COVENTRY UNIVERSITY

Faculty of Engineering, Environment and Computing

School of Computing, Electronics and Mathematics

MSc. Data Science

7150CEM Data Science Project

A Comparative Analysis of Machine Learning and Deep Learning Models for Electric Range Prediction, CAFV Eligibility Classification, and Anomaly Detection in Washington's Electric Vehicle Data

Author: Diveen Changappa Nellamakkada Robin

SID: 14511011

1st Supervisor: Dr. Anup Pandey
2nd Supervisor: Dr. Mark Johnston

Submitted in partial fulfilment of the requirements for the Degree of Master of Data Science

Academic Year: 2024/25

Declaration of Originality:

I affirm that, apart from appropriately referenced sources, this project is entirely my original work and is not derived from any other sources. Consequently, every occurrence of previously published material from books, journals, magazines, the internet, etc., has been acknowledged through citations in the main report's References or Bibliography sections. Moreover, I consent to the storage and utilization of an electronic version of this project for the purposes of plagiarism prevention and detection.

Statement of copyright:

I understand that Coventry University owns the copyright for this project report and for any model developed as a result of the project. Support, including the funding, is available for commercializing the products and services that are developed by staff and students. Any revenue that is generated will be slitted with the inventor/s of the product or service. For further information please see www.coventry.ac.uk/ipr or contact ipr@coventry.ac.uk

Statement of ethical engagement

I affirm that this research has been submitted to the Coventry University ethics and monitoring website (https://ethics.coventry.ac.uk), and the application information are provided below. User information is utilized in accordance with ethical standards and never used for any activity outside those criteria. Also, the ethical certification has been attached here in the document (appendix 1).

Signed: Diveen Changappa Nellamakkada Robin Date:09/12/2024

First Name:	Diveen Changappa
Last Name:	Nellamakkada Robin
Student ID number	14511011
Ethics Application Number	P181509
1st Supervisor Name	Dr. Anup Pandey
2nd Supervisor Name	Dr. Mark Johnston

Table of Contents

SECTION I

SECTION II

SECTION III

SECTION IV

SECTION V

SECTION VI

SECTION VII

Table of Figures

Abstract

Electric vehicles (EVs) are revolutionising the automobile industry by providing environmentally friendly alternatives to conventional combustion-engine vehicles. Despite their increasing popularity, issues such as range anxiety, Clean Alternative Fuel Vehicle (CAFV) eligibility classification, and anomaly detection prevent broad implementation. This study tackles these issues by contrasting machine learning (ML) and deep learning (DL) models for analysing EV data obtained from the Washington State Department of Licensing.

The study aims to fill gaps in evaluating the performance of ML and DL models for tasks such as electric range prediction, CAFV eligibility classification, and anomaly detection. Classification models include Random Forest, XGBoost, and Feedforward Neural Networks (FNN), whereas regression models include Random Forest Regressor and Deep Neural Networks (DNN). Isolation Forest is used to find anomalies. The dataset for Battery Electric Vehicles (BEVs) and Plug-in Hybrid Electric Vehicles (PHEVs) provides a comprehensive testing ground.

This study uses a comparative analysis to evaluate which models are most suited for certain activities, providing insights into their strengths and limits. The findings are likely to increase forecast accuracy, classification approaches, and anomaly detection frameworks, hence promoting the adoption of sustainable transportation solutions.

SECTION I.

1. Introduction

In the global search for ecologically friendly and sustainable transportation, electric vehicles (EVs) have emerged as a transformative innovation. Their adoption marks a shift away from conventional internal combustion engine vehicles toward cleaner alternatives, driven by advancements in battery technology, growing environmental consciousness, and favourable government policies. EVs are widely regarded as a critical component of efforts to combat climate change and reduce reliance on fossil fuels. However, the seamless integration of EVs into traditional transportation systems is hampered by several persistent challenges. Among the most pressing are range anxiety, effective classification for Clean Alternative Fuel Vehicle (CAFV) eligibility, and the detection of anomalies in operational data. Addressing these challenges is vital to building trust among stakeholders and consumers, ensuring the growth and sustainability of the EV industry.

The primary purpose of this research is to utilize robust data-driven approaches, combining advanced machine learning (ML) and deep learning (DL) methodologies, to tackle these critical challenges. Specifically, this study investigates the most efficient models for analysing EV datasets, which include attributes such as vehicle type, electric range, and geographical location. By systematically comparing the performance of these models, the research aims to identify the best strategies for predicting electric range, classifying CAFV eligibility, and detecting anomalies. To achieve this, the study leverages Random Forest Regressor and Deep Neural Networks (DNN) for regression tasks, Random Forest, XGBoost, and Feedforward Neural Networks (FNN) for classification tasks, and Isolation Forest for anomaly detection.

Making sure the models selected fit the dataset's properties which include high-dimensional and heterogeneous data is a major goal of this research. In previous studies, Random Forest and XGBoost have been thoroughly tested for classification tasks due to their interpretability and capacity to handle a variety of datasets (Liu, Ting, & Zhou, 10 February 2009). On the other hand, Deep Neural Networks are especially well-suited for regression issues such as estimating electric range since they are excellent at capturing complex, non-linear correlations (Shahriar Afandizadeh, 2023). Isolation Forest provides a dependable technique for anomaly detection, allowing outliers to be found in big, complicated datasets. Although the effectiveness of each of these models is acknowledged separately, a comparative study customized for EV datasets is necessary to ascertain their relative advantages and disadvantages.

The incorporation of machine learning methodologies holds practical importance. Accurate range projections mitigate consumer apprehensions over EV use, while allowing manufacturers to enhance battery performance and helping infrastructure planners to strategically position charging stations. Classifying vehicle types assists governments in customizing incentives, informs utility providers for energy forecasting, and facilitates environmental evaluations to measure the carbon impact of various EVs. Anomaly detection improves data integrity, guaranteeing dependable inputs for fleet management and infrastructure planning.

This report's scope includes assessing and contrasting the above stated models to determine how well they handle different facets of EV data analytics. The study explores several important procedures, such as feature selection, model training, data preparation, and performance assessment. The trade-offs between interpretability, accuracy, and computing efficiency are also covered, offering stakeholders from legislators to business professionals' actionable insights. This research recognizes that external factors including customer behaviour, environmental conditions, and the availability of charging infrastructure may also have an impact on EV adoption and performance, even though its primary focus is on data-driven techniques.

This study's dependence on a static dataset from the Washington State Department of Licensing is one of its limitations; it might not accurately reflect the dynamic nature of real-world circumstances. Furthermore, even though the selected models are optimal for the data at hand, their performance might alter if the scale or properties of the data change. However, the study makes the assumption that the dataset is reflective of common EV usage patterns and offers enough variation to adequately evaluate the models.

This study guarantees a methodologically sound and human and AI driven strategy to discovering the most effective models for EV analytics by customizing the model selection to the unique features of the dataset. This study is well-positioned to advance our knowledge of model performance and aid in the creation of novel solutions for the EV sector. The study's conclusions are ultimately intended to guide the development and application of data-driven initiatives that boost customer trust, streamline regulatory structures, and quicken the shift to environmentally friendly transportation options.

1.1 Project Aim:

The objective of this project is to do an extensive study of electric vehicle (EV) data utilizing sophisticated data-driven techniques, such as classification, regression, anomaly detection, and exploratory data analysis. The study aims to assess the efficacy of classical machine learning (ML) and deep learning (DL) models in important tasks, including Clean Alternative Fuel Vehicle (CAFV) eligibility categorization, electric range prediction, and anomaly detection. This initiative aims to reveal actionable insights that enhance EV analytics, promote sustainable transportation development, and influence industry practices.

1.2 Objectives:

Investigate Advanced Analytical Techniques: Examine cutting-edge machine learning and deep learning methodologies to tackle issues in EV categorization, regression, and anomaly detection.

Replicate complex data scenarios: Formulate ways to evaluate algorithm robustness in demanding real-world scenarios to guarantee model reliability.

Acquire and Enhance Dataset: Obtain and preprocess a varied dataset for electric vehicle study, assuring its suitability for high-performance predictive modelling.

Develop Classification Models: Develop and enhance models, such as Random Forest, XGBoost, and Feedforward Neural Networks (FNN), to accurately classify CAFV eligibility.

Forecast Electric Vehicle Range Precisely: Construct regression models, including Random Forest Regressor and Deep Neural Networks (DNN), to provide accurate electric range forecasts.
Analyse and Evaluate Models: Execute a methodical evaluation of performance between

machine learning and deep learning models, taking into account criteria such as accuracy, efficiency, and interpretability.

Guarantee Data Integrity: Utilize Isolation Forest methodologies to identify abnormalities and assess the integrity of electric vehicle datasets.

Contribute to Electrical Vehicle Advancements: Deliver insights that guide policy formulation, enhance customer confidence, and refine electric vehicle adoption tactics.

1.3 Research Questions:

1. To what extent can machine learning and deep learning models accurately classify Clean Alternative Fuel Vehicle (CAFV) eligibility based on vehicle attributes?
2. What is the significance of anomaly identification in maintaining the integrity of electric vehicle datasets, and how effective is the Isolation Forest method in this situation?
3. What is the extent of the impact of feature selection on the performance of models predicting electric range and identifying CAFV eligibility in electric vehicle datasets?
4. In what ways do environmental elements like geographic location, climate, and seasonal variations influence the precision of electric vehicle range forecasts, and can models be modified to incorporate these variables?

1.4 Report Overview

The report's structure carefully delineates and arranges each portion, providing readers with a coherent progression of the project and specifying the actions conducted in each segment. The following is the comprehensive structure:

Section 1: Presents the Introduction, encompassing the Problem Statement, scope, and aims and objectives of the research.

Section 2: Provides the Literature Review, encompassing prior studies and research that established the groundwork for this effort.

Section 3: Evaluates the Methodology, including the strategy and rationales for the selection of methodologies utilized in this project.

Section 4: Addresses the Requirement Gathering process, encompassing the finished datasets and particular requirements essential for the research.

Section 5: Elucidates the Dataset, including its overall structure, characteristics, and preprocessing procedures.

Section 6: Details the Project Design, including a summary of the project’s structure and progression.

Section 7: Discusses the techniques and models utilized in the analysis, encompassing machine learning and deep learning procedures.

Section 8: Offers Implementation Details, clarifying the application of each technique or model, along with findings and validations.

Section 9: Examines the Results and Validation, concentrating on the assessment and evaluation of regression, classification, and anomaly detection tasks.

Section 10: Emphasizes Project Management, encompassing timeframes, task management, and essential procedures.

Section 11: Provides a Critical Assessment of the project, highlighting its strengths, weaknesses, and opportunities for enhancement.

Section 12: Concludes the paper with conclusions and future work, summarizing findings and suggesting prospective research avenues.

SECTION II.

2. Literature Review

2.1 Comparative Analysis of Machine Learning Models for Predicting Electric Vehicle Range.

The adoption of electric vehicles (EVs) is essential for addressing climate change, and a major problem in this area is the precise forecasting of EV range. An accurate range prediction model fosters consumer confidence and facilitates wider acceptance. Gregorius Airlangga's study, Comparative Analysis of Machine Learning methods for Predicting Electric Vehicle Range, highlights the significance of diverse machine learning (ML) (Airlangga, 2024) methods in tackling this issue. The research underscores the efficacy of ensemble methods, such as Random Forest Regressor and Gradient Boosting Regressor, in contrast to basic linear models for elucidating intricate, non-linear relationships inside electric vehicle datasets.

The study assesses five machine learning models: Linear Regression, Ridge Regression, Lasso Regression, Random Forest Regressor, and Gradient Boosting Regressor, employing Mean Squared Error (MSE) as the principal evaluation criterion. Linear models, despite their computing efficiency and interpretability, had difficulties with the dataset's complexity, underscoring their inadequacy in modelling non-linear interactions successfully. Conversely, Random Forest and Gradient Boosting approaches exhibited enhanced performance, adeptly capturing complex feature relationships, with Random Forest attaining the minimal error.

Airlangga's research emphasizes the significance of preprocessing and feature engineering in improving model performance. It utilizes a comprehensive preparation workflow, encompassing imputation, encoding, and feature modification, essential for managing the varied forms of EV data. These measures guaranteed that the models could process categorical and numerical data effortlessly, enhancing their prediction efficacy.

This study (Airlangga, 2024) offers significant insights into model selection for electric vehicle range prediction, highlighting the benefits of ensemble approaches in managing non-linear complications. The study emphasizes the necessity of integrating real-time data with external variables, including weather and driving patterns, to enhance prediction accuracy. These findings immediately inform the current study, substantiating the use of Random Forest Regressor and Gradient Boosting methodologies inside the analytical framework.

2.2 Feature Analyses and Modeling of Lithium-Ion Battery Manufacturing Based on Random Forest Classification

(Lei Tong, 2021) They have examined the utilization of Random Forest (RF) for the classification and ranking of features in lithium-ion battery production, essential for electric cars (EVs). The research examines RF's capacity to handle high-dimensional data and its interpretability, establishing it as an effective instrument for predictive analyses related to battery performance and electric vehicle range.

The study employed a dataset comprising several manufacturing and operational characteristics, such as charge cycles, battery capacity, and temperature stability. Random Forest was selected for its ensemble characteristics, enabling it to model intricate, non-linear interactions. The identification of crucial factors influencing battery longevity, including charge density and material composition, allowed manufacturers to prioritize these elements in production. The study examined RF's resilience to overfitting attributed to its decision tree ensemble, ensuring constant accuracy across validation datasets.

An important element was the preprocessing pipeline, which involved addressing missing values and converting categorical data into numerical formats appropriate for RF. The feature importance ranking provided practical insights, facilitating dimensionality reduction and enhancing computational efficiency while maintaining prediction accuracy. The paper recognized constraints, including the computational cost of RF with extensive datasets, and proposed that hybrid methods integrating RF with deep learning might improve performance.

This article (Lei Tong, 2021) emphasizes the utility of RF in feature analysis and categorization within battery production and electric vehicle applications. The discoveries on feature ranking and robustness directly guide this project's methodology for preprocessing and selecting characteristics for regression and classification tasks. The focus on handling high-dimensional data renders RF a crucial element of EV analytics.

2.3 Isolation Forest for Anomaly Detection in Raw Vehicle Sensor Data

The authors (Hofmockel, 2018) here investigated the application of the Isolation Forest algorithm for anomaly detection in unprocessed vehicle sensor data. The project sought to tackle issues related to the management of extensive data produced by vehicle fleets, including the transmission costs and inefficiencies linked to the transfer of extraneous data to backend systems. Anomaly detection was proposed as a method for recognizing and transmitting only pertinent data, hence reducing data transfer while preserving critical events.
The research evaluated the efficacy of Isolation Forest in comparison to the Replicator Neural Network (ReplNN) for anomaly identification. The Isolation Forest method was trained on normal data derived from raw sensor data of vehicle fleets and subsequently assessed on datasets that included anomalies such as accidents, emergency braking, and sensor breakdowns. Isolation Forest functions by randomly segmenting data through binary trees, isolating anomalies more efficiently since they necessitate fewer divisions.

The results indicated that Isolation Forest surpassed ReplNN in nearly all cases, attaining a superior Area Under the Curve (AUC) score with an average of 0.908. Conversely, ReplNN exhibited tendencies of overfitting, impairing its capacity to generalize across varied datasets. The authors emphasized the efficiency of Isolation Forest with high-dimensional data, noting that characteristics such as the number of trees and subsample size affect its efficacy. The program successfully detected anomalies using a 95% quantile threshold, while preserving a minimal false positive rate. Furthermore, data transfer was minimized by as much as 93%, guaranteeing that only pertinent data was conveyed, hence conserving bandwidth and storage resources.

A recognized shortcoming was the algorithm's challenge in recognizing minor anomalies, such as mild over speeding, due to their resemblance to regular data. The authors proposed improving the method by integrating temporal data and time-series feature extraction to enhance anomaly identification.

This research (Hofmockel, 2018)highlights the efficacy of Isolation Forest for scalable anomaly identification in electric vehicle analytics. Its exceptional capability in detecting critical abnormalities justifies its integration into this project for ensuring data quality and reducing superfluous data transfer. Additional improvements, including the integration of temporal data, could broaden its applicability to real-world electric vehicle statistics.

2.4 Classification of Potential Electric Vehicle Purchasers: A Machine Learning Approach

This study (Javier Bas, 2021) investigates machine learning (ML) methodologies for categorizing individuals according to their likelihood of adopting electric cars (EVs). The research utilizes data from a choice-based poll to identify factors affecting electric vehicle uptake and assesses various machine learning models. This study is significant for comprehending the influence of socioeconomic, attitudinal, and vehicle-specific factors on consumer choices, offering essential insights for electric vehicle market strategies and policy development.

The study utilized data obtained from a stated-choice survey, encompassing characteristics such as tax incentives, range, price, fast-charging durations, and social effect elements. Three separate groups of features were identified: socioeconomic, attitudinal, and vehicle-related characteristics. Machine learning models, such as Support Vector Machines (SVMs), Artificial Neural Networks (ANN and DNN), and ensemble techniques like Gradient Boosting and Extremely Randomized Trees (XRT), were assessed for their efficacy in classifying electric vehicle adopters and non-adopters.

