UNDERSTANDING FACTORS INFLUENCING MOTORCYCLE OWNERSHIP IN NARAYANGANJ: A MACHINE LEARNING & STATISTICAL APPROACH

Tanzim Rashid*[1], Farear Jamil Nehal[2], Tanjina Mannan Risha[3], Simanto Biswas Shaisab[4], Md Aminul Islam[5]

[1] *Student, Bangladesh University of Engineering & Technology, Bangladesh, e-mail: tanzimrashid@iut-dhaka.edu*
[2]*Junior Engineer, BCL Associates Ltd., Bangladesh e-mail: farear.nihal@gmail.com*
[3]*Lecturer, Northern University Bangladesh, Bangladesh, e-mail: tanjina.risha@gmail.com*
[4] *Student, Bangladesh University of Engineering & Technology, Bangladesh, e-mail: simantobiswas11809101@gmail.com*
[5] *Transport Modeler, SMEC International Pty Ltd, Bangladesh, e-mail: m.aminulce07@gmail.com*

***Corresponding Author**

ABSTRACT

With rapid urban expansion in developing countries such as Bangladesh, understanding the determinants of motorcycle ownership is increasingly critical for sustainable transportation planning. However, existing studies in Bangladesh predominantly rely on traditional statistical approaches and have not utilized advanced machine learning techniques capable of capturing complex, nonlinear relationships influencing motorcycle ownership. This study addresses this research gap by applying both a machine learning method Random Forest (RF) and a traditional statistical approach Binary Logistic Regression (BLR) to model motorcycle ownership patterns in Narayanganj. Survey data was collected from 3,330 households in Narayanganj city corporation area which incorporated demographic, socioeconomic, housing, and travel behavior characteristics. Motorcycle ownership was taken as the dependent variable and modeled as a binary outcome (yes/no), with key independent variables including household income level, housing type, employment status and gender distribution among household members. An 80/20 train–test split was applied meaning 80% of the dataset was used to train the model while the other 20% was used to test the models' predictive abilities. The Random Forest model achieved 96% predictive accuracy, outperforming logistic regression which had an accuracy of 87%. Further evaluation using the F1-score and confusion matrix confirmed the robustness of the model. The RF model attained an F1-score of 98% for non-motorcycle owners (class 0) and an F1-score of 87% for motorcycle owners (class 1). While the BLR model achieved an F1-score of 93% for class 0 it only achieved an F1-score of 44% for class 1. In the RF model Gini-based feature importance helped identify household income, housing type, number of unemployed household members and number of males in the household as the most influential predictors. In contrast, the binary logistic regression (BLR) model provided coefficients and Z-values, which allowed for determining both the direction and magnitude of each variable's effect on the likelihood of motorcycle ownership. Results for the BLR model showed household income as the most influential variable which had a positive effect on motorcycle ownership. The models were developed using Python for Random Forest and STATA for binary logistic regression, ensuring methodological consistency across analytical frameworks. Overall, the findings not only deepen understanding of household motorcycle ownership dynamics in rapidly urbanizing cities like Narayanganj but also establish a foundation for exploring scenario-based forecasting and policy impact simulations to support evidence-driven transport planning.

Keywords: *Motorcycle Ownership, Random Forest (RF), Binary Logistic Regression (BLR), Machine Learning (ML), Vehicle Ownership Prediction.*

1. INTRODUCTION

Motorcycle ownership plays a crucial role in shaping urban transportation systems, particularly in rapidly developing cities where motorcycles offer an affordable, flexible, and time-efficient mode of travel. In densely populated urban areas like Narayanganj, the rising trend in motorcycle ownership has significant implications for traffic flow, transportation planning, road safety, and environmental sustainability (Herwangi, 2014). As urbanization intensifies, the growing reliance on motorcycles contributes to challenges such as congestion, increased accident risk, and deteriorating air quality. Understanding the underlying factors that drive motorcycle ownership is therefore essential for policymakers aiming to design mobility strategies that support both individual accessibility and long-term sustainability goals.

