Co2 Emission Prediction

Major Project Report

Submitted in partial fulfilment of the requirement of the degree of

Master of Computer Application

to

K.R Mangalam University

by

SURAJ SAURAV (2201560024)
NIKHIL KUMAR (2201560039)
ABHISHEK PATHAK (2201560041)

Under the supervision of

Ms. Ruchika Bhakhar

Department of Computer Science and Engineering

School of Engineering and Technology

K.R Mangalam University, Gurugram- 122001, India

June 2024

Co2 Emission Prediction

Major Project Report

Submitted in partial fulfilment of the requirement of the degree of

Master of Computer Application

to

K.R Mangalam University

by

SURAJ SAURAV (2201560024)
NIKHIL KUMAR (2201560039)
ABHISHEK PATHAK (2201560041)

Under the supervision of

Ms. Ruchika Bhakhar

Department of Computer Science and Engineering

School of Engineering and Technology

K.R Mangalam University, Gurugram- 122001, India

June 2024

CERTIFICATE

This is to certify that the Project Synopsis entitled, **"CO2 EMISSION PREDICTION** submitted by **"NIKHIL KUMAR (2201560039), ABHISHEK PATHAK (2201560041) and SURAJ SAURAV (2201560024)** to **K.R Mangalam University, Gurugram, India,** is a record of bonfire project work carried out by them under my supervision and guidance and is worthy of consideration for the partial fulfilment of the degree of **Master of Computer Application** in **Computer Science and Engineering** of the University.

17/5/24

Signature of supervisor

Ms. Ruchika Bhakhar

Assistant Professor, SOET

Signature Dean SOET

Dr Pankaj Agarwal, Dean SOET

Date: 3rd JUNE 2024

DECLARATION

We declare that this written submission represents our ideas in our own words and where other's ideas or words have been included, we have adequately cited and referenced the original sources. We also declare that we have adhered to all the principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in our submission. We understand that any violation of the above will cause disciplinary action by the Institute and canal so evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed. We further declare that if there is any violation of the intellectual property right or copyright, my supervisor and university should not be held responsible for the same. Student Name (Roll No.) (Signature)

Abhishek Pathak 2201560041 Abhishek Pathak

Nikhil Kumar 2201560039 Nikhil Kumar

Suraj Saurav 2201560024 Suraj Saurav

Place: K.R. MANGALAM UNIVERSITY, GURUGRAM

Date: 17.5.2024

ACKNOWLEDGEMENT

It gives us immense pleasure to express my deepest sense of gratitude and sincere thanks to my highly respected and esteemed guide "**Ms. Ruchika Bhakhar**" the Assistant Professor, School of Engineering and Technology, for their valuable guidance, encouragement and help for completing this work. Their useful suggestions for this whole work and co-operative behaviour are sincerely acknowledged.

I would like to express my sincere thanks to "**Ms. Ruchika Bhakhar**" for giving me this opportunity to undertake this project. We would also like to thank my HOD "**Shweta Bansal**" and School Dean (SOET) "**Dr. Pankaj Agarwal**" for their wholehearted support.

I also wish to express my indebtedness to my parents as well as my family member whose blessings and support always helped me to face the challenges ahead. At the end would like to express my sincere thanks to all my friends and others who helped me directly or indirectly during this project work.

Place: - K.R. Mangalam University

Date: - 17.5.2024

Abhishek Pathak 2201560041

Nikhil Kumar 2201560039

Suraj Saurav 2201560024

INDEX

S. No.	Title	Page No.
1.	Abstract	5-6
2.	Chapter 1: Introduction	7-9
3.	Motivation	10
4.	Chapter 2: Literature Review	11-13
5.	Gap Analysis	14
6.	Problem Statement	15
7.	Objectives	16
8.	Tools / Platform Used	17-18
9.	Chapter 3: Methodology	19-22
10.	Code & Output	23-35
11.	Chapter 4: Implementation & Results	36-39
12.	Conclusion & Future Work	40-42
13.	References	43
14.	Annexure I: Responsibility Chart	44
15.	Annexure II: Progress Reports duly signed by Mentors.	45
16.	Annexure III Research Paper (Published/Submitted)	

ABSTRACT

CO_2 Emissions became one of the major causes of Green House effect and Global Warming. There are many causes of CO_2 emissions. Due to Vehicles, Industries / Manufacturing, Electricity. The Motto is predicting the CO_2 Emissions and Identifying some of the main causes of it. So that Necessary actions can be taken. In this project, we want to put into practice a machine learning algorithm that will forecast CO_2 emissions and pinpoint some of the primary contributors. So that Necessary actions can be taken. Here the administrator trains the machine learning algorithm with a huge amount of data sets to predict C02 Emissions and features more responsible for Emissions so as to reduce their usage in future.

