EcoML: Harnessing Machine Learning for Electric Vehicle Adoption and Environmental Impact Analysis

Kajal Singh
StudentId: x23192461
MSc in Data Analytics
Email: x22112600@student.ncirl.ie
URL: www.ncirl.ie

***Abstract*—This aim of this research is to analyse the application and performance of machine learning methods across three different dataset : Air Quality Index (AQI) with geographic data, Electric Vehicle Population Data, and Total Emissions Per Country (2000-2020). By using variety of machine learning algorithm over the dataset like logistic regression, random forest, K nearest neighbour,clustering and linear regression model the author will answer few specific questions related to environmental impacts and sustainable practises. Each model used in the practise was trained to predict different outcomes for each of the datasets with high degree of accuracy and precision. The research also highlights which factors have major impacts on those target variable in each of the model. From Exploring the influence of geographical factor on air quality or predicting the category of pollutant emission trends of different types to finding out the type of EV adoption rates across different locations. Primarily results showing varying degree of effectiveness of the machine learning models, with a major focus on environmental policies and sustainable urban planning. This research will offers practical insights for policymakers and stakeholders in making data driven decisions for future sustainability.**

***Index Terms*—Linear regression, Random Forest, K- nearest neighbour, Clustering, , Logistic Regression,EDA(Exploratory data analysis)**

I. Introduction

"The best way to predict the future is to always analyse the past" in this era which is completely dominated by data driven decision making methods around, machine learning acts as an bridge between both predictive analysis and raw data. This research explores the application of machine learning algorithms on 3 different datasets: Air Quality Index (AQI) and Lat Long of Countries, Electric Vehicle Population Data, and Total Emissions Per Country (2000-2020). Each of the mentioned dataset provides a unique lens through which the author can analyse the relationship between environmental data and technology enhancements. The urge to address the environmental challenges and enhanced sustainable practise was always an concerning and debatable topics. Air quality majorly impacts the health as well as ecological system, proving it to be a topic of global monitoring and governance. At the same time adaptation of EV vehicles has also been equally concerning and crucial step to reduce urban pollution and transition to newer renewable energy resource. This research is a somewhere vital for policy makers and will also impact environmental planning.

This research basically aims to evaluate and analyse different machine learning algorithms for datasets using different machine learning methods to uncover patterns and predict future outcome. The selected datasets each contain over ten thousand rows and multiple attributes, offering a robust foundation for analysis. For this research the author has used five different machine learning models including regression model such as linear regression and logistic regression, as well as classification model as K- nearest neighbour, random forest and clustering methods. The models were trained to predict Air quality index category in the first dataset(Air Quality Index (AQI) values with geographical coordinates), type of electric vehicle in the second dataset(Electric Vehicle Population Data), and emission item in the last dataset(Total Emissions Per Country (2000-2020). These predictions not only address to specific research questions but also improve the performance of the model. This research will help the author to grasp technological adoption, providing a comprehensive view of its efficiency and scope.

The insights resulted from this project will be beneficial for a diverse set of stakeholders, environmental policy makers and government agencies that can leverage the finding to enhance the air quality regulations by analysing the pollution emission of each country and shape a sustainable urban planing strategies. By considering the AQI and emission they can implement more targeted interventions. Automotive industries will get an brief and insights into adaptation pattern of the consumer, and then they can plan marketing strategies to align with consumer trends and choices further.

II. Related Work

Machine learning techniques have been used in many different research's before such as : a. Air Quality Index (AQI) and Lat Long of Countries Research on air quality index prediction usually requires many machine learning algorithm and statistical methods to estimate and forecast values. D. A. Vallero the writer of the book "Fundamental of Air pollution"

[1] inspired the author to choose the research topic on the first place this book studies the effects and science of air pollution. It covers causes, consequences, and mitigation strategies and places a strong emphasis on diverse methodologies. The primary conclusion highlights how urgently productive and statistically grounded initiatives are needed to solve air quality problems and protect both human and environmental health. After having an grounded knowledge what pollution is exactly and how dangerous it is for the mankind. The author was concerned to find out which particular matter was responsible for the pollution alot. And then came across two different research paper what actually highlight the same issue.The research of Dixian Zhu, Tianbao Yang, and Xun Zhou [3] analysed the hourly concentration of different air pollutants (e.g., ozone, particle matter (PM 2.5) and sulfur dioxide) and how they have affecting the air quality. They utilized urban data and machine learning techniques to predict fine particulate of matter (PM2.5) concentrations, same as pooja bhalghat who studied to determine the quality of air Concentration of Gases such asso2, no2,co2, rspm, spm. etc in India to predict the same analysis.But this time she did that analysis in India only.From both these analysis the author somewhere found out which gases are more likely for pollution. Hence, the author selected the dataset Air Quality Index (AQI) and Lat Long of Countries which has a target column as AQICategory. This target variable shows the quality of air in that particular location. Through this research the author performed an predictive analysis by making an model that will further conclude the Air quality of that place into different category like(”Good”, ”Moderate”, ”Healthy”, ”Very unhealthy”,”Hazardous”). According to these studies, improving prediction accuracy and reliability requires combining a variety of data sources and advanced modelling techniques.

