**Analysis of Electric Vehicle Population Data**

This project proposes a data-driven analysis of Electric Vehicle (EV) population data of Washington State to inform sustainable transportation strategies and industry stakeholders while contributing to a greener future. We're genuinely thrilled about diving deep into this dataset, for our Python class. EVs aren't just a buzzword; they represent the future of mobility and a significant leap toward environmental conservation. With the global spotlight on sustainable living and pollution control, the surge in EV manufacturing is undeniable.

Now, let's talk data: Our dataset from Washington, a state with an eclectic mix of buzzing urban centers and serene countryside, is a goldmine of insights. Through Python's powerful data analysis tools, we aim to dissect the registrations of Battery Electric Vehicles (BEVs) and Plug-in Hybrid Electric Vehicles (PHEVs). This uncleaned data set containing 15000 records of Washington state's EV population will be analyzed, with several important features such as Vehicle identification number, County, City, Make, Model, Year, EV type, Clean Alternative Fuel Vehicle (CAFV) Eligibility, Range, Electric utility, Legislative District, DOL Vehicle ID. By leveraging Python's capabilities, we're trying to decode trends, gauge the BEV vs. PHEV preference, and derive actionable insights. This isn't just an academic exercise; it's our endeavor to understand and potentially influence the future of sustainable transport. (State of Washington 2023 Electric Vehicle Population Data)

Exploratory data analysis (EDA) provides a comprehensive view of the underlying patterns, relationships, anomalies, and structures within data. In our study, we delve into an electric vehicle population dataset comprising 15,000 records, encompassing details like VIN, county, city, and specifics about the vehicle model, type, and more. Our team undertook the task of data extraction from the "data.wa.gov" platform, procuring the dataset directly from the specified URL. Upon successful retrieval, we imported the CSV file into a designated software application, most likely Microsoft Excel, for an in-depth analysis. A significant portion of our project was dedicated to the meticulous data cleansing process, wherein we identified and rectified discrepancies like zeros and NaN values by substituting them with contextually relevant data or mean values, where applicable. Recognizing the importance of a streamlined dataset for effective modeling, we strategically eliminated superfluous columns that did not contribute to our modeling objectives. In our commitment to delivering a coherent dataset, we meticulously restructured and labeled the data, ensuring that column headers were self-explanatory, and values were comprehensible for future analyses. Through visualizations, we will assess the frequency distribution of electric vehicle ranges, compare the makes of electric vehicles with their respective ranges, investigate the distribution ratios between battery, hybrid plug-in, and fuel electric vehicle types, and analyze the distribution of electric vehicle attributes using a violin plot.

A graph with blue bars

Description automatically generated

The histogram depicts the distribution of electric ranges for a set of vehicles. Most vehicles have an electric range of 0 to 50 miles, with fewer vehicles in the subsequent range categories, and a notable smaller peak in the 200 to 300 miles range.

A screenshot of a graph

Description automatically generated

The bar chart displays the maximum electric range of various vehicle makers. "Tesla" has the highest electric range, surpassing 300 miles, while several other makers have ranges distributed between 150 to around 300 miles.

A screenshot of a graph

Description automatically generated

The pie chart illustrates the distribution of two types of electric vehicles: Battery Electric Vehicle (BEV) and Plug-in Hybrid Electric Vehicle (PHEV). The majority, at 77.6%, are Battery Electric Vehicles, while the remaining 22.4% are Plug-in Hybrid Electric Vehicles.

A diagram of an electric vehicle type

Description automatically generated

The violin plot provides a visual comparison of the electric range distribution between two types of electric vehicles: Plug-in Hybrid Electric Vehicle (PHEV) and Battery Electric Vehicle (BEV). While PHEVs have a more concentrated range distribution around 50 miles, BEVs display a broader distribution, with peaks around 50 and 200 miles, indicating varied electric ranges for these vehicles.

A screenshot of a computer screen

Description automatically generated

The scatter plot illustrates the relationship between "CAVF" on the x-axis and "Electric range" on the y-axis, using three distinct colors for data points. Most of the data points, represented in green, cluster on the far-right side indicating high CAFV points correspond to varied electric range points, while the lone red and blue points suggest lower CAFV values.