Previous research on motorcycle ownership has traditionally focused on socioeconomic variables such as income, employment status, household structure, and travel needs using statistical techniques including logistic regression (Bray & Holyoak, 2015). While such models have provided important insights, they are often limited by assumptions of linearity and reduced capacity to capture complex interactions among variables. To overcome these limitations, machine learning (ML) methods such as Random Forest (RF) offer a more robust analytical framework by efficiently handling diverse datasets and modelling nonlinear relationships that conventional statistical approaches may overlook (Shirgaokar, 2012).

In the context of Narayanganj, where motorcycle ownership continues to rise, applying machine learning techniques alongside traditional regression models enables a more comprehensive understanding of ownership determinants without relying solely on predefined variable relationships. By jointly applying Binary Logistic Regression and Random Forest models to household motorcycle ownership this study not only identifies the key predictors and their relationships but also combines interpretability of behavioural effects with improved predictive accuracy. To our knowledge, it is among the first studies in Bangladesh to integrate machine learning with conventional regression for motorcycle ownership analysis, providing evidence directly relevant for policy-oriented transport planning in rapidly urbanizing cities and supporting strategies to enhance urban mobility and promote sustainable transportation practices.

2. LITERATURE REVIEW

Binary logistic regression has traditionally been employed to model motorcycle ownership patterns. However, logistic regression assumes a linear relationship between explanatory variables and the log-odds of ownership, which limits its ability to capture complex, non-linear interactions commonly present in transport-related behavioural data (Bray & Holyoak, 2015). When multiple interacting socioeconomic and travel behaviour variables influence ownership, this method may provide incomplete or less accurate predictions.

Machine learning models, particularly Random Forest (RF), offer a robust alternative by efficiently handling high-dimensional datasets and modelling non-linear relationships without relying on restrictive assumptions (Olaniran et al., 2025). One of the key strengths of RF and similar machine learning models is their ability to assess the relative importance of predictors, providing interpretable insights into which socioeconomic or travel behaviour factors most strongly influence ownership patterns. Additionally, machine learning models can capture interactions among multiple predictors that are difficult to pre-specify in traditional regression frameworks, enabling a more nuanced understanding of ownership determinants.

RF has been applied in various transportation planning contexts, including travel behaviour modelling, mode choice prediction and other classification tasks, showing reliable performance even with complex patterns or missing data (Law et al., 2023). Several recent studies in South and East Asia have applied machine learning techniques to transport behaviour modeling. For example, Zisan et al. (2025) employed Random Forest and logistic regression models to analyse household car ownership in Narayanganj, Bangladesh demonstrating the superior predictive performance of machine learning methods and highlighting the importance of income and employment characteristics. Similarly, Yao et al. (2022) applied Random Forest models to household vehicle ownership analysis in a rapidly urbanizing Asian context showing that machine learning approaches can effectively capture nonlinear relationships overlooked by traditional regression models. In the domain of other transportation behaviour modelling studies, a study from Ashik et al. (2024) applied machine learning techniques to model commute mode choice in Dhaka, while Ghosh and Nagaraj (2024) conducted a comparative analysis of machine learning and traditional statistical models for travel behaviour in Bengaluru. These studies highlight the increasing regional adoption of machine learning approaches to capture complex transport-related decision-making processes in rapidly urbanizing cities.

Combining machine learning with conventional statistical models, such as logistic regression, allows for comparative evaluation of model performance and predictive accuracy, while leveraging the interpretability of regression coefficients alongside the flexibility of RF. This integrated approach enhances the robustness of ownership modelling and supports scenario-based analyses, offering a foundation for future studies on household vehicle adoption (Ma & Ye, 2019).

3. METHODOLOGY

3.1 Data Collection

A total of 3,330 household samples were collected through a structured survey conducted in Narayanganj. The survey gathered comprehensive socio-demographic and travel behaviour information, including monthly household income level, housing type, number of unemployed household members, number of household members employed, at a job, number of business persons in the household, number of males and females in the household. To avoid clustering, no two households within a 100-meter radius were selected.

