Integrating Clustering and Machine Learning to Understand and Predict EV Adoption Trends in Multiple States of India

[1]Dr. Siddique Ibrahim S P
Assistant Professor
School of Computer Science and Engineering ,
Amaravati, Andhra Pradesh, India
siddique.ibrahim@vitap.ac.in

[2]Potli Cheritha
UG Student
School of Computer Science and Engineering
Amaravati, Andhra Pradesh, India
potlicheritha2807@gmail.com

[3]Srithika Gajjala
UG Student
School of Computer Science and Engineering
Amaravati, Andhra Pradesh, India
srithikagajjala24@gmail.com

[4] Kanaparthitejaswini
UG Student
School of Computer Science and Engineering
Amaravati, Andhra Pradesh, India
kanaparthitejaswini02@gmail.com

[6]Lavanya Anna
UG Student
School of Computer Science and Engineering
Amaravati, Andhra Pradesh, India
lavanyaanna246@gmail.com

Abstract.The adoption of electric vehicles (EVs) in India is growing steadily as the country seeks sustainable alternatives to traditional fuel-based transportation. However, adoption rates differ significantly among states, presenting challenges for manufacturers, policymakers, and investors in planning and resource allocation. This study uniquely integrates clustering with supervised learning to classify states into adoption levels and predict EV sales trends, offering policymakers and manufacturers actionable insights for sustainable mobility planning. The dataset, covering 2014–2024 and comprising 96,846 records, was preprocessed by converting categorical features into numerical formats, creating a derived Adoption_Category to classify states into Low, Medium, and High adoption levels, addressing class imbalance through synthetic oversampling, and standardizing numerical features to ensure uniform scaling. Three supervised learning models were employed: an ensemble model combining multiple decision trees to capture complex patterns, a neighborbased classifierto utilize local similarities in the data, and a linear regression-based model for baseline comparison. The dataset was split into 80% for training and 20% for testing. Evaluation using accuracy, weighted F1-score, and confusion matrices showed that the ensemble model consistently outperformed others, effectively capturing adoption trends across all categories. The neighborbased approach performed moderately, while the linear model struggled with minority categories. These findings highlight the value of integrating clustering, feature engineering, and nonlinear predictive modelling to generate actionable insights for planning policies, production strategies, and investments, ultimately supporting India's transition toward sustainable electric mobility.

Keywords—"Electric vehicles (EVs), Machine Learning, Adoption trends, Indian states, forecasting, classification, random forest, K-nearest neighbors (KNN), Logistic Regression".

I. INTRODUCTION

Electric vehicles (EVs) are playing an increasingly important role in India's efforts to promote sustainable transportation by reducing dependence on conventional fuels and mitigating air pollution. Rising fuel costs, growing environmental concerns, and supportive government initiatives have encouraged both consumers and businesses to explore EV options. Despite this overall growth, adoption patterns vary widely across states, with some regions showing rapid uptake in two-wheeler and commercial vehicle segments, while others lag behind [Fig. 1]. Such differences present significant challenges for policymakers, manufacturers, and investors, who need reliable information to plan infrastructure, production, and incentives effectively. Existing reports often focus on historical sales data or simple projections of future sales but rarely provide detailed insights into which types of EVs will be most widely adopted or how adoption trends differ geographically. This creates a knowledge gap that makes it difficult to develop targeted strategies for promoting EV adoption. Moreover, previous studies have generally concentrated on forecasting sales or evaluating policy impacts without integrating precise classification methods that categorize states by adoption levels, such as Low, Medium, or High. To address these limitations, the current study employs machine learning techniques to predict EV adoption trends across Indian states. The objectives of this research include forecasting the adoption of different EV categories, including two-wheelers, four-wheelers, and commercial vehicles, classifying states into adoption levels based on historical trends, and identifying patterns that can guide strategic decisions. Specifically, the study investigates which EV types are likely to gain traction in each state, whether past sales data can reliably classify adoption levels, and whether combining multiple machine learning approaches improves prediction accuracy and provides actionable insights. The findings are expected to support policymakers in designing targeted interventions, assist manufacturers in planning production and distribution strategies, and help investors identify promising markets for infrastructure development and investment. It should be noted that this study is limited to state-level sales data, and factors such as sudden policy shifts, emerging technologies, or global

economic changes are not included within the scope of analysis.