b. Electric Vehicle Population Data The popularity of electric vehicles is influenced by multiple factors including economic incentives, infrastructure and technological advancements. The research ”Intelligent hybrid vehicle power control—Part I: Machine learning of optimal vehicle power” [9] has focused on using machine learning to control the operation of vehicle power systems and the prediction of user behavior for electric cars (EVs). By using machine learning to boost vehicle power use, Murphey, Yi Lu and Park (2012) pioneered the development of intelligent control systems for hybrid automobiles by presenting a research that offers a strong foundation for intelligently controlling the power of hybrid cars, resulting in improved performance and fuel efficiency as well as lower emissions. This strategy makes significant contributions to the growth of sustainable automobile technology in addition to promote the use of hybrid technologies. After going through this research the author wanted to know more about the user behaviour about this trend thus author go through ”Ensemble machine learning-based algorithm for electric vehicle user behavior prediction” by Chung, Yu-Wei and Khaki, Behnam and Li, Tianyi and Chu, Chicheng and Gadh, Rajit this research introduced an an machine learning-based algorithm specifically tailored for predicting electric vehicle user behavior. [10] Their research underscores the significance of understanding user patterns to increase energy consumption and battery utilization in electric vehicles. By harnessing various machine learning models, the study provides insights into user behavior, which can help in designing more efficient and user-friendly EV systems. And at last the research paper ”A machine learning method for predicting driving range of battery electric vehicles” [11]by Sun, Shuai and Zhang, Jun and Bi, Jun and Wang, Yongxing and others. This research highlights the use of machine learning techniques to forecast battery electric vehicle (BEV) driving range, an essential aspect for consumer trust and vehicle economy .Their strategy resolves range anxiety, one of the most common worries among EV users, by forecasting the driving range using historical data and real-time analytics. Taken as a whole, these studies show the various ways that machine learning may be applied to improve the efficiency and environmental friendliness of electric and hybrid cars. Thus when the author selected the dataset ”Electric Vehicle (EV) adoption metrics” for the research, author choosed the Target variable as Electric vehicle type to predict which type of EV users are more trusting. By going through the past research’s the author got an idea about the choice and have an idea about the issue. Thus this research will help the stakeholder to know the consumer preferences and thus plan there next step according to that.

c.Total Emissions Per Country (2000-2020) Predicting and analyzing emissions trends is vital for environmental policy and planning. Machine learning methods have been extensively used to forecast emissions from various sources. From the research of ”Energy consumption prediction and diagnosis of public buildings based on support vector machine learning: A case study in China” [13] the author Liu et al. (2020) demonstrated the efficacy of using support vector machines (SVM) to predict and diagnose the energy consumption patterns of public buildings in China. This study emphasizes the potential of regression model on sustainable energy used in the building sector, offering insights into the factors influencing energy inefficiency and providing a foundation for remedial actions. This research gave the author an idea of using machine learning model on the chosen dataset ”Total Emissions Per Country (2000-2020)” to identify what is the major reason of pollution across the world. Similar to this, Javanmard and Ghaderi (2022) created a hybrid model to anticipate greenhouse gas emissions using energy market data by combining machine learning methods with optimization approaches. This approach help with both strategic planning for emission reduction in urban environments and with examining emission trends. In order to reduce environmental effects, this model highlights how policy creation and predictive analytics may work together. [14] Collectively, these studies show how machine learning may be used in a variety of contexts to improve energy efficiency, predict environmental effects, and support sustainable practices in a range of industries. By providing innovative approaches and useful implications for controlling energy use and emissions, each research study improves the

foundation of knowledge and supports the transition to more ecologically friendly and sustainable behaviors.

d. Conclusions for related work The research included provide insightful information about how machine learning might be used for environmental analytics, but sometimes often encounter flaws with the data quality, model adaptability, and clarity of the results of machine learning. For example, different environmental situations and underlying data distributions may mean that machine learning models that work well in one geographic location do not always move well to other settings. In addition, policymakers may find it difficult to understand and clarify the decision-making procedures of certain machine learning techniques, especially deep learning, due to their rigid nature.

