
Time Series Regression to Analyzing and Forecasting Tourist Visitation in Batang Regency

Fatma Abiyya Fawasi¹, Wiyli Yustanti²

^{1,2}*Universitas Negeri Surabaya, Surabaya, Indonesia*

fatmaabiyya.22002@mhs.unesa.ac.id, wilyliyustanti@unesa.ac.id

ABSTRACT

This study aims to develop a forecasting model for the number of tourists at five tourist destinations in Batang Regency using historical tourist visit data and external factors, as well as to implement the best model into an application. The methods used are Random Forest Regressor, XGBoost Regressor, GRU, and LSTM, with stages based on Knowledge Discovery in Databases (KDD), feature selection, hyperparameter tuning, and evaluation using MAPE, RMSE, MAE, and R2. The results show that the best model was obtained from XGB-4, with an average MAPE of 4.96%, RMSE of 526.93, MAE of 428.07, and R2 of 0.972. The features that significantly affect the model include Exponential Moving Average (EMA), trend, lagging, volatility, temperature max, and rain. The best model was implemented in an application called LARAS BATANG, which stands for Layanan Ramalan Wisata Batang. The application displays tourism forecasting results with several destination options and forecasting periods. In addition, the application can show a forecasting prototype that includes a summary, a time series line chart, a donut chart of visitor data proportions, and a feature importance bar chart.

Keyword: Forecasting, Tourist Visit, XGBoost, Machine Learning, Time series Regression

Article Info:

Article history:

Received April 16, 2026

Revised May 20, 2026

Accepted June 24, 2026

Corresponding Author

Fatma Abiyya Fawasi

Universitas Negeri Surabaya, Surabaya, Indonesia

fatmaabiyya.22002@mhs.unesa.ac.id

1. INTRODUCTION

Tourism in Indonesia has often achieved recognition at the international level. Based on the travel and Tourism Development Index (TTDI) released by the World Economic Forum (WEF) in 2024, Indonesia ranked 22nd out of 119 countries, with a score of 4.46. This achievement shows a significant improvement compared to the 2022 Travel and Tourism Development Index (TTDI), where Indonesia ranked 32nd with a score of 4.4 [1]

The increase in the number of domestic tourist in 2024 indicates that tourism in Indonesia has continued to recover after the COVID-19 pandemic in 2020. Data from April 2024, which recorded 128.58 million domestic tourists, also shows an increase of 79.55% compared to April 2020, which recorded 71.62 million. In addition, the cumulative number of domestic tourist visits from January to April 2024 reached 355.09 million, showing an increase of 102.07% compared to the same period from January to April 2020, which recorded 175.72 million visits [2].

The number of tourists in Indonesia has generally increased after the COVID-19 pandemic. However, Batang Regency showed a decline. In 2023, the number of tourists in Batang Regency reached 635.851, placing it 26th out of 35 regencies. Meanwhile in 2020, the number of tourists reached 679.117, placing it 13th [3]. This condition shows that the success of tourism is not only related to the number of tourists attractions, but also to the influence of external factors. One important factor is considered by tourists is weather conditions.

Tourism is a sector that is often affected by weather conditions, such as temperature, rainfall, sunny weather, and other related factors. Tourism is also vulnerable to changes climate and weather [4]. A total of 69.5% of tourist respondents at Silgar and A Lanzada beaches in Spain stated that they decided to visit the beach based on weather forecast, especially sunny weather (43%) and the absence of rain (37.5%).

Forecasting the number of tourists at Jiuzhaigou Mountain, China, showed the lowest evaluation value using the LSTM method, with a MAPE value of 6.67 [6]. Another study conducted at Jiuzhaigou and Siguniang Mountains found that the VMD-GRU model was the best forecasting model [7]. Meanwhile, the XGBoost model showed the highest R^2 value of 0.82 in predicting customers for a tourism company in Switzerland [8].

The urgency of this study is strengthened by the declining trend in the number of tourist the COVID-19 pandemic in Batang Regency. Based on the number of tourists in Batang Regency from 2020 to 2023, there was a decrease of 6.37% [3]. This condition highlights the need for a system that can predict the number of tourists and provide recommendations for decision-making related to post-pandemic tourist recovery. The use of predictive methods, such as Random Forest Regressor, XGB Regressor (eXtreme Gradient Boosting Regressor), Gated Recurrent Unit (GRU), and Long Short-Term Memory (LSTM), can be innovation to support data-based decision-making.