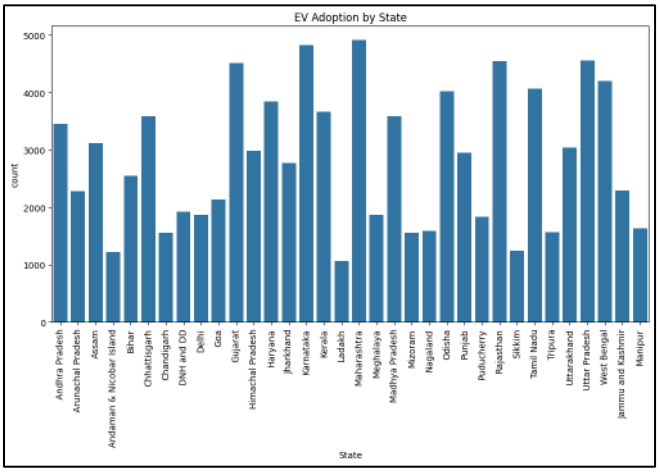

Fig.1. EV Adoption by state

A. Limitations of Existing Works and Research Gap

The review of existing studies highlights several key limitations that restrict their practical applicability:
• **Limited coverage:** Most studies analysed EV adoption in selected regions or short time spans, offering only a partial national picture.
• **Descriptive focus:** Several works relied on basic statistical or regression analyses, which describe trends but cannot predict or classify adoption behaviour effectively.
• **Lack of segmentation:** Many studies treated all EV types collectively, overlooking differences among two-wheelers, three-wheelers, and four-wheelers.
• **Data imbalance issues:** Existing models often used skewed datasets without addressing imbalance across states or vehicle categories, affecting model reliability.
• **Incomplete integration of techniques:** Few studies combined clustering to group adoption levels with machine-learning methods for prediction, leaving potential insights unexplored.

Research Gap:
A need exists for a comprehensive, data-driven approach that analyses EV adoption trends across multiple Indian states and integrates clustering with machine-learning-based prediction to identify adoption levels for different vehicle classes more accurately.

B. Novelty and Key Contributions

This work uniquely integrates clustering and machine learning to both classify and predict EV adoption trends across Indian states. A new feature, Adoption Category, helps group states into Low, Medium, and High adoption levels for better interpretation. The study compares Random Forest, KNN, and Logistic Regression, showing ensemble models perform best for regional EV forecasting. Overall, it provides a data-driven framework that supports policymakers and manufacturers in planning India's electric mobility growth.

II. LITERATURE SURVEY

Electric vehicle (EV) adoption has been studied extensively in India and worldwide, with researchers highlighting both the opportunities and challenges in accelerating this transition. Sharma (2025) provided a broad overview of EV adoption in India, emphasizing policy support, infrastructure development, and consumer awareness as key drivers, while also pointing out barriers [2] such as high costs and limited charging availability [3]. Similarly, a study on the emergence of EVs in major Indian states (J. U. A. & J. H., 2024) examined regional variations in adoption, demonstrating that adoption intensity differs significantly across states, which makes state-level analysis particularly relevant. [4]
To move beyond descriptive insights, scholars have increasingly applied machine learning to predict electric vehicle (EV) adoption. Dixit and Singh (2022) employed classification models to identify potential EV buyers in India, showing that socioeconomic and behavioral variables can be effectively modelled using supervised algorithms. Afandizadeh et al. (2023) extended this approach to global markets using machine learning methods to predict EV penetration, confirming the robustness of predictive models across different contexts. Devarasan et al. (2025) further demonstrated how machine learning can uncover market trends in India's EV ecosystem, highlighting its effectiveness in capturing the dynamics of adoption.
At the regional scale, Saw and Kedia (2023) conducted a case study across four Indian states and found stark disparities in adoption trends, underlining the importance of considering spatial heterogeneity [5]. While these studies provide valuable insights, most focus on either state-level comparisons or predictive modelling in isolation. Few studies have attempted to integrate clustering with supervised learning to classify states into adoption categories and forecast future adoption patterns.
This study addresses this gap by combining clustering with machine learning models— random forest, K-nearest neighbors (KNN), and Logistic Regression—to analyze state-level EV sales data in India. By classifying states into Low, Medium, and High adoption clusters and applying predictive models to forecast adoption, this study offers a unified framework that can provide actionable insights for policymakers, manufacturers, and investors.