III. Information Harvesting Methodology

This research follows to the CRISP-DM (Cross industry standard process for data mining) methodology. This method includes many different phases such as Business understanding, Data understanding, Data preparation,modeling, evaluation and deployment phases of the research. This particular project objective is to determine the factors that influence Air Quality of a location(lat and longitude) by analysis the total emission per country to find quantity hazardous gases and also find the reason responsible for this pollution with this it also focuses more on trend of Electric vehicle adaptation and see how these variables will help to predict future trends and how they contributing to environmental sustainability efforts. The methodology section of this report outlines the systematic approach adopted to carry out the investigation and compile necessary information. It acts as a guide, outlining the methods, approaches, and resources that are used in the research to ensure the accuracy, consistency, and validity of the results. [8]

A. Data Collection:

i. Air Quality Index Prediction and Analysis: Data collection: The dataset AQI of countries was collected from kaggle and thus the dataset consists of 16394 rows and 14 columns in it. The columns are:

-Country: This is the name of the nation where the air quality is being measured.

-City: The nation-specific city or place where the air quality data are collected.

-AQI value.: An numerical representation of the air quality index. In general, lower readings signify higher-quality air.

-AQI Category: A categorical representation of the AQI score, usually varying from ”Good” to ”Hazardous” according to the seriousness of the health risk.

-(CO AQI) value: The level of carbon monoxide pollution is represented by the Carbon Monoxide Air Quality Index.

-Category CO AQI: The CO AQI values effect on health impact category.

-Ozone AQI: An AQI is a measurement of the concentration of Ozone.

-Ozone AQI Category: The category for ozone acidity index (AQI) its a value that corresponds to health impact.

-NO2 AQI Value: Nitrogen Dioxide AQI value, indicating the level of NO2 pollution.

-NO2 AQI Category: The health impact category for the NO2 AQI value.

-PM2.5 AQI Value: Particulate Matter (2.5 micrometers and smaller) AQI value.

-PM2.5 AQI Category: The health impact category for the PM2.5 AQI value.

-lat (Latitude): The latitude of the monitoring location.

-lng (Longitude): The longitude of the monitoring location.

Once the dataset was downloaded from kaggle it was cleaned and and preprocessed to ensure that it was ready for analysis. When the author mentioned prepossessing that step includes removing duplicate values, dealing with missing values, encoding categorical and numeric values. And then transforming the categorical value to numeric by label encoding. The preprocessed data was further used for analysis that includes steps like feature selection, model selection and model evaluation. ii. Electric Vehicle Population Data Similar

	row	col
AQI.Category	4	3
PM2.5.AQI.Value	11	3
AQI.Value	3	4
PM2.5.AQI.Value.1	11	4
PM2.5.AQI.Category	12	4
CO.AQI.Category	6	5
CO.AQI.Value	5	6
Ozone.AQI.Category	8	7
Ozone.AQI.Value	7	8
AQI.Value.1	3	11
AQI.Category.1	4	11
PM2.5.AQI.Category.1	12	11
AQI.Category.2	4	12
PM2.5.AQI.Value.2	11	12

Fig. 1. List of columns in the dataset

to the AOI dataset the data for Electric vehicle was also collected from kaggle, a popular platform for sharing dataset. The dataset contains 135039 rows and 17 columns.And the dataset is in CSV format. -VIN (1–10): Stands for the first ten characters of the vehicle identification number, which can provide details about the create, model year, and other characteristics of the car.

-County: The county in which the car is registered, which is useful in examining local trends in the adoption of electric vehicles.

-City: The registered vehicle’s city, which is helpful for further deeper investigation within regions. State: The vehicle’s registration state, which reveals more general regional patterns in the acceptance of electric vehicles.

-Postal Code: The vehicle’s registered postal code, which is helpful for complete geographic studies and linking to local infrastructures like charging stations.

-Model Year: The year the car was produced, indicating how old it is and possibly pointing at future technological developments in more advanced versions.

-Make: The car’s manufacturer, which can be used to determine buyer tastes for specific electric vehicle brands.

-Model: The precise model of the car is important for understanding customer tastes for various kinds of electric vehicles.

-Electric Vehicle Type: Indicates whether the car is a Plug-in Hybrid Electric Vehicle (PHEV) or a Battery Electric Vehicle (BEV), which is important information to know while comparing adoption rates and usage trends.

-Base MSRP: Manufacturer's Suggested Retail Price, giving an idea of the cost of the vehicle, which can be correlated with adoption rates. Legislative District: The legislative district where the vehicle is registered, useful for policy analysis and targeting by government incentives.

