International Journal on Perceptive and Cognitive Computing (IJPCC) Vol 11, Issue 1 (2025)
https://doi.org/10.31436/ijpcc.v11i1.539

Predictive Analytics for Sustainable Tourism Development: A Data-Driven Approach

Irfan Qayyim Abdul Mohaimin, Aida Najihah Mohd Marzuki, Raini Hassan*
Department of Computer Science, Kulliyyah of Information and Communication Technology,
International Islamic University Malaysia, Gombak, Malaysia

*Corresponding author hrai@iium.edu.my
(Received: 30th July 2024; Accepted: 7th December, 2024; Published on-line: 30th January, 2025)

Abstract— Tourism plays a significant role in Malaysia's economic and social development, with efforts increasingly aligned to Sustainable Development Goals (SDGs), particularly SDG 8 (Decent Work and Economic Growth). This project addresses the lack of predictive analytics for sustainable tourism by employing a structured methodology encompassing data collection and preparation, exploratory data analysis (EDA), descriptive and predictive analytics, and feature engineering to identify key factors influencing sustainable tourism in Malaysia. The results show trends and patterns in tourism that inform the development of robust machine learning models to forecast sustainable tourism outcomes, in which 92% and 100% accuracy were achieved with Gradient Boosting and Support Vector Machine respectively. These models aim to support data-driven decision-making and promote long-term sustainability in Malaysia's tourism industry.

Keywords— Predictive analytics, sustainable tourism, machine learning, data science, Malaysia.

I. Introduction

Tourism is a rapidly growing industry and a key driver of economic development, fostering regional and national growth, foreign exchange, and social advancement. In Malaysia, the government has consistently supported the sector since its formal recognition in 1959, with initiatives like the Sustainable Tourism Recovery Project (2022) aligning with the National Tourism Policy 2020-2030 to enhance community resilience and promote eco-tourism [1].

However, despite Malaysia's emphasis on sustainable tourism and its alignment with Sustainable Development Goals (SDGs), the sector lacks robust predictive analytics tools to assess and plan for sustainable tourism outcomes effectively. For instance, in 2019, Malaysia welcomed 26.1 million international tourists, contributing RM86.14 billion to the economy [2]. Yet, the COVID-19 pandemic led to a significant decline in tourist arrivals and receipts, highlighting the sector's vulnerability to external shocks and the need for predictive tools to enhance resilience.

A. *Project Objectives*

- Identify key factors influencing sustainable tourism development in Malaysia using statistical and machine learning techniques.
- Provide data-driven insights to support informed decision-making and strategies for integrating sustainability into the tourism sector.
- Develop a machine learning model to predict sustainable tourism outcomes based on analysed factors.

B. *Project Questions*

- What are the critical factors driving sustainable tourism in Malaysia?
- How can predictive analytics enhance decision-making in sustainable tourism?
- What actionable strategies can be derived to balance economic, environmental, and social sustainability?

C. *Contributions to Data Science and Machine Learning*

- This study demonstrates the application of ML in sustainable tourism, offering methods to analyse complex relationships between tourism activities and sustainability outcomes.
- It will equip policymakers and businesses with tools to forecast trends, optimize resources, and minimize environmental and community impacts.
- It integrates predictive analytics into sustainable tourism practices, expanding applications of machine learning in this underexplored domain.
- Supports informed, proactive strategies for long-term resilience and sustainability in Malaysia's tourism industry.

This project addresses challenges in integrating advanced technologies into the tourism sector, aiming to balance growth with sustainability and delivering impactful tools and insights for policymakers, businesses, and researchers.

II. Related Works

Recent studies on sustainable tourism have explored key factors, advanced analytics, and machine learning

https://doi.org/10.31436/ijpcc.v11i1.539

techniques to better understand and predict trends. These approaches offer valuable insights into tourism patterns and help create models to guide decision-making. Table 1 showcases some of the notable works, highlighting their methods, findings, and practical contributions to the field.

TABLE I
REVIEW OF PREVIOUS WORKS

Year	Authors	Research Problem	Techniques Used	Result
2020	Mai, A., Thi, K., Thi, T., & Le, T.	Factors for sustainable tourism in Vietnam	SPSS, Smart-PLS-SEM	Social engagement is most important
2020	Nasir, N. F., Nasir, M. A., Nasir, M. N. F., & Nasir, M. F.	Understanding of domestic tourism in Malaysia	Qualitative study	Public health and safety are top priorities for travel
2022	Agrawal, R., Wankhede, V. A., Kumar, A., et al.	Role of BDA in sustainable tourism	Literature review, network analysis	Insights into trends and collaborations
2022	Hoffmann, F. J., Braesemann, F., & Teubner, T.	Predicting sustainability using TripAdvisor data	Machine learning, grid search	Identified patterns of sustainability
2024	Louati, A., Louati, H., Alharbi, M., et al.	Predicting tourist spending	Decision Trees, Random Forest, KNN, SVM, ARIMA	Spending trends identified

III. METHODOLOGY

The framework of this project is based on the data science lifecycle which usually consists of data collection, data preprocessing, exploratory data analysis (EDA), model building and machine learning, model evaluation, and data visualization. The key steps involved in the research process are outlined in the flowchart below (see Figure 1).

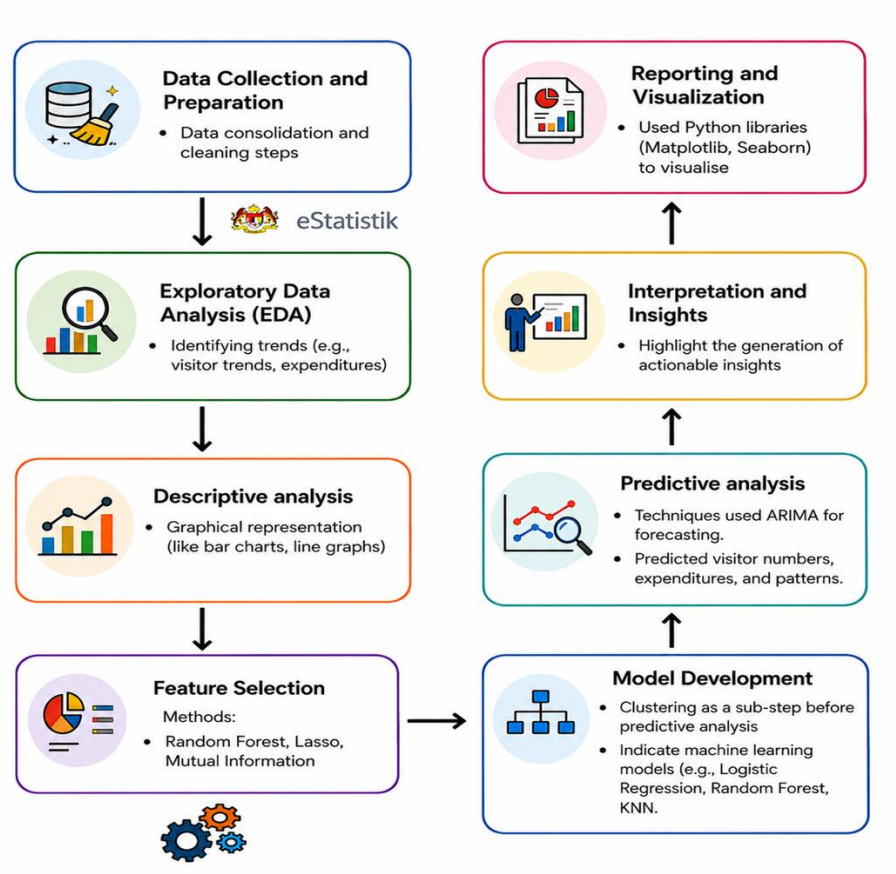

Fig. 1 The methodology flowchart

https://doi.org/10.31436/ijpcc.v11i1.539

A. *Data Collection*

The Domestic Tourism Survey (DTS), published by the Department of Statistics Malaysia (DOSM) on eStatistik, provides annual data on domestic tourism by state from 2011 to 2023. It includes metrics such as visitor numbers, trips, expenditures, and demographic profiles. DOSM also offers the data in Excel sheets for easier analysis. For this project, data from 108 Excel sheets and reports were consolidated, covering 41 features and 112 entries related to social, economic, and tourism factors (see Figure 2).