3.2 Data Processing

The raw data was processed by encoding categorical features such as income level and housing type. Although motorcycle ownership represents a minority class (14.29%), no resampling or artificial class balancing techniques were applied. Instead, model performance was evaluated using class-specific metrics such as precision, recall, and F1-score to ensure reliable assessment of both owner and non-owner households. After processing, relevant predictors for motorcycle ownership were selected based on previous literature and domain knowledge. The following Table 1 shows the frequency and percentage (%) of the responses of each variable.

Table 1: Frequency and Percentage Distribution of Variables in Dataset

Variable	Responses	Frequency	Percentage (%)
Housing Type	Tin Shade / Semi-Pacca Building	1,020	30.63%
	Single Floor / Duplex Building	890	26.73%
	Multistoried Building	1,420	42.64%
Monthly Household Income	<20,000 BDT	698	20.96%
	20,000–40,000 BDT	1,881	56.49%
	40,000–60,000 BDT	487	14.62%
	>60,000 BDT	264	7.93%

Variable	Responses	Frequency	Percentage (%)
Number of Unemployed Household Members	0	211	6.34%
	1	1,349	40.51%
	2 or more	1,770	53.15%
Number of Employed Household Members	0	1,295	38.89%
	1 or more	2,035	61.11%
Number of Business Person in Household	0	1,911	57.39%
	1 or more	1,419	42.61%
Number of Males in Household	0	28	0.84%
	1	1,844	55.38%
	2 or more	1,459	43.81%
Number of Females in Household	0	187	5.62%
	1	2,080	62.46%
	2 or more	1,063	31.92%
Motorcycle Ownership	No	2,854	85.71%
	Yes	476	14.29%

3.3 Model Development

3.3.1 Binary Logistic Regression (BLR)

To quantify the direction and magnitude of the influence of household characteristics on motorcycle ownership, Binary Logistic Regression (BLR) was employed alongside machine learning. The dependent variable was defined as a binary outcome, where $Y = 1$ indicates that a household owns at least one motorcycle and $Y = 0$ indicates otherwise.

BLR is suitable for modelling a Bernoulli-distributed dependent variable with only two possible outcomes. Most explanatory variables in this study are categorical or discrete, making logistic regression an appropriate and widely used approach (Kilic, 2015). The probability of motorcycle ownership for a household given its characteristics can be expressed as:

$$P\,(Y = 1) = \frac{1}{1 + e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k}} \tag{1}$$

Where, $P\,(Y = 1)$ represents the probability of a household owning at least one motorcycle and $X_1, X_2, \ldots, X_k$ represent the predictor variables. Here, β_0 is the intercept and $\beta_1, \beta_2, \ldots, \beta_k$ are the model coefficients. Coefficients were estimated using maximum likelihood estimation (MLE), and exponentiating the coefficients yields odds ratios, allowing interpretation of the effect of each variable on the likelihood of motorcycle ownership.

3.3.2 Random Forest (RF)

Random Forest (RF) is an ensemble learning method that constructs multiple decision trees and aggregates their predictions to enhance accuracy and robustness. For a classification problem, each tree generates a prediction $h_i(x)$ and the final RF prediction $H(x)$ is obtained through majority voting:

$$H(x) = \arg\max_{y} \sum_{i=1}^{k} I(h_i(x) = y) \tag{2}$$

Where k is the number of trees, $h_i(x)$ is the prediction of the i^{th} tree and $I(.)$ is an indicator function equal to 1 when the tree predicts class y and 0 otherwise.

Default hyperparameter settings were used for the Random Forest model, as preliminary testing showed stable performance without extensive tuning. The model was implemented using 500 decision trees, which was found to provide stable predictive performance while avoiding overfitting. Each tree was built using a bootstrap sample drawn with replacement from the training set. At each node, a random subset of predictors is considered for splitting and the best split is chosen based on the Gini impurity:

$$G(t) = 1 - \sum_{j=1}^{C} p(j|t)^2 \tag{3}$$

Where $p(j|t)$ is the proportion of observations of class j at node t, and C is the total number of classes. Trees are fully grown without pruning and the ensemble of trees forms the final RF model.

