A logo with a yellow and blue design

AI-generated content may be incorrect.

**National College of Ireland**

**Project Submission Sheet**

|  |  |  |  |
| --- | --- | --- | --- |
| **Student Name:** | MANSI SHARMA……………………………………………………………………………………………… | | |
| **Student ID:** | ……………[x23410396@student.ncirl.ie](mailto:%782341%303%39%36@st%75%64e%6e%74%2e%6e%63%69%72l%2e%69e)………………………………………………………………………………………………… | | |
| **Programme:** | …Msc Data Analytics …………………………………………………………… | **Year:** | …2025…………………… |
| **Module:** | Data Mining and Machine Learning ……………………………………………………………………………………………………… | | |
| **Lecturer:** | ……………………………………………………………………………………………………………… | | |
| **Submission Due Date:** | 5.05.2025  ……………………………………………………………………………………………………………… | | |
| **Project Title:** | AIR QUALITY AND CAR ELIGIBILITY …………………………………………………………………………………… | | |
| **Word Count:** | ……………………………………………………………5318………………………………………………… | | |

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the references section. Students are encouraged to use the Harvard Referencing Standard supplied by the Library. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action. Students may be required to undergo a viva (oral examination) if there is suspicion about the validity of their submitted work.

|  |  |
| --- | --- |
| **Signature:** | ……………………………………………………………………………………………………………… |
| **Date:** | …………………………02.05.2025…………………………………………………………………………………… |

**PLEASE READ THE FOLLOWING INSTRUCTIONS:**

1. Please attach a completed copy of this sheet to each project (including multiple copies).

2. Projects should be submitted to your Programme Coordinator.

3. **You must ensure that you retain a HARD COPY of ALL projects**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. Please do not bind projects or place in covers unless specifically requested.

4. You must ensure that all projects are submitted to your Programme Coordinator on or before the required submission date. **Late submissions will incur penalties.**

5. All projects must be submitted and passed in order to successfully complete the year. **Any project/assignment not submitted will be marked as a fail.**

|  |  |
| --- | --- |
| **Office Use Only** | |
| Signature: |  |
| Date: |  |
| Penalty Applied (if applicable): |  |

AI Acknowledgement Supplement

# [Insert Module Name]

# [Insert Title of your assignment]

|  |  |  |
| --- | --- | --- |
| **Your Name/Student Number** | **Course** | **Date** |
|  |  |  |

This section is a supplement to the main assignment, to be used if AI was used in any capacity in the creation of your assignment; if you have queries about how to do this, please contact your lecturer. For an example of how to fill these sections out, please click [here](https://libguides.ncirl.ie/useofaiinteachingandlearning/studentguide).

# AI Acknowledgment

This section acknowledges the AI tools that were utilized in the process of completing this assignment.

|  |  |  |
| --- | --- | --- |
| **Tool Name** | **Brief Description** | **Link to tool** |
|  |  |  |
|  |  |  |

# Description of AI Usage

This section provides a more detailed description of how the AI tools were used in the assignment. It includes information about the prompts given to the AI tool, the responses received, and how these responses were utilized or modified in the assignment. **One table should be used for each tool used**.

|  |  |
| --- | --- |
| **[Insert Tool Name]** | |
| [Insert Description of use] | |
| [Insert Sample prompt] | [Insert Sample response] |

# Evidence of AI Usage

This section includes evidence of significant prompts and responses used or generated through the AI tool. It should provide a clear understanding of the extent to which the AI tool was used in the assignment. Evidence may be attached via screenshots or text.

# Additional Evidence:

[Place evidence here]

# Additional Evidence:

[Place evidence here]

AIR QUALITY AND CAR ELIGIBILTY

PROJECT

Mansi sharma

[x23410396@student.ncirl.ie](file:///C:\Users\mansi\Downloads\x23410396@student.ncirl.ie)

*Abstract*—This research includes two datasets which are ‘AIR QUALITY’ AND ‘CAR ELIGIBILITY PROJECT’ includes machine learning model which perform supervised and unsupervised method of machine learning in it to bring insight and better prediction for future reference. It includes regression and classification models which aim for ‘AIR QUALITY -DATA VALUE’ which predict the pollution and most affected area and ‘CAR ELIGIBILITY’ includes the government policy which is Clean Alternative Fuel Vehicle (CAFV) Eligibility, which is based on other columns such as model, year, electric range of car. It is based on regression models such as linear regression and random forest regression and classification models such as Logistic classification and KNN.

