Edu Komputika 11 (1) (2024)

Edu Komputika Journal

https://journal.unnes.ac.id/journals/edukom/

Feature Extraction Implementation in the Forecasting Method to Predict Indonesian Oil and Gas Exports and Imports

Michael Anggun Kado Pradana✉, Irfan Pratama

Department of Information System, Faculty of Information Technology, Universitas Mercu Buana Yogyakarta, Indonesia

Article Info

Article History:
Received: 22 June 2024
Revised: 31 July 2024
Accepted: 28 August 2024

Keywords:
CatBoost, Data Mining, Exponential Smoothing Forecasting, SARIMAX, XGBoost

Abstract

Future export and import predictions can use data mining and forecasting applications of data mining. Then, normalisation is carried out using datasets taken at the centre of the statistical agency using a mix-max scaler. The normalisation results are then calculated using several forecasting methods, such as Exponential Smoothing, SARIMAX, XGBoost, and CatBoost. The accuracy of this method can be improved by using feature extraction decomposition. They are decomposing, such as trend, residue, and seasonal. The results of the decomposition then become new features that are entered into the prediction model. The prediction results are evaluated using the root mean square error (RMSE). The smaller the RMSE, the better the results. The prediction results without using the method obtained by the Exponential Smoothing method have the best level of accuracy with an average RMSE value of 0.111 and the SARIMAX method with an average RMSE value of 0.146. Meanwhile, the prediction results using the CatBoost and XGBoost feature extraction methods have the best level of accuracy with an RMSE value of 0.046. From the results of the comparison of predictions, the addition of decomposition features to most forecasting methods can significantly increase the accuracy of the calculation.

✉ Correspondence address:
Fakultas Teknologi Informasi, Universitas Mercu Buana Yogyakarta, Gg. Jemb. Merah No.84C, Condongcatur, Yogyakarta, 55281, Indonesia
E-mail: anggunkado@gmail.com

ISSN 2252-6811
E-ISSN 2599-297X

INTRODUCTION

It is important to predict exports and imports in the future to know the expected value and volume of exports and imports so that stakeholders and related parties can make decisions. More people can play a role if they see the future of exports and imports in Indonesia by utilising information technology and data mining. Data mining is a collection of systematic processes used to explore and search for values and complex relationships stored in a database. Then, new information is extracted by extracting valuable data from the dataset. So that it can be used as one of the steps in making better decisions (Gaol et al., 2019).

Oil and gas import export data, whose data source comes from the central statistics agency, is processed with data mining by implementing feature extraction to improve accuracy results and can provide different results between forecasting methods without feature extraction and forecasting methods using feature extraction. The best results can be used to analyse future export and import development trends in the Indonesian oil and gas sector so that it is expected to provide helpful information for stakeholders in making decisions related to foreign trade. It will also give an overview of the oil and gas trade in the future.

Forecasting can be used to predict future times based on trends that have occurred in the past (Setiyani et al., 2020). Forecasting is an application of data mining to estimate the amount of demand in the future. Predictions are systematically made based on data or information from the past and present. Of course, predictions do not provide definite results, but they are used as a factor in decision-making (Wildan & Asy'ari, 2023).

Various methods can be used in time series analysis, such as ARIMA, Random Forest Regression, Fuzzy, Naive Bayes, Prophet, and XGBoost Regression (Sitepu et al., 2021). Forecasting is a method of applying data mining that can estimate the amount of demand for goods in the future. Forecasting is a systematic calculation process regarding the possibilities of what will happen in the future based on information or data collected in the past or present to reduce errors experienced. Forecasting certainly only provides results that will occur in the future. However, forecasting determines what will happen in the future so that the information can be used to make good decisions (Wildan & Asy'ari, 2023).

In a study conducted by Aditya Pratama et al. (2022), the study compared two methods, Moving Average and Exponential Smoothing, using the Indonesian export and import values dataset in 2020. The results of the study showed that the best prediction results for Indonesian export values used the Exponential Smoothing method with a MAPE value of 8.17% with a predicted value for the following month of 14164.39 million US$ (Pratama et al., 2022). Another study conducted by Devilia Rahmawati (2021), the study used two models in predicting Wana Wisata Sayang Ka'ak Ciamis visitors; the two methods used were Box-Jenkins and Exponential Smoothing, and the results of the test showed that the Exponential Smoothing method had better results than the Box-Jenkins method with calculations in the 11th to 14th week giving an average error value of 7.77% while the Box-Jenkins method gave an average error value of 9.90%. The results of the calculation using the Exponential Smoothing method in the 15th week were 192.8377 (Rahmawati, 2021).

Other methods that can be used include the SARIMAX method in a study conducted by Nur Latief et al. (2022); the study used the SARIMAX method in its testing by taking data on rainfall and air temperature in the city of Makassar in 2007-2020, which were obtained from the Meteorology, Climatology and Geophysics Agency, Central Statistics Agency of South Sulawesi in. The MAPE value obtained was 17.75% (Latief et al., 2022). In addition, Ilham Julianto et al. (2021) research used data from West Java tourist visits using Arimax and SARIMAX modelling. The results of the study obtained a MAPE value of 16.61%. From the results of this study, the level of MAPE decline in ARIMAX modelling is relatively tiny, with a declining value of 1-2% compared to the SARIMAX method, which has a MAPE decline rate of 1-3% (Julianto et al., 2021).