B. *Exploratory Data Analysis (EDA)*

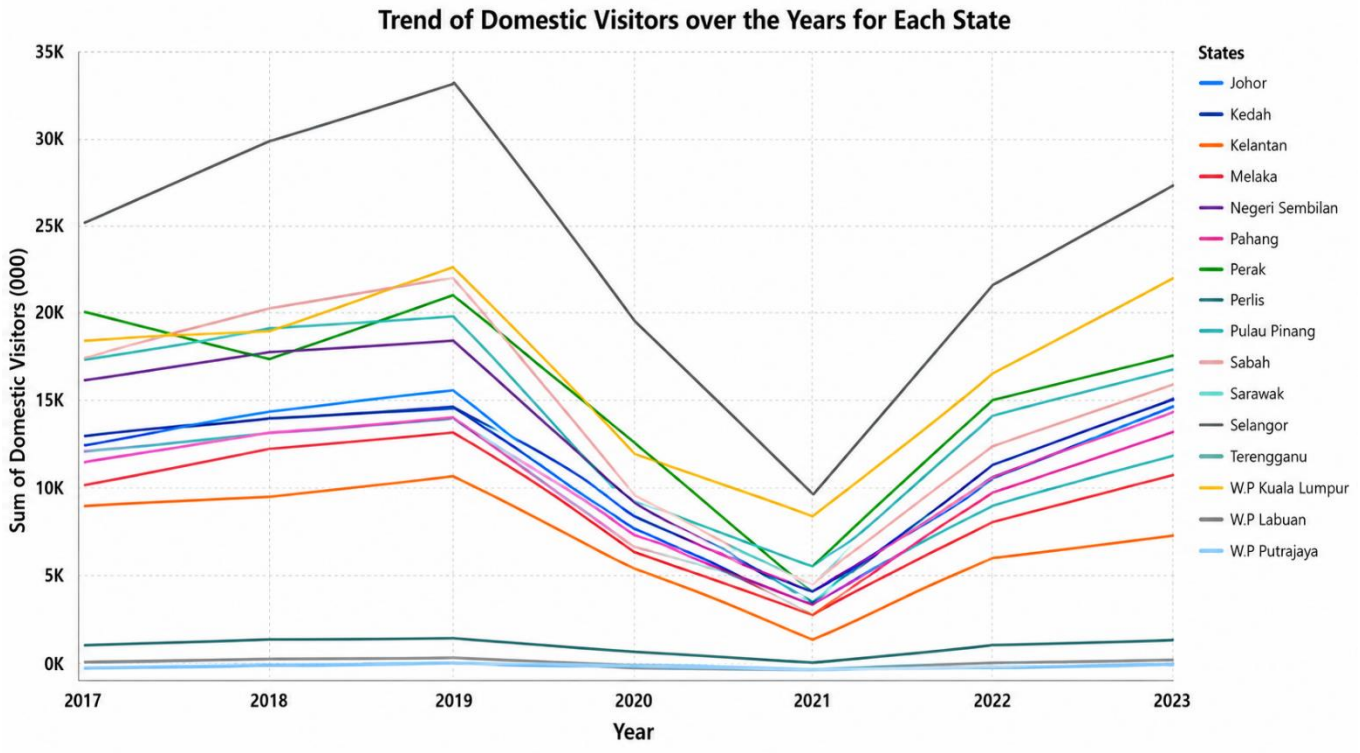

Fig. 2 Trend of domestic visitors over the years for each state

Domestic visitors increased from 2017 to 2019 across most states but dropped sharply in 2020 due to COVID-19. Recovery began in 2021 but remained below 2019 levels. Selangor had the most visitors, while Perlis, W.P. Putrajaya, and W.P. Labuan consistently recorded the lowest numbers, showing minimal pandemic impact due to their smaller size (see Figure 3).

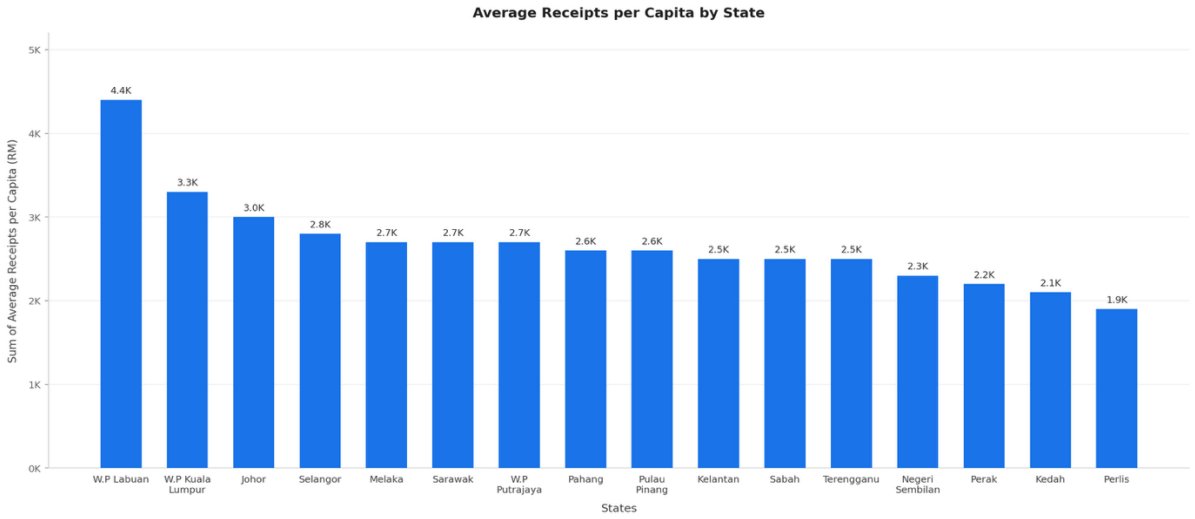

Fig. 3 Top 3 average receipts per capita by state

https://doi.org/10.31436/ijpcc.v11i1.539

W.P. Labuan leads in average receipts per capita, surpassing RM600. W.P. Kuala Lumpur and Johor rank second and third, respectively. Perlis records the lowest average receipts, under RM300, reflecting varying spending patterns across states (see Figure 4).