Among these models, the Support Vector Machine with a polynomial kernel attained the highest accuracy at 83.45%, closely followed by the Artificial Neural Network and Gradient Boosting. These models accurately classified adopters but encountered difficulties with non-adopters, indicating biases in their prediction capabilities. Deep Neural Networks (DNNs) exhibited suboptimal performance, presumably due to the survey dataset's restricted complexity and size, underscoring the algorithm's reliance on high-dimensional, non-linear data. The research also tackled data pretreatment issues, including the management of missing values using Multiple Imputation by Chained Equations (MICE) and the identification of the most pertinent features by Random Forest-based priority ranking. Factors such as tax incentives, charging infrastructure, and environmental attitudes were prioritized, whereas demographic characteristics like marriage status and gender were considered less significant.

A notable constraint observed was the erroneous categorization of two responder groupings. One group demonstrated strong pro-EV sentiments yet exhibited minimal environmental worry, whereas the other group showed environmental concern but had limited interest in EVs.

This inconsistency hindered precise forecasts. Furthermore, the stated-choice structure of the sample constrained the generalizability of the findings to real-world contexts.

This study (Javier Bas, 2021)offers significant insights for this project, especially regarding feature selection and model assessment. The utilization of Random Forest for feature ranking enhances its effectiveness in pinpointing the most significant variables for classification tasks such as Clean Alternative Fuel Vehicle (CAFV) eligibility. The research underscores the necessity of dataset balancing to mitigate biases favouring majority classes. The subpar performance of DNNs in this context indicates that simpler ML models may be more appropriate for datasets of restricted size or complexity, along with the project's emphasis on selecting models according to data properties.

This study emphasizes the necessity of contextual comprehension of factors affecting electric vehicle uptake and warns against excessive dependence on deep learning for moderately complex datasets. Furthermore, implementing rigorous preprocessing methods and analysing misclassified instances might improve the model's precision and relevance.

2.5 A Review of the Data-Driven Prediction Method of Vehicle Fuel Consumption

(Dengfeng Zhao, 2023)The publication offers an extensive analysis of data-driven techniques for forecasting vehicle fuel use, highlighting the significance of precise predictions to improve fuel efficiency and mitigate environmental damage. The research categorizes predictive methodologies into physical and data-driven models, emphasizing the shortcomings of conventional techniques while demonstrating the advantages of machine learning and neural networks in managing intricate, nonlinear correlations among fuel consumption indicators.

The review delineates a distinct comparative framework between conventional machine learning techniques (e.g., random forest, support vector machines) and neural networks, highlighting the latter's flexibility and enhanced efficacy in handling intricate datasets as shown in **Fig 1.** Random forests are recognized for their capacity to handle high-dimensional data and facilitate dimensionality reduction, but deep learning models such as feedforward neural networks (FNN) and deep neural networks (DNN) exhibit superior feature extraction abilities and generalization potential.

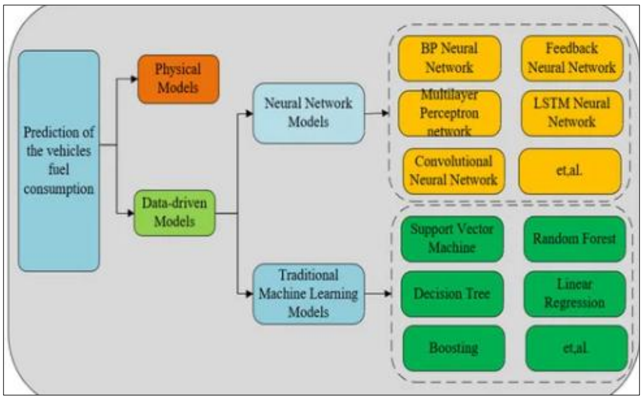


Fig. 1 - Classification of vehicle fuel consumption forecasting methods

Source- *https://www.mdpi.com/1996-1073/16/14/5258*

The authors here critically examine the shortcomings of single-model methodologies, including overfitting and susceptibility to noisy data, advocating for hybrid models as a remedy. These hybrid models include the advantages of multiple approaches, including random forests and gradient boosting, to attain elevated accuracy and resilient predictions. The review references practical cases where hybrid models attained coefficients of determination surpassing 0.90 and RMSE values below 0.40, demonstrating their effectiveness.

The paper emphasizes the significance of feature engineering and dimensionality reduction methods, such as PCA and sensitivity analysis, in enhancing model robustness and computational efficiency. These approaches align closely with our research, especially in forecasting fuel range through deep neural networks and regression problems.

(Dengfeng Zhao, 2023)This review is pertinent to the project, providing practical insights into utilizing sophisticated machine learning techniques such as deep neural networks and hybrid models for predictive analytics. It underscores the significance of feature selection and model generalization, consistent with the project's emphasis on predictive modelling for electric vehicle characteristics. The study focuses on gasoline usage, although its techniques and findings can be used to forecasting vehicle range and other electric vehicle-specific measures. This research establishes a solid basis for enhancing model selection and incorporating hybrid methodologies into the project architecture.

2.6 Machine Learning for Real-Time Fuel Consumption Prediction and Driving Profile Classification Based on ECU Data

This study (Canal, 2024)here emphasizes the utilization of machine learning (ML) models to assess real-time fuel usage and categorize driving behaviours through data obtained from a vehicle's Electronic Control Unit (ECU). This methodology incorporates Industry 4.0 ideas, employing cloud computing and data transfer technologies for real-time analysis. The project seeks to improve operating efficiency in the automobile sector, decrease fuel consumption, and encourage environmentally sustainable driving practices.

The research underscores the efficacy of incorporating machine learning into automobile systems. It utilizes multiple models including K-Means, Logistic Regression, and XGBoost for the classification of driving behaviour, as well as Ridge Regression, Support Vector Regression (SVR), and XGBoost for forecasting fuel usage. In comparison to conventional approaches, the ML models demonstrated enhanced accuracy and computational efficiency.

A comprehensive feature selection procedure was executed utilizing Scikit-Learn tools, identifying characteristics most pertinent to fuel consumption and driving efficiency. The research highlights the need of real-time data capture via cloud integration, especially for practical applications in dynamic settings.

This research is congruent with my topic, as it employs XGBoost and regression models, both of which are fundamental to my work. The feature selection approach and the utilization of real-time data are especially pertinent, offering insights for enhancing model efficiency and relevance. The differentiation between aggressive and economical driving styles provides a valuable framework for examining categorical outputs in my dataset.

This study presents (Canal, 2024) a comprehensive methodology for predicting fuel usage and classifying driving behaviour, pertinent to my research on electric car range and classification models. Its emphasis on feature selection and real-time data processing provides essential strategies for the implementation and optimization of predictive models, hence augmenting the resilience of my project.

2.7 Multilayer Feed-Forward Artificial Neural Networks for Class Modeling

The Journal here (Marini, 2007) present a new methodology for class modelling utilizing multilayer feed-forward artificial neural networks (FFNNs). This strategy, in contrast to traditional discriminative techniques, emphasizes modelling distinct classes to discern commonalities among data points within a category. The study emphasizes the adaptability of FFNNs in identifying nonlinear correlations and its capacity to surpass conventional methods such as SIMCA and UNEQ, especially for intricate datasets.

The research presents a novel approach in which class spaces are delineated utilizing feedforward neural networks structured as auto-associators. The network is trained to reproduce input data, with the residual variance between actual and reconstructed data acting as the criterion for class acceptance. This methodology facilitates robust class modelling independently of information from other categories, hence diminishing bias and mitigating the hazards of overfitting.

The research (Marini, 2007)illustrates the effectiveness of the technique on both synthetic datasets (X-OR issue) and real-world datasets (Italian CDO wines). In the X-OR dataset, FFNNs surpassed SIMCA and UNEQ, achieving a classification accuracy of 94%, demonstrating their efficacy in managing significant nonlinearity. Likewise, in the wine dataset, FFNNs attained superior specificity without sacrificing sensitivity, demonstrating their dependability even for less complex, linearly structured issues. These findings highlight the versatility of FFNNs in various contexts.

This study highlights the significance of model setup, encompassing feature selection, appropriate hidden layer size, and cross-validation, in relation to this project. The proposed FFNN-based class modelling demonstrated distinct advantages; yet, the research recognized problems such as the computational complexity of training and the necessity for systematic optimization to mitigate overfitting.

This work (Marini, 2007) provides significant insights into utilizing FFNNs for classification problems. The emphasis on delineating resilient class spaces corresponds with this project's concentration on feed-forward neural networks for vehicle type classification. The methodology for feature selection and bias reduction via auto-association can be modified to enhance model generalization and precision. The study's computational constraints and dependence on cross-validation underscore the necessity for scalability concerns when applied to extensive EV datasets.

2.8 Weather Impact on Energy Consumption for Electric Trucks: Predictive modelling with Machine Learning

This master's thesis (Carlsson, 2024)examines the influence of weather conditions on the energy usage of electric trucks, utilizing sophisticated machine learning methodologies like Random Forest, XGBoost, LSTM, and Convolutional Neural Networks. This corresponds effectively with initiatives such as forecasting electric car range or analysing energy consumption trends, which are influenced by external environmental variables.

The (Carlsson, 2024)paper's extensive data-driven methodology for analysing energy consumption across different weather situations yields significant insights. The scientists employed strong models for prediction by integrating historical meteorological data with real-time operating data from electric trucks. The study highlights the importance of critical factors such as temperature, precipitation, humidity, and wind speed, with temperature exhibiting a significant negative association to energy usage. The incorporation of SHAP values for feature significance improves interpretability, a methodology applicable in projects utilizing Random Forest and XGBoost models.

The application of LSTM and hybrid LSTM-CNN architectures is particularly pertinent for regression tasks, providing valuable insights for properly addressing time-series data. Their findings indicate that addressing temporal dependencies is essential for forecasting energy use across different weather conditions. The authors' selection of gradient boosting methods, such as XGBoost, enhances its established efficacy in regression and classification tasks, offering a different perspective for evaluating electric car performance.

This (Carlsson, 2024)study provides significant methodology and findings that might guide your project, especially with the identification of key variables affecting energy usage and improving model interpretability. Although their emphasis is on trucks, the fundamental ideas of integrating environmental data and utilizing hybrid machine learning models can significantly improve the predicted accuracy of electric car range and energy models. Nonetheless, modifying their methodologies for larger datasets or consumer cars may necessitate consideration of changes in operational settings and scaling factors.

2.9 Estimation of Energy Consumption of Electric Vehicles Using Deep Convolutional Neural Network to Reduce Driver's Range Anxiety

This study (Modi, 2020) examines an innovative method to tackle the significant issue of range anxiety among electric vehicle (EV) owners. The authors present a method for real-time energy consumption prediction by utilizing a deep convolutional neural network (D-CNN) and incorporating parameters such as vehicle speed, road elevation, and tractive effort. This emphasis corresponds with the increasing necessity for precise and actionable forecasts in electric vehicle technology to bolster user confidence and maximize performance.

The methodology of this work is quite novel, utilizing a deep learning architecture to analyse energy usage patterns, providing considerable advantages above conventional regression and simulation methods. This D-CNN framework utilizes exterior metrics readily obtainable through GPS and GIS systems, in contrast to methods that necessitate complex internal vehicle data, such as motor efficiency curves or battery specifications. The application of time-series-to-image encoding as input for the neural network is a distinctive novelty. Three encoding techniques Gramian Angular Field (GAF), covariance, and eigenvector methods—are examined to enhance model performance. The experimental findings illustrate the effectiveness of the CNN7 architecture, which surpassed other evaluated models (CNN9, conventional regression) in accuracy, exhibiting an error deviation of merely 5.09% on real-world datasets. A comparison examination with cutting-edge methodologies, including those put out by Yang et al. and Galvin, further substantiates the robustness of the suggested method. The model's principal merits are its capacity to generalize across diverse road conditions and its computational efficiency for real-time applications.

This study's (Modi, 2020) contributions are directly pertinent to initiatives utilizing machine learning for electric vehicle analytics, including range prediction and fuel consumption estimation. The implementation of deep learning methods for real-time forecasts underscores the potential for incorporation into user-focused electric vehicle applications. The study emphasizes the necessity of thorough data preprocessing, the choice of interpretable feature descriptors, and the investigation of architectures tailored for real-time implementation. It underscores the importance of reconciling precision with computational requirements to attain practical applicability in real-world contexts. This study is a crucial reference for improving prediction models using scalable, deep learning techniques.

2.10 Predicting Popularity of Electric Vehicle Charging Infrastructure in Urban Context

The work (Straka, 2020) examines the elements affecting the popularity of electric vehicle (EV) charging infrastructure, a crucial element for enhancing EV adoption. Given the rising demand for sustainable mobility, comprehending the efficacy of charging infrastructure is essential for urban planning and the expansion of the electric vehicle market. This study presents a data-driven analytical approach that employs Geographic Information Systems (GIS) and extensive charging transaction data to forecast the popularity of charging stations. The study aims to enhance infrastructure deployment strategies by considering social, demographic, and regional characteristics.

This work (Straka, 2020) makes a substantial contribution to the electric vehicle literature by introducing an innovative methodology for assessing the popularity of charging stations. It utilizes the EVnetNL dataset, comprising over one million charging transactions in the Netherlands, in conjunction with GIS data to investigate urban and demographic determinants. The authors utilize machine learning models logistic regression with L1 regularization, random forests, and gradient-boosted regression trees to categorize and forecast the highest-performing charging pools. The results emphasize key determinants, including proximity to commercial zones, sports facilities, and urban activity hubs, all of which are positively associated with the popularity of charging pools. In contrast, residential zones and locations with diminished socioeconomic level exhibited adverse effects. This is very pertinent to my project, since it underscores the significance of integrating spatial and socioeconomic data into prediction models.

The research's application of machine learning corresponds with my emphasis on methods such as random forests and gradient boosting. Its classification methodology for identifying "popular" charging pools aligns with my categorization task for vehicle kinds. The dataset's granularity and incorporation of various urban elements provide a thorough view on infrastructure design, yielding applicable findings for my project.

The research highlights the importance of data-driven methodologies in enhancing electric vehicle infrastructure, with practical implications for urban planning and policy formulation. The implementation of logistic regression with L1 regularization and ensemble techniques illustrates the resilience of machine learning in practical scenarios. Employing analogous methodologies, especially in feature selection and spatial analysis, may improve my project's model efficacy and contextual pertinence. Nonetheless, meticulous consideration of dataset variability and generalization constraints, as shown in the research, is crucial for ensuring scalability across regions.

2.11 Range Prediction Based on Battery Degradation and Vehicle Mileage for Battery Electric Vehicles

The research (Surabhi, 2024) examines the critical matter of range prediction in battery electric vehicles (BEVs), emphasizing the impacts of battery deterioration and vehicle mileage. This research highlights the increasing significance of precise range estimations to mitigate range anxiety, a primary barrier to the extensive deployment of electric vehicles. The study systematically correlates mileage with battery degradation to elucidate the limitations of lithium-ion batteries in BEVs.

This study (Surabhi, 2024) used a linear regression model to examine the correlation between battery deterioration and vehicle miles. It employs extensive datasets, incorporating empirical data from Tesla Model S consumers, to monitor the degradation of battery capacity over time. The research indicates that the battery deterioration curve is nonlinear, with notable performance declines happening beyond 200,000 miles. These findings correspond with industry apprehensions over the long-term sustainability of lithium-ion batteries and their effects on consumer trust. The research additionally investigates exogenous factors affecting battery degradation, including charging habits, environmental conditions, and driving behaviour. It indicates that regular utilization of fast-charging stations and severe temperatures expedite depreciation. This element is particularly pertinent to my project, which entails employing machine learning models such as Random Forest Regressor and Deep Neural Networks to forecast vehicle range. Incorporating these elements as variables in predictive models could markedly improve their precision and practical relevance.

The report also suggests implementing innovative battery chemistries and optimizing charging procedures to prolong battery lifespan. These findings are essential for the regression aspect of my study, which centres on fuel range prediction. Integrating the data from this research could enhance the model's robustness in managing long-term forecasts.

This work enhances the field of EV analytics by providing a comprehensive analysis of battery deterioration dynamics and their effects on range estimation. The findings about nonlinear degradation patterns and their affecting elements establish a robust basis for improving machine learning models in electric vehicle applications. This research emphasizes the necessity of integrating real-world usage data and accounting for environmental and behavioural aspects to create robust predictive models.

2.12 Prediction of Electric Buses Energy Consumption from Trip Parameters Using Deep Learning

The research paper “Prediction of Electric Buses Energy Consumption from Trip Parameters Using Deep Learning” (Teresa Pamuła, 2022)gives an idea on how deep learning models can be used to guess how much energy electric buses will use. This study corresponds with the contemporary emphasis on sustainable transportation and the application of machine learning (ML) methodologies for data-driven insights. The study utilizes realistic parameters like distance, elevation, and trip time to illustrate the viability of accurate energy consumption forecast, offering essential insights for enhancing operational efficiency. This research is directly related to the ongoing project, which centres on electric vehicle range prediction and categorization utilizing machine learning and deep learning models.

The research (Teresa Pamuła, 2022) utilizes sophisticated models such as autoencoders, LSTMs, and MLPs, attaining a prediction accuracy of 93%. The rigorous methodology emphasizes the significance of choosing relevant features and preparing data to improve predictive accuracy. These methods are very useful for changing to the goals of the current project, especially for regression tasks like predicting the electric range. Nevertheless, the article predominantly emphasizes public transportation, resulting in little exploration of individual electric vehicle applications. This gap provides an opportunity for the current effort to apply analogous approaches to personal electric vehicles, yielding deeper insights into electric vehicle analytics.