-DOL Vehicle ID: A unique identifier assigned by the Department of Licensing, useful for tracking individual vehicle records. Vehicle Location: The geographic coordinates (longitude and latitude) where the vehicle is typically located or registered.

-Electric Utility: The electric utility provider for the vehicle's location, relevant for studying the impact of different utility policies on electric vehicle adoption.

-2020 Census Tract: The census tract code based on the 2020 Census, useful for detailed demographic and socioeconomic analysis related to vehicle adoption.

-Electric Range: The maximum distance a car can go on electricity, which influences the choice of the buyer. The vehicle's market price is indicated by the base MSRP, or manufacturer's suggested retail price. Legislative District: For the purpose of conducting a policy effect analysis, the legislative district where the car is registered.

-DOL Vehicle ID: An exclusive tracking number issued by the Department of Licencing.

-Vehicle Location: The usual location of the vehicle's geographic coordinates.

-Electric Utility: The supplier of electricity at the vehicle's location, pertinent to utility policy research. -2020 Census Tract: Census tract code for socioeconomic and demographic research, derived from the 2020 Census.

-Eligibility for Clean Alternative Fuel Vehicles (CAFVs): This indicates if the car qualifies for any tax breaks or lower rates due to its reducing impact on the environment.

The author has applied the same data preprocessing and model selection techniques on this dataset same as mentioned above.

iii.Total Emissions Trend Analysis This dataset was also downloaded from kaggle this contains 58766 rows and 25 columns. This dataset is also in csv format.

-Area: This column identifies the geographic location of the country for which the emissions data is reported.

-Item: This specifies the source or category of emissions. This differentiation helps in understanding the specific contributions of emission.

-Element: Details the nature of the emissions different gas. And there direct and indirect source of emission.

-Unit: Indicates the unit of measurement used in the dataset, typically kilotons, which is a standard unit for reporting large-scale emissions.

-Yearly Columns (2000-2020): Each column from 2000 to 2020 represents the amount of emissions recorded in that particular year for each 'Item' and 'Element' combination.

B. Data Pre-processing

The collected data was then preprocessed to ensure the data is clean and ready for analysis. This method involves removing duplicate values,missing values and encoding categorical values using one hot encoding. This process is useful to conclude that the data is in usable format for the machine learning models. Here as an over view of the data preprocesing step for each of the datasets are: There are few common steps that we need to apply on eah of the datasets such as:

i. Importing Libraries: The required libraries for data cleaning, manipulation and vizulization are imported. These libaries include tidyverse,caret, caTools, ggplot2, geosphere, scales,lubridate, dplyr, readr, data.table, and tidyr.

ii. Importing the dataset: The dataset was imported using read,csv() function.

iii. Data cleaning: The author used process for data cleaning like finding missing, null values and duplicate values which are then removed from the dataset to get a better prediction of model further.

Now after the common steps the author will apply few data transformation steps for each of the model individually:

A. Air Quality Index Prediction and Analysis After the implementation of basis data cleaning and transforming, here is the further analysis: -Selecting the relevant features: In this step the relevant feature are selected based on their impact on target variable which is AQICategory for this dataset. It is determined that logistic regression feature will be used for predictions instead of random forest method because of its high accuracy and more relevant and persistent error values. As shown below: - After the data cleaning process the author found while checking that there are multiple values in the target variable (AQICateogry) to apply logistic regression as it is often suitable for classification tasks when the responses variable is categorical in this case we transformed the categorical values to 0,1,2,3,4 respectively.

```
+   mutate(`AQI.Category` = case_when(
+     `AQI.Category` == "Good"                            ~ 0,
+     `AQI.Category` == "Moderate"                        ~ 1,
+     `AQI.Category` == "Unhealthy for Sensitive Groups"  ~ 2,
+     `AQI.Category` == "Unhealthy"                       ~ 3,
+     `AQI.Category` == "Very Unhealthy"                  ~ 4,
+     `AQI.Category` == "Hazardous"                       ~ 5,
+     TRUE                                                ~ NA_integer_  # Catch-all
for unexpected categories
+   ))
>
```

Fig. 2. Conversion of categorical values into numerical values

Using logistic regression rather than random forest because logistic works best for smaller datasets and their easy interpretabilty, robustness and the evaluation of metric from this analysis like accuracy, precision, recall, F1 score, and AUC gives a perfect reason that the model is not over fitting for this dataset.

-Exploring the effect of specific features on the AQIcategory: It is explored using Histogram and GG plot. This is done

to get an better understanding of the relationship between the features and target variable. Overall preprocessing steps are taken to transform the dataset more suitable for analysis and modeling, and to improve the accuracy for the predictions.

