National College of Ireland

Optimizing placement of new EV charging stations in India using machine learning methodology

MSc Research Project
Data Analytics 2023-2024

Abhijit Singh
Student ID:x22157271

School of Computing
National College of Ireland

Supervisor: Jorge Basilio

National College of Ireland
Project Submission Sheet
School of Computing

National College of Ireland

Student Name:	Abhijit Singh
Student ID:	x22157271
Programme:	MSc Data Analytics
Year:	2023-2024
Module:	MSc Research Project
Supervisor:	Jorge Basilio
Submission Due Date:	12/08/2024
Project Title:	Optimizing placement of new EV charging stations in India using machine learning methodology
Word Count:	7637
Page Count:	21

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Abhijit Singh
Date:	12th August 2024

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Optimizing placement of new EV charging stations in India using machine learning methodology

Abhijit Singh
x22157271

Abstract

Automobiles have always been a crucial part of daily life of individuals. There is exponential increase in the daily commuters majorly in populated countries like India which has highlighted the limitations of conventional fuel-based vehicles in the future due to depleting resources. These fuel based vehicles are also harmful to the environment. The Government of India aims for 30% EV adoption by 2030 through initiatives like FAME-1 and FAME-2 but there is a poor adoption rate of EVs due to high costs and insufficient charging infrastructure that present challenges in adoption. This research proposes a machine learning framework using K-means clustering and HDBSCAN to optimize the placement of new EV charging stations across India by considering the existing infrastructure and socio-economic factors like per capita income, population density, and national highway density. The goal is to minimize waiting times and enhance user convenience, to support the government's EV adoption targets.

1 Introduction

1.1 Background

The growth of electric vehicles (EVs) is changing the transportation field, with a focus on developing and deploying charging infrastructure. As India aims for faster EV adoption through programs like FAME-1 and FAME-2, strategically placing charging stations is important for ease of consumers. Optimizing site selection in India faces challenges due to the socio-economic factors that are neglected in research analysis. EVs that are powered by lithium-ion batteries are known for their high energy density and efficiency that offer alternative to traditional fuel based vehicles. These batteries are preferred due to their high energy density, long life and relatively low self-discharge rates that makes them ideal for the EV use Jaguemont et al. (2016). Advancements in battery technology are continuously improving the energy density and efficiency of the battery Zhang et al. (2024).

1.2 Importance

As per the International energy agency (IEA) the automobile transport on fuel contributes about 24% of global CO2 emissions which makes cleaner alternatives like EV important Solaymani (2019). There is a need for a strong methods to identify optimal EV charging locations in India which a big and populous country with different socioeconomic factors,

high population density and different level of road connectivity. Addressing the above challenges is critical for a healthy EV adoption rate in consumers.To support widespread EV adoption in India, a well-distributed network of charging stations is essential for accessibility. EV do not produce emissions and if their source of electricity is of renewable in nature, the entire process will produce 50% fewer greenhouse gas emission than internal combustion engine vehicles Cornell (2017). According to the study of European Environment Agency (EEA) the wide spread adoption of EV can lead to considerable decrease in urban air pollution that will contribute to improved air quality and public health Soret et al. (2014). The International Energy Agency (IEA) reports that increasing the share of EVs in the global vehicle fleet is essential for achieving the targets set out in the Paris Agreement to limit global warming Bibra et al. (2021).

1.3 Research question and Objectives

The main aim of the research is to explore whether the machine learning methodology clustering algorithms can be effectively used to predict optimal location for new EV charging station to be setup in India based on the longitude and latitude data of the existing charging stations and socio-economic factors. Hence the research question which shall be used for analysis would be:

Can Clustering algorithms that considers factors such as per capita income, existing charging station location and population density, accurately predict the optimal location for a new EV charging station in India?

To answer the research question the data would be analysed for existing EV charge station in India along with different socio-economic factos like population density, per capita income and road network by implementing K means and Hierarchical Density-based spatial clustering of the application with noise (HDBSCAN) to locate and identify the charging location in India .

1.4 Contribution

The contribution that the results of this research will give is a data-driven approach for infrastructure planning for deployment of EV charging station in India by using the clustering algorithms involving socio-economic data.This will help the government for effective EV infrastrucure planning.

1.5 Structure of the Report

The research report will consist of meaningful sections that cover the in sights of the steps taken for the research:

1. Literature Review: The section would be providing an overview of the research on EV infrastructure planning and application of clustering algorithms performed in the domain.

2. Methodology: The details of the data sources, the steps taken for the preprocessing of data and the design of the clustering algorithm used for the study shall be documented in the section covering the entire structure and flow of the project

3. Results: The outcomes of the final clustering algorithm that includes the predicted location for the new EV charging station and their respective evaluation shall be presented under this section.

4. Discussion: The results obtained from the final algorithm shall be interpreted to understand their usefulness in the context of the aim of research and the current state of art and discussion of the practical implications.

5. Conclusion: This subsection will summarize all the key findings and focus on identifying the limitations of the research and suggest directions of the future work.

2 Related Work

2.1 Machine learning models for EV charging coordination

The author Shibl et al. (2020) compares seven different machine learning models which are decision tree (DT), random forest (RF), Support Vector Machine (SVM), Naïve Bayes (NB), K-Nearest Neighbours (KNN), Deep neural networks (DNN) and Long short-term memory (LSTM) to predict the power usage of the EV charging station (EVCS). The results highlight LSTM model achieves the highest accuracy (95%) in predicting power rating for the EVCS showing this model is best performing with a error rate of only $\pm 0.7\%$ and LSTM can be a very effective to handle the complex multi class classification problem in the EV charging field. Similarly, author Adhya et al. (2022) focuses on predicting the state of charge (SOC) of EV using three ML techniques CG-SVM, bagged tree regressor and boosted tree regressor. The research is based in Agartala, Indian and the author has used vehicle flow data from there. The CG-SVM model is best in performance in terms of mean squared error (MSE), mean absolute error (MAE), R-squared value, training time and root mean squared error (RMSE). The CS-SVM model has high prediction speed so it most suitable for predicting the EV charging demand.

2.2 Review on ML Applications in Charging Infrastructure

The paper Deb (2021) review the application of ML for the charging infrastructure planning that focus on station placement, demand predictions, and scheduling. Using three case studies in Helsinki, Finland the review highlights ML's effectiveness in optimizing decision-making and resource allocation of the EV infrastructure for good charging network

Lo Franco et al. (2023) emphasizes importance of accurate power demand forecasting for EV charging stations to address challenges like user behavior, infrastructure quality and initial state of charge. The methodology combines statistical pattern analysis, probabilistic modeling and ML behavior modeling to forecast power demand at individual and station levels. The results show differences in power demands across urban regions showing the model's accuracy and usefulness for grid integration and infrastructure planning.