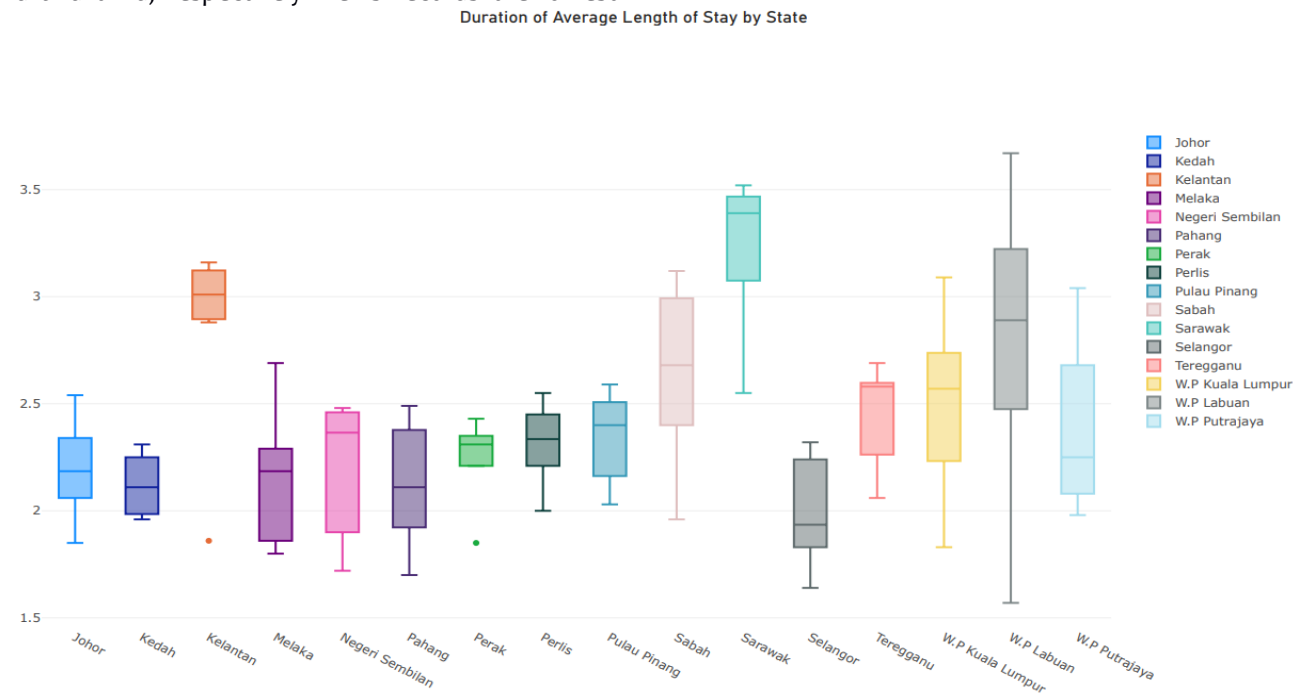

Fig. 4 Distribution of average length of stay by state

Sarawak has the highest median stay (3–3.5 days), while Kelantan shows consistent durations with minimal variability. W.P. Labuan exhibits significant variability with a wide range of stay lengths, reflecting diverse travel behaviours (see Figure 5).

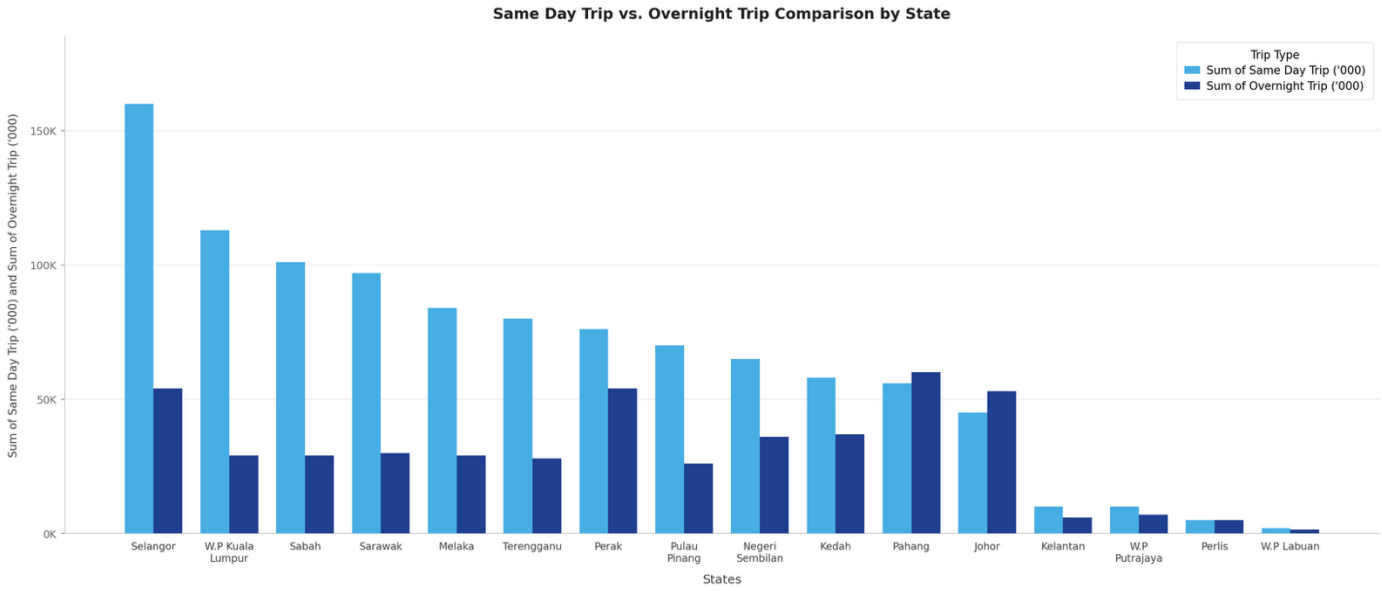

Fig. 5 Same day trip vs. overnight trip comparison by state

Selangor dominates both same-day and overnight trips, with same-day trips outnumbering overnight stays. Conversely, Johor, Pahang, and W.P. Labuan see more overnight trips, indicating longer stays. Kelantan and Pahang display balanced distributions between the two trip types (see Figure 6).

https://doi.org/10.31436/ijpcc.v11i1.539

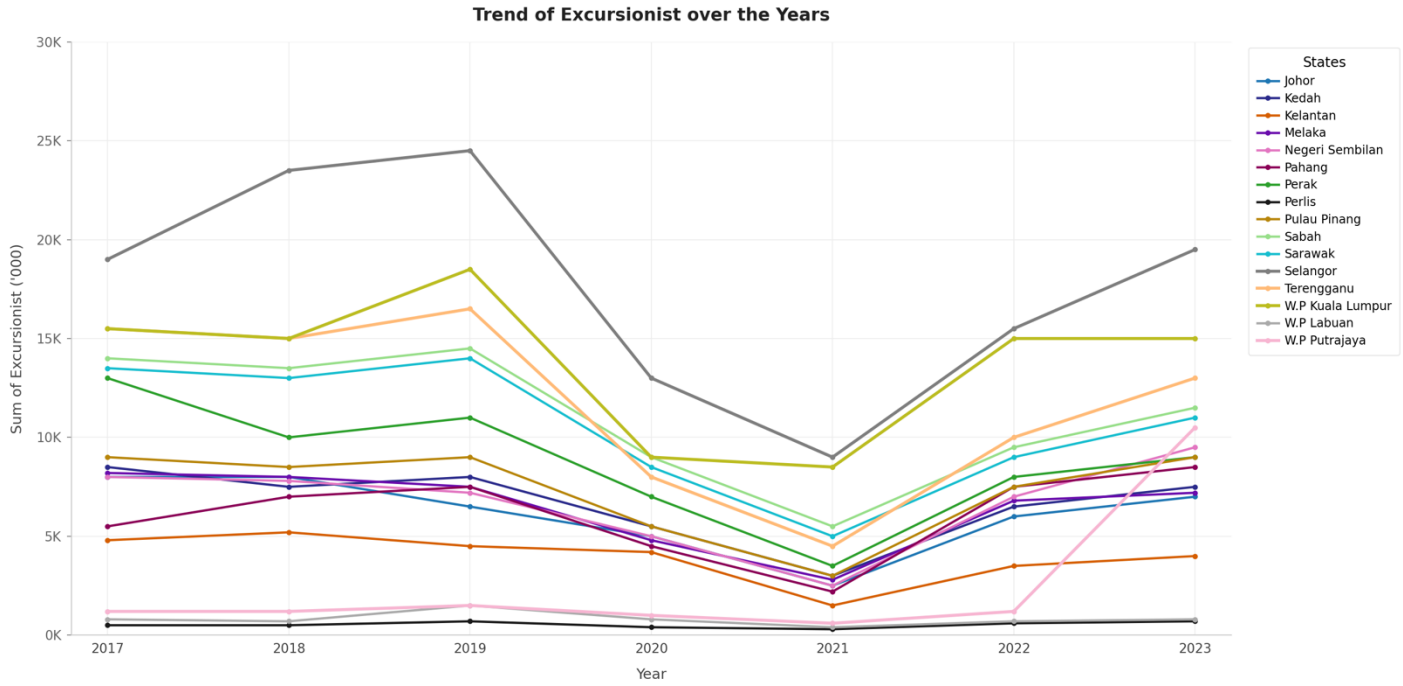

Fig. 6 Trend of excursionists over the years

Excursionist numbers declined in 2020 due to the pandemic but began recovering by 2022. Selangor and W.P. Kuala Lumpur consistently recorded the highest numbers, with Sarawak also showing relatively high but fluctuating figures (see Figure 7).