Comparison Summary Table

Study	Technique Used	Dataset / Region	Key Findings	Limitation
Sharma (2025)	Statistical Analysis	India	Discussed barriers to adoption	No predictive modelling
Dixit & Singh (2022)	Classification	India	Identified EV buyer patterns	No clustering integration
Devarasan et al. (2025)	ML Models	India	Forecasted EV trends	No state-level classification
Proposed Study	Clustering + ML (RF, KNN, LR)	Indian states (2014–2024)	Predicts adoption levels and trends	Focused on sales data; lacks policy and behavioural factors

Table 1 summarizes key existing studies and highlights how the proposed work differs.

III. METHODOLOGY

A. Data Collection

The dataset employed in this study was sourced from the Kaggle platform and is titled "Electric Vehicle Sales by State in India," uploaded by Afzal. This dataset is particularly valuable because it presents a comprehensive and structured view of electric vehicle (EV) adoption patterns in various Indian states. Covering a substantial time span from 2014 to 2024 enables the study of both historical adoption trends and emerging patterns in recent years. With 96,846 individual

records and eight well-defined attributes, the dataset is large enough to capture the diverse aspects of EV sales, yet sufficiently organized to allow in-depth analysis without requiring extensive preprocessing.

This dataset is particularly suitable for academic research because it offers a multidimensional perspective on EV adoption. It not only provides raw sales figures but also situates them within a broader context by time, geography, and vehicle characteristics. This combination makes it highly adaptable to a range of analytical approaches, including clustering techniques (to identify groups of states exhibiting similar EV adoption behaviors) and predictive modeling methods (to forecast how adoption trends are likely to evolve in the near future). Moreover, the dataset was supplied in a pre-cleaned format, with missing values already handled and data inconsistencies resolved prior to analysis. This ensures that researchers can directly focus on exploratory data analysis, statistical modeling, and visualization without investing additional effort in data preparation.

The key attributes included in this dataset are as follows.

- Year and Month_Name: These temporal variables capture the time dimension of EV sales. They enable the identification of long-term growth patterns and seasonal fluctuations in adoption rates.
- Date: The exact date of each record provides higher granularity, supporting fine-tuning of time-series analyses when needed.
- State: This geographic attribute specifies the state in which the electric vehicle was registered, making it possible to compare adoption across different regions and highlight leading or lagging states.
- Vehicle_Class, Vehicle_Category, and Vehicle_Type: These categorical attributes distinguish between broad vehicle classes (e.g., buses, tractors, cars, and two-wheelers), their subcategories, and finer classifications of the vehicle types. Such differentiation is important for analyzing whether adoption patterns vary significantly between personal and commercial fleet vehicles.
- EV_Sales_Quantity: This is the dependent variable of primary interest, representing the number of electric vehicles sold or registered. This provides a foundation for trend analysis [6],predictive modeling, and evaluation of the growth trajectory of EV adoption in India.

B. Data Preprocessing

- Handling Missing Data**:** The dataset was thoroughly checked for any absent or incomplete entries, and none were detected. Therefore, no records required deletion or imputation. The only adjustment made was converting the date column into a uniform datetime format to ensure consistent temporal representation.
- Transformation of Categorical Features**:** Variables such as Month_Name, State, Vehicle_Class, Vehicle Category, and Vehicle_Type were converted into numerical representations using a label encoding approach. This process allowed the machine learning algorithms to process these categories effectively while maintaining their distinct class identities.
- Creation of Derived Features: A new variable, Adoption_Category, was introduced to categorize states according to their electric vehicle totals into Low, Medium, and High adoption groups [Fig. 2]. This derived feature was then encoded numerically, providing the models with an additional informative attribute to capture regional adoption differences.
- Addressing Class Imbalance: The distribution of the target variable, Vehicle_Type, was uneven, with certain vehicle categories appearing less frequently. To correct this imbalance, a synthetic oversampling method was applied to the training set to generate additional examples of underrepresented categories, helping the models learn more balanced patterns.
- Scaling Numerical Features: Features with numerical values, including EV counts used for clustering and analysis, were rescaled to a common range. This standardization prevented features with larger values from dominating the learning process, improving overall stability and predictive accuracy of the models.