2.3 Optimization Techniques and Infrastructure Planning

The author Reda et al. (2024) analysed the use of self-organizing maps (SOM) techniques for optimizing the placement and capacity of EV charging stations for the bus route

in Addis Ababa, Ethiopia. The author emphasized the application of SOM over the clustering algorithms for optimal locations and predictive models to forecast the demand of charging station so there is efficient resources allocation and infrastructure . The results show that the ML approach can improve planning of the EV charging network for better coverage of the area and user satisfaction. The approach is good for urban planning and development beacuse it increases the spread of charging infrastructure with the EV users. Hafeez et al. (2023) emphasises on the development of adaptive scheduling algorithms using machine learning to manage the load on EV charging stations by predicting state of charge peak demand and adjusting charging schedules so that the system can reduce the over use and improve the availability of the charging infrastructure.

2.4 Behavioral and Demand Prediction Models

In Shahriar et al. (2020) the author analyzed the use ML techniques to analyze EV user behavior which focused on charging patterns and their preferences by using the historical data and making predictive model to forecast user behavior that enables better demand management and personalized services. The result show potential of ML to make user defined charging solutions which are very essential for improving satisfaction and efficiency.

The authors Mazhar et al. (2023) proposed a demand response management system for EV charging that uses ML model to predict and manage charging demand for shifting loads away from peak hours to reduce stress on the grid. The findings show that ML based solution can balance supply and demand and improve grid stability and energy efficiency which is very important for EV adoption increases.

2.5 Clustering Techniques for EV Infrastructure

The author in Srividhya et al. (2024) introduces a novel approach for establishing smart charging hubs network for EV using dynamic K-means clustering method. The study analyze challenges like predicting EV usage patterns, inefficient load management at charging stations and integration of renewable energy sources. By the continuous analyzing of real-time charging data , dynamic clustering method adapting to the changing EV demand and effectively balancing EV distribution by reducing congestion and improving resource utilization the author aims to achieve user acceptance and satisfaction. In Kalakanti and Rao (2022), the author emphasizes on charging station placement problem and requirement estimation by comparing various machine learning and simulation-based solutions like K-means and Gaussian mixture models, with traditional methods. The research uses case studies from Austin, US and Bengaluru, India showing that these algorithms were able reduce the average distance between EVs and charging stations and offer urban planners effective strategies for optimal station placement and better demand estimation of EV users.

2.6 Conclusion

The literature review shoes a significant advancements in using machine learning for optimizing EV infrastructure that include power usage prediction, state of charge estimation and infrastructure planning. There is progress in developed regions but there is a gap in applying these methodologies to India's unique socio-economic and geographical

conditions. Hence the research aims to solve such problem by using K-Means and HDB-SCAN clustering algorithms and integrated exisitng charger data with socio-economic and geographic factors to predict optimal locations for EV charging stations in India. The results of previous research will help develop a ML framework for India that offers valuable guidance for urban planning and policymakers to boost EV adoption.

Prediction Topic	Approach	Accuracy (%)	RMSE	R squared	MSE	MAE	Training speed
Power Rating for EVCS	LSTM	95					
State of Charge (SOC)	CG-SVM		0.6144	1	0.3776	0.4887	2241
Charging Infrastructure Planning	Various ML techniques	95.17					
Power Demand Forecasting	Gradient Boosting Regression (GBR)		0.0174	0.9959	0.0003		
Bus Arrival Time Prediction	Random Forest, Gradient Boosting, Linear Regression, Lasso Regression, KNN, SVR		0.233	0.999	0.054	0.137	
State of Charge (SoC) Estimation	LSTM		0.49				
User Behavior and Charging Patterns	LSTM, ANN, k-NN, PSF, SVR, XGBoost, RF	98	0.44 (LSTM)			0.29 (LSTM)	
Response Management	LSTM						
Integration with Renewable Energy	SVM, LSTM	98.09 (SVM), 4% (LSTM vs. RF)					

Figure 1: Summary of related research results

3 Methodology

The methodology that has been implemented for the research follows the CRISP-DM (cross industry standard process for data mining). It starts with collecting the data related to research requirements, understanding and preprocessing the data, followed by exploratory data analysis, clustering analysis and finally visualization and evaluation of results. This process shall ensure a comprehensive and iterative process to achieve the optimal results for location new EV infrastructure location. The flow of the entire research can be seen in the figure 1 below.

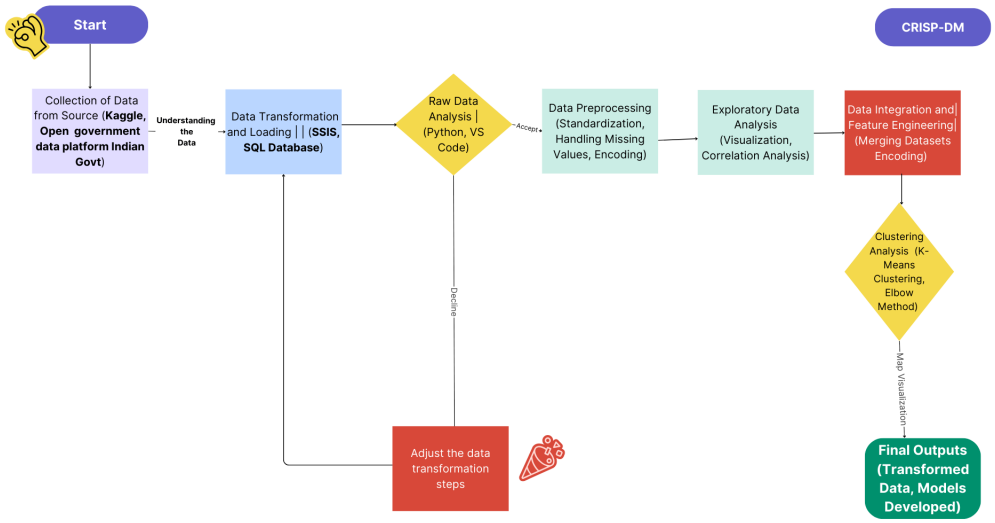


Figure 2: Research Methodology

3.1 Collection and loading of the Data from sources

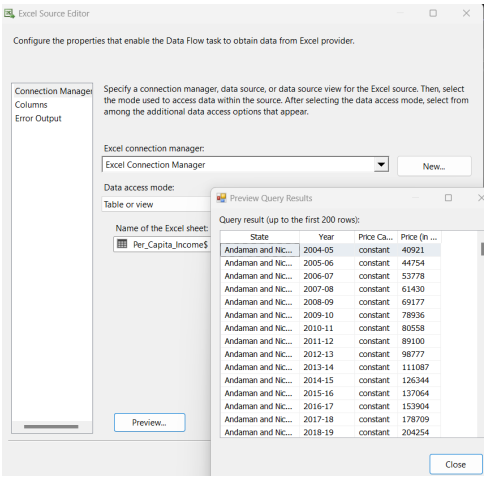
The datasets were gathered from different public data repositories. There were in total of 5 datasets that were gathered from various sources as stated1 below:

1. **Kaggle**: The existing location data i.e. longitude and latitude data for EV charging infrastructure in India consisting of 4800 unique records was gathered in the form of two CSV files from public data repository Kaggle.