Keywords— Air Quality, Data Value, Random Forest Regression, Linear Regression, KNN, Electric Vehicle Eligibility, CRISP-DM

# Introduction

With Efforts in controlling environmental hazard and activities which lead to growth of air pollution, it is equally important to keep in mind that manual daily life and activity parallel to such efforts can bring fine results. Risk related to air pollution is affecting daily life and respiratory problems are causing concern while population is increasing and causing more exhaustion of resources. It leads to more surveillance towards Electric Vehicle (EVs). Both the datasets are socially relevant for such great concern over the environment and human activity. Both the dataset combines to work on the same perspective to figure out predictive data values for region which is affected the most by air pollution. It shows the eligibility of a government-based scheme ‘Clean Alternative Fuel Vehicle (CAFV) Eligibility’. This project includes a structured data analysis and machine learning model implementation with steps include data cleaning, preprocessing, encoding such as label encoding and one hot encoding, mapping of categories and converting columns into integer format to process machine learning model.

The ’AIR QUALITY’ data set includes columns such as ‘Unique ID’, ‘Indicator ID’, ‘Name’, ‘Measure’, ‘Measure Info’, ‘Geo Type name’, 'Geo Join ID’, ‘Geo Place Name’, Time Period’, ‘Start date’, ‘Data Value’, ‘Message’. These columns are dynamic in terms of data type and values which they contain. It is useful in getting insights and evaluating the target variable to bring meaning out of the dataset collected. It includes 18862 rows and 12 columns in the entire dataset. The data columns help in deciding the preprocessing. This dataset includes many variables which hold no value to the main goal to determine the data value for pollution and regions. Hence, many columns were dropped down and encoding was done in order to make the dataset more meaningful and numerical so that the machine learning models can be performed in it.

The ‘Electric Vehicle’ includes columns 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’, 2020 Census Tract’. These columns contain 239747 rows in total and 17 columns. The data type and values it holds help in determining the eligibility of the CAFV eligibility criteria for vehicles to be given Incentives while purchasing. Basically, the main idea behind this is to increase the purchase power for certain types of vehicles so that people show willingness to buy them. It is done to boost EV adoption on the road and people can also buy it easily. This dataset was imbalanced with the target variable. Further, preprocessing and encoding supports the dataset to perform models and produce insights.

# Literature Review

According to previous studies by Zhang et al. (2019) that used linear regression for PM2.5 prediction and research by Li & Wang(2020), the study showed analysis of air pollution in his dataset where he struggles to handle the non-linearity of model [1]. This model performs linearity and shows only relations which are less scattered but, he was not able to derive in complex data handling and non-linear regression problem in his model. This work also did not account seasonal part of analysis and multicollinearity in data. Another model which includes analysis of data without feature engineering and leads to vague analysis of cluster as it might be affected by dimensionality and leads to complex and less interpretable models to understand the model.

Here it has been done in this model to predict the non-linear relation of pollution using regression methods and analysis of better results by comparing it with KMean clustering that is regional separation of dataset and bringing better results and conclusion. This project resolves the issue which were lacking in his study of pollution as one of the leading factor in environmental concern.

Both kMean clustering Random Forest regression techniques bring solution to the previous research problems been done and it bring better insights and decision making analysis. In this research data preprocessing, feature engineering and encoding and scaling methods are applied in order to bring results as per prediction method is concerned. it also handle data imbalance by using SMOTE and dimensionality reduction is done with help of PCA [2] [3].

The Air Quality dataset is complex in terms of prediction as it follows prediction of data with respect to seasonal trend , region. There were many columns and parameter are given in order to make a decision which not only covers the region which is polluted the most. It adhere the other aspect by pre processing and encoding columns out of raw dataset. It involves two main machine learning models which are performed in this data in order to get the predicted value with residual gap to be minimum with respect to test axis. Since this data is not clean hence , data cleaning , followed by pre processing , Encoding label and one hot Encoding for data processing and performance for machine learning model. Since model which was used here is “Linear Regression”.

According to study by Study by Chen et al. (2021) there is an imbalance dataset in Electric Vehicle analysis which leads to poor performance of the entire model. The imbalance data set was a main problem in the entire project which can be overcome by SMOTE.

The Electric Vehicle dataset include columns which affect the main column in terms of classification. It was not easy but, decision and analysis is made with help of data column feature and its parameter decision which shows that car is eligible for government incentives. It includes data columns such as Model , year , Electric Range. These parameter affects the most to decide the vehicle decision. When feature selection was done the most affecting parameter was ‘Electric Range’. It contain SMOTE before performing KNN [4] [5]which brings better result and accuracy in model building and performance.

