = \frac{100 * \sum \frac{|yi - yi'|}{yi}}{n} \quad (3)$$

In addition, SMAPE (Symmetric Mean Absolute Percentage Error) is also used to evaluate the accuracy of a model in forecasting time series values [31]. SMAPE measures the average percentage error of predictions symmetrically between the predicted and actual values. This metric is useful in assessing the model's accuracy over a specific time range. A smaller SMAPE value indicates that the forecasting model performs better in predicting time series values. The SMAPE formula involves the mean absolute value of the relative differences between actual and predicted values, and the result is expressed as a percentage. SMAPE can be represented as formula 4.

$$SMAPE = \frac{200}{n} * \sum \left| \frac{yi - yi'}{\frac{|yi| + |yi'|}{2}} \right| \quad (4)$$

Root Mean Square Error (RMSE) is a model evaluation technique in machine learning [32]. It calculates the average difference between predicted and actual values. The RMSE formula is defined as formula 5.

$$RMSE = \sqrt{((\frac{1}{n}) * \Sigma(yi - yi')^2)} \quad (5)$$

RMSSE, which stands for Root Mean Square Scaled Error, is a model evaluation method in machine learning [30]. RMSSE measures the average difference between predicted values and actual values and then takes the square root of the result. The RMSSE formula is defined as formula 6.

$$RMSSE = \sqrt{((1/n) * \Sigma((yi - yi') / yi)^2)} \quad (6)$$

[DOI: 10.14710/JSUNIMUS.12.1.2024.34-47]

where yi = actual value,

yi' = predicted value by model,

n = total of data

The importance of using various evaluation metrics is to provide a comprehensive picture of how well a forecasting model can predict future values, taking into account aspects such as scale, percentage errors, and symmetry in the comparison between predicted and actual values.

4. RESULTS AND DISCUSSION

4.1. Classical Decomposition of SO2

With PyCaret, the classical decomposition of SO2 can be identified. The classical decomposition of SO2 that can be plotted in PyCaret includes the graph of the actual data of SO2, the seasonal data of SO2, the trend, and the residual data of SO2. The plot of this feature can be seen below.

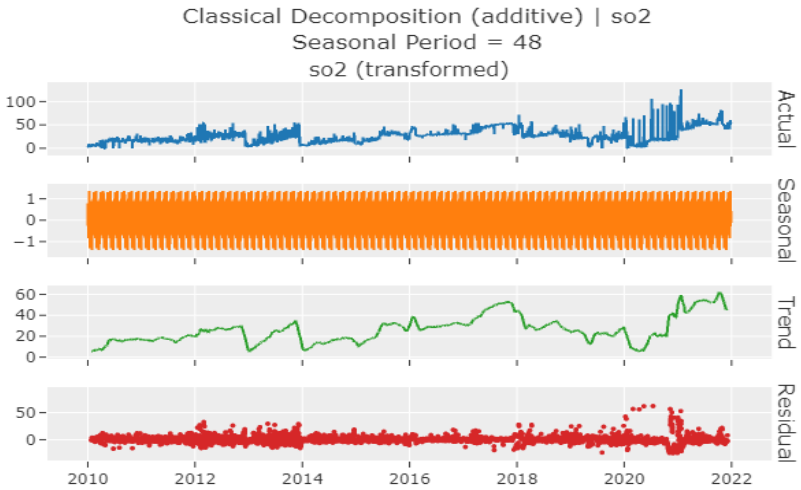

Fig. 2 Classical Decomposition (additive) of SO2

Figure 2 is a detailed time series decomposition chart with four distinct panels, each showing a specific aspect of the transformed data labeled as "SO2 (transformed)" with a seasonal pattern repeating every 48 units. In the top panel, you can observe the actual data points for the variable "SO2" over time, represented by the blue line. This depiction showcases the natural fluctuations in the data, marked by both upward and downward movements, with some notably high spikes that stand out as outliers.

The second panel, colored in orange, illustrates the recurring seasonal pattern inherent in the data. This pattern exhibits a high degree of regularity and follows a defined cycle, suggesting a robust and predictable seasonal effect influencing the variable. Moving down to the third panel, you'll find the green representation of the overall trend or direction of the data throughout the observed time period. It's quite evident from this panel that there is a discernible upward trend, implying that the "SO2" variable has been steadily increasing over time.