In order to better understand the effects of vehicle carbon dioxide (CO_2) emissions on the environment, a study paper is being written about them. It looks at how fuel burning in automobiles releases CO_2, one of the greenhouse gases that contribute to climate change. This research examines a range of vehicle types, including cars, trucks, and buses, considering variables that affect emissions, such as driving habits, fuel efficiency, and technological improvements. It evaluates the effects of elevated CO_2 levels on the environment, including air pollution and global warming, and looks into possible countermeasures. The study attempts to provide insights on lowering CO_2 emissions from automobiles, addressing the urgent need for sustainable transportation and a healthy environment. It does this by analyzing techniques such alternative fuels, hybrid technology, and legislative interventions.

One of the main contributors to the Green House effect and global warming is now CO_2 emissions. The reasons of CO_2 emissions are numerous. Because of transportation, manufacturing, and electricity. Predicting CO_2 emissions and pinpointing some of the primary drivers of them is the motto. in order for the required actions to be performed. In this project, we want to put into practice a machine learning algorithm that will forecast CO_2 emissions and pinpoint some of the primary contributors. Here, the administrator uses a vast array of data sets to train the machine learning algorithm to forecast carbon dioxide emissions and identify more responsible qualities that can lower future emissions.

forecasts to create effective preventative actions. By employing cutting-edge artificial intelligence to model and forecast vehicle CO_2 emissions, this study significantly advances the field. Several parameters, including vehicle class, engine size (L), cylinder transmission, fuel type, fuel consumption city (L/100 km), fuel consumption hwy (L/100 km), fuel consumption comb (L/100 km), fuel consumption comb (mpg), and CO_2 emissions (g/km) were used to build the model using the CO_2 emission by vehicles dataset from Kaggle. A deep learning long short-term memory network (LSTM) model and a bidirectional LSTM (BiLSTM) model were created in order to predict the CO_2 emissions that are generated by automobiles. Each of the models is effective. Four statistical assessment metrics were used by the researchers during the investigation: the determination coefficient (R^2), Pearson's correlation coefficient (R%), mean square error (MSE), and root mean square error (RMSE). The LSTM and BiLSTM models were developed and put

into use using the experiment datasets that Kaggle had collected. This paper examines the precise algorithm of carbon emissions from coal-fired power plants using Guangdong Province, one of the national first batch of low-carbon pilot provinces, as an example, in an effort to realize low-carbon development and strive to complete the carbon emissions reduction targets of the "13th Five-Year." This algorithm's computation fuel characteristic coefficient formula The fuel's proximate analysis data fits β, and the fuel characteristic coefficient is used to determine the CO2 level in flue gas. Next, the total amount of smoke is used to determine the carbon emissions. Lastly, the algorithm is validated using a thermal power plant in Guangzhou as an example, and it is contrasted with the coal consumption rate of power technique and the IPCC methodology.

CO2 emission prediction is crucial for effective climate change mitigation strategies and sustainable development planning. This abstract outline research focusing on predictive modelling techniques applied to CO2 emission trends. Utilizing historical data, machine learning algorithms, and statistical methodologies, the study aims to forecast future CO2 emissions trajectories. Existing literature review highlights methodologies, data sources, and limitations of CO2 emission prediction models. The proposed approach integrates various predictive algorithms like neural networks, regression analysis, and time series forecasting methods.

Factors influencing CO2 emissions, including economic indicators, population growth, energy consumption patterns, and technological advancements, are considered. Extensive validation exercises compare the proposed model's accuracy with existing methodologies. Sensitivity analysis evaluates the model's response to different scenarios, including policy interventions and socio-economic changes. Findings demonstrate the model's efficacy in accurately capturing CO2 emission trends across regions and sectors. The research contributes to advancing predictive capabilities in climate science, offering insights for policymakers, businesses, and environmental stakeholders.

