



PLANNING MALAYSIA:

Journal of the Malaysian Institute of Planners

VOLUME 23 ISSUE 6 (2025), Page 332 – 347

RIDERSHIP PREDICTION SYSTEM FOR RAPID BUS KUANTAN AND PENANG USING MULTI-FEATURE ANALYSIS

**Nur Nabilah Suzelan Amir¹, Nur Sabihah Anuar², Batrisyia Ismail³,
Nur Ainul Nasuha Mohd Azreen⁴ and Nor Azuana Ramli⁵**

^{1,2,3,4,5}Centre for Mathematical Sciences,
UNIVERSITI MALAYSIA PAHANG AL-SULTAN ABDULLAH

Abstract

Public transportation systems in Malaysia, particularly bus services in urban areas such as Kuantan and Penang, often struggle with inefficient scheduling and underutilized resources due to a lack of accurate demand forecasting tools. This study addresses the problem by developing a Ridership Prediction System for Rapid Bus Kuantan and Penang, employing a multi-feature analysis approach to predict bus ridership patterns based on various influencing factors, including weather conditions, weekends, and public holidays. The system aims to enhance public transportation efficiency through data-driven decision-making. A dataset covering ridership data from 2022 to 2024 was compiled from multiple sources, including Malaysia's Official Open Data Portal, Open-Meteo, and Office Holidays. The study employs several machine learning models, including Multiple Linear Regression (MLR), Ridge Regression (RR), Support Vector Regression (SVR), Artificial Neural Networks (ANN), and Random Forest (RF), to analyze ridership trends and forecast demand. Evaluation results indicate that the Random Forest model outperforms other models, achieving the highest accuracy with an R-squared value of 0.9940, making it the optimal choice for prediction. A Graphical User Interface (GUI) is also developed to visualize ridership forecasts, enabling transit operators to make informed scheduling and resource allocation decisions. The system offers valuable insights for enhancing public transportation services and improving the passenger experience. Future improvements include integrating real-time Internet of Things (IoT)-based data collection and exploring advanced deep learning techniques to further refine prediction accuracy.

Keywords: Ridership, Predictive Modelling, Machine Learning, Public Transportation.

⁵ Corresponding author. Email: azuana@umpsa.edu.my

INTRODUCTION

Public transportation plays a vital role in promoting urban mobility by providing a cost-effective, accessible, and environmentally sustainable alternative to private vehicles. However, accurately forecasting transit demand remains a challenge due to fluctuations driven by external factors such as weather conditions, public holidays, and socio-economic activities. Reliable ridership prediction is therefore essential for efficient service planning, optimal resource allocation, and improved passenger satisfaction (Rosni et al., 2025). In Malaysia, Rapid Bus services, operated by Prasarana Malaysia Berhad, form the backbone of urban public transport in several major cities, including Kuantan, Penang, and the Klang Valley. Among these, Rapid Kuantan and Rapid Penang were selected for this study due to their contrasting ridership patterns and strategic importance in the national transit network. Penang consistently reports high passenger volumes, reflecting its status as a densely populated urban and economic hub (Aziz & Mohamad, 2020). In contrast, Kuantan experiences more modest ridership levels, despite being a key administrative and commercial center on the East Coast. These disparities present a unique opportunity to examine how different urban contexts, regional demographics, and service characteristics influence bus usage.

Recent advancements in machine learning (ML) have demonstrated superior predictive capabilities over traditional statistical models by capturing complex, non-linear relationships in public transport datasets (Rocco di Torrepadula et al., 2024). International studies have effectively applied ML techniques such as Neural Networks (NN) (Jing et al., 2020; Farahmand et al., 2023), Long-Short Term Memory (Siswanto et al., 2024), support vector regressor (SVR) (Li et al., 2020), ensemble learning (Ge et al., 2024), and others in forecasting transit demand while incorporating factors like weather, holidays, and temporal variations. Despite these advancements, a gap remains in research focusing on Malaysia's public transportation system, particularly in developing a multi-feature ridership prediction model tailored to local transit conditions.