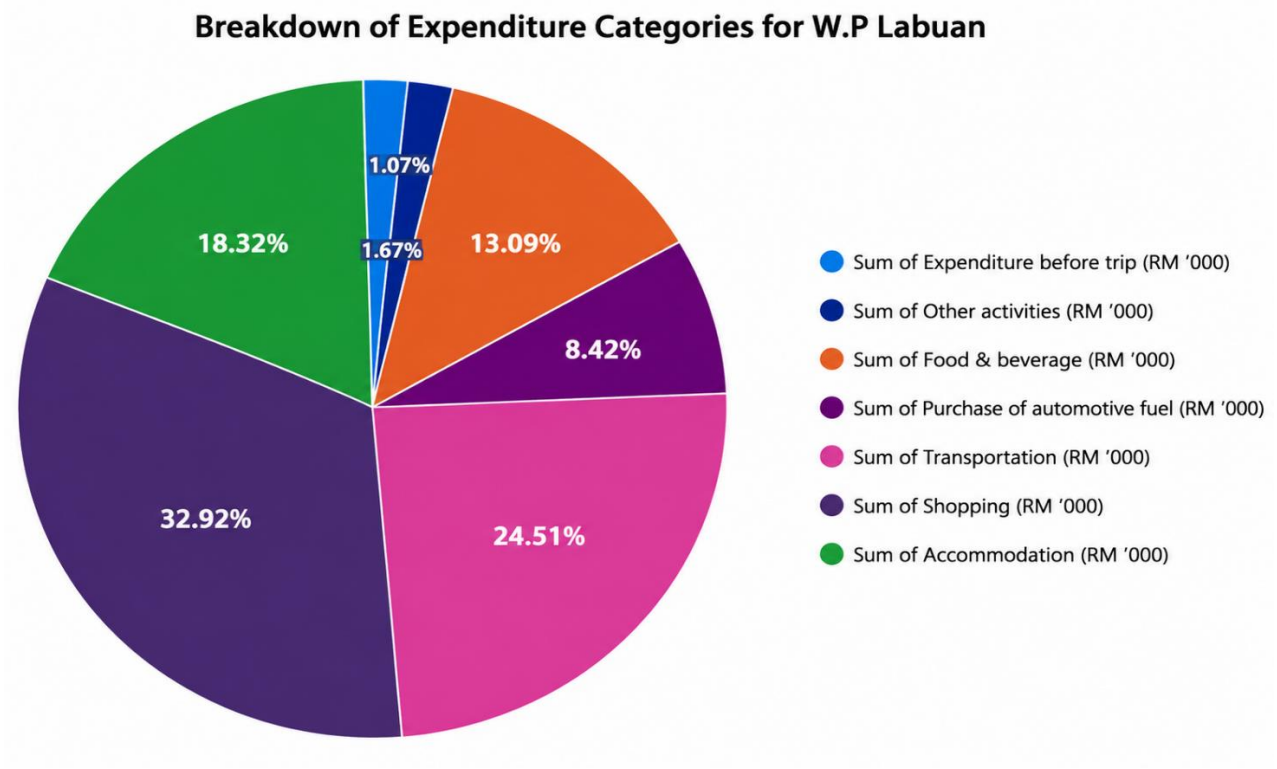

Fig. 7 Breakdown of expenditure categories for W.P Labuan

https://doi.org/10.31436/ijpcc.v11i1.539

W.P. Labuan's spending patterns are unique, with the highest share allocated to transportation (31.4%), followed by accommodation (26.2%), and lower spending on food and beverages (13.0%). This contrasts with other states, where shopping dominates expenditure categories (see Figure 8).

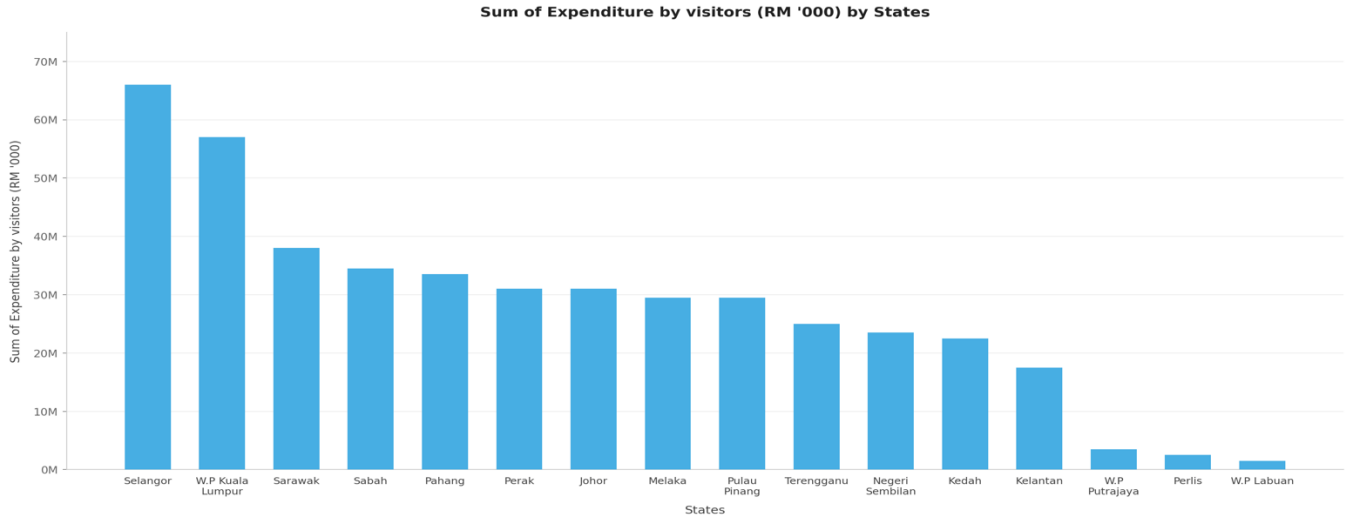

Fig. 8 Total expenditure by visitors for each state in Malaysia

Selangor leads in total visitor expenditure, surpassing RM5 billion, followed by W.P. Kuala Lumpur. W.P. Labuan, Perlis, and W.P. Putrajaya record the lowest expenditures, each under half a billion ringgit.

C. *Descriptive Analytics*

W.P. Labuan is the most expensive tourist destination in Malaysia, with the highest average receipt per capita and per trip. Tourists spend primarily on transportation, accommodation, and food & beverages due to its appeal as a luxury getaway offering activities like diving and island hopping. Limited-service availability further raises costs. Labuan attracts more overnight tourists, with high-end accommodations like Dorsett Grand Labuan catering to affluent visitors.

Selangor leads Malaysia's tourism sector, generating over RM50 billion in revenue over six years. It ranks first in domestic visitors, trips, and same-day visits, driven by its strategic location near W.P. Kuala Lumpur and well-connected transport networks. Attractions range from cultural landmarks like Batu Caves to urban hubs like Sunway Lagoon, supported by a vibrant culinary scene and growing staycation popularity.

Sarawak ranks third in excursionist and same-day trip numbers despite its geographic disconnection from mainland Malaysia. Affordable air travel and efficient infrastructure facilitate short trips. Unique attractions like the Mulu Caves, Dayak cultural experiences, and adventure activities make Sarawak a top destination for eco-tourism and memorable excursions.

D. *Data Preprocessing*

A structured data preprocessing approach was implemented to ensure data quality and consistency for analysis.