KEYWORDS: Machine learning (ML), Data Mining (DM), Orange Application (Orange Data Mining), Predicting, CO2, Carbon Dioxide, Vehicles, multiple linear regression analysis.

CHAPTER 1

INTRODUCTION

The transportation sector stands as a major contributor to CO2 emissions worldwide. Cars, trucks, buses, airplanes, and ships collectively release vast amounts of CO2 as they burn fossil fuels for power. Factors like the number of vehicles on the road, the types of fuel used, and how efficiently these vehicles operate all influence the quantity of CO2 released.

As a result of the possible effects on next generations, climate change has become a hot topic in modern conversation. Globally, scientists have carried out copious research on the subject, demonstrating that extreme weather events and other environmental problems are already manifestations of climate change [1].

The number of automobiles on the planet is growing yearly, and as a result, so are their emissions. The emissions cause harm to people's health in addition to polluting the air. Since carbon dioxide is one of the primary causes of global warming but has no direct impact on people, it is the most underappreciated of all the emissions from cars. [2].

The impact of vehicle CO2 emissions extends beyond environmental concerns. It affects air quality, human health, and the stability of ecosystems. Additionally, countries around the world are setting targets to reduce CO2 emissions from vehicles to mitigate climate change and meet international agreements aimed at limiting global temperature rise. In today's bustling world, vehicles are an indispensable part of our lives, powering our daily commutes, transporting goods, and connecting communities. However, there's a hidden cost to this convenience: carbon dioxide (CO2) emissions. When vehicles burn fuel like gasoline or diesel to generate power, they release CO2—a greenhouse gas—into the atmosphere.

Imagine your car as a factory on wheels, producing an invisible exhaust called CO2. Just as smokestacks release gases when factories produce goods, vehicles emit CO2 as they operate. These emissions trap heat in the Earth's atmosphere, contributing to what scientists call the "greenhouse effect." While this natural process keeps the planet warm enough to support life, human activities, especially the burning of fossil fuels, have significantly increased the concentration of greenhouse gases, including CO2, leading to global warming and climate change.

"Carbon dioxide (CO2) emissions are the gases released when human are burn fossil fuels like coal, oil, and natural gas for energy. These emissions enter the atmosphere and contribute to the greenhouse effect, trapping heat and causing the planet's temperature to rise. This increase in temperature leads to climate change, impacting weather patterns, sea levels, and the overall health of our planet. Understanding and managing CO2 emissions is crucial to mitigate their harmful effects and work towards a more sustainable future."

As one of the primary contributors to climate change, greenhouse gas emissions, particularly those of carbon dioxide (CO_2), are considered to be among the world's most significant environmental issues.

This paper attempts to investigate the relation between CO_2 emissions and economic growth, industry structure, urbanization, research and development (R&D) investment, actual use of foreign capital, and growth rate of energy consumption in China between 2000 and 2018. China has committed to reaching carbon neutrality by 2060 and peaking its CO_2 emissions by 2030, therefore this study is crucial to the country.

The Carbon dioxide (CO_2) is a colourless, odourless and non-poisonous gas formed by combustion of carbon and in the respiration of living organisms and is considered a greenhouse gas. The release of greenhouse gases and/or their precursors into the atmosphere over a predetermined area and time period is referred to as emissions. Emissions of carbon dioxide, often known as CO_2 emissions, are produced when fossil fuels are burned and cement is made. These emissions also include carbon dioxide released when solid, liquid, and gas fuels are consumed and when gas is flared.

There CO_2 Emissions became one of the major causes of Greenhouse effect and Global Warming. There many causes of CO_2 emissions. Due to Vehicles, Industries / Manufacturing, Electricity, Livestock. The Motto is predicting the CO_2 Emissions and Identifying some of the main causes of it.

The issue of global warming is now one that affects all countries equally. The Intergovernmental Panel on Climate Change reported that scientists were more than 95% certain that most of global warming is caused by increasing concentrations of greenhouse gasses and other human (anthropogenic) activities That balance between earth and atmosphere affected by an increase in acid gasses charcoal or carbon dioxide (CO_2), methane (CH_4), nitrous oxide viii (N_2O), hydrofluorocarbons (HFC) and perfluorocarbons (PFC), are referred to as greenhouse gases more widely. In particular, CO_2 is a major cause of Global Warming. About eight billion tons per year of carbon in the form of CO_2 emitted globally through burning fossil fuels for transport and for the production of heat and electricity around the world.