# Choice of methods:

For ‘AIR QUALITY’ dataset the targeted variable is ‘Data Value’ to understand the continuous data points which predict the value of a straight line (y = mx + c). It is used to find the linearity between the two variables but, considering the mean square error (MSE) which is 26.25081 and r2 score is 0.7435. It shows that the model is MODERATLEY strong with respect to relation of variable with target. This linear relation shows that homoscedasticity that is variance of residual in the entire target and actual value and no multicollinearity is found there as we can see through heatmap about the co-related relationship of variables amongst each other. Multicollinearities occur when two or more predicted shows high co-relation with each other. It shows that variable Convery similar information which can make linear regression model confusing. In the dataset given several columns such as ‘Vehicle miles travel’ and components which are extracted from raw datasets which undergo one-hot encoding. It raises the issue of which questions the stability and interpretation of model. Hence, ridge regression and lasso regression were applied to bring more clarity about the dataset prediction.

The Ridge (L2 Regularization) and Lasso Regression (L1 Regularization) both were applied which contribute to the model building as they clear whether multicollinearity is affecting or not. So based on further analysis the Mean square error (MSE) is 26.25 and r2 score is 0.7435 which shows that the model is identical as Linear regression. Further, Lasso Regression shows Mean square error is 30.32 and r2 core is 0.7036. It improves model simplicity and interpretation but, on the same hand it led to sacrifice the prediction value.

A graph with blue dots

AI-generated content may be incorrect. A graph with blue dots

AI-generated content may be incorrect. A graph with blue dots

AI-generated content may be incorrect.

The above visualizations shows the three different plot of Which include Linear regression , lasso regression and Ridge regression. It shows the linearity and relation of Actual and Predicted poiints in the dataset. The mdel are represented using scatter plot where x-axis contains the ACTUAL point and y-axis predict contain PREDICTIVE VALUES.

The linear regression shows positive relation as the actual value increase , the predictive value also goes up but, the problem arise it do not consider precision or predictiion for complex points. It shows sign of ‘Underfitting’ which is a sign that it capture insufficient data point , lack complexity.

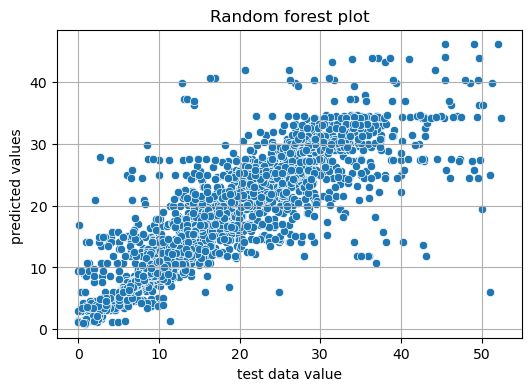
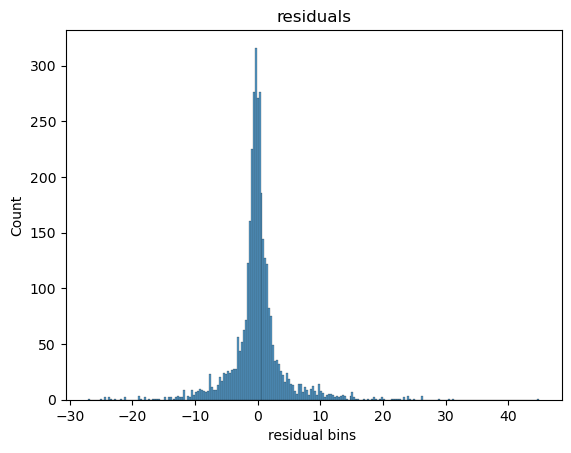
Lasso Regreesion shows that it shrinks the corfficient too much as flattened (clipped) is applied. For x-axis values which range from 40-50 shows prediction at 30-35 which shows it did now go upward trend. It shows model fail to capture the actual data and regualrization.

Ridge Regreesion visualization is better as compared to Lasso as it predicts complexity in prediction as it is less vertically spread. And prediction is closer to actual value. but, when it turns to higher prediction values it is unpredicted. As compared to Lasso it is a better mdoel.

The above three analysis show that linear regression generate positive co-relation for prediction but, sign of underfitted model . Another Lasso regression show shrink in coefficient value and loose prediction because of the same. While woth Ridge regresion analysis the prediction was better as it do not show wide spread and less vertical spread but , it underperform with higher actual predictive values.

