p-ISSN 2338 – 3216
e-ISSN 2528 – 1070

JOURNAL OF STATISTICS
UNIVERSITY OF MUHAMMADIYAH SEMARANG
12(1) 2024: 24-33

Jurnal Statistika
Universitas Muhammadiyah Semarang

FORECASTING THE OCCUPANCY RATE OF STAR HOTELS IN BALI USING THE XGBOOST AND SVR METHODS

Aviolla Terza Damaliana[1*], Amri Muhaimin[2], Dwi Arman Prasetya[3]
[1,2,3] Data Science, Faculty of Computer Science, UPNVJT, Indonesia

***E-mail**: *aviolla.terza.sada@upnjatim.ac.id*

Article Info:
Received: May 31, 2024
Accepted: June 6, 2024
Available Online: June 16, 2024

Keywords:
Forecasting, Hotel, SVR, Tourism, XGBoost.

Abstract: The hotel occupancy rate indicator has become a concern in recent years as it goes hand in hand with the rapid growth of the global tourism industry. A way to maintain or improve this indicator is to carry out managerial planning using forecasting methods. The forecasting methods used in this research are XGBoost and SVR. The advantage of this modeling is that it achieves high accuracy and processing speed. Meanwhile, the benefit of SVR is that it will produce good predictions because it can overcome overfitting. The steps in this research are exploring data, separating training data and testing data, transforming data, modeling data, forecasting data, and evaluating forecasting results using RMSE, MAE, and MAPE. The results show that the MAPE value from both methods is smaller than 10%, which means that both methods can accurately predict the occupancy rate of star hotels in Bali. Apart from that, the SVR method has smaller values for all model evaluation criteria than the XGBoost method, which means that the SVR method is better than XGBoost for predicting the occupancy rate of star hotels in Bali.

1. INTRODUCTION

Bali is one of the islands in Indonesia that is wealthy in culture and has many beautiful natural resources [1]. This island has many tourist attractions, including beach tourism, arts and culture tourism, natural tourism such as mountains and forests, shopping tourism, and various other interesting tourist attractions to visit. Unsurprisingly, the island of Bali is one of the destinations most visited by foreign tourists [2]. According to BPS [3], Bali's Ngurah Rai Airport is the main entry point for foreign tourists to Indonesia, with 5.25 million visits or 44.95% of the total number of foreign tourists entering Indonesia. Of course, the increasing number of tourists is one sign that the tourism industry in a region has developed [4].

To increase foreign exchange, the government must continue developing facilities and infrastructure that support tourism [5]. One of the infrastructures that needs to be developed is a place for tourists to stay or hotels. This is because the more crowded or popular a destination's tourism industry is, the better the condition of the tourism industry [5]. The hotel occupancy rate is a key indicator in the hospitality industry that reflects the extent to which a property can attract and retain guests [6]. Therefore, this indicator is one of the main focuses for hotel owners and governments that can provide an overview of an area's economic performance and attractiveness for tourists [6].

[doi: 10.14710/JSUNIMUS. 12.1.2024.24-33]

The hotel occupancy rate indicator has become a concern in recent years as it goes hand in hand with the rapid growth of the global tourism industry. A hotel's success in maintaining high room occupancy indicators can positively impact not only the hotel's revenue but also the local economy and the tourism industry as a whole. A way to maintain or improve this indicator is to carry out managerial planning using forecasting methods [6]. Forecasting hotel occupancy rate indicators has become critical in hospitality management strategies. The success of a hotel in predicting and managing occupancy rates can be an important foundation to improve operational efficiency and financial stability, as well as make hotel management make the right decisions to optimize room rates according to demand.

Several researchers have forecasted hotel occupancy rates, namely [6] and [7]. Darmawan et al [6] carried out the forecasting method using the Naïve Bayes and Decomposition method by producing MAPE values of 10.85% and 10.78%. The forecasting method carried out by Tamasoleng et al [7] uses the triple exponential smoothing method and the Winter method, which produces MAPE values of 23.35% and 27.38%. These two studies show that the MAPE value is still above 10%, which means that prediction accuracy has not been categorized as very accurate.