Artificial Intelligence (AI) has the potential to transform several industries including healthcare and medicine, education, marketing, food technology, banking and financial services, travel, real estate, entertainment, and most importantly in logistics and transportation that also focuses on developing such solutions that are environmentally friendly.

The remaining by-product of burning water (H_2O) and carbon monoxide gas (CO), also known as carbon dioxide (CO_2), is the emission of carbon dioxide, which is a greenhouse gas. Carbon footprint is a notion used as a reference in the measuring of CO_2 emissions. Artificial Intelligence (AI) has the potential to transform several industries including healthcare and medicine, education, marketing, food technology, banking and financial services, travel, real estate, entertainment, and most importantly in logistics and transportation that also focuses on developing such solutions that are environmentally friendly.

The use of AI in smart cities will not only help in achieving efficiency but also mitigate environmental issues.

One of the main reasons of climate change and changing earth temperature is emission of CO2. Presence of gas in atmosphere varies time to time. Some gases stay for few years and some thousands of years. Some gases are more hazardous than others. According to the survey conducted by Inter governmental Panel on Climate Change the temperature of land and ocean is increased by average 0.85°C.
As a result, there have been increased gas emissions over the past three decades. It has ix decreased number of cold days and nights and also increased number of hot days and nights.
Carbon dioxide (CO2) is more meticulous for Green House effect. There are different factors for CO2 emission such as living organisms, fermentation, burning of coal, natural gas, and oil, trees, wood products, solid waste. Carbon footprint has become a buzzword widely.
We can see the online website for the home of carbon foot printing, carbon calculators, CO2 reduction, carbon offsetting and caring for the climate. One of the main reasons of climate change and changing earth temperature is emission of CO2. Presence of gas in atmosphere varies time to time. Some gases stay for few years and some thousands of years.

Some gases are more hazardous than others. According to the survey conducted by Inter governmental Panel on Climate Change the temperature of land and ocean is increased by average 0.85°C. As a result, there have been increased gas emissions over the past three decades. It has decreased number of cold days and nights and also increased number of hot days and nights. Carbon dioxide (CO2) is more meticulous for Green House effect. There are different factors for CO2 emission such as living organisms, fermentation, burning of coal, natural gas, and oil, trees, wood products, solid waste. Carbon footprint has become a buzzword widely. We can see the online website for the home of carbon foot printing, carbon calculators, CO2 reduction, carbon offsetting and caring for the climate.
In order to preserve the health of our planet, a smart city must be both sustainable and technologically sophisticated, with zero CO2 emissions. There will be a significant increase in the use of online delivery and self-driving automobiles. In light of these considerations, this research suggests a CO2 emission predictive analysis system that uses artificial intelligence (AI) to forecast the amounts of CO2 released by cargo-transporting cars based on fuel efficiency and distance driven. By making smarter and more ecologically conscious decisions, the intention is to assist both drivers and customers in becoming more responsible members of society.
Because they may select the least polluting route, users of online delivery services have more control over how much CO2 their orders' delivery cars will emit. Our suggested approach intends to give locals the capacity to order environmentally friendly goods and have them delivered to their house in an environmentally acceptable manner. In the past, people could only get around by walking or by riding horses, camels, and donkeys. After the invention of the wheel, people and commodities were eventually transported across greater distances by use of wheeled vehicles Shacharit's,carts,andwagons.
During the Industrial Revolution of the 18th and 19th centuries, railroads and steam-powered ships were introduced, resulting in enormous improvements to transportation. These means of transportation had a significant impact on the expansion of trade and industry as well as the speed and efficiency of travel and the delivery of products. The transportation industry underwent a significant transformation in the 20th century due to the extensive use of cars and airplanes. The invention of the vehicle transformed personal transportation and allowed individuals to move swiftly and conveniently to new destinations.

2.Motivation

In the realm of scientific inquiry and technological innovation, there exists a profound opportunity to drive meaningful change. Today, as we embark on the journey of predicting CO2 emissions, we stand at the forefront of a movement that seeks to safeguard our planet and secure a sustainable future for generations to come.