Through this study, it is expected that the best method for predicting the number of tourists based on historical data can be identified. The model is also expected to improve the accuracy of tourist number predictions in the coming years. In addition, the model approach can help identify external factors that influence the number of tourists. Therefore, this study can serve as an initial step for the Batang Regency Tourism Office to strengthen data-based decision-making in an effort to increase the number of tourists after the pandemic.

2. METHODS

This study is a quantitative study with a data mining approach to develop a forecasting model for the number of tourist visitors by following the Knowledge Discovery in Databases (KDD) stages. The data used in this study consist of historical tourist visitor data and weather data from five destinations, namely *Kolam Renang Bandar*, *Taman Hiburan Rakyat Kramat*, *Pantai Ujungnegero*, *Pantai Sigandu*, and *Resort Pagilaran*. The research stages include data selection, data preprocessing, data transformation, data mining, evaluation, identification of influential factors, and model implementation. In the preprocessing stage, several processes were carried out, including temporal resampling of weather data, identification of missing values, duplicate data, outliers, and data exploration to understand trend patterns and relationships between variables. The transformation stage was conducted through feature engineering by adding lagging, trend, volatility, and Exponential Moving Average (EMA) features. In the modelling stage, this study compared Random Forest Regressor, XGB Regressor (eXtreme Gradient Boosting Regressor), Gated Recurrent Unit (GRU), and Long Short-Term Memory (LSTM) in four scenarios, namely the use of all features, hyperparameter tuning, feature selection, and a combination of feature selection and hyperparameter tuning. The model evaluation was carried out using MAPE, RMSE, MAE, and R^2 . The best model was

then used to analyze the factors influencing the prediction and was implemented into a web-based application using Streamlit.

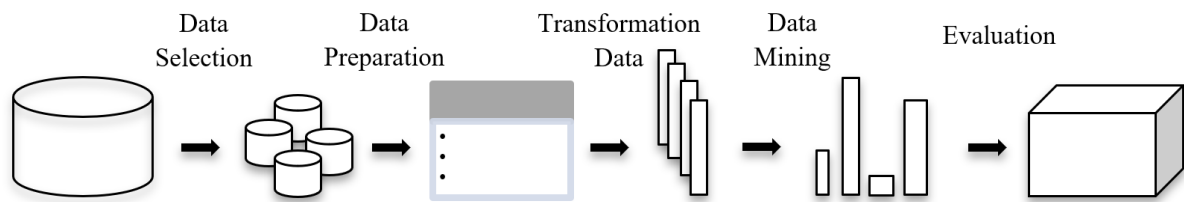


Figure 1. KDD Research Framework

3. RESULTS AND DISCUSSION

This section presents the research results and discussion based on the systematic methods and stages explained in the previous section. The discussion includes data selection, data preprocessing, data transformation, modelling, model evaluation, selection of the best model, analysis of influential features, and the implementation of the model into a web-based application.

3.1 Data Selection

This study used historical data on tourist visitors and weather in Batang Regency, Central Java Province. The tourist visitor data included the number of visits to five tourists destinations. This data was arranged in a monthly time series pattern.

Table 1. Tourist Destination Data

Destination	Rentang Waktu	Number of Visitor Data Rows
Kolam Renang Bandar (KRB)	01/01/1999 s.d 01/12/2024	312
Taman Hiburan Rakyat Kramat (THRK)	01/01/1999 s.d 01/12/2024	312
Pantai Ujungnegoro (PU)	01/01/2004 s.d 01/12/2024	252
Pantai Sigandu (PS)	01/01/2004 s.d 01/12/2024	252
Resort Pagilaran	01/03/2004 s.d 01/12/2024	250

The weather data has a daily time series pattern and consists of features X_1 , X_2 , X_3 , X_4 , X_5 , and X_6 . These features were used in this study to predict the number of tourist visitors.