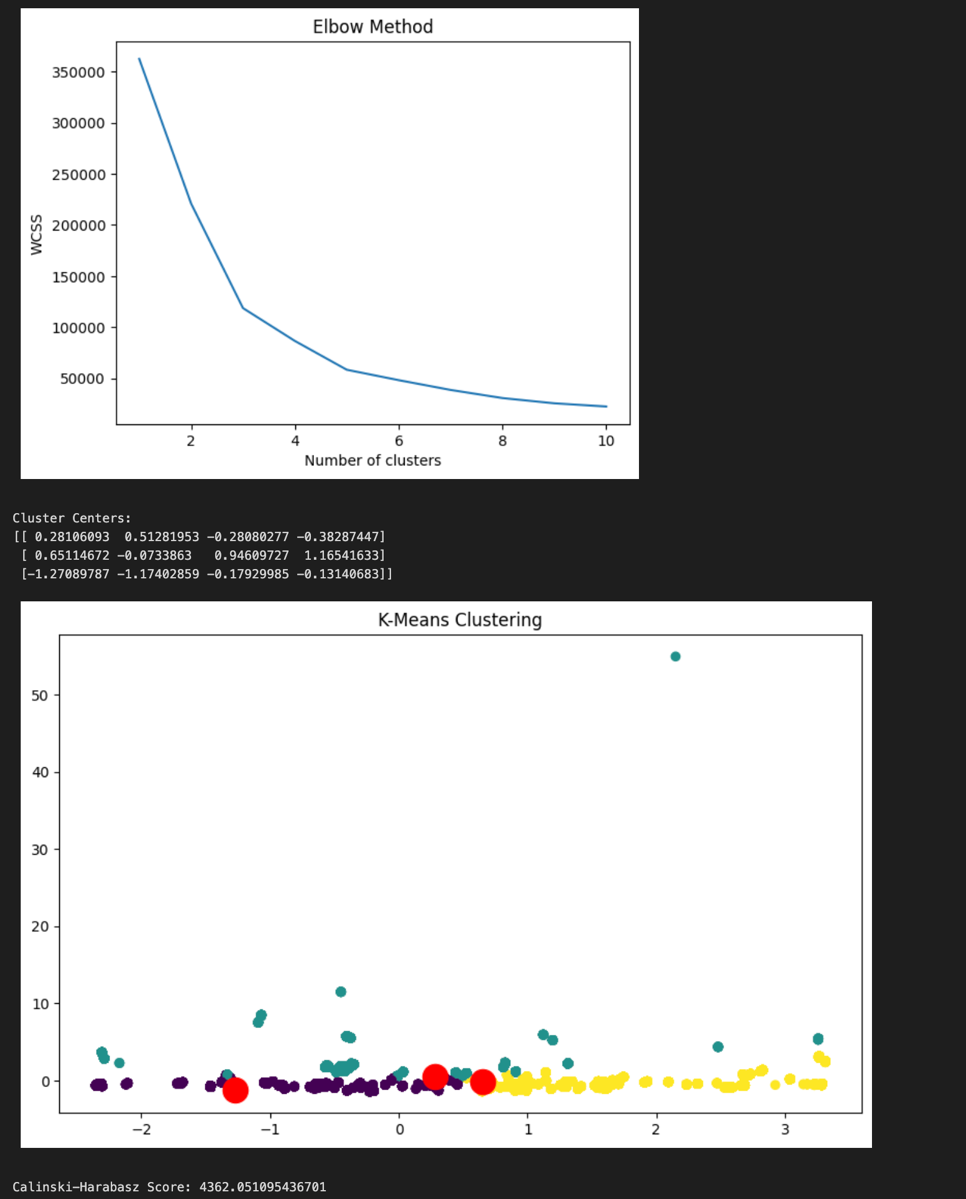
This data can be analyzed in many ways and answer several questions like determining popular EV models, evaluating adoption trends, comparing BEVs and PHEVs registrations, identifying high-adoption areas, etc. But the primary questions we will be analyzing are:

The primary goal of Model 1, which employs a decision tree algorithm, was to classify electric vehicle types based on features like City, County, and State. As part of the preprocessing, the categorical variables were transformed using label encoding. The dataset was divided into training and testing sets, with 80% used for training and 20% reserved for testing. Upon evaluation, the model achieved an accuracy rate of approximately 77.5%. However, a closer examination of the classification report and confusion matrix reveals disparities between the two classes: while the "Battery Electric Vehicle (BEV)" class displayed commendable precision, recall, and f1-score values, the "Plug-in Hybrid Electric Vehicle (PHEV)" class exhibited subpar metrics. This insight underscores the need for potential model improvements, especially for PHEV classifications.