Therefore, another method is needed to forecast hotel occupancy rates so that predictions can be made very accurately. The Decomposition, triple exponential smoothing, and Winter methods are forecasting methods for linear data [8]. If the time series data turns out to be nonlinear, the accuracy of the forecasting results becomes poor [9]. Several methods exist for nonlinear data, namely XGBoost and Support Vector Regression (SVR).

XGBoost is a method that results from the development of gradient-boosted decision tree algorithms that have reached high accuracy and processing speed results [10]. This method is commonly used for classification or prediction tasks, but with a unique processing scheme then, we can use it as a univariate forecasting method [11]. Some researchers have used the XGBoost method for univariate forecasting. One of the researchers who did this forecasting was [12]. They used the XGBoost method to forecast the Moroccan stock market by producing a MAPE value of 1.095%, which means the prediction accuracy is very accurate.

The SVR method is a development of the SVM method that can predict linear and nonlinear time series data [13]. The advantage of SVR is that this method can overcome overfitting and make predictions that will produce good precision accuracy [9]. This SVR method has been used to conduct univariate forecasting. One of the researchers who did this forecasting was Pradnyandita et al [14]. They use the SVR method to predict electronic money transactions. The evaluation result of the MAPE value obtained is 4.782%, which means that the prediction accuracy is very accurate.

Forecasting hotel occupancy rates using the XGBoost and SVR methods has never been done by other researchers. So, based on the MAPE results obtained by the XGBoost and SVR methods in other forecasting cases, our research aims to forecast star hotel occupancy rates using these two methods to get more accurate results and compare which method produces higher accuracy. Apart from that, this research is expected to provide predictive information that can help hotel management determine room rates dynamically to maximize revenue based on predicted demand.

[doi: 10.14710/JSUNIMUS. 12.1.2024.24-33]

2. LITERATURE REVIEW

2.1. XGBoost

XGBoost is a kind of ensemble method that combines several algorithms to reach better predictive performance compared to others. This method is a development of the CART method. The difference between the XGBoost method and the CART method is that the XGBoost method applies the second-order Taylor expansion to the loss function and simultaneously implements the first and second derivatives [10]. XGBoost will continue to add weak trees of different weights to the set but still make the Trees in the set as close to the rest of the predictions as possible [10]. Here are the similarities:

$$\hat{y}_i = \sum_{c=1}^{C} f_c(x_i), \quad f_c \in F \tag{1}$$

In equation (1), $\hat{y}_i$ is the predicted value, f_c is the result of the regression of the c-th tree, F is the overall result of the regression of the tree, and C is the number of regressions of the tree. Here is an equation that makes the predicted value in equation (1) expected to be as close to its actual value as possible without losing much information:

$$Obj^{(t)} = \sum_{i=1}^{n} L\left(y_i \hat{y}_i^{(t)}\right) + \sum_{i=1}^{t} \Omega(f_i) + K \tag{2}$$

In equation (2), $L\left(y_i \hat{y}_i^{(t)}\right)$is a loss function representing the difference between the predicted and actual values. The form of the function is a derivative form of the second order. $\Omega(f_i)$ is a form of regularization that defines the complexity of the model. Here are the similarities:

$$\Omega(f_i) = \gamma T + \frac{1}{2}\lambda \|\omega\|^2 \tag{3}$$

In equation (3), T is the number of leaf nodes, and ω is the weight of the leaf nodes. Using the second-order Taylor expansion, XGBoost obtained the following objective functions:

$$\begin{aligned} Obj^{(t)} &\backsimeq \sum_{i=1}^{n} \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \sum_{i=1}^{t} \Omega(f_i) \\ &= \sum_{j=1}^{T} \left[\left(\sum_{i \in I_j} g_i \right) \omega_j + \frac{1}{2} \left(\sum_{i \in I_j} h_i + \lambda \right) \omega_j^2 \right] + \gamma T \end{aligned} \tag{4}$$

In equation (4), g_i is the derivative of the first order with the formula, $g_i = \partial_{\hat{y}(t-1)} L\left(y_i \hat{y}_i^{(t-1)}\right)$ and h_i is the derivative of the second order with the formula, $h_i = \partial^2_{\hat{y}(t-1)} L\left(y_i \hat{y}_i^{(t-1)}\right)$

2.2. The SVR

SVR is a development method of SVM that is used for nonlinear regression tasks. The SVM model consists of 2 components: the kernel and the optimization method. The kernel will convert the nonlinear data into high-dimensional space, separating the data linearly [15]. The decision function for this SVR model can be expressed as [16] :