A study by Moosavi et al. (2023) employed chi-square automatic interaction detection (CHAID), Random Forest (RF), and gradient-boosted tree (GBT) models to determine which method is most accurate for predicting bus travel times on high- and low-frequency bus routes. This study also aims to determine the most applicable route construction approach. The dataset for this study was obtained from RapidKL Bus, consisting of raw AVL data that records the time-stamped locations of buses every 5 seconds. Based on the study's results, CHAID was the most accurate model for predicting travel time on high-frequency routes, while GBT showed the best performance for low-frequency service. This study also introduced the "key stop-based" route construction method, which is an accurate, reliable, and applicable method (Moosavi et al., 2023).

Lim et al. (2024) conducted a study focusing on public bus journey times, employing a methodology distinct from that of Moosavi et al. (2023). In this study, a basic artificial neural network (ANN) was developed and trained using a GPS dataset from a bus fleet in Johor, Malaysia. The results indicated that the model's accuracy improved compared to baseline approaches by incorporating factors such as bus stops, time of day, and travel distance (Lim et al., 2024). There have been few studies conducted on predicting ridership in Malaysia. A study by Zahir and Haron (2019) identified the factors that affect the operation of the myBAS service in Perlis. Three routes were selected as the case studies since they have the most access to popular cities. The authors utilized Atlas Ti Version 7.0 to analyze data and identify 11 factors affecting service operation. The results from the study indicated that LOS for headway pattern during peak hour on routes T101, F101, and CF109 is in the range of E (30min to 60 min) with one bus arriving per hour based on the United States Highway Capacity Manual 2010 (USHCM 2010) (Zahir & Haron, 2019). Although the findings from this study may benefit the local authority in achieving the status of a sustainable city by 2030, in line with the United Nations' declaration that focuses on green and healthy cities, the study did not apply any machine learning techniques in its prediction.

Finally, a preliminary analysis study was conducted to examine the factors influencing bus transit ridership. This study utilized correlation and multiple linear regression (MLR) analyses, focusing on ridership during peak and off-peak hours. Key factors investigated included bus commercial speed, service frequency, in-vehicle travel time, bus stop distance, and rainfall. The findings indicate that all these factors are significant and correlated with ridership. However, the regression models revealed that distance and travel time are not statistically significant in affecting ridership during peak and off-peak hours (Karunakaran et al., 2023).

Most existing local studies focus on travel time prediction (e.g., using Random Forest, GBT, or ANN models) or operational analysis, with limited attention given to direct ridership forecasting. Furthermore, these studies often rely on traditional statistical techniques or single-feature analyses without incorporating multi-dimensional external influences such as weather conditions, holiday schedules, and service-related variables. Additionally, there is a geographical imbalance, as most research focuses on the Klang Valley and Johor, leaving Kuantan and Penang relatively underexplored, despite their significance. Moreover, few studies conduct a comparative evaluation of multiple ML models under Malaysian transit conditions to identify the most effective approach. There is also a lack of practical, user-friendly tools, such as an interactive Graphical User Interface (GUI) that translates predictive models into actionable insights for transit operators and policymakers.

This study addresses these gaps by developing a multi-feature ridership prediction system for Rapid Bus services in Kuantan and Penang. It evaluates the performance of five ML models: MLR, Ridge Regression (RR), SVR, ANN, and RF using the datasets compiled from Malaysia's Official Open Data Portal, Open-Meteo, and Office Holidays for the period from 2022 to 2024. The system incorporates external variables such as weather and public holidays. It provides a GUI for visualization, offering a practical tool for developing support to inform data-driven scheduling and resource management decisions.

RESEARCH METHODOLOGY

This research encompasses five primary processes aimed at enhancing the effectiveness and accuracy of ridership predictions for Rapid Bus services in Kuantan and Penang. The initial phase involves data collection through various platforms, including data.gov.my for ridership statistics, Open-Meteo for meteorological data, and Office Holidays for holiday schedules spanning the years 2022 to 2024. Subsequently, the acquired data undergoes processing during the data pre-processing stage. Jupyter Notebook facilitates the integration of holiday and weather data relevant to Kuantan and Penang. At the same time, pgAdmin is employed to consolidate the necessary CSV files and encapsulate ridership, holiday, and weather data into a unified database. The next step is data cleaning, conducted in Jupyter Notebook, to address missing values, inconsistencies, and noise, thereby ensuring the data's suitability for subsequent analysis.

