1. ***Introduction to Irland Transportation:***

Ireland, a picturesque island nation in Northwestern Europe, boasts a diverse and well-connected transportation network that plays a crucial role in facilitating the movement of people and goods across the country. The transportation infrastructure in Ireland encompasses a mix of modern roadways, railways, airports, and ports, contributing to the overall economic development and connectivity of the nation.

The road network, managed and developed by Transport Infrastructure Ireland (TII), is a vital component of Ireland's transportation system. TII, established through the merger of the National Roads Authority and the Railway Procurement Agency, focuses on providing an integrated approach to the future development and operation of the national roads network. This network not only ensures efficient travel within cities and towns but also connects remote regions, fostering economic growth and enhancing accessibility.

In addition to the extensive road network, Ireland maintains a well-developed rail system. The railways, managed by Irish Rail, serve both urban and intercity transportation needs. The network connects major cities and towns, offering a reliable and sustainable mode of transport. Ireland's commitment to improving rail infrastructure aligns with its goals of reducing congestion, promoting environmental sustainability, and enhancing the overall efficiency of the transportation system.

Furthermore, Ireland's air transportation system is supported by a network of airports strategically located across the country. Dublin Airport, Shannon Airport, and Cork Airport are key hubs facilitating domestic and international flights, connecting Ireland to various destinations worldwide. These airports not only cater to passenger travel but also play a significant role in the transport of goods, supporting trade and commerce.

The maritime sector also contributes substantially to Ireland's transportation landscape, with several ports serving as gateways for shipping and ferry services. Ports such as Dublin Port and Cork Port play a crucial role in facilitating trade and ensuring the smooth flow of goods to and from Ireland.

Ireland's transportation infrastructure is a dynamic and interconnected system that plays a pivotal role in shaping the nation's economic landscape and improving the overall quality of life for its residents. As the country continues to invest in the development and enhancement of its transportation networks, Ireland is poised to maintain and strengthen its position as a well-connected and accessible destination.

***Problem Statement:***

By harnessing the power of machine learning algorithms, Ireland can usher in a new era of smart transportation, where predictive analytics, real-time data processing, and adaptive systems work in concert to address current challenges and anticipate future needs. This initiative not only aligns with global advancements in transportation technology but also positions Ireland at the forefront of intelligent and sustainable transportation solutions.

***Objective:***

The integration of machine learning objectives aims to create a data-driven transportation ecosystem that adapts to dynamic conditions, improves decision-making processes, and ultimately provides a more seamless and user-friendly experience for both commuters and logistics providers. Additionally, there is a need to explore innovative ways in which machine learning can contribute to sustainability efforts, such as optimizing routes to reduce fuel consumption and emissions.

In the realm of Irish transportation, there exists a pressing need to leverage machine learning techniques to address critical challenges and optimize the efficiency of the existing infrastructure. Key issues include predicting and managing traffic congestion in urban areas, optimizing public transportation schedules, enhancing the accuracy of travel time predictions, and developing intelligent systems to ensure the safety and reliability of transportation networks.

***Exploratory Data Analysis:***

Exploratory Data Analysis help us to analyse the pattern from the data. This EDA analysis help us to Understand the data and make us to take a suitable decision. EDA is a Visualization representation.

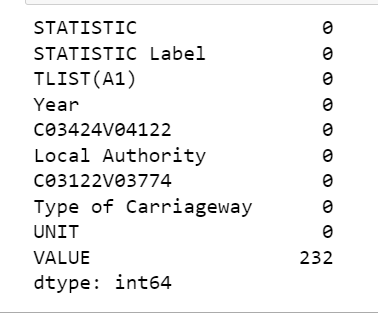


Figure1. Missing Values from the dataset

summary of a dataset, and it appears to be related to transportation or some form of statistical data. Let's break down the information:

STATISTIC and STATISTIC Label: These columns seem to represent some statistical measures or metrics. The fact that both have 0 missing values is positive, indicating that all entries have values for these columns.

TLIST(A1): The dataset contains 0 missing values in this column, suggesting that each entry has a value in the "TLIST(A1)" column.

Year, C03424V04122, Local Authority, C03122V03774, Type of Carriageway, UNIT: Similar to the STATISTIC and TLIST(A1) columns, these seem to represent different attributes or features in the dataset. It's good that there are no missing values in these columns.