In another forecasting method study conducted by Yoan Purbolingga et al. (2023), the study used a comparison of the XGBoost and CatBoost methods by taking data from the UCI Machine Learning Repository, with 918 data between the CatBoost and XGBoost methods, the CatBoost method obtained the calculation results of the CatBoost method getting a higher level of accuracy of 0.884058, compared to XGBoost with an accuracy level value of 0.847826 (Purbolingga et al., 2023). Other research on the CatBoost method, such as the research conducted by Andrian Istianto et al. (2024), used the CatBoost method in predicting rainfall by taking data sources from the Kaggle website for ten years, with a total of 145,460 data. Then, the data was separated between test data and test data with a ratio of 70% and 30% of the results of the rainfall prediction calculation of 94% of the attributes in the data (Istianto et al., 2024).

Research conducted by Dwika Ananda et al. (2022), the study predicted ratings on the

Playstore application using three stages, namely Data collection, Data preprocessing, and Data Modeling, with a dataset of 10,840 rows and 9660 different rows and has 13 features; the results of this study are, after testing six different methods including Decision Tree, K-Nearest Neighbors, Gradient boosting Tree, Random Forest, LightGBM and, XGBoost, the XGBoost method is at the first level with an accuracy level of up to 77.5% (Ananda et al., 2022). In another study conducted by Indah Sari et al. (2023), the study compared two algorithms, namely the Random Forest Algorithm and XGBoost, using the Metro Interstate Traffic Volume dataset consisting of 48204 rows and nine columns with a period of 2012-2018. The study shows no significant difference between the two algorithms in traffic volume data. However, the RMSE value of XGBoost is lower than that of Random Forest, where smaller values give better results and data processing time between the two algorithms. XGBoost has a data processing speed of 532% faster than Random Forest's. This shows that using XGBoost in prediction is appropriate because it has a quicker and more accurate processing time (Sari et al., 2023).

When carrying out prediction calculations to increase accuracy, feature extraction can be used using decomposition techniques.Decomposition is a prediction method based on patterns that predict what will happen again in the same pattern by identifying patterns in the data and extracting three essential components, namely trend, seasonality, and randomness (Aswi et al., 2024). Feature extraction is a standard method in which researchers try to develop transformations while retaining most information. Feature extraction and selection methods, such as forecast accuracy, are used separately or in combination to improve performance. However, its use must go through the feature selection stage because the features used could be useless. Feature selection is the process of selecting the best features that are useful for distinguishing classes. Feature extraction can be used in this context to reduce complexity and provide a simple representation of the data that represents each variable (Khalil, 2014).

Based on several previous studies, we can use the Moving Average, Exponential Smoothing, SARIMAX, XGBoost, and CatBoost methods to predict. Previous studies show that the Moving Average method better predicts data such as export and import values and tourist visitors. Previous studies using the SARIMAX method on rainfall, air temperature, and tourist visits show that this method effectively handles complex data and can provide reasonably accurate predictions. However, the results vary depending on the context and parameters used, and data must have a seasonal character for this method to be used.

Newer forecasting methods such as XGBoost and CatBoost in previous studies have shown a high level of accuracy with fast processing time in predicting large amounts of data. Still, each method has a level of accuracy and advantages and disadvantages. Therefore, in this study, feature extraction decomposition will be implemented by extracting three parameters: trend, seasonal, and residual. Then, the extraction results will become new features in the Moving Average, Exponential Smoothing, SARIMAX, and XGBoost methods so that they can compare the accuracy level between methods that use feature extraction and methods without feature extraction.

RESEARCH METHODS

This research uses the Cross Industry Standard Process for Data Mining (CRISP-DM) methodology.

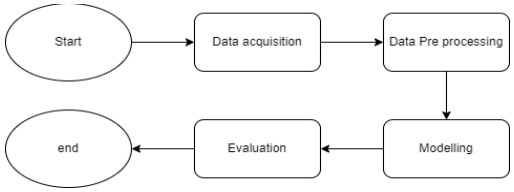

Figure 1. Research flow

Figure 1 shows that the research flow for implementing feature extraction in the forecasting method is Data acquisition, Data preprocessing, Modeling, and Evaluation.

A. Data Acquisition

The dataset in this research is obtained from the official website of the Central Bureau of Statistics of Indonesia which can be accessed via https://www.bps.go.id/id. This dataset has a time range starting from January 2012 to February 2024 with specific attributes.

B. Data Pre-Processing

Data processing is one of the essential things in this stage, including data cleaning, handling missing values, coding categorical variables, and data scaling (Sunarko et al., 2023). In the stage 1, existing data is cleaned from missing and null values. In the stage 2, the dataset is normalised using a mix-max scaler to make the value more stable between 0-1. Stage 3 seasonality checking is carried out because it will use the SARIMAX method, so the data must be seasonal. In the stage 4, the normalised dataset is

subjected to decomposition feature extraction; there are several developments in decomposition, such as additive decomposition and multiplicative decomposition. This study uses additive decomposition to decompose period data using trend, seasonality, and residual components. Additive is used because all elements are added to get results (Anggraeni et al., 2022). The final dataset is then separated into training and testing datasets. Training data is used to train the model. Testing data is used to test the performance of the model that has been trained using training data. Data division can be taken for training data of at least 80% of the data and 20% testing data. The division of training data must be greater than the testing data so that the model learning performance is better. The Training and testing data division ratio difference can affect model performance (Nugraha & Syarif, 2023).

C. Modelling

The resulting data from the preprocessing stage will proceed to the modelling stage using the forecasting training and testing methods that have been prepared previously. The forecasting methods include XGBoost, CatBoost, Exponential Smoothing, and SARIMAX.