The study's results (Teresa Pamuła, 2022)underscore the significance of feature engineering and domain-specific parameter selection, which can facilitate the discovery of influential variables in this project. The project enhances current research by resolving deficiencies in categorical feature integration and broadening the focus to individual electric vehicles, thereby contributing to a more thorough understanding of electric vehicle performance analytics.

2.13 Literature Gap

Predictive modelling approaches like machine learning and deep learning have been thoroughly investigated in the present collection of literature on electric vehicles (EVs) for applications such as estimating fuel range, classifying vehicle types, and detecting anomalies. Research utilizing models like as Random Forest, XGBoost, and Deep Neural Networks (DNNs) has proven to be useful in managing intricate, high-dimensional datasets, especially for applications related to electric vehicles. In addition, there is hope for better prediction accuracy with the incorporation of external elements including environmental impacts, charging habits, and battery deterioration. But there are still major holes in the existing literature.

Instead of looking at how different tasks interact with one another; most studies focus on classification or regression alone. There aren't many studies that look at mixed frameworks that combine classification tasks (like predicting the type of vehicle) with regression tasks (like estimating the range), even though these tasks are often linked in real-life EV situations. The impact of spatial and temporal dependencies on multi-task modeling using EV data has also been little investigated, despite some prior work in this area. It also often works on its own, without directly improving or adding to the results of classification and regression.

Filling this void with a unified framework that can process EV data for anomaly detection, classification, and regression all at once would be a huge step forward for the industry. This will improve operational efficiency, sustainability, and strategies for electric vehicle adoption with complete and scalable solutions for all parties involved.

SECTION III.

3. METHODOLOGY

The **Waterfall Model** was chosen as the software development lifecycle (SDLC) for this project because of its systematic and linear methodology. Each step establishes a definitive basis for the ensuing stages. This is especially advantageous for this research, as objectives and deliverables were obvious from the beginning.

Reason for Selecting the Waterfall Model:

Organized Framework: The sequential advancement guarantees that each phase, from data preprocessing to model assessment, is systematically performed.

First Aim Details: The ultimate aims, encompassing electric car range forecasting and vehicle categorization, were explicitly articulated at the project's commencement.

Systematic Information Flow: Outputs from each phase serve as inputs for the following phase, facilitating a seamless transition between data preparation, model creation, and evaluation.

The following is a modified workflow for this research (Fig2):

Requirement Analysis: Evaluating the dataset and objectives.
System Design: Organizing the model architectures and workflows.
Execution: Formulating and instructing models.
Assessment: Analysing model efficacy.
Documentation: Reporting and assessing outcomes.

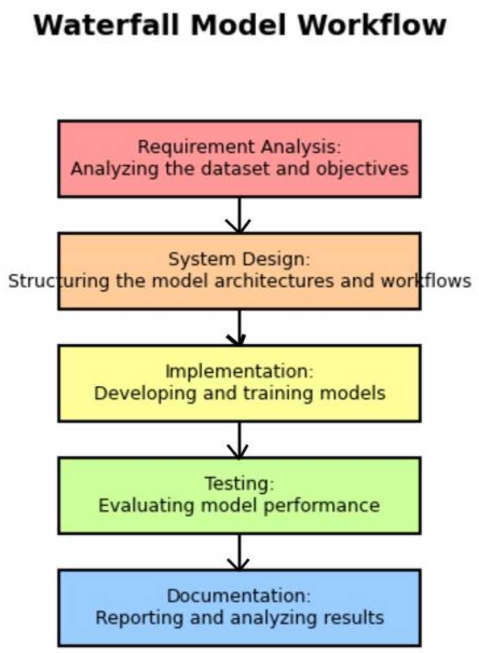

Waterfall Model Workflow
Requirement Analysis:
Analyzing the dataset and objectives
System Design:
Structuring the model architectures and workflows
Implementation:
Developing and training models
Testing:
Evaluating model performance
Documentation:
Reporting and analyzing results

Fig. 2 - Depicts the Waterfall Model

4. REQUIREMENTS

The requirements section establishes the fundamental elements necessary for the project's effective execution. This section provides a definitive framework for the research by thoroughly delineating the data, system, functional, non-functional, and technological requirements. The following are the explicit requirements for this project.

4.1 Data Requirements

The success of machine learning and deep learning models depends on the availability of comprehensive training data. Sufficient and relevant data is essential to improve model accuracy and performance. For this project, the dataset is obtained from the Washington State Department of Licensing, consisting of registered BEVs and PHEVs. The dataset includes critical attributes such as vehicle type, electric range, and CAFV eligibility. Thorough preprocessing, such as handling missing values and feature scaling, is performed to ensure data readiness for neural network training.

4.2 System Requirements

Using machine learning and deep learning models can be very complex and costly, so we need a system with the right hardware. The following are some of the project's specifications:

- Type of System: x64-based PC
- Operating System: Microsoft Windows 11
- Random Access Memory (RAM): 16GB
- Processor: Intel i5-11500, 6 cores, 12 threads, 2.7GHz base clock (4.6GHz boost)
- Graphics Card: NVIDIA GeForce GTX 3050, 4GB GDDR6

4.3 Functional requirements

- Preprocessing the dataset, including handling missing values, feature scaling, and class balancing.
- Analysing the correlation of features to select meaningful variables for model training.
- Splitting the dataset into training and testing sets.
- Implementing machine learning models (Random Forest and XGBoost) for classification tasks.
- Training and evaluating deep learning models, such as Feedforward Neural Networks (FNNs) and Deep Neural Networks (DNNs), for regression and classification tasks.
- Visualization tools for interpreting model outcomes, such as feature importance and prediction accuracy.
- Predicting electric range and classifying electric vehicle types (BEVs vs. PHEVs).
- Assessing model performance using metrics such as accuracy, precision, recall, and RMSE.

4.4 Non-Functional Requirements

- Utilizing Google Collaboratory for coding and model training to leverage its GPU and TPU support.
- The system should have minimal latency for efficient processing of large datasets.
- Ensuring data security to protect sensitive vehicle registration information.
- The system should handle variations in data distribution without performance degradation.

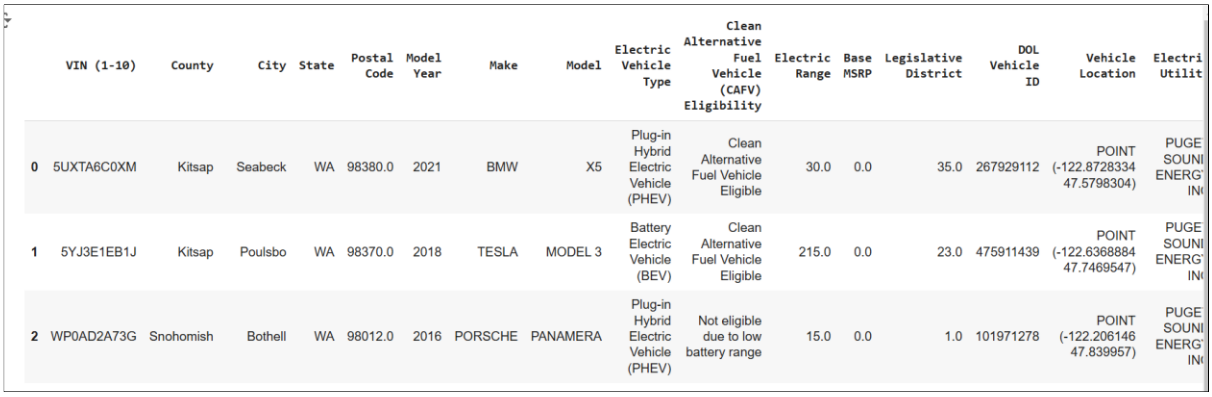
4.5 Technical Requirements

- Implementation of frameworks and libraries such as TensorFlow, Scikit-learn, and Matplotlib for model building and data visualization.
- The use of Python for coding and analysis due to its rich ecosystem of machine learning libraries.
- Regular data backups and recovery strategies to prevent data loss during experimentation.
- GPU acceleration via NVIDIA GTX 3050 to handle deep learning workloads efficiently.

SECTION IV.

5. Dataset

The dataset was carefully chosen to help the creation of machine learning and deep learning models that can predict the range of electric vehicles (EVs), sort vehicles into types (BEV vs. PHEV), and find out which vehicles are eligible for Clean Alternative Fuel Vehicles (CAFVs). This dataset lets us look at all the things that affect the use and success of electric vehicles. It can be used to help make policy, study the environment, and do consumer analytics. The data set is obtained from the Washington State Department of Licensing and includes complete records of all EV entries. It includes important details like Vehicle Type, Electric Range, and CAFV Eligibility. There are about 210,165 rows and 17 columns in the data, which show details about geographical, technical, and demographic factors.



	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	Electri Utilit
0	5UXTA6C0XM	Kitsap	Seabeck	WA	98380.0	2021	BMW	X5	Plug-in Hybrid Electric Vehicle (PHEV)	Clean Alternative Fuel Vehicle Eligible	30.0	0.0	35.0	267929112	POINT (-122.8728334 47.5798304)	PUGE SOUN ENERG IN
1	5YJ3E1EB1J	Kitsap	Poulsbo	WA	98370.0	2018	TESLA	MODEL 3	Battery Electric Vehicle (BEV)	Clean Alternative Fuel Vehicle Eligible	215.0	0.0	23.0	475911439	POINT (-122.6368884 47.7469547)	PUGE SOUN ENERG IN
2	WP0AD2A73G	Snohomish	Bothell	WA	98012.0	2016	PORSCHE	PANAMERA	Plug-in Hybrid Electric Vehicle (PHEV)	Not eligible due to low battery range	15.0	0.0	1.0	101971278	POINT (-122.206146 47.839957)	PUGE SOUN ENERG IN

Fig. 3-A snippet of the dataset.

Collecting and preparing the data

- The dataset is sourced from https://catalog.data.gov/dataset/electric-vehicle-population-data
- Attributes come from official EV registration records, which include information about the car, its location, and economic indicators.
- As part of the preprocessing steps, missing values are solved with, duplicates are removed, and categorical variables are changed into forms that computers can read.
- Oversampling or SMOTE are two methods used to fix data imbalances in groups like "Electric Vehicle Type."

To make sure the model works in real life, the information is split into Training, Testing, and Validation sets.

6. DESIGN

Step 1: Obtaining Data and Preprocessing
The dataset, obtained from the Washington State Department of Licensing, was uploaded to Google Drive and imported into Google Colab for processing. The data was subjected to thorough pre-processing, encompassing the management of absent values, detection of outliers, and normalization of features for uniformity. Categorical variables were encoded, and the dataset was partitioned into training and testing subsets. Pre-processing guaranteed data consistency, enhanced model performance, and reduced noise.

Step 2: Analysis of Features and Detection of Outliers
Feature analysis was conducted utilizing correlation matrices and feature importance metrics to ascertain the most influential variables for regression and classification tasks. Outlier identification approaches, including visualization techniques, guaranteed that abnormalities in the dataset did not negatively affect model performance. This stage also facilitated the comprehension of feature relationships.

Step 3: Modeling for Classification and Regression
Three models were employed for classification tasks to distinguish between Battery Electric Vehicles (BEVs) and Plug-in Hybrid Electric Vehicles (PHEVs): Random Forest, XGBoost, and a Feedforward Neural Network (FNN). A Random Forest Regressor and a Deep Neural Network (DNN) were employed for regression to forecast electric range. Feature selection and hyperparameter optimization were employed to enhance model accuracy.

Step 4: Anomaly Identification
The Isolation Forest model was utilized to identify anomalies in the dataset, including false or extreme values in significant characteristics. This guaranteed the preservation of data integrity throughout the procedure.

Step 5: Evaluation of the Model and Results
The performance of each model was assessed using relevant measures, including accuracy, precision, recall, RMSE, and F1 scores. The outcomes were analysed to determine the optimal model for each activity. Overfitting was assessed and mitigated using regularization methods and cross-validation.

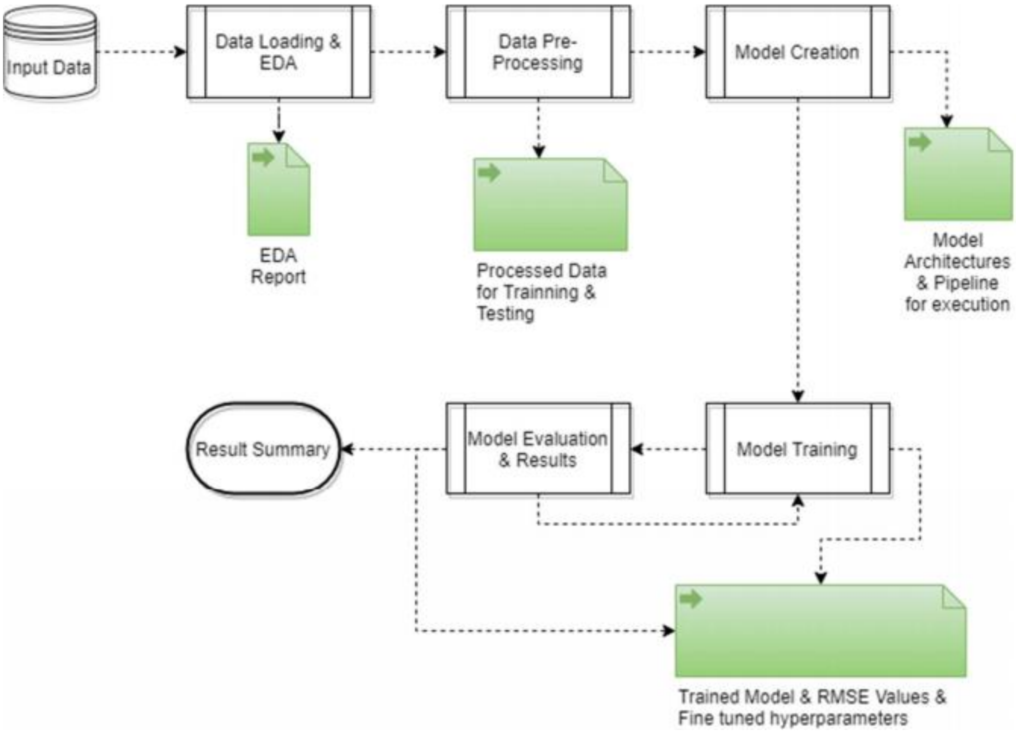

Input Data
Data Loading & EDA
Data Pre-Processing
Model Creation
EDA Report
Processed Data for Trainning & Testing
Model Architectures & Pipeline for execution
Result Summary
Model Evaluation & Results
Model Training
Trained Model & RMSE Values & Fine tuned hyperparameters

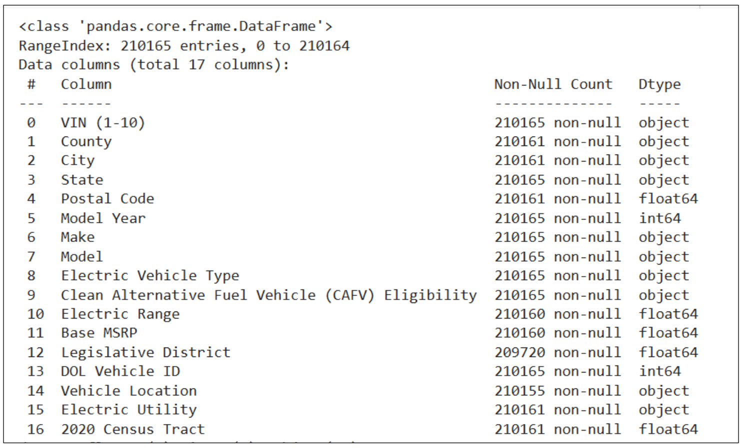
Fig. 4 - Block Diagram for the Design Flow.

SECTION V.

7.TECHNIQUES (Exploratory Data Analysis)

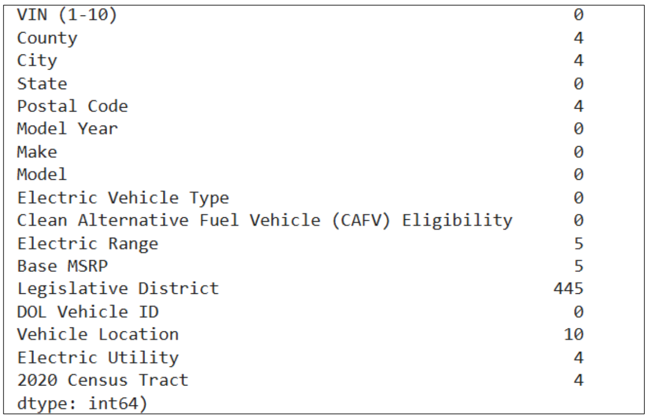
7.1 Data Overviewing

Once the dataset has been loaded properly, the next step is to get a general idea of how it is organized and what it contains. This is done by looking at the shape of the dataset to see how many rows and columns it has and then looking at the data types of each column to see what kinds of characteristics are there, such as numerical and categorical. Using Python's pandas library, tools like info() as shown in fig 5 and head() give a short summary of the dataset by showing sample records and non-null counts fig 6.