After the data is cleaned, it is partitioned into two portions: 80% for model training and 20% for model testing. The models utilized in this study include MLR, Ridge Regression (RR), SVR, ANN, and RF. Subsequently, all the models are assessed using metrics such as R-Squared (R^2), Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and Mean Squared Error (MSE) to evaluate their performance. If the performance metrics are inadequate, the process will revert to the development phase for hyperparameter tuning. Once the model meets the performance standards, it is integrated into a Graphical User Interface (GUI) for deployment and use. The entire flow of this research is illustrated in Figure 1.

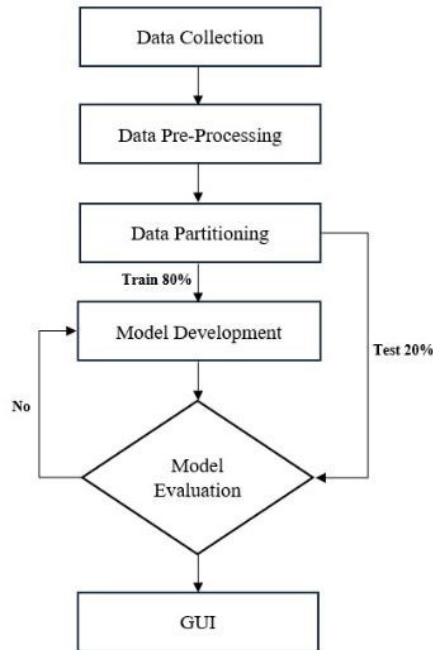


Figure 3: Research Flow Chart

Data Collection

Data for this project was collected through web scraping and downloading datasets from multiple sources. We employed scraping techniques to gather information about public holidays in Kuantan and Penang for the years 2022, 2023, and 2024 from the “Office Holidays” website, utilizing Jupyter Notebook. Additionally, a dataset with weather information for the years 2022, 2023, and 2024 was retrieved from the Open-Meteo website. We also gathered similar data on the number of public transport passengers on Prasarana Rapid Bus from 2022 to 2024 from the data.gov.my platform (Data Catalog). Table 1 describes each variable used in this study.

Table 4: Description of Each Variable

Variable	Data Type	Example	Description
date	Datetime	"2022-01-01"	The date in YYYY-MM-DD format, daily recorded
ridership_day	String	"Saturday"	Day of ridership
bus_rkn	Integer	"1722"	Amount of ridership in Rapid Bus Kuantan
bus_rkp	Integer	"39460"	Amount of ridership in Rapid Bus Penang

Variable	Data Type	Example	Description
ridership_location	String	"Kuantan"	Location of ridership
holiday_day	String	"Sunday"	Day of holiday
holiday_name	String	"Hari Raya Aidilfitri"	Name of holiday
holiday_type	String	"National Holiday"	Type of holiday
weather_day	String	"Monday"	Day of the weather
weather_code	Integer	"57"	Code for weather
temperature	Float	"23.96"	Temperature of the day in °C
rain	Float	"1.64"	Amount of rain in a day in millimeters (mm)

Data Preprocessing

In data preprocessing, multiple datasets were combined and reorganized into one dataset using Jupyter Notebook and pgAdmin. The 2022, 2023, and 2024 holiday data were first merged into a single file named "Holiday." Similarly, the weather data collected from the two areas, Kuantan and Penang, was consolidated into one file called "Weather." Finally, pgAdmin was utilized as a tool to combine the Ridership, Holiday, and Weather data into one file named "Combined Data." This transformation process ensured that all essential data was aggregated and organized, facilitating further analysis and modelling. After combining and organizing the data and thoroughly understanding the datasets during the description phase, the next step is to proceed with data cleaning. This involved processes such as formatting dates, dropping and renaming columns, checking for missing values and duplicates, and addressing outliers. Following this, exploratory data analysis was conducted, where graphs were plotted to examine the average ridership trend, the distribution of ridership, and the correlation between ridership in Kuantan and Penang.

Model Development

This section explores and explains several machine learning methods for regression problems relevant to our project, including Artificial Neural Networks (ANN), Support Vector Regression (SVR), Multiple Linear Regression (MLR), Ridge Regression (RR), and Random Forest (RF).