1. Exponential Smoothing

Exponential Smoothing modelling is a method widely used in time series analysis. It was developed by Box and Jenkins in 1976 and involves a set of adaptive coefficients. This method undergoes regular improvements on the newest forecasting objects and focuses on exponential decreases in older objects (Rahmawati, 2021). In other words, this method takes time series data in a decreasing (exponential) manner by calculating the average value of past refinements (Pratama et al., 2022). This method requires forecasters to determine whether there are seasonal trends and patterns in historical data and then select the appropriate form of those attributes (Chandra & Budi, 2020).

$$F\hat{}t + 1 = \alpha xt + (1 - \alpha)F\hat{}t \qquad (1)$$

In equation 1, xt is the estimated demand in the period, and Ft+1 is the demand for period t+1. The exponential constant α is a smoothing factor with values between 0 and 1.

2. SARIMAX

SARIMAX is an abbreviation for Seasonal AutoRegressive Integrated Moving Average with exogenous regressors, a model derived from its predecessors, the SARIMA and ARIMA models. The concept of the SARIMA model is almost the same as that of ARIMA. The only difference is the additional element of seasonality (Nagakusuma et al., 2022). This time series model includes seasonal variations, which are patterns that appear repeatedly over a period and have a fixed nature. This model is designed to capture regular seasonal fluctuations in the data, thus allowing for more accurate predictions when dealing with data with such characteristics (Julianto et al., 2021). The SARIMA model is denoted as SARIMAX (p, d, q) (P, D, Q) m with the following notation.

$$\phi p(B)\ \Phi P(Bm)\ (1 - B)d(1 - Bm)Dyt = \theta q(B)\ \Theta Q(Bm)\ \varepsilon t \qquad (2)$$

In Equation 2, the time series model consists of several components. p represents the number of non-seasonal autoregressive (AR) components, d is the non-seasonal integration order, and q indicates the number of non-seasonal moving average (MA) components. For the seasonal component, p represents the number of seasonal autoregressive (AR) components, d is the seasonal integration order, and Q indicates the number of seasonal moving average (MA) components. The seasonal period in this model is determined by m. The autoregressive parameters are denoted by ϕp for non-seasonal and ΦP for seasonal, while the moving average parameters are denoted by θq for non-seasonal and ΘQ for seasonal. In addition, B is the lag operator used in this model (Maysofa et al., 2023).

3. XGBoost

XGBoost is a development algorithm from gradient tree boosting and modelling that uses assemble logic. This method is suitable for handling large-scale machine-learning activities. XGBoost can complete various activities such as classification, regression, and ranking because XGBoost can adapt to multiple situations and is very flexible in past calculations. The XGBoost calculation flow is carried out by collecting trees consisting of various trees previously usually called classification and regression trees (CART) (Herni Yulianti et al., 2022). Classification and regression tree (CART) is a general term for classification tree and regression tree.

The CART classification tree introduces the Gini coefficient to replace the information gain. The CART regression tree will look for the mean or median of the ends to predict the outcome. Cost complexity pruning (CCP) is used to cut non-leaf nodes with a minor gain error and delete child nodes with non-leaf nodes to avoid overfitting (Lv et al., 2021). XGBoost is an

ensemble of decision trees built additively. The model in an iteration is expressed as follows.

$$obj(\emptyset) = \sum_i L\,(y\hat{}_i, y_i) + \sum_i L \cap (f_k), f_k \in F \qquad (3)$$

In Equation 3, L represents the difference in standard deviation between the predicted value and the original value ∩ and represents the model complexity regularisation function to avoid overfitting. F represents the function in the functional space F, and F represents the set of all trees created (Sari et al., 2023).

4. CatBoost

CatBoost is a categorical boosting developed by Yandex by implementing gradient boosting modelling, which uses logical decision trees as base predictors (Purbolingga et al., 2023). CatBoost can automatically handle categorical data using statistical methods, so it can overcome data overfitting by optimising various input parameters. CatBoost performs random execution on category data rather than calculating the average value of labels and binary replacement (Istianto et al., 2024). CatBoost adds a decision tree model based on the gradient of the loss function to improve the previous prediction. The prediction is updated with the following formula.

$$Ft + 1(x) = Ft(x) + \eta ht(x) \qquad (4)$$

From Equation 4, $Ft(x)$ is the prediction at iteration t, η is the learning rate. This factor controls how much influence the new model has on the final prediction, $\eta ht(x)$ is the decision tree created at iteration t.

D. Evaluation

At this stage, it is related to evaluating the method that has been built and testing how well the model works on data that has never been seen before by comparing the original and predicted data. The Root Mean Square Error (RMSE) model is also evaluated at this stage. RMSE results from the difference between the expected and original values. The expected result is that the smaller the RMSE obtained. RMSE is formulated in Equation 5 as follows.

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}} \qquad (5)$$

In Equation 5, you can see the formula for calculating the RMSE value (Hudzaifah & Rismayadi, 2021).

RESULT AND DISCUSSION

The results of each stage in the methodology section will be explained next in the Results and Discussion section. This section discusses the results of Data Acquisition, Pre-Processing, modelling, and evaluation.

A. Data Acquisition

The dataset used in this research covers January 2012 to February 2024. It has approximately 145 rows of data and four types of data: Oil and Gas Export Value, Oil and Gas Export Volume, Oil and Gas Import Value, and Oil and Gas Import Volume. All the data types are Time Series. A sample of the dataset that will be used in this research can be seen in Table 1.