RF also provides a measure of feature importance by quantifying each predictor's contribution to reducing impurity across all trees:

$$\text{Imp}(f) = \frac{1}{k}\sum_{i=1}^{k}\sum_{t\in T_i(f)} \Delta i(t) \tag{4}$$

Where $Ti(f)$ is the set of nodes in tree i where feature f is used for splitting and $\Delta i(t)$ is the decrease in node impurity at t. This allows identification of the most influential factors affecting motorcycle ownership.

3.4 Model Evaluation

The dataset was divided into training (80%) and testing (20%) subsets to assess the predictive performance of both models. This separation ensures that models are trained on a substantial portion of the data while retaining a distinct subset to evaluate performance on unseen observations, reducing the risk of overfitting.

Model performance was assessed using several metrics. Accuracy, measuring the overall proportion of correctly predicted motorcycle owners and non-owners:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \tag{5}$$

Precision and recall were used to evaluate the reliability of positive predictions and the model's ability to correctly identify actual motorcycle owners, respectively. The F1-score, the harmonic mean of precision and recall, was computed to provide a balanced measure:

$$\text{F1} = \frac{2\cdot\text{Precision}\cdot\text{Recall}}{\text{Precision}+\text{Recall}} = \frac{2\cdot TP}{2\cdot TP+FP+FN} \tag{6}$$

This combination of logistic regression and Random Forest enables comparison between traditional statistical and machine learning approaches while identifying the key predictors of motorcycle ownership in Narayanganj.

4. RESULTS AND DISCUSSIONS

4.1 Binary Logistic Regression Results

The binary logistic regression model was used to predict motorcycle ownership. The variable number of females was omitted in the model, since the number of males and number of females were found to be collinear. The number of males was retained because it showed stronger statistical significance.

Table 2: Results of the Binary Logistic Regression Model

Variable	Coefficient	Std. Error	Z	P-value
Constant	-6.4492	0.339	-19.011	<0.001
Housing Type	0.1949	0.096	2.037	0.042
Household Monthly Income	1.8265	0.098	18.605	<0.001
Number of Unemployed People in Household	-0.4979	0.105	-4.754	<0.001
Number of Employed People in Household	-0.8295	0.166	-5.006	<0.001
Number of Business Persons in Household	0.1509	0.185	0.814	0.415
Number of Males in Household	0.6064	0.143	4.230	<0.001

Table 3: Performance Metrics of the Model

Performance Metric	Value
Log-Likelihood (Model)	-719.07
Log-Likelihood (Null)	-1093.32
McFadden R^2	0.3423
Model Accuracy	0.8769
F1 Score (No)	0.93
F1 Score (Yes)	0.44

The regression coefficients indicate both the direction and strength of the relationships between the predictors and motorcycle ownership. Positive coefficients increase the likelihood of ownership, while negative coefficients reduce it. The Z values show how far each estimate is from zero in standard deviation units, with larger absolute values indicating stronger evidence of a real effect. P values complement this by measuring statistical significance, where values below 0.05 are typically considered statistically significant in studies, values below 0.01 indicate very strong significance and sometimes values below 0.1 are considered marginally significant.

Among the predictors, household monthly income is the strongest and most significant factor, with a coefficient of 1.8265 and a p value below 0.001. This implies that higher income substantially increases the probability of motorcycle ownership. The number of unemployed household members shows a significant negative effect, suggesting that financial strain lowers the likelihood of owning a motorcycle. The number of employed members also has a significant negative coefficient, which may reflect household mobility patterns or reliance on employer-provided or alternative transport modes.

Housing type shows a small but significant positive effect, indicating that households living in more permanent structures are slightly more inclined to own motorcycles. The number of business persons in the household is not statistically significant, meaning it does not meaningfully influence ownership in this dataset. In contrast, the number of males in the household has a significant positive effect, which aligns with gendered mobility norms commonly observed in South Asian urban areas.

Model performance metrics show that the regression provides a meaningful fit to the data. The log-likelihood of the fitted model (-719.07) is substantially higher than that of the null model (-1093.32), indicating that the included predictors collectively explain a significant portion of the variation in motorcycle ownership compared to a model with no explanatory variables. The McFadden R^2 of 0.3423 reflects a moderate level of explanatory power. This suggests that although the socio-demographic variables account for a meaningful share of the likelihood of motorcycle ownership, a substantial portion of variability remains unexplained.