Table 2. Research Feature Description

Features	Description
n month	Time Series Index
y	Monthly Tourist Visits
X_1	Average Daily Minimum Temperature
X_2	Average Daily Maximum Temperature
X_3	Clear Days Total
X_4	Cloudy Days Total
X_5	Drizzle Days Total
X_6	Rainy Days Total

3.2 Preprocessing Data

Data preprocessing was conducted to prepare the raw data before it was used in the modelling process. In this stage, the daily weather data was transformed into monthly data through temporal resampling. Then, the tourist visitor data and weather data were combined based on the same time period and destination name. The missing value identification showed that there were 163 missing values in the y feature. These missing values occurred due to several factors, such as extreme weather, the Covid-19 pandemic, agrarian conflict, renovation, and unrecorded data. To handle this problem, the missing values were filled using linear interpolation. The outlier identification process found 77 outliers in the dataset. The highest number of outliers appeared during the Eid al-Fitr period, with 42 outliers. Other outliers were found during school holidays, national holidays, and the year-end holiday period.

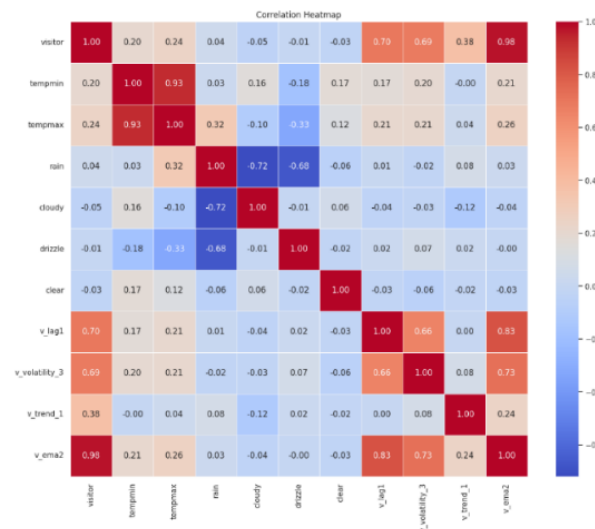


Figure 2. Correlation Heatmap

Based on data exploration using a correlation heatmap, X_{10} had the highest correlation with tourist visitors, with a correlation value of 0.98. It was followed by X_7 and X_9 , with correlation values of 0.70 and 0.69. Meanwhile, X_2 , X_1 , and X_6 showed low correlations with tourist visitors, with values of 0.24, 0.20, and 0.04. The lowest correlations were found in X_4 , X_3 , and X_5 , which showed negative correlation values of -0.05, -0.03, and -0.01.

3.3 Transformation Data

The transformation stage included several processes, namely feature engineering, data normalization, sliding window, and data splitting.

Table 3. Feature Engineering Results

n bulan	destinasi	y	X_7	X_8	X_9
1	KRB	2500	500	900	1001,7
2	KRB	1500	2500	1000	1000
3	KRB	2800	1500	300	680,7
4	KRB	2200	2800	700	650,6

Feature engineering was conducted by adding temporal information to the model. In this process, several new features were added, including X_7 as lagging, X_8 as trend, X_9 as volatility, and X_{10} as Exponential Moving Average.

```

1 scaled
array([[0.05352586, 0.34015608, 0.3715148 , ..., 0.7741935 , 0.35000002,
        0.03846154],
       [0.043835 , 0.5817175 , 0.6082139 , ..., 0.53333336, 0.42166668,
        0.15897436],
       [0.043835 , 0.6321492 , 0.54483795, ..., 0.04301075, 0.45 ,
        0.6153846 ],
       ...,
       [0.21030247, 0.64900637, 0.46571255, ..., 0.08602151, 0.25 ,
        0.7307693 ],
       [0.20857964, 0.6395745 , 0.48362207, ..., 0.08888889 , 0.11166668,
        0.8346154 ],
       [0.14074361, 0.44852448, 0.43255472, ..., 0.04301075, 0.35000002,
        0.6923077 ]], dtype=float32)

```

Figure 3. Normalization Result

Data normalization was applied using MinMaxScaler to change all feature values into a range between 0 and 1. After that, the time series data was transformed into a supervised learning format using the sliding window method. The window size used in this study was 6 or 12, depending on the scenario.

Table 4. Data Splitting Results

Destination	Length Train Data	Length Test Data
KRB	298	12
PS	238	12
PU	238	12
RP	236	12
THRK	298	12

Data splitting was performed using a time-based split to divide the dataset into training data and testing data. For all five tourist destinations, the testing data covered 12 months, from January 2024 to December 2024. Meanwhile, the length of the training data was different for each destination because each destination had a different data time range.

3.4 Data Mining and Evaluation

The data mining stage was the main process for applying prediction models in this study. The implementation was carried out using the Python programming language and Google Colab. Several modelling scenarios were used as the basis for the experiments.