Table 1. Oil and Gas Export-Import Volume Value Data

Date	Oil and Gas export value	Oil and Gas Export Volume	Oil and Gas Import Value	Oil and Gas Import Volume
2012-01-01	3142.6	4187.4	3019.3	3168.2
2012-02-01	3355.5	4236.8	3492.7	3689.60
2012-03-01	3486.1	4525.5	4008.9	3813.22
2012-04-01	3560.7	4382.9	4120.4	3986.74
2012-05-01	3724.9	4631.8	3442.1	3435.94
…	…	…	…	…
2023-10-01	1370.4	2190.6	3206.8	4308.70
2023-11-01	1282.9	2035.2	3488.7	5004.20
2023-12-01	1478.9	2671.3	3372.4	5011.40
2024-01-01	1397.6	2360.3	2698.3	4045.30
2024-02-01	1216.9	1865.8	2979.6	4298.80

B. Data Pre-processing

Data management is one of the important things in this stage because the existing data is cleaned from missing values, null values , and so on. This stage begins with cleaning the data, then checking whether the data is seasonal or not, followed by data transformation. The purpose of processing the data is to ensure quality and that the data is clean and ready to use so that the results obtained from the study are relevant. The explanation of the calculation steps is as follows:

1. Checking missing values and attribute types.

Table 2. Attribute Types

Column	Non-Null Count	Dtype
Date	146 non-null	datetime64[ns]
Oil and Gas export value	146 non-null	float64
Oil and Gas Export Volume	146 non-null	float64
Oil and Gas Import Value	146 non-null	float64
Oil and Gas Import Volume	146 non-null	float64

After going through the stages of checking missing and null values, the result was that the dataset used did not have the two conditions mentioned. So, the data is ready to be used for the following processing stage.

2. Data Exploration

The existing dataset is then analysed further to understand the characteristics and patterns in the data. The goal is to understand the data before further processing it. At this stage, data exploration is intended to determine the characteristics of the data. Information that can be generated includes whether the dataset has a seasonal pattern component. This information can be found in several ways, including analysing the Autocorrelation Function (ACF) results. Using the function contained in the Python programming language, information is obtained that the dataset has seasonal characteristics. An example of an ACF plot can be seen in Figure 3.

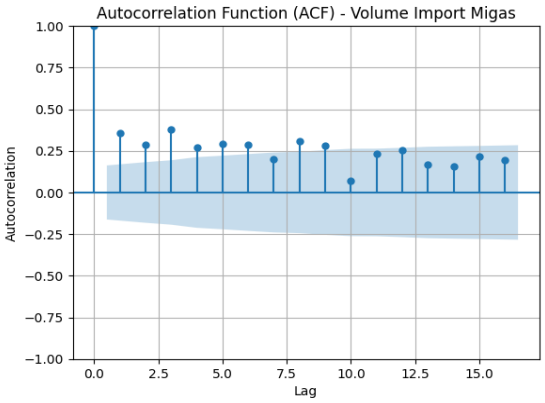

Figure 3. Plot ACF

The presence of seasonal characteristics in the dataset provides options regarding method use. For example, the SARIMAX method is one of the methods used in the modelling stage. This method assumes the data is seasonal, so it must first be proven whether the dataset has these characteristics.

3. Normalization

Even though the data used are the same units, the maximum and minimum values of the data are not the same, so data normalisation is needed so that all data attributes have the same range of maximum and minimum values (Kharis, 2023). The normalisation process helps make the forecasting method more stable and convergent. This research will use a mix-max scaler to normalise the original data. A mix-max scaler is a data normalisation method that carries out linear transformations on original data with a range of 0 to 1 to balance the values on unbalanced attributes (Rifatama et al., 2023). The results of data normalisation can be seen in Table 3.

Table 3. Data Normalization

Date	Oil and Gas export value	Oil and Gas Export Volume	Oil and Gas Import Value	Oil and Gas Import Volume
2012-01-01	0,8160	0,8607	0,6219	0,3308
2012-02-01	0,8832	0,8762	0,7465	0,5148
2012-03-01	0,9245	0,9667	0,8825	0,5585
2012-04-01	0,9481	0,9220	0,9118	0,6197
2012-05-01	1,0000	1,0000	0,7332	0,4253
…	…	…	…	…
2023-10-01	0,2558	0,2346	0,6713	0,7334
2023-11-01	0,2282	0,1859	0,7455	0,9790
2023-12-01	0,2901	0,3853	0,7149	0,9815
2024-01-01	0,2644	0,2878	0,5374	0,6404
2024-02-01	0,2073	0,1328	0,6114	0,7299

The normalisation results shown in Table 3 will be used as a dataset in the modelling and evaluation stages. This normalisation process can be cancelled if needed (especially in the actual forecasting process), and the data values will return to their original form. Thus, the normalised data will only be used to determine the estimated data. However, the final data will still be presented in its original format.

4. Decomposition

Decomposition is a method since what usually happens will repeat or reoccur with the same pattern (seasonal). Usually, the pattern is quite

complex; for example, a value increase fluctuates and is irregular. Analysis and forecasting are generally complicated, so to overcome this, decomposition (fraction) is carried out into three components: trend and seasonal residue (Hudzaifah & Rismayadi, 2021). In short, decomposition is the sequential collection of data in time by analysing the pattern of relationships between variables. The decomposition result variables will be used as additional features in each dataset (export and import). Table 4 shows an example of the final dataset form used in the modelling stage.