Finally, the bottom panel, distinguished by its red color, provides insight into the residuals – these are the disparities between the actual data points and what would be predicted based on the seasonal and trend components. The scattered dots around the zero line in this

[DOI: 10.14710/JSUNIMUS.12.1.2024.34-47]

panel suggest that the forecasting model has effectively captured the majority of the data's underlying behavior. What remains in the form of these scattered dots is mostly attributed to random noise, indicating that it is not accounted for by the model's components and is essentially unpredictable. This type of chart is commonly used to break down time series data into its main components to better understand the underlying patterns and behaviors. It can help in predicting future values by analyzing these components separately.

4.2. Diagnostics Plot for SO2

Time series diagnostics plots are a collection of visualizations used to examine whether a time series model fits the data well and to identify characteristics of the data that may not have been captured by the model. These plots are critical for identifying issues with the model, such as remaining trends or correlations, and for verifying the statistical assumptions underlying the model. With PyCaret we can plot this diagnostic plots to better understand the correlation between the data and the model. These are the diagnostics plot for DKI Jakarta SO2 Index in 2010-2021.

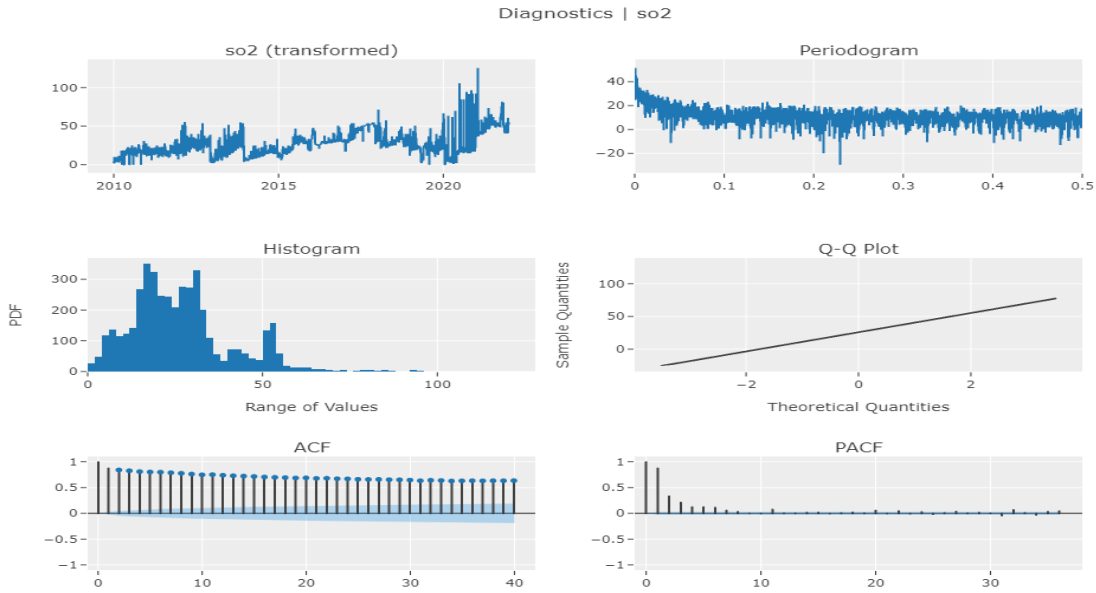

Fig. 3 Diagnostics plot of so2

Figure 3 is a collection of diagnostic plots for a transformed time series, labeled "so2 (transformed)". Each plot reveals specific statistical aspects of the data. The first plot, which focuses on the variable "so2," provides a visual representation of the transformed time series data spanning from 2010 to approximately 2020. This plot reveals fluctuations in the values of "so2" over this time period, with a notable trend of increasing values towards the end of the observed period. It suggests that there may be a significant temporal evolution or trend in the "so2" variable during this timeframe.

Moving on to the second plot, we delve into the examination of potential consistent cycles or frequencies within the time series. However, upon inspecting the periodogram, we notice the absence of prominent peaks. This lack of clear peaks indicates that there are no distinct cycles or recognizable periodic patterns that stand out in the data, suggesting a more aperiodic nature. The third plot takes the form of a histogram, illustrating the distribution of data values and their corresponding frequencies. A closer look at this histogram reveals that the majority of the data points cluster around lower values, while there are some extreme higher

[DOI: 10.14710/JSUNIMUS.12.1.2024.34-47]

values. This skewed distribution suggests that the data is concentrated towards the lower end of the scale, with occasional occurrences of higher values.