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 210165 entries, 0 to 210164
Data columns (total 17 columns):
 #   Column                                             Non-Null Count   Dtype
---  ------                                             --------------   -----
 0   VIN (1-10)                                         210165 non-null  object
 1   County                                             210161 non-null  object
 2   City                                               210161 non-null  object
 3   State                                              210165 non-null  object
 4   Postal Code                                        210161 non-null  float64
 5   Model Year                                         210165 non-null  int64
 6   Make                                               210165 non-null  object
 7   Model                                              210165 non-null  object
 8   Electric Vehicle Type                              210165 non-null  object
 9   Clean Alternative Fuel Vehicle (CAFV) Eligibility  210165 non-null  object
 10  Electric Range                                     210160 non-null  float64
 11  Base MSRP                                          210160 non-null  float64
 12  Legislative District                               209720 non-null  float64
 13  DOL Vehicle ID                                     210165 non-null  int64
 14  Vehicle Location                                   210155 non-null  object
 15  Electric Utility                                   210161 non-null  object
 16  2020 Census Tract                                  210161 non-null  float64
```

Fig. 5 – Represents the info of the dataset



```
VIN (1-10)                                            0
County                                                4
City                                                  4
State                                                 0
Postal Code                                           4
Model Year                                            0
Make                                                  0
Model                                                 0
Electric Vehicle Type                                 0
Clean Alternative Fuel Vehicle (CAFV) Eligibility     0
Electric Range                                        5
Base MSRP                                             5
Legislative District                                445
DOL Vehicle ID                                        0
Vehicle Location                                     10
Electric Utility                                      4
2020 Census Tract                                     4
dtype: int64)
```

Fig. 6 – Shows the number of values in each column of the dataset

Also, describe() is used to make statistical summaries for numerical fields, like mean, median, and standard deviation. This step also looks for possible problems, like missing numbers or strange things. These insights help with the preprocessing step by showing where cleaning, feature engineering, or estimation are needed. This basic understanding makes sure that further research is structured and focused.

Table 1: Description of the Columns

Column Name	**Description**
VIN (1-10)	The first 10 characters of a vehicle's unique identification number for tracking and identification.
County	The registered county associated with the vehicle's owner, covering in-state and out-of-state vehicles.
City	The city in which the vehicle's registered owner resides.
State	The state or region associated with the vehicle's registration details.
Postal Code	The postal (ZIP) code of the registered vehicle owner's address.
Model Year	The year of manufacture of the vehicle, derived from the Vehicle Identification Number (VIN).
Make	The manufacturer or brand of the vehicle (e.g., Tesla, Nissan).
Model	The specific model of the vehicle (e.g., Model S, Leaf).

Electric Vehicle Type	Differentiates between BEVs (Battery Electric Vehicles) and PHEVs (Plug-in Hybrid Electric Vehicles).
Clean Alternative Fuel Vehicle (CAFV) Eligibility	Indicates if the vehicle qualifies under the Clean Alternative Fuel Vehicle standards as defined by House Bill 2042.
Electric Range	The maximum distance a vehicle can travel using only its electric charge, measured in miles.
Base MSRP	The Manufacturer's Suggested Retail Price (MSRP) for the vehicle's base trim level.
Legislative District	Specifies the legislative district in Washington State where the registered vehicle owner resides.
DOL Vehicle ID	A unique identification number assigned to each vehicle by the Department of Licensing (DOL).
Vehicle Location	GPS coordinates or geographical details representing the vehicle's registration location.
Electric Utility	The electric power provider for the vehicle's registered address, including cooperatives and municipal services.
2020 Census Tract	A geographic identifier combining state, county, and census tract codes as per the 2020 US Census.

7.2 Data Handling

Data handling is an essential phase in any data science workflow, particularly for machine learning initiatives, since the caliber of input data directly influences the precision and efficacy of prediction models. Our study's dataset comprises 210,165 records across 17 columns, and the analysis indicated the existence of missing values in several variables, as illustrated in Fig. 6. Attributes including Postal Code, Electric Range, Base MSRP, and Legislative District display numerical missing values, whereas County, City, Electric Utility, and Vehicle Location contain categorical null entries. Resolving these concerns guarantees the dataset's integrity and dependability for future study.

Missing value treatment encompasses two methodologies:

Imputation: For numerical columns, absent values were substituted with the mean of their corresponding columns. This guarantees the dataset's statistical integrity and averts data skewing.

Row Elimination: For categorical characteristics, rows with absent values were discarded to prevent the introduction of biases during encoding and model training. This strategy is efficacious when absent data represents a minor proportion of the dataset.

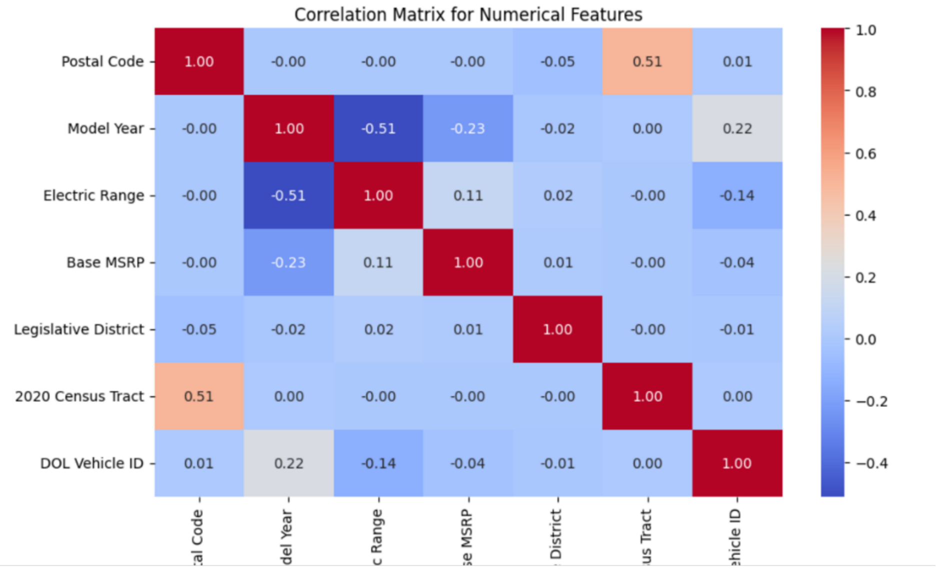
Reasons for Data Management Essential

Data inconsistencies, including absent values, generate noise that impairs the efficacy of machine learning algorithms. (Kotsiantis, 2006) assert that rigorous preprocessing guarantees the dataset is devoid of abnormalities and enhances the effectiveness of prediction algorithms such as Random Forest and XGBoost. Furthermore, addressing missing values is crucial for deep learning models, which are susceptible to data deficiencies.

Principal Advantages:

- Enhances model robustness by reducing noise.
- Maintains dataset integrity by the application of suitable imputation methods.
- Mitigates the possibility of skewed or incomplete predictions, hence augmenting the overall validity of the investigation.

7.3 Correlation Matrix



	al Code	del Year	c Range	e MSRP	District	us Tract	hicle ID
Postal Code	1.00	-0.00	-0.00	-0.00	-0.05	0.51	0.01
Model Year	-0.00	1.00	-0.51	-0.23	-0.02	0.00	0.22
Electric Range	-0.00	-0.51	1.00	0.11	0.02	-0.00	-0.14
Base MSRP	-0.00	-0.23	0.11	1.00	0.01	-0.00	-0.04
Legislative District	-0.05	-0.02	0.02	0.01	1.00	-0.00	-0.01
2020 Census Tract	0.51	0.00	-0.00	-0.00	-0.00	1.00	0.00
DOL Vehicle ID	0.01	0.22	-0.14	-0.04	-0.01	0.00	1.00

Fig. 7 – Heat Map for the numerical features

Figure 7 illustrates the correlation matrix, emphasizing the pairwise correlations among the numerical features within the dataset. Utilizing Matplotlib and Seaborn, these packages enable the creation of clear and aesthetically pleasing visualizations. The Seaborn heatmap is highly successful, employing a color-coded gradient to signify the magnitude and direction of correlations, with red indicating strong positive correlations and blue indicating strong negative correlations.

Important Relationships: A correlation of -0.51 between Electric Range and Model Year indicates a moderate negative association, suggesting that older car models often had shorter ranges due to less sophisticated battery technologies. A robust correlation of 0.51 between Postal Code and the 2020 Census Tract indicates regional linkages.

Weak Relationships: Weak correlations (around 0) between features like Base MSRP and Legislative District suggest that they are independent and unlikely to have a direct impact on one another.

Insights on Feature Selection:

Strongly linked variables, such as Postal Code and 2020 Census Tract, indicate potential redundancy, necessitating dimensionality reduction.

Weakly linked variables, such as Legislative District and Electric Range, suggest that these properties offer independent information advantageous for model performance.

This heatmap is a vital component of exploratory data analysis, facilitating the identification of a high degree of redundant features, and independent variables crucial for model development and predictive accuracy.

7.4 CAFV Eligibility Distribution

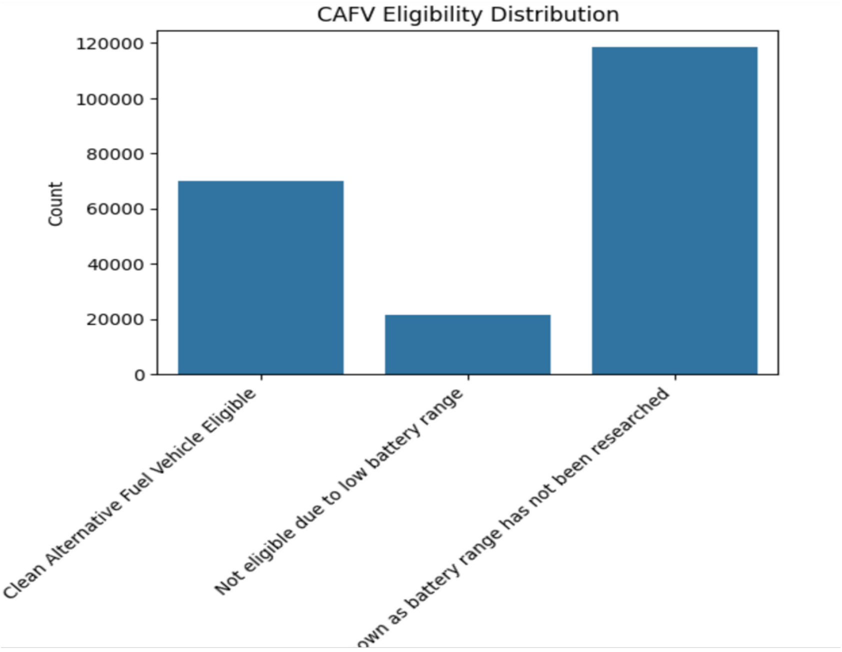


Fig 8 – Distribution graph for CAFV Categories

Figure 8 depicts the distribution of Clean Alternative Fuel Vehicle (CAFV) eligibility within the dataset, offering a detailed classification of vehicles as eligible, ineligible owing to inadequate battery range, or unclassified due to insufficient data. This visualization was created utilizing the Seaborn library with a count plot, successfully illustrating the distributions of category variables.

The graph indicates that most vehicles are classified as "Eligibility Not Researched," signifying deficiencies in data concerning battery range validation. A considerable quantity of automobiles qualifies as CAFV, consistent with Washington's initiatives to advance sustainable transportation. The smallest category comprises vehicles deemed ineligible due to restricted battery range, underscoring a domain for possible legislative or technological enhancement.

Recognizing the distribution of CAFV eligibility is essential to this research, as it offers insights into the dataset's class imbalance, which directly influences the training and performance of classification models. This approach also highlights data gaps, facilitating the prioritization of future data gathering or imputation methods. This plot is crucial for illustrating the complexities of identifying and predicting CAFV eligibility, facilitating a clear understanding of the dataset.

7.5 Electric Vehicle Type Distribution

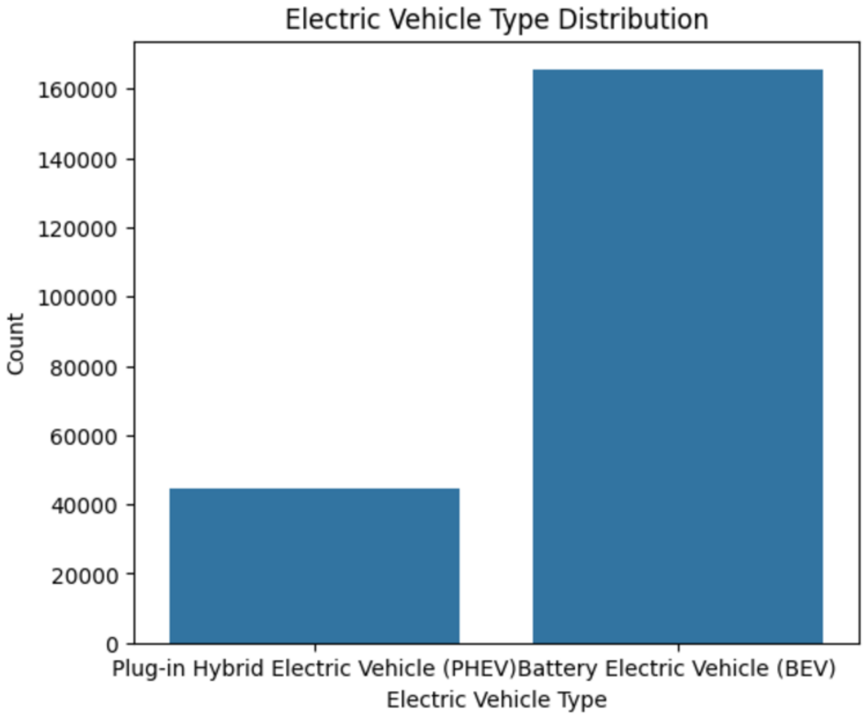


Fig 9 – Represents the class balance for Electric Vehicle Type

Figure 9 depicts the class distribution for the Electric Vehicle Type feature, indicating a notable imbalance in the dataset. The predominant type of cars is Battery Electric cars (BEVs), totalling 165,554 times, whilst Plug-in Hybrid Electric Vehicles (PHEVs) account for merely 44,611 occasions. This signifies that BEVs comprise roughly 79% of the entire dataset.

The gap shows the increasing use of fully electric vehicles, propelled by innovations in battery technology and heightened environmental consciousness. This lopsided distribution can result in biased predictions favouring BEVs in machine learning models. Rectifying this imbalance using methods like as oversampling, under sampling, or class weighting is crucial for developing strong and equitable classification models.

7.6 Feature Scaling: Standard Scaler

Standard scaling is an essential phase in data preprocessing that standardizes numerical features by eliminating the mean and adjusting to unit variance. This guarantees that all features contribute uniformly to the machine learning models, preventing those with greater magnitudes from overshadowing the training process. The Standard Scaler, a component of the Scikit-learn toolkit, normalizes data by focusing it at zero with a standard deviation of one. This is especially beneficial for algorithms that are sensitive to the distribution of feature values, such neural networks and support vector machines.

In line with this project, the Electric Range and Base MSRP columns were standardized to normalize the input values, hence enhancing convergence speed during training and augmenting the model's performance. This modification is particularly advantageous when employing models such as Random Forest for classification and Deep Neural Networks (DNN) for regression.

Standard scaling ensured data consistency, hence boosting the models' capacity to generalize on unfamiliar data. In the absence of scaling, models may erroneously perceive features of greater magnitude as more important.

7.7 One-hot Encoding

One-hot encoding is a technique for transforming categorical information into numerical representation by generating binary columns for each distinct category. This removes any unwanted numeric relations created by methods such as label encoding. The Electric Utility and City columns in your dataset were one-hot encoded utilizing Pandas' get_dummies() function, generating dummy variables that enabled machine learning models to read categorical data efficiently.

This encoding technique inhibited the establishment of artificial hierarchy, hence ensuring strong classification and regression efficacy. For example, Random Forest and XGBoost utilized these encoded factors to enhance predictions while accurately interpreting categorical correlations.

7.8 Label Encoding

Label encoding transforms categorical variables into numerical representation by allocating a distinct integer to each category. Although efficient, it is best suited for features with an actual linear relationship. In your dataset, the attributes Make and Model were label-encoded, enabling regression models to utilize these categorical variables without increasing the dataset size.

In regression tasks, label encoding increased processing speed while maintaining feature ability to be translated. But it was utilized selectively to prevent unfairness in classification models.

7.9 SMOTE (Synthetic Minority Over-sampling Technique)

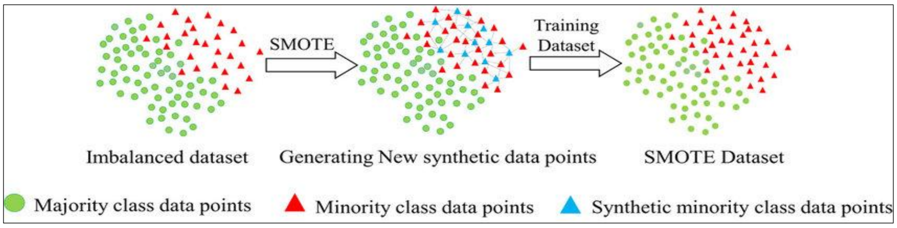

SMOTE
Training Dataset
Imbalanced dataset
Generating New synthetic data points
SMOTE Dataset
Majority class data points
Minority class data points
Synthetic minority class data points

Fig 10 – SMOTE Technique

Source-https://pub.aimind.so/synthetic-minority-over-sampling-technique-smote-empowering-ai-through-imbalanced-data-handling-d86f4de32ea3

SMOTE (Synthetic Minority Over-sampling Technique) is a common method employed to rectify class imbalance in datasets. It operates by producing synthetic data points for the minority class via interpolation. Rather than just replicating current minority class samples, SMOTE examines the feature space of minority class instances and generates new examples by amalgamating existing samples with their nearest neighbours. This mitigates the risk of overfitting typically linked to simplistic oversampling.

