PREDICTION OF AVIATION NON-VOLATILE PARTICULATE MATTER EMISSIONS USING MACHINE LEARNING

RAHUL BALASAHEB MOGAL

(Student No- C22081745)

October 2023

School of Mathematics,

Cardiff University

A dissertation submitted in partial fulfilment of the requirements for **MSc Data Science and Analytics** by taught programme, supervised by **Dr Andrew Crayford.**

Executive Summary

In today's rapidly evolving aviation sector, recognizing and comprehending the ecological consequences of aircraft engine emissions is of outmost significance. Our study delves deep into the specifics of non-volatile particulate matter (nvPM) emissions, bridging the gap between engine operations, environmental conditions, and the emissions they produce.

Our analytical work started with the integration of two foundational datasets(*ICAO Aircraft Engine Emissions Databank*) one detailing gaseous emissions and the other centered on nvPM emissions. By meticulously combining these datasets, we curated a unified dataset that was broad in scope and profound in its details. This fusion of data provided a holistic view of the different engine operational parameters and the associated emission outputs, priming us for further investigative endeavors. Recognizing that data quality is the bedrock of any analytical study, we embarked on a thorough data-cleaning process. We streamlined the dataset by removing redundant columns and addressing missing data points with appropriate techniques, preserving the data's authenticity. Leveraging both specialized domain expertise and insights extracted directly from the data, we pinpointed specific attributes for predictive modeling, ensuring our models were fed with the most influential predictors.

Our visual analytics, ranging from intricate boxplots to informative scatter plots, unveiled the underlying structure and nuances of our data. These graphical representations not only highlighted central data trends and variability but also identified potential anomalies. This in-depth visual analysis set the stage for our subsequent modeling endeavors, promoting an informed and strategic approach.

Our modeling phase was an eclectic mix of statistical and machine-learning techniques. After establishing a baseline with the multitarget regression model, we expanded our horizons by experimenting with decision trees and their ensemble counterparts like random forests and gradient-boosting regressions. Each of these models, equipped with their own unique strengths, unraveled subtle patterns and relationships in our data. In addition, we tapped into the K-Nearest Neighbors model, reinforcing the principle that diverse modeling approaches often lead to more robust insights.

Each model was rigorously validated using an array of performance metrics, emphasizing not just in-sample fit but also out-of-sample predictive prowess. The Gradient Boosting Regressor, renowned for its iterative error reduction, emerged as a particularly potent tool, showcasing its capability to handle complex analytical challenges.

Our research offers an analytical deep dive into prediction of nvPM emissions, unveiling the intricate interplay between engine operations, ambient conditions, and the resulting emissions. The insights harvested from this study are poised to shape aviation policies, inspire innovations in engine design, and play a pivotal role in the global pursuit of sustainable aviation.

Abstract

It is crucial to understand the intricate web of aviation emissions in an era of growing environmental concerns and increased public health awareness. This investigative report undertakes a granular analysis of aviation emissions, shedding light on their multifaceted repercussions on both ecological systems and human well-being.

Leveraging cutting-edge data analytics methodologies paired with the latest advancements in machine learning, our study dives deep into aviation emissions, unraveling inherent patterns and dynamics. But our exploration doesn't stop at mere pattern recognition; we venture further, aiming to pinpoint the origins and root causes of these emissions. This encompassing approach is instrumental in devising informed strategies that can effectively mitigate the negative impacts of these emissions.

Central to our research is the significance of predictive modeling. As our findings underscore, understanding emission data is a complex endeavor, rife with intricate variations and interdependencies. Accurate, dependable models, as demonstrated in our study, transcend academic significance. They possess the transformative potential to sculpt sustainable aviation paradigms for the future, ensuring that the aviation industry aligns with global sustainability objectives.

Table of Contents

List Of Figures

List of Tables

1. Introduction

The amazing achievements of aviation, which represent the desire of humans to overcome geographical barriers and the unwavering drive for progress, have consistently fascinated our collective imagination. As time progresses, the sector has expanded its reach to encompass more extensive geographical areas, facilitating connections between distant regions and promoting global cohesion. Nevertheless, this considerable expansion highlights a pressing issue: the ecological impact of our increasingly active airspace. As society witnesses the rising presence of airplanes traversing the skies, it becomes imperative to confront the environmental implications that accompany their operations.

At the core of this environmental issue lies a specific type of pollution known as non-volatile particulate matter (nvPM)(Ge et al. 2022). Although their physical dimensions may be small, the consequences they carry are immense. When these minuscule particles are discharged into the Earth's atmosphere, they can exert significant impacts on the whole climate system and, notably, on human well-being. Due to their dimensions and inherent characteristics, these particles have the ability to persist in the atmosphere for extended durations, serving as catalysts in the process of cloud formation and conceivably exerting an influence on meteorological phenomena. Moreover, upon inhalation, these substances present potential health hazards, particularly to the respiratory system.

Aircraft engines, which are remarkable feats of engineering, consist of a complex network of components that operate in synchronization. As the combustion process occurs and the fuel interacts with the surrounding atmosphere, it results in the generation of a diverse range of pollutants. One notable emission that stands out is nvPM(non-volatile particulate matter(Quadros et al. 2022). The measurement and prediction of this phenomenon provide significant difficulty, mostly attributed to its minuscule scale and the multitude of elements that influence its generation. Given the projected increase in air travel and the introduction of more congested flight paths, it is crucial to comprehend and address the issue of non-volatile particulate matter (nvPM) emission. Can the complicated nature of non-volatile particulate matter (nvPM) emissions be understood and predicted by looking closely at how engines work and how they interact with outside environmental factors? This work is the result of a lot of hard work. This endeavor goes beyond the realm of academia. If our research is successful, the aviation sector may be provided with predictive tools that would enable them to anticipate and potentially mitigate emissions prior to the commencement of an aircraft's flight.

We started our study by building on a large dataset(*ICAO Aircraft Engine Emissions Databank*), which is a collection of data about engine states, environmental factors, and the non-volatile particulate matter (nvPM) emissions that happen as a result. Utilizing contemporary data science and machine learning methodologies, our objective was to uncover

intricate patterns, correlations, and valuable insights that may evade traditional analytical approaches. Through the utilization of influential research in the respective discipline, our inquiry thoroughly deconstructs various techniques, employs cutting-edge analytical tools, and unveils noteworthy findings. This study demonstrates our unwavering dedication to leveraging the potential of data science to craft a more sustainable future for the aviation sector.

There are numerous ramifications of our findings for the aviation industry. The predictive capabilities of our model can play a fundamental role in defining emission control strategies. By relying on robust data, these policies can effectively target the most significant sources of emissions. Moreover, the findings derived from our model can serve as a guiding tool for research and development teams, stimulating technological advancements that target engine characteristics that have the most significant impact on non-volatile particulate matter (nvPM) emissions. In an era characterized by heightened awareness within the aviation industry regarding its environmental impact, the forecasts and perspectives we offer serve as illuminating beacons, emphasizing the importance of employing evidence-based approaches to navigate a path towards a more ecologically sustainable aviation sector. The present study showcases a comprehensive approach that incorporates many modeling methodologies, with particular emphasis on the Gradient Boosting Regressor. This approach exemplifies the significance of ongoing refinement and scientific rigor in addressing complex issues such as nvPM emissions. Our objective is that this research not only contributes to the advancement of scientific knowledge, but also fosters a narrative that prioritizes sustainable aviation.

2. Background Research

The aviation industry, renowned for its role in contemporary transportation and as a tribute to human ingenuity, has emerged as a significant catalyst for the process of globalization. The aviation industry has effectively responded to the increasing demands of globalization by enabling efficient transportation of individuals and goods, thereby facilitating the quick movement of people and commodities around the globe. However, as technology continues to achieve unprecedented levels of success, it inadvertently poses a significant threat to the natural environment, prompting critical deliberations over the consequences of its emissions.

Throughout history, the emissions generated by the aviation industry have exerted a significant influence on the worldwide environment.

The emissions encompass a variety of pollutants, such as Hydrocarbon (HC) ,carbon dioxide (CO2), nitrogen oxides (NOx), particulate matter, and other substances. Although carbon dioxide (CO2) is a well-known and significant contributor to the greenhouse effect, it is important to recognize that other pollutants, like nitrogen oxides (NOx), also play a crucial role in this phenomenon. While the impact of NOx emissions may be less apparent, they have an equally detrimental effect by contributing to the production of ozone at higher altitudes during aircraft cruising. The presence of generated ozone not only contributes to the greenhouse effect but also has implications for air quality at lower altitudes, hence exerting an influence on public health(Emami et al. 2018).

The environmental impact of the aviation industry extends beyond the atmosphere. Ground-based aviation activities like taxiing, takeoff, and landing have a significant impact on an airport's emissions profile. In light of these complex problems, it is evident that there exists an imperative requirement for accurate and all-encompassing data pertaining to emissions within the aviation sector. The comprehensive collection and rigorous analysis of such data can play a pivotal role in informing and guiding policy decisions that have the potential to bring about dramatic changes, steering the industry towards a more sustainable trajectory.

Black carbon (BC) is a significant pollutant among the many emissions originating from airplane engines(Teoh et al. 2019). This particular pollutant not only has a significant impact on the phenomenon of global warming as a result of its capacity to absorb heat, but it also plays a role in the development of contrails. The presence of heightened concentrations of fine particulate matter in the atmosphere, which can be attributed to the aviation sector, has been unequivocally associated with health implications, such as an elevated susceptibility to premature mortality.

Based on an analysis point of view, recent progress in predictive modeling, specifically the use of methods like convolutional neural networks (CNNs), has shown promise in estimating

aircraft emissions(Ge et al. 2022). Using models that have been trained on large datasets is a key part of getting useful information about emissions like non-volatile particulate matter (nvPM), which in turn helps people come up with good ways to reduce their effects. One tool that has attracted much attention is the Smoke Correlation for Particle Emissions CAEP11 (SCOPE11)(Agarwal et al. 2019). The estimation of BC mass and number emissions has been a primary emphasis of SCOPE 11, rendering it an invaluable instrument that provides a comprehensive understanding of BC emissions. This detailed perspective is essential for academics and policymakers alike.

Given the estimates indicating substantial growth in the worldwide aviation fleet by the year 2036, it becomes increasingly imperative to comprehend and address the issue of aviation emissions in a proactive manner. The emissions originating from aircraft engines, particularly those in close proximity to airports, have the capacity to significantly increase concentrations of particulate matter (PM) and ozone, hence yielding noteworthy environmental and health consequences.

The SCOPE11 methodology provides a systematic approach for estimating emissions from aircraft engines, seeing them as notable sources of particulate matter with black carbon as a key component. A link was found between the smoke number (SN) and the mass concentration of black carbon (BC) by SCOPE11. This shows that rigorous scientific methods can be used in real life. This correlation is based on comprehensive data collected from several aviation engine models.

When comparing machine learning models to SCOPE11 and similar empirical methods, it is evident that they possess a combination of distinct advantages. The deterministic nature of SCOPE11 ensures that it is clear, but the adaptive nature of machine learning makes it easier to find small patterns in data, especially when correlations are not necessarily linear. Achieving a balance between the two, however, is a challenging endeavor. The particular goals and constraints of a given project frequently determine the choice of either one option or a combination of both.

3. Methodology and Approach

3.1Data Integration and Exploration.

The foundation of our analytical project was established by obtaining robust and comprehensive data. We sourced our data from the EASA ICAO databank(*ICAO Aircraft Engine Emissions Databank* [no date]), a reputable reservoir of aviation-related emission statistics, ensuring the credibility and relevance of our initial dataset. To fully grasp a subject, it's vital to have a complete set of data that covers all its aspects. With this in mind, our first task was to bring together data from two different sections, one about gaseous emissions and the other about nvPM emissions. We combined these datasets using the 'UID no' as a common link, ensuring a smooth merge.

Using the data given in the CSV files, we started our work by merging these different datasets into one. We matched the data using the UID No Detail, creating a full set of information about emission figures and related details. We then took a deep dive into this data, calculating basic statistics and creating visuals to understand its makeup, patterns, and connections. This first look at the data helped shape the next steps in our study.

3.2Exploratory Data Analysis.

