**WATT WHEELS-SURGE IN ELECTRIC VECHICLES**

**Abstract and Highlights**

The transition to electric vehicles (EVs) is a pivotal aspect of achieving sustainable and environmentally friendly transportation. However, this paradigm shift presents multifaceted challenges that need thorough examination and resolution. The project's goal is to give significant insights into the complicated terrain of electric car adoption, allowing stakeholders to make educated decisions and contribute to the development of sustainable transportation solutions.

For the project, we have developed and deployed various models for both exploratory and predictive purposes. We have used SAS, Tableau, and excel to execute and develop our models. We have done cluster analysis, using ward clustering, centroid and average clustering, this was done in SAS for exploratory purposes. For the clusters both automatic and user-specified methods were used. For performance evaluation, we have evaluated the stability of the clusters by checking the number of clusters before and after sampling. For the predictive part, we have developed and executed three models in the SAS, them being, decision tree, regression, and neural network. We have developed multiple versions for each of the above-mentioned supervised data mining techniques to get the best model possible. We then compared the best model from the three supervised datamining techniques to arrive at the best model. We used the fit statistics to evaluate the performance of the above models. We have scored the best model on our score data set to find the predictive power of our model. Some of the recommendations from the analysis of data we have done is to focus more on the data mining approaches to completely analyse and solve the complications that influence the adoption of electric cars (EVs), with a special emphasis on Battery Electric cars (BEVs) and Plug-in Hybrid Electric Vehicles (PHEVs). The project team, "Data Detectives," conducted in-depth research utilising the Electric Vehicle Population dataset collected from Kaggle, with a focus on the Washington State region.

The predicted results will shed light on the variables impacting EV adoption, uncover trends in EV distribution across different areas, assess the influence of government regulations, and give insights into anticipating charging station demands. This study is critical for promoting sustainable transportation options, assisting stakeholders in making informed decisions, and advancing the development of electric mobility.

**Problem Description**   
  
**Description and some background about the focal problem addressed in the project:**

This project is to analyse the electric vehicles using the data set that is sourced from Kaggle. As, the world shifts towards cleaner and more sustainable mobility solutions by understanding this adapting the electric vehicles is becoming important. This project centres on leveraging data mining techniques to analyse the Electric Vehicle Population dataset, focusing on Washington State, to unearth patterns, trends, and factors affecting the growth and sustainability of EV usage.

**Why it is an important problem to be addressed.**

The significance of addressing this problem lies in the transformative potential of electric vehicles in mitigating environmental impacts, reducing dependency on fossil fuels, and promoting sustainable transportation. As governments, industries, and consumers increasingly embrace EVs, informed decision-making is essential for effective policy formulation, charging infrastructure planning, and overall market growth. Identifying disparities in adoption, assessing the impact of government policies, and predicting electric vehicle range contribute to creating a robust foundation for a sustainable and scalable electric mobility ecosystem (Singhal, n.d.).

**Well-Articulated Questions:**

The predictive analytic question driving this project is based on factors such as the existing population of electric vehicles, population density, and economic indicators.

1. **Predictive Analysis Question:**

How accurately can we anticipate the need for charging stations in specific geographic areas, considering criteria such as the existing population of electric vehicles and economic indicators?

1. **Exploratory Analysis Question:**

What is the correlation between the rate of electric vehicle adoption in urban versus and the availability of public charging infrastructure, and how has this relationship evolved over time?

**Data Exploration, Preparation, and Visualization**

**Description of Data Set and Data Source**

In this project we used dataset about electric vehicle population in Washington state. The dataset displays registered Battery Electric Vehicles (BEVs) and Plug-in Hybrid Electric Vehicles (PHEVs) recorded by the Washington State Department of Licensing (DOL) from year 2020 to 2024. In the United States, BEV sales experienced an impressive nearly 57% year-over-year increase, marking the highest growth among the top three electric vehicle markets. The dataset is obtained from Kaggle. The link of the data is, (Singhal, n.d.).

<https://www.kaggle.com/datasets/yashusinghal/electric-vehicle-population-dataset?resource=download>

**Analytical Methodology Overview**

The project plan includes a structured timetable with milestones for project commencement, data exploration and selection, variable optimisation, model creation, performance assessment, model comparison, finalisation, and results distribution. SAS Enterprise Miner will be used to carry out both exploratory and predictive analytics, while Tableau or Excel will help with data preparation and visualisation.

Before embarking on data preparation, we conducted thorough data cleaning in Excel to enhance the accuracy and reliability of our dataset. This involved addressing missing values, eliminating duplicates, and rectifying any inconsistencies. We also streamlined the dataset by reducing the number of variables. Once this initial cleaning phase was complete, we converted the Excel file into a SAS file for exploratory modelling. We created a copy of the SAS dataset, which was then divided into two separate datasets using SAS for both scoring and modelling purposes in our predictive analysis.