Given the above limitation with linear model performance in regression another model ‘Random Forest Regression’ is performed. So that it consider the complex prediction , accuracy and prediction.

Random Forest Regression is an ensemble which make multiple decision tree to predict the value. It predict by averaging the value of different trees. Each tree give prediction and and average of all tree is final output. This mdoel shows a spike from overall all the models. It gave Mean square Value(MSE) as 19.99 and r2 score as 0.8046 aprox 80% variability in data which shows a strong model performance.it shows a strong linear trend and cover complexity as well , it predicts better for the data value.

The above are predicting values visualizationa and residual plot of Random Forest Regression. The plotting depicts that all the points are showing a linearity and cluster with each other. It adhere the complexity by predicting points which are above 40 , it shows diagnality in between x-axis and y-axis. Prediction is more accurate and better as compared to Lassos and Ridge regression. It is relatively narrow and compact with each other. It handle non-linear trend, features, iregular trend.it do not show underfittingand overfitting in trend.

The residual value which is gap in between test data and prediction data , it shows that most of the gap is near to zero that means it is accurate. It shows symmetry in plotting , that means data set is well predicted for all points it is nearly same. It shows that residual value is near to centre zero which means prediction is accurate.

A diagram of a graph

AI-generated content may be incorrect.

Average Pollution Level per Cluster:

cluster Data Value

0 0 32.480193

1 1 20.095772

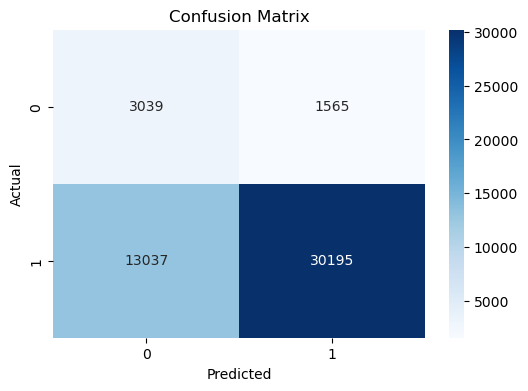
2 2 8.480852

After performing prediction value through Random Forest Regression, another model KMeans clustering is also there to bring some insightful analysis of the data value points in order to know the cluster of region which are affected the most with pollution level. The K-Mean clustering is performed so that policy making and environment concern should be there with those region which are affected the most. It raise concern for eco-friendly environment to draw the reasona and cause of air pollution. There are three clusters in it. For this PCA is performed to bring the clustering in a 2D format which shows clearity in the data generated. The region with ‘0’ is most affected region. It need growing concern and survillance, region with ‘1’ is moderately affected and point ‘2’ is least affected. It shows concern for regiona and data value relation. The value for 0 is 32.48 , 1 is 20.09 and 2 is 8.4. The value varied with respect to cluster which give clear idea.

Both the above model Random Forest provide prediction and cover complexity in it and K-mean clustering give a segmented idea for policy making and planning.

For Dataset ‘VEHICLE ELIGIBILITY MODEL’ , it target the variable ‘Clean Alternative Fuel Vehicle (CAFV) Eligibility’ . this variable is a parameter of classification model. It is the deciding factor wheather a vehicle is eligible for govrnment basied incentives and benfits based on parameter suhc as model ,year ,electric range. Classification models is based on Logistic regression first. It is to define the two type of decision 0 and 1.

Logistic regression is supervised classification machine learing model which is use in binary. It have accuracy score as 69% , it suggest it actually correctly predict for government shceme by 69.47%. The precision is 0.95 which shows model perform precisely with most of the dataset. In recall it state 0.66 that means model indicates 66 of actual data set with class 0 as to be non-eligible for the scheme. This logotic machine learning model is imbalance in nature due to which it shows that it is require SMOTE and other model to determine the best possible decision on basis of classification models.

 A blue line graph with white text

AI-generated content may be incorrect.

The above two are confusion matrix and ROC curve which are drawn from logistic Classification, in order to understand the model performance, these two diagrams play a crucial role.