Table 5. Modelling Scenarios

(Feature, Tuning)	Random Forest Regressor	XGB Regressor	Gated Recurrent Units (GRU)	Long ShortTerm Memory (LSTM)
X1 -X10, No	RF-1	XGB-1	GRU-1	LSTM-1
X1 -X10, Grid Search	RF-2	XGB-2	GRU-2	LSTM-2
Feature Selection, No	RF-3	XGB-3	GRU-3	LSTM-3
Feature Selection, Grid Search	RF-4	XGB-4	GRU-4	LSTM-4

These are the models for each methods referred to table 5.

1. Random Forest Regressor

The Random Forest Regressor method was implemented by using the RandomForestRegressor() function from the sklearn.ensemble library. The modelling process was divided into four scenarios, from RF-1 to RF-4. In RF-1 and RF-3, the model was trained using fixed parameters, namely n estimators = 100, max depth = 3, and min

samples leaf = 2. Meanwhile, in RF-2 and RF-4, hyperparameter tuning was performed using Grid Search to find the best parameter combination for each tourist destination.

Table 6. Random Forest Regressor Evaluation Results

Model	R2	RMSE	MAE	MAPE
RF 1	0,83	1.322,92	1.093,95	21,68
RF 2	0,72	1.574,51	1.244,23	32,10
RF 3	0,85	1.251,70	1.058,73	20,66
RF 4	0,78	1.467,10	1.214,83	28,32

The average evaluation results show that the best Random Forest scenario is RF-3 with MAPE 20.66%, R2 0.853, RMSE 1,251.70, MAE 1,058.73.

2. XGB Regressor (eXtreme Gradient Boosting Regressor)

The XGBoost Regressor method was implemented by using the XGBRegressor() function from the xgboost library. In XGB-1 and XGB-3, the model was trained using the initial parameters, namely n estimators = 500, max depth = 8, min child weight = 4, and reg lambda = 0. Meanwhile, in XGB-2 and XGB-4, hyperparameter tuning was applied using Grid Search to obtain the best parameters for each tourist destination.

Table 7. XGB Regressor Evaluation Results

Model	R2	RMSE	MAE	MAPE
XGB 1	0,934	750,72	609,11	10,38
XGB 2	0,958	596,19	481,69	8,05
XGB 3	0,954	647,87	520,57	8,49
XGB 4	0,972	526,93	428,07	4,96

The results show that the best scenario for the XGBoost Regressor model is XGB-4 with MAPE 4.96%, R2 0.972, RMSE 526.93, MAE 428.07.

3. Gated Recurrent Units (GRU)

The Gated Recurrent Unit (GRU) method was applied using a Sequential architecture. The architecture consisted of GRU, Dropout, Dense, and Dense (1) output layers for regression. In GRU-1, the model used an input shape of (12, 10), a GRU hidden layer with 64 units, a dropout value of 0.2, and a dense hidden layer with 32 neurons. GRU-2 used all features and applied hyperparameter tuning to find the best combination of window size, number of units, dropout value, and dense layer size. The tested window sizes were 6 and 12, the number of units was 32 or 64, the dropout values were 0.2 and 0.3, and the dense layer sizes were 16 and 32 neurons. The evaluation results of the GRU model are presented in Table 8.

Table 8. Gated Recurrent Units (GRU) Evaluation Results

Model	R2	RMSE	MAE	MAPE
GRU 1	0,637	1260,87	1125,92	31,87
GRU 2	0,860	815,32	684,51	19,25
GRU 3	-0,366	1015,64	885,95	26,38
GRU 4	0.942	610.60	503.87	14.22

The results show that the best GRU scenario is GRU-4 with MAPE 14.22%, R2 0.942, RMSE 610.60, MAE 503.87.

4. Long Short-Term Memory (LSTM)

The Long Short-Term Memory (LSTM) method was also applied using a Sequential architecture. The architecture consisted of LSTM, Dropout, Dense, and Dense

(1) output layers for regression. In LSTM-1, the model used an input shape of (12, 10), an LSTM hidden layer with 64 units, a dropout value of 0.2, and a dense hidden layer with 32 neurons. LSTM-2 used all features with hyperparameter tuning to find the best combination of window size, number of units, dropout value, and dense layer size. The tested window sizes were 6 and 12, the number of units was 32 or 64, the dropout values were 0.2 and 0.3, and the dense layer sizes were 16 and 32 neurons. Meanwhile, LSTM-3 used selected features without hyperparameter tuning, while LSTM-4 used selected features with hyperparameter tuning.