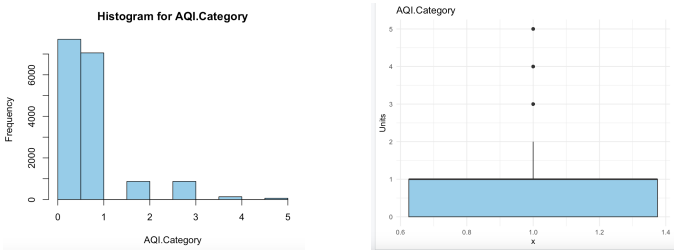

Fig. 3. Histogram AQI Category

B. Electric Vehicle Population Data After performing the basic EDA function now we will apply advance function over the model. This piece of code perform transformation by transforming the categorical value into numeric values, then by removing outlier, forming confusion metrics, splitting of data into test and training, plotting graphs and lastly calculating values of different features of the model. There it comes the part of Visualization which consist of histogram.

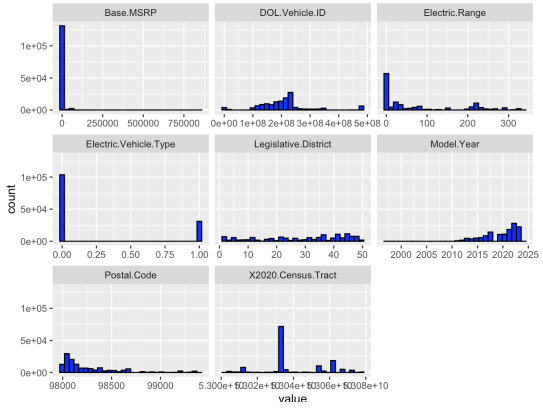

Fig. 4. Histogram for the dataset

-Overall the above points demonstrate the importance of data prepossessing in preparing data for analysis and modelling. By performing these steps the author can improve the quality of their analysis and produce more accurate and reliable results.

C. Total Emissions Per Country (2000-2020): By applying all the basic function over the dataset we received a clean and filter dataset. Which clearly separated the categorical and numeric values such as: The following steps were taken for

```
> categorical_features
[1] "Area"    "Item"    "Element" "Unit"
> numeric_features
 [1] "X2000" "X2001" "X2002" "X2003" "X2004" "X2005" "X2006" "X2007" "X2008"
[10] "X2009" "X2010" "X2011" "X2012" "X2013" "X2014" "X2015" "X2016" "X2017"
[19] "X2018" "X2019" "X2020"
```

Fig. 5. Categorization of columns

this dataset apart from the basic such as visualization tool is used to plot the target variable before and after the data preprocessing:

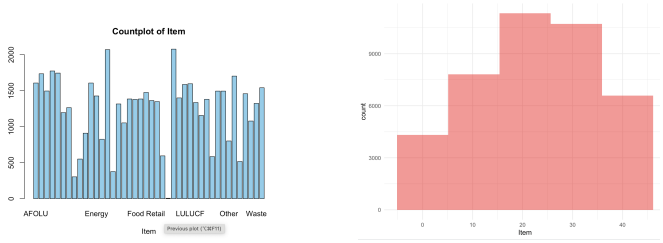

Fig. 6. Histogram Plot for target variable

IV. Model Selection:

Multiple different columns were target variable in this research such as AQICategory, Electric vehicle type and Item column from the other two datasets. Different machine learning models such as logistic regression, linear regression, random forest, clustering and KNN is used and chosen based on their ability to perform very well on both larger and smaller datasets and their suitability for regression models. [7]

A. 1. Air Quality Index Prediction and Analysis

Bases on the dataset, logistic regression model was implemented and evaluated for the AQI dataset. The evaluation metrics used for comparison were RAE,MSE,RSE. And values of F1 score and AUC score for advance implementation. i. Logistic regression Logistic regression is a statistical model used for categorical classification tasks, although it can be extended to multiclass clasificatio also. Few reasons are mentioned below to choose logistic regression over any other for this model are: 1.Multinomial outcomes: Logistic regression is very well suited for categorical variables. Making it versatile choice to predict categories outcome for our target variables such as good, moderate, unhealthy etc. 2. This model works very well with small datasets. 3. The linear regression model had an r square value of on the train and test, indicating that it explains about of the variability in data. This model fitted so well in the model that it seems to be the best choice for this dataset. This model gave the best fitted model with an accuracy of 0.89, Precision of 1 and recall of 0.77.