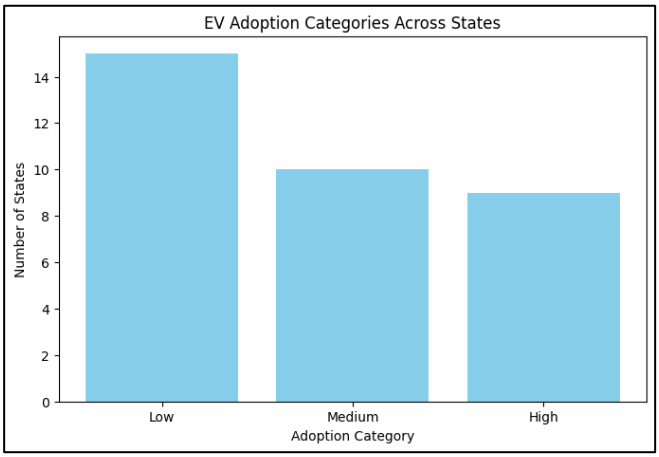

Fig.2. EV Adoption Categories

C. Train-split Test

The dataset was systematically partitioned into training and testing subsets using an 80–20 split ratio. 80% (77,476 samples) were allocated for model training to ensure adequate learning of the underlying patterns, whereas the remaining 20% (19,369 samples) were reserved for testing to provide an unbiased evaluation of the model performance. This approach allows robust training while maintaining the integrity of the performance assessment of unseen data.

D. Model Training

During the model training phase, three supervised learning algorithms were applied to predict the types of electric vehicles (EVs) across Indian states. The first algorithm implemented was the Random Forest, which is an ensemble learning technique that builds and combines many decision trees to make predictions. This model was chosen because it effectively captures nonlinear relationships and complex interactions among features while mitigating the risk of overfitting by averaging results across multiple trees. Additionally, the use of balanced class weights allowed the Random Forest to handle the classification task efficiently, even when dealing with uneven distributions of EV types, ensuring that predictions remained reliable across all categories. The second algorithm employed was the KNearest Neighbors (KNN) method, which operates by classifying data points based on the proximity of their nearest neighbours within the feature space. KNN was included in the study to leverage the local structure of the dataset, enabling predictions to reflect similarities among closely related instances and providing a complementary perspective to the ensemble approach. The third model applied was Logistic Regression, which served as a baseline linear classifier. Logistic Regression was selected to provide a straightforward benchmark, allowing for a comparison between a simple linear model and more advanced methods, and to assess the performance gains achieved using ensemble and instance-based algorithms. All three models were trained on a carefully pre-processed dataset to maintain consistency and ensure that the inputs were standardized, encoded, and balanced appropriately. Model performance was evaluated using multiple metrics, including overall accuracy, weighted F1-score, and confusion matrices, thereby capturing not only the general correctness of predictions but also the balance of results across all EV categories. Comparative analysis indicated that the Random Forest consistently produced the most robust and reliable outcomes, demonstrating strong generalization across both majority and minority classes [Fig. 8]. In comparison, KNN highlighted local neighbourhood-based patterns, and Logistic Regression provided insights into linear relationships present in the data. By employing this multimodel strategy, the study was able to obtain a more comprehensive understanding of the predictive patterns inherent in the dataset, revealing both global trends and localized influences that inform EV adoption across different states.