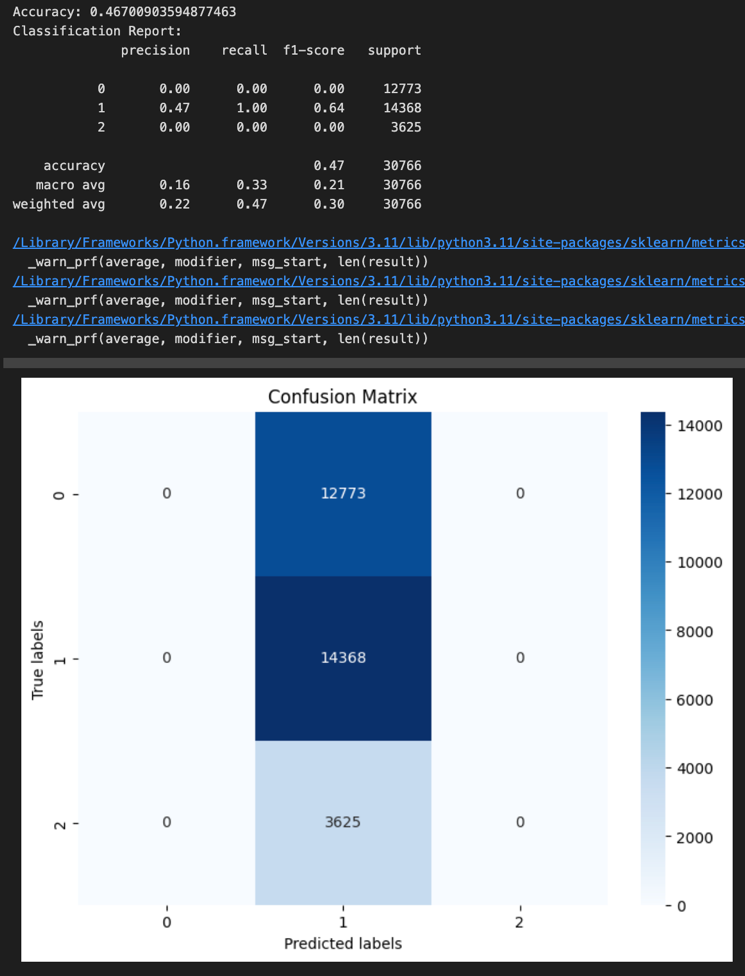
**A screenshot of a graph

Description automatically generated**

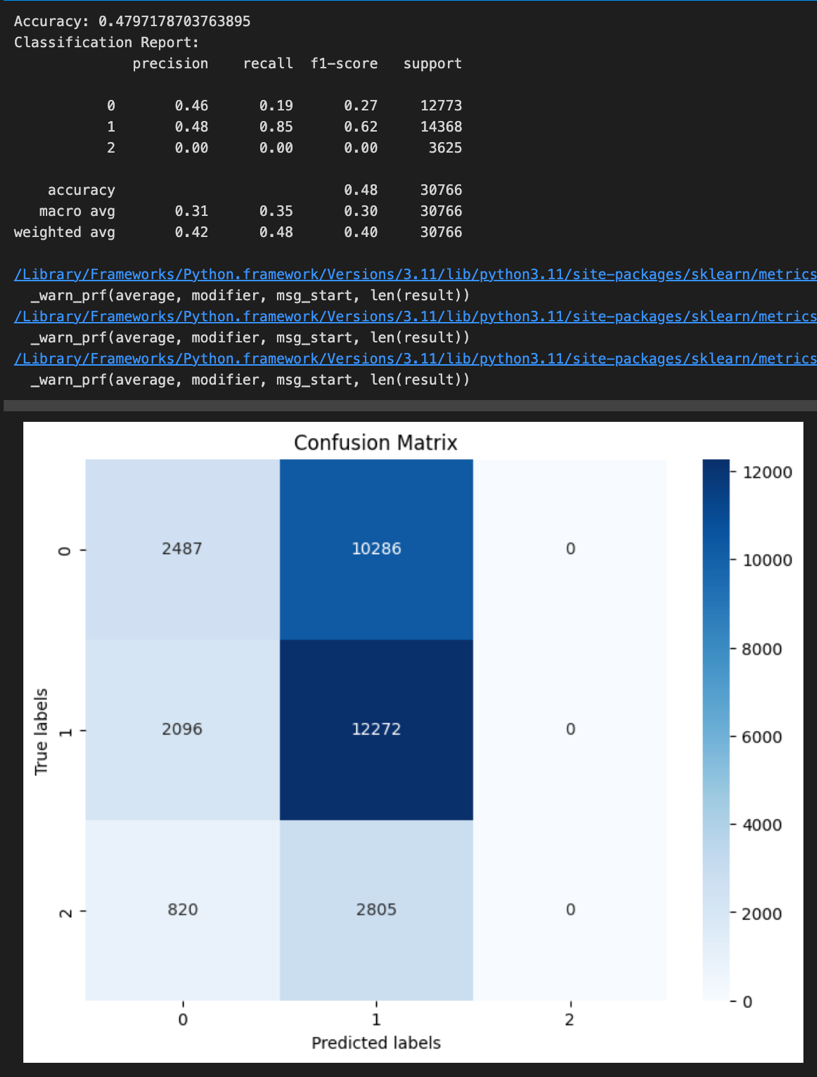
Model 2 analysis showcases the application of K-means clustering on vehicle data, particularly focusing on features like 'Make', 'Model', 'Base MSRP', and 'Electric Range'. The Elbow Method, depicted in the first graph, suggests an optimal cluster number around 3, where the within-cluster sum of squares (WCSS) begins to plateau. Using this insight, the K-means algorithm was applied with three clusters. The subsequent scatter plot displays these clusters in different colors, plotted on the two principal components derived from PCA. Notably, most data points are concentrated towards one area, but distinct clusters still emerge. The reported Calinski-Harabasz Score, a metric to evaluate cluster quality where higher values indicate better clustering, further quantifies the quality of this clustering.