ii. Electric Vehicle Population Data: For this dataset two regression model were applied named as random forest and clustering. Both these model have different aspect and different way of analysis for this data set.

i. Clustering: Clustering such as k mean or hierarchical clustering, is a form of independent learning, as the name suggest it is used to group a set of objects in such a way that objects in the same group are more similar to each other. applying this method to this dataset offer valuable insights like identifying patterns in the dataset, helps to calculate patterns and trends, it can also directly impact the niche market or customer segments which can help business further to tailor there products, manufactures and policy makers. For this

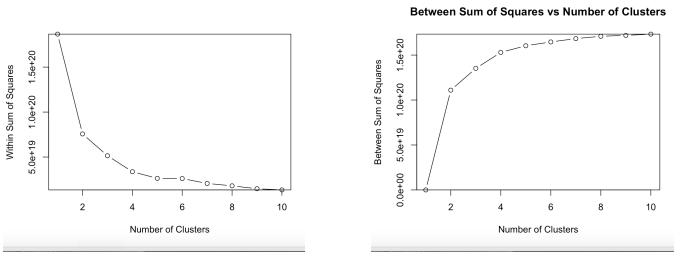

Fig. 7. Cluster plots for the dataset

particular dataset the author applied K cluster method, elbow method, and BSS value which give plots mentioned above:

ii. Random forest: After the analysis for this model i analysed that random forest will be the best choice for analysing data and to find the significant errors also.And also gives the perfect value of accuracy, precision, and recall to support the choice of model. Here are the reasons mentioned below: a. Non linear Relationships: Random forest are considered to be very good at handling non linear relationships between variables, which is likely given the diverse range. b. Multiple decision tree: Random forest builds multiple trees and merge their outcomes to get more stable and accurate predictions. c. Classification and regression: Random forest can be used for both classification and regression based on the target variable, making it an versatile tool for this dataset.

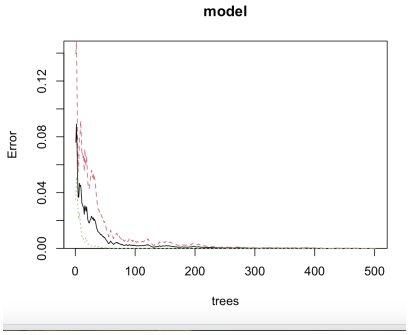

Fig. 8. Random forest model

Random Forest is generally better for your needs if the primary goal is predictive accuracy and model understanding in a controlled learning context, Random Forest is usually a better fit. It will enable you to forecast particular results and to understand how different factors affect such forecasts. By highlighting basic groups in the data that would not otherwise be labeled, clustering can improve this research and provide light on how various car kinds or user types stick together according to usage patterns or other traits. Random Forest, on the other hand, is usually more suitable and effective for direct prediction jobs, particularly when you have labeled data and need strong, dependable results. Hence this model is not came out to be perfectly fitted model for the dataset as it gives the accuracy value as 0.99, precision as 1 and recall as 0.99. Which proves to be that the selected model is perfectly fitting the dataset.

iii. Total emission per country: Same as the last dataset the author used two different model over it named as linear regression and KNN model. For this dataset the target variable was item column which basically shows that which type of pollution is there at a location.In the process of data preprocessing the author analysed a confusion matrix as shown below :

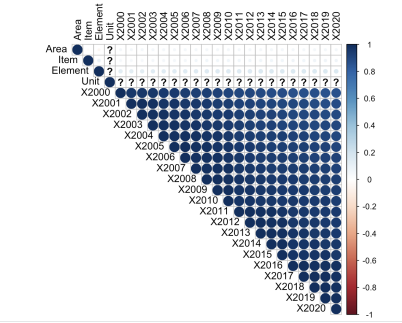

Fig. 9. Confusion Matrix

B.

i. Linear regression Firstly we applied linear regression over the model. To support the choice of model below is few reason for that: For the dataset mentioned above the author has used linear regression at first due to below mentioned reasons such as: 1. This model perfectly fits for dataset which shows lineartrend with them. moreover this model is particularly chosen for datasets where the change occurs at constant or neg-liable rate across time. 2. Linear regression provides a quantifiable way to forecast future based on historical data. This model also gave an MSE,RSE,MAE value which did not supported the model working at its best. 3. It also acts as a base line model to compare complex models ahead. 4.We also applied few hypothesis testing like breusch- pagan test and durbin- watson test whic helps to confirm the assumption of homoscedasticity and independent residual. Which increase the effectiveness and standard of this model. ii. K- Nearest neighbour: The