In our EDA, we are determined to delve deep into different types of emissions, both gaseous and particulate (nvPM). By visualizing these emissions, we can identify and analyze patterns, trends, and anomalies more effectively.

3.2.1 Gaseous Emissions: Using a combination of bar plots and line graphs, we'll illustrate the magnitude of different gaseous emissions across various flight phases. By segmenting the emissions based on flight phases, we aim to pinpoint the phases that are especially emission-intensive. This segmentation offers clarity on where interventions might be most effective.

3.2.2 Non-volatile Particulate Matter (nvPM) Emissions: Combination of bar charts & line charts has employed to represent nvPM emissions across flight phases. With the aviation industry's increasing focus on particulate emissions, understanding the nvPM landscape across flight phases is crucial. Such a visualization will shed light on the most critical phases regarding particulate emissions.

To not just understand the magnitude but also the relationships between different attributes, regression plots can be invaluable.

3.2.3 Regression Plots for Emissions vs. Engine Parameters: We'll use scatter plots with a regression line to visualize how emissions relate to different engine parameters like thrust and fuel flow. These plots will allow us to discern if there's a linear relationship between emissions and engine parameters. The presence or absence of such a relationship can guide engine design and operational strategies.

3.3.4 Correlation Heatmaps: To get an overall sense of how different variables (including emissions and engine parameters) correlate with each other, we'll employ heatmaps. Heatmaps can provide a quick, visual representation of the relationships between multiple variables simultaneously. This is particularly useful to identify pairs of variables that have strong positive or negative correlations, guiding further detailed analysis.

3.2.5 Detailed Scatter Plots for Identified Strong Correlations: For pairs of variables that exhibit strong correlations in the heatmap, we'll create detailed scatter plots. These scatter plots will allow a granular look at the nature of the relationships between variables, identifying potential causative factors or outliers that might be influencing the relationship.

3.3Data Preprocessing.

The efficacy of machine learning models is heavily contingent upon the quality and structure of the data they're trained on. Given this, refining our dataset becomes paramount to ensuring models can derive accurate and meaningful insights

3.3.1 Handling Missing Values: Data often comes with gaps or missing values which can distort the model's perception and lead to biased outcomes. Utilizing our domain expertise, we confronted these gaps. Where meaningful, we can impute missing values using the median, as it's less sensitive to outliers. In instances where data absence was significant, strategic omissions were made to maintain data quality. Addressing missing values ensures a comprehensive and representative dataset, minimizing potential bias in our models.

3.3.2 Feature Engineering:

Feature engineering serves as the bridge between raw data and effective machine learning models.

While raw datasets provide a wealth of information, they often don't capture the intricate relationships, patterns, or nuances inherent in the data. This is where domain knowledge becomes valuable. By leveraging expertise and understanding of the specific field, in this case, engine emissions and aerodynamics, one can derive new features or attributes that might

otherwise remain hidden. Such derived features can unveil latent patterns, trends, and relationships that are critical for predictive accuracy.

For instance, in the context of predicting engine emissions, domain experts might know the subtle effects of specific engine parameters on emissions under varying conditions. They might be aware of how different combustion techniques or fuel types can influence particulate emissions, or how certain atmospheric conditions can amplify or mitigate emission levels. By incorporating this domain knowledge into feature engineering, one can create attributes that capture these nuanced relationships, making them explicit and accessible to machine learning models.

3.3.3 Data Encoding: Machine learning algorithms require numerical input. However, our dataset contained categorical variables which are non-numerical. We employed label encoding to transform these categorical variables into a format compatible with machine learning algorithms, ensuring the essence of the data remained intact. This transformation is crucial for the applicability of many machine learning algorithms, ensuring they can process the data effectively.

3.3.4 Outlier Detection: Outliers or extreme values can disproportionately influence model training, leading to models that are not generalizable. We utilized visual tools, such as boxplots, coupled with statistical methods to identify potential outliers. Once flagged, these outliers were then handled appropriately, either through adjustment or removal. Addressing outliers ensures our models learn from a representative dataset and aren't unduly influenced by extreme values.

3.3.5 Data Splitting: To assess the performance of our models, we need to test them on unseen data. The dataset was divided into training and testing subsets. The training subset is used to train the models, while the testing subset is reserved to validate their performance. This step is vital for evaluating the robustness and generalizability of our models. Training and testing on separate data subsets ensures we get a genuine measure of a model's predictive capabilities.

3.4Predictive Modeling

Equipped with a refined dataset, our journey ventured into the heart of predictive analytics. Harnessing the power of diverse regression models, our objective was to discern patterns and make accurate predictions.

3.4.1 Model selection:

Linear regression:

It is a fundamental tool in predictive modeling, particularly when we assume that relationships between variables are linear(*Linear Regression in Machine learning - Javatpoint*). When it comes to predicting nvPM (non-volatile particulate matter) emissions from aircraft engines, linear regression becomes valuable by helping us establish connections between different engine parameters (predictors) and emission levels

In simple terms, the mathematical formula for simple linear regression, which involves just one predictor, looks like this:

$Y = \beta 0 + \beta 1X + \epsilon$

Here's what these symbols mean:

- Y represents the nvPM emission level (the thing we want to predict).
- X is the independent variable or predictor (it could be any engine parameter).
- $\beta 0$ is the y-intercept, the point where the regression line intersects the Y-axis.
- $\beta 1$ is the coefficient associated with X, indicating how much Y changes for a one-unit change in X.
- ϵ is the error term, which accounts for the differences between the actual emissions and the emissions predicted by the model.

The goal of linear regression is to find the values of $\beta 0$ and $\beta 1$ that minimize the sum of squared differences (errors) between the observed emission levels and the levels predicted by the model.

When we have multiple engine parameters as predictors, the linear regression model extends to:

$$Y = \beta 0 + \beta 1X1 + \beta 2X2 + \cdots + \beta nXn + \epsilon$$

In this situation:

- Each Xi represents a distinct engine parameter.

- βi represents the coefficient for the corresponding Xi, telling us how Y (nvPM emissions) changes when we change Xi by one unit while keeping all other predictors constant.

These coefficients (β values) are essential for understanding how individual engine parameters affect nvPM emissions. Typically, we calculate these coefficients using methods like the Least Squares technique, which focuses on minimizing the sum of squared differences between the actual and predicted emission values.

In our efforts to predict nvPM emissions, utilizing linear regression provides a straightforward yet insightful model, especially when the relationships between engine parameters and emissions follow linear patterns.

Tree-Based Regression

Tree-based regression techniques can naturally capture non-linear relationships and variable interactions without specific model alterations, making them apt for predicting nvPM emissions from aircraft engines, where relationships might be non-linear or hierarchical. At their core, these methods recursively divide data into uniform subsets based on feature values until a set criterion is achieved. The tree structure represents internal nodes as decisions on feature values, branches as decision outcomes, and leaf nodes as prediction values. Their strengths in predicting nvPM emissions encompass handling non-linearities, modeling variable interactions, and offering visual interpretability.

Given the complex interplay of aircraft engine parameters, trees can elucidate intricate parameter interactions and rank features by importance, shedding light on influential engine parameters. They also intuitively model hierarchical decisions where one parameter's influence may be conditional on another. Enhancing predictive accuracy, ensemble techniques like Random Forest Regressor(*Machine Learning Random Forest Algorithm - Javatpoint*) and Gradient Boosting Machines (GBM) can be employed. The former averages predictions from multiple trees trained on random data subsets, while the latter builds trees sequentially to rectify preceding errors. In the context of nvPM emissions prediction, tree-based models, with their capacity to unravel non-linearities and interactions, coupled with ensemble methods, offer a holistic, interpretable, and robust predictive framework.

K-Nearest Neighbors (KNN)

K-Nearest Neighbors (KNN) is a non-parametric, instance-based learning algorithm that predicts based on the values of its 'k' nearest data points in the feature space. Its simplicity and intuitiveness make KNN a favored choice for various predictive tasks, including the estimation of nvPM emissions from aircraft engines. The mechanism of KNN involves identifying 'k' closest data points from the training set to a given prediction point(*K-Nearest Neighbor(KNN) Algorithm* 2017). For regression tasks, such as estimating nvPM emissions, the prediction usually constitutes the mean of the target values of these 'k' neighbors.

When predicting nvPM emissions, KNN's adaptability stands out as it doesn't assume any specific functional form, allowing it to adjust to non-linear relationships between engine parameters and emissions. Its local decision-making capability ensures it captures nuanced data patterns that might elude global models, and it naturally addresses feature interactions without necessitating explicit model changes. However, there are challenges to consider. KNN is notably sensitive to feature scales, necessitating proper normalization or standardization, especially given the varied nature of engine parameters. With extensive datasets, determining the 'k' nearest neighbors can be computationally taxing. Moreover, the choice of 'k' holds considerable weight in predictions, with cross-validation often employed to determine the optimal value.

Support Vector Machines (SVM)

Support Vector Machines (SVM) is a supervised learning algorithm primarily used for classification and regression. For predicting nvPM emissions, the focus is on Support Vector Regression (SVR), the regression variant of SVM(*Machine Learning Algorithm Explained* 2020). SVR works by identifying a hyperplane in a higher-dimensional space that best fits the data, ensuring that most data points lie within an ε (epsilon) margin from this hyperplane. Points outside this margin are labeled support vectors, and the model aims to limit their deviations.

SVR's strengths in predicting nvPM emissions include its ability to handle non-linear relationships by using kernel functions to project data into higher-dimensional spaces. Furthermore, it is robust, less sensitive to outliers, and its regularization parameter aids in preventing overfitting. However, challenges arise in parameter tuning, such as selecting the right kernel and adjusting parameters like C and ε. Optimal values are often found using cross-validation. Additionally, SVM can be computationally intensive for large datasets, particularly with non-linear kernels, and its interpretability may be lower than more transparent models like decision trees. In the realm of nvPM emissions prediction, both KNN and SVM have their merits. While KNN excels at capturing local patterns, SVM, with the right kernel functions, can decipher intricate, non-linear relationships. Utilizing both can yield a holistic understanding and precise predictions of nvPM emissions based on various engine parameters.

3.4.2 Ensemble Techniques:

Ensemble techniques leverage the power of multiple models to deliver superior predictive performance compared to individual models. By aggregating the predictions of multiple models, ensemble methods often achieve better generalization and robustness to noise and outliers. In the context of predicting nvPM emissions from aircraft engines, ensemble techniques like Random Forests and Gradient Boosting Regressors play a pivotal role(Singh 2018).

The implications of our study in non-volatile particulate matter (nvPM) emission prediction are far-reaching. The utilization of ensemble approaches enhances forecast accuracy by leveraging the combined strengths of varied models, while simultaneously reducing the intrinsic unpredictability typically observed in individual models. The decrease in variability serves to enhance the stability of predictions, hence reducing the impact of idiosyncrasies inherent in different models. In addition to ensuring accuracy and stability, these strategies provide us with a holistic understanding of the data. The panoramic picture presented in this analysis captures comprehensive and overarching trends, while also identifying subtle and nuanced patterns, so facilitating a comprehensive and multifaceted comprehension of non-volatile particulate matter (nvPM) emissions.

3.4.3 Empirical Modeling:

Beyond tree-based models, Empirical models, which are grounded in observational data rather than theoretical frameworks, offer unique perspectives in predictive analytics(9781107185142_excerpt.pdf.). One such empirical technique we ventured into is the K-Nearest Neighbors (KNN) model. Unlike traditional algorithmic models, KNN operates on the principle of similarity and proximity, providing a complementary viewpoint to our existing analytical tools.

The multidimensional nature of the broader ramifications of our endeavors in nvPM emission prediction is evident. The incorporation of the K-nearest neighbors (KNN) algorithm brings forth a unique standpoint by employing proximity-based decision-making, hence providing a distinctive viewpoint that complements the knowledge acquired via tree-based and linear models. The empirical basis of the K-nearest neighbors (KNN) algorithm provides a level of flexibility in modeling, allowing it to effectively handle and adjust to the numerous nuances and complexity present in the dataset. From a critical perspective, in scenarios where individual engine parameters tend to converge or display distinct localized patterns, the localized decision-making approach of KNN proves to be an important tool, effectively addressing these unique characteristics of the data

3.4.4 Final Model Selection and Insights for nvPM Emission Prediction:

Upon rigorous validation and comparative analysis of the different models employed, our methodology converged on the selection of the Gradient Boosting Regressor for predicting nvPM emissions from aircraft engines.(Masui 2022)

This decision was driven by several factors:

1. **Predictive Accuracy:** Among the models we explored, including linear regression, tree-based models, KNN, SVM, and ensemble techniques, the Gradient Boosting Regressor consistently exhibited superior accuracy in capturing the nuances of the data and predicting nvPM emissions.