The overall accuracy of 0.8769 (87.69%) and the F1 score of 0.93 for non-owners demonstrate that the model performs well in identifying households that do not own a motorcycle. However, the model shows weaker predictive performance for actual motorcycle owners. This is reflected by lower F1 score for the "Yes" class. This highlights a limitation of logistic regression. While it captures general trends effectively, it may struggle with minority outcomes in the dataset, potentially underestimating ownership probability in some households. In contrast, machine learning approaches such as Random Forest can capture complex, nonlinear relationships and interactions between variables, often providing better predictive performance for both owner and non-owner classes. Therefore, combining logistic regression for interpretability with machine learning methods for predictive accuracy can offer complementary insights for understanding and forecasting motorcycle ownership in urban settings.

Overall, the results highlight the key socio-economic and demographic factors shaping motorcycle ownership in Narayanganj while also reflecting the usual limitations of logistic regression, including its assumption of linearity in the log odds.

4.2 Random Forest Results

The following Table 4 presents the results from the Random Forest model, while Figure 1 illustrates both the confusion matrix. Figure 2 shows the feature importance of the predictors. The confusion matrix provides a detailed breakdown of the model's classification performance. It basically shows the number of correctly and incorrectly predicted households for both motorcycle owners and non-owners. This helps assess where the model performs well and where misclassifications occur. The feature importance chart indicates the relative contribution of each predictor in determining motorcycle ownership. The importance values sum up to 1, meaning each feature's contribution is expressed as a proportion of the total predictive power. Higher importance values highlight the variables that have the greatest influence on the model's predictions.

Table 4: Random Forest (RF) Model Results

Metric	Non-Owner (Class 0)	Non-Owner (Class 1)
Precision	0.99	0.79
Recall	0.96	0.98
F1-Score	0.98	0.87
Accuracy	0.96	0.96

The Random Forest model was applied to the same dataset as the BLR model with the same 80-20 split so the models can be directly compared to each other. The results demonstrated strong predictive performance. It achieved an overall accuracy of 96%, indicating a notable improvement compared to the logistic regression model accuracy of 87%. The model performed particularly well in identifying households without a motorcycle, with precision and recall both above 0.95. For motorcycle-owning households, the recall was very high at 0.98, while the precision was slightly lower at 0.79, resulting in an F1-score of 0.87. In comparison, while the BLR model only achieved an F1-score of only 0.44 for motorcycle-owning households, which is the main target of prediction in this study. This indicates that the Random Forest model delivered a substantially better result when predicting motorcycle owners. These metrics suggest that the Random Forest model provides a well-balanced classification and effectively captures both owner and non-owner households.

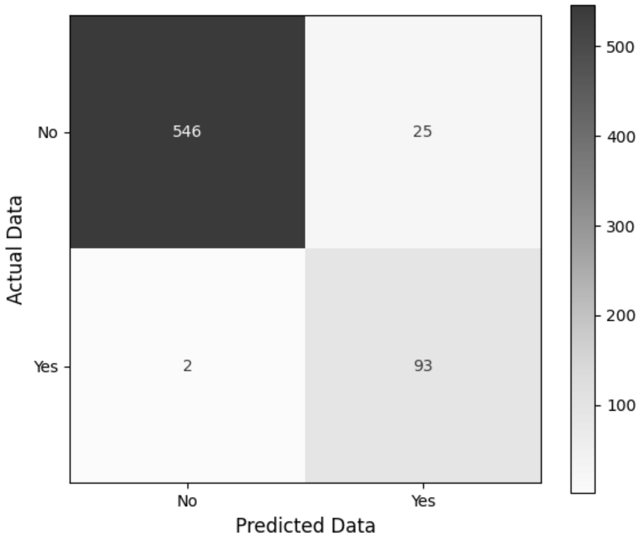

Figure 1: Confusion Matrix of the RF Model

The confusion matrix illustrates the model's predictive accuracy across the two classes. Out of 571 actual non-owner households, 546 were correctly classified as non-owners, while 25 were misclassified as owners. For motorcycle-owning households, 93 out of 95 were correctly identified, with only 2 misclassified as non-owners. This demonstrates the model's excellent ability to distinguish between owner and non-owner households, particularly improving the detection of motorcycle owners compared to the BLR model.