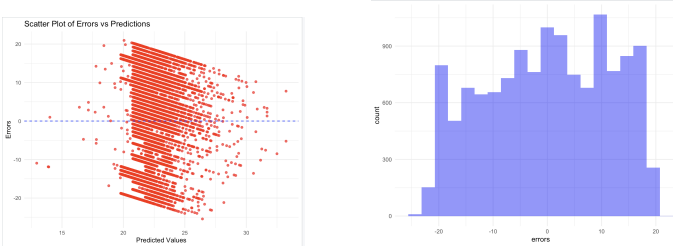

Fig. 10. Linear Regression Model

second method which i have chosen for this was KNN because this classification works with an multidimensional feature space. There are several features for choosing KNN mentioned below: a. It is very simple to implement and understand. It is

Model	rse	mae	F1-score	AUC-score	
Logistic Regression	0.89	0.344	0.79	0.90	0.96

TABLE I
LOGISTIC REGRESSION

beneficial to use when we have complex multiple categorical values in the dataset as we have seen above author has tired to use the linear algorithm but it wont work perfectly because there was an abrupt changes in the dataset through time period because of external intervention. b. This method uses local information of each data point to find the nearest neighbour and predict values for it. c. The number of neighbours impact the result alot. A smaller k makes the model sensitive to noise, whereas a larger k makes it expensive and potentially over smooth.

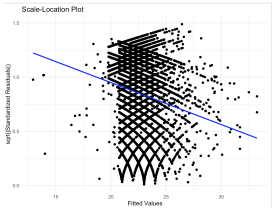

Fig. 11. Scatter plot

Inshort KNN is more suitable for emission dataset if the task involves complex dataset that are not linear. For this particular dataset the target column was 'item' which basically represent different category of emission sources and types.This column is without a natural order thus The selection of target model clarifies it that KNN is more accurate and appropriate for the model.The accuracy of the model was 0.87, the precision was 0.5 and the recall was 0.5. Which shows the prefect fit of this model. The author has received error values for both the model but the value from KNN model makes more sense than the above one because of the mentioned reasons:

V. RESULT

A. Air Quality Index Prediction and Analysis:

The table provides the performance of a machine learning model, named as logistic regression.To perform a model testing the dataset is then divided into two parts as train and test model. The train dataset is used by the model to build and improvise its working and once the model performs is qualified the test data is now used to check the functionality of the dataset. The model were evaluated based on r square calculated on both training and test sets, root mean square error, absolute mean error. Based on the results, the model achieved the best performance with r square value of

Each score of the model depicts something about the model like mse(0.79) means that models prediction is quite close to the actual values, rse(0.89) shows the amount of prediction which got deviated from the actual value on average and at last the mae(0.34) indicates that the model is quite accurate in terms of average error magnitude. When analysing the advance

Model	accuracy	Precision	Recall	F1- score
Random forest	0.99	1	0.99	0.99

TABLE II
RANDOM FOREST REGRESSION

Model	mse	rse	mae	mean
Linear regression	131.6873	11.47	9.7	22.26

TABLE III
LINEAR REGRESSION

error F1- score(0.34) means the model is quite accurate in terms of average error and at last the AUC- Score(0.96) means its that the model is excellent and has high degree of separability between positive and negative classes.

The author only applied one model over this dataset because after the analysis it was discovered that logistic regression suits the best for this model. The values which we have drawn from this model itself is very accurate and precised. Thus, inlcnclusion the model out performed in predicting the response variable for this dataset.

B. Electric Vehicle Population Data

The table above shows the performance of 2 different model i. Clustering and random forest, on the given dataset. The evaluation metric used are R- squared(both train and test), root mean square error(RMSE) and mean absolute error. ii. Random forest regression: This model was chosen first for the model by the author. For the model we can conclude that an accuracy of (0.99) shows that the model correctly predicts the true class label almost prefect for every instance, the precision(1.0) shows that every instance predicted as positive by the model was actually positive, recall(0.99) shows that model is able to identify almost the actual positive case, missing very few values. And at last for the advance errors the F1(0.99) means that not only model is precise but also robust in terms of finding the relevant case. But Clustering method gave the best visualization graph. In summary the clustering model outperformed it wheres random forest model perform best fit.

C. Total Emissions per Country

The table presents result of 2 different model one as linear regression and thier hypothesis and other as K- nearest neighbour method both used to evaluate different errors. The author used Linear regression and their hypo testing firstly which give the following errors as:

i. The linear regression model error measure performance of the model such as mse(131.68) gives an idea about the magnitude of error, lower the magnitude indicates better fit. Thus it shows that this model is not suitable for this dataset. Now coming upto the other error rse(11.47) means that the prediction is 11.47 untis away from the actual data point. mae(9.7) describes that models prediction is far related to the true data point, and at last mean(22.56) shows the baseline around which the other measurement are centered.