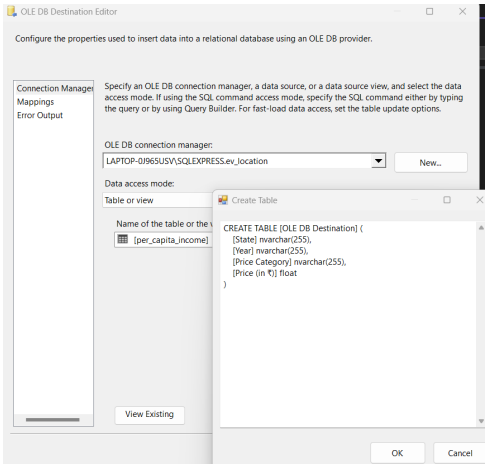
2. **Open Government Data Platform**: Data related to the state wise national highway, per capita income and population density was obtained by Government of India open data repository website (Open government data platform).

3.2 Data Transformation and loading

Using SQL server integration services which is a Microsoft powered Extract, transform and load (ETL) tool, a data pipeline was constructed using the SSIS to extract raw data from excel and feed it in the pipeline to later insert it into SQL tables for all the four data sets.Transformed data was then inserted into an MS SQL database tables ev_location, charger_one, charger_two, pop_density, national_highway, and per_capita_income.

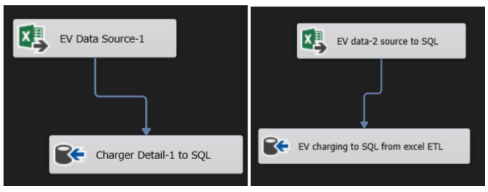


(a) Extraction from raw CSV files using SSIS

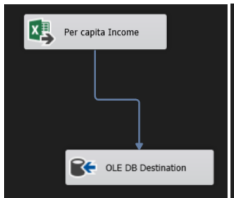


(b) Creation of SQL table and Mapping CSV with SQL table

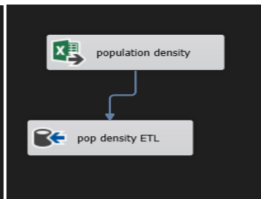
Figure 3: Data Processing Steps

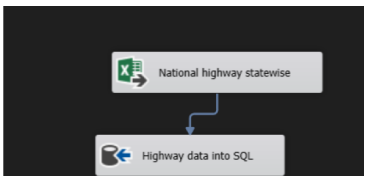


(a) SSIS Pipeline1



(b) SSIS Pipeline2





(c) SSIS Pipeline3

Figure 4: The SSIS pipelines for ETL process and loading data to SQL

3.3 Raw Data Analysis

Python was used to fetch the data from the SQL tables so that further processing can be applied to the data like EDA and suitable model application.

1. After the execution of the SSIS project the transformed data from CSV sources were migrated to SQL database in their respective table.

2. The SQL tables was accessed using Python programming language my the means of Vs code IDE

3. Python was used to pull the data from the SQL tables using the correct connection string from the database

4. The library pyodbc was used to establish connection between SQL and python for loading the data and perform actions on the data. It simplifies the process of connecting Open database connectivity(ODBC) and is suitable for SQL servers ,postgresql etc.

3.4 Data Prepossessing

The data after being extracted from the SQL tables was checked for inconsistency in feature names. Columns were renamed for the consistency across the dataset. State names were converted to upper case to maintain a uniformity. Numerical columns were converted to appropriate data types. Missing values in numeric columns were filled with the mean of the respective columns. Summary statistics were computed to verify the cleaning and check the correctness of data. Encoding of categorial features using label encoding like 'vendor_name', 'city', 'country', 'staff', 'payment_modes', 'station_type', 'zone', 'power_type', and 'vehicle_type'. The data was split into train and test dataset for model training and evaluation. Handling duplicate values in the data set by dropping the same values

3.5 Exploratory Data Analysis

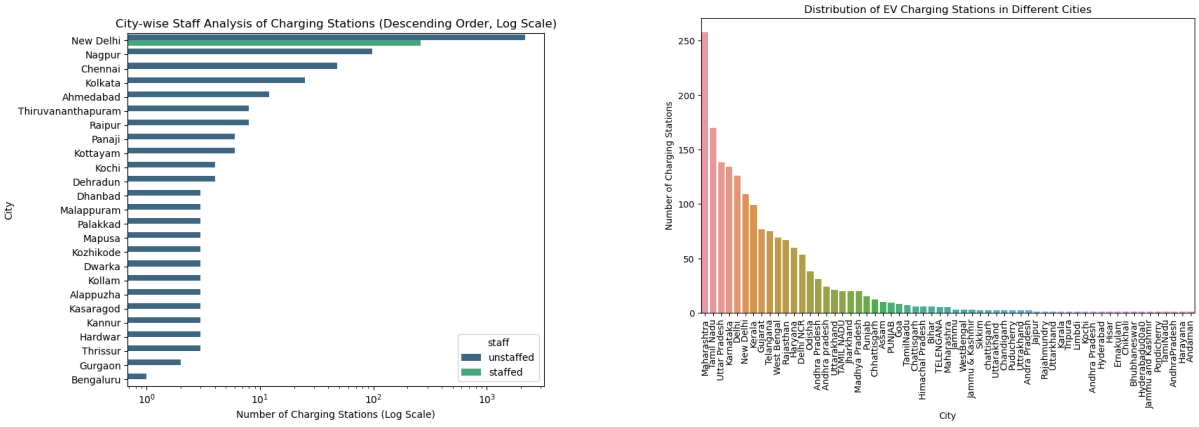
The datasets were then processed through individual exploratory data analysis stages to find the better insights from the datasets. EDA of the dataset provided insights into various aspects of EV charging stations across different states and city in India.

3.5.1 Distribution of charging stations across different regions in India:

The distribution of EV charging station spread across different cities and states in India was visualized using bar chart to understand the present number of charging station state and city wise. City wise New Delhi has the greatest number of charging station followed by Nagpur and Chennai. State wise Maharashtra has the greatest number of charging stations.