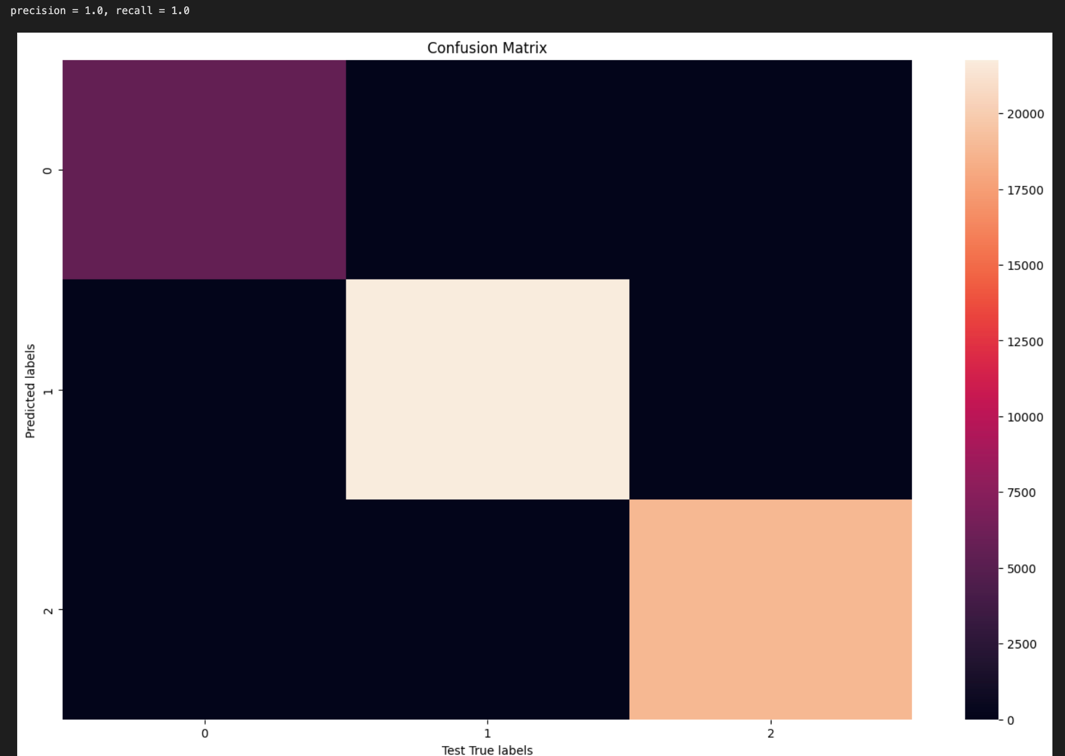
Model 3 utilizes a logistic regression approach to predict the 'Clean Alternative Fuel Vehicle (CAFV) Eligibility' of vehicles based on their 'Legislative District'. The model's overall accuracy stands at approximately 46.7%. The classification report reveals that the model is only effectively predicting for the class labeled '1', with a precision of 0.47, recall of 1.00, and an F1-score of 0.64. For the classes '0' and '2', the precision, recall, and F1-score are all 0. The confusion matrix visually confirms these observations, showing that all predictions fall into the '1' category. This indicates that the model might be suffering from a class imbalance problem, making it inefficient at predicting the other two classes.