2. **Model Robustness:** The iterative nature of Gradient Boosting, where each tree corrects the errors of its predecessor, ensures that the model is robust against potential overfitting and anomalies in the dataset.

3. **Handling Complex Relationships:** Given the intricate interplay of various engine parameters and their potential non-linear relationships with nvPM emissions, the Gradient Boosting Regressor, with its ability to capture such complexities, emerged as the most apt choice.

4. **Future Predictions and Insights:** Leveraging the Gradient Boosting Regressor, our predictions are anchored in rigorous data analytics, providing valuable insights into future nvPM emission trends. This knowledge can be pivotal for various stakeholders in the aviation sector.

3.5Model Evaluation for nvPM Emission Prediction:

The efficacy of a predictive model is gauged not just by its ability to fit training data but by its performance on unseen data. For our nvPM emission prediction models, we employed a robust evaluation framework, ensuring that our model selections were empirically sound(*3 Regression Metrics You Must Know: MAE, MSE, and RMSE* 2022).

Mean Squared Error (MSE): It measures the average squared differences between the predicted values and the actual values.

Formula:

$$\mathbf{MSE = (1/n) * \Sigma(actual - prediction)^2}$$

where:

- **Σ** – summation
- **n** – sample size
- **actual** – the actual data value
- **forecast** – the predicted data value

A lower MSE indicates a better fit of the model to the data. However, since it squares the errors, it gives more weight to larger discrepancies.

Mean Absolute Error (MAE): It calculates the average of the absolute differences between the predicted and actual values.

Formula:

$$\mathbf{MAE = (1/n) * \Sigma| actual - Prediction |}$$

where:

- **Σ** – summation
- **n** – sample size
- **actual** – the actual data value

- **forecast** – the predicted data value

MAE provides a linear penalty to each error, making it more interpretable and less sensitive to outliers compared to MSE.

R-squared (Coefficient of Determination): It quantifies the proportion of the variance in the dependent variable that is predicted by the independent variables.

Formula:

R-Squared = 1 - (Sum of Squares of Residuals (SSR) / Total Sum of Squares (SST))

R-squared values range from 0 to 1, with higher values indicating a better fit. An R-squared of 1 means the model perfectly predicts the target variable, while a value of 0 indicates the model doesn't improve prediction over simply using the mean of the target variable.

4. Implementation

This segment outlines the techniques used to apply the discussed approach and characteristics using Python. The initial part provides insights into the resources utilized for execution and evaluating the effectiveness of the models. Subsequent sections delve into the strategies employed for crafting the machine learning models.

4.1System Preparation

Designing a machine learning model is often demanding on the hardware it runs on. This hardware-intensive nature makes it challenging to develop an efficient machine learning model on a regular laptop. This software benefits immensely from a GPU's virtual memory and processing prowess. Google Colab(*Google Colab - Introduction*) offers a solution by leveraging cloud technology, enabling the development and execution of machine learning models using Python. As a web-based development environment, it allows users to write Python code in a Jupyter notebook-style interface without any local software installations. For this project, Google Colab was indispensable. It offers high-performance computing resources, including GPUs and Tensor Processing Units (TPUs), significantly boosting machine learning tasks' speed. Moreover, Google Colab seamlessly integrates with Google Drive, allowing users to mount their Drive contents directly onto the Colab notebook. This integration was particularly useful to deal with any type of data files, could be stored on the Drive and accessed in the notebook via a file path. Regarding the implementation, the suggested models were developed using Python 3, powered by the Scikit-learn and Keras modules. All experiments took place in the Google Colaboratory environment, boasting 12 GB of RAM and 106 GB of storage. This cloud-based solution offers complimentary access to GPUs with minimal setup. Also, the personal laptop utilized in the research was an AMD Ryzen 7 4800H with Radeon Graphics, clocking 2.90 GHz, 16 GB RAM, and ran on a Windows platform.

4.2Descriptive Analysis

The foundation of our analytical journey was established by obtaining robust and comprehensive data. We sourced our data from the EASA ICAO databank(*ICAO Aircraft Engine Emissions Databank*), a reputable reservoir of aviation-related emission statistics, ensuring the credibility and relevance of our initial dataset.

Gaseous Emissions Dataset (gaseous_emission_df):

Overview: This dataset was a treasure trove of gaseous emission specifics, boasting a vast 105 columns. It provided an expansive range of information, capturing various facets of gaseous emissions from aircraft engines.

The dataset has 896 individual records, the dataset encompassed a rich diversity of engines and their respective emission metrics.

Preliminary examination revealed missing data in several columns. For instance, the "Combustor Description" and "Remark" columns exhibited conspicuous data gaps. Such omissions could arise from varied data collection methodologies or certain engine parameters not being universally applicable.

The dataset's granularity was evident in its detailed engine-specific columns such as B/P Ratio, Pressure Ratio, and Rated Thrust (kN). These metrics were pivotal in offering insights into the correlation between engine performance attributes and their respective emission levels.

Non-volatile Particulate Matter Emissions Dataset (nvpm_emission_df):

Overview: This dataset, although more concise than its gaseous counterpart, was integral to our study. It comprised 76 columns and 214 records, focusing on the nuanced details of non-volatile particulate matter emissions. A notable observation was the emphasis on engines manufactured by "Rolls-Royce plc", providing a unique lens to analyze emissions from this specific manufacturer. As with the gaseous dataset, missing values were evident, especially in columns such as "Remark". Such data gaps underscored potential areas for refining data collection processes or highlighted parameters that were not universally relevant across all engines.

Merging Datasets:

The unification of our two primary datasets was anchored by the "UID No", a unique identifier present in both datasets. This merging exercise enriched our analytical base, ensuring a comprehensive view of engine emissions.

Merging was not without its set of challenges. Columns with identical names in both datasets could have caused ambiguity. To counter this, we introduced suffixes to columns originating from the nvPM dataset, ensuring clarity. Post-merge, to enhance dataset manageability and coherence, redundant columns were identified and pruned. The outcome was a streamlined dataset with 214 rows and 148 columns.

Our merged dataset was emblematic of synergy, amalgamating the depth of both original datasets. It provided a panoramic perspective on engine emissions, setting the stage for meticulous and informed subsequent analysis.

4.3 Exploratory Data Analysis (EDA)

In our EDA(*Importance of EDA in Machine learning* 2018), we are determined to delve deep into different types of emissions, both gaseous and particulate (nvPM). By visualizing these emissions, we can identify and analyze patterns, trends, and anomalies more effectively. We have used visual tools and techniques to unearth patterns, trends, and anomalies in the data.

4.3.1 Hydrocarbon Emission (HC):

HC emissions are highest during the take-off phase. Climb Out also contributes significantly to HC emissions, while the Approach and Idle phases have minimal HC emissions.

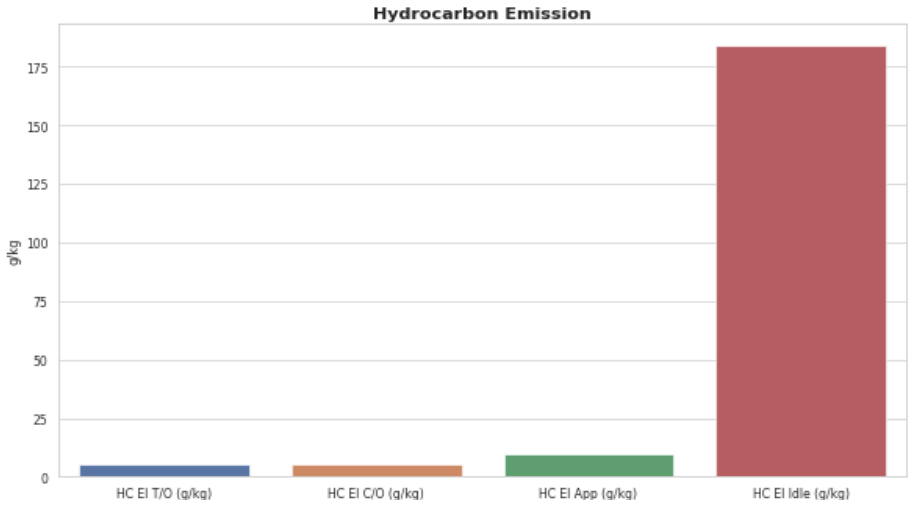

Figure 1 – Hydrocarbon Emission pattern.

The Take Off phase stands out as a critical period for HC emissions, suggesting the need for targeted intervention to reduce emissions during this phase. While Climb Out contributes significantly, the magnitude of emissions is lower than in Take Off. Approach and Idle phases can be considered as periods of relatively low HC emissions.

4.3.2 Carbon Monoxide Emission (CO):

Take-off is the primary contributor to CO emissions, followed closely by the Climb Out phase. Approach and Idle phases exhibit lower CO emissions.

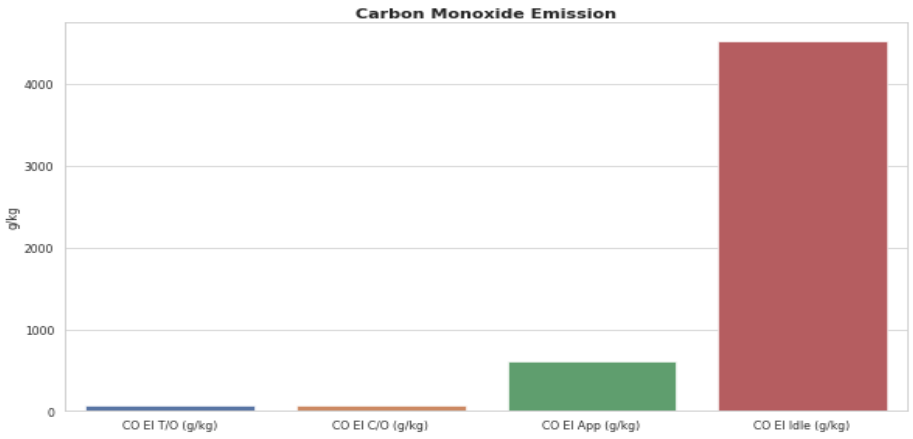

Figure 2- Carbon Monoxide emission pattern

The dominance of Take-off in CO emissions emphasizes the importance of addressing CO emissions during the initial flight phase. Strategies targeting Climb Out can also contribute to emission reduction. Approach and Idle phases, with lower emissions, may be considered as periods of comparatively less environmental concern.

4.3.3 Oxides of Nitrogen (NOx):

Take Off and Climb Out phases are the primary contributors to NOx emissions. Approach and Idle phases contribute less but are still significant.

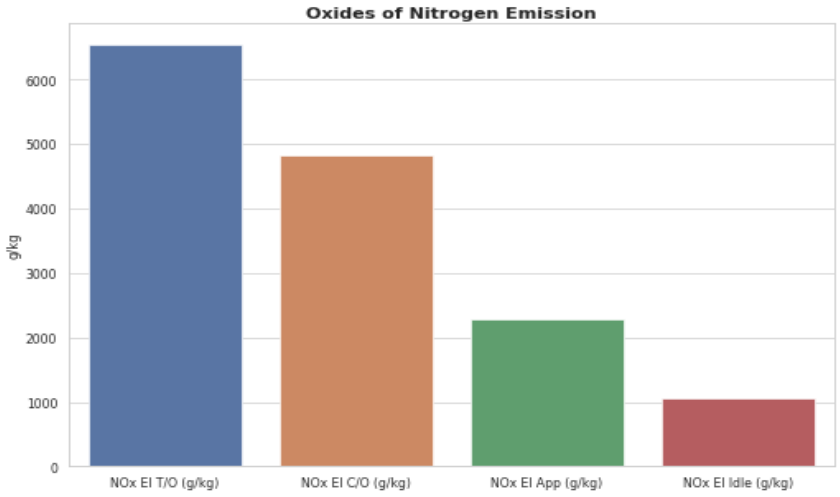

Figure 3- Notrates of Oxides emmison pattern.