We load the original dataset into Tableau for visualization. After the visualization is done, we got the visualizations to understand the data into useful insights. The analysis is performed in SAS enterprise miner. The Exploratory analysis phase of our research, we applied significant clustering methodologies, including ward clustering, average clustering, and centroid clustering. These sophisticated techniques played a pivotal role in unveiling intricate patterns within the dataset. By grouping similar data points together, these clustering methods contributed valuable insights into the inherent structure of our data. This analytical approach not only facilitated a comprehensive understanding of the dataset but also laid the foundation for further in-depth exploration and interpretation of meaningful clusters in our research. To enhance the precision and robustness of our clustering outcomes, we conducted multiple iterations of the clustering process using a 0% sample of the data. The repetition of clustering techniques on a substantial subset aimed to reinforce the reliability of our results and ensure greater accuracy in identifying underlying patterns within the dataset. This deliberate approach in refining the clustering methodology contributes to the overall improvement of the quality of our research findings and strengthens the validity of our clustering outcomes.

Predictive analysis is carried out, we utilized regression, decision trees, and neural networks in our analysis to predict outcomes and unveil correlations among variables within the dataset. The application of these methodologies allowed us to create a detailed score report, facilitating in-depth analysis and informed decision-making by comparing various models.

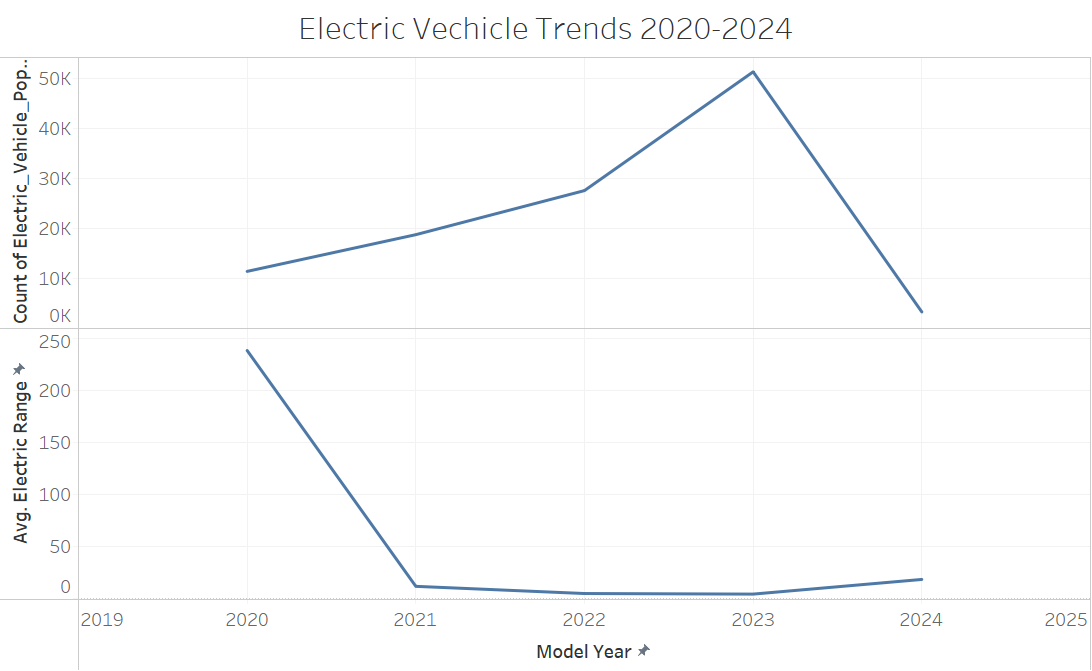
Our project is designed to offer a comprehensive understanding of the dataset by emphasizing both exploratory and predictive analyses. The synergy between these approaches enables the discovery of insights, identification of trends, and accurate predictions within the gaming industry. The project showcases a meticulous and systematic approach to data analysis, incorporating processes such as data cleaning, visualization, clustering, and predictive modelling. This methodological rigor ensures that the insights and conclusions derived from the analysis are founded on robust methodologies and trustworthy data, enhancing the overall credibility of our findings.

**Cleaning and modifying the Data Set:**

Initially, the unprocessed dataset contained multiple variables such as VIN (1-10), County, City, State, Postal Code, Model Year, Make, Model, Electric Vehicle Type, Clean Alternative Fuel Vehicle (CAFV) Eligibility, Electric Range, Base MSRP, Legislative District, DOL Vehicle ID, Vehicle Location, Electric Utility, and 2020 Census Tract. We selectively narrowed down our focus to essential variables for effective analysis. We specifically considered City, State, Make, Model, Electric Vehicle Type, CAFV Eligibility, Legislative District, Electric Utility, are categorical variables. with the objective variable being 'Electric Range,' which represents the fuel tank capacity that determines an EV's total range on a single charge. Predictor factors include manufacturer, model, year of manufacturing, country of origin, vehicle type, and range. Model Year and Electric Range were designated as numerical variables, and Vehicle Location is a nominal variable. This deliberate selection aimed to optimize our analytical results and provide accurate answers to our research questions. As data often does not come in a ready to use format, often data miners need to change the format of data, remove duplicates as it can often skew the results because of duplicates. By doing this step we have ensure that the data is accurate, non-skewed and ready to be deployed.