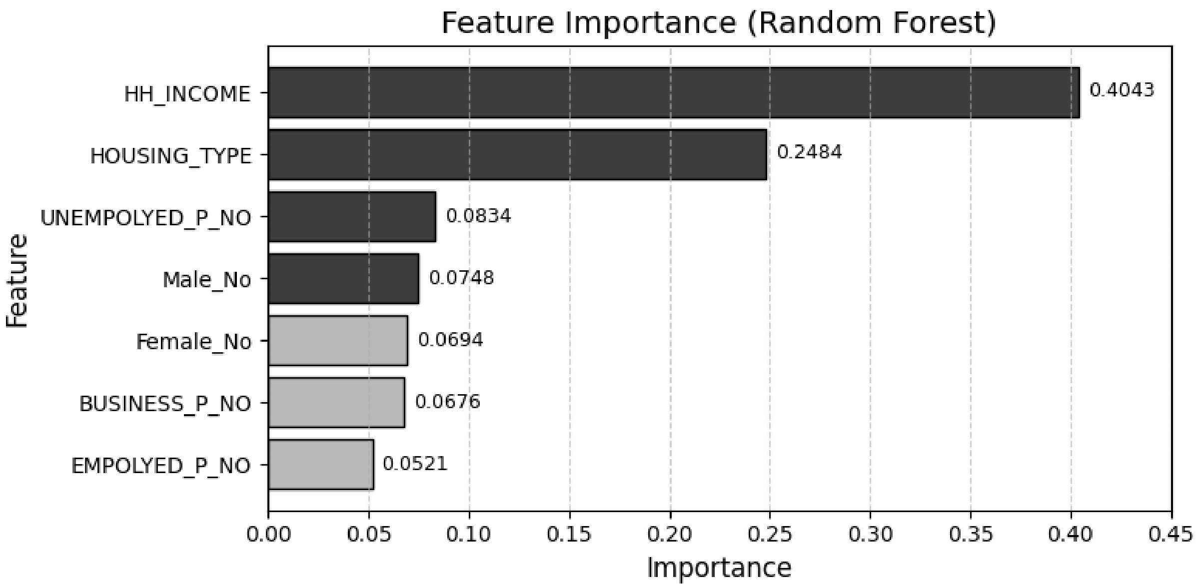

Figure 2: Feature Importance of RF Model

The feature importance plot from the Random Forest model quantifies the relative contribution of each predictor to the model's decisions. Consistent with the previous statistical analysis, household monthly income emerges as the most influential factor, with the highest importance score of 0.4043. This is followed by housing type at 0.2484, confirming that these two variables are the dominant determinants of motorcycle ownership in this dataset. The remaining features, including the number of unemployed members, employed members, males, females, and business persons, together account for the remaining 0.3473. This indicates that they also play a meaningful but comparatively smaller role. While the Random Forest captures more complex, nonlinear relationships than the regression model, it does not provide information on the magnitude or direction of the effects. It only shows the relative importance of each predictor. Overall, the similarity in factor influence between the two models reinforces the robustness of the identified key determinants.

5. CONCLUSIONS

This study examined the key factors influencing motorcycle ownership in Narayanganj using both Binary Logistic Regression (BLR) and Random Forest (RF) models. The findings show that socio-demographic and household characteristics strongly shape motorcycle ownership patterns.

The BLR model provided insights into the direction and magnitude of these effects. Household monthly income emerged as the strongest positive predictor. The number of unemployed and employed household members had significant negative effects on ownership. Housing type and the number of males in the household also showed positive influences, though smaller in magnitude. The model performed well overall but struggled with minority outcomes. For example, the F1-score for motorcycle-owning households was only 0.44.