Artificial Neural Network

One of the machine learning models that can predict complex data patterns is the Artificial Neural Network (ANN). ANNs consist of various processing nodes, or neurons, that imitate the activity of human brain cells. The core concept of ANN is to simulate the human brain, enabling it to learn from large datasets (Paul et

al., 2021). The network learns to make predictions resembling human thought by recognizing connections and patterns in the input data. In this study, the ANN model is trained on input data, including weather and holidays, to predict ridership. The developed model is designed to predict ridership during bad weather on public holidays. This projection aims to support Rapid Bus operations in Kuantan and Penang by enhancing the reliability of bus services, optimizing scheduling efficiency, and ensuring the effective management of bus resources. This strategy enhances public transport operational decisions by leveraging ANN's capability to learn intricate patterns and model non-linear relationships.

Support Vector Regressor

Support Vector Regression (SVR) is a machine learning method suitable for addressing regression problems. Compared to conventional linear regression algorithms, it is particularly effective for datasets with complex or non-linear patterns. This is because SVR approximates complex non-linear relationships by mapping data into higher dimensions using kernel functions. One of the benefits of using SVR is its capacity to simulate intricate relationships between different ridership variables, such as weather and holidays. For instance, the total ridership count may fluctuate when weather conditions change, such as temperature or rainfall, and on days when travel demand increases. The potential for improved accuracy lies within the SVR algorithm, which is designed to address the unique nonlinear nature and interdependencies of these issues.

Multiple Linear Regression

Multiple Linear Regression (MLR) is a statistical method in which two or more independent variables influence a dependent variable. In MLR, several explanatory variables are utilized to derive the predicted value, rather than relying on a single variable as in standard linear regression. The formula is presented in Equation 1. In this context, MLR contributes to making complex predictions and serves as an excellent technique for estimating and simulating intricate relationships. The application of MLR has shown its effectiveness in conducting efficient and reliable ridership flow analyses, as evidenced by its use in previous studies.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \varepsilon \quad (1)$$

where y is the response/dependent variable, β_0 is the y -intercept, $\beta_1, \beta_2, \dots, \beta_n$ are the slope coefficients, $x_1, x_2, \dots, x_n$ are the independent variables/predictors, and ε is the error term.

Ridge Regression

Ridge regression is a type of linear regression that utilizes regularization to mitigate multicollinearity and prevent overfitting. It is especially beneficial when the input dataset exhibits multicollinearity, as it introduces a penalty term (L2 regularization) to the loss function. As a result, this penalty term keeps the regression coefficients from being overly influenced by data noise, enabling the model to generalize better to new data (Schreiber-Gregory, 2018). The relationship between ridership, weather, and holidays can be modelled with a ridge regression approach. Ridge regression is a method to mitigate the overfitting of regression coefficients, which is possible due to the high correlation among features, such as ridership trends.

Random Forest

Random Forest is an ensemble method in machine learning applicable to both regression and classification problems. This technique utilizes Bootstrap Aggregation (bagging), leveraging the combined strength of various decision trees. Unlike a single decision tree, Random Forest combines predictions from multiple trees to arrive at the final decision, thereby enhancing the overall model's accuracy. Each decision tree is trained on different groups of rows and features from the dataset to promote model diversity. This transformation enables the model to handle complex and non-linear relationships, making it a valuable tool for managing the intricacies of diverse data (Salman et al., 2024).

Model Evaluation

In this study, all proposed models were evaluated using statistical metrics for performance assessment: R-squared (R^2), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Mean Squared Error (MSE). Generally, MSE is the mean square of the difference between the prediction value and actual value, RMSE is the root square of MSE, MAE is the absolute value of the difference between the prediction value and actual value, and R-squared is a statistical measure that represents the proportion of variance in a dependent variable that is explained by the independent variables in a regression model. Equations 2 to 5 define the mathematical equations of all metrics used in this study, where y_t is the actual value of testing data, while is $\hat{y}_t$ the predicted value generated from the machine learning model.

$$MSE = \frac{1}{T} \sum_{t=1}^T (y_t - \hat{y}_t)^2 \quad (2)$$

$$RMSE = \sqrt{\frac{1}{T} \sum_{t=1}^T (y_t - \hat{y}_t)^2} \quad (3)$$

$$MAE = \frac{1}{T} \sum_{t=1}^T |y_t - \hat{y}_t| \quad (4)$$

$$R^2 = 1 - \frac{\sum_t (y_t - \hat{y}_t)^2}{\sum_t (y_t - \bar{y})^2} \quad (5)$$