**Exploring the Data set by Visualization**

In the line graph in Fig 1.1 depicts the trends of average of electric range and count of Electric Vehicle Population for model year. The view is filtered on Model year which ranges from 2020 to 2024.In count of Electric Vehicle Population 2023 it peaks at 50k in the year 2023 followed by next surge in between 20k to 30k in 2022.The least was seen in 2024.The count of Electric Vehicle Population was 10k in year 2020.In average electric range graph at peak at 250 in the year 2020, it has been almost constant from the year 2021 to 2024.

 Fig.1.1 Electric vehicle Trends 2020-2024

In the horizontal bars in Fig 1.2 shows the average of electric range for clean alternative fuel vehicle (CAFV) Eligibility broken down by electric vehicle. Colour shows details about Clean alternative fuel vehicle (CAFV) Eligibility. The average electric range is peaked at almost 200 in battery electric vehicle at Clean alternative fuel vehicle (CAFV). Not eligible alternative fuel was between 20 and 40 for the same battery electric vehicle. For plug in hybrid graph the Clean alternative fuel vehicle (CAFV) has no major difference with Not eligible alternative fuel as in battery electric vehicle. The average range is CAFV 40 and not eligible alternative fuel is at 20.

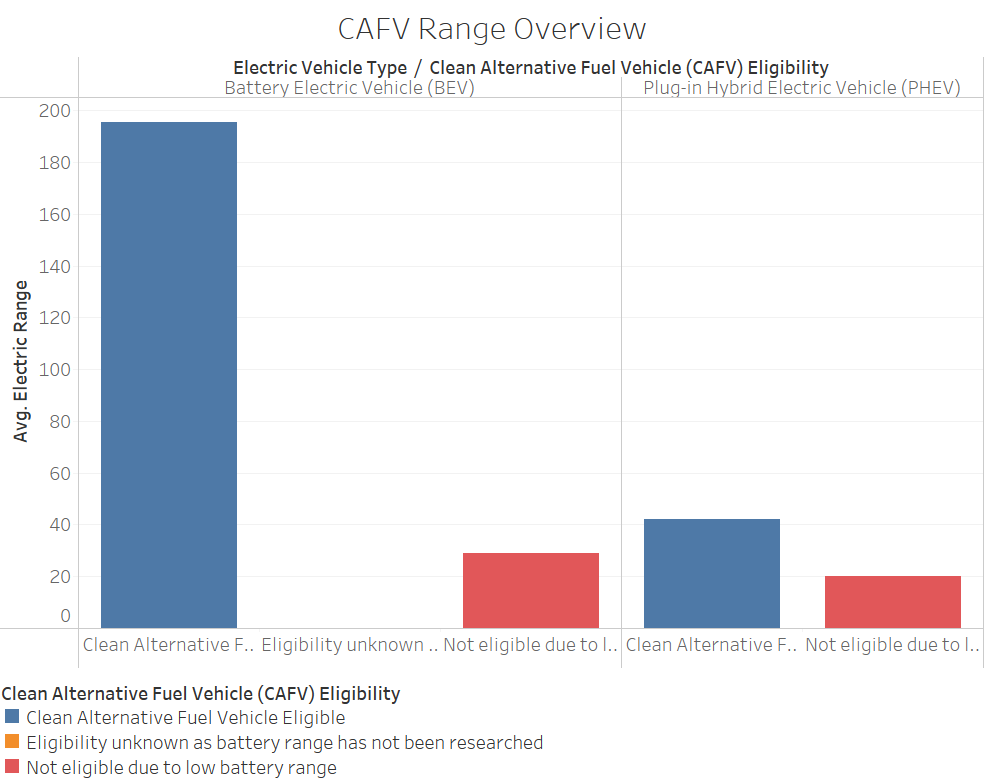


Fig.1.2 CAFV Range Overview

In Packed bubbles in Fig 1.3,model and model year.The colour shows details about electric vehicle type,size shows average of electric range.The marks are labelled by model and model year.The view is filtered at model year which ranges from 2020 to 2024.On overall, we can observe that electric range is higher for BEV than PHEV.The Significant majority of models choosen by customers are Model S 2020, Model Y, Model X, Model 3.