The classification task of forecasting Electric Vehicle Type (Battery Electric Vehicle (BEV) versus Plug-in Hybrid Electric Vehicle (PHEV)) encountered considerable class imbalance. Failure to rectify this imbalance may lead models such as Random Forest, XGBoost, or Neural Networks to exhibit bias towards the majority class (BEV), hence diminishing predicted accuracy for the minority class (PHEV). SMOTE equilibrated the dataset by generating additional PHEV instances, guaranteeing that the model obtained an equitable representation of both classes during training.

Utilizing SMOTE enhanced the model's generalizability for the minority class (PHEV), which is essential for achieving credible predictions. This enhanced the model's precision, recall, and F1-score for the minority class, resulting in more equitable evaluation measures overall. Furthermore, SMOTE alleviated the risk of overfitting linked to repetitive duplication by including heterogeneity via synthetic samples. This was crucial for improving the interpretability of the data and guaranteeing that the classification models operated effectively in real-world situations.

Although SMOTE is effective, it must be employed carefully, particularly with high-dimensional datasets or ones with overlapping classes. It is crucial to assess its influence on overall model efficacy and confirm that the produced synthetic samples are relevant within the dataset's context.

7.10 Train-Test Split

Train-test splitting is a crucial procedure in machine learning for assessing a model's capacity to generalize to novel data. Conventionally, an 80-20 division is employed, allocating 80% of the data for training and 20% for testing. This ratio achieves equilibrium, enabling the model to train efficiently while preserving adequate data for impartial review. The train_test_split() function from Scikit-learn is a widely utilized utility that guarantees the random division of data into categories.

This study employed train-test splitting for both classification problems, such as predicting Electric Vehicle Type, and regression tasks, such as predicting Electric Range. The training set facilitates the model's acquisition of patterns from the data, whereas the testing set evaluates its performance on novel data. Train-test splitting reduces overfitting concerns and guarantees accurate performance measures by assessing on unique data, rendering it an essential procedure in developing robust, generalizable models.

8. MODELLING

8.1 Random Forest

Random Forest is an extensive ensemble learning technique proficient in managing both classification and regression assignments. It generates several decision trees as shown in fig 11 during training and consolidates their outputs using majority voting for classification and averaging for regression to improve predictive accuracy and mitigate overfitting. This method leverages the advantages of individual trees while addressing their shortcomings, resulting in resilient models appropriate for diverse applications.

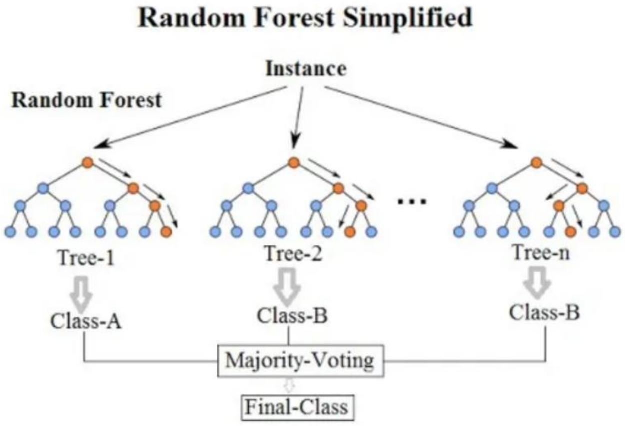


Fig 11 – Random Forest (Decision Trees)

Source: *https://williamkoehrsen.medium.com/random-forest-simple-explanation-377895a60d2d*

Random Forest Classification: In classification contexts, Random Forest demonstrates superiority by aggregating predictions from multiple decision trees to ascertain the most likely class for a specific input. This ensemble approach efficiently handles high-dimensional data and intricate relationships among characteristics. A study by (Iranzad, 2024) demonstrates the effectiveness of Random Forest in feature selection for classification tasks, emphasizing its capacity to discern pertinent variables and enhance model performance.

Random Forest Regression: In regression tasks, Random Forest forecasts continuous outcomes by averaging the predictions of individual trees. This approach elucidates complex interactions between predictors and responses, providing adaptability in modeling non-linear dependencies. The (Polyzos, 2023)authors demonstrated the efficacy of Random Regression Forests in optimal lag selection for time series data, highlighting its effectiveness in financial research.

Random Forests are distinguished by their robustness against overfitting, particularly in the context of extensive datasets and numerous variables. They offer insights on feature significance, facilitating the interpretation of model predictions. Nonetheless, they can be computationally demanding, and the ensemble characteristic may hinder the interpretability of individual decision pathways.

8.2 XG Boost

XGBoost (Extreme Gradient Boosting) is a cutting-edge gradient boosting technique extensively employed for supervised learning problems. It improves conventional gradient boosting by the incorporation of innovative methods like as tree cutting parallelized computation, and regularization, resulting in enhanced computational efficiency and resilience to overfitting (Chen, 2016). XGBoost is an ensemble technique that incrementally builds decision trees, with each successive tree aimed at reducing the faults of its predecessors. Its scalability and flexibility enable smooth integration with structured datasets in classification and regression contexts.

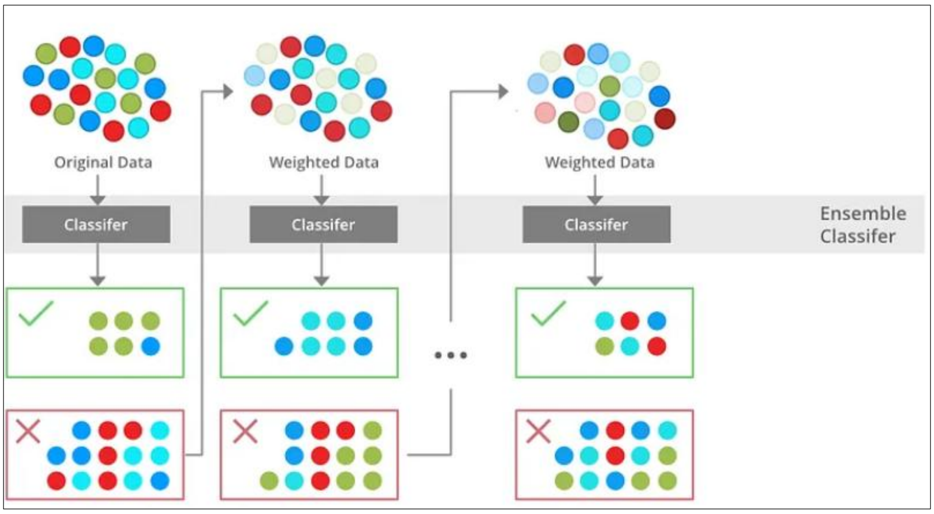


Fig.12 – XGBoost Working

Source- https://www.geeksforgeeks.org/xgboost/

Features of XGBoost:

Gradient-oriented Optimization employs a second-order Taylor approximation to calculate gradients, hence enhancing the model's convergence rate.

Regularization: XGBoost employs L1 (Lasso) and L2 (Ridge) regularization to manage model complexity and mitigate overfitting.

Handling Missing Data: The algorithm can autonomously handle absent values by optimizing the partitions according to the training data.

Feature Importance: XGBoost offers a hierarchy of features according to their predictive efficacy, facilitating feature engineering.

This research employs XGBoost for classification, specifically in predicting Electric Vehicle Type. The model classifies by recognizing patterns that differentiate BEVs from PHEVs through the analysis of complex relationships within the dataset.

8.3 Feedforward Neural Networks (FNNs)

Feedforward Neural Networks (FNNs), illustrated in Fig. 13, constitute one of the most fundamental types of artificial neural networks. They comprise three fundamental components: an input layer, one or more hidden layers, and an output layer. In a Feedforward Neural Network (FNN), information progresses unidirectionally—from the input layer, through the hidden layers, to the output layer. This architecture lacks cycles or feedback loops.

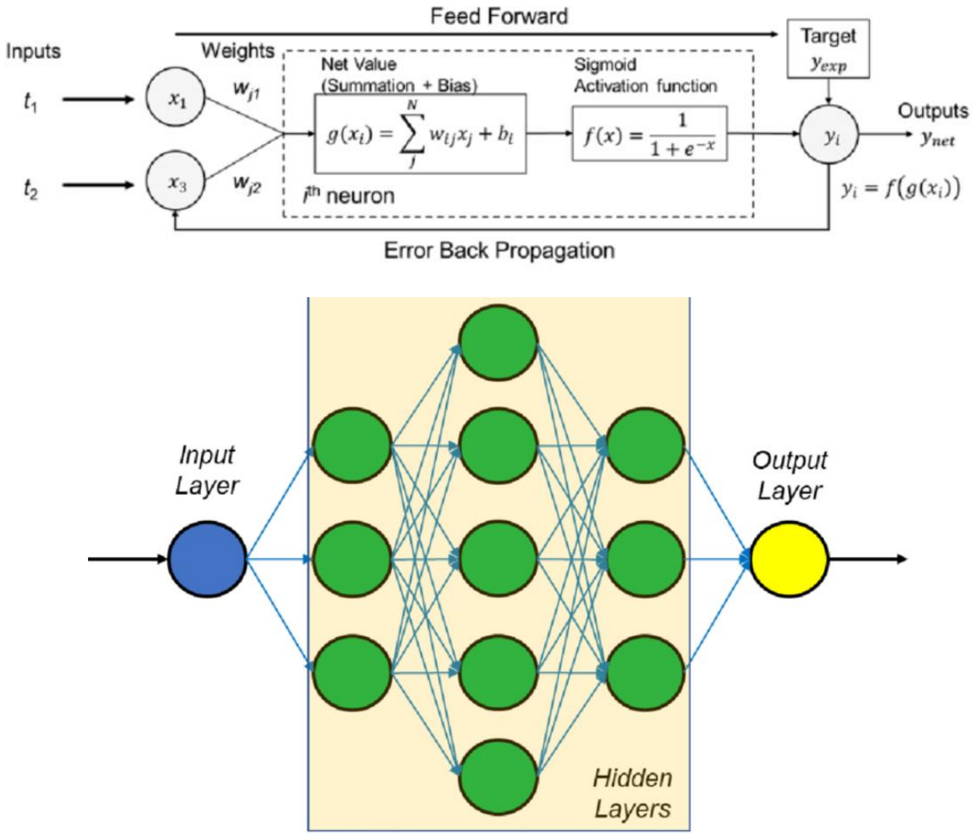


Fig. 13 – FNN Architecture

Source- https://www.researchgate.net/figure/Schematic-of-FNN-architecture-with-a-back-propagation-algorithm-14_fig4_363667901

Each neuron in a layer is connected to every neuron in the subsequent layer via weights (Goodfellow, 2016). The input values (x1,x2,...,xnx_1, x_2, ..., x_nx1,x2,...,xn) are multiplied by these weights (wijw_{ij}wij) and summed up, with a bias term (bbb) added. This sum is passed through an activation function (f(x)f(x)f(x), such as Sigmoid or ReLU) to introduce non-linearity, generating an output. This process is mathematically represented as:

$$y = f\left(\sum_{i=1}^{N} w_i x_i + b\right)$$

The output layer provides the final prediction, while the error between predicted and actual values is propagated back through the network (backpropagation) to adjust weights and minimize the loss.

This project uses the FNN classifier to predict the type of electric vehicle (classification problem). The architecture's simplicity is appropriate for this issue, as it adeptly captures correlations among characteristics following preprocessing and scaling. Input parameters such as Electric Range and Base MSRP are utilized to categorize automobiles as BEV or PHEV. FNNs are effective for structured data and comparatively simple to build. Nevertheless, they are incapable of capturing spatial or temporal connections in data, rendering them inappropriate for picture or sequential data. Their success significantly depends on preprocessing processes like as feature scaling and encoding, in addition to the selection of hyper parameters.

8.4 Deep Neural network (DNN)

A Deep Neural Network (DNN) is a form of artificial neural network consisting of numerous layers of interconnected neurons. It is engineered to find complex patterns and correlations among data by transmitting information across its numerous layers, each of which extracts progressively advanced features. This study uses a Deep Neural Network (DNN) for regression to forecast the Electric Range of electric vehicles.

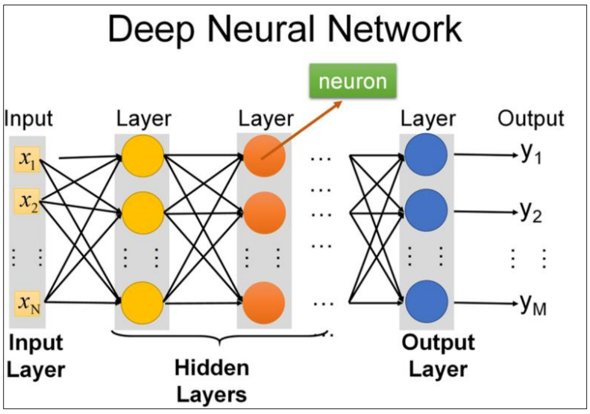


Fig. 14- DNN Architecture

Source- https://www.researchgate.net/figure/A-typical-architecture-of-DNN_fig2_335845675

The architecture of the DNN for this dataset are:

1. **Input Layer**: Accepts the scaled features (X_train_scaled) from the dataset. The input layer ensures that the data is passed into the network.
2. **Hidden Layers**:
    - **First Layer**: A dense layer with 128 neurons and ReLU (Rectified Linear Unit) activation function. It captures the initial relationships in the dataset.
    - **Second Layer**: A dense layer with 64 neurons and ReLU activation, further processing the extracted features.
    - **Third Layer**: A dense layer with 32 neurons and ReLU activation for deeper feature extraction.
3. **Output Layer**: A single neuron with no activation (linear output) is used for regression tasks, predicting a continuous value such as the electric range.

This model refers to the electric vehicle dataset. Identifies non-linear associations among attributes such as car manufacturer, model, and electric utility. Analyse extensive datasets to discern complex patterns in the input variables, facilitating precise predictions.

Layers of the code:

Dense Layers: Fully connected layers in which each neuron from one layer is interconnected with every neuron in the subsequent layer.
Activation Functions: The ReLU function introduces non-linearity, allowing the network to acquire intricate patterns.

8.5 Anomaly Detection

The Isolation Forest is a machine learning technique particularly designed for anomaly identification. In contrast to conventional methods, it isolates anomalies rather than profiling standard data. It functions on the basis that anomalies are rare and distinct, facilitating their isolation. The model separates data by the construction of a succession of randomized decision trees (Isolation Trees) utilizing random feature splits. The fewer divisions need to isolate a data point, the greater its likelihood of being an oddity. This method is cost-effective and exceptionally successful for high-dimensional datasets.

Anomaly detection is essential for spotting anomalies in data, which may signify errors, fraud, or unforeseen behaviour. Within the electric vehicle dataset, anomalies may reveal variations, such as abnormally elevated MSRP values for specific car ages or electric ranges.

The Isolation Forest detected irregularities in the properties of Electric Range and MSRP for this project. These abnormalities may indicate data entry inaccuracies, uncommon car models, or exceptional situations necessitating additional scrutiny. An electric vehicle with a manufacturer's suggested retail price surpassing $800,000 or an abnormally short range may indicate data discrepancies or distinctive product attributes.

In conclusion, utilizing Isolation Forest for anomaly identification enhanced the data analysis by targeting outliers for focused investigation. This methodology guaranteed data integrity and provided insights into irregular patterns, rendering it a crucial instrument for enhancing model precision and dependability.

SECTION VI.

9.EVALUATION/RESULTS

9.1 Classification

In machine learning, assessing model performance is essential, particularly for novel data, as dependence on accuracy alone may result in misinterpretation. This project employs many measures to thoroughly evaluate the classification models. These encompass accuracy, confusion matrix, precision, recall, F1 score, and ROC-AUC.

Accuracy
One measure of accuracy is the proportion of correct predictions relative to the total number of forecasts. In a nutshell, it shows how well the model is doing. On the other hand, accuracy by

itself could be deceiving in datasets that are uneven, since it might not reveal that the model is failing to accurately forecast minority classes.

The Confusion Matrix

Prediction outcomes across many categories are graphically represented by the confusion matrix:

True Positive (TP): Instances of positive predictions that were accurate.

Positively anticipated unfavourable occurrences: **True Negatives (TN).**

Incorrectly predicted positive occurrences (Type I error) are known as **false positives** (FP).

Incorrectly projected negative occurrences (Type II error): this is known as a **false negative (FN).**
Analysing model flaws and performance in depth is made possible by this matrix.

Precision

The percentage of correct predictions relative to the total number of positive forecasts is called precision. Important situations when the cost of false positives is considerable, such as when wrong classifications impact operational or business choices, make this a must-have.

Recall

The percentage of true positives that the model properly identifies is called recall. In anomaly detection and other situations where false negatives (positive cases not detected) might have serious implications, this is of the utmost importance.

F1 Score Finding the harmonic mean of recall and precision yields the F1 score. It gives a fair assessment of the model's accuracy in both positive and negative outcomes, which is very helpful in datasets that aren't balanced.

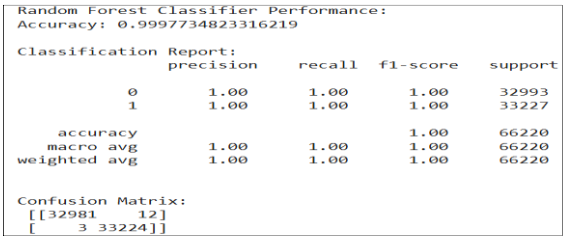
AUC and ROC Curve

The ROC curve compares the **True Positive Rate (TPR)** with the **False Positive Rate (FPR),** showing how the two metrics are related. The AUC measures how well the model can differentiate between different classes on the whole. The ability to discriminate is better demonstrated by a greater AUC.

For purposes like projecting the types of electric vehicles (BEV vs. PHEV), these criteria are essential for a comprehensive assessment of the categorization models.

9.1.1 Final Results

1) Random Forest