VALUE: There are 232 missing values in the "VALUE" column. This column likely represents the numerical values associated with the dataset, and the missing values might need attention depending on the analysis or modelling you plan to perform.

STATISTIC: 1 unique value

STATISTIC Label: 1 unique value

TLIST(A1): 8 unique values

Year: 8 unique values

C03424V04122: 27 unique values

Local Authority: 27 unique values

C03122V03774: 5 unique values

Type of Carriageway: 5 unique values

UNIT: 1 unique value

VALUE: 122 unique values

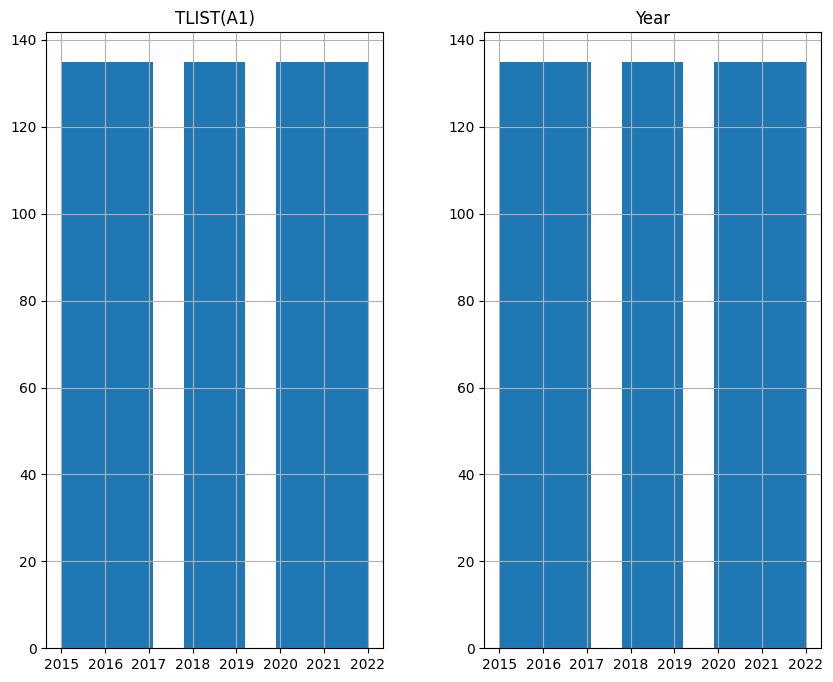


Figure 2. TLIST(A1) and Year

From the above visualization we can see that this dataset consist of 7 years of TLIST data are available. And the data is completely balanced.

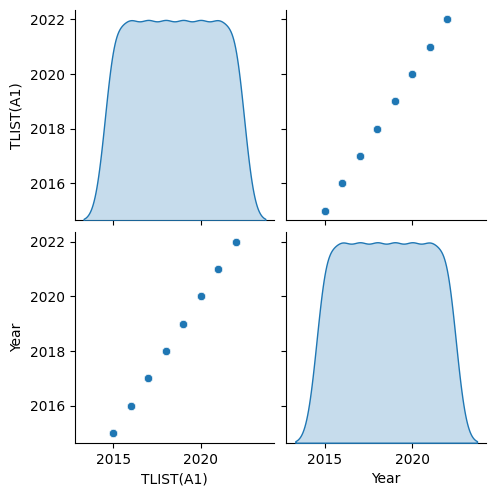


Figure 3. Scatter Plot

A scatter plot is a graphical representation of the relationship between two continuous variables namely TLIST(A1) and Year column. It is a way to visually explore and observe patterns, trends, and relationships between the variables. In a scatter plot, each data point is represented as a dot, and the position of the dot on the graph corresponds to the values of the two variables being compared.

***Positive Correlation:*** If the points generally slope upwards from left to right, it indicates a positive correlation—increasing values on one variable correspond to increasing values on the other.

***Negative Correlation:*** If the points generally slope downwards from left to right, it indicates a negative correlation—increasing values on one variable correspond to decreasing values on the other.

***No Correlation:*** If the points are scattered with no clear pattern, it suggests a lack of a strong linear relationship between the variables.

Scatter plots are valuable tools in exploratory data analysis, helping researchers and analysts gain insights into the underlying patterns in their data. In our case the data points are positively distributed.