The Q-Q (Quantile-Quantile) plot, in the fourth position, serves to compare the actual distribution of the data with an idealized normal distribution. In an ideal normal distribution, data points should closely follow a straight reference line. This particular Q-Q plot indicates a tendency towards normality, with most points following the reference line. However, some deviations are noticeable, particularly at the extreme values, indicating some departure from a perfectly normal distribution. The fifth plot, the ACF (Autocorrelation Function), provides insights into how the values within the time series correlate with each other over time. This plot reveals a consistent correlation at early time intervals, gradually decreasing as time progresses but remaining statistically significant. This suggests the presence of temporal dependency or autocorrelation within the time series data.

Lastly, the sixth plot, the PACF (Partial Autocorrelation Function), is another measure of correlation within the time series, but it accounts for correlations at earlier time intervals. Within this plot, we observe some prominent partial correlations at early lags, indicating strong correlations at specific time points. However, other lags show no significant partial correlation, signifying more isolated or uncorrelated periods in the time series. Together, these charts provide a broad overview of the statistical characteristics of the transformed time series, which is invaluable for determining the appropriate time series analysis model and for model validation purposes.

4.3. Difference Plots for SO2

Difference plots are a type of diagnostic tool used in time series analysis. They are used to visualize the effect of differencing, which is a method for making a time series stationary. Stationarity is a property of a time series that means its statistical properties, like mean and variance, do not change over time, which is often a necessary condition for many time series modeling techniques. These are the difference plots for DKI Jakarta SO2 in 2010-2021.

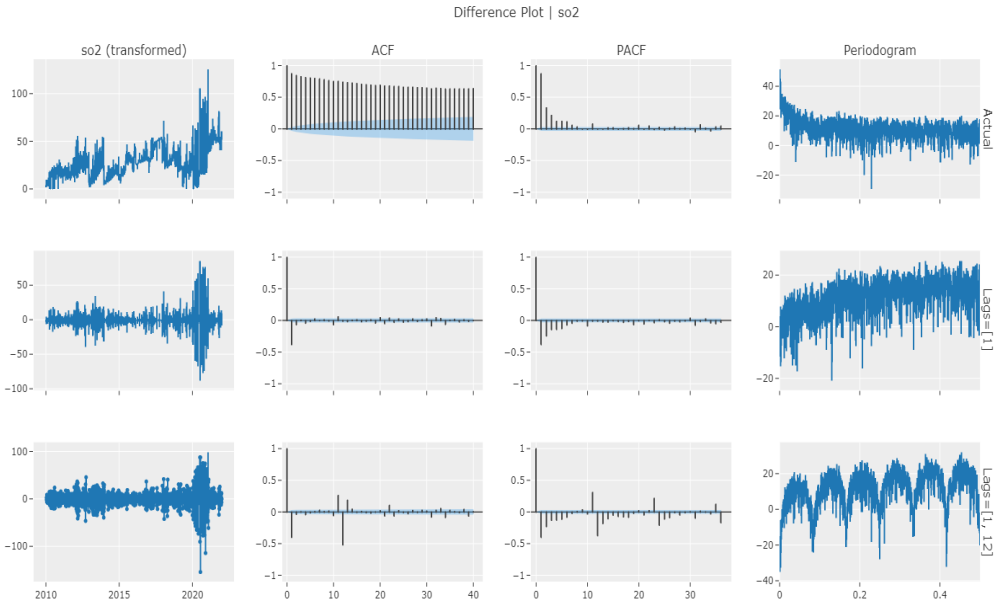

Fig. 4 Difference plots for SO2

[DOI: 10.14710/JSUNIMUS.12.1.2024.34-47]

Figure 4 is a set of time series diagnostic plots or difference plot. In this section, the transformed "SO2" variable through three different time series plots. The first plot showcases the original data over time, revealing an evident upward trend. The second plot unveils the data after applying a first difference, effectively removing the trend and exposing fluctuations around a constant mean. The third plot appears to represent the data after a second difference or another filtering technique, resulting in data that exhibits characteristics of stationarity with a consistent variance across time.

