Sri Lanka Institute of Information Technology

Optimizing Electric Vehicle (EV) Charging Station Placement

Statement of Work

Mini Project By

The_Classifiers

Fundamentals of Data Mining (IT3051)

2025

Table of Contents

1. Background

The global shift towards sustainable transportation has led to a rapid increase in electric vehicle (EV) adoption. As governments and consumers prioritize reducing carbon emissions and dependence on fossil fuels, EVs are becoming a cornerstone of modern mobility. However, one of the primary barriers to widespread EV adoption is the lack of adequate charging infrastructure. Range anxiety, the fear of running out of battery power without access to a charging station—remains a significant concern for potential EV owners. According to the International Energy Agency (IEA), the number of EVs on the road is expected to reach 145 million by 2030, necessitating a proportional expansion of charging networks to support this growth.

Traditional approaches to infrastructure planning often rely on broad demographic data or reactive measures, such as installing stations after demand surges. These methods can lead to inefficient resource allocation, with some areas overserved while others remain underserved. Data-driven solutions offer a more proactive alternative. By analyzing patterns in EV ownership, usage, and geographic distribution, machine learning (ML) can identify high-density areas of EV owners and predict optimal locations for new charging stations. This not only enhances user convenience but also promotes equitable access to sustainable transport options.

Machine learning has proven effective in spatial analysis and predictive modeling within the transportation sector. Unlike conventional statistical tools, ML algorithms can handle large datasets with geospatial features, such as vehicle locations, postal codes, and census tracts, alongside attributes like vehicle type, model year, and electric range. Techniques like clustering (e.g., K-Means) can group EV owners by location to reveal hotspots, while regression models can forecast future demand based on trends in adoption. The explainability of these models ensures stakeholders can understand why certain cities are prioritized, fostering trust in the recommendations.

The dataset selected for this project, the Electric Vehicle Population Data (containing records on VIN, county, city, state, model year, make, model, EV type, electric range, and more), is ideal for this challenge. It provides a comprehensive view of EV registrations in Washington State, including location-based features like cities, counties, and vehicle coordinates. The dataset's mix of categorical (e.g., make, model) and numerical (e.g., electric range, model year)

data allows for real-world data mining tasks such as geospatial aggregation, outlier detection, and feature correlation analysis. While focused on Washington, the insights can be generalized to other regions experiencing EV growth.

By leveraging this data with machine learning, this project aligns with the broader goal of sustainable infrastructure development. Predicting suitable cities for new charging stations based on EV owner density not only offers practical experience in data mining and geospatial analysis but also has real-world applications in urban planning, energy policy, and environmental conservation. Although not intended to replace expert infrastructure assessments, this project demonstrates the role of data mining and ML in optimizing EV ecosystems, supporting green initiatives, and accelerating the transition to electric mobility.

We selected this project because it intersects critical areas: sustainable energy, data science, and urban development, where tangible impact is achievable. EV infrastructure gaps hinder adoption, but data-driven predictions can guide efficient expansion. Our aim is to build skills in data preprocessing, modeling, and evaluation while contributing to a socially relevant field. This project provides an opportunity to explore technology's role in eco-friendly transportation, empower decision-makers with insights, and inspire future innovations in smart cities.

1.1 Project Summary

This mini project aims to leverage the Electric Vehicle Population Data to predict optimal cities in Washington State for the placement of new EV charging stations, based on the density and distribution of EV owners. With the rapid rise in EV adoption driven by sustainability goals, the lack of adequate charging infrastructure poses a significant barrier. The dataset, containing detailed records of EV registrations including city, county, vehicle type, and location coordinates, provides a rich foundation for geospatial analysis. Using machine learning techniques such as clustering, the project will identify high-density areas of EV ownership to guide infrastructure planning.

The project will process and analyze the data to generate actionable insights, focusing on aggregating EV locations by city and visualizing density hotspots. This data-driven approach will enable prioritization of cities with the greatest need and potential impact, supporting urban planning and environmental initiatives. While not intended to replace comprehensive infrastructure studies, this project offers a practical learning experience in data mining, geospatial analysis, and predictive modeling, with real-world applications in enhancing EV accessibility.