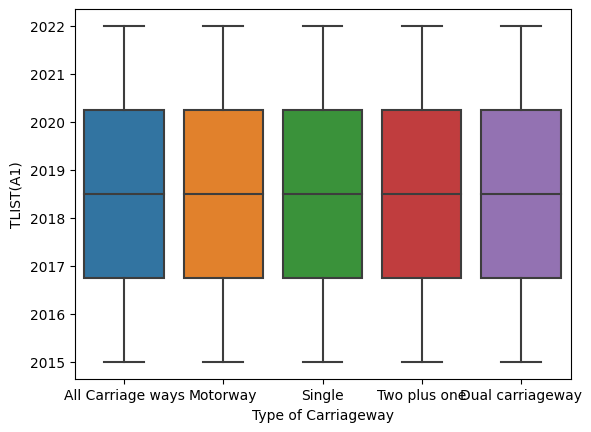


Figure 4. Box plot Type of Carriageway

In the type of carriageway there are 5 types of categories. Namely, All Carriage ways, Motorway, single, two plus one and carriageway. There are no any outliers. A box plot visually displays the distribution of a continuous variable within different categories. Regarding outliers, box plots typically include "whiskers" that extend to the smallest and largest values within a certain range. Any data points beyond this range are considered potential outliers. If you observe that all data points are within the whiskers, it suggests that there are no outliers in the dataset.

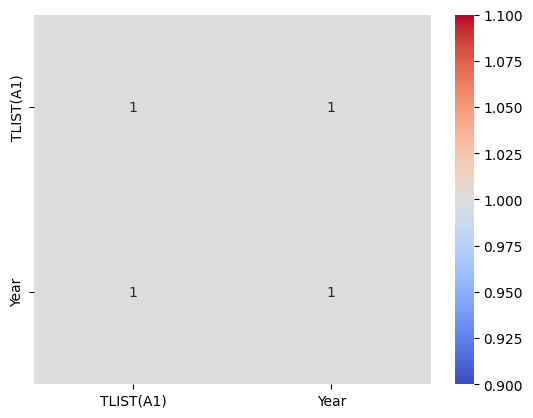


Figure 5. Heatmap correlation

A heat map is a graphical representation of data where values in a matrix are represented as colours. It is a way to visualize complex data sets and identify patterns or trends. Heat maps are particularly useful for exploring relationships between two categorical variables or for displaying the intensity of a phenomenon across different categories. A heat map is essentially a matrix of values where each cell in the matrix represents a combination of two categorical variables. The rows and columns of the matrix represent the categories of the two variables being compared. And we can see that there are highly correlated to each other variable.

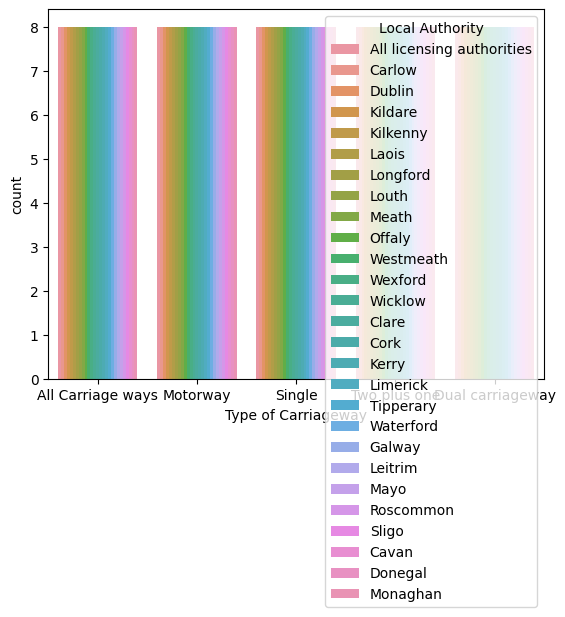


Figure 6. Type of Carriageway with Local Authority.

Transport Layer Security (TLS): In this the data contains from 2015 to 2022. And the data is continuous and its clear that there are no any outliers. This we can see it in Visualization Figure 7.