Table 9. Long Short-Term Memory (LSTM) Evaluation Results

Model	R2	RMSE	MAE	MAPE
LSTM-1	0,552	1465,34	1093,00	29,03
LSTM-2	0,472	1536,78	1267,69	31,66
LSTM-3	-0,455	1581,57	1343,91	39,49
LSTM-4	0,507	1114,92	943,24	33,25

The results show that the LSTM scenario with the lowest average evaluation value is LSTM-1 with MAPE 29.03%, R2 0.552, RMSE 1465.34, MAE 1093.

3.5 Best Forecasting Model Scenario

The best forecasting model was selected by comparing the performance of all models based on the lowest average evaluation value.

Table 10. Best Model Performance

Model	R2	RMSE	MAE	MAPE
XGB-4	0,972	526,93	428,07	4,96

Based on Table 10, the best method for predicting the number of tourist visitors at five tourist destinations in Batang Regency was the XGBoost Regressor with the XGB-4 scenario. The results show that all scenario with the lowest average value is XGB Regressor with scenario model XGB-4 MAPE 4.96%, R2 0.972, RMSE 526.93, MAE 428.07. This proves that the combination of SHAP (Shapley Additive Explanations) feature selection and Grid Search tuning may produce the best prediction model with the lowest evaluation value.

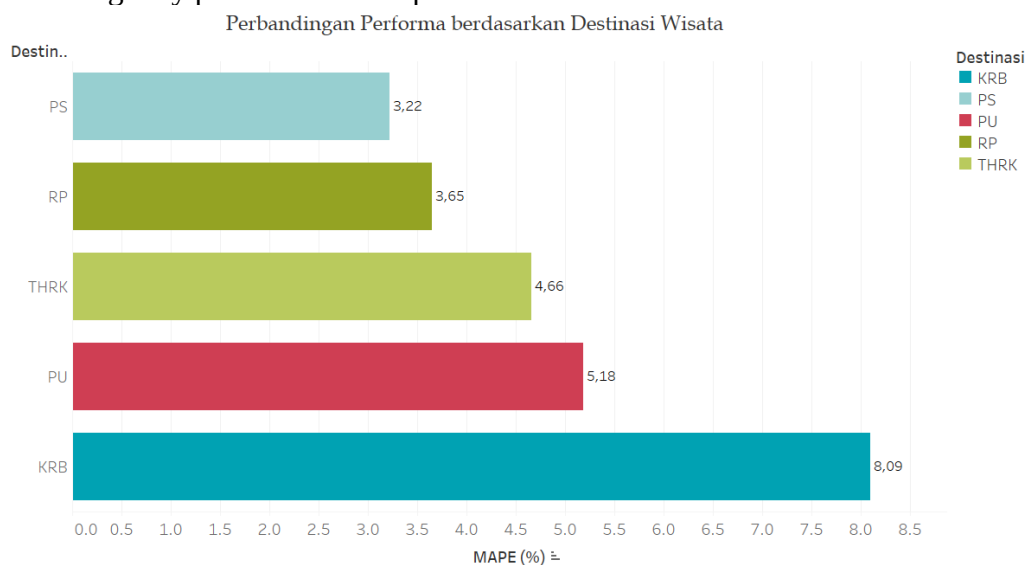


Figure 4. XGB-4 Performance by Destination

Figure 4 shows the performance of the XGB-4 model for each tourist destination. The best result was obtained at Resort Pagilaran (RP), which had the lowest MAPE value of

3.22%. Pantai Sigandu (PS) with a MAPE evaluation value of 3.65%. Next, Taman Hiburan Rakyat Kramat (THRK) obtained a MAPE value of 4.66%. The highest evaluation results were found at Pantai Ujungnegoro (PU) and Kolam Renang Bandar (KRB), with MAPE values of 5.18% and 8.09%.

3.6 Feature Importance Results of Forecasting Model

The influential factors were analyzed using feature importance and feature selection results.

Table 11. Dominant Features of XGB-4

Destination	Dominant Factor
KRB	X ₁₀ , X ₈ , X ₇ , X ₉ , X ₂ .
PU	X ₁₀ , X ₈ , X ₇ , X ₉ .
PS	X ₁₀ , X ₇ , X ₈ , X ₉ , X ₂ , X ₆ .
RP	X ₁₀ , X ₈ , X ₇ , X ₉ .
THRK	X ₁₀ , X ₇ , X ₈ .