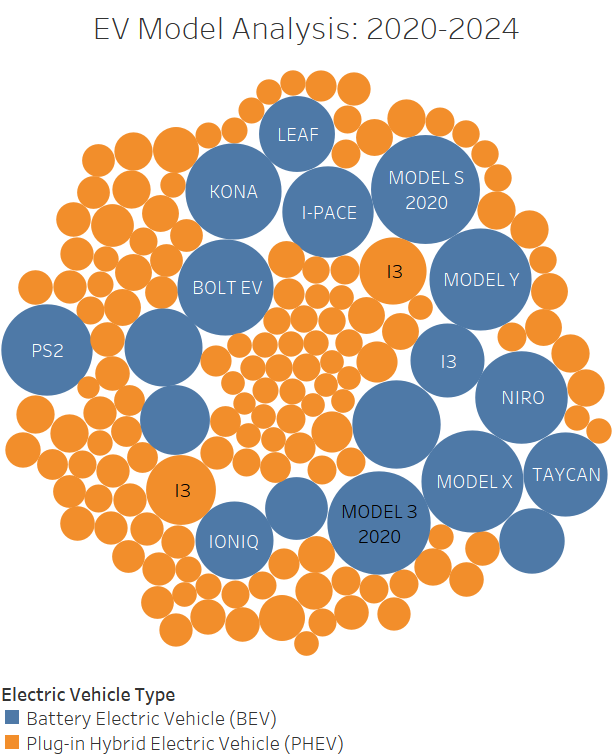


Fig.1.3 EV Model Analysis 2020-2024

In this analysis we can see that the darker the colour, more the average electric range in Fig 1.3. In that particular year and also decides whether the vehicles is eligible for Clean Alternative Fuel Vehicle (CAFV). It also represents Electric Vehicle Type (BEV) OR (PHEV) in their respective areas. From this we conclude that, in the year 2020 BEV has highest avg electrical range and least is in 2021 and the vehicles type is plug-in Hybrid Electric Vehicle.

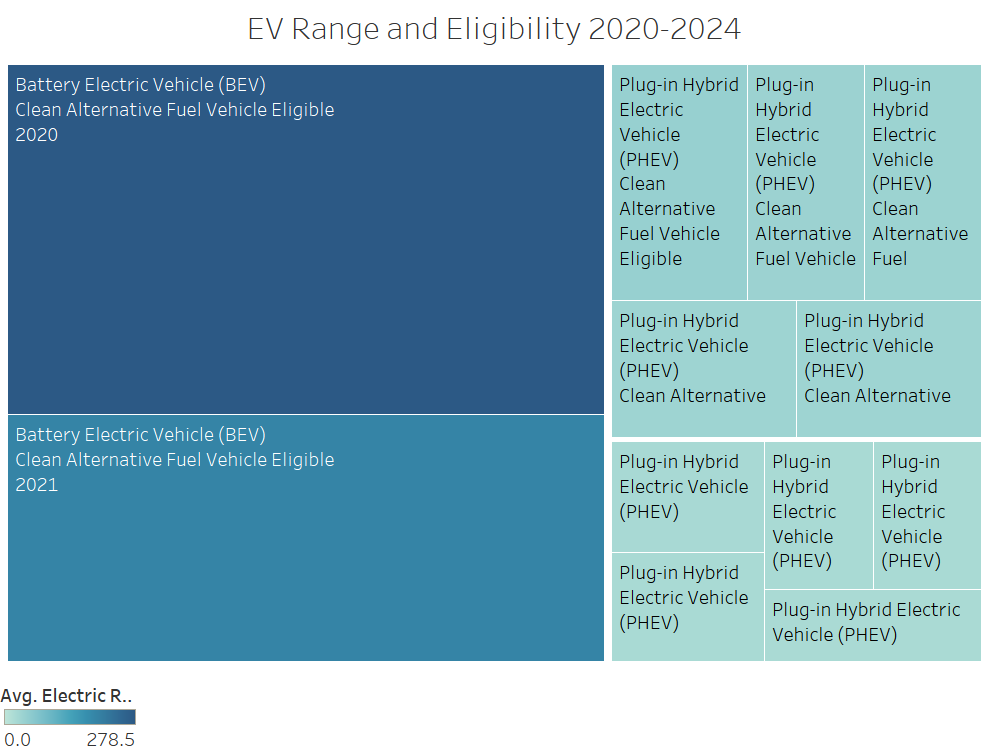


Fig.1.4 EV Range and Eligibilty 2020-2024

**Exploratory Analysis:**

In exploratory analysis we explore the Requirement of charging stations according to the electric vehicle adoption over the time.

Exploratory analysis is an unsupervised data mining technique which utilizes clustering methods to find patterns and group the data. In unsupervised data mining technique, we selected target variable as Electric Range of the vehicle. we rejected the city and state and other two variables as rejected as they are unique variables and those are categorical. We explore the need for charging stations in relation to the uptake of electric vehicles over time through exploratory analysis.

A computer screen shot of a diagram

Description automatically generated

Fig.2.Exploratory Analysis Diagram

In clustering there are two types, K-Means clustering and hierarchical clustering. Hierarchical clustering is unsupervised in that the number of clusters are created automatically. In hierarchical clustering we have got 5 clusters for Ward clustering, 4 clusters for centroid clustering and 4 for average clustering.