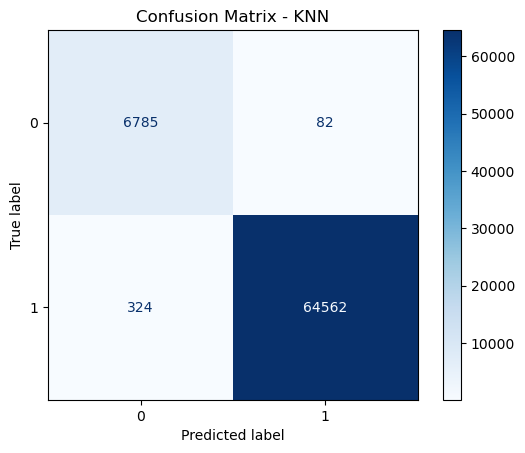
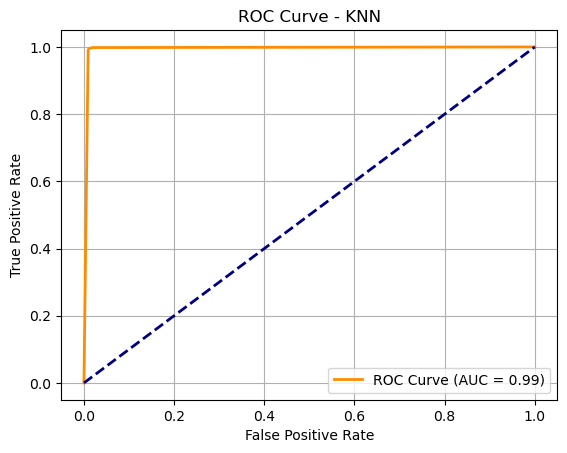
The confusion Matrix states that model shows from top left it states 3039 values which are not eligible for scheme, and from top right it shows incorrectly predict 1565 values as not eligible for scheme. For False negative part it predicts 13037 as to be not eligible when they were eligible. For bottom right, it shows true positive value that 30195 are eligible and it is as per prediction.

This “confusion matrix” made by using logistic classification shows that there are a greater number of values in ACTUAL 1 part than other, hence it is imbalance in nature. Hence, SMOTE can make this analysis better in future Model.

The KNN model is better for data distribution and flexible as well.

Another plot shows a ROC curve that is Receiver operating characteristic curve. It is a curve which depicts the nature of models as to be biased on not. It shows ROC value is 0.68. It is stating that the model is randomly choosing eligibility by 68% chance.

This is a clear depiction that model is biased, and SMOTE technique can make it better with another model which is KNN.

Another model is KNN, it is model which is followed by Logistic regression. It classifies cases based on majority, hence SMOTE is performed before implementation and a balance of both X\_train and y\_train is done so that it may lead to better performance with good accuracy of the model. KNN model perform extremely well because of the same and it shows that there were only 82 as False positive and 324 as False negative cases. With the better accuracy as 99.43%, it shows that the model is well defined in terms of accuracy and model precision since SMOTE was applied. It can be stated that the data is Balanced and highly structured as well.it is generalized and stable in nature now.

Logistic model in this dataset follows a problem with interpretation and it leads to struggle with imbalance dataset problems. On the other hand, KNN achieves better accuracy and with SMOTE it can overcome problem of data imbalance in nature. It is, hence, a perfect model to predict for government schemes.

## Methodology:

The KDD methodology is used to guide the project work for both the dataset that is ‘AIR QUALITY’ and ‘ELECTRIC VEHICLE ELIGIBILITY’.

It is entirely based on SELECTION, PREPROCESSING, TRANSFORMATION, DATA MINING, INTERPRETATION/EVALUATION.

1. *SELECTION*

The data selection was a complex part for both the datasets target a particular column which is used for prediction and classification. It is important to understand the meaning of the columns and question after which the model will be framed. The ‘AIR QUALITY’ dataset includes two main columns which was ‘DATA VALUE’ AND ‘REGION’ as criteria to perform machine learning model. The prediction of regression that is continuous type of model has been done on DATA VALUE which shows the pollution and air quality affected by numerical values and for clustering the data set to get insight out of it. The region has been used to perfectly relate the data value with the region so that it can bring real life solutions based on the dataset. On the other hand, the ‘ELECTRIC VEHICLE DATASET’ it includes two main issues that is ‘Clean Alternative Fuel Vehicle (CAFV) Eligibility’ and ‘Electric Range’, since both are highly co-related, hence it is quite dicey to select the dataset and co-relate the column meaning. But eligibility value as 0 and 1in scheme is the main area of selection. The dataset is from the USA. It is highly relatable to study environmental issues with respect to prediction and classification

1. *PREPROCESSING*

Both the datasets include pre-processing techniques in it. It was complex and a solution with respect to its value has been made as per encoding and type of category is concerned.