Model4 analysis illustrates the performance of a Gradient Boosting Classifier used to predict 'Clean Alternative Fuel Vehicle (CAFV) Eligibility' based on 'Legislative District'. The achieved accuracy for this model is approximately 47.9%, which indicates that the model's predictions are correct about half of the time. The classification report reveals that the classifier performs well for the class labeled '1', with a recall of 0.85, but struggles with the class labeled '2', where both precision and recall are 0. The confusion matrix further validates these observations, showing a high number of false positives for the '0' class and no correct predictions for the '2' class. The warnings above the confusion matrix suggest potential issues with some classes having no predicted samples, which aligns with the poor performance for class '2'. Overall, while the model performs decently for one class, there's significant room for improvement, especially for the underrepresented classes.



Model 5 analysis reflects the implementation and performance of a Linear Regression model to predict the "Electric Range" of vehicles based on their "Model Year". The chosen feature, "Model Year", implies an exploration into how newer vehicle models might impact electric range. The resultant Mean Squared Error (MSE) is 7426.20, which is a measure of the average squared difference between the observed actual outturn values and the values predicted by the model. A high MSE indicates that the model may not be a perfect fit for the data. The R-squared value, which provides a measure of how well the model's predictions match the actual data, is approximately 0.1925, or 19.25%. This suggests that only about 19.25% of the variability in the "Electric Range" can be explained by the "Model Year", implying that there are likely other significant factors affecting the electric range not accounted for in this simple linear model.  
  
Model 6 analysis utilizes a Random Forest classifier to predict the 'CAFV' (Clean Alternative Fuel Vehicle) status based on the 'Electric Range' of vehicles. The purpose is to understand if a vehicle's electric range can determine its classification as a CAFV. The model was trained on 70% of the data and validated on the remaining 30%, ensuring a robust evaluation. The precision and recall values, both of which are reported as 1.0, indicate a perfect classification by the model. The presented confusion matrix visualizes these results, where the diagonal elements represent correct predictions and off-diagonal elements would indicate misclassifications. However, given the perfect precision and recall, this suggests that the 'Electric Range' may be a very strong indicator of CAFV status, at least for the dataset in question.



In conclusion, our electric vehicle population data analysis six models presented employ a range of algorithms, from decision trees to clustering and regression, to extract patterns from vehicle-related data. Model 1's decision tree algorithm managed a decent accuracy for classifying electric vehicle types, though improvements are needed for the PHEV classifications. Model 2's K-means clustering, on the other hand, provided valuable insights into vehicle grouping based on features like 'Make' and 'Electric Range', with the clusters being clearly distinguished. Both Models 3 and 4 aimed to predict CAFV eligibility based on 'Legislative District', but both struggled with imbalanced classes or poor predictions for certain classes. Model 5's linear regression revealed that "Model Year" only explains a modest portion of the variability in "Electric Range", pointing towards other influential factors. Finally, Model 6's near-perfect predictions using a Random Forest classifier underscore the power of ensemble methods and suggest that 'Electric Range' is pivotal in determining CAFV status. Collectively, while some models showcased significant potential, others highlighted the nuances and challenges associated with machine learning, emphasizing the importance of diverse feature consideration and addressing class imbalances.

# **References**

State of Washington. (September 15, 2023). *Electric Vehicle Population Data*. Data.gov.<https://catalog.data.gov/dataset/electric-vehicle-population-data>

Licensing, W. S. D. of. (2023, September 14). *Electric Vehicle Population Data: Data.WA: State of Washington*. Data.WA.gov. https://data.wa.gov/Transportation/Electric-Vehicle-Population-Data/f6w7-q2d2