A screenshot of a computer

Description automatically generated

Fig.2.2. Ward Clustering

A screenshot of a computer

Description automatically generated

Fig.2.3. 85% Ward Clustering

A screenshot of a computer

Description automatically generated

Fig.2.4. Centroid Clustering

A screenshot of a computer

Description automatically generated

Fig.2.5. 85%Centroid Clustering

A screenshot of a computer

Description automatically generated

Fig.2.6. Average Clustering

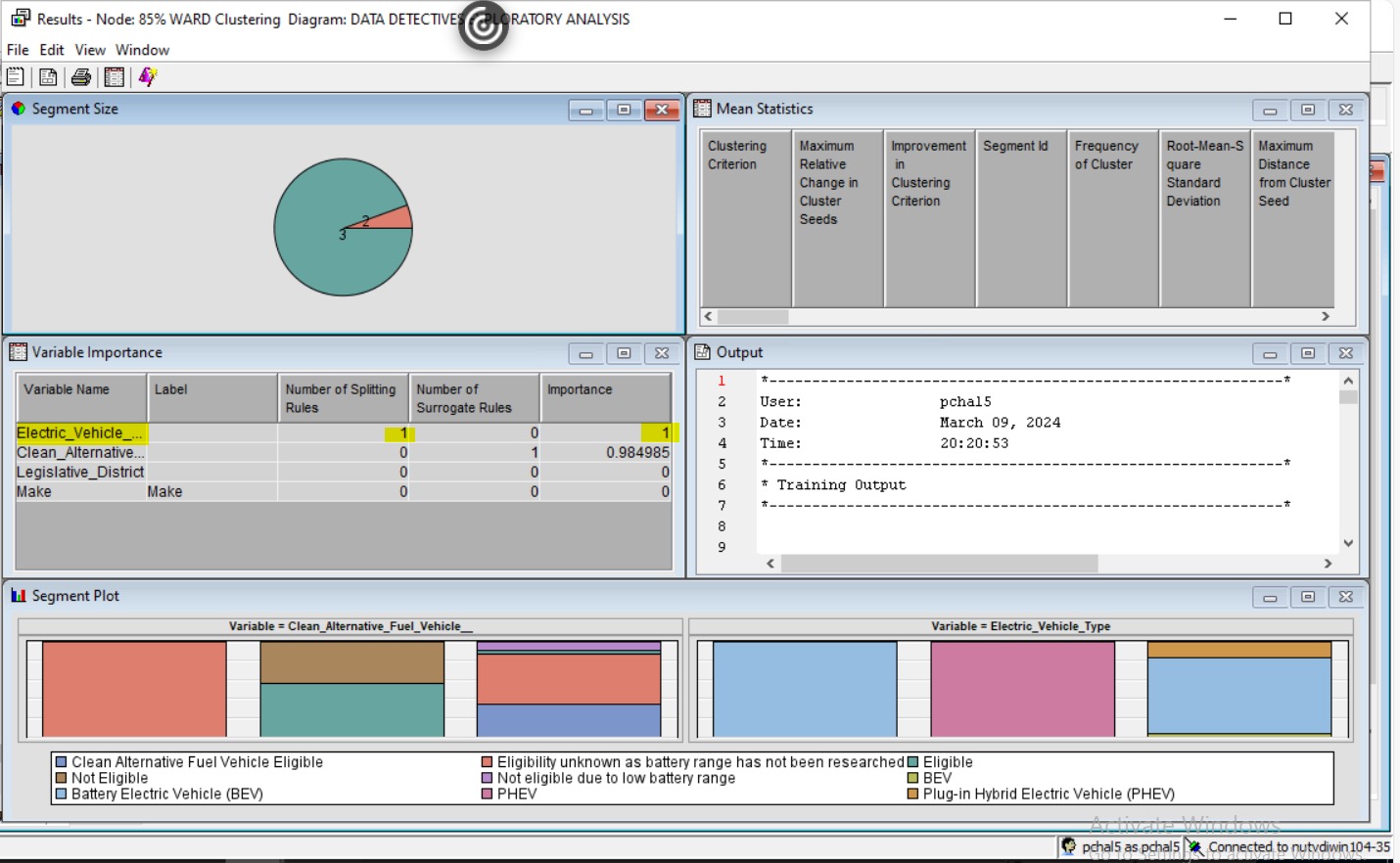
A screenshot of a computer

Description automatically generated

Fig.2.7. 85% Centroid Clustering

For the 85% sample, we have got 3 clusters for 85% ward clustering, 3 for 85% centroid clustering and 3 for 85% average clustering. We can infer from the above that the number of clusters before and after sampling for all clustering algorithms (ward, centroid and average clustering) are within +/- 2 range. But the number of clusters are not greater than 6 this makes it easier to interpret. We did the K-means clustering for Ward, centroid and average clustering. (K. P. Sinaga and M.S. Yang, 2020). From the results of 85%ward clustering window. Going to the view tab, then cluster profile and variable importance we get the above result. It shows that the most important variable is Electric Vehicle and the least important is legislative district and make which we can see in the results window in above figure.