3.5.2 Distribution of charging station types

There are different types of charging stations based on the capacity of the chargers and other configurations. To understand the dominant type of charging stations a pie chart



(a) EV distribution city wise

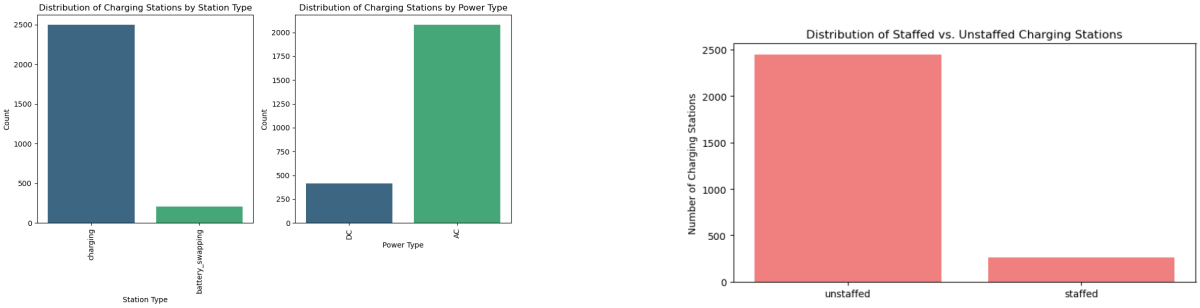
(b) EV distribution state wise

Figure 5: Distribution of EV infrastucture

was generated to display types of charging stations which showed that conventional charging is more widely used than battery swapping with 92.3% being charging station. Additionally supply based AC charging stations are more dominantly present than the DC charging units.

3.5.3 Staffed vs. Unstaffed Charging Stations

The plot to understand the number of staff and unstaffed charging station was created to understand the staff facility in general in India. In general, there were more unstaffed stations then staffed ones. This can be an important addition of introducing staff at the stations so that new users can be facilitated in case of any charging-based issues and general maintenance of charging infrastructure can be performed.



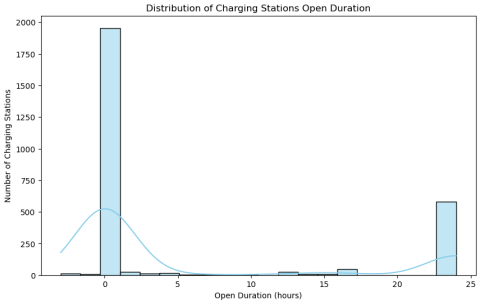
(a) Charger Type

(b) Staffed vs Unstaffed EV charging station

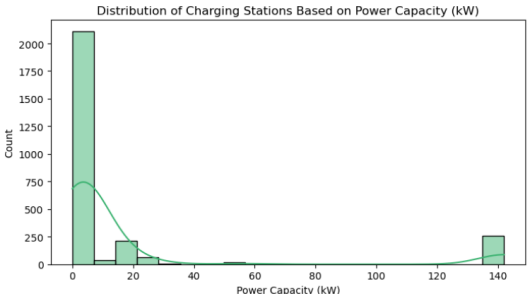
Figure 6: Existing Chargers

3.5.4 Power Capacity and Open Duration of Charging Stations

The distribution of power capacity (kW) of charging stations shows that most stations have a capacity of either 3.3 kW or 15 kW and a few having significantly higher capacities. A histogram was plotted that shows that the distribution of open duration (hours) of charging stations. Most stations were open 24 hours a day, with few variations.



(a) Open hours for EV

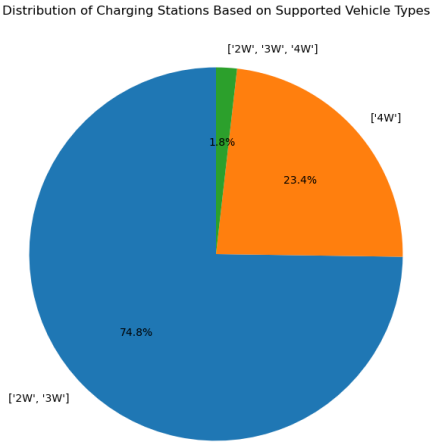


(b) Station Capacity Distribution

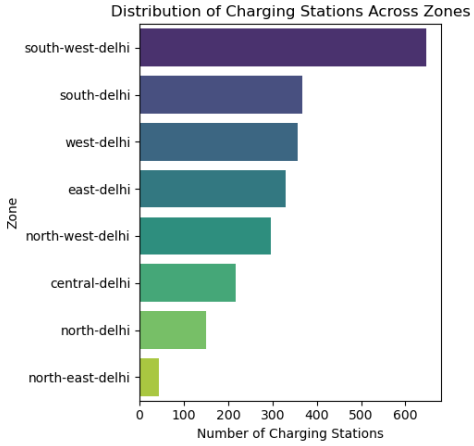
Figure 7: Station Capacity and Open hours Distribution

3.5.5 Supported Vehicle Type

To understand which are the types of vehicle which can be charged at the available charging station a pie chart of existing charging stations and their supported vehicle type was plotted. The pie chart reflected that majority of stations supported 4 wheelers followed by 2 and 3 wheelers.



(a) Distribution of supported vehicle type



(b) Distribution of type and charging staion in Delhi

Figure 8: Charging Station Distribution in New Delhi

3.5.6 Distribution of EV chargers in National Capital

To understand how well the EV infrastructure is distributed within the national capital of India New Delhi a bar plot was used to check the number of station available in different regions within Delhi that shows that south west Delhi had highest number of station.

3.5.7 Socio-Economic Factor- Per Capita Income, Population Density and State-wise Road contruction

The three graphs in Figure 9 and 10 provide the overview of India's economic and geographical division. The first graph shows difference in per capita income with states like Delhi, Goa, and Chandigarh leading in it highlighting economic disparities. The second graph shows that high population density in states like Bihar and West Bengal does not

go hand in hand with higher income, as it is seen between Delhi and Bihar. This indicates a economic difference suggesting that targeted policies are needed particularly in high population and low-income areas. High-income states having purchasing power are main targets for EV adoption while the densely populated regions are good for setting up EV infrastructure.The expanding road network as shown in the third graph, supports EV infrastructure growth to improve accessibility and connectivity for new charging station important for healthy EV adoption.