The task before us is daunting yet exhilarating. We are tasked with harnessing the power of data, machine learning, and predictive analytics to unveil the intricate patterns and trends underlying CO2 emissions. Through our relentless pursuit of knowledge and our unwavering commitment to excellence, we have the potential to revolutionize how we understand and mitigate the impact of human activity on our environment.

But our mission extends beyond mere prediction; it is a call to action, a rallying cry for change. With every line of code, we write, every model we refine, we inch closer to a future where renewable energy sources reign supreme, where carbon emissions are a relic of the past, and where the delicate balance of our ecosystems is preserved for generations to come.

As stewards of this planet, we hold a sacred responsibility to use our talents and expertise for the greater good. We are not merely data scientists or engineers; we are architects of possibility, builders of a brighter tomorrow. Our work today lays the foundation for a world where sustainability is not just a buzzword, but a way of life. So, let us press forward with determination and purpose, knowing that our efforts have the power to shape the course of history. Let us draw inspiration from the boundless potential that lies within us, and let us dare to dream of a future where our actions today echo through the annals of time.

Together, we have the power to defy expectations, to surpass limitations, and to forge a path towards a more sustainable and equitable world. So, let us rise to the challenge before us, united in our pursuit of a better future for all. The journey may be long and arduous, but the destination – a world where CO2 emissions are but a distant memory – is well worth the effort.

Our project is not just about data analysis; it's a testament to our commitment to harnessing technology for the greater good. By accurately forecasting CO2 emissions, we empower policymakers, industries, and individuals to make informed decisions that reduce our carbon footprint and mitigate climate change.

In the face of mounting environmental challenges, our project serves as a beacon of hope, demonstrating that through innovation and collaboration, we can overcome even the most daunting obstacles. Let us embrace this opportunity with passion and determination, knowing that our efforts today pave the way for a greener, healthier planet for future generations. Together, let's drive change and build a brighter tomorrow.

Chapter 2

3. LITERATURE REVIEW

During a literature survey, we collected some of the information about CO2 Emission prediction mechanisms currently being used.

CO2 Emissions Prediction Survey Employing Machine Learning Techniques:

The CO2 Emissions has become one of the main causes of Greenhouse gases and Global Warming. There many causes of CO2 emissions. Due to Vehicles, Industries / Manufacturing, Electricity, Livestock. The Motto is to predict the CO2 Emissions and Identifying some of the main causes of it So that Necessary actions can be taken.

Artificial Intelligence (AI) has the potential to transform several industries including healthcare and medicine, education, marketing, food technology, banking and financial services, travel, real estate, entertainment, and most importantly in logistics and transportation that also focuses on developing such solutions that are environmentally friendly.

This creates very huge amounts of data at any given time as more data needs to undergo analysis before a prediction can be made. Each of the techniques listed under regression has its own advantages and limitations over its other counterparts. One of the noteworthy techniques that were mentioned was decision trees.

During the last decade, the growth of energy consumption and CO2 emissions be more important than industry growth trends, it means that the reduction of energy consumption and CO2 emission or slightly does not cause a decrease in growth industries. The industrial sector is the main contributor of the total CO2 emissions in the world.

It is to be expected that there is significant worry regarding carbon dioxide emissions given the expansion of the manufacturing sector. The primary cause of emissions is energy use. This accounts for over 65% of total gas emissions, and if precise action is not taken, emissions are predicted to increase.

Coal and Electricity have been a preferred form for energy consumption and has consistently registered a higher growth rate than other forms of energy. Higher electrical power usage is more closely related to higher CO2 emission levels. Recently, many efforts have been put on the complex system of economy environment and relevant CO2 emission reduction issues. A new manufacturing method called as "green manufacturing," which is appropriate for a sustainable development strategy, is required in an effort to minimize CO2. Green manufacture is an economically-driven, system-wide, and integration approach to the reduction and elimination of all waste streams associated with the design, manufacture, use and disposal products and materials.

Zero probable safety issues, zero health risks for product consumers and operators, zero environmental contamination, and as much recycling and waste disposal as possible during the production process are the requirements to achieve green manufacturing. The way decision trees work is that they Decision trees use multiple algorithms to decide to split a node into two or more sub-nodes. The creation of sub-nodes increases the homogeneity of resultant sub-nodes. Stated differently, we might assert that the node's purity rises in relation to the target variable. One of the major issues the world is currently experiencing is climate change. CO2 emission is the largest contributor, and it is mainly released by human activities. About two thirds of greenhouse gas (GHG) emissions in Europe are attributed to the energy industry, and the sources used to produce power can have a significant impact on how much CO2 is released over time.