ANALYSIS AND DISCUSSION

The primary objective of this study is to develop a predictive model for ridership on Rapid Bus services in Kuantan and Penang. Before modelling, some exploratory analyses need to be conducted to observe trends, data distribution, and correlations between variables. Figure 2 illustrates Kuantan and Penang's monthly total ridership patterns from January 2022 to October 2024. Penang, represented by the orange line, has a significantly higher ridership than Kuantan, ranging from 800,000 to 1.4 million. Ridership in Penang fluctuates greatly, with frequent peaks and troughs. Sharp declines occur in the middle of 2023 and the middle of 2024, likely due to seasonal factors, service interruptions, or holidays. Penang maintains a substantial ridership despite these variations, demonstrating that public transport is widely utilized. Kuantan, represented by the blue line, has consistently had low ridership, averaging less than 200,000 during this period. Its ridership pattern remains relatively steady, with slight variation, indicating a consistent but low demand for public transportation in Kuantan. This comparison emphasizes a critical distinction between the two locations. Penang exhibits rapid and strong ridership patterns, while Kuantan shows smaller, more stable usage.

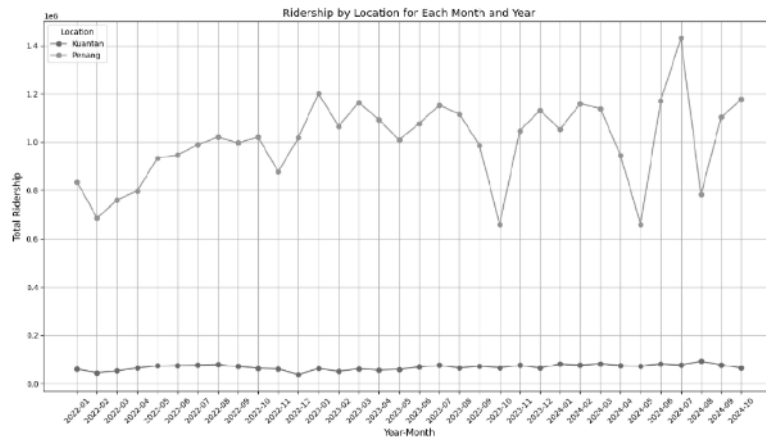


Figure 2: Comparison of Average Ridership in Kuantan and Penang by Month from 2022 until 2024

The distribution of ridership frequency across various ridership levels is shown in Figure 3. With over 700 instances of ridership below 10,000, a significant portion of the data concentrates on this lowest point, indicating numerous days with low ridership. This likely correlates with locations or times of low demand, such as Kuantan. Conversely, the distribution reveals another peak in the 30,000–40,000 ridership range, with frequencies declining as ridership exceeds 50,000. This suggests a more manageable, consistent set of high-ridership instances, likely in line with data from locations like Penang, which have robust public transportation systems. Overall, the distribution is dual, displaying two distinct ridership patterns: one with low demand and another with considerably higher levels, reflecting different trends in transportation use across the analyzed locations.

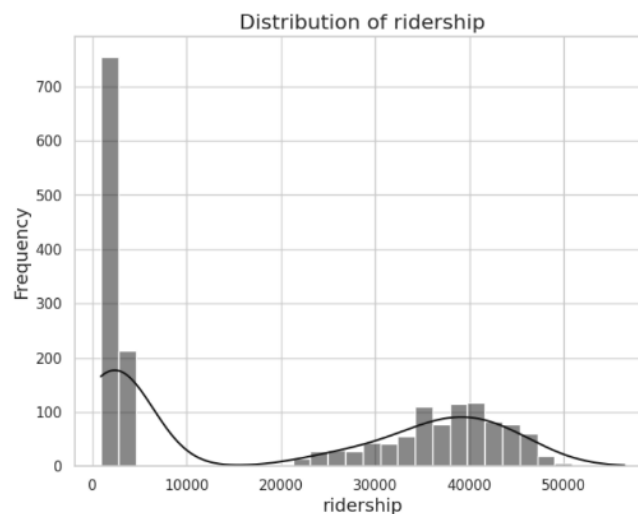


Figure 4: Distribution of Ridership

The heatmaps in Figure 4 illustrate the correlation between ridership and weather-related factors, specifically temperature, rain, and weather code for Kuantan and Penang. In Kuantan, ridership displays a weak positive correlation with temperature ($r = 0.25$), indicating a slight increase in ridership with rising temperatures. However, the correlation between ridership and rain ($r = 0.026$) as well as weather code ($r = -0.09$) is negligible, suggesting that rain and weather conditions have minimal impact on ridership. Additionally, there is a strong positive correlation between rain and weather code ($r = 0.68$), while temperature exhibits moderate negative correlations with both rain ($r = -0.39$) and weather code ($r = -0.57$).