[doi: 10.14710/JSUNIMUS. 12.1.2024.24-33]

$$z = w.\psi(x) + a \tag{5}$$

In equation (5), x is the input, w and a is a constant vector, and $\psi(x)$ is a nonlinear function. The purpose of the SVR algorithm is to determine the best parameters, namely w and a, with optimization as follows:

$$\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}(\beta_i - \beta_i^*)(\beta_i - \beta_i^*)\mathcal{K} + \sum_{i=1}^{n}(\beta_i - \beta_i^*) - \sum_{i=1}^{n} y_i(\beta_i - \beta_i^*)$$
$$\text{Conditionally } \textstyle\sum_{i=1}^{n}(\beta_i - \beta_i^*) = 0\ , 0 \le \beta_i, \beta_i^*\ \le C, i = 1,2, \dots, n \tag{6}$$

In equation (6), $\mathcal{K}$ is a kernel function that can be calculated by the equation $\mathcal{K}(i,j) = \psi(x_i)^T\psi(x_j)$. From equation (6), we will get the SVR nonlinear decision function, which can be seen as follows:

$$g(x_i) = \sum_{i=1}^{n}(\beta_i - \beta_i^*)\,\mathcal{K}(x_i, x_j) + a \tag{7}$$

2.3. Model Evaluation

Model evaluation is used to see and determine whether the forecasting model created can predict very accurately. There are several model evaluation criteria, including MAE, RMSE, and MAPE. The formula used is as follows:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{8}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{9}$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \tag{10}$$

MAE and RMSE criteria are used to compare accuracy between forecasting models. The best method will be the forecasting method with the smallest MAE or RMSE. The MAPE is used to determine whether the forecasting results are accurate or not. According to Trimono et al [17], a percentage of MAPE less than 10% indicates the forecast results are very accurate, 10% to 20% of the prediction results are accurate, 21% to 50% means reasonable forecast, and above 50% means the prediction results are not accurate.

3. METHODOLOGY

The data used for this study is the occupancy rate of star-rated hotels in Bali Province. This dataset comes from the Central Bureau of Statistics (BPS). This monthly time series data was taken from January 2008 to January 2019 with 144 observations. In this study, we will process and analyze this data using the Python programming language

The measures used to forecast the occupancy rate of star hotels in Bali are as follows:

1. We are exploring time series data. At this stage, we will create and interpret a time series plot and a time series decomposition plot.

[doi: 10.14710/JSUNIMUS. 12.1.2024.24-33]

2. Perform the data modeling stage. Before the data is modeled, we will separate the data that becomes training data and the testing data. The testing data used is from January 2019 to December 2019, while the training data is the rest of the testing data, as many as 132 observations. After that, data transformation is performed with the MinMaxScaler function in Python. After the data is ready, then it is auctioned using XGBoost and continued with SVR.
3. In this step, we will forecast for next year based on data modeling training.
4. Perform a model evaluation. We will return the transformation data to the original data at this stage. After that, we will create a time series plot that compares the 3rd step with the testing data. In addition to building plots, we will also calculate and interpret MAE, RMSE, and MAPE.

4. RESULTS AND DISCUSSION

Hotel occupancy rate forecasting is used as a managerial strategy to optimize room rates according to high and low demand. The pattern of time series data on Hotel occupancy rates in Bali from 2008 to 2019 can be seen in the figure below:

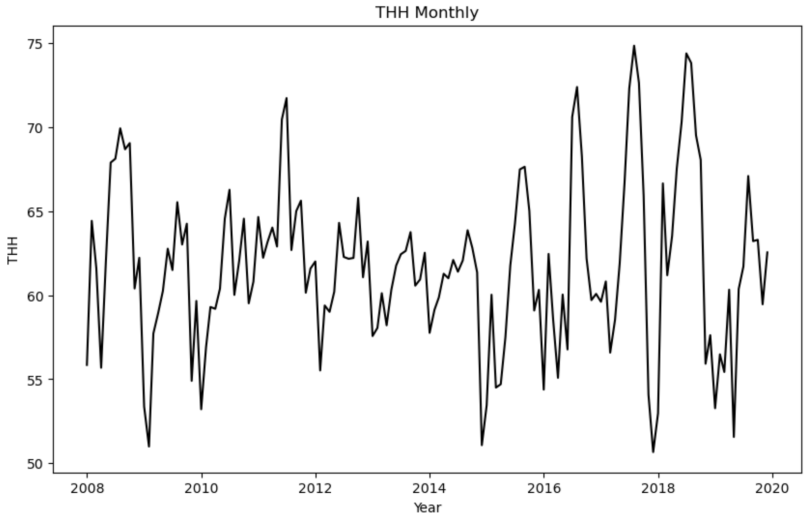

Fig 1. Time Series Plot of Hotel Occupancy Rate in 2008 – 2019

Based on Fig 1. Before 2015, the hotel occupancy rate had no trend or seasonal pattern. However, from 2015 to 2019, the data became seasonal. Based on that, we can't definitively determine the overall pattern of the data, so we need a decomposition plot. If we look at Fig 1, then we use an addictive model in creating decomposition plots because the variation in data changes is quite stable.

In Fig 2, the time series plot is separated into three components, namely trend, seasonal, and residual components. In the trend component, the pattern of hotel occupancy data shows no trend. Meanwhile, there is a pattern of data on seasonal components. As for the residual component, it can be seen that the data points are unstable in the middle of the plot, so we can guess that the data pattern is nonlinear. Based on the decomposition plot in Fig 2, we can conclude that the data on Hotel occupancy rates from 2008 to 2019 only have seasonal patterns and are not linear (non-linear).

[doi: 10.14710/JSUNIMUS. 12.1.2024.24-33]

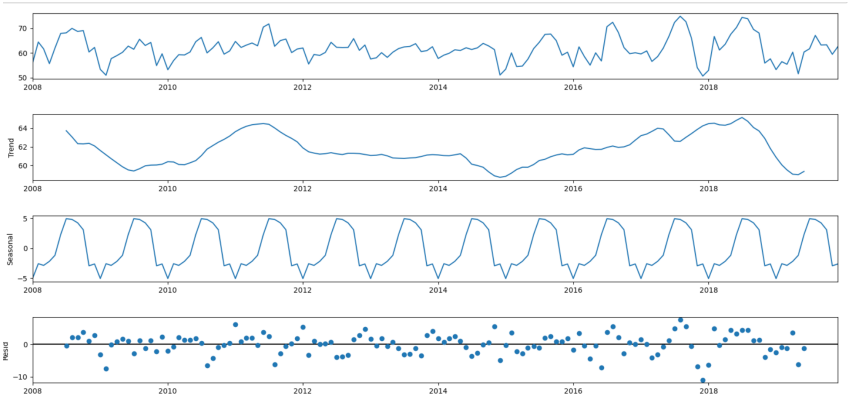

Fig 2. Decomposition Plot of Hotel Occupancy rates

The high or low interest in star-rated hotel occupancy from 2008 to 2019 can be seen per month using a box plot as shown below:

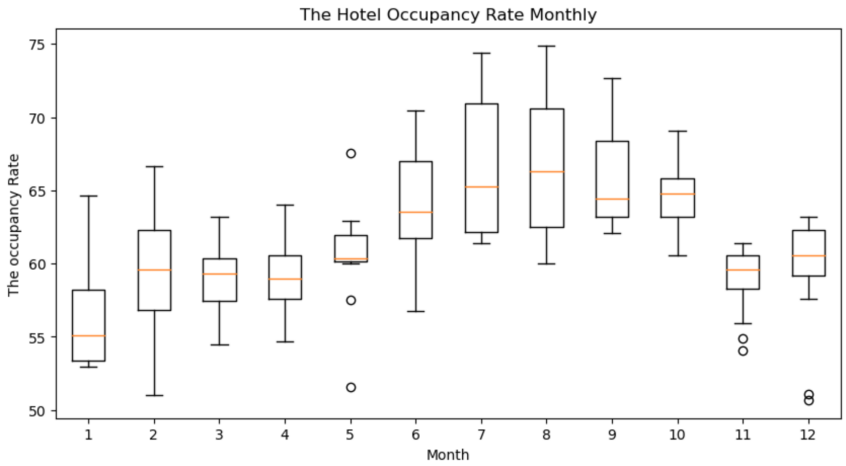

Fig 3. Box Plot of Hotel Occupancy Rate Per Month

Fig 3 shows July and August have the highest average hotel occupancy rates compared to other months. This means that star hotel enthusiasts are high in that month, and it's time for hotels to maximize their stay rates. Apart from using plots, we will look at the characteristics of the data using descriptive statistics. The results can be seen in Table 1.