We selected this project because it bridges data science with sustainable transportation, addressing a pressing need for efficient charging networks. By building skills in data preprocessing, model development, and interpretation, the team aims to contribute to eco-friendly urban development while exploring technology's role in shaping future mobility solutions.

1.2 Objectives (SMART)

- **Specific**: Identify the top 10 cities in Washington State with the highest density of electric vehicle (EV) owners based on current registration data to recommend locations for new charging stations.

- **Measurable**: Achieve a clustering accuracy of at least 85% using a machine learning model (e.g., K-Means or DBSCAN) to group cities by EV owner density, validated through silhouette scores or similar metrics.

- **Achievable**: Utilize the provided dataset and standard data mining tools (e.g., Python with Pandas, Scikit-learn, and Geopandas) to preprocess data and develop a predictive model within the project timeline, leveraging the team's existing skills.

- **Relevant**: Support sustainable transportation initiatives by providing data-driven insights that address the growing need for EV charging infrastructure, aligning with environmental and urban planning goals.

- **Time-bound**: Complete the identification and ranking of high-density cities, along with model development and evaluation, within a 5-week period starting from September 20, 2025, with a final deliverable due by October 25, 2025.

1.3 Functional Requirements

- **Data Ingestion**: The system must import and process the Electric Vehicle Population Data CSV, including fields such as City, County, Vehicle Location, Model Year, and Electric Range.
- **Data Aggregation**: The system must aggregate EV ownership data by city to calculate the density of EV owners, enabling identification of high-density areas.
- **Geospatial Analysis**: The system must generate geospatial visualizations (e.g., heatmaps or density plots) using vehicle coordinates to highlight regions with concentrated EV ownership.
- **Clustering Model**: The system must implement a clustering algorithm (e.g., K-Means or DBSCAN) to group cities into categories based on EV owner density, identifying the top 10 priority cities.
- **Prediction Output**: The system must produce a ranked list of cities with the highest EV owner density, accompanied by supporting metrics (e.g., number of EVs per city).
- **Visualization Interface**: The system must provide a simple dashboard or report (e.g., using Streamlit or Matplotlib) to display clustering results, density maps, and city rankings.
- **Data Filtering**: The system must allow filtering of data by attributes such as Model Year, Electric Vehicle Type, or County to refine analysis as needed.

1.4 Non – Functional Requirements

- **Performance**: The system must process the Electric Vehicle Population Data and generate clustering results within 5 minutes on a standard laptop with 8GB RAM and a multi-core processor.
- **Usability**: The dashboard or report interface must be intuitive, allowing users with basic data analysis knowledge to interpret city rankings and density visualizations within 10 minutes of interaction.
- **Reliability**: The system must achieve a minimum uptime of 95% during testing, ensuring consistent results across multiple runs with the same dataset.
- **Scalability**: The system should handle an increase in dataset size up to 500,000 records without significant performance degradation, using optimized algorithms and data structures.
- **Security**: The system must anonymize location data (e.g., aggregate to city level) to protect individual privacy, complying with basic data protection principles.
- **Maintainability**: The code must be well-documented and structured (e.g., using comments and modular functions) to allow team members to modify or debug it within 30 minutes.
- **Portability**: The system must run on common operating systems (Windows, macOS, Linux) using standard Python libraries, requiring no specialized hardware.

2. Scope of work

The scope of this mini project encompasses the development of a data-driven solution to predict optimal cities in Washington State for the placement of new electric vehicle (EV) charging stations, utilizing the Electric Vehicle Population Data. The project will focus on analyzing EV ownership distribution, applying machine learning techniques, and delivering actionable insights to support infrastructure planning.

Inclusions:

- Analysis of EV owner density by city, county, and postal code using the provided dataset.
- Geospatial analysis and visualization to identify high-density hotspots for EV ownership.
- Development of a clustering model to rank the top 10 cities for potential charging station placement.
- Creation of a prototype dashboard to present findings and recommendations.
- Ethical considerations, including data privacy and equitable distribution of infrastructure.