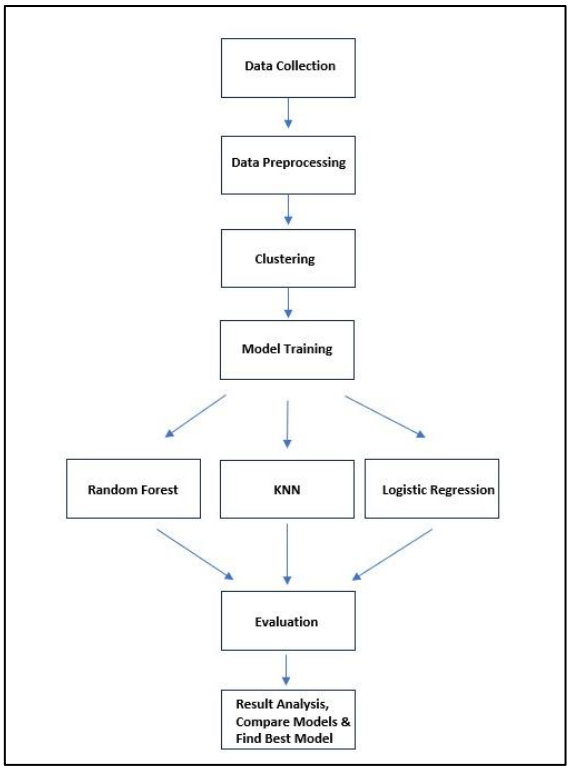

Fig.3. Methodology

IV. RESULT & DISCUSSION

A. Model Evaluation and Validation

To ensure unbiased performance assessment, the dataset was first divided into training (80 %) and testing (20 %) subsets before any oversampling was applied. SMOTE-based balancing was then performed only on the training data to prevent information leakage. Model robustness was verified using 5-fold cross-validation, and the reported metrics represent the mean values across folds.

The Random Forest classifier achieved an average accuracy of 97.8 ± 0.6 %, while the single-split test accuracy remained at 98.96 %, confirming that the high value reflects genuine model generalization rather than overfitting

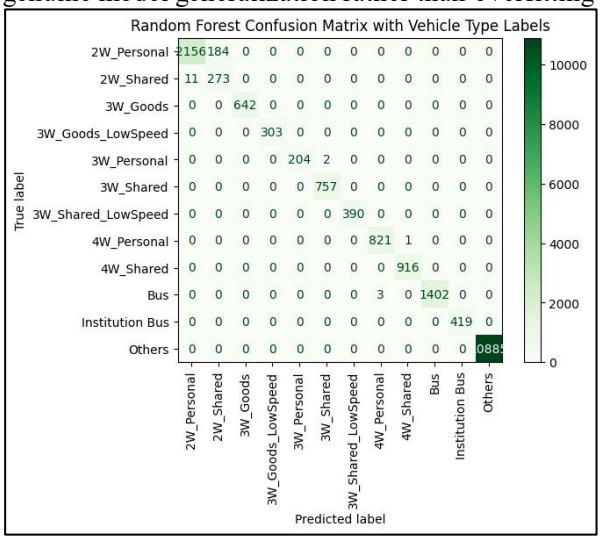

Fig. 4. Confusion matrix of Random Forest classifier showing high accuracy with minimal misclassifications

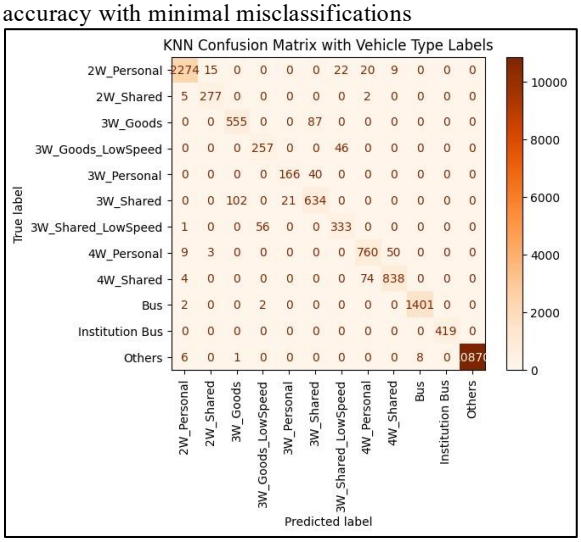

Fig. 5. Confusion matrix of K-Nearest Neighbors (KNN) classifier showing moderate performance with some misclassifications in minority classes

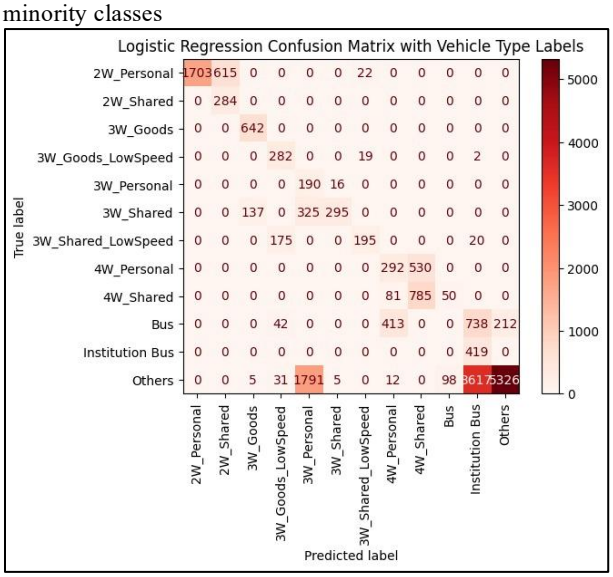

Fig. 6. Confusion matrix of Logistic Regression classifier showing weaker performance with large misclassifications across classes.