1) *Handle missing data*

The dataset contained several columns representing percentages related to tourism activities. Upon analysis, it was found that approximately 14.29% of the data in these columns were missing. To address this, mean imputation was applied, a technique that replaces missing values with the mean of the respective column. This approach was chosen as it assumes that the data's missingness is random and ensures that no rows or columns are dropped, thereby preserving the dataset's integrity. This approach ensured that the dataset remained reliable and complete, ready for subsequent analysis and modelling.

2) *Feature selection*

To identify the most relevant features, multiple feature selection techniques were employed to ensure a robust and comprehensive analysis which are as below:

- *Random Forest Feature Importance*
- *Mutual Information Regression*
- *Lasso Regularization*
- *Recursive Feature Elimination (RFE)*
- *Correlation Heatmap*

This multi-method approach ensures the selection of a highly predictive and diverse set of features, capturing both statistical and model-driven insights for downstream analysis and modelling. After analysis, the final selected features were States, Year, Domestic Visitors ('000),

https://doi.org/10.31436/ijpcc.v11i1.539

Average Receipts per Capita (RM), and Average Length of Stay (see Figure 9).

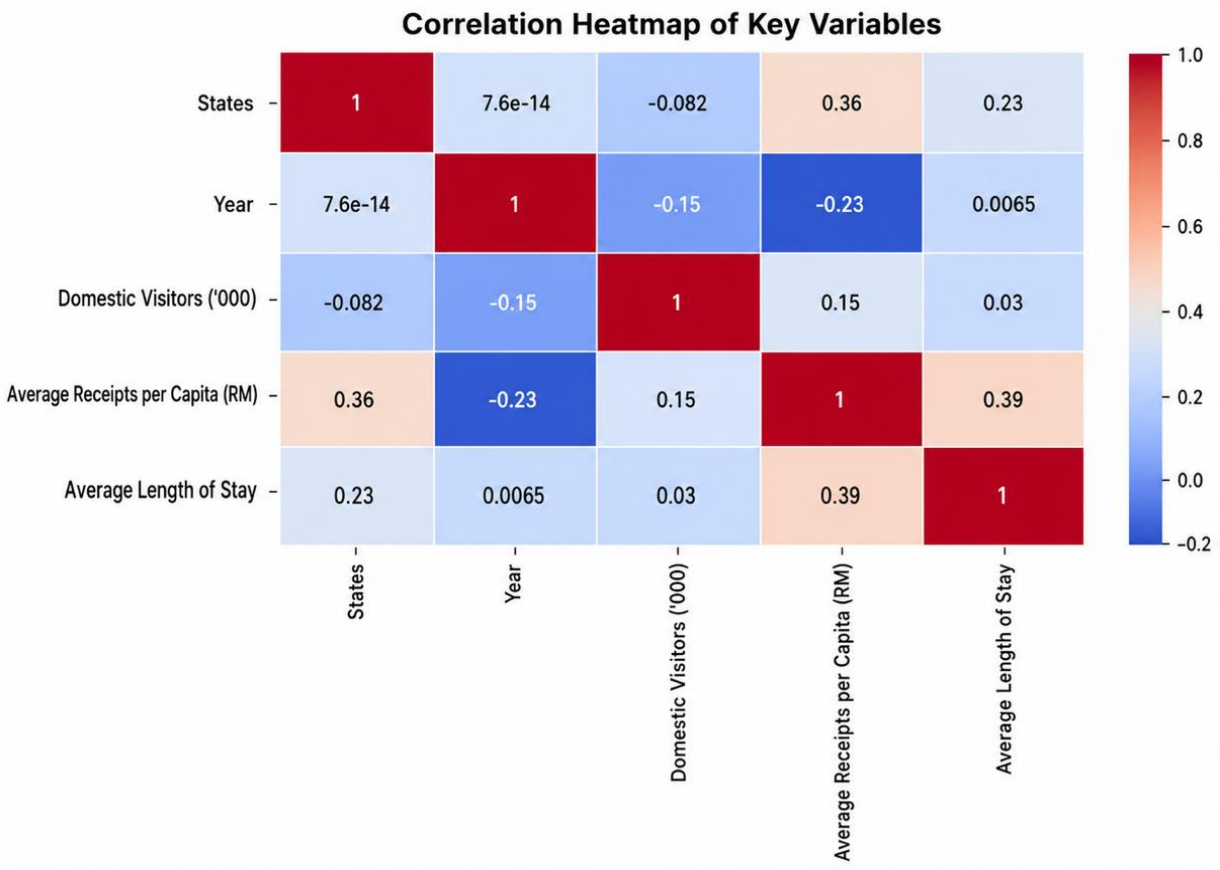

Fig. 9 Correlation heatmap of selected features

3) Clustering

Clustering was performed to uncover underlying patterns and group the data into meaningful clusters. The clustering analysis aimed to segment the dataset into homogeneous groups, which would later serve as the target variable for supervised learning tasks. Additionally, the ideal result for clustering would be that each entry for each state belong to their respective clusters only. This is to ensure consistency and accuracy in further analysis. Hence, cluster reassignment will be performed to data entries of each state that do not align with their supposed designated clusters. The approach employed K-Means clustering, a widely used method for partitioning data into non-overlapping clusters based on their similarity.

K-Means clustering was chosen for its simplicity, scalability, and ability to efficiently handle large datasets. It partitions data into distinct, non-overlapping clusters based on similarity, providing clear and interpretable groupings for analysis. Alternative methods like hierarchical clustering were less suitable due to their computational intensity and limited scalability for large datasets, making K-Means the optimal choice.

a) Features

- Domestic Visitors ('000): The number of domestic tourists visiting each state.
- Average Receipts per Capita (RM): The revenue generated per tourist.
- Average Length of Stay: The average number of days spent by tourists. These features capture the economic and behavioural aspects of tourism, making them ideal for segmentation.

b) Preparation

- Dataset was split to before, during, and after COVID-19
- Pre COVID-19: 2017-2019
- During COVID-19: 2020-2021
- Post COVID-19: 2022-2023

c) Elbow Method

https://doi.org/10.31436/ijpcc.v11i1.539

The optimal number of clusters was determined using the Elbow Method, which involves plotting the within-cluster sum of squares (WCSS) against the number of clusters. A "bend" in the curve was observed, suggesting that five clusters (before and after COVID-19) and four clusters (during COVID-19) were optimal (see Figure 10-15).

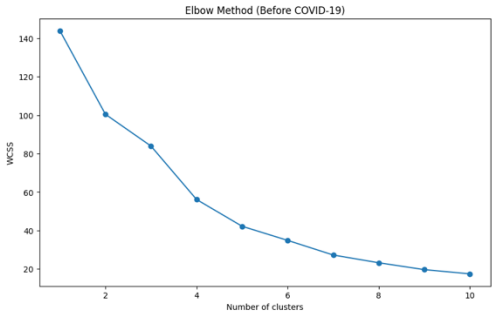

Fig. 10 Elbow method for before COVID-19

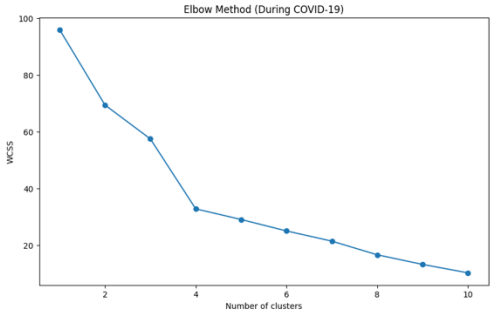

Fig. 11 Elbow method for during COVID-19

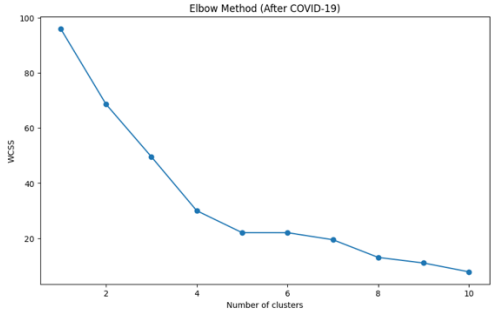

Fig. 12 Elbow method for after COVID-19

d) Clustering Execution

- Before and After COVID-19: The data from 2017-2019 (pre-COVID) and 2022-2023 (post-COVID) were clustered separately, each with five clusters.
- During COVID-19: Data from 2020-2021 were clustered using four clusters, reflecting the irregular tourism patterns during the pandemic.
- W.P. Labuan (Cluster 3) was identified as an outlier due to its distinct characteristics, including low domestic visitors and high receipts per capita with an extended length of stay.