Exclusions:

- Integration of real-time data or external APIs for current charging station locations.
- Development of a fully functional web application for public use (limited to a prototype).
- Detailed analysis of traffic patterns, energy grid capacity, or cost estimates for station installation.
- Expansion beyond Washington State data to other regions.

Link	https://www.kaggle.com/datasets/yanghu583/electric-vehicle-population-data-2025
Labels	VIN (1-10), County, City, State, Postal Code, Model Year, Make, Model, Electric Vehicle Type, Clean Alternative Fuel Vehicle (CAFV) Eligibility, Electric Range, Base MSRP, Legislative District, DOL Vehicle ID, Vehicle Location, Electric Utility, 2020 Census Tract
Class/y	Target
No of Rows	250660
Prediction	Optimal Charging Station Locations Based on clustering and geospatial analysis, the model will predict which areas (cities, neighborhoods, or grid coordinates) are the best candidates for installing new EV charging stations. Future Demand for Charging Stations Using regression or time-series modeling, the system will predict the expected number of EVs (demand) in a given area, which translates into how many charging stations will be needed.

3. Activities

3.1 Project Setup and Planning

- Select the Electric Vehicle Population Data dataset that meets the project requirements for geospatial analysis.
- Backup datasets and obtain approval from the instructor.
- Identify the real-world problem statement: optimizing EV charging station placement based on owner density.
- Prepare the Statement of Work (SOW) for submission.

3.2 Data Selection & Understanding

- Explore and study the dataset variables and features (e.g., City, County, Vehicle Location, Electric Range).
- Study dataset documentation (data dictionary, variable meanings) to understand data structure.
- Identify the target variable (e.g., EV Owner Density Rank or City Priority Score derived from clustering).

3.3 Data Pre-processing

- Handle missing values (e.g., zero Electric Range entries marked as "Eligibility unknown").
- Encode categorical variables (e.g., Make, Model, Electric Vehicle Type).
- Data Transformation: Normalize/standardize numerical features (e.g., Model Year, Electric Range).
- Handle data imbalances (e.g., uneven city representation if present).

3.4 Exploratory Data Analysis (EDA)

- Visualize distributions and correlations (e.g., EV density by city using heatmaps).

- Identify trends, outliers, and patterns in EV ownership distribution.
- Generate summary statistics and insights (e.g., top cities by EV count).

3.5 Feature Engineering & Selection

- Create new derived features (e.g., EV density per city, average electric range by region).
- Apply feature selection methods (e.g., correlation filter, geospatial relevance) to prioritize key variables.

3.6 Model Development

- Train a baseline model (e.g., K-Means clustering) to group cities by EV density.
- Train and compare advanced models (e.g., DBSCAN, hierarchical clustering) for robustness.
- Tune hyperparameters (e.g., number of clusters) for best performance.

3.7 Model Evaluation & Validation

- Apply k-fold cross-validation where applicable (e.g., for predictive models if used).
- Evaluate using metrics like silhouette score for clustering or MSE if regression is explored.
- Compare results across models to select the most effective approach.

3.8 Model Expandability

- Generate feature importance plots or cluster centroids to interpret results.
- Use SHAP or LIME to explain why certain cities are prioritized.
- Provide suggestions for charging station placement based on model output.

3.9 Software Solution Development

- Implement final pipeline (data preprocessing → clustering → visualization).

- Use Google Colab notebooks and Python scripts for development.
- Provide a simple interface or demo script (e.g., Streamlit dashboard) for results.

3.10 Reporting & Documentation

- Create final report (background, methods, results, conclusion) tailored to EV infrastructure.
- Prepare dataset description and documentation specific to the EV dataset.
- Maintain GitHub repo with clear README and requirements file.

3.11 Presentation & Delivery

- Prepare a 10-minute video presentation with a demo of the clustering results.
- Rehearse and prepare for viva to explain the methodology and findings.
- Submit all deliverables and submissions on time.