Based on Table 11, the XGB-4 scenario with SHAP feature selection in XGBoost showed that historical visitor features had a higher contribution than weather variables. The results show that X₁₀, X₈, and X₇ were selected in all five destinations. This was followed by X₉, which was selected in four destinations. Meanwhile, the least selected features were X₂ and X₆.

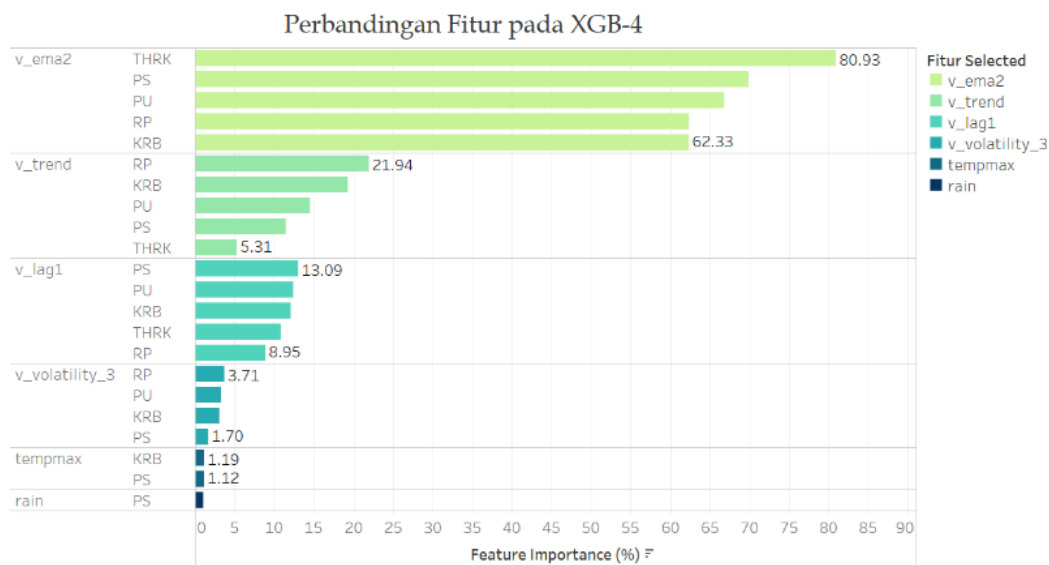


Figure 5. Feature Comparison of XGB-4

Based on Figure 5, X₁₀ had the highest contribution among all features. The highest contribution of X₁₀ was found at Taman Hiburan Rakyat Kramat (THRK), with a value of 80.93%. The X₈ feature showed the highest contribution at Resort Pagilaran (RP), with a value of 21.94%. In addition, X₇ contributed to Pantai Sigandu (PS), with a value of 13.09%. Meanwhile, the features with the lowest contribution were X₉, X₂, and X₆.

3.7 Forecasting Model Result Implementation

The best model, XGB-4, was implemented in a web-based application using Streamlit. The application was named LARAS BATANG, which stands for *Layanan Ramalan Wisata Batang*. This application was developed to support *Dinas Pariwisata dan Olahraga Kabupaten Batang* in preparing strategies to increase the number of tourist visitors in Batang Regency.

ISSN:

1. Home Page



Figure 6. Homepage of LARAS Batang

The home page presents the description, objectives, and main features of the LARAS BATANG application.

2. Destination Page

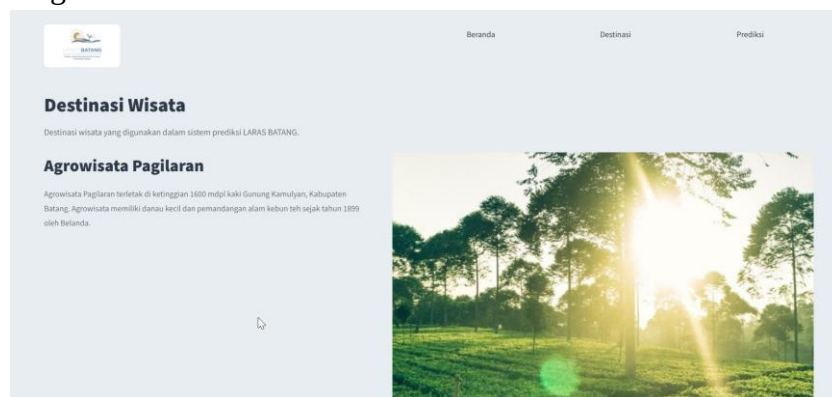


Figure 7. Tourist Destination Page

The destination page provides photos and brief descriptions of the tourist destinations.