In Penang, ridership also demonstrates weak correlations with all-weather attributes. The correlation between ridership and temperature ($r = 0.13$) and between ridership and rainfall ($r = 0.067$) is very weak, indicating that weather conditions have a minimal influence on ridership. The correlation between ridership and weather code is almost negligible ($r = -0.00047$). Similar to Kuantan, rain and weather codes are strongly positively correlated ($r = 0.76$), while temperature has moderate negative correlations with rain ($r = -0.44$) and weather code ($r = -0.49$). Overall, the analysis highlights that weather factors, including temperature, rainfall, and weather codes, have a limited impact on ridership patterns in both Kuantan and Penang. While subtle differences exist between the two locations, the overarching impact of weather on ridership remains minimal.

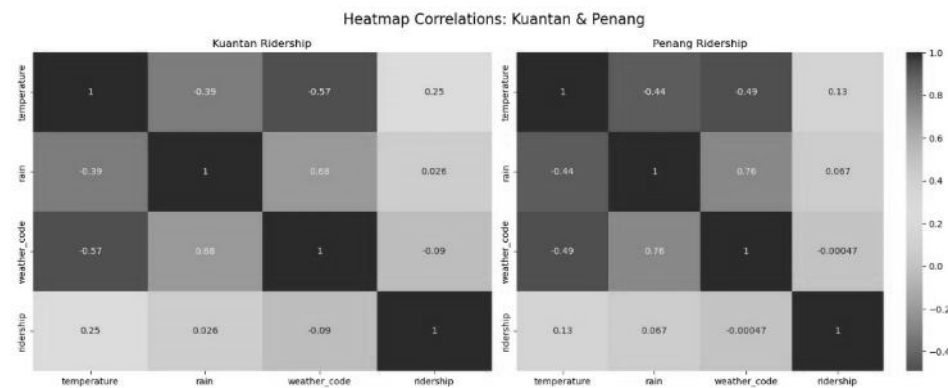


Figure 4: Heatmap Correlation: Kuantan & Penang

Although not all factors highly correlate with ridership, all factors were included in the model development. The results obtained as tabulated in Table 2, where the performance of five models: Random Forest (RF), Support Vector Regression (SVR), Multiple Linear Regression (MLR), Ridge Regression (RR), and Artificial Neural Network (ANN) was evaluated based on R^2 , Mean Squared Error (MSE), Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE). Among these models, RF achieved the best performance, with an R^2 of 0.994, indicating that it explains 99.4% of the variance in the data. It also has the lowest error values, making it the most reliable model for predictions.

Table 2 Model Comparison

Model	R ²	MSE	MAE	RMSE
RF	0.994008	2.07E+06	814.4755	1437.274
SVR	0.986775	4.56E+06	1569.635	2135.299
MLR	0.968617	1.08E+07	2590.035	3289.354
RR	0.944507	1.91E+07	2815.668	4374.062
ANN	0.020401	3.38E+08	16842.27	18377.64

Source: Author's Calculation

SVR ranks as the second-best model, achieving an R² of 0.987, which reflects its ability to capture 98.7% of the variance. However, its error values exceed RF's, indicating somewhat less accurate predictions. MLR and RR demonstrate moderate performance levels, with MLR recording an R² of 0.969 and RR at 0.944, meaning these models explain 96.8% and 94.5% of the variance, respectively. Their error metrics are significantly higher than those of RF and SVR, suggesting a challenge in accurately modelling the data's complexity. ANN exhibits the poorest performance, achieving a mere R² of 0.020, highlighting its failure to capture variance in the data. It also shows the highest error values (MSE: 3.38e+08, MAE: 16842.27, RMSE: 18377.64), rendering it unsuitable for this predictive task. ANN's underperformance likely stems from limited data and insufficient feature engineering. In contrast, RF and SVR's high accuracy and lower error rates suggest a preference for traditional machine learning models over deep learning methods in this context.

Based on these results, RF is the most appropriate model for predicting ridership, as it provides the highest accuracy and lowest error rates. Consequently, it is utilized for model deployment in our Graphical User Interface (GUI), as shown in Figure 5. The GUI was developed in Python with the Tkinter library.