```
Random Forest Classifier Performance:
Accuracy: 0.9997734823316219

Classification Report:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00     32993
           1       1.00      1.00      1.00     33227

    accuracy                           1.00     66220
   macro avg       1.00      1.00      1.00     66220
weighted avg       1.00      1.00      1.00     66220


Confusion Matrix:
 [[32981    12]
 [    3 33224]]
```

Fig. 15 – Performance Metrics of Random Forest Classifier

The picture illustrates the performance metrics of the Random Forest Classifier utilized for classifying Electric Vehicle Types (BEV versus PHEV). The attained accuracy is 99.98%, signifying exceptional predictive capabilities.

The classification report offers comprehensive insights:

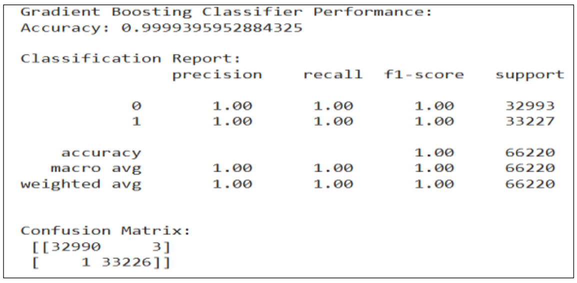
Precision (1.00 for both classes) signifies that all positive predictions generated by the model are exceptionally accurate, with little false positives.

Recall (1.00 for both classes) indicates the model's capacity to accurately detect all true positives without omission.

The F1-Score (1.00) equilibrates precision and recall, indicating robust overall performance. The support column indicates the sample count for each class, facilitating an accurate assessment of balanced data.

The **confusion matrix** reveals that genuine predictions (32,981 and 33,224) predominate, with negligible false positives (12) and false negatives (3), demonstrating the model's robustness for this dataset. The results indicate the classifier's efficacy in properly detecting EV kinds.

2) XGBoost



```
Gradient Boosting Classifier Performance:
Accuracy: 0.9999395952884325

Classification Report:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00     32993
           1       1.00      1.00      1.00     33227

    accuracy                           1.00     66220
   macro avg       1.00      1.00      1.00     66220
weighted avg       1.00      1.00      1.00     66220


Confusion Matrix:
 [[32990     3]
 [    1 33226]]
```

Fig.16 – Performance Metrics of XGBoost

The picture shows **Gradient Boosting Classifier** (XGBoost) performance measures for the classification of Electric Vehicle Types (BEV against PHEV). With an overall accuracy of 99.99%, the model shows very dependability and precision.

The classification report notes the following:

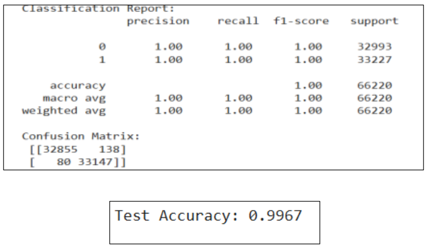
Precision 1.00 for both classes indicates that almost all positive forecasts were true, hence reducing false positives.

Recall (1.00 for both classes) guarantees that the model can detect all real positive cases, so guaranteeing no missed classifications.

Validating consistent performance, **F1-score** (1.00) shows a perfect mix of accuracy and recall.

Emphasizing the balanced evaluation across 32,993 samples for Class 0 and 33,227 samples for Class 1, the support column. With only 3 wrong positives and 1 false negative, the **confusion matrix** shows that actual predictions (32,990 and 33,226) virtually rule. This outstanding performance emphasizes how well the XGBoost classifier can generalize, hence producing high trust in predicts for the particular dataset.

3) Feed Forward Neural Network



```
Classification Report:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00     32993
           1       1.00      1.00      1.00     33227

    accuracy                           1.00     66220
   macro avg       1.00      1.00      1.00     66220
weighted avg       1.00      1.00      1.00     66220

Confusion Matrix:
 [[32855   138]
 [   80 33147]]
```

```
Test Accuracy: 0.9967
```

Fig.17 – Performance Metrics of feed forward neural network

With a test accuracy of 99.67% (Fig. 17), the Feedforward Neural Network (FNN) for binary classification in this research shows really good performance. Additional metrics are included

by the classification report:

Precision, Recall, and F1-Score: Both classes—0 for BEV and 1 for PHEV—achieved perfect values of 1.00 across these measures. Recall measures the fraction of true positive predictions among all positive predictions; precision denotes the harmonic mean of recall and F1-score offers a harmonic mean of precision and recall. These findings show the great dependability of the model in categorizing several kinds of vehicles.

The column for support shows every class in the test set's actual count. Class 0 has 32,993 cases; class 1 has 33,227 cases, so showing a fair distribution of classes.

The **confusion matrix** demonstrates:

Correct classifications for both classes allow True Positives (32,855 and 33,147). Minimal misclassifications, suggesting great predictive power, define False Positives and False Negatives (138 and 80).

The great test accuracy and favourable performance measures highlight the FNN's efficiency in managing the dataset, most likely because of well-structured layers (input, hidden, and output) and efficient Adam optimizer use.

9.1.2 Evaluation of Models Using ROC and AUC Metrics

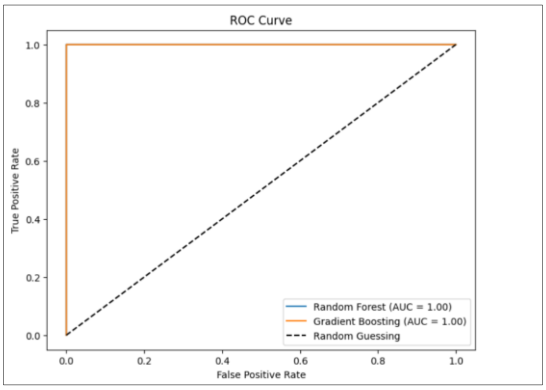


Fig. 18 – ROC curve for RF and XGB

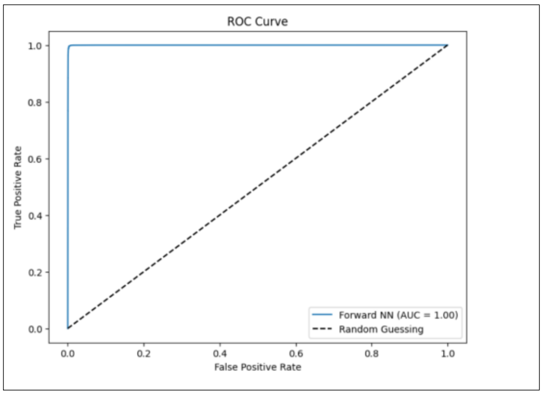
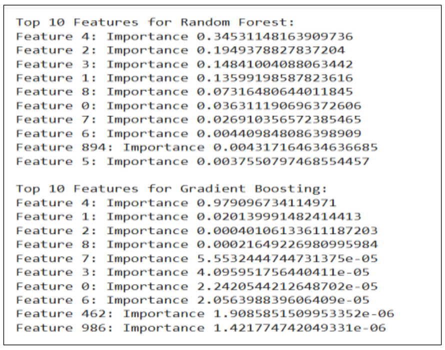


Fig.19 – ROC curve for Feed forward Neural network

In this dataset, all the three models as shown in the above fig.18 and fig.19 had an AUC of 1.0, indicating perfect performance with no misclassification. These results reflect the dataset's quality and the models' ability to leverage its characteristics for classification tasks.

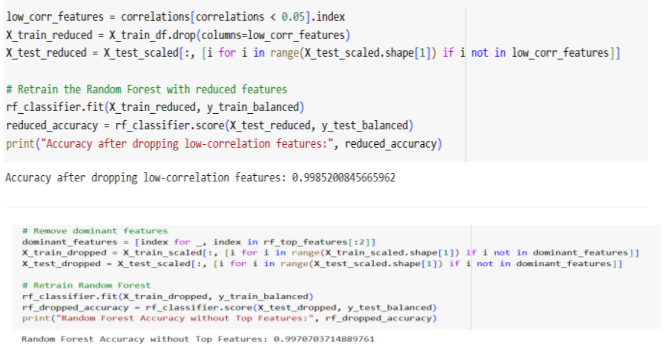
Model Comparison: While all models achieved good accuracy, **Random Forest** stood out for its robustness and ease of handling huge datasets without overfitting. **Gradient Boosting**, while similarly effective, requires iterative learning, which makes it computationally costly. **Feedforward Neural Networks** (FNN) fared marginally lower but remained competitive, especially in regression tasks. Random Forest emerges as the best fit due to its balance of efficiency and predictive capability, making it appropriate for the dataset.

9.1.3 Feature Importance and Model Simplification Analysis



```
Top 10 Features for Random Forest:
Feature 4: Importance 0.34531148163909736
Feature 2: Importance 0.1949378827837204
Feature 3: Importance 0.14841004088063442
Feature 1: Importance 0.13599198587823616
Feature 8: Importance 0.07316480644011845
Feature 0: Importance 0.036311190696372606
Feature 7: Importance 0.026910356572385465
Feature 6: Importance 0.004409848086398909
Feature 894: Importance 0.004317164634636685
Feature 5: Importance 0.003755079746855445 7

Top 10 Features for Gradient Boosting:
Feature 4: Importance 0.979096734114971
Feature 1: Importance 0.020139991482414413
Feature 2: Importance 0.00040106133611187203
Feature 8: Importance 0.00021649226980995984
Feature 7: Importance 5.55324447447313 75e-05
Feature 3: Importance 4.095951756440411e-05
Feature 0: Importance 2.242054421264870 2e-05
Feature 6: Importance 2.0563988396064 09e-05
Feature 462: Importance 1.908585150995335 2e-06
Feature 986: Importance 1.42177474204933 1e-06
```



```
low_corr_features = correlations[correlations < 0.05].index
X_train_reduced = X_train_df.drop(columns=low_corr_features)
X_test_reduced = X_test_scaled[:, [i for i in range(X_test_scaled.shape[1]) if i not in low_corr_features]]

# Retrain the Random Forest with reduced features
rf_classifier.fit(X_train_reduced, y_train_balanced)
reduced_accuracy = rf_classifier.score(X_test_reduced, y_test_balanced)
print("Accuracy after dropping low-correlation features:", reduced_accuracy)

Accuracy after dropping low-correlation features: 0.9985200845665962
```

```
# Remove dominant features
dominant_features = [index for _, index in rf_top_features[:2]]
X_train_dropped = X_train_scaled[:, [i for i in range(X_train_scaled.shape[1]) if i not in dominant_features]]
X_test_dropped = X_test_scaled[:, [i for i in range(X_test_scaled.shape[1]) if i not in dominant_features]]

# Retrain Random Forest
rf_classifier.fit(X_train_dropped, y_train_balanced)
rf_dropped_accuracy = rf_classifier.score(X_test_dropped, y_test_balanced)
print("Random Forest Accuracy without Top Features:", rf_dropped_accuracy)

Random Forest Accuracy without Top Features: 0.9970703714889761
```

Fig.20 – Top 10 features for the model

Fig.21 – Snippet of the code performing Feature importance test

Boosting models. It emphasizes that a limited number of factors substantially affect categorization performance, whereas others contribute negligibly or not at all. This is apparent in the feature importance scores, where a select few features prevail while the others are insignificant.

A study was performed to assess the stability of the Random Forest model by excluding the two most significant features. Remarkably, the model's accuracy persisted at a high level (about 99.7%), suggesting that alternative predictive traits offset the absence of the predominant ones. Additionally, when features indicating poor correlation (<0.05) with the target variable were eliminated, the accuracy remained over 99.85%. These findings emphasize the dataset's intrinsic separability and the model's robustness.

The results indicate that low-correlation characteristics provided minimal predictive value, and their elimination improved computational efficiency without compromising performance. This illustrates the Random Forest model's ability to identify intricate patterns across several features, highlighting its appropriateness for the dataset. The analysis indicates chances to simplify the data set while preserving high accuracy, hence ensuring efficiency and clarity in classification operations. This method confirms the lack of overfitting while assuring an effective model for implementation.

9.2 Regression

R² Score (Coefficient of Determination): The R² score measures how well the regression model explains the variance in the target variable. Its value ranges from 0 to 1, where:

- $R^2 = 1$: The model perfectly predicts all target values.
- $R^2 = 0$: The model performs no better than a simple mean prediction of the target variable.
- Negative R^2 values indicate that the model performs worse than the mean predictor. In this case, a near-perfect $R^2 = 0.99987$ indicates that the Random Forest Regressor explains almost all the variability in the target variable.

Mean Absolute Error (MAE): MAE quantifies the average magnitude of errors in a regression model without considering their direction (positive or negative).

It is calculated as:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

Mean Squared Error (MSE): MSE penalizes larger errors more than MAE since errors are squared before averaging. It is computed as:

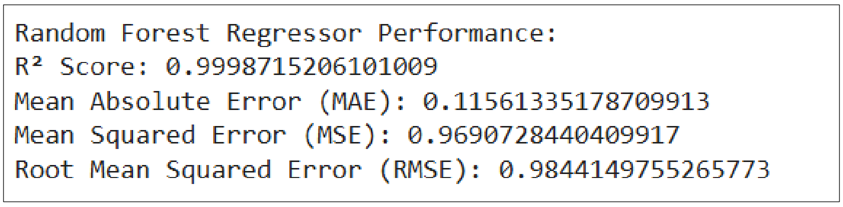
$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Root Mean Squared Error (RMSE): RMSE is the square root of MSE and provides an interpretation in the same units as the target variable:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

9.2.1 Final Results

1)Random Forest regressor:



```
Random Forest Regressor Performance:
R² Score: 0.9998715206101009
Mean Absolute Error (MAE): 0.11561335178709913
Mean Squared Error (MSE): 0.9690728440409917
Root Mean Squared Error (RMSE): 0.9844149755265773
```

Fig. 22 - Performance of Random Forest Regressor

Fig. 22 illustrates the performance metrics of the Random Forest Regressor utilized on the Electric Vehicle dataset. An outstanding R^2 score of 0.99987 signifies that the model accounts for almost all the variance in the dataset. Low values of Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE) indicate great predictive accuracy and negligible error.

Fig. 23 depicts the correlation between actual and forecasted EV ranges. The proximity of points to the diagonal "Perfect Prediction" line validates the model's efficacy in capturing the data's non-linear trends and properly forecasting EV ranges.

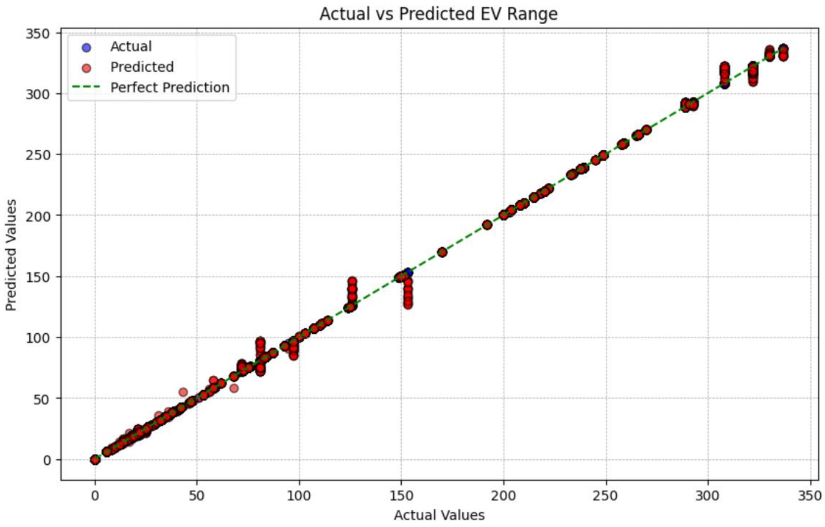
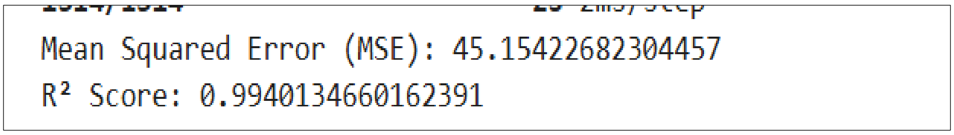


Fig. 23 - Actual vs predicted range for RF

2) Deep Neural Network

Figure 24: The graph depicts the training and validation loss of the DNN model throughout its 50 epochs of training. The blue curve illustrates the training loss, which declines swiftly, signifying that the model is proficiently acquiring the fundamental patterns inside the dataset. The orange curve illustrates validation loss, which first declines but subsequently stabilizes after several epochs, indicating the model's capacity to generalize to novel data. The disparity between the two curves indicates possible overfitting, which is alleviated in this instance by consistent validation loss throughout the epochs.