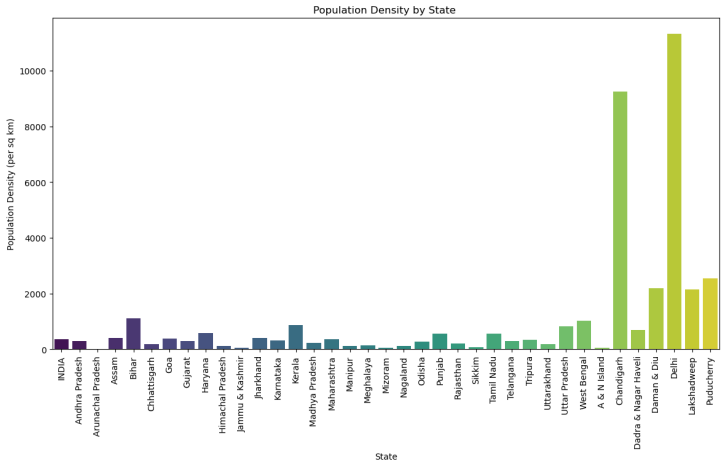
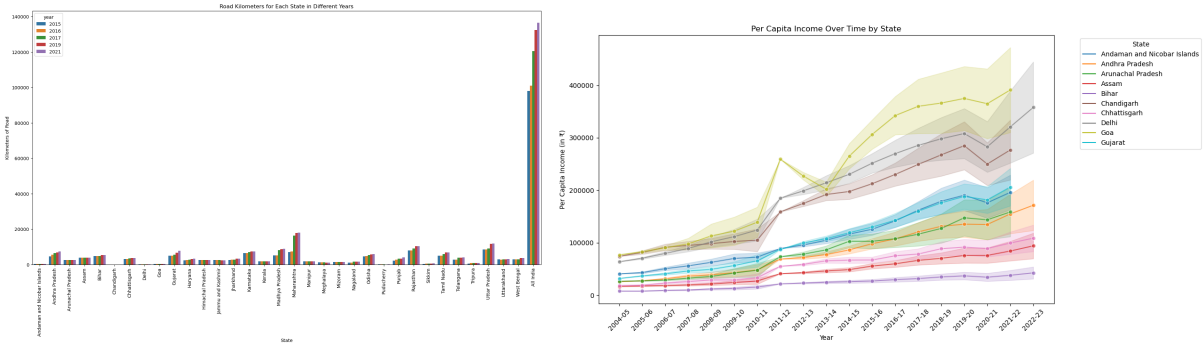


Figure 9: State-wise Population Density India



(a) Per-capita income statewise

(b) Year by Year Road expansion state-wise

Figure 10: Station Capacity and Open hours Distribution

3.6 Data Integration and Feature Engineering

3.6.1 Data Integration

In data integration phase the multiple datasets that were obtained from various sources of the EV charger placement analysis were merged to create a single dataset for analysis. The data sources included **EV Charging Station Data** which was data from two tables, charger_one and charger_two, containing information about different EV charging stations. **Population Density Data** which was data providing population numbers for various states in India.**National Highway Data** about the road network and its expansion year by year. **Per Capita Income Data** providing the per capita income for different states over year by year.

3.6.2 Data Cleaning and Standardization

The integration process involved the standardization of column names in all the data sets so that they are in the consistent format like making the feature name in lower case and removing the space in the name and replacing it with a _. Renaming of big column names like in population density and national highway table. It also involved the dropping of irrelevant columns such as contact number, URL etc.

3.6.3 Data Merger

After the cleaning of data merging of data from all the different tables were merged on the basis of their common table like ev_data1 and ev_data2 were merged on city and state column. Later this was merged with population density and national highway data on state column. Income data was aggregated by state population and income per person column was added and later merged with the existing combined data on the state column. Missing values were also removed from the merged data.

3.6.4 Feature Engineering

From the merged data new features were derived to have a deeper insight by the use of combined data and column.

- **Creating EV Per Capita**: Calculation of the number of EV stations per capita by dividing the total number of stations by the population of each state was done to understand the existing percentage of charger in ratio with the population
- **Handling Latitude and Longitude**: The lattitude and longitude columns were filled with non-null values wherever possible
- **Converting Data Types**: It was made sure that numeric fields were correctly typed to avoid errors while the calculations and visualizations.

3.7 Clustering Analysis

For the analysis considering the existing charging infrastructure location and socio-economic factors a algorithm that recognized meaning pattern to club similar data together was required so clustering algorithm was used to fulfil the requirement. Clustering technique is a fundamental technique in the field of unsupervised machine learning that aims to identify natural groupings and patterns in a data Kodinariya et al. (2013). It includes organizing the data points into clusters based on similarities, with the objective of maximizing intra-cluster similarity and inter-cluster separation. Clustering methods, such as K-means, hierarchical clustering, and density-based clustering are widely used in many real life domains including biology, psychology and economics. The main challenge in clustering is often determining the appropriate number of clusters, which is a critical analyzed by researcher in paper Kodinariya et al. (2013) .The goal of the clustering analysis is to identify optimal locations for setting up new EV charging stations based on socio-economic and geographical factors. The steps that were involved were as below:

- **Selecting Features**: Relevant features for clustering were identified and selected that included latitude, longitude,income per person and EV per capita feature that was generated by feature engineering and population density.

- **Scaling Features**: To bring the features in a standard form for modelling they were standardized using StandardScaler to ensure that each feature contributes equally to the distance calculations in clustering algorithms.

- **KMeans Clustering**: KMeans modelling technique was used for prediction of new EV location.Using the KMeans algorithm the task of dividing data into clusters was performed. The optimal number of clusters was determined using the elbow method.

3.8 Advanced Clustering for Detailed Analysis

Apart from the basic clustering technique of KMeans, an advance clustering technique of **HDBSCAN** was used for prediction of the new EV infrastructure. HDBSCAN is a density based clustering algorithm method that does not require every data point to be assigned to the cluster because it identifies the dense clusters and considers the points that is not assigned to a cluster as outliers or noise as highlighted by the author in paper Stewart and Al-Khassaweneh (2022). HDBSCAN algorithm offers a hierarchical representation of the clusters and it is also capable of identifying clusters of varying density, making it effective technique for cluster analysis and outlier detection. The outliers were identified by examining the cluster labels produced by HDBSCAN. Data points that did not belong to any cluster were labeled as -1 and they were considered outliers. These points were isolated and analyzed to understand their characteristics. Outliers were handled by HDBSCAN algorithm by not assigning them to any cluster and labeling them as noise. This prevents outliers from affecting the cluster formation for the EV charging station locations. In general the data realted to the socio-economic factors like population density had different ranges of value which were high and low giving rise to outliers.

.To give an advanced approach to the clustering analysis, the HDBSCAN algorithm was used. HDBSCAN is a more very advanced clustering technique that can handle noise and changing cluster densities that makes it suitable for complex datasets like EV location data set.

The steps that were involved were as follows:

- **Perform HDBSCAN Clustering**: HDBSCAN was applied to the scaled features like ev percapita, longitude and latitude to identify clusters and noise points within the data.

- **Extract Cluster Centers**: On appying the HDBSCAN clusters were given in the output. For each cluster identified by the HDBSCAN the mean of the cluster points was calculated to determine the cluster centers.