The significant contributions of Take Off and Climb Out to NOx emissions highlight the need for interventions during the early flight phases to effectively reduce NOx emissions. While the Approach and Idle phases have lower emissions individually, their cumulative impact remains notable, making them important considerations in emission reduction strategies.

4.3.4 Smoke Number (SN):

Take-off is pivotal for SN emissions, followed by Climb Out. Approach and Idle phases contribute less but are still essential.

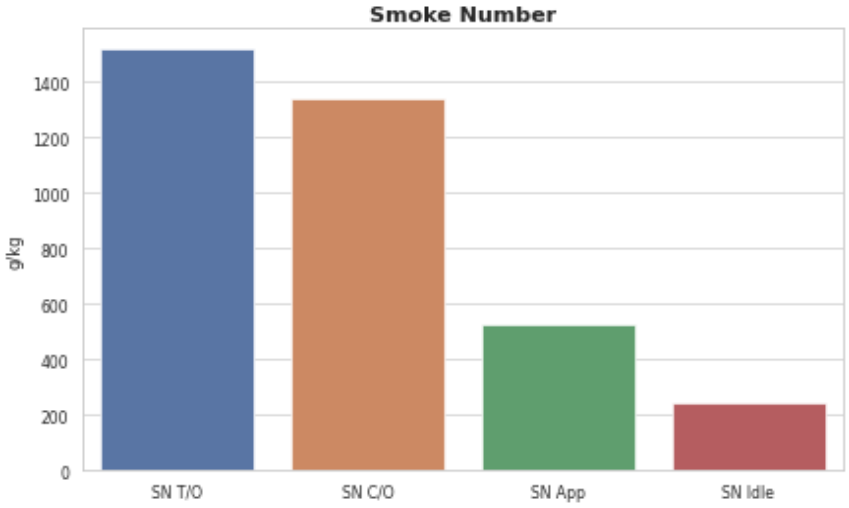

Figure 4- Smoke emission pattern.

Targeting SN emissions during the take-off phase is crucial due to its prominent role. Climb Out also warrants attention as a significant contributor to emissions during the early stages of flight. Approach and Idle phases, with lower emissions, remain integral parts of the overall emission landscape.

4.3.5 Smoke Number vs. nvPm EI:

There isn't a strong linear relationship between Smoke Number (SN) Max and the two nvPM metrics, highlighting the need for more advanced modeling techniques.

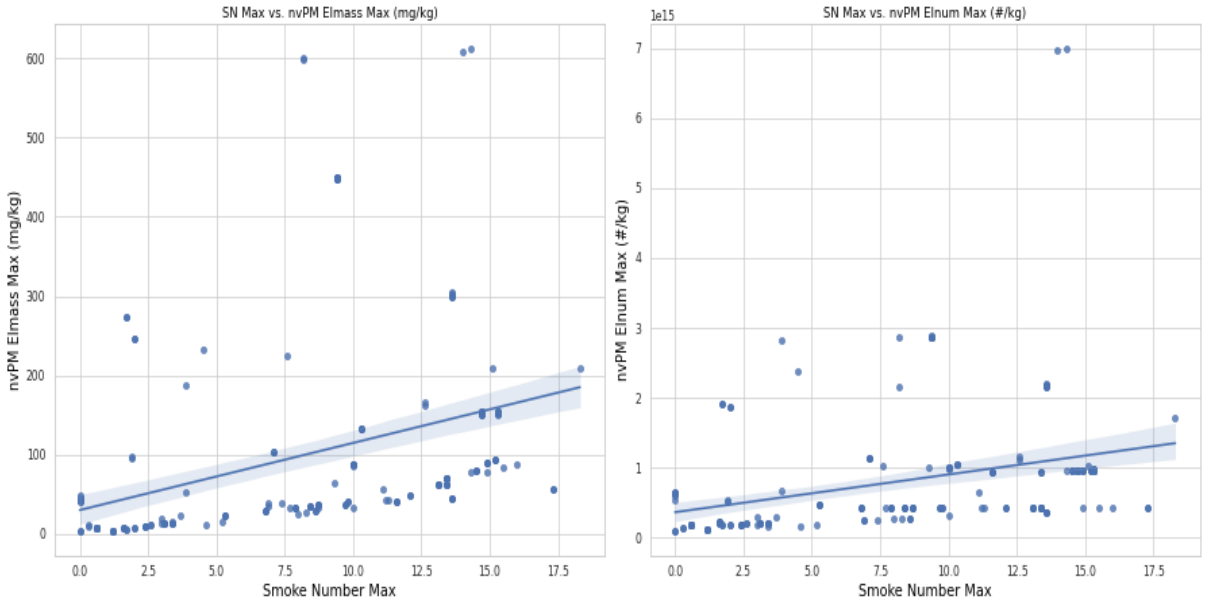

Figure 5- Smoke vs nvPm relationship.

The absence of a strong linear relationship between SN Max and nvPM metrics underscores the complexity of their interactions. Advanced modeling approaches may be necessary to capture these intricate relationships effective.

4.3.6 Non-volatile Particulate Matter Emission Index Mass (nvPM EImass):

Take-off is the primary contributor to nvPM mass emissions, followed by Climb Out. Approach

To reduce nvPM mass emissions significantly, targeted interventions during the take-off and Climb-out phases are crucial. While the Approach and Idle phases have lower contributions

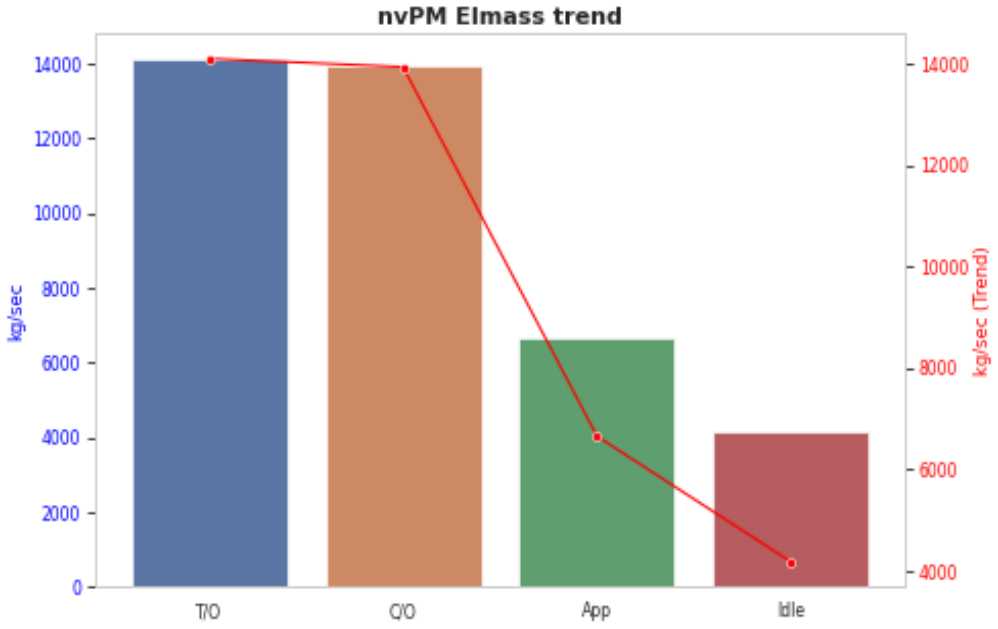

Figure 6- nvPM EI (mass)emission pattern.

individually, addressing emissions during these phases remains important for a comprehensive emission control strategy.

4.3.7 Non-volatile Particulate Matter Emission Index Number (nvPM EInum):

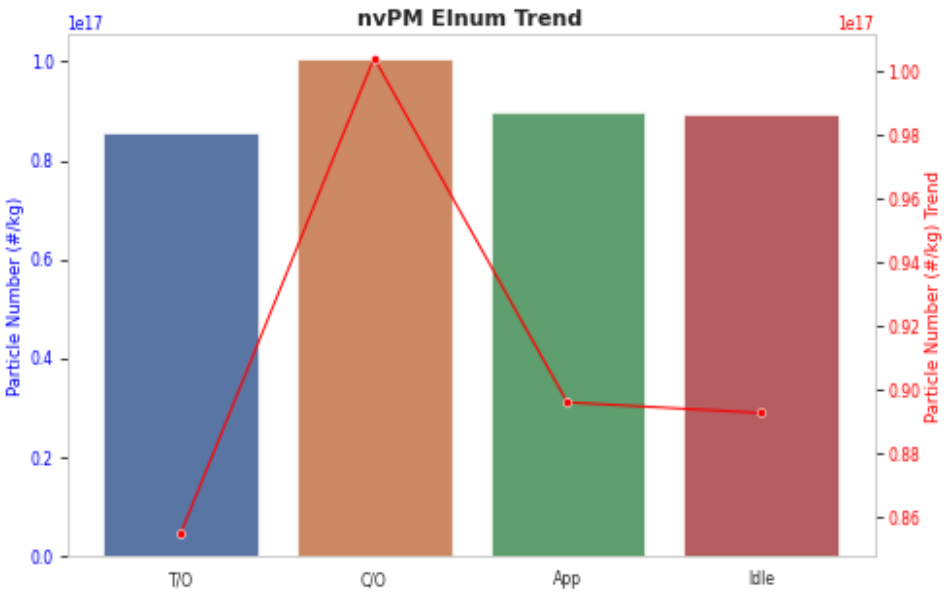

Figure 7- nvPM EI (num) emission pattern.

Take-off is undeniably the primary contributor to nvPM number emissions, followed by Climb Out. Approach and Idle phases consistently show the lowest emissions.

Targeted interventions during the take-off phase are essential for mitigating nvPM number emissions effectively. Climb Out, while a significant contributor, is overshadowed by the dominance of the Take Off phase. Approach and Idle phases, with consistently low emissions, emphasize the need for emission control during these phases.

4.3.8 All Emission Index with Heatmap:

The Take Off phase consistently exhibits the highest emission levels across most categories. Smoke Number remains relatively consistent but lower than other emissions. NvPM Mass emissions are moderate and consistent across all flight phases.

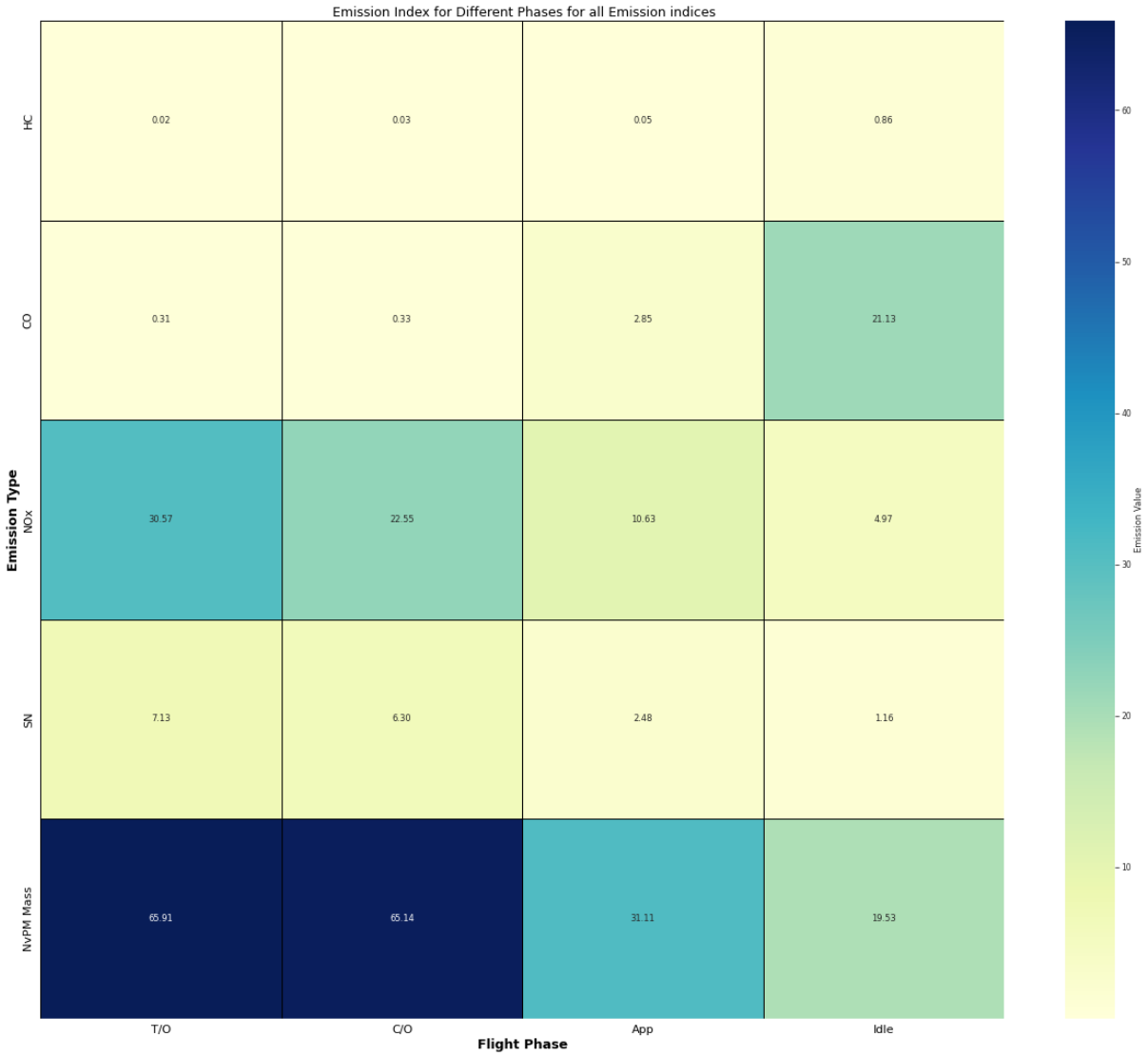

Figure 8- Heatmap of all emission categories.