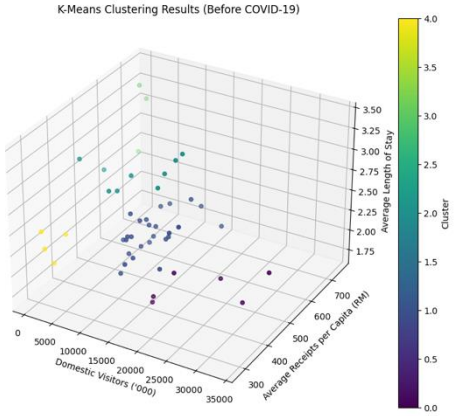

Fig. 13 K-Means clustering results (before COVID-19)

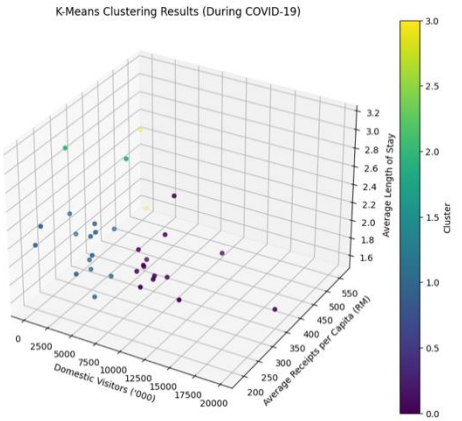

Fig. 14 K-Means clustering results (during COVID-19)

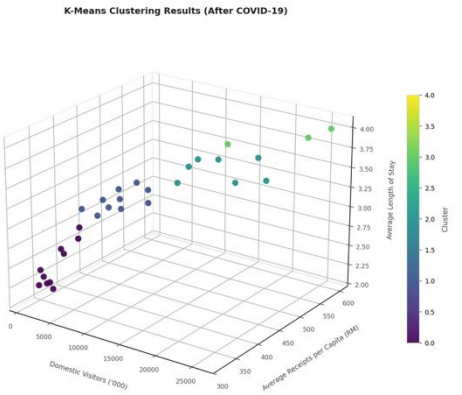

Fig. 15 K-Means clustering results (after COVID-19)

e) Sankey Diagram

- Sankey diagrams were used to determine which entries from each state belong to which cluster.
- Before and After COVID-19: Most data entries consistently belong to respective cluster for each state

https://doi.org/10.31436/ijpcc.v11i1.539

- During COVID-19: A large amount of data belongs to cluster 0 or 1 only, showing that states with supposed different characteristics are now grouped up together, indicating unreliability
- During COVID-19 data will be disregarded from the dataset for future machine learning purposes as its massive inconsistency in cluster assignment is deemed unreliable
- Before and After COVID-19 data is combined (see Figure 16-18)

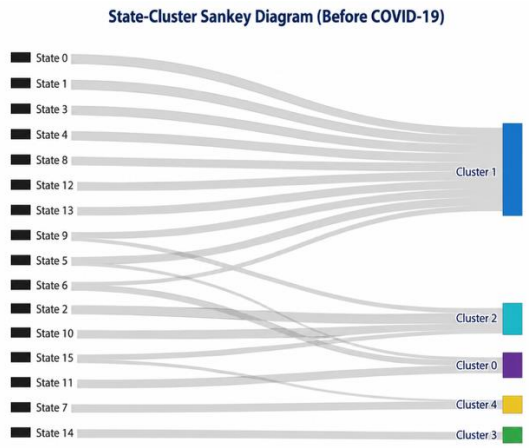

Fig. 16 State-cluster Sankey diagram (before COVID-19)

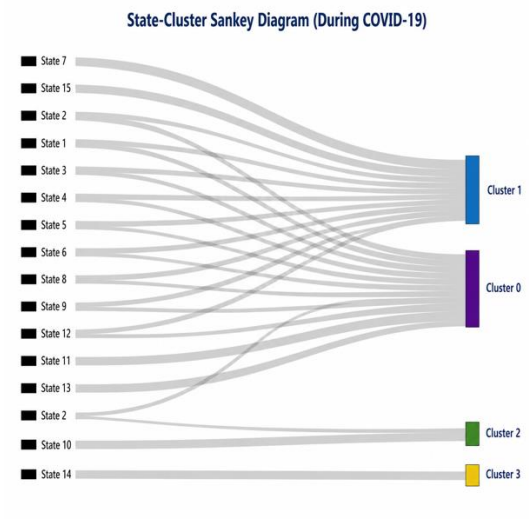

Fig. 17 State-cluster Sankey diagram (during COVID-19)

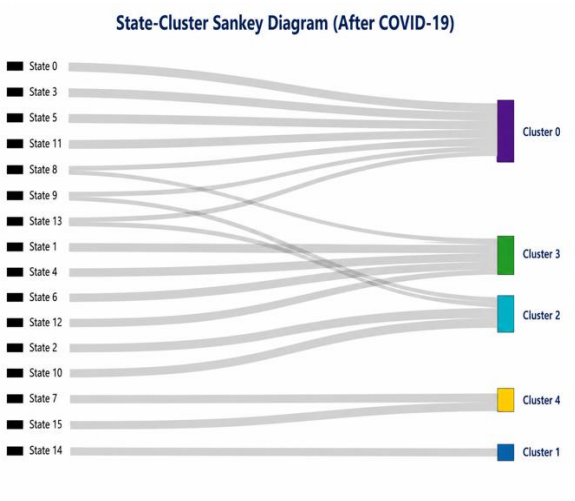

Fig. 18 State-cluster Sankey diagram (after COVID-19)

f) Cluster Reassignment

- States with a singular data entry grouped into a different cluster will be assigned to its majority cluster (see Figure 19-21)
- This includes states 3, 6, 9, 10, 13, 14
- States with many data entries grouped into different clusters will undergo cluster reassignment by Euclidean distancing
- This includes states 0, 8, 12, 15

States	Distance
0	3718.111823
8	3418.959858
12	2885.900126
15	6796.179174

Fig. 19 Euclidean distancing for states 0, 8, 12, and 15

States	0
0	0
8	2
12	2
15	1

Fig. 20 Cluster reassignment for states 0, 8, 12, and 15

https://doi.org/10.31436/ijpcc.v11i1.539

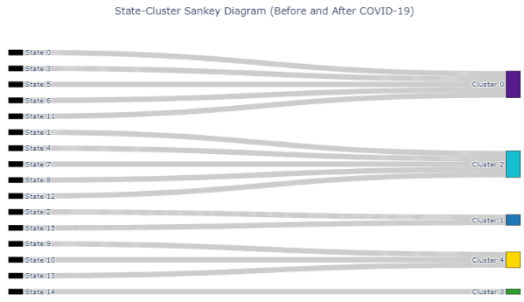

Fig. 21 State-cluster Sankey diagram after removing before COVID-19 and cluster reassignment

g) Cluster Characteristics

Cluster 0 (Purple):

- Mid to high domestic visitors
- Mid to high average receipts per capita
- Low average length of stay
- States: Pahang, Selangor, Melaka, Perak, Johor

Cluster 1 (Blue):

- Low domestic visitors
- High average receipts per capita
- Mid average length of stay
- States: Kelantan, W.P Putrajaya

Cluster 2 (Cyan):

- Low to mid domestic visitors
- Mid average receipts per capita
- Low to mid average length of stay
- States: Kedah, Negeri Sembilan, Perlis, Pulau Pinang, Terengganu

Cluster 3 (Green):

- Low domestic visitors
- High average receipts per capita
- Mid to high average length of stay
- States: W.P Labuan

Cluster 4 (Yellow):

- Low to mid domestic visitors
- High average receipts per capita
- Low to mid average length of stay
- States: Sabah, Sarawak, W.P Kuala Lumpur

h) Potential Limitations of K-Means Clustering

- K-Means depends on initial centroid placement, which can lead to suboptimal results. This was addressed using "k-means++" for better initialization
- The algorithm assumes spherical clusters, which may not match real-world data. Outliers like W.P. Labuan challenged this assumption
- K-Means is sensitive to outliers, as seen with W.P. Labuan, consistently grouped in a separate cluster (Cluster 3)

E. Model Development

The clustering labels (Cluster 0 to Cluster 4) obtained from the prior K-Means analysis were used as the target variable for classification tasks. Special attention was given to Cluster 3 (W.P. Labuan), identified as a potential outlier due to its unique characteristics.