Within this section, we encounter three ACF (Autocorrelation Function) plots, each corresponding to one of the three transformed "SO2" time series. The ACF plot for the original time series illustrates high initial correlations that do not quickly dissipate, indicating a degree of temporal dependence within the data. The second and third ACF plots demonstrate a significant reduction in temporal correlation, particularly in the third plot where nearly all correlations fall within the confidence bounds. This suggests that the data may have achieved stationarity, with reduced autocorrelation.

Much like the ACF, this section presents three PACF (Partial Autocorrelation Function) plots. The first PACF plot shows several lags with significant partial correlations. After differencing (as seen in the second and third plots), nearly all lags fall below the confidence bounds. This indicates that the data has been effectively decomposed, successfully removing autocorrelation dependence.

In the last section, we examine three distinct periodograms, each corresponding to the original time series, the time series after the first difference, and the time series after the second difference. These periodograms provide insight into the strength of frequencies within the data over time. The first plot does not exhibit clear frequency peaks, suggesting the potential absence of strong cyclic components. However, the other two periodograms, which pertain to the differenced data, display more pronounced variations in frequency strength. This may indicate the presence of certain patterns or cycles in the transformed data, which were not as evident in the original series.

Overall, these plots are used to evaluate whether the data transformation and differencing processes have successfully rendered the data stationary, which is a critical assumption in many time series analysis models. This assessment is crucial to ensure that any models to be constructed will fit the data well and that the resulting predictions will be accurate.

4.4. The Best Model

SO2 index that has downloaded will be merged into one timeline with Microsoft Excel and then processed with PyCaret. In this process the SO2 index will be runned and forecasted by various models in PyCaret. The objective of this process is to know how is the performance of these models in Jakarta SO2 index in 2010-2021 by calculating the MASE, RMSSE, MAE, RMSE, MAPE, SMAPE, and R2 and rank them. We can get the output in the form of table. And the best model for Jakarta SO2 Index 2010-2021 is the one in the top of the table, Naive Forecaster model. Table 3 and 4 are the PyCaret model comparison and the naive model performance.

[DOI: 10.14710/JSUNIMUS.12.1.2024.34-47]

Table 3. PyCaret Model Comparison for The Sulpur Dioxide Compositions

Model	MASE	RMSSE	MAE	RMSE	MAPE	SMAPE
Naive Forecaster	0,3864	0,3098	2,8857	3,7735	0,0593	0,0623
Exponential Smoothing	0,4171	0,3341	3,1109	4,0660	0,0648	0,0676
Huber	0,4253	0,3414	3,1704	4,1542	0,0673	0,0685
Extreme Gradient Boosting	0,4273	0,3346	3,1859	4,0691	0,0685	0,0691
ETS	0,4284	0,3371	3,1944	4,1029	0,067	0,0695
Theta Forecaster	0,4294	0,3215	3,2022	3,9129	0,0674	0,0698
Gradient Boosting	0,446	0,3345	3,3232	4,0680	0,0718	0,0718
Light Gradient Boosting	0,4649	0,3424	3,4647	4,1638	0,0751	0,0754
Lasso	0,5364	0,3939	3,9948	4,7894	0,087	0,0862
Regressor	0,5364	0,3939	3,9942	4,7889	0,087	0,0862
Random Forest	0,5472	0,3979	4,0714	4,8344	0,0896	0,0879
Extra Trees	0,5504	0,4035	4,0930	4,9010	0,0904	0,0881
Elastic Net	0,5592	0,4023	4,1641	4,8916	0,0909	0,0898
Linear	0,5826	0,4191	4,3379	5,0958	0,0947	0,0934
Ridge	0,5826	0,4191	4,3380	5,0958	0,0947	0,0934
Bayesian Ridge	0,5861	0,4195	4,3638	5,1007	0,0953	0,094
Orthogonal Matching Pursuit	0,5898	0,4216	4,3918	5,1253	0,0961	0,0947
Croston	0,6253	0,4377	4,6464	5,3146	0,1033	0,1003
Decision Tree	0,7993	0,5848	5,9492	7,1062	0,1313	0,1275
Polynomial Trend Forecaster	0,8479	0,5718	6,3206	6,9562	0,1358	0,1489
K Neighbors	0,8980	0,6349	6,6715	7,7076	0,1492	0,1387
AdaBoost	1,3311	0,8507	9,8953	10,3265	0,229	0,2026
STLF	1,6829	1,1223	12,5045	13,6247	0,282	0,2627
ARIMA	2,1101	1,4970	15,6827	18,1732	0,3613	0,2904
Grand Means Forecaster	2,4563	1,5287	18,2894	18,5782	0,4074	0,514
Seasonal Naive Forecaster	2,7665	1,8425	20,5714	22,3764	0,4763	0,3717