In all the clusters we can see the results as that only the Segment Id 2 is having the greater number of records 166792 in Average and Centroid clusters and the ward cluster is having the number of records as 166787 in segment Id 2, After the 85% sample the number of records changes for segment 2 and segment id 3 but same for all the three clusters 85% ward, Average and centroid clusters. Segment id 2- 7504 and Segment Id 3-134274 records

 Fig.2.1. Variable Importance

In our project we have used all the major clustering algorithms: Ward clustering, Centroid Clustering and Average clustering. Based on this exploratory analysis, focusing on the Clean Alternative Fuel type and EV type most of the vehicles would fall into the cluster 2 and cluster 3. this tells us the new vehicles will be produced with these similarities comes under these two clusters. based on this data we can decide how many fuel type source stations would be needed in Washington state.

**Predictive Analysis:**

In the phase dedicated to predictive modelling, we implemented and applied the three primary supervised data mining techniques: decision trees, logistic regression, and neural networks. This process involved utilizing a dataset that had been divided into distinct segments for modelling and scoring purposes.

From our predictive analysis, the variable’s role that are shown below. The variable Electric range is taken as a **target variable**, when considering electric vehicles, electric range typically relates to the miles or kilometres the car can drive on a single charge. Since we have chosen Electric Range as our target, it can have any value within a specified range because it is a numerical outcome and a continuous variable. **Regression prediction** is the methodology we have selected to precisely evaluate and forecast electric range in considering these criteria. Using the assistance of this method, we can effectively arrange the data and make it easier to understand the connections between various components and the electric range.

A computer screen shot of a diagram

Description automatically generated

Fig.3.1. Predictive Analysis Diagram

For decision trees, we employed models with different configurations: breadth 2, depth 6; breadth 2, depth 4; breadth 2, depth 2; and breadth 3, depth 6. Through the model comparison node, we determined that the most optimal model is the one with breadth 3 and depth 6 (B3D6). To evaluate the effectiveness of this model, we examined its accuracy and misclassification rate. These metrics provide valuable insights into the model's performance and its ability to accurately classify instances within the dataset.

A screenshot of a computer

Description automatically generated

Fig.3.2. RegTree B3D6 Results

We used a variety of regression approaches, such as stepwise regression, forward regression, backward regression, and exhaustive regression, for the Multiple Linear Regression model. Following an extensive review, the Exhaustive Regression model was found to be the best model. The average squared error and root average squared error were used to evaluate the models' performance. By evaluating the model with the lowest error in both metrics, the best model was identified.

A screenshot of a computer

Description automatically generated

Fig.3.3. Exhaustive Regression Results

We selected to use the back-propagation approach and a neural network model with three hidden units for our prediction. The neural network's productive management of irregular patterns and complex data connections served as the base for this choice. The weights of the network could be continuously adjusted due to the back-propagation technique, which reduced the error between expected and actual results. By utilizing this combination, we hope to improve prediction performance across a variety of datasets by finding a balance between model complexity and generalization ability. The neural network architecture was ultimately chosen based on its ability to adapt to different issue domains and yet achieve strong predicted accuracy in our analysis tasks. For Neural networks, the best performance model resultedNeural Network 3HU.