1) Data Preparation

The dataset underwent preprocessing to ensure consistency and reliability. Two separate datasets were used:

- With Cluster 3 (C3): Included all clusters from the K-Means analysis
- Without Cluster 3 (No C3): Excluded Cluster 3 to evaluate its impact as an outlier

2) Machine Learning Methods

Six supervised learning algorithms were employed for cluster classification, each chosen for their strengths in handling specific data characteristics: Logistic Regression (LR), a probabilistic model effective for linearly separable data, was selected due to the somewhat linear relationships among features in the dataset, despite its tendency to overfit in high dimensions. Decision Trees (DT), Random Forest (RF), and Gradient Boosting (GB)were included for their robustness and ability to handle the distinct characteristics and varied data patterns across states, making them well-suited for the complex and heterogeneous nature of tourism data. K-Nearest Neighbors (KNN) was selected as a distance-based algorithm that could serve as a benchmark, while Support Vector Machines (SVM) was chosen for its effectiveness in high-dimensional spaces and its ability to capture complex decision boundaries, providing a versatile comparison to tree-based methods.

3) Grid Search for Hyperparameter Tuning

To optimize model performance, GridSearchCV was used to systematically explore hyperparameters for each model. The hyperparameters were chosen based on best practices and exploratory data analysis. The grid search process identified the best parameter combinations for improving accuracy and generalizability.

F. Predictive Analytics

ARIMA is a popular method for forecasting time series data, combining autoregression, integration, and moving average components. Autoregression links current values to past values, integration stabilizes the series and moving average smooths fluctuations. By analysing historical data, ARIMA optimizes parameters to forecast future values,

https://doi.org/10.31436/ijpcc.v11i1.539

making it effective for predicting complex, unstable patterns.

IV. Results

The results of the machine learning models offer critical insights into the application of predictive analytics in Malaysia's sustainable tourism sector. The evaluation of each model, conducted on datasets with and without Cluster 3 (representing W.P Labuan), highlights the impact of outliers on model performance and the effectiveness of hyperparameter optimization through Grid Search.

A. *Performance of Machine Learning Models*

TABLE II
Machine Learning Model Accuracies

Acc.	Machine Learning Model					
	LR	DT	RF	GB	KNN	SVM
C3	0.83	0.88	0.88	0.92	0.79	0.75
No C3	0.87	0.74	0.74	0.78	0.83	0.87
GC3	0.88	0.71	0.83	0.75	0.83	0.96
No GC3	0.91	0.78	0.74	0.78	0.83	1.00

B. *Insights from Results*

The SVM model achieved the best overall performance with an accuracy of 100% on the dataset without Cluster 3 after tuning. The optimal parameters for this model were {'C': 10, 'gamma': 'scale', 'kernel': 'rbf'}. The Gradient Boosting model also performed exceptionally well, achieving an accuracy of 92% on the dataset with Cluster 3 before tuning. Its best parameters were {'learning_rate': 0.1, 'max_depth': 10, 'min_samples_leaf': 2, 'min_samples_split': 2, 'n_estimators': 100, 'subsample': 0.8}.

The performance of SVM is highly influenced by hyperparameter tuning and the inclusion of Cluster 3 (C3). Without C3 and with hyperparameter tuning, SVM achieves perfect classification accuracy (100%), as the absence of noisy and overlapping data simplifies the dataset. Including C3 with tuning slightly reduces accuracy to 96%, as the added complexity introduces noise. Without tuning, performance drops significantly to 87% for datasets without C3 and 75% for those with C3, highlighting the critical role of tuning in optimizing SVM to the dataset's characteristics. Default parameters fail to handle overlapping features and outliers effectively, leading to suboptimal results.

Gradient Boosting demonstrates its strength in handling complex datasets like C3 due to its iterative learning approach. With default parameters, it achieves the highest accuracy (92%) for C3, effectively capturing patterns in noisy data. However, hyperparameter tuning for C3 reduces accuracy to 75%, likely due to overfitting to noise and outliers. When C3 is excluded, the dataset becomes simpler, and Gradient Boosting performs moderately, achieving 78% accuracy regardless of tuning. This suggests that the model's ability to handle complex relationships is less impactful on simplified datasets.

C. *ARIMA*

To forecast future domestic tourism demand, an ARIMA model was used, identifying 'Domestic Tourism Trips ('000)' as a key variable. Selangor, with the highest historical trip volumes, was selected for analysis to predict domestic tourist numbers for the next 10 years. The results provide valuable insights for tourism stakeholders in Selangor, aiding in planning and development decisions, while also highlighting the impact of including or excluding COVID-19 years in the analysis. The decision to exclude COVID-19 data was made to avoid skewing the predictive models with irregular and non-representative trends.

1) *Forecasting excluding COVID-19*

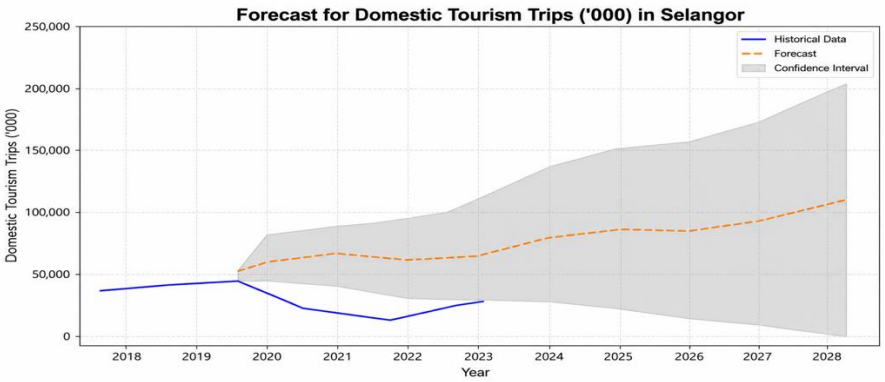

Fig. 22 Forecast for domestic tourism trips ('000) in Selangor starting from the year 2019

https://doi.org/10.31436/ijpcc.v11i1.539

The ARIMA model's forecast for Selangor starting from 2019 demonstrates a growth in domestic tourism over the next decade. The confidence interval for this forecast is narrower in the early years (2020-2023), reflecting greater certainty in the predictions. As the forecast progresses, the confidence interval gradually widens, showing increasing uncertainty but maintaining a generally optimistic upward trajectory (see Figure 22).

Policymakers can use these insights to invest in better infrastructure, promote hidden gems like Kampung Kuantan Firefly Park and Farm in The City. The RM200,000 grant allocated under the Visit Selangor Year (VSY) 2025 [8] campaign can support local tourism players in developing innovative products that cater to both domestic and international tourists. Furthermore, Selangor's digital campaign portal, *visitselangor2025.my*, can serve as a platform for promoting these initiatives and engaging with tourists in real-time [9].