Table 1. Data Characteristics

	Value
Mean	61.7161
Min	50.6600
Max	74.8600

Table 1 shows that the average percentage of tourists staying in star hotels is 61.72% with the minimum percentage being 50.66% and the maximum being 74.86%. Based on this, we can suggest that stakeholders maintain and continually improve the quality of star hotel occupancy so that the percentage continues to increase.

After exploring the data, the next step is modeling. In this step, we first separate the training data and testing data. Then, we perform MinMax Scaler transformation for each

[doi: 10.14710/JSUNIMUS. 12.1.2024.24-33]

training and testing data to get normal data. In modeling using the XGBoost method, the optimal parameters used to forecast the occupancy rate of star hotels in Bali are as follows:

Table 2. XGBoosts Model Parameters

Parameters	Value
Objective ($Obj^{(t)}$)	Reg:squarederror
Eta	0.3
Depth	3

While the optimal parameters to perform SVR modelling on the occupancy rates hotel in Bali is as follows:

Table 3. SVR Model Parameters

Parameters	Value
kernel	'RBF'
γ	0.5
C	10
ϵ	0.05
toll	0.001

Visualization of the comparison of forecasting results using the XGBoost method with testing data can be seen in the figure below

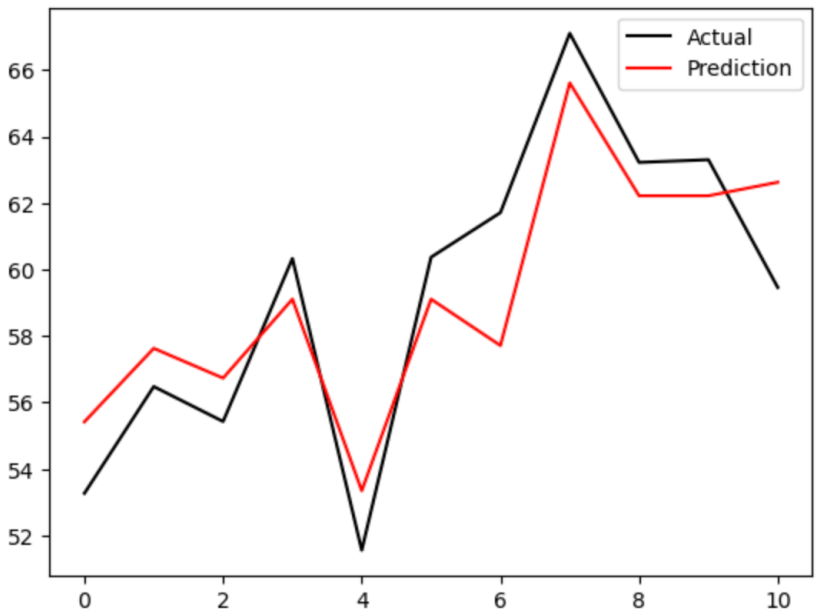

Fig 4. Visualization of Forecasting Results with Actual Value XGBoost Method

[doi: 10.14710/JSUNIMUS. 12.1.2024.24-33]

Fig 4 shows that the forecasting results using the XGBoost method produce prediction values that are almost the same as the actual values, which can be seen in the forecasting values that capture actual data patterns. Meanwhile, the comparison of forecasting results using the SVR method with testing data can be seen in the figure below. Fig 5 shows that the forecasting results using the SVR method have also produced prediction values that are almost the same as the actual values, which can also be seen in the forecasting values, which can also capture actual data patterns. However, let's compare Fig 4 and Fig 5. We can see that the SVR method produces better prediction values than the XGBoost method because the difference in forecasting results with the actual data is smaller.