Understanding the consistent dominance of the Take Off phase in emissions across various categories highlights the potential benefits of focusing on this phase for emission reduction strategies. Additionally, the relatively lower Smoke Number suggests fewer visual pollutants during flight. NvPM Mass emissions, while moderate, remain consistent across flight segments, warranting a comprehensive approach to emission control.

4.3.9 Nvpm EImass & EInum with Rated Thrust(kN):

Both nvPM metrics tend to increase with rated thrust, indicating the importance of engine thrust in emission studies.

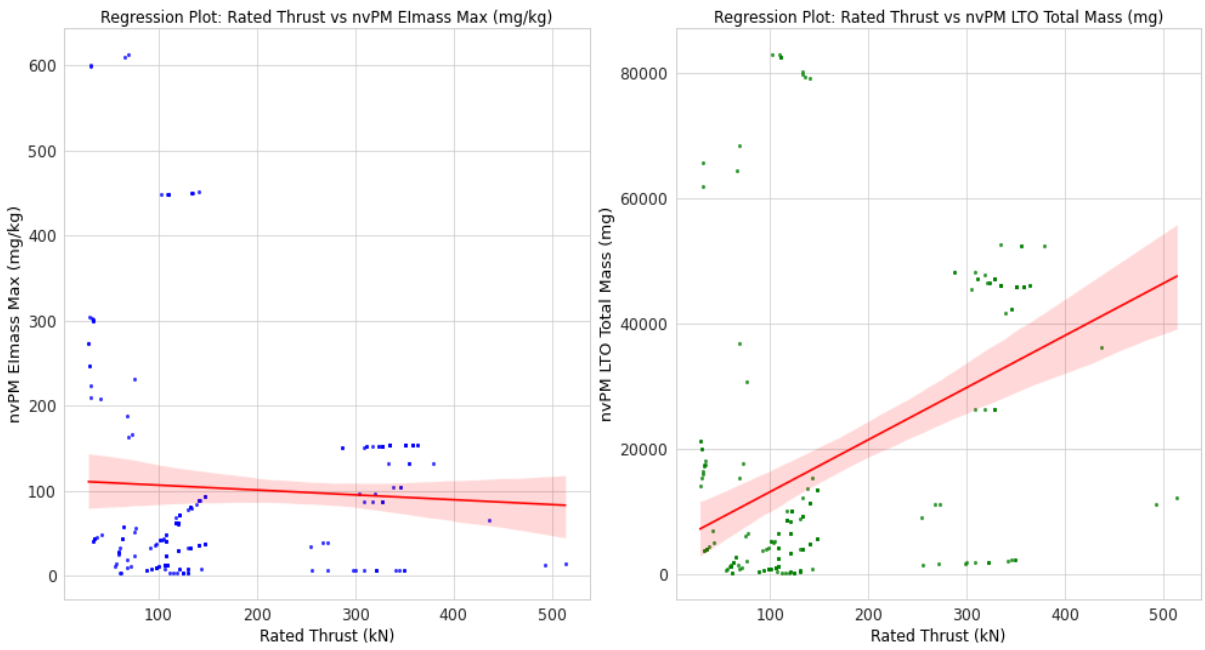

Figure 9- nvPm vs Rated Thrust

The positive correlation between nvPM metrics and rated thrust underscores the pivotal role of engine thrust in emission studies. Strategies aimed at reducing emissions should consider engines with higher thrust values as potential sources of significant emissions.

4.3.10 Nvpm EImass & EInum with Fuel Flow:

Fuel flow and nvPM emissions show some correlation, highlighting the impact of fuel consumption on emissions.

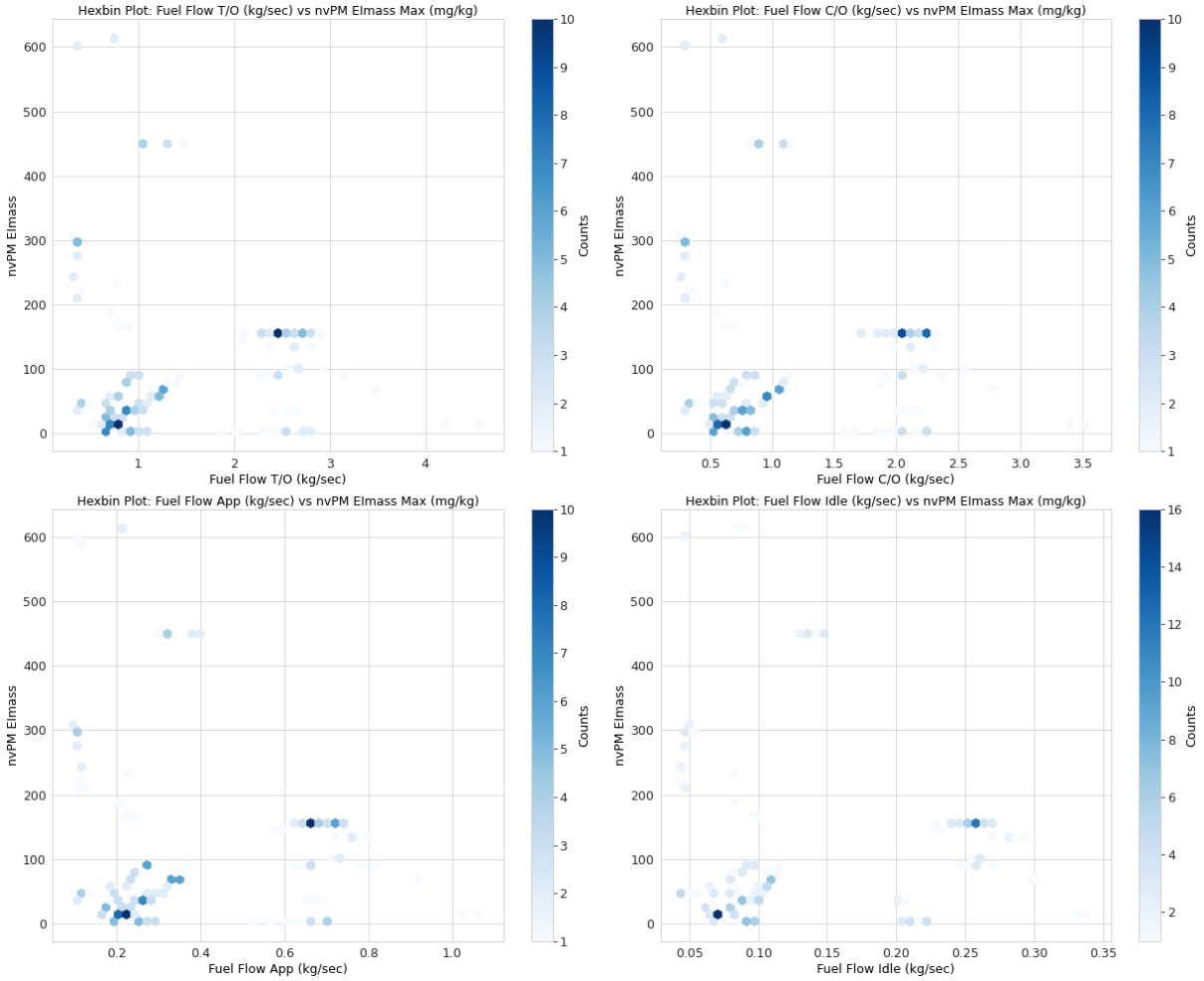

Figure 10- nvPM vs Fuel flow.

The observed correlation between fuel flow and nvPM emissions underscores the importance of fuel consumption management in emission reduction efforts. Strategies aimed at optimizing fuel usage can have a positive impact on reducing emissions.

These detailed observations and implications provide valuable insights into the emissions data and can guide targeted interventions and strategies for mitigating aircraft emissions effectively across various flight phases.

4.3.11 Nvpm EImass & EInum with Fuel Content :

The following figure show how fuel composition parameters affect NvPM mass emission

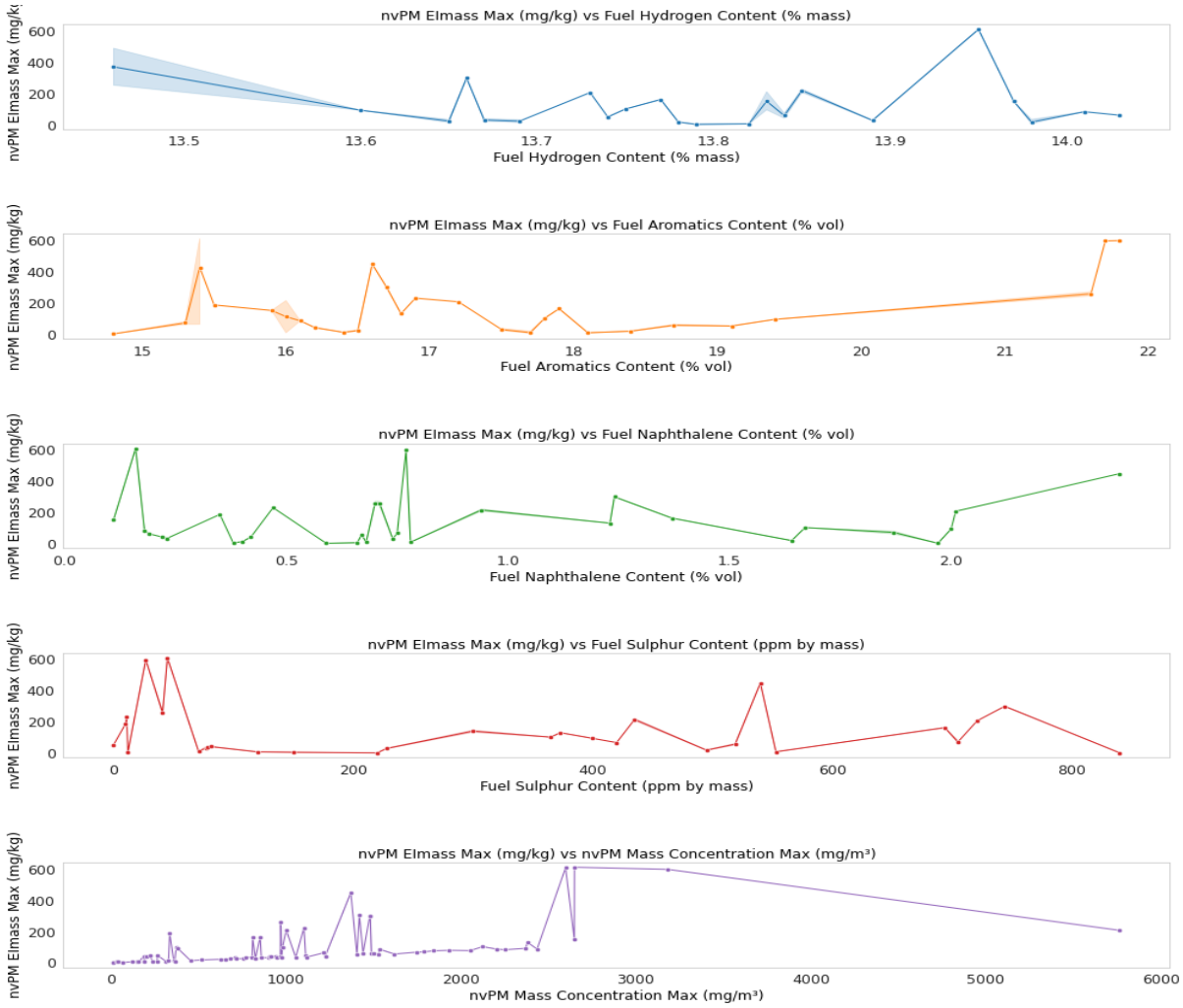

Figure 11- nvPM emission vs Fuel content.