Table 4. Naive Model Performance for The Sulfur Dioxide Compositions

cutoff		MASE	RMSSE	MAE	RMSE	MAPE	SMAPE
0	19/11/2021	0,0968	0,0828	0,7143	1	0,0168	0,0171
1	26/11/2021	0,1542	0,1168	1,1429	1,4142	0,0267	0,0273
2	03/12/2021	0,4799	0,4164	3,5714	5,0568	0,0718	0,0769
3	10/12/2021	0,4015	0,2614	3	3,1848	0,0688	0,0662
4	17/12/2021	0,7993	0,6716	6	8,2115	0,1124	0,1242
Mean		0,3864	0,3098	2,8857	3,7735	0,0593	0,0623
SD		0,2519	0,216	1,8939	2,6438	0,0344	0,0383

5. CONCLUSION

Based on this research, the conclusion is the best compatible model for DKI Jakarta SO2 index in 2010-2021 is the naive model. In this research, There are six factors that can influence the performance of the model. These are RMSE, MAE, MASE, RMSSE, MAPE, and SMAPE.

[DOI: 10.14710/JSUNIMUS.12.1.2024.34-47]

The best three models with RMSE value are naive forecaster (3.7735), exponential smoothing (4.0660), and gradient boosting (4.0680). Next is MAE, the best three models with MAE value are naive forecaster (2.8857), exponential smoothing (3.1109), and huber (3.1704). After that is MASE, the best three models with MASE value are naive forecaster (0.3864), exponential smoothing (0.4171), and huber (0.4253). An then, the best three models with RMSSE value are naive forecaster (0.3098), theta forecaster (0.3215), and gradient boosting (0.3345). And the best three models with MAPE value are naive forecaster (0.0593), exponential smoothing (0.0648), and ETS (0.067). Lastly, The best three models with SMAPE value are naive forecaster (0,0623), exponential smoothing (0.0676), and huber (0,0685). The result shows that naive forcaster model has the best performance for those parameters with MASE of 0.3864, RMSSE of 0,3098, MAE of 2.8857, RMSE of 3.7735, MAPE of 0.0593, and SMAPE of 0.0623. The recommendations that can be given for the next research is to ensemble some of the machine learning models that has been compared to make a new model that has better performance than naive model.

REFERENCES

[1] R. Arissa and A. A. Kiswandono, "Kajian Indeks Standar Polusi Udara (ISPU) Pm10, So2, O3, Dan No2 Di Kota Bandar Lampung," *Anal. Anal. Environ. Chem.*, vol. 2, no. 2, pp. 38–46, 2017.

[2] E. E. Saragih, D. R. Jati, and S. Pramadita, "Analisis Polutan Udara (CO, NO2, SO2, PM10, PM2,5 dan TSP) di Industri Galangan Kapal serta Pengaruhnya terhadap Lingkungan Kerja," *J. Teknol. Lingkung. Lahan Basah*, vol. 10, no. 2, p. 129, 2022, doi: 10.26418/jtllb.v10i2.56051.

[3] G. M. Tampa, S. S. Maddusa, and O. R. Pinontoan, "Analisis Kadar Sulfur Dioksida (SO2) Udara di Terminal Malalayang Kota Manado Tahun 2019," *Journalof Public Heal. Community Med.*, vol. 1, no. 3, pp. 87–92, 2020.

[4] Q. Mutawakkilah, I. Trina, R. Rinawati, and A. A. Kiswandono, "Analisis Sulfur Dioksida (So2) Sebagai Polutan Udara Di Kabupaten Lampung Barat Dan Way Kanan Pada Tahun 2019-2020," *Anal. Anal. Environ. Chem.*, vol. 8, no. 1, p. 11, 2023, doi: 10.23960/aec.v8i1.2023.p11-20.