Table 4. Training Data Samples with Decomposition

Date	Oil and Gas export value	Trend	Seasonal	Residual
2012-03-01	0,9245	0,9144	0,0050	0,0051
2012-04-01	0,9481	0,8990	0,0121	0,0370
2012-05-01	1,0000	0,8715	-0,0036	0,1322
2012-06-01	0,7392	0,8270	-0,0118	-0,0760
2012-07-01	0,7455	0,7771	-0,0017	-0,0298

After all the preprocessing stages are done, the dataset will go through the final stage, namely splitting. At this stage, the dataset will be divided into data used for model training (training) and model testing (testing). The ratio of each part is 132 data and 14 data. Because the data used is time series data, the data division mechanism is carried out by paying attention to the data sequence (not done randomly) to maintain data continuity.

C. Modelling

After all pre-processing stages have been carried out, the dataset will be tested using several forecasting methods. So, we know which method is most suitable for use on datasets of a similar type. The research uses several forecasting methods: XGBoost, CatBoost, SARIMAX, and Exponential Smoothing. In this research, two schemes are used at the modelling stage: testing using a dataset without new features without decomposition and a testing method using a dataset with the addition of new features resulting from decomposition.

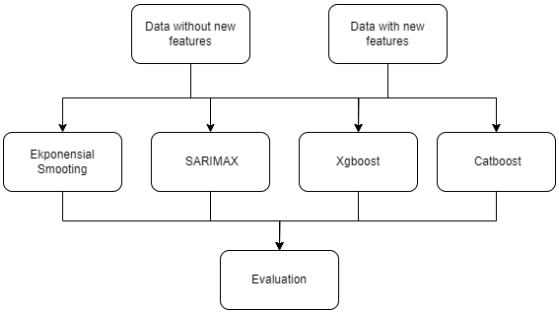

Figure 2. Model testing scenarios

The following are the results of each method test, presented in graphical form to show how close the actual data is compared to the predicted data produced by each method from the previously presented schemes.

1. Exponential Smoothing

In Figure 4 (a), the exponential method prediction result graph experiences fluctuations because some prediction points are close to the actual value points. In Figure 4 (b), the exponential method with decomposition prediction results shows that the graph tends to be more stable.

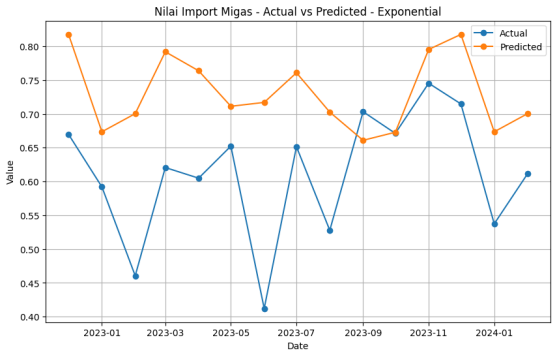

Figure 4. (a) Exponential smoothing method test graph

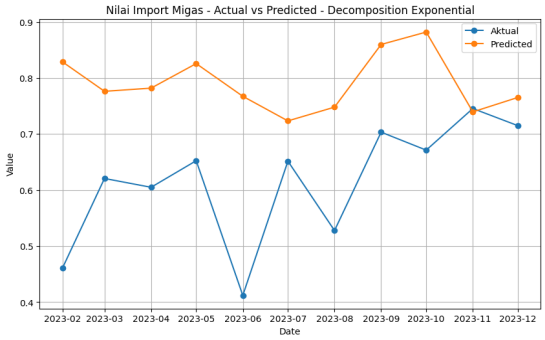

Figure 4. (b) Exponential smoothing method test graph with additional features

However, the model with decomposition in the exponential method does not capture sharp fluctuations in the actual value. Based on Figure 4, the Exponential Smoothing method tends to have predictions closer to the actual value at some points, even though both models have difficulty capturing sharp fluctuations in the actual value. These results show that the decomposition process's additional features do not significantly help with the Exponential Smoothing method test results.

2. SARIMAX

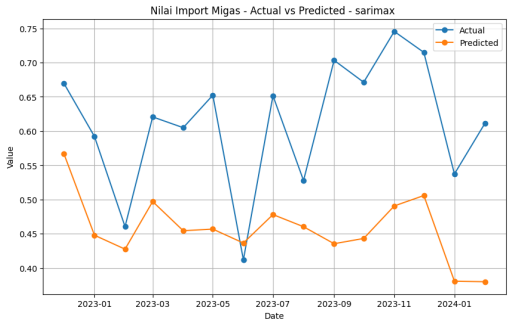

Figure 5. (a) SARIMAX method test chart

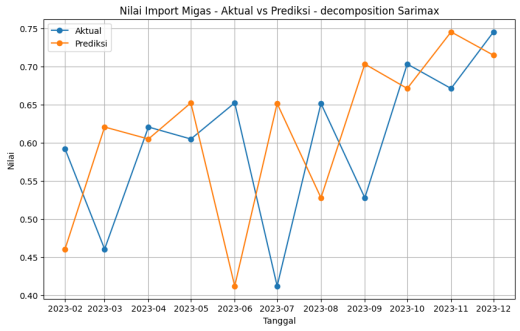

Figure 5. (b) SARIMAX method test chart with additional features

In Figure 5 (a), The prediction result graph using SARIMAX shows that the prediction point tends to be stable but cannot follow the fluctuation of the actual value. However, some prediction points are close to the actual value point. In Figure 5 (b), the prediction result of the SARIMAX method using decomposition shows more synchronisation with the actual value but is still lacking in capturing sharp fluctuations. From both images, Figure 5 (b) shows better results in capturing the actual value pattern to provide more accurate results. Some points are closer to the actual value compared to Figure 5 (b) From these results, the additional features of the decomposition process appear to help the prediction results of the SARIMAX method so that it produces a prediction value that is relatively close to the actual value of the dataset. It is a seasonal model of a time series