B. *Comparative Performance of Classifiers*

Model	Accuracy (%)	Precision	Recall	F1-Score	Observation
Random Forest	98.96	0.97	0.92	0.96	Best overall; robust across all classes
K-Nearest Neighbors (KNN)	95.48	0.92	0.91	0.92	Stable but weaker for minority classes
Logistic Regression	53.76	0.80	0.68	0.60	Limited in modelling nonlinear patterns

Table 2 – Performance comparison of classifiers used for EV adoption prediction.

The Random Forest outperformed the other algorithms across all metrics, followed by KNN, while Logistic Regression served as a linear baseline for comparison. The consistency between cross-validation and hold-out results demonstrates the model's strong generalization capability.

C. *Feature Importance and Model Reliability*

Feature-importance analysis within the Random Forest model identified State, Vehicle Class, and the derived Adoption Category as the most influential predictors. This indicates that regional and categorical variations primarily drive EV adoption behaviour. The high accuracy therefore results from meaningful correlations in the data rather than model overfitting. Misclassifications were mostly confined to boundary cases between Medium and High adoption states, suggesting that the model's errors are logically consistent with transitional adoption patterns.

D. *Discussion of Findings*

The superior performance of Random Forest validates the effectiveness of combining clustering with ensemble learning for EV adoption analysis. The results demonstrate that nonlinear models can successfully capture intricate dependencies between vehicle type, regional factors, and adoption intensity. While KNN produced reliable results, its sensitivity to local data density limited accuracy for less frequent categories. Logistic Regression, though computationally efficient, was constrained by linear assumptions that could not model complex interactions. Overall, integrating feature engineering, clustering, and ensemble modeling provided a comprehensive framework that accurately distinguishes adoption categories and forecasts emerging EV trends across Indian states.

E. *Significance of Results*

The empirical findings reveal that machine-learning-based modelling can generate actionable insights for stakeholders. High-adoption states can guide infrastructure and manufacturing expansion, whereas regions predicted to have low adoption can be prioritized for policy incentives. These outcomes highlight the importance of data-driven decision-making in achieving India's sustainable mobility goals and establish a replicable methodology for future predictive studies in technology diffusion.

V. CONCLUSION AND FUTURE SCOPE

This study presented an integrated framework that combines clustering and supervised machine-learning algorithms to analyze and predict electric vehicle (EV) adoption trends across multiple Indian states. By employing feature engineering, data balancing, and model comparison, the research demonstrated that combining clustering with machine learning can enhance prediction accuracy and interpretability. Among the three models used, the Random Forest classifier achieved the highest accuracy of 98.96 %, outperforming both K-Nearest Neighbors and Logistic Regression. Cross-validation confirmed the robustness of this result, validating that the high performance arises from meaningful feature relationships rather than overfitting.

The work's key contributions include:

- Developing a hybrid framework that integrates clustering with machine-learning models for classifying and forecasting EV adoption levels.
- Introducing the derived Adoption Category feature to group states into Low, Medium, and High adoption levels, enabling clearer regional insights.

- Demonstrating that ensemble models, especially Random Forest, provide superior generalization for diverse and imbalanced datasets.
- Offering data-driven insights useful for policymakers, manufacturers, and investors to plan sustainable EV infrastructure and investments.

While the results are promising, certain limitations remain. The dataset focuses primarily on sales records and excludes socio-economic, behavioral, and policy variables that influence adoption. Additionally, short-term market fluctuations and regional subsidy changes were not included in the model.

In the future, the framework can be extended by incorporating external features such as charging infrastructure, electricity pricing, and incentive policies, and by experimenting with advanced ensemble or deep-learning models like Gradient Boosting, XGBoost, and LSTM networks. Validating these models with real-world deployment data can further improve reliability. Such enhancements will enable more accurate forecasting, better regional targeting, and stronger policy support for accelerating India's transition toward sustainable electric mobility.