The ‘AIR QUALITY’ data set includes columns with no duplicated values, but one of the column names ‘Messages’, contains all NAN values which was dropped. It is include pre processing with many columns which were extracted from the other one. The ‘start date’s column contains a mix of years and months. So, both the columns were extracted from it. In same manner seasons such as winter, summer and annual average were extracted from season columns and given numerical value by using mapping function in it. The mapping function is used to determine the category of columns and assign numerical values based on number of columns. One of the main is one-hot encoding technique been done on Name column. It shows that there are many columns starting by name because label encoding could be misleading for machine learning models to categorize the name of all the air pollution and values which it holds to analyze might get affected. Hence, there are many columns which are extracted by one-hot encoding.

These outliers were also detected in this data which were removed with help of IQR – ‘Inter quartile range’ and filtering lower bound and upper bound from it.

On the other hand, the ‘ELECTRIC VEHICLE’ DATASET is also preprocessed with help of removing- all null values, duplicated values. In this dataset, Label Encoding is mainly done with many columns to determine the type. The Label Encoding was done with Postal Region, City, Country, Make, Model as main label encoding technique which labels all the unique values with a numerical value. Also, mapping the categorial values which contain only two or three types in their column was done. Columns such as ‘electric vehicle type’ and ‘CAFV’ eligibility were encoded using map function. Certain columns which hold no value in analysis were removed named as vehicle location, 2020 census tract, postal code, vehicle ID. These columns hold no value in determining the machine learning problem solution. Also, outliers were removed from the dataset to improve performance.

1. *TRANSFORMATION*

It is the main part of the entire process to prepare a model for performing. Both the datasets include transformation of columns from string to integer values and a co-relation matrix was made to study the values that have been made.

Majorly Transformation take place in ‘AIR QUALITY’ dataset. The transformation with season, month, year from date column is done to study better analysis of data value with respect to different times of the year and it shows the trend as well. This dataset. New categorical columns with air name were also encoded to compare the data value with respect to the target variable. It includes data scaling as well after splitting. In order to make a common scale for analysis.

PCA was performed in order to study the clustering behavior and dimensionality reduction was done in this step.

PCA stands for PRINCIPAL COMPONENT ANALYSIS. The high dimensional datasets are transformed into two-dimensional components with help of PCA. It is used in KMean clustering to transform and analyze better. There are two components named PCA1 and PCA2 which direct maximum variance. Later, a new data frame is created with help of PCA, and clustering K-Means is obtained with help of algorithm. The scatter plot below depicts the PCA and KMean cluster technique which is used to transform the data set. It is made with help of Python library Seaborn, and it shows a clear cluster of regions which are affected by data value



Transformation with ‘ELECTRIC VEHICLE’ takes place as it includes transformation of raw dataset with help of data Label Encoding technique, and it label the column separately. It is because the number of columns were more if it undergoes one hot encoding, it helps in reducing the number of columns as it can raise more complexity and cause difficulty in analysis of data set. It includes SMOTE.

SMOTE is a technique which is used for imbalance dataset. After Logistic analysis, it was clear that it was biased towards the majority as it gave confusion matrix based on biased result hence, before moving for another classification model. SMOTE technique was done in order to analyze and balance the dataset.

SMOTE stands for “synthetic minority oversampling technique”. It generates minority classes synthetic classes in Clean Alternative Fuel Vehicle (CAFV) Eligibility, this target column contains more ineligible cases which lead to rise of SMOTE and make working model more reliable for test data set. It brings synthetic classes which raise the accuracy and performance of KNN machine learning model.

1. *DATA MINING*

There are two data set which are ‘AIR QUALITY’ and ‘ELECTRIC VEHICLE’. These data sets involve four types of machine learning models

For ‘AIR QUALITY’ it involves regression model

1. Linear Regression- it is a great model which is easy to interpret for regression technique. It involves linearity of data set which is simple to predict value. But it is also sensitive to outlier or complexity if there is non-linear dataset then it may cause overfitting and not reliability. As per the air quality data set. The linear model shows mean square error in linear regression 26.250812379905263 and r2 score in linear regression: 0.743500108191007. hence, it Lasso and Ridge regression was applied in order to understand regression deep in it and figure out the best fit model for it. For Lasso regression technique -L1, it shrinks some of the coefficient to zero and it automatically select more relevant feature and reduce complexity. This is clear when r2 score using lasso regression: 0.7036662471422448 and mean square error using lasso regression: 30.327504987389688 And Ridge regression model includes all the features and reduce overfitting in its nature. This entire analysis of models claims that the model is overfitted and cannot be linear as it cannot consider the complexity.
2. Random forest regression - it is used for non-linearity of data set, and it handles outliers as well. It is performed to consider overfitting problems and complexity of the dataset. It gave r2 score in random forest: 0.8046818935774944 and mean squared error for random forest: 19.989322139416984 which shows better results and prediction as it covers complexity and better accuracy.
3. KMean – it is to understand pattern and segment of cluster with respect of data value. Basically, with more understanding and analysis of the dataset this helps in better policy making. It includes PCA as it tends to reduce dimension and make it easier to understand the values with respect to clusters.