Nur Nabilah Suzelan Amir, Nur Sabihah Anuar, Batrisyia Ismail, Nur Ainul Nasuha Mohd Azreen
and Nor Azuana Ramli
RIDERSHIP PREDICTION SYSTEM FOR RAPID BUS KUANTAN AND PENANG USING MULTI-
FEATURE ANALYSIS

Figure 5: Ridership Prediction Rapid Bus Kuantan and Penang for Weekday.

On April 19, 2024, the Ridership Prediction System for Rapid Bus in Kuantan and Penang predicted the number of passengers under the same conditions. For Kuantan, the predicted ridership was 2,591, while for Penang, it was higher at 40,559. The day was a weekday, and the passenger count was slightly lower than during public holidays. People may ride the bus to work or school to avoid road congestion. On the other hand, the Ridership Prediction System for Rapid Bus in Kuantan and Penang estimated different passenger numbers for the two locations under similar conditions on January 1, 2024, which is a public holiday. As shown in Figure 6, for Kuantan, the predicted ridership was 2,728, while for Penang, it is much higher at 45,549. Due to the New Year celebration, this date falls into the holiday category, which typically sees more passengers than weekdays or weekends. Penang is a well-known tourist destination, particularly during holidays such as New Year's. Tourists may visit Penang for celebrations or sightseeing, and many of these visitors are likely to rely on public transport. Penang may have a larger population, with more people using public transport and better service compared to Kuantan.

The figure displays two side-by-side screenshots of a web application titled "Ridership Rapid Bus Kuantan & Penang Predictor". Both interfaces feature input fields for Temperature (°C), Rain (mm/day), Weather Code, Year, Month, Day, and Ridership Location (0 = Kuantan, 1 = Penang). Buttons for "Get Weather Description", "Determine Day Type", and "Predict" are present. The left screenshot shows a predicted ridership of 2728, while the right screenshot shows a predicted ridership of 45549, indicating a significant increase in ridership for a public holiday.

Figure 6: Ridership Prediction Rapid Bus Kuantan and Penang for Public Holiday.

CONCLUSION

This research successfully developed a ridership prediction system for Rapid Bus services in Kuantan and Penang. The project employed multi-feature analysis supported by machine learning techniques, including Artificial Neural Networks, Multiple Linear Regression (MLR), Ridge Regression, Support Vector Regression (SVR), and Random Forest (RF). The system demonstrated its ability to accurately predict ridership patterns based on influencing factors, such as weather conditions and day type. Among the evaluated models, the RF model emerged as the top performer, achieving the best performance across R-squared, MAE, and RMSE metrics. In addition, the developed graphical user interface (GUI) enhanced accessibility and user interaction, providing a user-friendly ridership predictor for Rapid Bus Kuantan and Penang.

The proposed ridership prediction system offers significant benefits to Rapid Bus management, particularly in strategic decision-making related to scheduling, route optimization, and resource allocation. It can help identify ridership trends, improve service efficiency, and reduce operational costs, thereby supporting sustainable public transport initiatives. Moreover, passengers could benefit from a more reliable schedule and accurate occupancy information, which would increase their confidence in and reliance on public transport services.

Despite these achievements, several limitations were encountered in this study. First, the dataset obtained from the data.gov.my open data platform

lacked a clear description of its sample size (i.e., the total number of records), making it difficult to assess data representativeness. Second, the dataset contained incomplete ridership records and omitted important socio-economic variables, such as population density, income levels, and employment distribution, which could significantly influence ridership behavior. These data constraints may have contributed to the extremely poor performance of the ANN model ($R^2 = 0.02$), indicating that deep learning methods require richer and more comprehensive datasets to perform effectively.

For future work, the system could be enhanced by integrating Internet of Things (IoT) sensors on buses to collect real-time ridership and environmental data, thereby improving prediction accuracy. Incorporating additional data sources such as passenger feedback, traffic conditions, and socio-economic indicators may also enrich model insights. Furthermore, exploring advanced deep learning architectures and hybrid modeling approaches could further strengthen the system's predictive capabilities and ensure it remains adaptive to the evolving demands of public transportation in Malaysia.

ACKNOWLEDGEMENT

This research was funded by a grant from the Ministry of Higher Education of Malaysia (FRGS Grant) RDU230127 (FRGS/1/2023/TK02/UMP/02/2) and the Research and Innovation Department, Universiti Malaysia Pahang Al-Sultan Abdullah.