The use of machine learning and artificial intelligence techniques to predict the Co2 emissions is an increasing trend. More and more researchers invest their time every day in coming up with results that can further reduce the emissions of Carbon dioxide. Due to the vast number of options available, there can be n number of ways on how to predict the carbon dioxide emissions, but all methods don't work the same way. The output varies for each algorithm even if the same data set is being applied.

In the cited paper the carbon dioxide prediction has been carried out by using the random forest algorithm is being used to predict the emissions of carbon dioxide. This is just one way of looking at the problem by approaching it using a predictive model, using the random forest to predict the carbon dioxide emissions. However, there are always other factors that influence the carbon dioxide emissions, such as rural and urban population, energy use precipitate, renewable energy use etc.

Accurately predicting the co2 emissions is a challenging task, but the modern web has proved to be a very useful tool in making this task easier. Because of the data's interconnected style, it is simple to extract particular sentiments, which facilitates the development of relationships.

This paper's regression model will be used to examine CO2 emissions. To create the model data is collected from the World Bank from 1960 to 2014. The CO2 is measured in kt. After collecting data, the data is normalized. The model will predict the CO2 emission based on previous years data. To cross-validate the model the evaluation technique is used by dividing dataset into training data and testing data.

CO2 EMMISIONS FORECAST USING HISTORICAL DATA ANALYSIS:

The co2 emission prediction process is filled with uncertainty and can be influenced by multiple factors. Therefore, the co2 emissions plays an important role in Environmental changes. We took the data from the world bank. It contains more historical data over the years so as to give us the best fit model among all the taken machine learning models.

Understanding the nexus CO2 emissions and economic growth helps economies in formulating energy policies and developing energy resources in xiv sustainable ways. The relationship between economic growth and CO2 emissions has been the subject of many in-depth studies in recent years, but there hasn't been much study done in these areas in terms of qualitative systematic review and meta-analysis. This review paper's primary goal is to provide a thorough assessment of the connection between CO2 emissions and economic expansion. Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA), a qualitative systematic and meta-analysis approach, has been proposed in this regard, and the Web of Science database has been selected. Therefore, a review of 175 published articles appearing in 55 scholarly international journals between 1995 and 2017 has been achieved to reach a broad review of the nexus between economic growth and CO2 emissions with other indicators.

As a consequence, the chosen papers have been arranged according to the following criteria: author, year of publication, length of data, various methodologies, method of data analysis, indicator name, country, scope (single country and multiple countries), journals, results, and outcome in which they appeared. The findings of this study showed that the relationship between CO2 emissions and economic growth supports the need for policy alternatives that both constrain economic expansion and reduce emissions.

Given the fact that bidirectional causality exists, as far as economic growth increases or decreases, further CO2 emissions are stimulated in higher or lower levels and consequently, a potential reduction of the emissions should have an adverse influence on economic growth.

Technical analysis is done using by applying machine learning algorithms on historical data of co2 emissions. Typically, the process entails compiling news articles and data from social media platforms to extract user sentiments.

Other data like previous emissions data are also considered. The model was able to make predictions about co2 predictions and feature analysis will be done based on the carbon dioxide emissions. In this project the data sets collected are made to work on different machine learning algorithms like decision trees and random forest to get the machine learning model that gives more accuracy among all the algorithms.

4. Gap Analysis

Data Availability and Quality:

Gap: The availability and quality of data significantly impact the accuracy of CO2 emission predictions. Current datasets often suffer from incompleteness, inconsistency, and lack of granularity, hindering the precision of predictive models.

Opportunity: Addressing this gap requires collaboration with governmental agencies, research institutions, and industry partners to improve data collection methodologies, enhance data sharing initiatives, and implement quality control measures. Additionally, leveraging emerging technologies such as remote sensing and IoT devices can supplement existing datasets, enriching predictive capabilities.

Model Complexity and Interpretability:

Gap: Existing CO2 emission prediction models often exhibit high complexity, making them challenging to interpret and deploy in real-world scenarios. This complexity hampers stakeholder understanding and confidence in model outputs, limiting their practical utility.