The Random Forest model outperformed BLR in predictive accuracy. It achieved an overall accuracy of 96% and an F1-score of 0.87 for motorcycle owners. This model captured complex, nonlinear relationships among variables that the regression could not. Feature importance analysis confirmed the dominance of household income and housing type. Other variables, such as employment and household composition, contributed meaningfully but to a lesser extent. RF shows only relative importance, not the magnitude or direction of effects.

Overall, the results reveal that higher household income, permanent housing structures and male household composition are key drivers of motorcycle ownership. Employment patterns and financial constraints reduce ownership likelihood. These findings are important for policymakers. They can inform strategies to manage motorcycle demand, improve public transportation and promote sustainable mobility.

Future research could include more detailed socio-economic and spatial data, such as household travel behavior, vehicle usage patterns, and neighborhood characteristics. This would improve predictive accuracy and allow for scenario-based transport planning. That can help policymakers assess the potential impacts of interventions like public transit improvements or traffic demand management.

DECLARATION OF USE OF AI

The authors declare that AI tools were used only to improve grammar and readability of the manuscript originally written by the authors. AI was not used in the research design, data analysis, interpretation of results or any other part of the study.

REFERENCES

Ashik, F. R., Sreezon, A. I. Z., Rahman, M. H., Zafri, N. M., & Labib, S. M. (2024). Built environment influences commute mode choice in a global south megacity context: Insights from explainable machine learning approach. *Journal of Transport Geography, 116*, 103828. https://doi.org/10.1016/j.jtrangeo.2024.103828

Bray, D., & Holyoak, N. (2015). *Motorcycles in Developing Asian Cities: A Case Study of Hanoi.* Paper presented at the 37th Australasian Transport Research Forum, Sydney, Australia.

Ghosh, T., & Nagaraj, N. (2024). Evaluating the determinants of mode choice using statistical and machine learning techniques in the Indian megacity of Bengaluru. *arXiv (Cornell University).* https://doi.org/10.48550/arxiv.2401.13977

Herwangi, Y. (2018). Motorcycle Dependency in Low-Income People: Modeling Spatial and Socio-Economic Factors in Urbanized Area of Yogyakarta. *Plano Madani: Jurnal Perencanaan Wilayah Dan Kota*, *7*(2), 196–208.

Kilic, S. (2015). Binary logistic regression analysis. *Journal of Mood Disorders*, *5*(4), 191–194. https://doi.org/10.5455/JMOOD.20151202122141

Law, T. H., Ng, C. P., Poi, A. W. H., Guo, B., & Hu, Q. (2023). Motorcycle-to-car ownership at different urbanization levels: Road mobility, accessibility, and income inequality. *Transportation Research Part D Transport and Environment*, *122*, 103879.

Ma, J., & Ye, X. (2019). Modeling Household Vehicle Ownership in Emerging Economies. *Journal of the Indian Institute of Science*, *99*(4), 647–671. https://doi.org/10.1007/S41745-019-00133-9

Olaniran, O. R., Olaniran, S. F., Alzahrani, A., Alharbi, N. M., & Alzahrani, A. A. (2025). Random Forest Adaptation for High-Dimensional Count Regression. *Mathematics*, *13*(18), 3041. https://doi.org/10.3390/math13183041

Shirgaokar, M. (2012). *The rapid rise of Middle-Class vehicle ownership in Mumbai*. https://escholarship.org/uc/item/84v2140d

Yao, G., Chen, Y., Cui, K., Xu, D., & Liu, J. (2022). A method of household car ownership prediction using ordered Probit model. *Proceedings of the 11th International Conference on Green Intelligent Transportation Systems and Safety, 775*, 759–771. https://doi.org/10.1007/978-981-16-5429-9_57

Zisan, M. F. M., Rashid, T., Bhuyia, M. I., & Islam, M. A. (2024). Modeling Determinants of Car Ownership in Narayanganj: A Machine Learning and Statistical Approach. *Proceedings of International Conference on Civil Engineering Research & Innovations,* 1254-1259. https://icceiruet.org/wp-content/uploads/2025/12/Proceeding-ICCEI-2025.pdf (ISBN: 978-984-37-0506-8)