NvPM EImass emission is closely linked to fuel content. Fuel hydrogen content is crucial. Fuel hydrogen content appears to linearly relate to NvPM EImass emission. As fuel hydrogen content increases, NvPM EImass emission falls. This shows that hydrogen-enriched fuels burn cleaner, reducing particulate matter emissions. However, gasoline aromatics behave differently. NvPM EImass emission increases with fuel aromatic content. Aromatic fuels may increase particle emissions due to incomplete combustion. Naphthalene, an aromatic hydrocarbon, enhances this association. Increased gasoline naphthalene content greatly increases NvPM EImass emissions, highlighting its significant impact on particle emissions.

Also important in fuel is sulphur. Sulphur content in fuel directly affects NvPM EImass emission. This is because gasoline sulphur produces sulphate particles, which increase NvPM emissions. Finally, NvPM EImass emission and maximum mass concentration are linearly related to environmental particulate matter concentration. Increased NvPM mass concentration in the environment increases mass-based emissions.

4.4Data cleaning

Data Cleaning plays a pivotal role in ensuring that the dataset is ready for modeling. In this section, we elaborate on the various preprocessing steps undertaken to prepare the dataset for subsequent analysis and modeling.

4.4.1 Feature Selection:

First we have chosen important parameters in emission index prediction

The selected columns primarily encompass parameters crucial for understanding aircraft engine emissions. These include details about fuel composition (like H/C ratio, aromatic content, hydrogen content, naphthalene, and sulfur), engine mechanics (such as combustor type, engine type, bypass ratio, and pressure ratio), and specific operational conditions (like rated thrust and different fuel flow rates). Additionally, environmental parameters like ambient pressure, temperature, and humidity can significantly influence emission characteristics. The columns related to the actual emission metrics for hydrocarbons, carbon monoxide, oxides of nitrogen, and smoke number provide direct measurements, offering a comprehensive dataset to predict engine emissions. These columns provide a holistic view of the factors and outcomes associated with aircraft engine emissions, making them vital for any predictive modeling in this domain.

Engine type, pressure ratio, and combustor description affect combustion. These are engine and combustion characteristics. Full and efficient combustion affects emission quantity and quality. Some engine types use fuel efficiently, reducing particle emissions.

Operational parameters are critical to system performance. These parameters, such as thrust settings or ambient conditions like temperature and humidity, greatly affect operation and results. Based on these parameters, engine efficiency and emissions might vary greatly. Lower ambient temperatures may increase air density, which can affect combustion efficiency and emissions.

Airflow and Aerodynamics: Bypass ratio and pressure ratio indicate engine aerodynamics and internal airflow. These traits may affect fuel-air mixing and combustion efficiency and non-volatile particulate matter emissions. A higher bypass ratio means more air bypasses the combustion chamber, which may affect engine emissions.

Manufacturer and engine identity, albeit seemingly unimportant category data, might include vital information. Manufacturers may use proprietary technologies or design philosophies that

affect emissions.

Historical and compliance data can provide context for engine testing and regulatory compliance, revealing its expected performance. Recently tested or special compliance engines may have well-documented emission patterns, which are crucial for accurate projections.

4.4.2 Handling Missing Values:

In the data preparation pipeline, effectively handling missing values is crucial for assuring the resilience and dependability of any future analysis or modeling. The presence of missing values within a dataset can introduce bias, diminish the statistical strength and efficiency of a model, and even yield misleading outcomes. The dataset under consideration had missing values in the "Combustor Description" column.

The column under "Combustor Description," which is of utmost importance, necessitated thorough attention to rectify any instances of missing information. Various ways exist for addressing missing data, including imputation, which involves replacing missing values with statistical estimates, and utilizing algorithms capable of directly handling missing values. However, it is important to note that each approach has its own specific considerations.

In the our scenario, it was determined that the most suitable approach was to eliminate rows that contained missing data in the "Combustor Description" column. The process of imputation has the potential to produce artificial biases. Considering the specific characteristics of the "Combustor Description" column, it was crucial to maintain its authenticity. By excluding these specific rows from our dataset, we have taken measures to maintain the authenticity of our data, so mitigating the potential for abnormalities or distortions that may arise from imputed or incomplete data.

4.4.3 Encoding Categorical Variables:

In many machine learning models, it is essential to convert categorical variables into a numerical format. This conversion allows the model to work with these variables effectively.

Encoding categorical data into numerical data is essential for machine learning algorithms that need numerical input. Many machine learning algorithms are mathematical and employ numbers, therefore categorical data, while rich in information, cannot be directly used.

The columns "Combustor Description" and "Eng Type" in our dataset were categorical. To use this data in our models, we had to numericize these categories. One-hot, mean, and label

encoding all have drawbacks.

Using one-hot encoding, each category would have its own column. Our dataset has several classifications, therefore this would have greatly increased dimensionality. High-dimensionality models are computationally expensive and can overfit to training data, reducing performance on unknown data.

However, label encoding gives each category a distinct numerical identifier. This approach keeps information in the categorical variable without adding dimensionality. Label encoding could introduce order or hierarchy (because categories get numbers), but our techniques, like gradient boosting, can manage such complexities.

4.4.4 Outlier Management:

Outliers are data points that deviate significantly from the rest of a dataset. These disparities may be due to data variability or data collection issues. Many statistical methods are sensitive to outliers, which can impair machine learning model training. It is important to remember that outliers are not always bad.

Outliers were thoroughly analyzed in our dataset. Outliers often accurately represent the subject. Certain engine types or operational scenarios can cause aviation emissions to differ significantly from the baseline. Outliers in a dataset can reveal harsh operating circumstances or engine performance.

Outliers were retained strategically after considering numerous aspects. Initially, these outliers were assessed for significance. An outlier is valuable if it can provide unique insights or help the model understand severe but realistic situations. Understanding the data's basic features is crucial. Eliminating outliers may oversimplify and misrepresent reality in datasets that anticipate extreme results due to the phenomenon under study. Finally, data loss is a constant concern. When eliminating outliers reduces data volume, preserving them is vital, especially if the dataset is small.

These preparation techniques prepare the dataset for modeling and analysis. Rows with missing values are removed to ensure complete data and avoid bias in our studies. Encoding categorical variables lets our machine-learning algorithms use the information in these columns to improve prediction. These data pretreatment steps ensure the dataset's quality and usefulness for modeling.

4.4.5 Data Splitting:

Predictive modeling aims to create a model that accurately describes the data and can apply its learnt patterns to new, unseen data. This emphasizes the need to divide the dataset into training and testing sections.

The training subset helps the model learn the dataset's core patterns and correlations. Ultimately, a model's efficacy depends on its ability to make accurate predictions on data not used in training. In this scenario, testing subset matters. We can determine the model's prediction power and extrapolation capabilities by testing it on the testing dataset.

Many criteria were considered before choosing the 80/20 split, which allocates 80% of data to training and 20% to testing. For effective learning with fewer datasets, the model needs enough data. Increase the training set's data share to achieve this. The split establishes a peaceful equilibrium by giving the model enough data for effective learning and ample data for thorough performance validation. Balance is crucial in a small dataset to maximize data point utility and gain reliable model effectiveness insight.

5. Model Development for nvPM Emission Prediction.

The essence of our analytical endeavor lies in model development, where we navigate the complex realm of regression models to predict non-volatile particulate matter (nvPM) emissions. In this segment, we will elucidate our methodology, emphasizing the implementation of different models, their significance, and our rationale for selecting them. The implementation of these models is made straightforward by the Python module scikit-learn.

5.1Model Selection and Its Significance:

5.1.1 Linear and Regularized Regression Models:

Linear regression is simple and assumes a linear connection between predictors and targets. High-dimensional data or multicollinearity (correlated independent variables) can make simple linear regression difficult. Predictive modelling relies on linear regression, which assumes dependent and independent variables are linearly related. It minimises the residual sum of squares, the difference between observed and forecasted values. Overfitting can occur with high-dimensional data or multicollinearity in linear regression. The model may match training data well, but it may not perform well on new data.

Regularized regression methods like Ridge regression reduce overfitting. Ridge regression adds a regularization term to linear regression. Overfitting can result from model complexity, hence this word discourages it. The hyperparameter α controls the level of regularization. A smaller α number preserves most traits, whereas a bigger one may eliminate others, ensuring a more broad model. Linear and Ridge regression are easy to implement, initialize the Linear Regression model, fit it using training data, and also specify the α parameter for Ridge regression. Both models can predict fresh data outcomes after training. As regards overfitting, the model performs

5.1.2 Tree-based and Ensemble Models:

Decision Tree is a flowchart with core nodes representing features or attributes, branches representing decision rules, and leaf nodes representing outcomes. Data is separated into segments based on established thresholds of various attributes in this intuitive, perceptual decision-making method. Individual trees may overfit, incorporating irrelevant or incorrect training dataset material.

Ensemble methods like Random Forest and Gradient Boosting address this issue. The Random

Forest algorithm uses a group of decision trees. The suggested method combines many decision trees to improve accuracy and reduce overfitting. Gradient Boosting builds trees sequentially. Every tree in the ensemble model corrects the mistakes of its predecessor, improving forecast accuracy. However, this iterative method may require careful tuning to avoid overfitting. The provided training data can initialize and train the Decision Tree model. In Random Forest and Gradient Boosting, the n_estimators option specifies the ensemble tree count. Before fitting models to training data, this specification is established. These models can predict novel datasets after training.

5.1.3 Chain-based Regression Models:

SVR and KNN regression analysis methods differ. Support Vector Regression (SVR) uses SVM techniques. It seeks the best hyperplane to divide a dataset into classes in a modified space. The simpler K-nearest neighbors (KNN) algorithm predicts a new data point by evaluating its proximity to its neighbors. Adjusting the number of neighbors assessed depends on the situation. The Regressor Chain ensemble method is designed for many outputs. Instead of predicting several outputs simultaneously, this strategy progressively links regressors, which consider each other.

The SVR model can be instantiated with a 'linear' kernel after importing the necessary classes. K-Nearest Neighbours (KNN) users can select the number of neighbors to consider by setting the n_neighbors argument. The Regressor Chain approach uses a Decision Tree as a base estimator. Fitting these models on training data follows initialization. After training, people can predict new data.

Model development serves as the heart of our analytical journey, where we delve into the realm of regression models to predict non-volatile particulate matter (nvPM) emissions. In this section, we comprehensively outline our approach to model selection, the significance of different models, and the final model selection process.

6. Evaluation and Results

The process of constructing a model serves as the initial step, but the crucial aspect is in its validation using real-world, unseen data. In the process of predicting non-volatile particulate matter (nvPM) emissions, a thorough evaluation of each model was conducted, with a focus on ensuring that our conclusions were supported by empirical evidence and adhered to rigorous scientific standards.

In order to assess the precision and dependability of our models, we utilized various evaluative criteria, each offering a distinct perspective to comprehend the model's capabilities and limitations:

The Mean Squared Error (MSE) is a metric that measures the average of the squared discrepancies between the expected and actual values. Although it provides a reliable metric, the squared term in this measure can significantly amplify larger errors. Essentially, a smaller mean squared error (MSE) indicates a model that is more aligned with the data. However, it is important to exercise caution regarding its susceptibility to extreme deviations.