Opportunity: Simplifying model architectures and enhancing interpretability through techniques such as feature selection, model visualization, and uncertainty quantification can bridge this gap. Moreover, fostering interdisciplinary collaborations between data scientists, domain experts, and policymakers can facilitate the development of transparent and actionable predictive models.

Temporal and Spatial Resolution:

Gap: Many CO2 emission prediction models operate at coarse temporal and spatial resolutions, overlooking fine-grained variations and localized trends. This limitation undermines the efficacy of localized mitigation strategies and inhibits the assessment of policy interventions' effectiveness.

Opportunity: Advancements in computational resources and modeling techniques enable the development of high-resolution spatiotemporal prediction models. By integrating satellite imagery, geographic information systems (GIS), and machine learning algorithms, we can enhance the granularity and accuracy of CO2 emission forecasts, empowering stakeholders with actionable insights at various scales.

5. Problem Statement

Data Quality and Availability:

The quality and availability of CO_2 emission data are critical determinants of prediction accuracy. However, existing datasets often suffer from incompleteness, inconsistency, and insufficient granularity. Inadequate data hinder the development of robust predictive models, limiting their applicability in guiding effective mitigation strategies.

Model Complexity and Interpretability:

Many CO_2 emission prediction models exhibit high complexity, rendering them difficult to interpret and deploy in practical settings. Model opacity undermines stakeholder trust and confidence, hindering the adoption of predictive insights for policy formulation and implementation. Simplifying model architectures and enhancing interpretability are essential for fostering transparent and actionable predictions.

Temporal and Spatial Resolution:

CO_2 emission prediction models commonly operate at coarse temporal and spatial resolutions, overlooking fine-grained variations and localized trends. This limitation impedes the assessment of regional disparities in emissions and undermines the effectiveness of localized mitigation efforts. Enhancing resolution through advanced modelling techniques is crucial for capturing nuanced emission dynamics and informing targeted interventions.

Incorporation of Socioeconomic Factors:

Socioeconomic drivers significantly influence CO_2 emissions, yet traditional prediction models often overlook their impact. Ignoring socioeconomic factors limits the comprehensiveness of predictive insights and hinders the development of tailored mitigation strategies. Integrating socioeconomic data into predictive frameworks is essential for capturing the complex interplay between human activities and emission trends.

Uncertainty Quantification and Risk Assessment:

Despite advancements in predictive modelling, accurately quantifying and communicating uncertainties remains a challenge. Decision-makers require probabilistic forecasts and risk assessments to navigate uncertain futures effectively. Improving uncertainty quantification techniques and enhancing transparency in uncertainty communication are essential for facilitating risk-informed decision-making and enhancing stakeholder confidence.

6. Objectives

- **To predict the carbon dioxide emissions using ML model.**

  The objective of predicting carbon dioxide (CO2) emissions using machine learning (ML) models is to develop accurate and reliable predictive frameworks that leverage historical data and relevant features to forecast future emissions levels. ML models offer a data-driven approach to understanding the complex relationships between various factors influencing CO2 emissions, including economic activities, energy consumption, industrial processes, transportation patterns, land use changes, and environmental factors.

- **To identify the features that are causing more carbon dioxide emissions.**

  The goal of identifying features that contribute to higher carbon dioxide (CO2) emissions is to understand the underlying drivers of emissions and inform targeted mitigation strategies. This process involves analysing various factors that influence CO2 emissions across different sectors and activities, such as energy production, transportation, industrial processes, land use changes, and socioeconomic factors.

- **To reduce their usage in future.**

To reduce the usage of fossil fuels and other high-emission activities in the future, several strategies can be implemented:

Transition to Renewable Energy: Promoting the adoption of renewable energy sources such as solar, wind, hydroelectric, and geothermal power can significantly reduce reliance on fossil fuels for electricity generation and heating. Incentives, subsidies, and investment in renewable energy infrastructure are crucial for accelerating this transition.

Energy Efficiency Improvements: Enhancing energy efficiency across sectors, including buildings, transportation, and industry, can reduce energy consumption and associated CO2 emissions. Implementing energy-efficient technologies, improving building insulation, and adopting fuel-efficient vehicles are examples of measures that can contribute to lower energy usage.