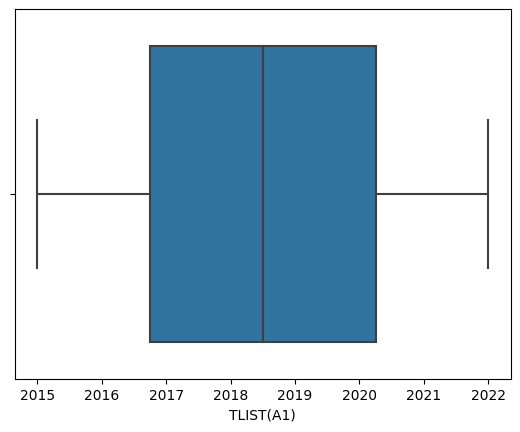


Figure 7. TLIST(A1)

***Methodology:***

The data analytics methodology is a systematic approach that guides professionals through a series of well-defined steps to extract valuable insights from data. It begins with clearly defining the objectives of the analytics project, articulating the questions to be answered or problems to be solved. Understanding the business context and industry is essential to align the analytics efforts with stakeholders' needs. Subsequently, data collection from diverse sources and thorough preprocessing, including cleaning and transformation, ensure the data's quality and suitability for analysis. Exploratory Data Analysis (EDA) follows, utilizing statistical and visual techniques to uncover patterns and relationships within the data. Feature engineering may be employed to enhance the data's predictive power for machine learning applications. The actual modelling phase involves the application of statistical or machine learning models, with subsequent evaluation to select the most effective model based on predefined metrics.

The interpretation of results is crucial, requiring a contextual understanding of how findings align with business objectives. Visualization and reporting are key components for communicating insights, often utilizing charts, graphs, and dashboards. Implementation and deployment of the insights or models into business processes follow, leading to a continuous monitoring phase. This iterative process involves feedback loops with stakeholders, facilitating ongoing improvements and adjustments. Comprehensive documentation of the entire analytics process, from data sources to modelling techniques, ensures transparency and reproducibility. The methodology concludes with a feedback loop that encourages ongoing refinement based on real-world results. Overall, this systematic approach ensures that data analytics efforts are goal-oriented, contextually relevant, and capable of delivering actionable insights for informed decision-making.