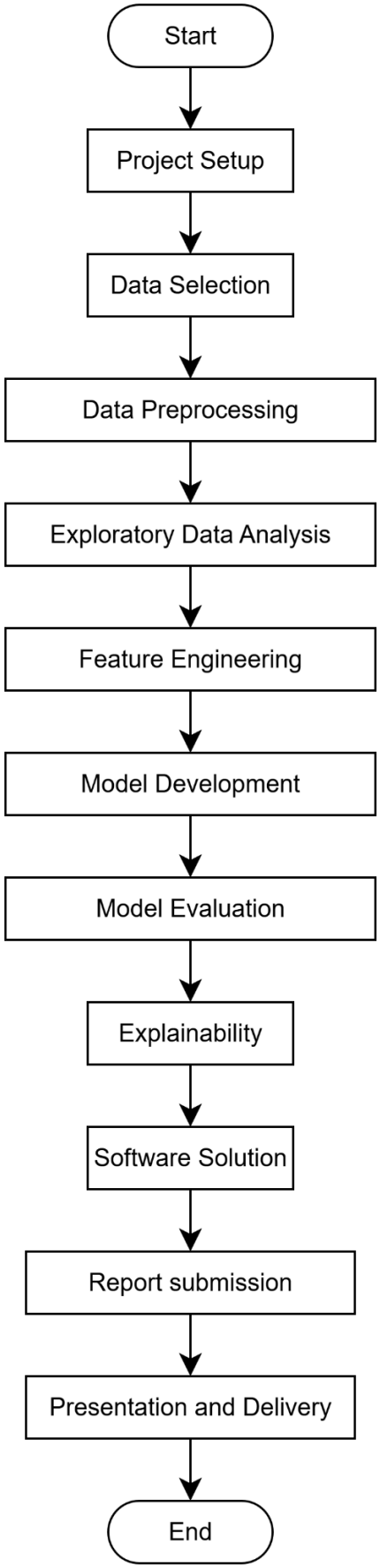

Figure 3-1 Activity Diagram

4. Approach

This project will follow the CRISP-DM (Cross Industry Process for Data Mining) methodology, which provides a structured framework for solving data driven problems. This ensures that the project workflow remains systematic, iterative, and reproducible.

4.1 Data selection

- Search for a dataset that is publicly available and addresses the real-world problem of optimizing EV charging station placement.
- Explore platforms like Kaggle, Hugging Face, OpenML, or Data.gov for suitable datasets.
- elect the Electric Vehicle Population Data dataset, which contains recent data and supports data mining, machine learning, preprocessing techniques, and algorithms.
- Check for usage restrictions and licensing of the dataset to ensure compliance.
- Prepare backups in case the primary dataset is unsuitable.
- Obtain approval of the dataset from the instructor.

4.2 Data Pre-processing

- Explore the full dataset to identify data quality issues (e.g., missing values in Electric Range).
- Data cleaning: Handle missing values using techniques like mean, median, mode, or predictive models like k-Nearest Neighbors or regression.
- Identify and remove duplicate records to ensure data uniqueness.
- Encoding: Encode categorical variables (e.g., Make, Model, Electric Vehicle Type) using one-hot encoding.
- Scaling: Scale numerical features (e.g., Model Year, Electric Range) using StandardScaler for normalization or standardization.
- Imbalance handling: Address uneven city representation (if present) using methods like oversampling or weighted clustering.

4.3 Exploratory Data Analysis (EDA)

- Detect and remove outliers for numeric variables (e.g., Electric Range) or replace them with capped values, visualized using boxplots or scatterplots.
- Measure the quality of the dataset, including accuracy, completeness, consistency, timeliness, believability, and interpretability.

4.4 Modeling

- **Baseline model**: Use K-Means clustering (simplest clustering model) to set benchmark performance, which is easy to implement, interpretable, and efficient.
- **Advanced models**: Explore DBSCAN, hierarchical clustering, or density-based methods for robustness.
- **Model comparisons**: Evaluate and compare all models to select the best-performing approach for identifying high-density EV ownership areas.