For ‘ELECTRIC VEHICLE’ involves classification model, it involves two types of main models, that is Logistic classification model and KNN model.

1. Logistic classification

This is based on classification of binary classification as the main target column is Clean Alternative Fuel Vehicle (CAFV) Eligibility as there are two types of categories as eligible and non-eligible, it is important to decide. After logistic regression technique in model. The data set was found to be imbalanced in nature hence, SMOTE takes place. It gave accuracy score for logistic regression is: 0.6947487248097667 and for two type of classes it is different like for class 1, the precision is 0.95 and for class 0 it is 0.19. The F1 score is more for class 1 as it is 0.81 and for class 0 includes a minority score of F1 as 0.29. it clearly shows that logistic classification fails with minority and works for majority only.

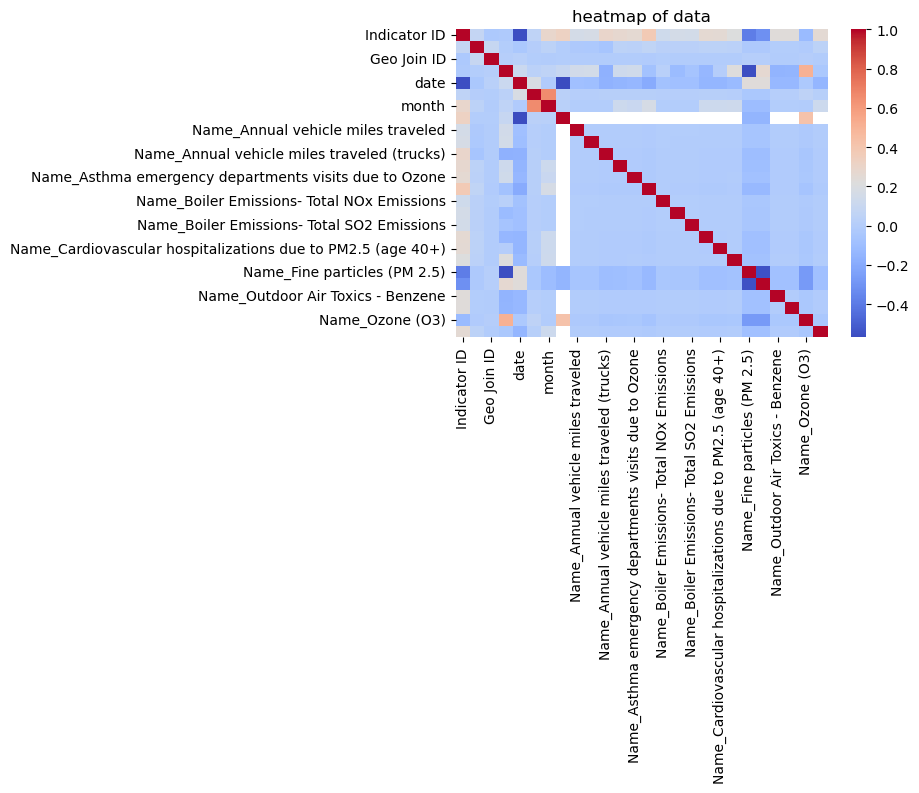
1. KNN

After performing SMOTE technique, it was considered to be the main method in order to perform classification model. KNN find nearest neighbor and closest data point. The model performs based on have 3 nearest neighbor values. Cross validation score was 5 which means it split the data into 5 folds and means accuracy is 1 and standard deviation is 0 which indicates that model is stable and consistent. It also shows effective SMOTE, or resampling is done and strong pattern.

1. . *INTERPRETATION /EVALUATION*

Both data sets were evaluated with imbalance value to dimension complexity.

The Air quality data involves regression models to predict air quality related metrics. The analysis and model performing reveal following insight:



The heatmap given shows the entire data set and correlation of variable, as far as machine learning models are concerned it shows the relationship of models between each other.