- **Transform Cluster Centers**: The cluster centers were transformed back to the original scale for meaningful interpretation so that the new ev charging location can be predicted by clusters.

3.9 Tools and Technologies

The predictive method for predicting the new EV infrastructure location utilized a variety of tools and libraries for the execution to handle various stages of data loading, processing,

visualization, and clustering. Below are some of the main key tools and technologies that were used for the research project:

- **Python**: Python was used for data manipulation, analysis, modeling, and visualization work in the VS Code IDE using its powerful libraries.
- **Microsoft SQL Server**: MS SQL was used as the primary database system for storing and managing the data. The data was stored within 5 tables which were charger_one, charger_two, pop_density, national_highwayL, per_capita income that were for Ev longitdue latitude, population density, national highway and per capita income data respectively.
- **SQL Server Integration Services (SSIS)**: SSIS which is a platform for data integration and workflow was used to import abd transform of data from various source CSV files for ev data, population density, national highway and per capita income data.

3.10 Final Outputs and Evaluation

The evaluation of performance of the models ensure that new predicted EV charging location were significant and optimal in their placement keeping in consideration the factors like population density, per capita income and exisitng EV infrastucutre along with the road network. The two clustering techniques of K-Means and HDBSCAN were used to identufy the new locations.This was done by the application fo K-means clustering which produced well defined clusters with a high Silhouette Score of 0.9648 which showed strong intra-cluster similarity and separated clusters. The low Davies-Bouldin Index also validated the effectiveness of the applied clustering, showing that the clusters were compact and distinct.The silhouette coefficient is the measure of how similar an object is to its own cluster compared to other clusters. It ranges from -1 to 1, where a high value indicates that the object is well matched to its own cluster and poorly matched to neighboring clusters Tambunan et al. (2020).Silhouette scores was used to evaluate the clustering results of new predicted EV charging stations to make sure that they were based on robust and well-separated clusters which gives effective infrastructure planning and implementation. The HDBSCAN algorithm which is advanced clustering algorithm which has ability to handle varying densities and noises was applied which provided additional insights more complex and diverse regions. While K-Means clustering was particularly effective in identifying clusters in urban and semi-urban settings the HDBSCAN had flexibility that allowed for identifying clusters in regions with more diverse characteristics. The Silhouette Score for HDBSCAN was even higher at 0.9951 that shows its capacity to identify tight and well-defined clusters but at the same time the Davies-Bouldin Index was higher that shows a broader spread of some clusters. The clusters generated by K-Means and HDBSCAN which were aligned with the socio-economic factors were mapped using the Folium library to make interactive visualizations of potential EV charging station locations across India.

4 Design Specification

The research project emphasises on a combination of data integration, feature engineering and advanced clustering techniques to predict the optimal locations for new EV charging

stations. The overall architecture of the project can be given as follows:

4.1 Data Integration

- The research uses multiple datasets including EV charging station data, population density data, national highway data, and per capita income data whicl are integrated and stored in **MS SQL Server**. **SQL Server Integration Services (SSIS)** was used to import, transform, and load data from CSV files into the SQL Server database. The **PyODBC** library in Python was used to connect the SQL Server and fetch the data to prepare it for preprocessing and modeling.

- Data is fetched from a CSV to SQL database using ETL process and SSIS tool. The data is then imported from SQL database into python using pyodbc library for further processing.

- Pandas library is used to merge these different datasets into a single dataset for data preprocessing and modelling

4.2 Data Preprocessing and Feature Engineering

To utilize the merged data in a better way feature engineering was performed on the merged data where new features such as EV per capita and income per person were derived to provide more deeper insights. Missing values in dataset were handled using imputation using the simple imputer. After this the process of encoding of feature was done to maintain a standard form using standard scaler to normalize the features. Apart from the general libraries for data processing like pandas, matplotlib and seaborn for data visualization and performing arithmetic operation on data like numpy and sklearn standard scaler for scaling the data. The main python libraries that were used for the execution the research project were as below:

- **Kmeans**: The Kmeans component of Sklearn was used for implementing KMeans clustering model on the EV dataset which is a method of partitioning data into clusters by minimizing variance within each cluster. Pedregosa et al. (2011).

- **HDBSCAN**: This library was used for applying advanced clustering analysis, HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise) on the EV dataset which is a clustering algorithm that can find clusters of varying densities and handle noise effectively. Rahman et al. (2016)

- **Folium**: This is a library in python that is used to create interactive maps that visualize the geographical distribution of the identified clusters from the clustering method hence providing a clear way to interpret the spatial data of longitude and latitude of new EV infrastructure.Rajamani and Iyer (2023)

4.3 Clustering Analysis

The **KMeans** clustering algorithm is used to identify optimized locations for the new EV charging stations on the basis of socio-economic factors and existing EV charging station. The main features of latitude, longitude, EV per capita, income per person, and population density were selected and standardized using StandardScaler to make a equal

contribution in the clustering process. KMeans was applied to the data with 1,500 clusters and the resulting clusters were created and scaled back to the original values. These were stored for analysis and visualization. These clusters were finally plotted on a map of India using the Folium library to check the location visually and provide data driven decision making for EV infrastructure development for policy makers. **HDBSCAN** which is advanced density based clustering method which was also applied to further analyze by considering the noise and varying data densities. The same main features were standardized and HDBSCAN was used with a cluster size of 50 to identify clusters and noise points. The cluster centers obtained were averaged and transformed back to the original scale and the results visualized on an interactive map using Folium.

5 Implementation

The final stage of the implementation emphasized on solving and finding answer to the research question: **"Can clustering algorithms that considers factors such as per capita income, existing charging station location, and population density accurately predict the optimal location for a new EV charging station in India?"** . The detailed description of the research steps and what are the tools and technologies used for the outputs that are generated are mentioned in the section below.

5.1 Final Output

5.1.1 Extraction and Transformation of Data

Data from multiple sources (existing charger data, population density, national highways, and per capita income) was integrated and imported from CSV to SQL using SSIS ETL tool and after that data was cleaned and prepossessed using Pandas and PyODBC.

5.1.2 Code in python

Python code was used for data extraction, transformation, cleaning, merging, modeling, and evaluation. Important libraries used were Pandas, NumPy, and Scikit-learn. Clustering algorithms of KMeans and HDBSCAN were implemented to identify optimal EV charging locations.