Model	accuracy	precision	recall
KNN	0.87	0.5	0.5

TABLE IV
CAPTION

Thus author choosed another model named as KNN for this dataset whose values are as:

After the output received from the linear regression the author applied knn on the model and received pretty accurate error such as accuracy(0.87) means n this model correctly that predicted outcome 87 percent of time, which is comparatively better. Precision(0.5) shows that when the model predicted the positive class its 50 percent correct of all the time. This shows that every two prediction one is correct. Recall(0.5) means model correctly identifies 50 percentage of all actual positive cases. This shows that model is missing half of potential cases.

This shows KNN method is far better than the regression model for this dataset. As it out performed in terms of accuracy, precision of the model, making it most suitable.

VI. Conclusion

The research explores how machine learning can support electric vehicle adoption and reduce environmental impact by analyzing three datasets: Air Quality Index (AQI), Electric Vehicle Population Data, and Total Emissions per Country. Models like logistic regression, random forest, k-nearest neighbors (kNN), clustering, and linear regression were applied to predict trends and classify data accurately. Based on the three sets of result, it is evident that the performance of the model varies depending upon the dataset and evaluation metrics used, In the first set of results the logitisc regression model has outperformed and the author didn't applied any other model because the precises answer the out depicted. For the second dataset Clustering model outperformed and for the last KNN model was chosen after the analysis.

References

[1] D. A. Vallero, *Fundamentals of air pollution*. Academic press, Massachusetts, 2014.

[2] K. Veljanovskal and A. Dimoski, "Air quality index prediction using simple machine learning algorithms," *International Journal of Emerging Trends & Technology in Computer Science (IJETTCS)*, 2018.

[3] H. Peng, "Air quality prediction by machine learning methods", University of British Columbia, 2015. DOI: 10.14288/1.0166787

[4] D. Zhu, C. Cai, T. Yang, and X. Zhou, "A machine learning approach for air quality prediction: Model regularization and optimization," *Big data and cognitive computing*, vol. 2, no. 1, p. 5, 2018.

[5] K. M. O. Nahar, M. A. Ottom, F. Alshibli, and M. M. A. Shquier, "Air quality index using machine learning–a Jordan case study," *Compusoft*, vol. 9, no. 9, pp. 3831–3840, 2020.

[6] N. N. Maltare and S. Vahora, "Air Quality Index prediction using machine learning for Ahmedabad city," *Digital Chemical Engineering*, vol. 7, pages 100093, 2023.

[7] C. Lesmeister, *Mastering machine learning with R*. Packt Publishing Ltd, 2015.

[8] A. Kassambara, *Machine learning essentials: Practical guide in R*. Sthda, 2018.

[9] Y. L. Murphey, J. Park, Z. Chen, M. L. Kuang, M. A. Masrur, and A. M. Phillips, "Intelligent hybrid vehicle power control—Part I: Machine learning of optimal vehicle power," *IEEE Transactions on Vehicular Technology*, vol. 61, no. 8, pp. 3519–3530, 2012.

[10] Y.-W. Chung, B. Khaki, T. Li, C. Chu, and R. Gadh, "Ensemble machine learning-based algorithm for electric vehicle user behavior prediction," *Applied Energy*, vol. 254, pp. 113732, 2019.

[11] S. Sun, J. Zhang, J. Bi, Y. Wang, et al., "A machine learning method for predicting driving range of battery electric vehicles," *Journal of Advanced Transportation*, vol. 2019, Hindawi, 2019.

[12] D. Pevec, J. Babic, and V. Podobnik, "Electric vehicles: A data science perspective review," *Electronics*, vol. 8, no. 10, pp. 1190, 2019.

[13] Y. Liu, H. Chen, L. Zhang, X. Wu, and X.-J. Wang, "Energy consumption prediction and diagnosis of public buildings based on support vector machine learning: A case study in China," *Journal of Cleaner Production*, vol. 272, pp. 122542, 2020.

[14] M. E. Javanmard and S. F. Ghaderi, "A hybrid model with applying machine learning algorithms and optimization model to forecast greenhouse gas emissions with energy market data," *Sustainable Cities and Society*, vol. 82, pp. 103886, 2022.

[15] Z. Zheng, X. Lin, M. Yang, Z. He, E. Bao, H. Zhang, and Z. Tian, "Progress in the application of machine learning in combustion studies," *ES Energy & Environment*, vol. 9, no. 2, pp. 1–14, 2020.