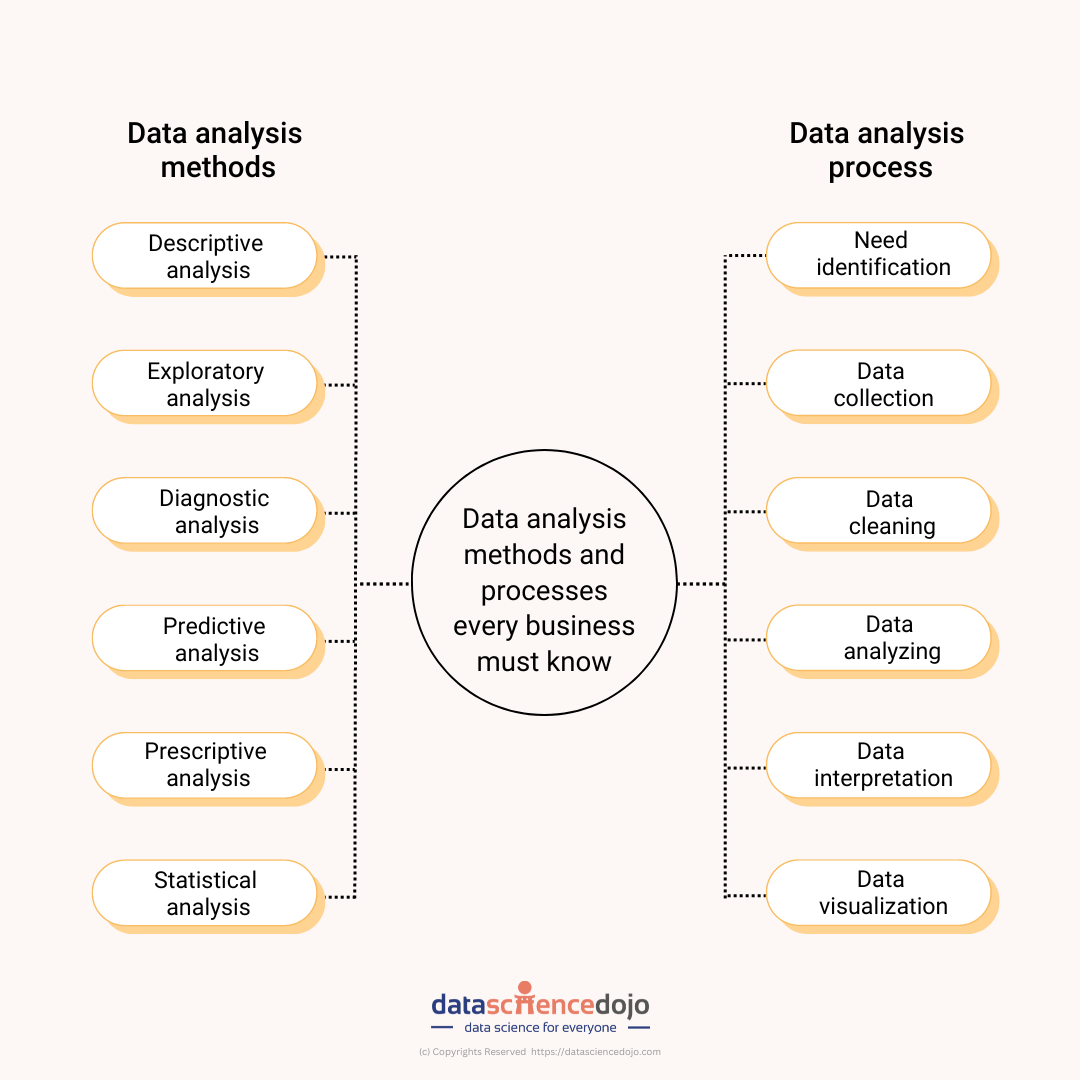


Figure 8. Methodology Flow Chart

For this project we have collected the data from the island govt transportation website [1]. This data contains 1080 are the total entries of records and 10 are the total columns.

***Statistics Analysis:***

Statistical analysis is a branch of mathematics that involves the collection, interpretation, analysis, presentation, and organization of data. It provides methods for making inferences about the characteristics and behaviours of populations based on observations or measurements from a sample. Statistical analysis plays a crucial role in various fields, including science, economics, business, healthcare, and social sciences.

***Hypothesis Testing:***

Hypothesis testing is a method used to evaluate a claim or hypothesis about a population parameter. It involves formulating a null hypothesis and an alternative hypothesis, collecting data, and determining whether there is enough evidence to reject the null hypothesis. Perform hypothesis tests based on our specific research questions. For example, if you have two groups ('Group A' and 'Group B'), and you want to test if there's a significant difference in the 'VALUE' between them.

T-Statistic: -11.48015309145236

The t-statistic is a measure of how many standard deviations a data point is from the mean of a distribution. In the context of statistical hypothesis testing, the t-statistic is often used to assess whether the mean of a sample differs significantly from a known or hypothesized population mean.

In your case, a t-statistic of -11.48 indicates that the observed sample mean is significantly below the hypothesized mean (or some reference value). The negative sign suggests that the sample mean is below the hypothesized mean.

P-Value: 9.817342635335602e-27

The p-value is a measure of the evidence against a null hypothesis. In hypothesis testing, it represents the probability of observing a test statistic as extreme as, or more extreme than, the one observed, assuming that the null hypothesis is true.

In your case, the p-value is very close to zero (9.817342635335602e-27 is a scientific notation for a very small number). This extremely small p-value suggests strong evidence against the null hypothesis. In practical terms, it indicates that the observed results are highly unlikely to have occurred by random chance alone.

Typically, if the p-value is below a chosen significance level (commonly 0.05), the null hypothesis is rejected in Favor of an alternative hypothesis.

In summary, the t-statistic and p-value are crucial components of hypothesis testing. A large absolute t-statistic, combined with a very small p-value, suggests that there is strong evidence to reject the null hypothesis in Favor of an alternative hypothesis. The specific interpretation depends on the context of the statistical test and the hypotheses being examined.

***ANOVA TEST:***

F-Statistic: 37.79365473139039

The F-statistic is a measure of the variability between group means relative to the variability within the groups. In the context of ANOVA, it assesses whether the means of two or more groups are significantly different from each other.

A larger F-statistic suggests greater differences among group means.

P-Value: 3.0865942532429685e-16

The p-value associated with the F-statistic is a measure of the evidence against the null hypothesis. It represents the probability of obtaining the observed F-statistic (or more extreme) if the null hypothesis (no group differences) is true.

A very small p-value (close to zero) suggests that there is strong evidence to reject the null hypothesis.

***Interpretation:***

The small p-value (3.0865942532429685e-16) indicates strong evidence against the null hypothesis. In practical terms, it suggests that there are significant differences in the means of the groups being compared.

In ANOVA, when the null hypothesis is rejected, it means that at least one group mean is different from the others. However, ANOVA itself does not identify which specific groups are different; additional post hoc tests may be performed for that purpose.