Businesses, especially hotels and travel agencies, should prepare for higher demand by improving services and offering tailored packages. Strategic moves, such as planning around peak travel times and encouraging public-private partnerships, can help make tourism both profitable and sustainable. These findings highlight how predictive analytics can guide smart decisions and support long-term tourism development (see Figure 23).

2) *Forecasting including COVID-19*

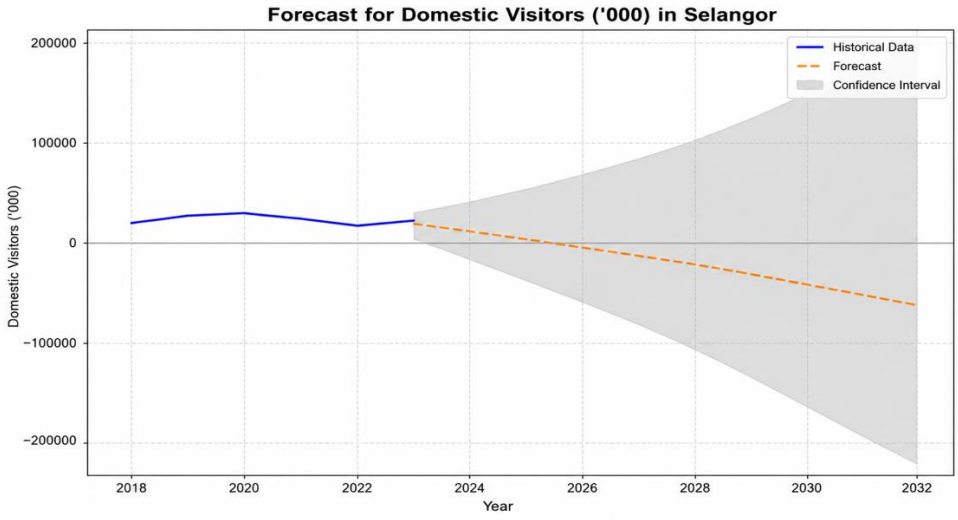

Fig. 23 Forecast for domestic tourism trips ('000) in Selangor starting from the year 2023

This forecast presents a more complex trajectory, with a gradual decline in domestic tourism trips. The confidence interval for this forecast is much wider immediately following 2023, indicating significant uncertainty caused by the inclusion of the COVID-19 years. Over time, the confidence interval expands rapidly, reflecting greater variability and difficulty in predicting long-term trends under disrupted conditions.

Unlike the 2019 forecast, where the trajectory was upward and the confidence interval symmetrical, the 2023 forecast highlights a declining trend, with the lower bound suggesting the possibility of sharp declines in domestic tourism. This highlights the lingering effects of COVID-19 on tourism recovery and the challenges in achieving stability.

To mitigate these risks, policymakers should diversify tourism offerings and invest in resilience strategies to stabilize the tourism sector. By integrating predictive insights with ongoing initiatives like VSY 2025, Selangor can achieve its goal of attracting eight million tourists and generating RM11.7 billion in tourism receipts by 2025 [10]. These data-driven strategies will ensure the long-term sustainability and resilience of Selangor's tourism industry.

V. Future Work

Future work will focus on expanding data collection to include monthly metrics, enabling detailed analysis of seasonal variations and short-term trends. Extending the study to incorporate international tourism data will provide a comprehensive view of global trends impacting Malaysia. Moreover, developing real-time predictive systems integrated with dashboards will enhance decision-making and allow timely responses to changing tourism patterns.

Additionally, incorporating more variables such as weather conditions, socio-economic factors, transportation accessibility, and major event schedules could improve the model's predictive accuracy. Integrating real-time data through IoT devices will enable the model to adapt

continuously to changing tourism behaviours, leading to more precise prediction and proactive data-driven strategies.

Apart from that, integrating Internet of Things (IoT) technology into tourism data collection can significantly enhance the accuracy and responsiveness of predictive models. Deploying IoT devices such as smart sensors at popular tourist sites can provide real-time data on visitor numbers, crowd density, and resource usage. This live data can be directly integrated into the ARIMA forecasting model to allow dynamic updates. These efforts will contribute to a more comprehensive and effective approach to forecasting and decision-making in sustainable tourism development.

This study has several limitations that need to be addressed for more robust outcomes. The dataset used covers only the period from 2017 to 2023, limiting the model's ability to capture long-term trends. Incorporating more extensive historical data and monthly datasets would improve model robustness.

VI. Conclusions

This study highlights the role of data science and machine learning in promoting sustainable tourism in Malaysia. By analyzing patterns and forecasting trends with models like SVM and ARIMA, it identifies key factors influencing tourism and provides actionable insights for sustainable development. Despite limitations, the research supports data-driven strategies for balancing economic, social, and environmental goals, offering a foundation for future advancements in predictive analytics and sustainable tourism practices.

Acknowledgment

The authors hereby acknowledge the review support offered by the IJPCC reviewers who took their time to study the manuscript and find it acceptable for publishing.

Conflict of Interest

The authors declare that there is no conflict of interest.

References

[1] Ministry of Tourism, Arts and Culture Malaysia (MOTAC), "National Tourism Policy 2020-2030," 2020. [Online]. Available: https://motac.gov.my/en/archives/2020/national-tourism-policy-2020-2030

[2] Tourism Malaysia, "Tourism contributes RM86.14 billion to Malaysia economy with 26.1 million tourists in 2019," 2019. [Online]. Available:https://www.tourism.gov.my/media/view/tourism-contributes-rm86-14-billion-to-malaysia-economy-with-26-1-million-tourists-in-2019

[3] A. Mai, K. Thi, T. Thi, and T. Le, "Factors influencing on tourism sustainable development in Vietnam," 2020. [Online]. Available: http://growingscience.com/beta/msl/3686-factors-influencing-on-tourism-sustainable-development-in-vietnam.html

[4] N. F. Nasir, M. A. Nasir, M. N. F. Nasir, and M. F. Nasir, "Understanding of Domestic Tourism in Malaysia," International Research Journal of Modernization in Engineering Technology and Science, 2020. [Online]. Available: https://www.irjmets.com/uploadedfiles/paper/volume2/issue_10_october_2020/4490/1628083177.pdf

[5] R. Agrawal, V. A. Wankhede, A. Kumar, S. Luthra, and D. Huisingh, "Big data analytics and sustainable tourism: A comprehensive review and network-based analysis for potential future research," International Journal of Information Management Data Insights, vol. 2, no. 2, p. 100122, 2022. [Online]. Available: https://doi.org/10.1016/j.jjimei.2022.100122

[6] F. J. Hoffmann, F. Braesemann, and T. Teubner, "Measuring sustainable tourism with online platform data," EPJ Data Science, vol. 11, no. 1, 2022. [Online]. Available: https://doi.org/10.1140/epjds/s13688-022-00354-6

[7] A. Louati, H. Louati, M. Alharbi, E. Kariri, T. Khawaji, Y. Almubaddil, and S. Aldwsary, "Machine Learning and Artificial Intelligence for a sustainable tourism: A case study on Saudi Arabia," Information, vol. 15, no. 9, p. 516, 2024. [Online]. Available: https://doi.org/10.3390/info15090516

[8] E. S. Journal, "State grant for tourism product development to roll out by end-Jan - tourism Selangor," Selangor Journal, 2025. [Online]. Available: https://selangorjournal.my/2025/01/state-grant-for-tourism-product-development-to-roll-out-by-end-jan-tourism-selangor/

[9] A. Sharon, "Malaysia: New Digital Portal Drives Tourism in Selangor," OpenGov Asia, 2024. [Online]. Available: https://opengovasia.com/2024/12/25/malaysia-new-digital-portal-drives-tourism-in-selangor/

[10] S. Online, "Selangor looking to sports tourism to boost figures," The Star, 2024. [Online]. Available: https://www.thestar.com.my/news/nation/2024/11/04/selangor-looking-to-sports-tourism-to-boost-figures