3. Forecasting Page

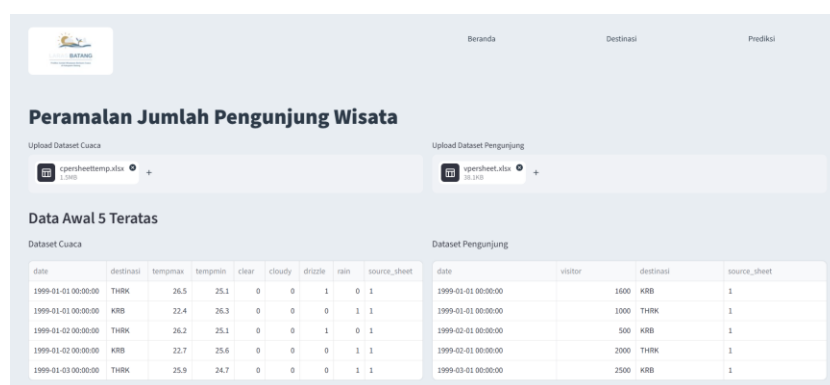


Figure 8. Forecasting Page

On the forecasting page, users can upload visitor and weather datasets. The prototype then displays a dataset summary, model evaluation results, and visualization outputs. These outputs include a line chart, donut chart, and feature importance bar chart.

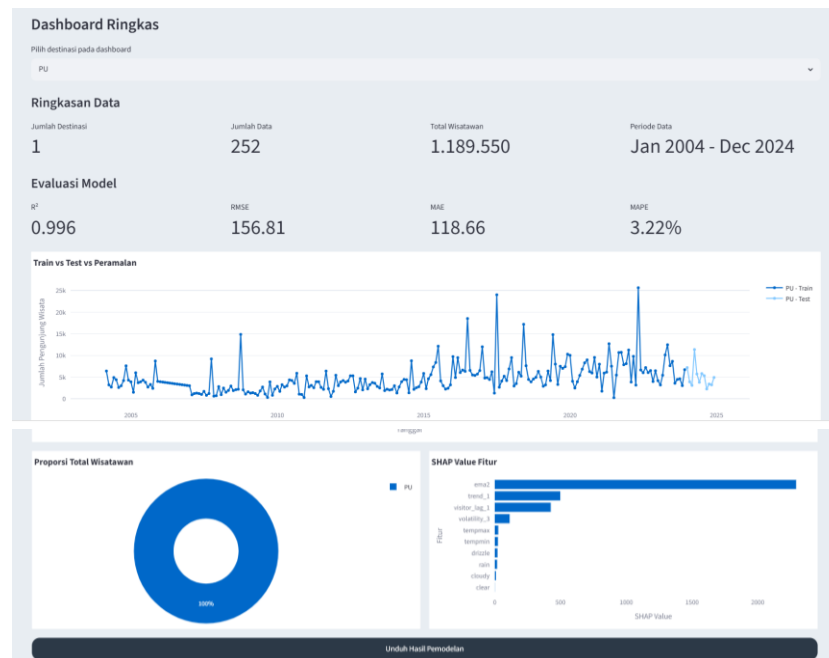


Figure 9. Forecasting Model Prototype Results

The prototype displays a dataset summary, modelling evaluation results in the form of a line chart, donut chart, and feature importance bar chart.

CONCLUSION

This study aimed to predict the number of tourist visitors in Batang Regency, Central Java Province. The results showed that the best model was XGBoost with the XGB-4 scenario, which used SHAP feature selection and hyperparameter tuning. This result indicates that the XGBoost model with SHAP feature selection and hyperparameter tuning produced better prediction performance than Random Forest, GRU, and LSTM on the dataset used in this study.

The most influential factors in predicting the number of tourist visitors were historical visitor features. Exponential Moving Average (EMA), trend, lag, volatility, maximum temperature, and the number of rainy days appeared as dominant features in most tourist destinations.