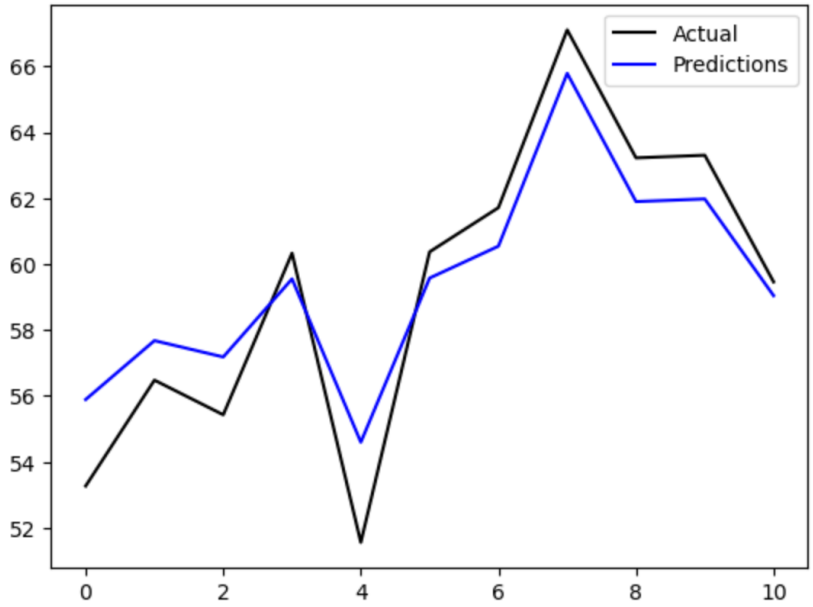

Fig 5. Visualization of Forecasting Results with Actual Value SVR Method

In addition to comparing visualizations such as Fig 4 and Fig 5, we will determine which method is best between XGBoost and SVR using the results of model evaluation criteria. The results of the model evaluation criteria can be seen in the table below:

Table 4. Model Evaluation Criteria

Criterion	XGBoost	SVR
MAE	2.0072	1.6139
RMSE	1.7833	1.4311
MAPE (%)	3.0301	2.4965

In Table 4, we can see that the RMSE, MAE, and MAPE values of the XGBoost dan SVR method are 1.7833, 2.0072, 3.03%, 1.4311, 1.6139, and 2.4965%. Based on these values, it shows that the SVR method has a value of all model evaluation criteria smaller than the XGBoost method, which means that the SVR method is better than XGBoost to forecast the occupancy rate of star hotels in Bali. The MAPE value of both methods also shows a value smaller than 10%, which means that both methods can predict the occupancy rate of star hotels in Bali very accurately.

[doi: 10.14710/JSUNIMUS. 12.1.2024.24-33]

5. CONCLUSION

Time series data on the occupancy rate of star hotels in Bali has seasonal patterns and nonlinear data types. This data shows that July and August have the highest average hotel occupancy rates, so hotels can maximize their stay rates. The forecasting results using XGBoost and SVR show that the MAPE value can predict the occupancy rate of star hotels in Bali very accurately. In addition, it can also be concluded that the SVR is the best method to forecast the occupancy rate of star hotels in Bali.

REFERENCES

[1] N. AISHAH, D. DEVIANTO, and M. MAIYASTRI, "PEMODELAN JUMLAH KUNJUNGAN WISATAWAN MANCANEGARA KE INDONESIA MELAUI BANDARA NGURAH RAI BALI DENGAN MODEL SARIMA-ARCH," *Jurnal Matematika UNAND*, vol. 10, no. 3, p. 248, Jul. 2021, doi: 10.25077/jmu.10.3.248-259.2021.

[2] B. G. Prianda and E. Widodo, "PERBANDINGAN METODE SEASONAL ARIMA DAN EXTREME LEARNING MACHINE PADA PERAMALAN JUMLAH WISATAWAN MANCANEGARA KE BALI," *BAREKENG: Jurnal Ilmu Matematika dan Terapan*, vol. 15, no. 4, pp. 639–650, Dec. 2021, doi: 10.30598/barekengvol15iss4pp639-650.

[3] BPS, "statistik-kunjungan-wisatawan-mancanegara". 2023.

[4] N. U. Clarissa, W. Sulandari, and R. Respatiwulan, "Peramalan jumlah kedatangan wisatawan mancanegara ke bali menggunakan metode hibrida SSA-WFTS," *Jurnal Ilmiah Matematika*, vol. 8, no. 1, p. 19, Apr. 2021, doi: 10.26555/konvergensi.v8i1.21460.