REFERENCES

- Aziz, A., & Mohamad, J. (2020). Public Transport Planning: Local Bus Service Integration And Improvements In Penang, Malaysia. *Journal of the Malaysian Institute of Planners*, 18(3), 179–189.
- Data Catalog. (n.d.). *Ridership headline*. https://data.gov.my/data-catalogue/ridership_headline
- Farahmand, Z. H., Gkiotsalitis, K., & Geurs, K. T. (2023). Predicting bus ridership based on the weather conditions using deep learning algorithms. *Transportation Research Interdisciplinary Perspectives*, 19, 100833.
- Ge, Z., Yang, L., Li, J., Chen, Y., & Xu, Y. (2024). Bus schedule time prediction based on an LSTM-SVR model. *Mathematics*, 12(22), 3589.
- Jing, Z., & Yin, X. (2020). Neural network-based prediction model for passenger flow in a large passenger station: An exploratory study. *IEEE Access*, 8, 36876–36884.
- Karunakaran, R., Yuen, C. W., Chuah, J. H., Zainol, R., & Goh, B. H. (2023). Correlation analysis and regression of transit ridership by means of public buses – A preliminary data analysis. *IOP Conference Series: Materials Science and Engineering*, 1289, 012044.
- Li, C., Wang, X., Cheng, Z., & Bai, Y. (2020). Forecasting bus passenger flows by using a clustering-based support vector regression approach. *IEEE Access*, 8, 19717–

- 19725.
- Lim, Y. T., Keoh, S. L., Chong, Y. W., Ibrahim, N. F., & Sharul Kamal, A. R. (2024). An ANN-based prediction model for public bus journey time. In *Proceedings of the 6th International Conference on Artificial Intelligence in Information and Communication (ICAIIC 2024)* (pp. 578–583). IEEE.
- Moosavi, S. M. H., Aghaabbasi, M., Yuen, C. W., & Jahed Armaghani, D. (2023). Evaluation of applicability and accuracy of bus travel time prediction in high and low frequency bus routes using tree-based ML techniques. *Journal of Soft Computing in Civil Engineering*, 7(2), 74–97.
- Office Holidays. (n.d.). *Malaysia public holidays 2024*. <https://www.officeholidays.com/countries/malaysia/2024>
- Open-Meteo. (n.d.). *Weather forecast API*. <https://open-meteo.com/>
- Paul, S., Kumar, V., & Jha, P. (2021). Artificial neural network and its applications: Unraveling the efficiency for hydrogen production. In J. Ren, W. Shen, Y. Man, & L. Dong (Eds.), *Applications of artificial intelligence in process systems engineering* (pp. 187–206). Elsevier. <https://doi.org/10.1016/B978-0-12-821092-5.00014-0>
- Rocco di Torrepadula, F., Napolitano, E. V., Di Martino, S., & Mazzocca, N. (2024). Machine learning for public transportation demand prediction: A systematic literature review. *Engineering Applications of Artificial Intelligence*, 137, 109166.
- Rosni, N. A., Qi, T. S., & Mohd Nor, M. (2025). University Students' Perceptions Of Public Bus Service Efficiency And Effectiveness in Influencing Ridership. *PLANNING MALAYSIA*, 23(36). <https://doi.org/10.21837/pm.v23i36.1712>
- Salman, H. A., Kalakech, A., & Steiti, A. (2024). Random forest algorithm overview (Trans.). *Babylonian Journal of Machine Learning*, 2024, 69–79. <https://doi.org/10.58496/BJML/2024/007>
- Schreiber-Gregory, D. N. (2018). Ridge regression and multicollinearity: An in-depth review. *Model Assisted Statistics and Applications*, 13(4), 359–365. <https://doi.org/10.3233/MAS-180446>
- Siswanto, J., Manongga, D., Sembiring, I., & Wiyono, S. (2024). Deep learning based LSTM model for predicting the number of passengers for public transport bus operators. *Jurnal Online Informatika*, 9(1), 18–28.
- Zahir, M. H., & Haron, S. (2019). The study of headway patterns and level of service in myBAS ridership. *Journal of Physics: Conference Series*, 1349, 012082.

Received: 2nd September 2025. Accepted: 14th September 2025