5.1.3 Models Developed

- **KMeans** clustering model was deployed to the data consisting of features like latitude, longitude of existing EV , EV per capita and per person income and which was scaled earlier using Standardscalers to identify 1500 optimal locations for new EV charging stations.
- **HDBSCAN** is Hierachical Density based spatial clustering of application with noise is a model was use as advanced clustering algorithm for identifying number of clusters based on the densities of data points which is useful in identifying the clusters in dataset with changing densities.Latitude, longitude, EV per capita, income per person, and population density features same as KMeans were selected for the HDBSCAN analysis. These features were standardized using StandardScaler

encoding. Unlike Kmeans there is no required of predefined number of clusters in HDBSCAN.

- Evaluation of clustering models and their result was done using **Silhouette scores** and **Davies-Bouldin** score to ensure the quality and precision of the clusters.The silhouette coefficient is the measure of how similar an object is to its own cluster compared to other the clusters. The score ranges from -1 to 1 in which a high value indicates that the object is well matched to its own cluster and poorly matched to neighboring clusters. The silhouette coefficient is used to determine the optimal number of clusters in the dataset Yuan and Yang (2019). The Davies-Bouldin index is the measure of average similarity between each cluster and its most similar cluster which is relative to the average dissimilarity between points in the cluster and points in its most similar cluster. The lower value of this score shows better the clustering.Tempola and Assagaf (2018)

- Interactive maps were created using Folium library that displayed potential EV charging station locations across India.

5.1.4 Insights and Reports

The clustering analysed and identified optimal EV charging locations in India which will help policymakers for targeting high-demand areas based on socio-economic factors. This approach can help to allocate EV resources efficiently and addressing gaps in existing infrastructure and guiding data-driven expansion of EV charging networks.

6 Evaluation

6.1 Case Study 1: KMeans Clustering

6.1.1 Results

The KMeans clustering method identified many optimal locations for new EV charging stations across India by grouping of the data points on the basis of socio-economic factors and existing locations. This method predicted majorly region from urban and semi-urban areas in India. These clusters were based on per capita income, population density, road infrastructure and existing charging stations.

6.1.2 Analysis

The KMeans method effectively highlighted regions with high economic activities and population density that shows these areas may have higher demand for EV infrastructure in upcomming time. The silhouette score of 0.9648 indicates well-defined clusters covering areas with possible high demand of EV infrastructure based on all the factors. However the clustering majorly identified the urban and semi urban areas which can be seen on map. Additionally, the Davies-Bouldin Index of 0.0394 reflects compact clusters with less amount scattering of points.

6.2 Case Study 2: HDBSCAN Clustering

6.2.1 Results

HDBSCAN identified clusters of potential EV charging station locations with a different clustering approach by handling different data densities and noise without the predefined number of clusters. The clusters formed by HDBSCAN showed new prediction from the data not only considering the urban areas with high population and income but also the normal rural areas.

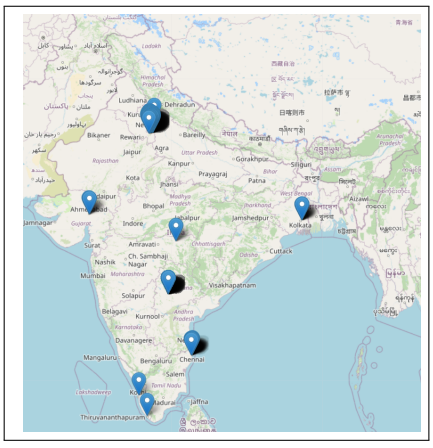
6.2.2 Analysis

HDBSCAN's flexibility allowed a detailed understanding of the spatial distribution for potential charging stations. HDBACAN results further validated the research problem by identifying new EV location apart from the ones identified my Kmeans by focusing on the rural areas as well. At the same time it achieved a silhouette score of 0.9951 meaning that the clusters are very tight and well-defined. On the other hand, a Davies-Bouldin Index of 0.6547 indicates that the clusters are more spread out and not as compact as those formed by KMeans.

6.3 Discussion

The experiments efficiently shows that there is a potential of clustering algorithms for predicting optimal locations of new EV charging stations in India. KMeans method gave a well-defined clusters in high-demand areas, while HDBSCAN offered a new pattern analysis. However, the analysis faced limitations due to its static nature, data quality, and geographical differences in each area. Future work should make use of dynamic data with improved data detail and multilevel-criteria decision analysis. Developing a adaptive models to respond to the changing socio-economic conditions can also be important for long-term planning for future work.

(a) Kmean output of predicted locations



(b) HDBSCAN output of predicted locations

Figure 11: Predicted Locations

Criterion	KMeans	HDBSCAN
Number of Clusters	2500	Variable (based on density)
Silhouette Score	0.9648	0.9951
Davies-Bouldin Index	0.0394	0.6547
Outlier Detection	Not directly handled	Directly identifies outliers
Handling of Density Variations	Less effective as requires predefined clusters	More effective (density-based)
Cluster Visualization	Tight clusters, many points	More new pattern clusters, fewer points
Computation Time	Faster Results	Slower in processing due to complexity

Figure 12: Results

7 Conclusion and Future Work

The research was aimed to determine if clustering algorithms by including factors like per capita income, existing charging stations,road infrastucture and population density can accurately predict optimal locations for new EV charging stations in India.The reseach was able to successfully identify potential locations for new EV charging stations using KMeans and HDBSCAN clustering algorithms. The results highlighted the regions in India with high demand potential based on socio-economic and geographic factors.

7.1 Key Findings

- **KMeans Clustering**:The KMeans algorithm identified several cluster centers in India showing the optimal locations for setting up new EV charging stations. These locations were majorly concentrated in urban and semi-urban areas due to their higher population densities, greater per capita income, and existing EV infrastructure. The high Silhouette Score (0.9647) and low Davies-Bouldin Index (0.0394) indicate that the clusters that were formed are well-defined having clear separation between clusters and high relationship within clusters.

- **HDBSCAN Clustering**: The HDBSCAN algorithm also identified potential locations for new EV charging stations. The clusters were less densely packed compared to KMeans and they were able to highlight more new patterns, by not considering the common locations that may not have been identified by KMeans. The high Silhouette Score for HDBSCAN (0.9951) suggests that the clusters are more well-defined compared to KMeans and higher Davies-Bouldin Index (0.6547) indicates more dispersion within the clusters.

7.2 Implications

The findings shall help the policymakers and urban planners in strategically expanding EV charging infrastructure thorughout India.The research shows that there is a difference in level of distribution of potential EV charging location among the rural and urban areas throughout India as the clustering algorithm considering the data and socio-economic factors majorly predicting the urbanized regions. This shows a potential challenge by geographical and infrastructural level in rural and less developed areas where there are low population and lack of infrastructure so the EV stations in such areas will be useless. The findings show that the clustering

algorithms can be effective in predicting optimal locations in areas of high demand but at the same time it shows that there is need to introduce more criteria an data input to capture all the different geographical contexts in India. The research has set up a stage for the future work that shall include dynamic data and more criteria from even rural segments to accurately predict EV infrastructure planning.