4.5 Validation and Evaluation

- Split the dataset into training and testing subsets if predictive modeling is used, with an 80% training and 20% testing split.
- Use k-fold cross-validation (e.g., k=5) where applicable for robust performance estimation.
- Evaluate using metrics like silhouette score for clustering or MSE if regression is explored; visualize results with plots or charts.

4.6 Explainability & Ethical Considerations

- Compute feature importance or cluster centroids to interpret which factors (e.g., city, EV type) influence density predictions.
- Analyze how geographic and vehicle factors influence recommendations, suggesting suitable locations for charging stations.
- Document limitations, biases (e.g., urban vs. rural data representation), and ethical concerns related to EV infrastructure planning.

4.7 Reproducibility & Code Management

- Use version control with a GitHub repository to maintain code history, branches,

and collaborations, ensuring clear commit messages for tracking changes.

- Environment Management: Use the same Python versions to avoid compatibility issues.
- Script organization: Use separate Google Colab notebooks for data processing, modeling, and visualization.
- Ensure reproducibility with fixed random seeds where possible.

4.8 Tools and Technologies

4.8.1 Tools

- Google Colab: for collaborative notebook-based coding and model training.
- Visual Studio Code (VS Code): for local development and script editing.
- WEKA: for experimenting with data mining algorithms and comparing with Python implementations.
- GitHub: for version control, code hosting, and collaboration.

4.8.1 Technologies/Libraries

- Programming Language: Python
- Libraries for Data Handling: pandas, NumPy
- Libraries for Machine Learning: scikit-learn
- Visualization Libraries: matplotlib, Seaborn
- UI Development: Streamlit (for simple interactive app/interface)

5. Deliverables

5.1 Project Plan and Documentation

5.1.1 Functional Requirements:

- Predict optimal cities for new EV charging stations based on EV owner density.
- Allow users to filter data by attributes (e.g., city, EV type) for customized analysis.
- Generate visual insights on EV distribution and density hotspots.

5.1.2 Non-Functional Requirements:

- Efficiently manage 100,000+ EV records from the dataset.
- Ensure model outputs are interpretable and explainable (e.g., via cluster centroids).
- Support prototype deployment on a web or local dashboard.
- Ensure data privacy by aggregating location data to city level.

5.2 Data Collection and Preparation

Raw Data Sets:

- Electric Vehicle Population Data CSV containing records with features like City, County, Vehicle Location, Model Year, and Electric Range.

5.3 Data Preprocessing Documentation:

- Handling of missing values.
- Categorical feature one-hot/label encoding.
- Numerical feature normalization.
- Feature engineering (e.g., EV density per city, average range by region).
- Manage uneven city representation (if present) using weighting or sampling techniques.

5.4 Exploratory Data Analysis (EDA)

Dashboard/Report

- Summary statistics for both categorical and numerical characteristics.
- Heatmaps showing the correlation between EV density and geographic factors (e.g., urban vs. rural).
- Distribution of demographics (e.g., county, model year).
- Visualization of EV ownership trends (e.g., by city, EV type) in relation to density hotspots.

5.5 Model Development

Model selection and Justification-

- Selection of K-Means as a baseline clustering model for its simplicity and interpretability.
- Exploration of advanced models like DBSCAN or hierarchical clustering for robustness.
- Justification based on performance metrics and suitability for geospatial data.

Model Training and Validation:

- Split the dataset into subsets for analysis (e.g., training and testing if predictive modeling is used).
- Evaluate metrics such as silhouette score for clustering or MSE if regression is explored.
- Adjust hyperparameters (e.g., number of clusters) using techniques like elbow method or grid search.

5.6 Model Evaluation

Evaluation report-

- Silhouette score, cluster visualization, and performance comparison across models.
- Analysis of feature importance or cluster centroids to explain city prioritization.

- Comparison of model performance to select the best approach.

Risk Status Prediction-

- Output of ranked cities for new charging station placement based on EV density.