3. XGBoost

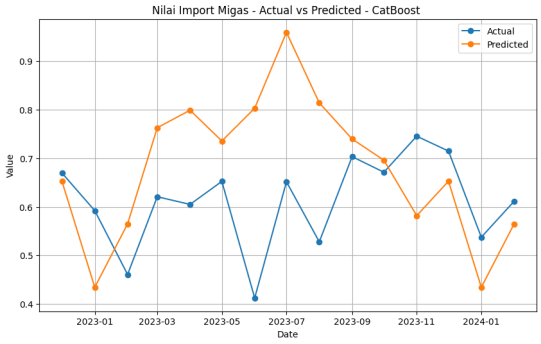

Figure 6. (a) XGBoost method test graph

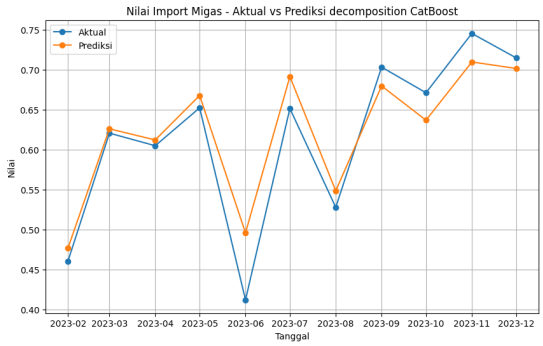

Figure 6. (b) XGBoost method test chart with additional features

In Figure 6 (a), the prediction results using the XGBoost method without decomposition show a relatively large fluctuation in the prediction value at several points. The prediction value cannot be applied to the actual value at several points. In Figure 6 (b), the prediction results using the XGBoost method with decomposition show that the model prediction is more accurate because more prediction points are close to the actual value, and some periods of prediction results are very close to the actual value. Overall, Figure 6 (b) provides more accurate results because it matches the predicted and actual values well. The application of decomposition to the XGBoost method provides significant results in prediction. From these results, the additional features of the decomposition process are more helpful significantly than the results of the two methods

previously discussed, where the results of the prediction graph in Figure 6 (b) are more like the actual data.

4. CatBoost

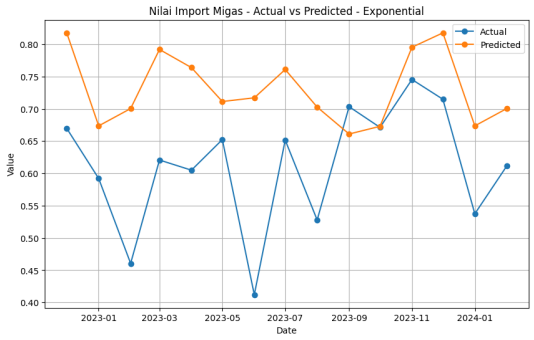

Figure 7. (a) CatBoost method test graph

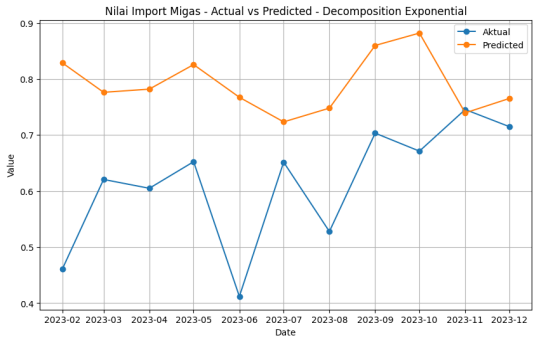

Figure 7. (b) CatBoost method test chart with additional features

Figure 7 (a) prediction results using the CatBoost method without decomposition show that the graph experiences quite large fluctuations because, at several points, the prediction results have many prediction points with actual value points that have significant differences even though some prediction points are close to the actual value points, indicating that the model has difficulty in following the actual value pattern. In Figure 7 (b), the CatBoost decomposition method's prediction results show that the graph shows more stable and neat results, and the predicted value points are closer to the actual value. They show good agreement between the predicted and actual values with minimal deviation. Overall, the CatBoost method with decomposition provides accurate results. Of the three previous methods, the CatBoost method graph results show the most significant performance increase after the decomposition process's new features are applied. This shows that feature extraction with the decomposition mechanism influences the forecasting results of most of the methods used for testing in this research.

D. Evaluation

The modeling process has been carried out, and the evaluation results of each model are obtained for each type of data using the RMSE measurement metric. The RMSE values obtained can be seen in Tables 4 and 5.

Table 5. RMSE Calculation Results Without Decomposition Method

Method	Oil and Gas export value	Oil and Gas Export Volume	Oil and Gas Import Value	Oil and Gas Import Volume	Avg
XGBoost	0,093	0,124	0,189	0,239	0,161
Exponen sial	0,036	0,096	0,146	0,167	**0,111**
Catboost	0,081	0,109	0,178	0,223	0,148
Sarimax	0,064	0,117	0,174	0,222	0,144

Table 6. RMSE Calculation Results Using Decomposition Method

Method	Oil and Gas export value	Oil and Gas Export Volume	Oil and Gas Import Value	Oil and Gas Import Volume	Avg
XGBoost	0,013	0,023	0,052	0,095	**0,046**
Exponen sial	0,044	0,090	0,207	0,253	0,148
Catboost	0,011	0,022	0,034	0,118	**0,046**
Sarimax	0,027	0,052	0,117	0,196	0,098

Based on the results in Table 5, the method with the smallest RMSE is Exponential Smoothing with an average RMSE value of 0.111. These results are better than those of the CatBoost and SARIMAX methods. These results are the results of testing with a scheme without new decomposition features. Table 6 shows that the most petite average RMSE methods are XGBoost and CatBoost, with an average RMSE value of 0.046. These results are tests with a new decomposition feature addition scheme. In Table 5, the exponential smoothing method gives the best calculation results. This is in line with research conducted by Aditya Pratama et al. (2022) that shows that the exponential smoothing method has a good level of accuracy in Indonesian export values (Pratama et al., 2022).