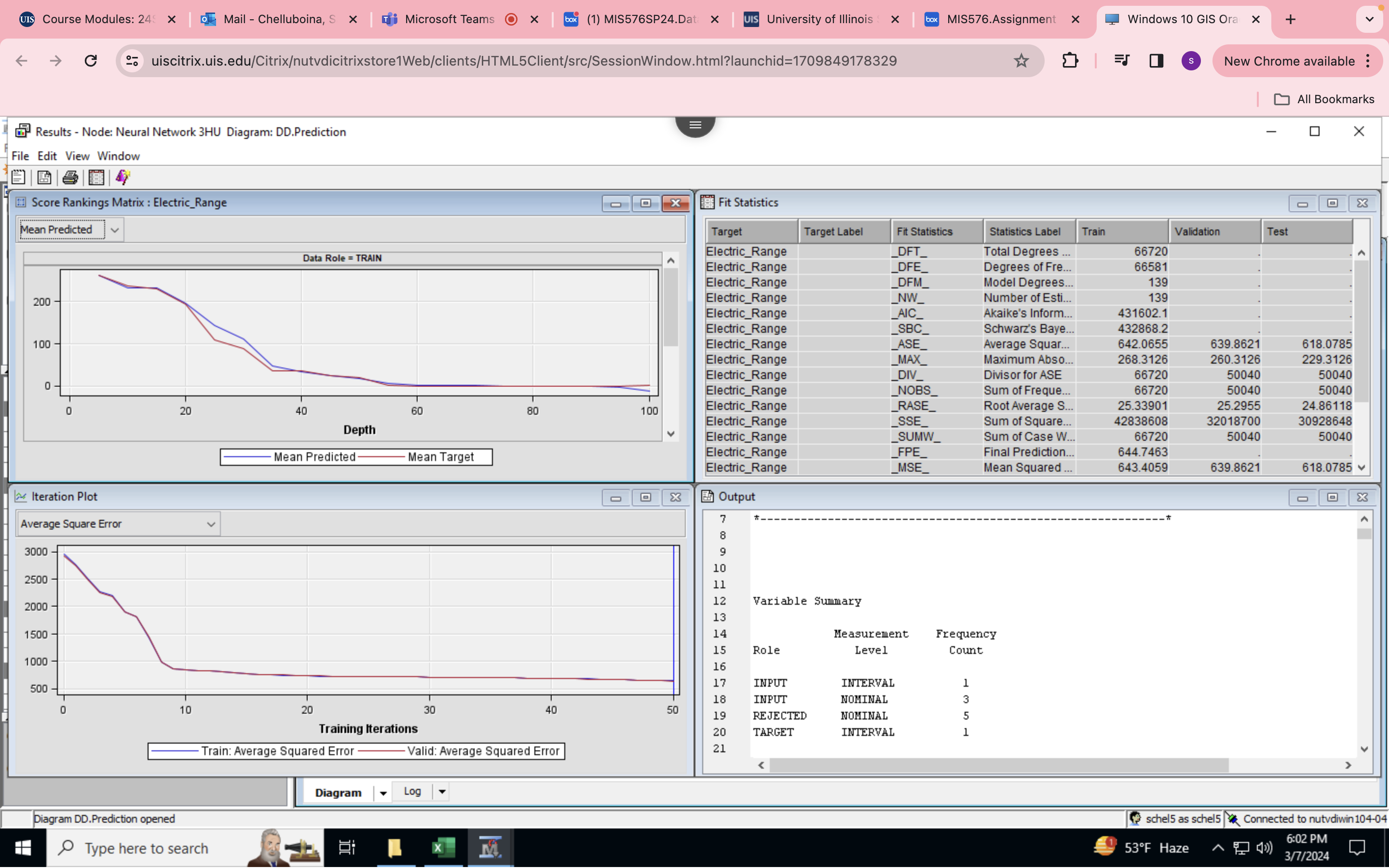
****

Fig.3.4. Neural Network Results

We then compare the three best model nodes, and the best node chosen by SAS turns out to be **Regression Tree B3D6.**

A screenshot of a computer

Description automatically generated

Fig.3.5.Final Predict Model ScoreRanking Result

A screenshot of a computer

Description automatically generated

Fig.3.6. Final Predict Model Output

**A screenshot of a computer

Description automatically generated**

Fig.3.7. Final Predict Model Fitstatistics

**Scoring the Model:**

In comparing our model with the scored data, we evaluated the predictions against the actual Electric\_Range labels. From the Excel sheet generated during scoring, we found that out of 166,800 total observations, our model made 117,678 correct predictions. This indicates that the accuracy of our model stands at approximately 70.59%. These findings provide valuable insights into the performance of our predictive model and its ability to accurately classify genres based on the provided data (Datacadamia, 2020).

A screenshot of a computer

Description automatically generated

Fig.3.8. Scoring Accuracy

Our predictive model achieved an estimated accuracy of approximately 70.59%, signifying its capability to forecast the need for charging stations with considerable reliability. This accuracy data shows how well the model can identify patterns and trends in the data, enabling stakeholders to decide on the deployment of charging stations with knowledge. These results highlight the need for proactive infrastructure design going forward to support the increasing use of electric vehicles and promote environmentally friendly transportation options.

**Conclusion and practical (actionable) recommendations**

The Whole Data set is about the Electric Vehicle population in three types, named hybrid electric vehicles (HEV), plug-in hybrid electric vehicles (PHEV), and battery electric vehicles (BEV). Which are currently registered through the Washington State Department of Licensing (DOL). These Vehicles are tentatively manufactured in the year 1997-2024. Where TESLA, RIVIAN, KIA, BMW, CHEVROLET and many other manufacturing companies are designed these Vehicles. These Vehicles with different models are being in the market with great competition. But the Electric Range which will be the target variable for this project Which is the number of miles of a particle model Electric vehicle, that plays crucial role to decide the performance of a Vehicle based on battery, price, feature, and country it is manufactured.

The aim is to predict the best vehicle which can give more miles and their charging stations that are available in those locations. Also, the benefits in using hybrid or battery vehicles. The sales of each Vehicle based on the availability has also been playing crucial part to decide on future predictions of the demand and improvements in keeping better service of a Vehicle. At present only the Washington registered vehicles data set is used to Analyse the data. Where Supervised and unsupervised techniques are used such as for exploratory analysis methods of Ward, average, centroid. Where Unsupervised for Decision Tree, Regression, Neural Network models. Which can be decided by its accuracy. Overfitting techniques, misclassification rate. From all these results we can see that In exploratory we can explain through number of clusters before and after sampling+/-2n formulae. Also for predictive we use misclassification rate and overfitting techniques. While scoring results are obtained by using the accuracy formulae i.e,. (TP+TN)/(TP+TN+FP+FN). All these results are compared after data cleaning and data visualization. These results also can be used to predict the number of vehicles in a particular geographic area with the visualization and The Urban usage and the charging stations required can be identified by the high frequency of vehicles and their Electric range based on the make, model, company. Also we can identify whether vehicle is eligible for clean Alternative Fuel Vehicle for an alternative fuel options in near future.