7.3 Limitations

- Dependence on data quality and scale for meassuring as new data is upcomming but there is still limited data.
- Geographic and infrastructural constraints of hilly terrains and deserts or other areas where setup is not possible should not be considered .
- The analysis carried out was static in nature performed on the historic data and there are needs to involve the dynamic changing data.

7.4 Future Work

- Incorporating Dynamic Data: For future developments , new ev infrastructure and population growth dynamic data can be considered.
- Enhanced Data Collection: Improve scale and coverage of data should be introduced for future for better quality of analysis.
- Multi-Criteria Decision Analysis: Combine clustering with additional factors like land availability and grid capacity.
- Integration with Transportation Models: Refine site selection based on travel patterns.

7.5 Conclusion

The research was successful in demonstrating that the clustering algortihms KMeans and HDBSCAN were able to effectively predict the optimal location for the new EV charging stations in India by integrating socio-economic and geographical factors of per capita income, population density and existing charging stations locations. The research finding make a strong allignment with the research question which shows that data driven approach is useful tool to plan the expansion og EV infrastucture in India strategically. The research highlights that the effectiveness of clustering algorithm in identifying new location in urban and sub urban areas where there is decent economic activity and population desisty. The research also shows the limitations of EV exapnsion in rural and less devloped areas as the clustering algorithm was not too effective as there were low population density, economic activity and insufficient infrastructure in these areas. This suggest Kmeans and HDBSCAN were powerful in urban areas prediction but not too effective in rural areas context. This opens a scope for future work that can integrate more dynamic data and multiple more criteria in the analysis for better decision making. The data collection methods can also be improved to aquire better quality data. This research was able to contribute new suggestion in devlopement strategies related to expansion of EV in India.

References

Adhya, D., Pal, A., Chakraborty, A. K. and Bhattacharya, A. (2022). Machine learning application for prediction of ev charging demand for the scenario of agartala, india, *2022 4th International Conference on Energy, Power and Environment (ICEPE)*, IEEE, pp. 1–5.

Bibra, E. M., Connelly, E., Gorner, M., Lowans, C., Paoli, L., Tattini, J. and Teter, J. (2021). Global ev outlook 2021: Accelerating ambitions despite the pandemic.

Cornell, R. P. (2017). *The environmental benefits of electric vehicles as a function of renewable energy*, PhD thesis.

Deb, S. (2021). Machine learning for solving charging infrastructure planning problems: A comprehensive review, *Energies* **14**(23): 7833.

Hafeez, A., Alammari, R. and Iqbal, A. (2023). Utilization of ev charging station in demand side management using deep learning method, *IEEE Access* **11**: 8747–8760.

Jaguemont, J., Boulon, L. and Dubé, Y. (2016). A comprehensive review of lithium-ion batteries used in hybrid and electric vehicles at cold temperatures, *Applied Energy* **164**: 99–114.

Kalakanti, A. K. and Rao, S. (2022). Charging station planning for electric vehicles, *Systems* **10**(1): 6.

Kodinariya, T. M., Makwana, P. R. et al. (2013). Review on determining number of cluster in k-means clustering, *International Journal* **1**(6): 90–95.

Lo Franco, F., Ricco, M., Cirimele, V., Apicella, V., Carambia, B. and Grandi, G. (2023). Electric vehicle charging hub power forecasting: a statistical and machine learning based approach, *Energies* **16**(4): 2076.

Mazhar, T., Asif, R. N., Malik, M. A., Nadeem, M. A., Haq, I., Iqbal, M., Kamran, M. and Ashraf, S. (2023). Electric vehicle charging system in the smart grid using different machine learning methods, *Sustainability* **15**(3): 2603.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V. et al. (2011). Scikit-learn: Machine learning in python, *the Journal of machine Learning research* **12**: 2825–2830.

Rahman, M. F., Liu, W., Suhaim, S. B., Thirumuruganathan, S., Zhang, N. and Das, G. (2016). Hdbscan: Density based clustering over location based services, *arXiv preprint arXiv:1602.03730* .

Rajamani, S. K. and Iyer, R. S. (2023). Use of python modules in ecological research, *Perspectives on the Transition Toward Green and Climate Neutral Economies in Asia*, IGI Global, pp. 182–206.

Reda, H., Mohapatra, S. K., Das, T. K. and Dash, S. K. (2024). Electric bus arrival and charging station placement assessment using machine learning techniques, *International Journal of Sustainable Engineering* **17**(1): 1–17.

Shahriar, S., Al-Ali, A.-R., Osman, A. H., Dhou, S. and Nijim, M. (2020). Machine learning approaches for ev charging behavior: A review, *IEEE Access* **8**: 168980–168993.

Shibl, M., Ismail, L. and Massoud, A. (2020). Machine learning-based management of electric vehicles charging: Towards highly-dispersed fast chargers. energies. 13,(2020).

Solaymani, S. (2019). Co2 emissions patterns in 7 top carbon emitter economies: The case of transport sector, *Energy* **168**: 989–1001.

Soret, A., Guevara, M. and Baldasano, J. (2014). The potential impacts of electric vehicles on air quality in the urban areas of barcelona and madrid (spain), *Atmospheric environment* **99**: 51–63.

Srividhya, V., Gowriswari, S., Antony, N. V., Murugan, S., Anitha, K. and Rajmohan, M. (2024). Optimizing electric vehicle charging networks using clustering technique, *2024 2nd International Conference on Computer, Communication and Control (IC4)*, IEEE, pp. 1–5.

Stewart, G. and Al-Khassaweneh, M. (2022). An implementation of the hdbscan* clustering algorithm, *Applied Sciences* **12**(5): 2405.

Tambunan, H. B., Barus, D. H., Hartono, J., Alam, A. S., Nugraha, D. A. and Usman, H. H. H. (2020). Electrical peak load clustering analysis using k-means algorithm and silhouette coefficient, *2020 International Conference on Technology and Policy in Energy and Electric Power (ICT-PEP)*, IEEE, pp. 258–262.

Tempola, F. and Assagaf, A. F. (2018). Clustering of potency of shrimp in indonesia with k-means algorithm and validation of davies-bouldin index, *International Conference on Science and Technology (ICST 2018)*, Atlantis Press, pp. 730–733.

Yuan, C. and Yang, H. (2019). Research on k-value selection method of k-means clustering algorithm, *J* **2**(2): 226–235.

Zhang, G., Chen, X., Ma, Y., Huo, H., Zuo, P., Yin, G., Gao, Y. and Fu, C. (2024). Recent advances and practical challenges of high-energy-density flexible lithium-ion batteries, *Frontiers of Chemical Science and Engineering* **18**(8): 91.