When compared between the best results of Table 5 and Table 6, the smallest RMSE value is the RMSE value from Table 6. This shows that adding new features to the decomposition can improve the accuracy of the forecasting results from most of the methods used in this study except exponential smoothing. In Table 6, the RMSE results for the exponential smoothing method are

much larger than in Table 5. This shows that the addition of new features to the dataset does not have a good effect on the performance of the forecasting method because the exponential smoothing method experiences overfitting due to the addition of trend, seasonal, and residual components resulting from decomposition into the model, making the model too complicated and causing the model to have difficulty capturing patterns in the test data. At the same time, the XGBoost method uses cost complexity pruning (CCP) to handle data overfitting. The CatBoost method performs automatic optimisation of various parameters entered so that adding features to these methods can provide better results.

CONCLUSION

The calculation results of the method without decomposition of the Exponential Smoothing method have the smallest average RMSE value with an average value of 0.111. However, in the calculation using the addition of the decomposition feature, the Exponential Smoothing Method has an RMSE value of 0.148. This value is higher than without the use of decomposition. However, the other three methods in using the decomposition feature provide better results such as the exponential method, the three methods provide more minor RMSE results, the XGBoost method with an average value of 0.046, CatBoost with an average value of 0.046 and SARIMAX with an average value of 0.098.

The results of using the decomposition feature significantly affect most of the study's methods, except for the exponential smoothing method. The Exponential Smoothing method experienced a decrease in performance with an increasing RMSE when the decomposition feature was added, indicating that this method is overfitting and unable to handle the additional complexity of the trend, seasonal, and residual components. Other methods can overcome overfitting through cost complexity pruning (CCP) in XGBoost and automatic optimisation of parameter addition in CatBoost.

This study's findings answer how adding new decomposition features affects the prediction model for Indonesian oil and gas exports and imports. The main objective of this study is to determine the most accurate prediction method and compare methods with new features and methods without new features. The best method is to use the XGBoost and CatBoost methods. So that it can be used to help stakeholders make better decisions, this study has limitations in terms of the data period of only 11 years and the lack of data details. In addition, the method used may have limitations in capturing more complex patterns. In addition, it is still limited to feature extraction with additive decomposition; hopefully, in the study, we can collect datasets with a longer and more detailed period to improve prediction accuracy. It can also use more sophisticated methods to handle more complex data and explore other feature extractions to improve the accuracy of a forecasting method. Another hope in the following study is to develop a system that can compare oil and gas value and volume predictions to help analyse development trends. It can also be used for decision-making in the oil and gas sector by stakeholders or related parties who need it.

REFERENCES

Ananda, D., Pertiwi, A., & Muslim, M. A. (2022). *Prediksi Rating Aplikasi Playstore Menggunakan Xgboost Prediksi Rating Aplikasi Playstore Menggunakan Xgboost*. *October 2020*.

Anggraeni, A. S., Utama, R. C., & Wati, D. C. (2022). Penghalusan eksponensial dan dekomposisi saham apple.inc. *Jurnal Sintak*, *1*(1), 24–30. https://journal.iteba.ac.id/index.php/jurnalsintak/article/view/25%0Ahttps://journal.iteba.ac.id/index.php/jurnalsintak/article/download/25/25

Aswi, A., Rahma, I., & S, M. F. (2024). *Penerapan Metode Hybrid Dekomposisi-Arima dalam Peramalan Jumlah Wisatawan Mancanegara*. *7*(1), 19–26. https://doi.org/10.12962/j27213862.v7i1.18738

Chandra, C., & Budi, S. (2020). Analisis Komparatif ARIMA dan Prophet dengan Studi Kasus Dataset Pendaftaran Mahasiswa Baru. *Jurnal Teknik Informatika Dan Sistem Informasi*, *6*(2), 278–287. https://doi.org/10.28932/jutisi.v6i2.2676

Gaol, I. L. L., Sinurat, S., & Siagian, E. R. (2019). Implementasi Data Mining Dengan Metode Regresi Linear Berganda Untuk Memprediksi Data Persediaan Buku Pada Pt. Yudhistira Ghalia Indonesia Area Sumatera Utara. *KOMIK (Konferensi Nasional Teknologi Informasi Dan Komputer)*, *3*(1), 130–133. https://doi.org/10.30865/komik.v3i1.1579

Herni Yulianti, S. E., Oni, S., & Yuana, S. (2022).

Penerapan Metode Extreme Gradient Boosting (XGBOOST) pada Klasifikasi Nasabah Kartu Kredit. *Journal of Mathematics: Theory and Applications*, *4*(1), 21–26. https://doi.org/10.31605/jomta.v4i1.1792

Hudzaifah, M., & Rismayadi, A. A. (2021). Peramalan Arus Lalu Lintas Berdasarkan Waktu Tempuh Dan Cuaca Menggunakan Metode Time Series Decomposition. *Jurnal Responsif : Riset Sains Dan Informatika*, *3*(2), 207–215. https://doi.org/10.51977/jti.v3i2.559

Istianto, A. F., Id Hadiana, A., & Rakhmat Umbara, F. (2024). Prediksi Curah Hujan Menggunakan Metode Categorical Boosting (Catboost). *JATI (Jurnal Mahasiswa Teknik Informatika)*, *7*(4), 2930–2937. https://doi.org/10.36040/jati.v7i4.7304

Julianto, I. R., Indwiarti, & Rohmawati, A. A. (2021). Prediksi Jumlah Kunjungan Wisatawan Di Jawa Barat Dengan Model Arimax Dan Sarimax Menggunakan Data Google Trends. *E-Proceeding of Engineering*, *8*(435), 4229–4241.