5.7 Deployment Plan

- Deploy the model using a web interface or local dashboard (e.g., Streamlit).
- User input form for filtering data (e.g., by county or EV type).
- Real-time city ranking and density visualization.
- Probability or confidence visualization for cluster assignments.
- Provide recommendations for charging station placement based on output.
- Monitor model performance and update with new data as needed.

6. Project Plan & Timeline

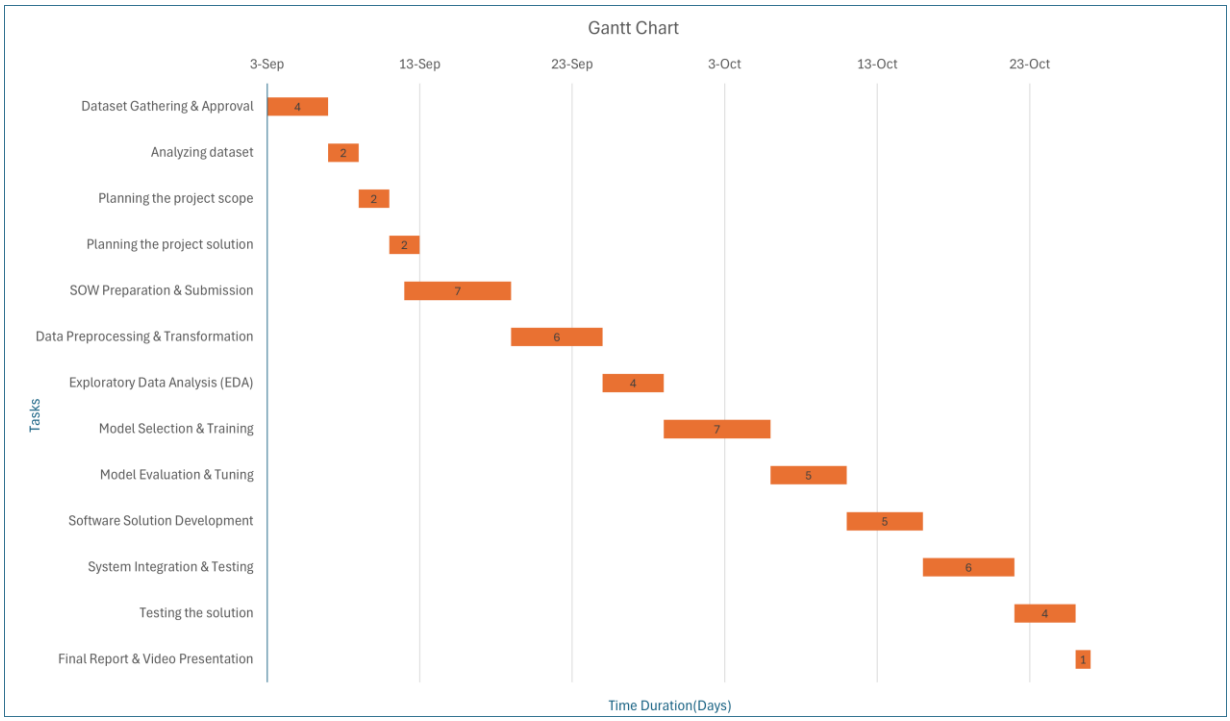

Figure 6-1 Ghantt Chart

7. Assumptions

5.7 The dataset is representative of EV populations in Washington State.

5.8 No access to real-time or additional external data.

5.9 Project focuses on density-based predictions; external factors like existing stations are not included.

5.10 Team has access to necessary computing resources.

8. Project team, roles, and responsibilities

Registration Number	Name	Responsibilities
IT23252554	L.G.D.P.M.Dissanayake	○ Build the data mining algorithm. ○ Back-end development ○ Documenting the findings
IT23274648	W.A.A.I.Wijesuriya	○ Analyze the data. ○ Build the data mining algorithm. ○ UI Design
IT23163386	A.M.S.Dulakshika	○ Analyze the data. ○ Build the data mining algorithm. ○ Evaluate the best fit model
IT23423992	K.H.Dissanayake	○ Build the data mining algorithm. ○ Back-end development ○ UI Design