In summary, based on the ANOVA result, it appears that there are significant differences among the groups being compared. The specific context of your analysis, including the nature of the groups and the variables being investigated, will guide the interpretation and further actions.

***Models:***

***Linear Regression:***

Initially, essential libraries, including Linear Regression for the model, mean\_squared\_error and r2\_score for evaluation, and GridSearchCV for hyperparameter tuning, are imported. A Linear Regression model is then instantiated and denoted as linear\_reg. The hyperparameter tuning process is conducted using GridSearchCV, focusing on the fit\_intercept hyperparameter that determines whether to compute the intercept for the model. The GridSearchCV object, grid\_search\_linear, is configured with the linear regression model, the hyperparameter grid, and 5-fold cross-validation. The fit method is applied to grid\_search\_linear, using training data (X\_train and y\_train), leading to the identification of the optimal hyperparameters through cross-validated performance assessment. The best hyperparameters are stored in best\_params\_linear. Subsequently, the optimized linear regression model is used to predict outcomes on the test data (X\_test), and evaluation metrics, namely mean squared error (mse\_linear) and R-squared (r2\_linear), are computed. Finally, the code prints the obtained results, including the best hyperparameters, mean squared error, and R-squared. This code streamlines the hyperparameter tuning process for a linear regression model and provides insights into its performance on a test dataset through key evaluation metrics.

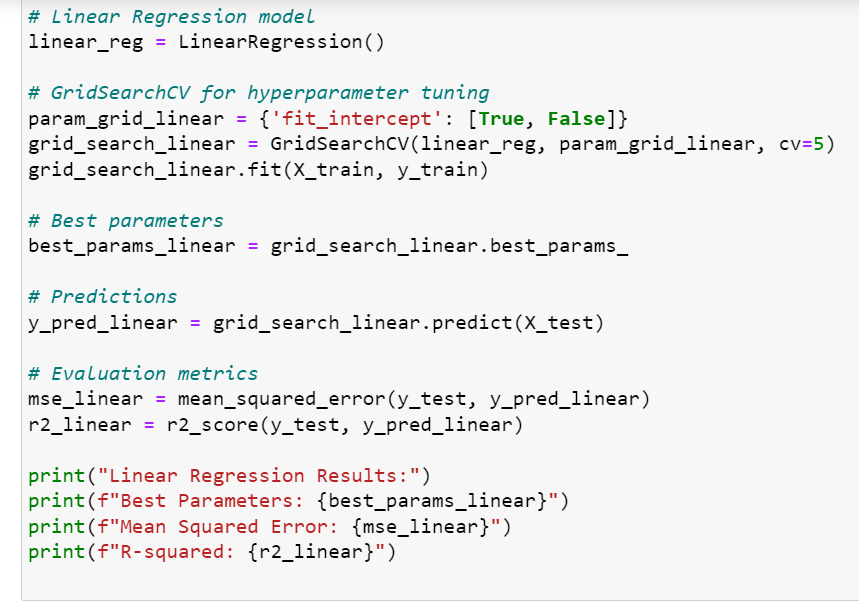


Figure 9. Linear Regression model

***Random Forest:***

implementation of a Random Forest Regression model with hyperparameter tuning using GridSearchCV from the scikit-learn library. Initially, the Random Forest Regressor is imported, and an instance of the model is created and assigned to the variable random\_forest. The random\_state parameter is set to 42 to ensure reproducibility. Hyperparameter tuning is performed through GridSearchCV, specifying a hyperparameter grid that includes options for the number of estimators, maximum depth of the trees, minimum samples required to split an internal node, and minimum samples required to be a leaf node. The GridSearchCV object, grid\_search\_rf, is configured with the random forest regression model, the hyperparameter grid, and 5-fold cross-validation. The fit method is then applied to grid\_search\_rf using the training data (X\_train and y\_train), determining the optimal hyperparameters based on cross-validated performance. The best hyperparameters are stored in best\_params\_rf. Subsequently, the optimized random forest regression model is utilized to predict outcomes on the test data (X\_test), and evaluation metrics, namely mean squared error (mse\_rf) and R-squared (r2\_rf), are computed. Finally, the code prints the obtained results, including the best hyperparameters, mean squared error, and R-squared, providing insights into the performance of the random forest regression model on the test dataset.