```
1514/1514                                    2s 2ms/step
Mean Squared Error (MSE): 45.15422682304457
R² Score: 0.9940134660162391
```

Fig.24 - Results of DNN Model

Figure 25 illustrates the performance metrics of the DNN model on the test set. The Mean Squared Error (MSE) is 45.15, signifying the average squared deviation between predicted and actual values. An R^2 Score of 0.994 indicates that the model accounts for 99.4% of the variance in the target variable, reflecting great accuracy and exceptional performance.

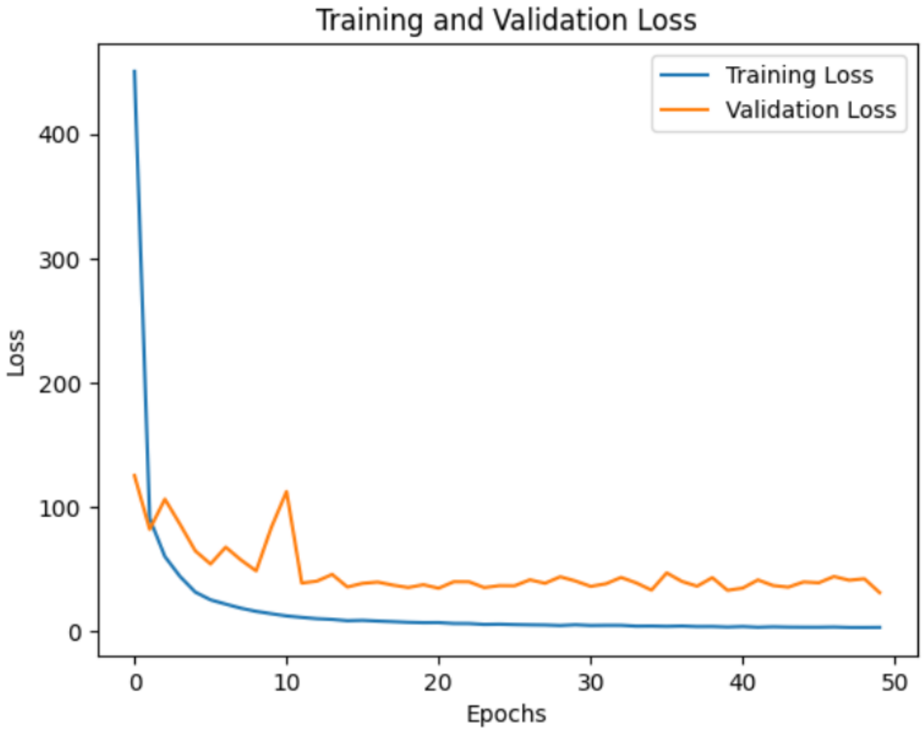


Fig .25-Graph representing the loss during trianing and validating

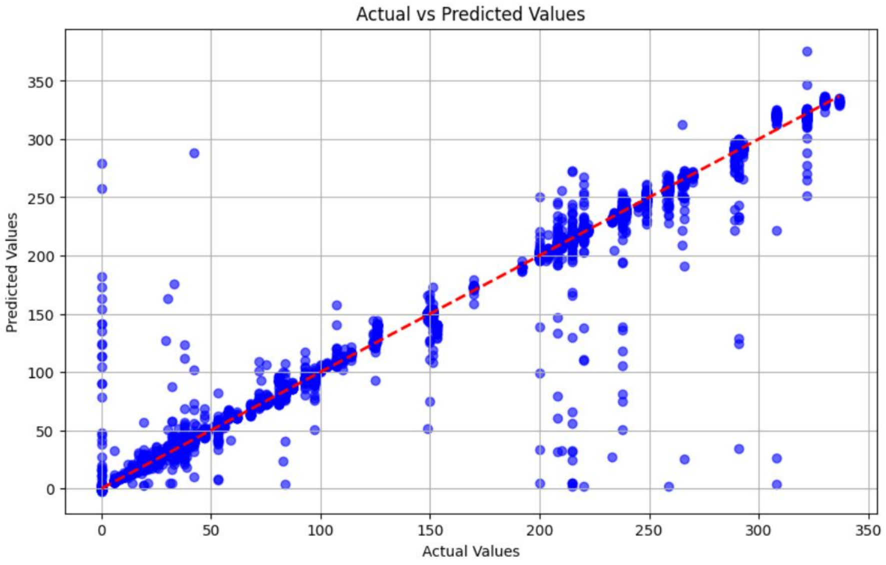


Fig. 26 -Actual vs Predicted values

9.3 Descriptive Analysis of Electric Vehicle Characteristics

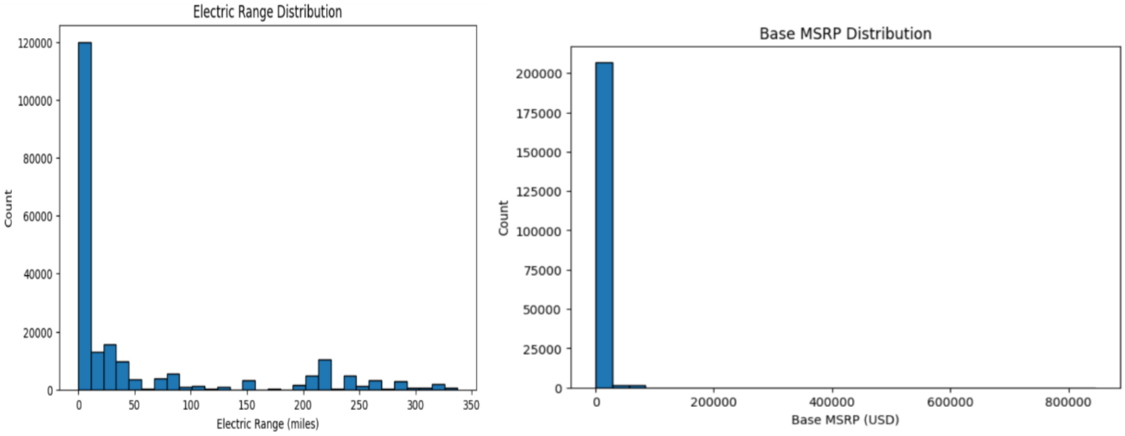
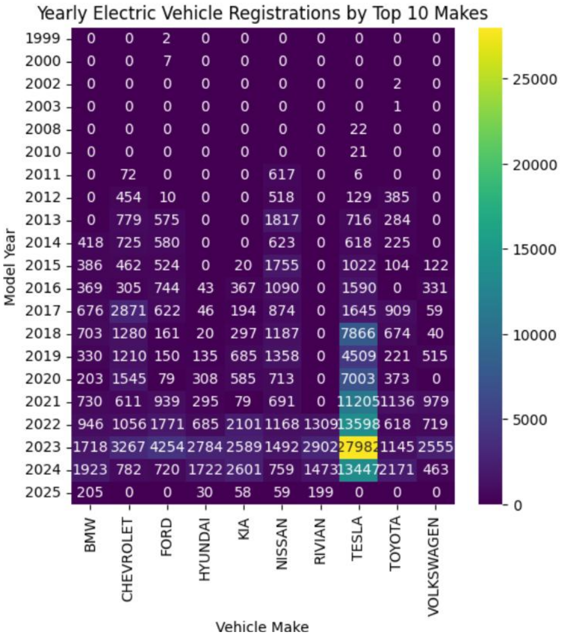


Fig.27 -Distribution of Electric Range and Base MSRP respectively.



	BMW	CHEVROLET	FORD	HYUNDAI	KIA	NISSAN	RIVIAN	TESLA	TOYOTA	VOLKSWAGEN
1999	0	0	2	0	0	0	0	0	0	0
2000	0	0	7	0	0	0	0	0	0	0
2002	0	0	0	0	0	0	0	0	2	0
2003	0	0	0	0	0	0	0	0	1	0
2008	0	0	0	0	0	0	0	22	0	0
2010	0	0	0	0	0	0	0	21	0	0
2011	0	72	0	0	0	617	0	6	0	0
2012	0	454	10	0	0	518	0	129	385	0
2013	0	779	575	0	0	1817	0	716	284	0
2014	418	725	580	0	0	623	0	618	225	0
2015	386	462	524	0	20	1755	0	1022	104	122
2016	369	305	744	43	367	1090	0	1590	0	331
2017	676	2871	622	46	194	874	0	1645	909	59
2018	703	1280	161	20	297	1187	0	7866	674	40
2019	330	1210	150	135	685	1358	0	4509	221	515
2020	203	1545	79	308	585	713	0	7003	373	0
2021	730	611	939	295	79	691	0	11205	1136	979
2022	946	1056	1771	685	2101	1168	1309	13598	618	719
2023	1718	3267	4254	2784	2589	1492	2902	27982	1145	2555
2024	1923	782	720	1722	2601	759	1473	13447	2171	463
2025	205	0	0	30	58	59	199	0	0	0

Fig. 28 - count (Top 10 makes) of EV registration.

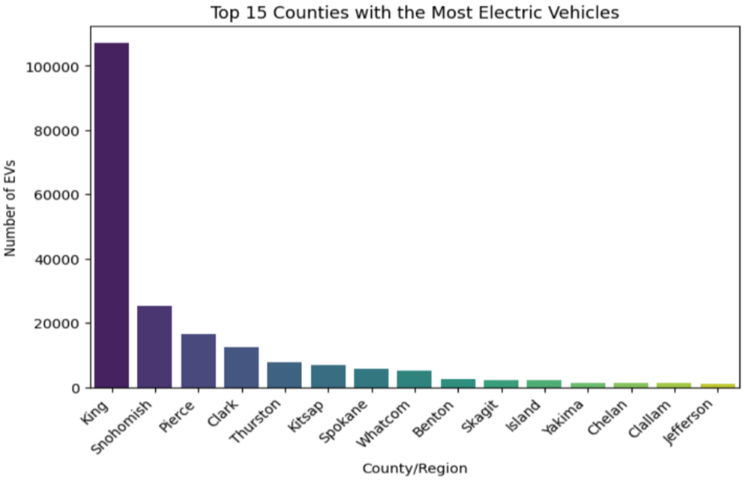


Fig. 29- County with Most EV's

Figure 26 illustrates the asymmetric distributions of Electric Range and Base MSRP, signifying the prevalence of vehicles with reduced ranges and lower price thresholds. These observations correspond with the affordability aspect affecting market adoption. Figure 27, the heatmap, illustrates annual electric car registrations by the leading 10 manufacturers, highlighting Tesla's preeminent market share, especially from 2021 onward. Figure 28 indicates that King County is at the forefront of electric car adoption, succeeded by Snohomish and Pierce, illustrating regional preferences and socio-economic influences on EV penetration. Collectively, these visualizations offer an extensive insight into trends in electric vehicle adoption and market distribution.

SECTION VII.

10. Project Management

Project management is the process of efficiently directing a team to accomplish a defined goal within a designated time period. This systematic method is essential for monitoring advancement and guaranteeing the punctual achievement of milestones. Project management includes several phases: initiation, planning, scheduling, monitoring, and execution. These stages facilitate a methodical process, particularly for initiatives such as the analysis of Electric Vehicle datasets, where precise data preparation and model evaluations are essential. This method enhances deliverable efficiency and ensures outcomes fit with established objectives.

10.1 Project Scheduling

An explicitly outlined timeline is essential for the success of any project. It establishes timelines for each task and guarantees the efficient use of resources. This project required scheduling tasks like data pretreatment, feature engineering, model training, and validation within a predetermined timetable. A Gantt chart (appendix 2), illustrating these phases, clearly monitored progress and emphasized dependencies among tasks. This guaranteed the alignment of all components, preventing delays. Compliance with these schedules was essential for ensuring consistency in the evaluation of machine learning models such as Random Forest and Deep Neural Networks. Tracking milestones using feedback loops promoted task execution.

10.2 Risk Management

Risk management is essential for recognizing and mitigating project challenges. This project was at risk from managing enormous datasets, which required a lot of processing time and computational power. Optimized data pretreatment and distributed computing frameworks reduced this danger. Data loss, health issues, and system breakdowns were also addressed. System crashes were managed via robust cloud storage and regular backups. Data loss was reduced by restricting access to authorized staff. These hazards were managed to protect sensitive Electric Vehicle data. By understanding and resolving these obstacles, the project-maintained progress and process without compromising deliverable quality**.**

Table 2: Risk Mitigation Plan Table

No	Risk	Impact	Risk Level	Mitigation Plan
1	Loss of Data	Data corruption may cause inaccuracies.	Medium	Regular backups and restricted access to data.
2	Resource Fatigue	Delayed tasks due to workload pressures.	Medium	Allocate buffer time in the schedule.
3	System Downtime	Temporary loss of coding or testing capacity.	Medium	Implement robust cloud backups for recovery.
4	Hardware Failure	Delays in accessing tools and data.	Medium	Store data securely on redundant systems.
5	Large File Processing Time	Prolonged runtime affecting overall timelines.	Medium	Optimize preprocessing and utilize scalable cloud infrastructure.

10.3 Quality Management

Quality management ensures products and processes meet or exceed stakeholder expectations through methodical methods. Quality management was crucial to preprocessing and analysing the electric car dataset to fulfil the highest predictive modelling and anomaly detection requirements in this project. Iterative adjustments were made based on supervisor feedback to match model results to expectations.

Identifying dataset errors, validating models, and fixing electric range or MSRP distribution anomalies were crucial. Regular evaluation meetings enabled adaptable modifications, assuring high-quality project outcomes. Figures 26–28 show how distribution visualizations and heatmaps verified data integrity and model performance to ensure quality. Continuous quality checks ensured prediction and insight accuracy, affecting reliability and relevance. This quality focus improved the model's performance and aligned with the project's goal of actionable, high-quality results.

10.4 Social, Legal, Ethical, and Professional Considerations

Social, legal, ethical, and professional factors are essential for the appropriate and legitimate implementation of machine learning or data science initiatives. Social factors guarantee that initiatives yield beneficial outcomes for society while avoiding detriment to any group. Legal considerations pertain to adherence to data protection regulations, including **GDPR (General**

Data Protection Regulation) and other pertinent laws, to ensure the protection of data privacy and usage rights. Ethical principles underscore the equitable, accountable, and transparent management of data to prevent bias, abuse, or violations of trust. Professionalism in data science requires upholding honesty, precision, and responsibility in every facet of the project.

In the framework of this study, the dataset was meticulously managed to address these considerations. The data utilized for this investigation complied with ethical standards, guaranteeing that no sensitive information was revealed. The analysis and modelling procedures adhered rigorously to legal and ethical norms to safeguard data integrity and avert illicit usage. The experiment complied with the principle of informed consent, guaranteeing that the information was utilized solely for research purposes.

To provide transparency, an ethical approval document outlining the compliance steps implemented has been incorporated in Appendix 2. These factors underscore the project's dedication to ethical and legal standards, promoting trust and responsibility.

11. CRITICAL EVALUATION

This section examines the project's strengths and weaknesses, providing a comprehensive view of its implementation and outcomes. The study's notable benefit is in its extensive application of modern data analysis methods to comprehend Electric Vehicle (EV) performance. The dataset, however varied, requires considerable preprocessing and feature engineering to improve its usefulness. A significant problem encountered during the research was the disparity in feature importance, which was addressed by methods such as correlation analysis and feature reduction. This guaranteed that the machine learning models employed solely the most significant variables for predictions.

The analysis demonstrated that Random Forest and Gradient Boosting models were exceptionally effective, achieving nearly flawless performance metrics. Nonetheless, this elicited apprehensions over possible overfitting, which were meticulously mitigated through cross-validation and evaluation on novel data. Moreover, the Deep Neural Network (DNN) model demonstrated competitive outcomes, highlighting its potential for scalability and adaptability to more extensive datasets.

A drawback noted is the dataset's intrinsic bias, as specific variables such as MSRP and Electric Range predominate the predictions. Future research should aim to gather a wider array of features to enhance the generalizability of the models. This assessment underscores the necessity for ongoing enhancement, advocating for further experiments utilizing varied datasets and hybrid modelling techniques to improve predicted precision and relevance in practical applications.

12. CONCLUSION

This project developed a machine learning framework to predict electric range, identify performance-affecting features, and address classification challenges like CAFV eligibility for Electric Vehicle (EV) data. Advanced algorithms, including Random Forest, Gradient Boosting, and Deep Neural Networks, produced remarkably accurate forecasts, with feature relevance evaluations highlighting critical qualities such as MSRP and Electric Range.

Stringent preprocessing measures guaranteed the models' resilience and reliability, while cross-validation and testing validated their capacity for effective generalization.

The research not only illustrated the significance of machine learning in electric vehicle analytics but also emphasized its practical implications. Range forecasts can inform the strategic positioning of charging stations and infrastructure development, whilst categorization results can aid governments and manufacturers in directing incentives for CAFV adoption. Despite concerns regarding overfitting, the study developed a scalable methodology for prospective improvements, like the incorporation of pricing and location data or the investigation of hybrid models to increase forecasts.