**Team Member Contributions:**

| **Project Kick-off** | **Define Problem Statement** | **Assigned To** | **Timeline** |
| --- | --- | --- | --- |
| **Project Initiation** | 1. Identify Project | Project Team | 01.31.23 |
|  | 2. Establish Project Team |  | 01.31.23 |
|  | 3. Develop Project Charter |  | 02.03.23 |
|  | 4. Conduct Initial Kick-off Meeting |  | 02.03.23 |
| **Data Exploration and Selection** | 1. Select Sample Dataset with Observations | Project Team | 02.17.23 |
|  | 2. Obtain and Validate Data |  | 02.17.23 |
|  | 3. Explore and Understand Data Characteristics |  | 02.17.23 |
|  | 4. Prepare Data for Modelling |  | 02.17.23 |
| **Variable Optimization** | 1. Conduct Variable Selection | Project Team | 02.26.23 |
|  | 2. Implement Dimension Reduction Techniques (e.g., PCA) |  | 02.26.23 |
| **Exploratory analysis** | 1. Convert Excel file into SAS format. | Haritha & Prasanth | 03.01.23 |
|  | 2. Perform exploratory analysis in SAS Enterprise Miner. |  | 03.01.23 |
|  | 3. Apply clustering methodologies (ward, average, centroid). |  | 03.01.23 |
|  | 4. Visualize original dataset in Tableau for initial insights. |  | 03.01.23 |
| **Iterative Clustering** | 1. Conduct multiple iterations of clustering using a 10% sample. | Haritha & Prasanth | 03.01.23 |
|  | 2.Refine clustering methodology. |  | 03.01.23 |
|  | 3.Strengthen reliability and accuracy of results. |  | 03.01.23 |
| **Predictive Analysis** | 1. Utilize regression, decision trees, and neural networks. | Satya and Nikhitha | 03.02.23 |
|  | 2.Predict outcomes and unveil correlations. |  | 03.02.23 |
|  | 3. Create a detailed score report. |  | 03.02.23 |
|  | 4.Compare various models for informed decision-making |  | 03.02.23 |
| **Finalization and Interpretation** | 1. Finalize Selected Model | Project Lead | 03.06.23 |
|  | 2. Interpret and Understand Model Results |  | 03.06.23 |
|  | 3. Prepare Final Model Report |  | 03.06.23 |
| **Results Dissemination** | 1. Communicate Results | Project Team | 03.07.23 |
|  | 2. Prepare and Submit Project Report |  | 03.07.23 |

**References:**

1. **Yashu Singhal (2024). Electric Vehicle Population Dataset. (2024, January 27). Kaggle.** [**https://www.kaggle.com/datasets/yashusinghal/electric-vehicle-population-dataset?resource=download**](https://www.kaggle.com/datasets/yashusinghal/electric-vehicle-population-dataset?resource=download)
2. **Washington Cities by population.Washington Demographic by cubit.** [**https://www.washington-demographics.com/cities\_by\_population**](https://www.washington-demographics.com/cities_by_population)
3. **Chen, T. D., Kockelman, K. M., & Khan, M. (2013). Locating Electric Vehicle Charging Stations: Parking-Based Assignment Method for Seattle, Washington. Transportation Research Record, 2385(1), 28-36.** [**https://doi.org/10.3141/2385-04**](https://doi.org/10.3141/2385-04)
4. **Yawei He, Kara M. Kockelman, Kenneth A. Perrine, Optimal locations of U.S. fast charging stations for long-distance trip completion by battery electric vehicles,**
5. **Journal of Cleaner Production,Volume 214, 2019, Pages 452-461, ISSN 0959-6526,** [**https://doi.org/10.1016/j.jclepro.2018.12.188**](https://doi.org/10.1016/j.jclepro.2018.12.188)**.**
6. **Bailey, J. (Ed.)., Khan, L. (Ed.)., Washio, T. (Ed.)., Dobbie, G. (Ed.)., Huang, J. Z. (Ed.)., & Wang, R. (Ed.). (2016). Advances in Knowledge Discovery and Data Mining: 20th Pacific-Asia Conference, PAKDD 2016, Auckland, New Zealand, April 19-22, 2016, Proceedings, Part I. Retrieved from https://i-share-uis.primo.exlibrisgroup.com/discovery/fulldisplay?docid=alma9982171640105898&context=L&vid=01CARLI\_UIS:CARLI\_UIS&lang=en&search\_scope=MyInstitution&adaptor=Local%20Search%20Engine&tab=LibraryCatalog&query=any,contains,khan%202016,&mode=advanced&offset=0**