[5] H. Christian Anderson Wint, A. Irma Purnama, and T. Suprapti, "PREDIKSI HUNIAN HOTEL MENGGUNAKAN ALGORITMA K-NEAREST NEIGHBORS (STUDI KASUS : HOTEL RUMAH KITA KOTA CIREBON)," 2024.

[6] R. N. Darmawan, J. C. A. Wijaya, and A. P. Putra, "Peramalan Tingkat Penghunian Kamar (TPK) pada Hotel Berbintang di Banyuwangi dengan Metode Naive dan Decomposition," *G-Tech: Jurnal Teknologi Terapan*, vol. 8, no. 1, pp. 114–124, Dec. 2023, doi: 10.33379/gtech.v8i1.3543.

[7] J. Dwi Putra Tamasoleng, I. Bagus Ary Indra Iswara, J. Tukad Pakerisan No, and P. Denpasar Selatan, "Analisis Perbandingan Metode Triple Exponential Smoothing dan Metode Winter Untuk Peramalan Tingkat Hunian Hotel Aston Denpasar," *Jurnal Nasional Komputasi dan Teknologi Informasi*, vol. 3, no. 1, 2020.

[8] L. Zhang, W. Bian, W. Qu, L. Tuo, and Y. Wang, "Time series forecast of sales volume based on XGBoost," in *Journal of Physics: Conference Series*, IOP Publishing Ltd, Apr. 2021. doi: 10.1088/1742-6596/1873/1/012067.

[9] F. Yulianto, W. Firdaus Mahmudy, and A. A. Soebroto, "Comparison of Regression, Support Vector Regression (SVR), and SVR-Particle Swarm Optimization (PSO) for Rainfall Forecasting," 2020. [Online]. Available: www.jitecs.ub.ac.id

[10] C. X. Lv, S. Y. An, B. J. Qiao, and W. Wu, "Time series analysis of hemorrhagic fever with renal syndrome in mainland China by using an XGBoost forecasting model," *BMC Infect Dis*, vol. 21, no. 1, Dec. 2021, doi: 10.1186/s12879-021-06503-y.

[11] Md. S. Rahman, A. H. Chowdhury, and M. Amrin, "Accuracy comparison of ARIMA and XGBoost forecasting models in predicting the incidence of COVID-19 in Bangladesh," *PLOS Global Public Health*, vol. 2, no. 5, p. e0000495, May 2022, doi: 10.1371/journal.pgph.0000495.

[12] H. Oukhouya and K. El Himdi, "Comparing Machine Learning Methods—SVR, XGBoost, LSTM, and MLP— For Forecasting the Moroccan Stock Market," MDPI AG, Jun. 2023, p. 39. doi: 10.3390/iocma2023-14409.

[13] D. Indra Purnama and S. Setianingsih, "Support Vector Regression (SVR) Model for Forecasting Number of Passengers on Domestic Flights at Sultan Hasanudin Airport Makassar Model Support Vector Regression (SVR) untuk Peramalan Jumlah Penumpang Penerbangan Domestik di Bandara Sultan Hasanudin Makassar," vol. 16, no. 3, pp. 391–403, 2020, doi: 10.20956/jmsk.v%vi%i.9176.

[14] I. Nengah Dharma Pradnyandita and A. A. Rohmawati, "Electronic Money Transactions Forecasting with Support Vector Regression (SVR) and Vector Autoregressive Moving Average (VARMA)," *Intl. Journal on ICT*, vol. 8, no. 1, pp. 69–85, 2022, doi: 10.21108/ijoict.v8i1.632.

[15] Z. Meng, H. Sun, and X. Wang, "Forecasting Energy Consumption Based on SVR and Markov Model: A Case Study of China," *Front Environ Sci*, vol. 10, Apr. 2022, doi: 10.3389/fenvs.2022.883711.

[16] L. Rubio and K. Alba, "Forecasting Selected Colombian Shares Using a Hybrid ARIMA-SVR Model," *Mathematics*, vol. 10, no. 13, Jul. 2022, doi: 10.3390/math10132181.

[17] T. Trimono, A. Muhaimin, and N. Selayanti, "Forecasting the number of traffic accidents in Purbalingga Regency on 2023 using time series model," vol. 8, pp. 419–427, 2024, doi: 10.11594/nstp.2024.4168.