Khalil, T. (2014). *A Survey of Feature Selection and Feature Extraction Techniques in Machine Learning*. 372–378.

Kharis, S. A. A. (2023). Prediksi Kelulusan Siswa pada Mata Pelajaran Matematika menggunakan Educational Data Mining. *Jurnal Riset Pembelajaran Matematika Sekolah*, *7*(1), 21–29. https://doi.org/10.21009/jrpms.071.03

Latief, N. H., Nur'eni, N., & Setiawan, I. (2022). Peramalan Curah Hujan di Kota Makassar dengan Menggunakan Metode SARIMAX. *STATISTIKA Journal of Theoretical Statistics and Its Applications*, *22*(1), 55–63. https://doi.org/10.29313/statistika.v22i1.990

Lv, C. X., An, S. Y., Qiao, B. J., & Wu, W. (2021). Time series analysis of hemorrhagic fever with renal syndrome in mainland China by using an XGBoost forecasting model. *BMC Infectious Diseases*, *21*(1), 1–13. https://doi.org/10.1186/s12879-021-06503-y

Maysofa, L., Syaliman, K. U., & Sapriadi. (2023). Implementasi Forecasting Pada Penjualan Inaura Hair Care Dengan Metode Single Exponential Smoothing. *Jurnal Testing Dan Implementasi Sistem Informasi*, *1*(2), 82–91.

Nagakusuma, J., Palit, H., & Juwiantho, H. (2022). *Prediksi Penjualan Pada Data Penjualan Perusahaan X Dengan Membandingkan Metode GRU , SVR , DAN*.

Nugraha, W., & Syarif, M. (2023). Teknik Weighting untuk Mengatasi Ketidakseimbangan Kelas Pada Prediksi Churn Menggunakan XGBoost, LightGBM, dan CatBoost. *Techno.Com*, *22*(1), 97–108. https://doi.org/10.33633/tc.v22i1.7191

Pratama, A. A., Agushinta R., D., & Mukhyi, M. A. (2022). Penerapan Metode Moving Average dan Exponential Smoothing untuk Prediksi Nilai Ekspor dan Impor Indonesia. *Jurnal Ilmiah FIFO*, *14*(1), 58. https://doi.org/10.22441/fifo.2022.v14i1.006

Purbolingga, Y., Marta, D., Rahmawatia, A., & Wajhi, B. (2023). Perbandingan Algoritma CatBoost dan XGBoost dalam Klasifikasi Penyakit Jantung. *Jurnal APTEK Vol. 15 No 2 (2023) 126-133*, *15*(2), 126–133. http://journal.upp.ac.id/index.php/aptek/article/download/1930/1163/4970

Rahmawati, D. (2021). Pemodelan Box-Jenkins dan Exponential Smoothing untuk Prediksi Pengunjung Daerah Wisata Sayang Ka'ak Ciamis. *Jurnal Riset Matematika*, *1*(2), 109–118. https://doi.org/10.29313/jrm.v1i2.375

Rifatama, M. I., Faisal, M. R., Herteno, R., Budiman, I., Itqan, M., & Mazdadi. (2023). Optimasi algoritma k-nearest neighbor dengan seleksi fitur menggunakan xgboost. *JIRE (Jurnal Informatika & Rekayasa Elektronika)*, *6*(1), 64–72.

Sari, F. I., Gunawan, E. L., Adhigiadany, C. A., & Lisanthoni, A. (2023). *Model Prediksi Kepadatan Lalu Lintas : Perbandingan Antara Algoritma Random Forest dan XGBoost*. *2023*(Senada), 296–303.

Setiyani, L., Wahidin, M., Awaludin, D., & Purwani, S. (2020). Analisis Prediksi Kelulusan Mahasiswa Tepat Waktu Menggunakan Metode Data Mining Naïve Bayes : Systematic Review. *Faktor Exacta*, *13*(1), 35. https://doi.org/10.30998/faktorexacta.v13i1.5548

Sitepu, F. T. B., Sirait Vince Amelia Prada, & Yunis, R. (2021). Analisis Runtun Waktu Untuk Memprediksi Jumlah Mahasiswa Baru Dengan Model Random Forest. *Paradigma - Jurnal Komputer Dan Informatika*, *23*(1), 99–105. https://doi.org/10.31294/p.v23i1.9781

Sunarko, B., Hasanah, U., Hidayat, S.,

Muhammad, N., Ardiansyah, M. I., Ananda, B. P., Hakiki, M. K., & Baroroh, L. T. (2023). Penerapan Stacking Ensemble Learning untuk Klasifikasi Efek Kesehatan Akibat Pencemaran Udara. *Edu Komputika Journal*, *10*(1), 55–63. https://doi.org/10.15294/edukomputika.v10i1.72080

Wildan, K., & Asy'ari, S. (2023). *Penentuan Metode Peramalan (Forecasting) Pada Permintaan Penjualan Di Cv. Lia Tirta Jaya Prigen*. *2*(11), 4077–4089.