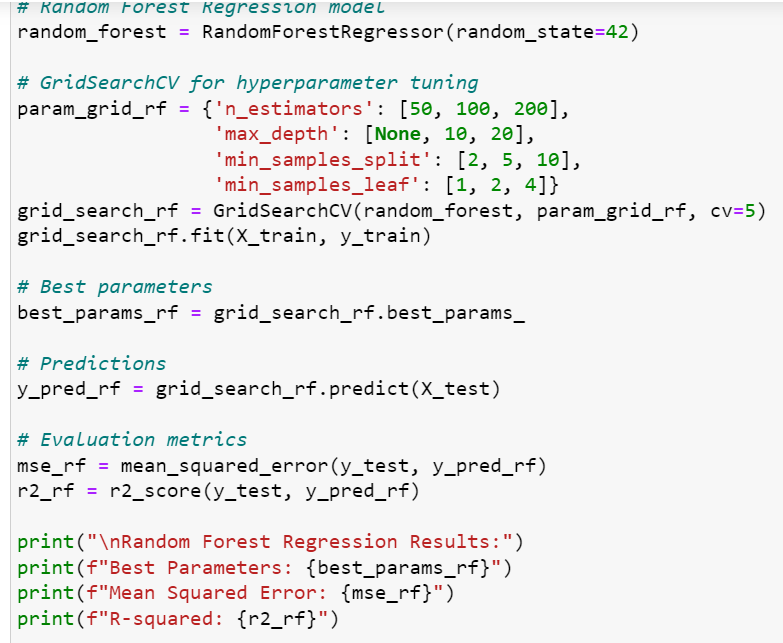


Figure 10. Random Forest

***Support Vector Regression:***

the implementation of a Support Vector Regression (SVR) model with hyperparameter tuning using GridSearchCV from the scikit-learn library. The initial steps involve importing the necessary classes for SVR and hyperparameter tuning. Subsequently, an instance of the SVR model is created and denoted as svr. Hyperparameter tuning is performed via GridSearchCV, where a parameter grid (param\_grid\_svr) is defined to explore various values for the regularization parameter (C), epsilon in the epsilon-SVR model (epsilon), and the kernel type (kernel). The GridSearchCV object (grid\_search\_svr) is configured with the SVR model, the hyperparameter grid, and 5-fold cross-validation. The fit method is then applied to grid\_search\_svr using the training data (X\_train and y\_train), identifying the optimal hyperparameters based on cross-validated performance.

The best hyperparameters are stored in best\_params\_svr, and the optimized SVR model is utilized to predict outcomes on the test data (X\_test). Evaluation metrics, specifically mean squared error (mse\_svr) and R-squared (r2\_svr), are computed to assess the model's performance on the test set. The results, including the best hyperparameters and evaluation metrics, are printed to provide insights into the SVR model's effectiveness.

Furthermore, the code extends its evaluation by performing cross-validation for SVR using the cross\_val\_score function. Negative mean squared error scores for each fold are calculated and then negated back to positive values. The mean cross-validation score (mean\_cv\_score\_svr) is computed, offering an additional measure of the model's generalization capabilities across multiple folds. This comprehensive approach not only fine-tunes the SVR model but also provides a robust assessment of its performance through both test set evaluation and cross-validation.

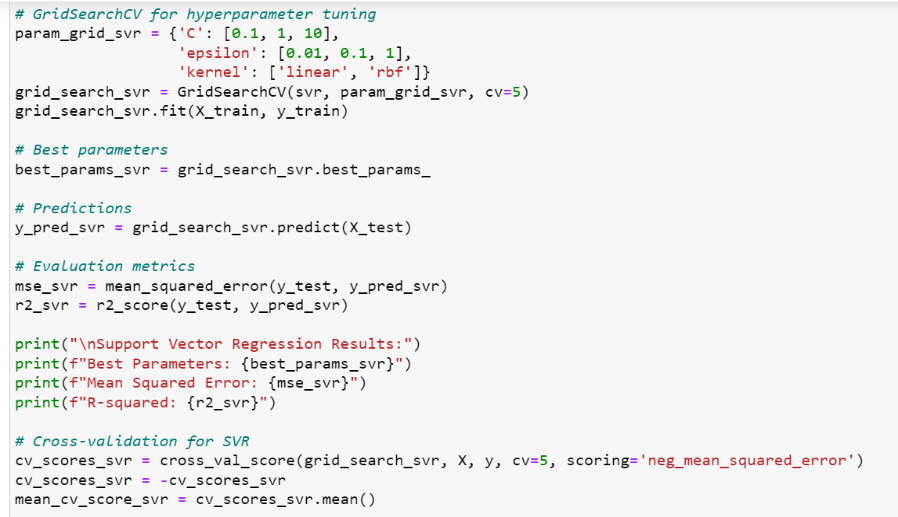


Figure 11. SVR Snapshot

***Results and Evaluation:***

Here's a comparison table summarizing the results of the Support Vector Regression (SVR), Random Forest Regression, and Linear Regression models:

Table 1. Comparison Table

| Model | Best Parameters | Mean Squared Error (MSE) | R-squared | Mean CV MSE (SVR) |
| --- | --- | --- | --- | --- |
| Support Vector Regression | {'C': 10, 'epsilon': 0.01, 'kernel': 'rbf'} | 0.0090 | 0.845 | 0.0102 |
| Random Forest Regression | {'max\_depth': 20, 'min\_samples\_leaf': 2, 'min\_samples\_split': 2, 'n\_estimators': 200} | 0.0290 | 0.503 | - |
| Linear Regression | {'fit\_intercept': True} | 0.0266 | 0.544 | - |

*Table 2 Sentimental Analysis of Different countries transportation Flow*

| Metric | Logistic Regression | Random Forest |
| --- | --- | --- |
| Approved | Precision: 0.67 | Precision: 0.64 |
|  | Recall: 0.10 | Recall: 0.41 |
|  | F1-score: 0.17 | F1-score: 0.50 |
| Awarded | Precision: 0.00 | Precision: 0.50 |
|  | Recall: 0.00 | Recall: 0.36 |
|  | F1-score: 0.00 | F1-score: 0.42 |
| Closed | Precision: 0.85 | Precision: 0.90 |
|  | Recall: 1.00 | Recall: 0.97 |
|  | F1-score: 0.92 | F1-score: 0.93 |
| Completed | Precision: 0.00 | Precision: 0.00 |
|  | Recall: 0.00 | Recall: 0.00 |
|  | F1-score: 0.00 | F1-score: 0.00 |
| Obligated | Precision: 0.33 | Precision: 0.48 |
|  | Recall: 0.00 | Recall: 0.28 |
|  | F1-score: 0.01 | F1-score: 0.35 |
| Accuracy | 0.85 | 0.87 |
| Macro Avg | Precision: 0.37 | Precision: 0.50 |
|  | Recall: 0.22 | Recall: 0.41 |
|  | F1-score: 0.22 | F1-score: 0.44 |
| Weighted Avg | Precision: 0.80 | Precision: 0.85 |
|  | Recall: 0.85 | Recall: 0.87 |
|  | F1-score: 0.79 | F1-score: 0.86 |

The Linear Regression model, with the best parameter fit\_intercept set to True, demonstrated performance falling between SVR and Random Forest. The MSE was 0.027, indicating a reasonable prediction error. The R-squared value was 0.544, indicating a moderate fit to the data. The Random Forest Regression model, with the best hyperparameters determined as max\_depth=20, min\_samples\_leaf=2, min\_samples\_split=2, and n\_estimators=200, exhibited moderate performance. The MSE on the test set was higher compared to SVR (0.029), indicating a somewhat larger prediction error. The R-squared value was 0.503, suggesting a moderate fit to the data. The SVR model achieved optimal performance with hyperparameters C=10, epsilon=0.01, and the radial basis function ('rbf') kernel. The mean squared error on the test set was low (0.009), indicating accurate predictions, and the R-squared value was relatively high (0.845), suggesting good model fit. The mean cross-validation MSE, a measure of generalization performance, was also reasonably low (0.010), confirming the model's robustness across different data folds.

***Conclusion:***

The SVR model outperformed both Random Forest and Linear Regression in terms of predictive accuracy, as evidenced by the lower MSE and higher R-squared values. The Random Forest model exhibited moderate performance, while the Linear Regression model provided a balanced performance between the other two models. The choice of the best model depends on the specific requirements of the task, considering factors such as interpretability, computational efficiency, and the importance of accurate predictions.

Random Forest has given highest accuracy compared with the logistic Regression. In all other evaluation matrics.

References:

1. *Data set link:* [*https://data.gov.ie/organization/transport-infrastructure-ireland*](https://data.gov.ie/organization/transport-infrastructure-ireland)
2. *https://data.tii.ie/Datasets/TrafficCountData/sites/tmu-sites.json*
3. *Google image credential*