The Mean Absolute Error (MAE) is a metric that focuses on the absolute differences between predicted values and actual values, in contrast to the Mean Squared Error (MSE). The implementation of this straightforward approach guarantees that every error, regardless of its extent, is handled in a consistent manner. As a result, the Mean Absolute Error (MAE) metric is sometimes considered to possess greater transparency, as it does not excessively penalize larger errors. This attribute renders it particularly valuable in scenarios when outliers are present.

The coefficient of determination, sometimes referred to as R-squared, is a measure that indicates the extent to which a model can explain the variability of the dependent variable. The measure quantifies the proportion of variance in the dependent variable that can be accounted for by the independent variables. When the values of R-squared are confined within the range of 0 to 1, a larger value signifies that the model has the ability to explain a greater amount of variability in the target variable. Nevertheless, it is uncommon to witness a perfect R-squared value of 1 in real-world scenarios, and in certain cases, it may suggest the presence of overfitting. On the other hand, a close-to-zero R-squared value indicates that the model may not be providing substantial predictive value beyond a simplistic mean-based approach.

6.1Regression Model Performance Metrics:

Model	MSE train	MAE train	R-squared train	MSE test	MAE test	R-squared test
Multitarget Regression	834.4993	17.39244	0.944672	2056.739356	26.705638	0.825663
Decision Tree Regressor	8.217301e-33	1.850372e-17	1.0	775.015476	4.854762	0.934307
Random Forest Regressor	333.9447	4.955512	0.977859	798.223246	9.15581	0.93234
Gradient Boosting Regressor	1.916463	0.9603046	0.999873	397.14294	5.431383	0.966337

Table 1- Perfoemance Metrics

6.2Evaluation of Model Performance:

1. Multitarget Regression: - MSE & MAE: The Multitarget Regression's high values for training and testing sets show that it may not be capturing data patterns as well as other models. Higher error numbers imply a large difference between projected and actual values. R-Squared The model looks to explain a lot of variance with a training data score of 0.944672. The R-squared score lowers to 0.825663 for test data, suggesting the model may not generalize well to unseen data.

2. Decision Tree Regressor: - MSE & MAE: If the training data values are near zero, the model fits it flawlessly. However, the higher test error numbers, notably the MSE, may indicate overfitting. R-Squared The ideal training data score of 1.0 indicates overfitting. The model memorized training data instead of learning. The test set's R-squared value of 0.934307 implies the model still generalizes well to fresh data.

3. Random Forest Regressor: - MSE & MAE: The ensemble Random Forest approach handles complicated datasets well due to its lower error rates. It has fewer training and test mistakes than Multitarget Regression, making it more promising. R-Squared With values of 0.977859 for training and 0.93234 for testing, the Random Forest model captures data variation and generalizes effectively to new data.

4. Gradient Boosting Regressor: - MSE & MAE: This model performs best on the test set with the lowest prediction errors. Its reduced mistakes, especially on test data, demonstrate its

strong predictive abilities. R-Squared For the training set, the Gradient Boosting Regressor performs well with a nearly perfect R-squared value. Its high test set value of 0.966337 shows its robustness and flexibility, making it the most trustworthy model examined.

Each model has pros and cons, but the Gradient Boosting Regressor accurately predicts nvPM emissions. Its evaluation measures indicate stability and robustness, making it a good candidate for actual implementations.

Model Evaluation is crucial for understanding its performance.

6.3The Best-performing Predicting model:

In our investigation, it was shown that the Gradient Boosting Regressor had superior performance compared to other models. The test data exhibits a notable R-squared value of approximately 0.963, which highlights the strong predictive capability of the model. In addition, the minimized mean squared error (MSE) and mean absolute error (MAE) numbers highlight the high level of accuracy in its predictive capabilities. The Gradient Boosting Regressor exhibited noteworthy outcomes within this particular context. However, it is crucial to recognize that the performance of the model can be inherently linked to the properties of the dataset. Therefore, in order to ensure maximum efficiency, it is recommended to conduct regular assessments and periodic revisions of the model.

Among the models that were assessed, Gradient Boosting stands out due to the following characteristics:

- Iterative refinement is a process in which forecasts are gradually improved by the incorporation of knowledge gained from previous blunders.
- The versatility of the model lies in its ability to effectively capture complex data patterns that may not be readily discernible by linear models.
- The demonstration of resilience towards potential data anomalies is exemplified through the allocation of additional attention to problematic circumstances.
- Adaptive regularization is a technique that incorporates factors such as learning rates in order to address the issue of overfitting.
- The utilization of shallow trees enhances the generalizability of the approach, hence expanding its potential applications.

In light of these characteristics, the comprehensive methodology of Gradient Boosting, coupled with its capacity to accommodate the intricacies of our dataset, makes it particularly adept for

forecasting nvPM emissions under our specific circumstances.

6.4Results:

This analysis entails a comprehensive evaluation of the predictions produced by the gradient boosting algorithm for the variable 'nvPM Eimass (mg/kg)'.

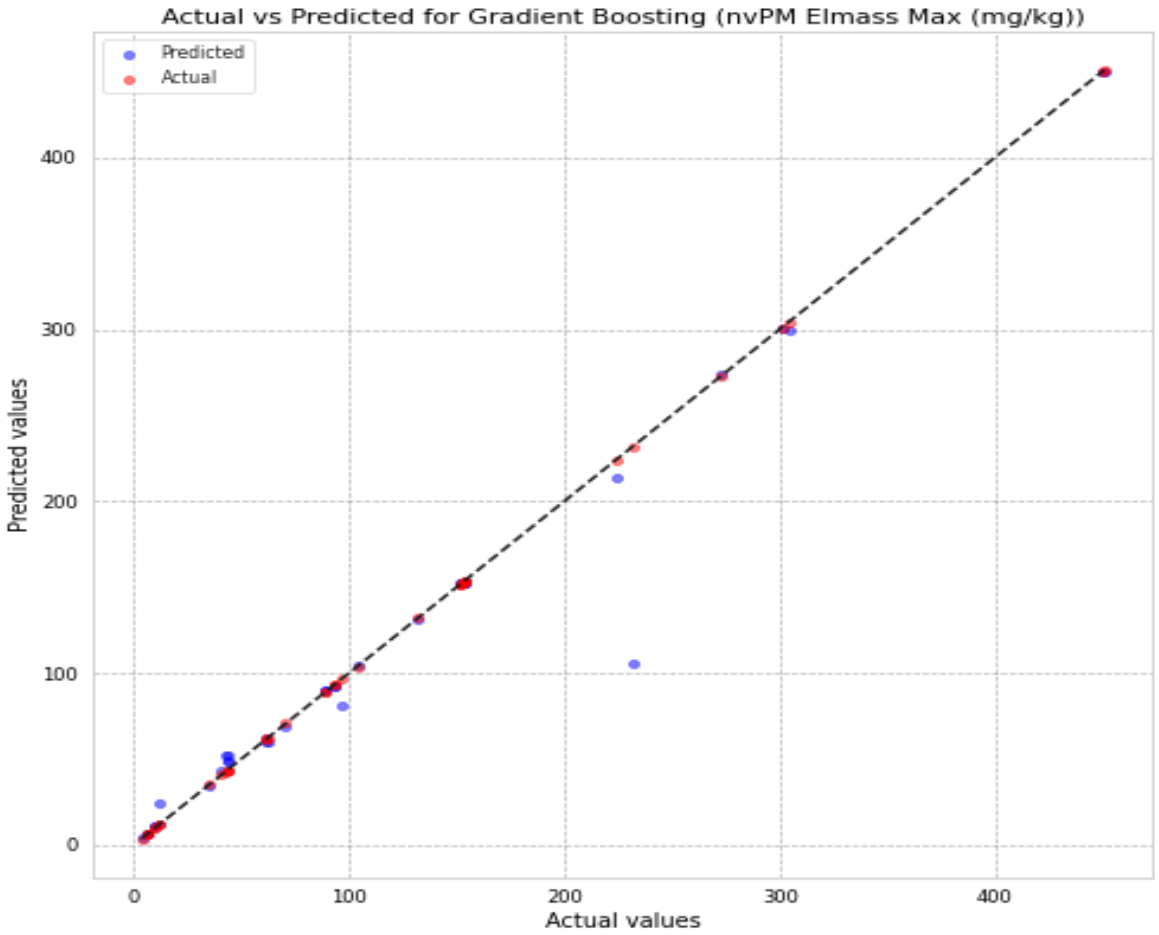

Figure 12- Prediction Analysis.

The visual representation of the model's predictions demonstrates its competence in accurately anticipating the desired outcomes. In the initial data point, the model exhibits a small undervaluation of 3.1%. Upon transitioning to the subsequent data presentation, it becomes apparent that the model's proficiency is clearly demonstrated by a mere variance of 0.78%. The final figure exhibits a disparity of 1.77%, which, although slightly greater, remains within an acceptable limit. The fourth image provides more evidence of the model's consistency, demonstrating a very imperceptible distinction. Nevertheless, the analysis of the fifth plot uncovers a significant overestimation of 10.7%, which serves as an indication of prospective opportunities for enhancing the model. The graphical analysis presented in this study highlights the significance of utilizing visual diagnostics in the field of machine learning. These visual

tools play a crucial role in assessing the dependability of models and identifying areas that require enhancement.

Actual vs Predicted values by Gradient bossting algorithm.

ata Points	Actual Values nvPM Eimass(mg/kg	Predicted Values nvPm EI mass (mg/kg)
1	70.8	68.633535
2	153.7	152.540013
3	304.3	298.94564
4	152.5	152.270419
5	43.9	48.604837

Table 2- Prediction Comparison

The provided table offers a comparative analysis between observed and projected values for the dependent variable 'nvPM Eimass (mg/kg)' utilizing the Gradient Boosting model. Upon initial examination, the predictions roughly align with the actual values, hence confirming the model's resilience.

The Gradient Boosting model demonstrated a high level of accuracy in its predictions, but with slight variations. In the initial data point, the model's forecast exhibited a deficit of 2.2 mg/kg, equivalent to 3.1% in relation to the true value. The second data point demonstrated a high level of precision, as indicated by a variation of only 0.78%. However, the disparity expanded to 1.77% in the third aspect. The fourth data point highlighted the model's high level of consistency, with a deviation of only 0.2 mg/kg. On the contrary, the fifth data point exhibited an overestimation of 10.7%. These discrepancies highlight the necessity of ongoing reviews and adjustments of models.

Reliability of the Model: The close alignment between the predicted values and the actual values indicates that the Gradient Boosting model exhibits reliability in relation to the specific job being addressed. Nevertheless, although the model demonstrates commendable performance in the majority of cases, sporadic aberrations (as exemplified by Data Point 5) underscore the significance of ongoing review and enhancement of the model.

Observations: The analysis of this comparison provides valuable insights into the strengths and potential areas for enhancement of the model. Frequent comparison between anticipated outcomes and observed data serves the dual purpose of validating the current effectiveness of

the model and offering insights for its improvement. The reliability of emission prediction models is of utmost relevance in the field of environmental science, considering their crucial role in accurately estimating emissions.

7. Discussion

The choice between the SCOPE11 empirical methodology and machine learning approaches is not a one-size-fits-all decision but rather a nuanced consideration that hinges on the specific nature of the problem at hand and the characteristics of the available data.

SCOPE11 Empirical Methodology:

- Limited Model Flexibility: The SCOPE11 methodology, like many empirical models, often relies on predefined functional forms (e.g., polynomial or exponential). While this simplifies the modeling process, it might not capture the inherent complexity or non-linearity present in real-world data.
- Extrapolation Risks: Empirical equations derived from SCOPE11 can provide accurate predictions within the data range used for derivation. However, extrapolating beyond this range can lead to significant inaccuracies, making it less suitable for scenarios with data outside the training range.
- Over-reliance on Data Quality: The accuracy and reliability of SCOPE11-derived equations heavily depend on the quality and representativeness of the initial data. Any bias or systematic error in the dataset will be reflected in the derived relationships, potentially leading to biased predictions.
- Lack of Adaptability: Once empirical relationships are derived, adapting or updating them to accommodate new data or insights can be challenging. This lack of adaptability can hinder the model's ability to incorporate new information.
- Assumption Limitations: The empirical methodology often involves assumptions about data distribution, error terms, or relationships between variables. If these assumptions do not hold in practice, the model's accuracy can be compromised.