[5] M. Faisal *et al.*, "Prosiding Forum Ilmiah Tahunan IAKMI (Ikatan Ahli Kesehatan Masyarakat Indonesia) ANALISIS KUALITAS UDARA BERBASIS INDEKS STANDAR PENCEMARAN UDARA (ISPU) DI PELABUHAN BONGKAR MUAT BATU BARA CIREBON TAHUN 2022," *Pros. Forum Ilm. Tah. IAKMI* , pp. 1–12, 2022.

[6] J. A. Brandt and D. A. Bessler, "Price Forecasting and Evaluation: An Application in Agriculture*," 1983.

[7] E. S. Gardner, "Exponential Smoothing: The State of the Art," vol. 4, no. August 1984, pp. 1–28, 1985.

[8] J. Fan, Y. Guo, and B. Jiang, "Adaptive Huber regression on Markov-dependent data," *Stoch. Process. their Appl.*, vol. 150, pp. 802–818, 2022, doi: 10.1016/j.spa.2019.09.004.

[9] P. Carmona, F. Climent, and A. Momparler, "Predicting failure in the U.S. banking sector: An extreme gradient boosting approach," *Int. Rev. Econ. Financ.*, vol. 61, pp. 304–323, 2019, doi: 10.1016/j.iref.2018.03.008.

[10] Wiharja, "Identifikasi Kualitas Gas So 2," *J. Teknol. Lingkung.*, vol. 3, no. 3, pp. 251–255, 2002, [Online]. Available: http://download.portalgaruda.org/article.php?article=62168&val=4561&title=

[DOI: 10.14710/JSUNIMUS.12.1.2024.34-47]

[11] T. Budiwati, "Analisis Hujan Asam Dan Co2 Atmosfer," *Pros. Semin. Nas. Penelitian, Pendidik. dan Penerapan MIPA, Fak. MIPA, Univ. Negeri Yogyakarta, 16 Mei 2009*, pp. 276–281, 2009.

[12] A. Augista Firstanti, A. Erlan Afiuddin, T. Azis Ramadani, P. Studi Teknik Pengolahan Limbah, J. Teknik Permesinan Kapal, and P. Perkapalan Negeri Surabaya, "Pola Sebaran Emisi SO2, NO2, dan Partikulat dari Cerobong Batu Bara Industri Kecap Menggunakan Software Screen View," vol. 5, no. 1, pp. 78–82, 2022.

[13] N. Aini, R. Ruktiari, M. R. Pratama, and A. F. Buana, "Sistem Prediksi Tingkat Pencemaran Polusi Udara dengan Algoritma Naïve Bayes di Kota Makassar," *Pros. Semin. Nas. Komun. dan Inform.*, vol. 3, pp. 83–90, 2019.

[14] F. Al Farisi, B. Budiyono, and O. Setiani, "Pengaruh Sulfur Dioksida (SO2) Pada Udara Ambien Terhadap Risiko Kejadian Pneumonia Pada Balita," *J. Kesehat. Masy.*, vol. 6, no. 4, pp. 438–446, 2018, [Online]. Available: https://ejournal3.undip.ac.id/index.php/jkm/article/view/21452

[15] R. E. Handriyono and A. W. S. Dewi, "Studi Kandungan Asam Pada Air Hujan Di Kawasan," *Tek. Lingkung.*, vol. 1, no. 2, pp. 52–55, 2018, [Online]. Available: http://jurnal.universitaskebangsaan.ac.id/index.php/envirosan/article/view/144/115

[16] A. H. Saputra, I. M. A. Satya, F. P. Sari, and A. Mulya, "Spatial rain probabilistic prediction performance using costsensitive learning algorithm," *E3S Web Conf.*, vol. 464, p. 19001, 2023, doi: 10.1051/e3sconf/202346419001.

[17] S. M. Robial, "Perbandingan Model Statistik pada Analisis Metode Peramalan Time Series (Studi Kasus: PT. Telekomunikasi Indonesia, Tbk Kandatel Sukabumi)," *J. Ilm. SANTIKA*, vol. 8, no. 2, pp. 1–17, 2018.

[18] G. I. Fajarini1 and I. Purnamasari2, "Prediksi Data Curah Hujan Dengan Menggunakan Statistika Non Parametrik Rainfall Data Prediction Using Non-Parametric Statistic," *Eksponensial*, vol. 32, no. 9, pp. 1805–1808, 1977.