This project effectively met its aims, establishing a robust platform for future research in EV analytics. It provides significant insights into sustainable mobility, aiding the overarching objective of enhancing electric vehicle adoption and streamlining the transition to clean energy.

13. ACHIEVEMENTS

This project aims to investigate the capabilities of machine learning models in evaluating Electric Vehicle (EV) data, with a specific emphasis on electric range prediction and categorization tasks. By employing complex techniques including Random Forest, Gradient Boosting, and Deep Neural Networks, I attained exceptionally precise outcomes, with R^2 scores approaching perfection. This illustrates the models' capacity to effectively capture both linear and non-linear patterns in the data.

A notable success of this study was the identification of the main features, including MSRP and Electric Range, which significantly impacted the model's projections. Through the assessment of feature significance and the elimination of low-correlation variables, I improved model interpretability and ensured computational economy while maintaining performance. This procedure enabled me to enhance the dataset for improved predictive results.

A significant achievement was rectifying class imbalance in the dataset and guaranteeing that the models exhibited strong generalization to novel data via rigorous validation methods. I also examined the impact of hyperparameter tuning and feature engineering to enhance the models further.

This study has illustrated the practical utility of machine learning in electric vehicle analytics and established a foundation for subsequent research. I am assured that these findings significantly enhance sustainable transportation and data-informed decision-making within the electric vehicle sector.

14. FUTURE WORK

This project can be expanded in future endeavours to yield more profound insights and practical applications for the adoption of electric vehicles (EVs) and the development of infrastructure. Integrating additional statistics, such the real-time positioning of charging stations, the frequency of electric vehicle charging across various locations, and historical charging trends, could substantially improve the analysis. Access to data regarding vehicle costs, user demographics, and income levels would facilitate a more robust correlation between affordability and electric vehicle adoption trends.

By including more detailed geographical data, such as traffic density or nearness to urban regions, the project might forecast high-demand zones for the installation of new charging stations. Additionally, integrating meteorological data may enable an examination of how seasonal or climatic fluctuations influence electric range and charging patterns.

Ultimately, the implementation of deep learning models such as Recurrent Neural Networks (RNNs) may facilitate the identification of temporal trends, hence enhancing predictive analytics for electric vehicle infrastructure requirements. This research, through these additions, could enhance forecast accuracy and significantly influence public policy and corporate investment in sustainable transportation.

15. STUDENT REFLEECTION (5R FRAMEWORK)

Reporting

The study involved the application of deep learning algorithms to examine electric car data. The primary objectives were to forecast electric range, categorize vehicle kinds, and detect anomalies. During the project, I performed data preprocessing, feature selection, and the implementation of sophisticated machine learning algorithms.

Responding

At first, I was worried by the magnitude of the dataset and the computational demands. The project required continual learning and adjustment. Through continuous feedback from my supervisor and incremental improvements, I developed confidence and enjoyment in attaining high model accuracy and actionable insights.

Relating

This initiative connected academic understanding with practical application. Practical application of concepts from machine learning and deep learning courses enhanced my comprehension. Working with my supervisor resembled a professional environment, providing insights into the use of data science in practice.

Reasoning

The obstacles encountered, including managing imbalanced datasets and computing delays, necessitated analytical reasoning and innovative problem-solving. I recognized the significance of job prioritization, selecting suitable models, and deriving relevant interpretations from outcomes. This rationale enhanced my capacity to address analogous real-world challenges in the future.

Reconstruction

This endeavour enabled me to reevaluate my methods of learning and problem-solving. I recognized the significance of patience and incremental enhancement. The acquired abilities will be helpful in my work as a data scientist. I am now better prepared to utilize modern analytics to address intricate difficulties, guaranteeing significant outcomes.

DATASET AND LINK FOR CODE:

https://catalog.data.gov/dataset/electric-vehicle-population-data - Data set link

https://drive.google.com/drive/folders/1-6Fx7lVYcEzkhh9kgas4sG9uODs_4mRm

16. Bibliography

Airlangga, G. (2024). Comparative Analysis of Machine Learning Models for Predicting Electric Vehicle Range. pp. Vol 9, No 1.

Canal, R. R. (2024). Machine learning for real-time fuel consumption prediction and driving profile classification based on ECU data. IEEE Access. DOI: 10.1109/ACCESS.2024.3400933.

Carlsson, H. &. (2024). Weather Impact on Energy Consumption for Electric Trucks. *Weather Impact on Energy Consumption for Electric Trucks. Master's Thesis, Chalmers University of Technology, Sweden.*, p. 49.

Chen, T. &. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.*

Dengfeng Zhao, H. L. (2023). A Review of the Data-Driven Prediction Method of Vehicle Fuel Consumption. vol 16 ,issue14.

Goodfellow, I. B. (2016). *Deep Learning.* https://www.deeplearningbook.org/.

Hofmockel, J. S. (2018). Isolation Forest for Anomaly Detection in Raw Vehicle Sensor Data.

Iranzad, R. &. (2024). A review of random forest-based feature selection methods for data science education and applications. *International Journal of Data Science and Analytics.*

Javier Bas, C. C. (2021). Classification of potential electric vehicle purchasers: A machine learning approach. *Technological Forecasting and Social Change*, Volume 168.

Kotsiantis, S. B. (2006). Data Preprocessing for Supervised Learning. *International Journal of Computer Science*, 1(2), 111–117.

Lei Tong, K. L. (2021). Feature Analyses and Modeling of Lithium-Ion Battery Manufacturing Based on Random Forest Classification. *Transaction on Mechatronnics*, Volume: 26 Issue: 6.

Liu, F. T., Ting, K. M., & Zhou, Z.-H. (10 February 2009). Isolation Forest. *2008 Eighth IEEE International Conference on Data Mining.* IEEE.

Marini, F. M. (2007). Multilayer feed-forward artificial neural networks for class modeling. . *Chemometrics and Intelligent Laboratory Systems*, Vol 88, issue 1.

Modi, S. B. (2020). Estimation of energy consumption of electric vehicles using deep convolutional neural network to reduce driver's range anxiety. ISA transactions, 98, 454-470.

Polyzos, E. &. (2023). Autoregressive Random Forests: Machine Learning and Lag Selection for Financial Research. *Computational Economics,*, 64, 225–262.

Straka, M. D. (2020). Predicting popularity of electric vehicle charging infrastructure in urban context. IEEE Access, 8, 11315-11327.

Surabhi, S. N. (2024). Range Prediction based on Battery Degradation and Vehicle Mileage for Battery Electric Vehicles. International Journal of Science and Research, 13, 952-958.

Teresa Pamuła, D. P. (2022). Prediction of Electric Buses Energy Consumption from Trip Parameters Using Deep Learning. *Energies* , 15(5).

https://reflection.ed.ac.uk/reflectors-toolkit/reflecting-on-experience/5r-framework#:~:text=The%205R%20framework%20for%20reflection%20will%20guide%20you%20through%20Reporting,sense%20of%20a%20learning%20experience.

https://www.tensorflow.org/guide/keras/sequential_model

https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html

https://matplotlib.org/stable/gallery/index.html

https://towardsdatascience.com/exploring-feature-importance

https://scikit-learn.org/stable/modules/model_evaluation.html

https://stackoverflow.com/questions/61503183/how-can-i-add-grid-lines-to-a-catplot-in-seaborn

Aula - Lecture 9- Class Imbalance, Missing Data, Ensembles

https://coventry.aula.education/#/dashboard/10488088-4949-4f7e-a9f5-cf3b75ac259b/journey/materials/fe09a019-8650-4257-b0bd-5eb9222fdadc

Aula - Week 7 - Data Visualisation

https://jakevdp.github.io/PythonDataScienceHandbook/-

https://www.w3schools.com/python/python_ref_string.asp

https://github.com/WillKoehrsen/prediction-documentation/blob/master/Deep%20Neural%20Network.ipynb

https://www.geeksforgeeks.org/feedforward-neural-network/

https://stackoverflow.com/questions/40758562/can-anyone-explain-me-standardscaler

https://www.datacamp.com/tutorial/isolation-forest

https://gist.github.com/pb111/cc341409081dffa5e9eaf60d79562a03

https://ieeexplore.ieee.org/abstract/document/4781136

https://ieeexplore.ieee.org/document/9314252/citations

https://pdfs.semanticscholar.org/251e/87bcd1c12ce4e54d3facdd8c34fe98ac1ae8.pdf

https://www.sciencedirect.com/science/article/pii/S0040162521001918

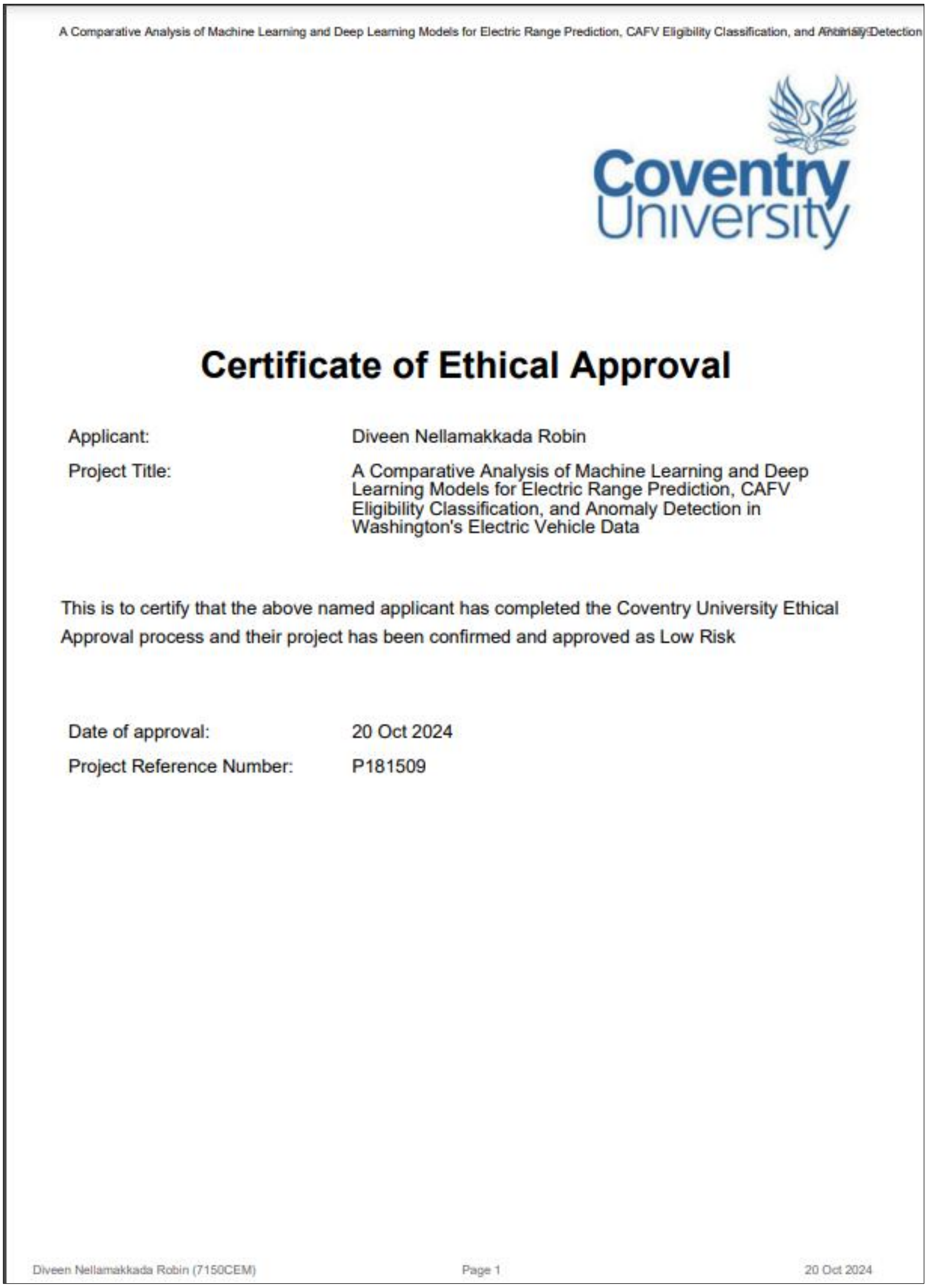
https://www.mdpi.com/1996-1073/16/14/5258

https://ieeexplore.ieee.org/abstract/document/10530543

https://www.diva-portal.org/smash/record.jsf?pid=diva2%3A1874010&dswid=-2520

https://www.researchgate.net/publication/220715764_Data_Preprocessing_for_Supervised_Learning

APPENDIX -1



A Comparative Analysis of Machine Learning and Deep Learning Models for Electric Range Prediction, CAFV Eligibility Classification, and Anomaly Detection

Coventry University

Certificate of Ethical Approval

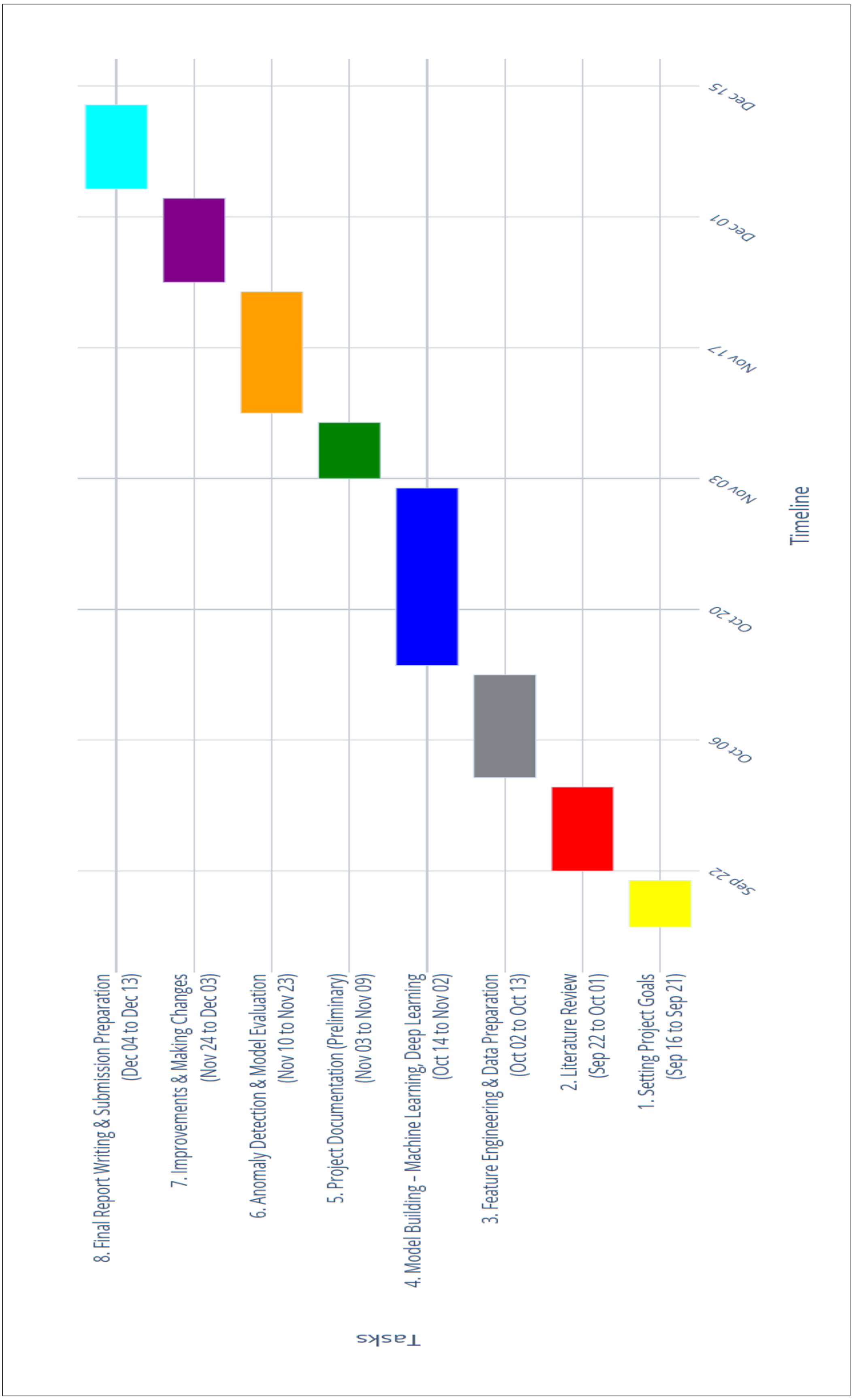
Applicant:	Diveen Nellamakkada Robin
Project Title:	A Comparative Analysis of Machine Learning and Deep Learning Models for Electric Range Prediction, CAFV Eligibility Classification, and Anomaly Detection in Washington's Electric Vehicle Data

This is to certify that the above named applicant has completed the Coventry University Ethical Approval process and their project has been confirmed and approved as Low Risk

Date of approval:	20 Oct 2024
Project Reference Number:	P181509

Diveen Nellamakkada Robin (7150CEM) Page 1 20 Oct 2024

APPENDIX – 2 (Gantt Chart)


Tasks
8. Final Report Writing & Submission Preparation (Dec 04 to Dec 13)
7. Improvements & Making Changes (Nov 24 to Dec 03)
6. Anomaly Detection & Model Evaluation (Nov 10 to Nov 23)
5. Project Documentation (Preliminary) (Nov 03 to Nov 09)
4. Model Building - Machine Learning, Deep Learning (Oct 14 to Nov 02)
3. Feature Engineering & Data Preparation (Oct 02 to Oct 13)
2. Literature Review (Sep 22 to Oct 01)
1. Setting Project Goals (Sep 16 to Sep 21)
Sep 22
Oct 06
Oct 20
Nov 03
Nov 17
Dec 01
Dec 15
Timeline