Machine Learning Approaches:

- Overfitting: Machine learning models, especially complex ones, can overfit the training data by fitting it too closely, capturing noise rather than the underlying trend. This can lead to poor generalization to new, unseen data.
- Black Box Nature: Some machine learning models, particularly deep learning models,

are notoriously hard to interpret. This "black box" nature can make it challenging to understand the model's underlying decision-making process or identify sources of error.

- Data Hungry: Many machine learning models, especially complex ones, require large amounts of data to train effectively. In scenarios with limited data, these models might not perform optimally and can be prone to overfitting.
- Sensitivity to Hyperparameters: Machine learning models come with hyperparameters that need to be carefully tuned. The model's performance can be highly sensitive to these hyperparameters, and tuning them requires expertise and computational resources.
- Computational Intensity: Training complex machine learning models, such as deep neural networks, can be computationally intensive and time-consuming, potentially limiting their practicality in resource-constrained environments.
- Risk of Data Leakage: If not handled properly, there's a risk of data leakage, where information from the test set inadvertently influences the model during training, leading to overly optimistic performance metrics.

In the context of predicting aircraft emissions, both methods have their merits. The SCOPE11 method offers a direct and interpretable approach based on observed empirical relationships. It's excellent for scenarios where stakeholders need clear, understandable equations, such as regulatory compliance. On the other hand, machine learning models, with their flexibility, might offer higher accuracy, especially when dealing with complex, multi-dimensional data. Their ability to capture intricate patterns, non-linearities, and variable interactions can lead to more accurate predictions, essential for operational or regulatory scenarios.

In conclusion, the choice between the SCOPE11 method and machine learning models isn't a binary decision but rather a matter of trade-offs. It depends on the specific requirements of the problem, the nature of the data, and the desired balance between interpretability, flexibility, and predictive accuracy. In practice, a hybrid approach that combines the strengths of both methodologies might offer the best solution.

8. Conclusion

In this research study, our objective was to conduct a thorough examination of aircraft engine emissions, employing a combination of conventional methodology such as SCOPE11 and contemporary machine learning approaches. The conducted analysis has unveiled discernible emission patterns that exhibit variability, even within engine models originating from the identical manufacturer. The complexities of the data were further emphasized while addressing missing values, encoding categorical variables, and managing outliers.

Significant historical and compliance data provided a lot of information about the engine's performance and how well it met legal requirements. The Gradient Boosting Regressor, which showed advanced machine learning models, showed what ensemble methods can do. However, the task of achieving optimal performance on our dataset posed hurdles, with a particular focus on striking a balance between the intricacy of the model and its practicality in real-world scenarios.

Although our research yielded significant insights, certain limits were apparent. The presence of computational limitations, interpretability concerns in modeling, and the task of preventing overfitting were pervasive..

In future research initiatives, it is advisable to explore more factors, use more sophisticated machine learning models, and potentially broaden the dataset to include a wider range of engines and climatic circumstances. The correlation established between empirical approaches and current analytics highlights the significant potential within this field, aiming to facilitate a more sustainable future for aviation.

9. Future Work

9.1Integration of Additional Emission Data:

Broader Emission Categories: While our focus has primarily been on Black Carbon (BC) and Non-Volatile Particulate Matter (nvPM), the inclusion of other significant aircraft emissions, such as Carbon Dioxide (CO2) and Nitrogen Oxides (NOx), could provide a more comprehensive view of an aircraft's environmental impact.

Temporal Analysis: Exploring emissions data over extended periods can help in identifying trends, variations, or cyclical patterns. This could be crucial in understanding the long-term effects and in developing strategies for mitigation.

9.2Enhanced Machine Learning Models :

Deep Learning Integration: The incorporation of neural networks and deep learning techniques might offer more nuanced insights, especially when dealing with vast datasets or complex emission patterns.

Feature Engineering: Further refinement in feature selection and engineering can potentially enhance the predictive power of our models. Exploring non-linear transformations or even domain-specific features can be pivotal.

Validation with Real-world Scenarios :Field Testing: Validating our predictions with actual field measurements can help in fine-tuning our models. Establishing partnerships with aviation authorities or research institutions for real-time data collection can be instrumental.

Comparative Analysis: Comparing the performance of the SCOPE11 method and our machine learning models in real-world scenarios will provide a practical benchmark of their accuracy and reliability.

9.3Refinement of the SCOPE11 Methodology :

Incorporation of Advanced Mathematical Models: While the current polynomial regressions serve the purpose, exploring other mathematical models might offer better fits or predictions in certain scenarios.

Automated Data Collection Mechanisms: Modernizing the data collection phase with automated tools or sensors can streamline the process, ensuring more frequent and accurate data inputs.

Collaborative Approaches :Hybrid Models: Combining the empirical strength of the SCOPE11 method with the predictive power of machine learning can result in hybrid models that harness the best of both worlds.

Stakeholder Collaboration: Engaging with aircraft manufacturers, airlines, and environmental agencies can offer diverse perspectives, refining our approaches and ensuring they are aligned

with industry needs and standards.

10. Reflection

Engaging in this research endeavor resembled an exploration of the complex and multifaceted domain of aviation emissions. The individuality of emissions behaviors is influenced by the distinctive features associated with each engine, model, and even the specific operational conditions. This understanding was brought to my attention. It became clear that, despite widespread celebration of the achievements of contemporary machine learning algorithms, more established approaches, such as SCOPE11, have an underappreciated elegance and dependability. The aforementioned well-established procedures serve as a fundamental basis upon which more recent analytical techniques might be built.

The data had characteristics that mirrored those of the real world, encompassing intricacies, occasional disorder, and subtle distinctions. The statement reaffirmed the notion that within specialized domains such as aviation, the utilization of data is never straightforward and readily applicable. However, it requires meticulous preprocessing, thorough validation, and a critical evaluation to assure its alignment with the real-world situations it aims to depict.

Nevertheless, the difficulties did not cease with the process of data preprocessing. The domain of machine learning, despite its considerable capabilities, is fraught with inherent intricacies. The pursuit of an optimal model encompasses not only its accuracy but also the imperative of ensuring the model's predictions may be effectively applied to diverse contexts. Furthermore, the selection of the model, its level of interpretability, and its compatibility with domain-specific intricacies are of utmost importance.

Upon critical reflection of the study process, it becomes evident that there were some junctures when alternative trajectories may have been pursued in order to yield more comprehensive and profound findings. A more comprehensive approach to data collecting, encompassing a wider range of engine types, operational scenarios, and regional differences, would have yielded a more comprehensive understanding of emissions. The process of model generation may have been enhanced through a more iterative approach, involving the exploration of a range of models and subsequent refinement based on ongoing feedback loops. The significance of stakeholder participation emerged as an additional insight. Incorporating the insights of domain experts, environmentalists, and pilots could have enriched the analysis by incorporating perspectives that may not be captured by a purely data-driven approach.

References:

3 Regression Metrics You Must Know: MAE, MSE, and RMSE. 2022. Available at: https://proclusacademy.com/blog/explainer/regression-metrics-you-must-know/ [Accessed: 9 October 2023].

9781107185142_excerpt.pdf. [no date]. Available at: https://assets.cambridge.org/97811071/85142/excerpt/9781107185142_excerpt.pdf [Accessed: 9 October 2023].

Agarwal, A. et al. 2019. SCOPE11 Method for Estimating Aircraft Black Carbon Mass and Particle Number Emissions. *Environmental Science & Technology* 53(3), pp. 1364–1373. doi: 10.1021/acs.est.8b04060.

Ge, F. et al. 2022. Predicting aviation non-volatile particulate matter emissions at cruise via convolutional neural network. *Science of The Total Environment* 850, p. 158089. doi: 10.1016/j.scitotenv.2022.158089.

Google Colab - Introduction. [no date]. Available at: https://www.tutorialspoint.com/google_colab/google_colab_introduction.htm [Accessed: 9 October 2023].

ICAO Aircraft Engine Emissions Databank. [no date]. Available at: https://www.easa.europa.eu/en/domains/environment/icao-aircraft-engine-emissions-databank [Accessed: 9 October 2023].

Importance of EDA in Machine learning. 2018. Available at: https://discuss.analyticsvidhya.com/t/importance-of-eda-in-machine-learning/64342 [Accessed: 9 October 2023].

K-Nearest Neighbor(KNN) Algorithm. 2017. Available at: https://www.geeksforgeeks.org/k-nearest-neighbours/ [Accessed: 9 October 2023].

Linear Regression in Machine learning - Javatpoint. [no date]. Available at: https://www.javatpoint.com/linear-regression-in-machine-learning [Accessed: 9 October 2023].

Machine Learning Random Forest Algorithm - Javatpoint. [no date]. Available at: https://www.javatpoint.com/machine-learning-random-forest-algorithm [Accessed: 9 October 2023].

Masui, T. 2022. *All You Need to Know about Gradient Boosting Algorithm − Part 1. Regression*. Available at: https://towardsdatascience.com/all-you-need-to-know-about-gradient-boosting-algorithm-part-1-regression-2520a34a502 [Accessed: 9 October 2023].

Singh, A. 2018. *A Comprehensive Guide to Ensemble Learning (with Python codes)*. Available at: https://www.analyticsvidhya.com/blog/2018/06/comprehensive-guide-for-ensemble-models/ [Accessed: 9 October 2023].

Teoh, R., Stettler, M.E.J., Majumdar, A., Schumann, U., Graves, B. and Boies, A.M. 2019. A methodology to relate black carbon particle number and mass emissions. *Journal of Aerosol Science* 132, pp. 44–59. doi: 10.1016/j.jaerosci.2019.03.006.

What is SVM? Machine Learning Algorithm Explained. 2020. Available at: https://www.springboard.com/blog/data-science/svm-algorithm/ [Accessed: 9 October 2023].

Quadros, F.D.A., Snellen, M., Sun, J. and Dedoussi, I.C. 2022. Global Civil Aviation Emissions Estimates for 2017–2020 Using ADS-B Data. *Journal of Aircraft* (DOI: 10.2514/1.C036763), pp. 1–11. doi: https://doi.org/10.2514/1.c036763.

Restori, M. 2021. *What is Exploratory Data Analysis*. Available at: https://chartio.com/learn/data-analytics/what-is-exploratory-data-analysis/.

Schulz, M. and McConnell, J.R. 2022. *Historical changes in aerosol*. Available at: https://www.sciencedirect.com/topics/earth-and-planetary-sciences/emission-inventory [Accessed: 9 September 2023].

Si, M., Tarnoczi, T.J., Wiens, B.M. and Du, K. 2019. Development of Predictive Emissions Monitoring System Using Open Source Machine Learning Library – Keras: A Case Study on a Cogeneration Unit. *IEEE Access* 7(Environ. Sci. Technol. 2019, 53, 21, 12865–12872), pp. 113463–113475. Available at: https://ieeexplore.ieee.org/abstract/document/8771122.

Statology. 2019. *What is a Good R-squared Value?*. Available at: https://www.statology.org/good-r-squared-value/.

Voigt, C. et al. 2021. Cleaner burning aviation fuels can reduce contrail cloudiness. *Communications Earth & Environment* 2(1), pp. 1–10. Available at: https://www.nature.com/articles/s43247-021-00174-y.

Wasiuk, D., Khan, M., Shallcross, D. and Lowenberg, M. 2016. A Commercial Aircraft Fuel Burn and Emissions Inventory for 2005–2011. *Atmosphere* 7(6), p. 78. doi: https://doi.org/10.3390/atmos7060078.

Wikipedia Contributors. 2019. *Regression analysis*. Available at: https://en.wikipedia.org/wiki/Regression_analysis.

Appendices

All the Codes along with their data is given in the Link Below.

Appendix 1 –

Link contains two folders:

https://drive.google.com/drive/folders/1FoeeU2dCLtjHTyz27Fq6dYE-04lj6XVp?usp=sharing

Folder 1- Data (Two datasets for code & Predicted dataset as Output)

Folder 2- Code(.pynb & .py files of similar code.)