The best model was implemented into a web-based application using Streamlit. The LARAS BATANG application can display the data processing stage, evaluation result visualization, important features, and tourist visitor predictions for future periods.

This study is expected to become a reference for *Dinas Pariwisata dan Olahraga Kabupaten Batang* in making data-based strategies to increase the number of tourist visitors more accurately. The application system can be improved by adding historical storage for prediction results, more interactive visualizations, and a more user-friendly interface. Future research is also expected to develop the XGB Regressor model further by using a longer data time range.

REFERENCES

- [1] F. Betti and I. Tussyadiah, *Travel & Tourism Development Index 2024*. Geneva, Switzerland: World Economic Forum, 2024.
- [2] Badan Pusat Statistik, *Perkembangan Pariwisata April 2024*. Jakarta, Indonesia: Badan Pusat Statistik, 2024.
- [3] Dinas Pariwisata Provinsi Jawa Tengah, *Buku Pariwisata Jawa Tengah Dalam Angka*. Semarang, Indonesia: Pemerintah Provinsi Jawa Tengah Dinas Kepemudaan, Olahraga dan Pariwisata, 2024.
- [4] Z. Pololikashvili, *Tourism and Rural Development: Understanding Challenges on the Ground Lessons Learned from the Best Tourism Villages by UNWTO Initiative*. Madrid, Spain: United Nations World Tourism Organization, 2023.
- [5] D. R. Toubes, N. Araújo-Vila, and J. A. Fraiz-Brea, "Influence of weather on the behaviour of tourists in a beach destination," *Atmosphere*, 2020.
- [6] J.-W. Bi, Y. Liu, and H. Li, "Daily tourism volume forecasting for tourist attractions," *Annals of Tourism Research*, 2020.
- [7] C. Zhang and Y.-X. Tian, "Forecast daily tourist volumes during the epidemic period using COVID-19 data, search engine data and weather data," *Expert Systems with Applications*, 2022.
- [8] S. Lionetti, D. Pfäffli, M. Pouly, T. Vor der Brück, and P. Wegelin, "Tourism forecast with weather, event, and cross-industry data," in *Proceedings of the 13th International Conference on Agents and Artificial Intelligence*, 2021.
- [9] M. I. Ananda, "Multivariate forecasting harga daging ayam dan sapi melibatkan faktor cuaca, ekonomi, dan kesehatan menggunakan GRU," *Jurnal Ilmu Komputer Agri-Informatika*, 2023.
- [10] M. Chhetri, S. Kumar, P. P. Roy, and B.-G. Kim, "Deep BLSTM-GRU model for monthly rainfall prediction: A case study of Simtokha, Bhutan," *Remote Sensing*, 2020.
- [11] P. L. Seabe, C. R. B. Moutsinga, and E. Pindza, "Forecasting cryptocurrency prices using LSTM, GRU, and bi-directional LSTM: A deep learning approach," *Fractal and Fractional*, 2023.
- [12] S. Kirsan, et al., "EksPy: A new Python framework for developing graphical user interface based PyQt5," 2024.
- [13] R. Struniawski and R. Kozera, "TfELM: Extreme Learning Machines framework with Python and TensorFlow," *SoftwareX*, 2024.
- [14] W. Yustanti and N. Rochmawati, "Analisis algoritma klasifikasi untuk memprediksi karakteristik mahasiswa pada pembelajaran daring," *Jurnal Edukasi dan Penelitian Informatika*, 2022.
- [15] N. Rochmawati, et al., "Analisa learning rate dan batch size pada klasifikasi Covid menggunakan deep learning dengan optimizer Adam," *Journal of Information Engineering and Educational Technology*, vol. 5, no. 2, 2021.
- [16] Y. Zhuo, et al., "Prediction model for spontaneous combustion temperature of coal based on PSO-XGBoost algorithm," *Scientific Reports*, vol. 15, no. 1, 2025.

ISSN:

- [17] N. Ganesh, P. Jain, A. Choudhury, P. Dutta, K. Kalita, and P. Barsocchi, "Random forest regression-based machine learning model for accurate estimation of fluid flow in curved pipes," *Processes*, vol. 9, no. 11, 2021.
- [18] E. Fitri, "Analisis perbandingan metode regresi linier, Random Forest Regression dan Gradient Boosted Trees Regression Method untuk prediksi harga rumah," *Journal of Applied Computer Science and Technology*, vol. 4, no. 1, 2023.