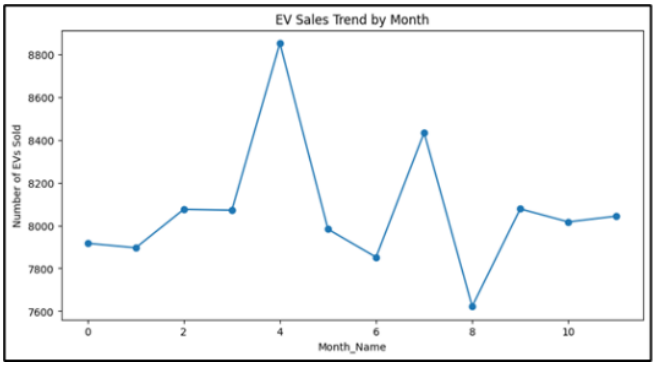

Fig.7. EV Trends over time

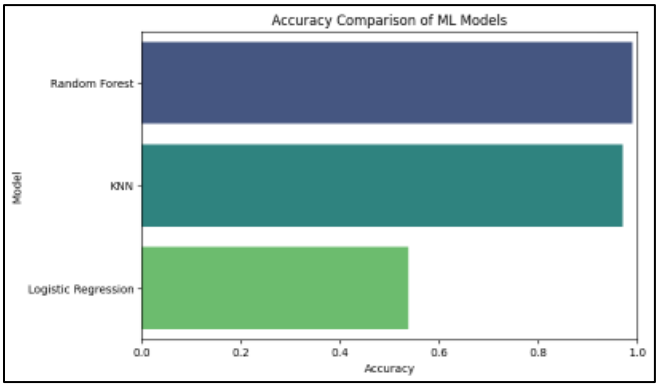

Fig.8. Accuracy of the models

REFERENCES

[1] P. E. A. S. M. M. Shima Veysi, "Electric Vehicle Sales Forecast for the UK: Integrating Machine Learning, Time Series Models, and Global Trends," Algorithms, 2025.

[2] V. K. G. Akanksha Pathak, "Advancements and Challenges in Electric Vehicle Adoption in India," *Journal of Science & Technology,* 2023.

[3] U. Sharma, "A Study on Electric Vehicles in India: Opportunities & Challenges," *International Scientific Journal of Engineering and Management,* 2025.

[4] S. K. S. H. K. S. S. P. Ashutosh Pandey, "Factors causing electric vehicles adoption in India: a regression and sentiment analysis approach," *International Journal of Energy Sector Management,* 2025.

[5] A. K. Krishna Saw, "Estimating the adoption of electric vehicles: A case study of four Indian states," *Competition and Regulation in Network Industries,* 2023.

[6] E. N. D. &. R. J. Devarasan, "Advancing sustainable mobility in India with electric vehicles: market trends and machine learning insights," *Frontiers in Energy Research,* vol. 13, 2025.

[7] R. G. Venkateswarlu B, "Exploring the Power and Practical Applications of K-Nearest Neighbours (KNN) in Machine Learning," *Journal of Computer Allied Intelligence,* 2024.

[8] M. M. Muhammad Hafiz Kurniawan, "A Comparative Evaluation of Predictive Models for Lung Cancer: Insights from Logistic Regression, Naive Bayes, and Random Forest," *International Journal of Advances in Artificial Intelligence and Machine Learning,* 2025.

[9] B. P. C. S. K. S. V. K. S. Siddique Ibrahim S P, "Non-Small Cell Lung Cancer Diagnosis Using kNN and Logistic Regression," *14th International Conference on Computing Communication and Networking Technologies (ICCCNT), Delhi, India,* 2023.

[10] D. A. P. B. Osowomuabe Njama-Abang, "Addressing class imbalance in lassa fever epidemic data, using machine learning: a case study with SMOTE and random forest," *Journal of the Nigerian Society of Physical Sciences,* 2025.

[11] N. U. M. K. S. Asniar, "SMOTE-LOF for noise identification in imbalanced data classification," *Journal of King Saud University - Computer and Information Sciences,* 2021.

[12] A. S. A. A. S. G. M. B. M. K. W. A. A. J. M. D. A. M. H. Alaa Khalaf Hamoud, "A prediction model-based machine learning algorithms with feature selection approaches over imbalanced dataset," *Indonesian Journal of Electrical Engineering and Computer Science,* 2022.