[19] S. Nakagawa and H. Schielzeth, "A general and simple method for obtaining R2 from generalized linear mixed-effects models," *Methods Ecol. Evol.*, vol. 4, no. 2, pp. 133–142, 2013, doi: 10.1111/j.2041-210x.2012.00261.x.

[20] A. K. MS, M. A. Sasmita, and A. H. Saputra, "Prediksi Particulate Matter (PM 2.5) di DKI Jakarta Menggunakan XGBoost," *J. Apl. Meteorol.*, vol. 2, no. 1, pp. 1–9, 2023, doi: 10.36754/jam.v2i1.355.

[21] F. Insani and S. I. Darlianti, "Pembentukan Model Regresi Linier Menggunakan Algoritma Genetika untuk Prediksi Parameter Indeks Standar Pencemar Udara (ISPU)," *J. CoreIT J. Has. Penelit. ...*, vol. 5, no. 2, pp. 110–117, 2019, [Online]. Available: http://ejournal.uin-suska.ac.id/index.php/coreit/article/view/9157

[22] N. H. Latief, N. Nur'eni, and I. Setiawan, "Peramalan Curah Hujan di Kota Makassar dengan Menggunakan Metode SARIMAX," *Stat. J. Theor. Stat. Its Appl.*, vol. 22, no. 1, pp. 55–63, 2022, doi: 10.29313/statistika.v22i1.990.

[23] P. Rani, "PyCaret based URL Detection of Phishing Websites," *Turkish J. Comput. Math. Educ.*, vol. 11, no. 1, pp. 908–915, 2020, doi: 10.17762/turcomat.v11i1.13589.

[24] T. Kämpke, "Efficient versioning for matrix structures," *Comput. Math. with Appl.*, vol. 27, no. 3, pp. 9–19, 1994, doi: 10.1016/0898-1221(94)90041-8.

[25] L. Gabet, "The decomposition method and distributions," *Comput. Math. with Appl.*, vol. 27, no. 3, pp. 41–49, 1994, doi: 10.1016/0898-1221(94)90045-0.

[26] S. L. Smith, "Three variables.," *Mil. Med.*, vol. 158, no. 12, pp. 1–7, 1993, doi: 10.1093/milmed/158.12.a8.

[27] T. Park, S. G. Yi, S. Lee, and J. K. Lee, "Diagnostic plots for detecting outlying slides in a cDNA microarray experiment," *Biotechniques*, vol. 38, no. 3, pp. 463–471, 2005,

doi: 10.2144/05383RR02.
[28] S. D. Ian Abramson (University of California, “Institute of Mathematical Statistics is collaborating with JSTOR to digitize, preserve, and extend access to The Annals of Statistics. ® www.jstor.org,” *Ann. Stat.*, 1991.
[29] Z. Huang and T. Zhao, “Predictive performance of ensemble hydroclimatic forecasts: Verification metrics, diagnostic plots and forecast attributes,” *Wiley Interdiscip. Rev. Water*, vol. 9, no. 2, pp. 1–30, 2022, doi: 10.1002/wat2.1580.
[30] M. Doszyn, “Biasedness of Forecasts Errors for Intermittent Demand Data,” *Eur. Res. Stud. J.*, vol. XXIII, no. Special Issue 1, pp. 1113–1127, 2020, doi: 10.35808/ersj/1874.
[31] V. Kreinovich, H. T. Nguyen, and R. Ouncharoen, “How to Estimate Forecasting Quality: A System-Motivated Derivation of Symmetric Mean Absolute Percentage Error (SMAPE) and Other Similar Characteristics,” *Dep. Tech. Reports*, vol. 7, pp. 1–11, 2014, [Online]. Available: https://scholarworks.utep.edu/cs_techrep/865%0Ahttp://digitalcommons.utep.edu/cs_techrephttp://digitalcommons.utep.edu/cs_techrep/865
[32] T. Chai and R. R. Draxler, “Root mean square error (RMSE) or mean absolute error (MAE)? -Arguments against avoiding RMSE in the literature,” *Geosci. Model Dev.*, vol. 7, no. 3, pp. 1247–1250, 2014, doi: 10.5194/gmd-7-1247-2014.