The linear regression model is simple due to its linearity that is based on y = mx + c, it struggles a lot with non-linearity as the dataset is complex and includes non-linear predictions which need to cover nonlinear approaches.

Next came Lasso regularization, which is L1, it reduces model complexity and shrinks the features. It cannot capture the pattern of the data set because it ignores or shrinks. After which L2 – ridge regression which gave r2 score as 0.84 and MSE as 19.98. it handles model complexity and penalizes the coefficient as well. But linear approach cannot be best fit for this as well.

The Random Forest Regression model gave the best prediction, and it is better than the linear regression model as it covers complexity and variability more effectively. It is a more suitable model for regression tasks. R2 score is 0.84 and Lowest MSE as 19.98. After this KMean model bring clarity of understanding in the dataset relation with region as bring cluster with it. It is an unsupervised learning model and divide data into clusters. It is used to better understand and make strategical decisions by government against region, climate and time period factors which leads to increase of pollution and air quality degrades as well. Although governments take continuous initiative, it brings precision and better understanding of pollution and air quality related factors in real life.

The heatmap if for ‘ELECTRIC VEHICLE’ DATASET: IT SHOWS CO-RELATION WTH ANNOTATIONS IN IT

A screenshot of a graph

AI-generated content may be incorrect.

The heatmap shows that all the above columns contain the highest value between relation of electric range column and main target column Clean Alternative Fuel Vehicle (CAFV) Eligibility. This shows complexity and on the same hand support to the variable has been decided as a target variable.

The Logistic regression model which is used in this model is to find binary classification of the column in order to help government with this initiative of EV vehicles on road and raise its production. The model struggles with imbalance and more class 1 dominated data which show major part to be eligible for the policy made by government and it shows overall accuracy to be 0.694 and precision was very low for this class. That is 0.19 for non-eligible class F1 score show 0.29 as majority class. Hence, after SMOTE it performed better with KNN as it shows perfect performance with accuracy to be 1.0 and standard deviation to be zero. It was better able to learn the pattern and distribution effectively after balancing the data set with synthetic samples.

## Conclusions and Future Work:

The “AIR QUALITY AND ELECTRIC VEHICLE ELIGIBILITY” project contains a lot of scope of improvement and discussion. The main objective of this project is to find the air quality and ESV model to direct government for better policy making and decision making with respect to the data set given. It is important to consider that AIR QUALITY is based on many factors not just on pollution, there is temperature, precipitation, humidity and geographical location of a place to see if a place is near to sea or it is urbanized place where car and vehicles run a lot in number. It is equally important to consider other deciding factors as well. But this project on the same hand contributes to determining and predicting which region and data value will get degraded as per prediction and seasonal factor is concerned. As we can see with increase in year, the data value increases because as urbanization is concerned and improvement in science and technology, these days vehicles are also increasing day by day. There are more different types of models.

This project holds more discussion value for the part where data value can vary with measurement units based on collection of data value, it is important to note that measurement unit should be same for all when it turns to air quality. The affected regions and areas are given by Postal Id, region, country and county. Although it is the same, for better understanding the longitudinal and latitudinal part and considering exact location or affected region of the data value Geo Analysis could make it better in order to bring clarity to locals and researchers. It is important to consider the complexity and scope of improvement with respect to more factor analysis approaches one can make much better and modify it.

For ELECTRIC VEHICLE data set. It is important to note the analysis which is based on year and model:

A graph of blue dots

AI-generated content may be incorrect.

This small analysis reveals a lot about today’s generation and how technology is advancing with time and scope of improvement. Government is brining initiative to bring EV on road and it is parallel that new Models are increasing with respect to car. It is important to take initiative with respect to the model and analysis based on certain factors which the government considers as in this data et model, year, electric range are important factor other factors such as financial part where someone is upper, middle or lower class. Things related to income of individual, affected regions based on pollution and cluster can be other factors which help in decision making of vehicle eligibility.

There are lots of model related improvements which one can make with respect to the model imbalance and consider leakage of data. The data set with respect to its imbalance nature is a great concern of fraud as well. Is there a problem in data collection or fraud entries? Government should show growing concern for this type of analysis as well as more eligible people for incentives without surveillance and considerations of their wealth as well can mislead the use of funds, although raising the demand can be done with other factors such as considering them with preference on road rules. Someone can get better credits and incentives by means of CNG, petrol and other aspects of Vehicle.

##### References