Shift to Low-Carbon Transportation: Encouraging the use of public transportation, cycling, walking, and electric vehicles can reduce greenhouse gas emissions from the transportation sector. Investing in public transit infrastructure, expanding bike lanes, and incentivizing electric vehicle adoption through tax credits and rebates are effective strategies.

7. Tools/platform Used

7.1Details of tools, software, and equipment utilized.

PLATFORM USED

For this project, we have used various latest technologies which will be evaluated in this chapter with every detail of why it is used.

PROGRAMMING LANGUAGE: **PYTHON**

We have used Python language as it is relatively new as compared to other languages like HTML, JAVASCRIPT, etc and comes with so many features. We can perform Machine Learning, Computer Vision, Artificial Intelligence, etc with python and construction of GUI application is also easily achieved in Python.

Python is a widely used general-purpose, high level programming language. It was created by Guido van Rossum in 1991 and further developed by the Python Software Foundation. It was designed with an emphasis on code readability, and its syntax allows programmers to express their concepts in fewer lines of code. Python is a programming language that lets you work quickly and integrate systems more efficiently. There are two major Python versions: **Python 2 and Python 3**

Reasons for Selecting this language:

1. Short and Concise Language.
2. Easy to Learn and use.
3. Good Technical support over Internet
4. Many Packages for different tasks.
5. Run on Any Platform.
6. Modern and OOP language

Some specific features of Python are as follows:

1. An interpreted (as opposed to compiled) language. Contrary to e.g. C or Fortran, one does not compile Python code before executing it. In addition, Python can be used **interactively**: many Python interpreters are available, from which commands and scripts can be executed.

2. A free software released under an **open-source** license: Python can be used and distributed free of charge, even for building commercial software.
3. **Multi-platform**: Python is available for all major operating systems, Windows, Linux/Unix, MacOS X, most likely your mobile phone OS, etc.

4. A very readable language with clear non-verbose syntax.

5. A language for which a large variety of high-quality packages are available for various applications, from web frameworks to scientific computing.

6. A language very easy to interface with other languages, in particular HTML and CSS.

7. Some other features of the language are illustrated just below. For example, Python is an object-oriented language, with dynamic typing (the same variable can contain objects of different types during the course of a program).

7.2 ENVIRONMENTAL SETUP

SOFTWARE REQUIREMENTS

Below are the requirements to run this software:

1. Windows/Linux/Mac OS any version, hence it can run on any platform.
2. Python3, it needs python to be installed in system to run successfully.
3. Packages in python -
   a. openCV
   b. skimage
   c. numpy
   d. tkinter

7.3 HARDWARE REQUIREMENTS

In terms of hardware requirements there is not much required at all but still below requirements are must:

1. Working PC or Laptop
2. Webcam with drivers installed
3. Flashlight/ LED if using this at night.

7.4 PLATFORMS ALREADY TESTED ON:

It is tested on Linux Mint, Linux Ubuntu, Windows 7 , Windows 10 and Windows 11.

Chapter 3

8: METHODOLOG

The flowchart of our research methods and algorithms used in this research paper can be seen in FIGURE3, emission in Ventry method was applied by collecting relevant data from various sources then a statistical technique (Regression analysis) was used to analyse the relationship between one or more independent variables and a dependent variable, such as CO2 emissions to help identify the factors that influence CO2 emissions and predict the impact of changes in these factors. Several machine learning algorithms such as neural networks, decision trees, and random forests were used to analyse our large datasets and identify patterns in the data. These algorithms predicted future CO2 emissions based on historical data and identified the factors that have the greatest impact on CO2 emissions.

8.1 Data Collection and Pre-processing

Two sources provided the dataset used in this work. The first data source is a vehicle regis traction dataset derived from Statistics data; model year, make, model, vehicle class, engine size and cylinders summarized in TABLE 1.

Table 1: Vehicle registration dataset features

Columns	Details
Model Year	The year the car's model was built
Make	The company of the vehicle
Model	The car's model
Vehicle Class	The class of vehicle depends on their utility, capacity and weight
Engine Size	The size of the engine used in Litre
Cylinders	The number of cylinders

The second source was gotten from the Government of Canada, and it contains fuel consumption rating data which includes the following features in TABLE 2: transmission, fuel type, fuel consumption city, fuel consumption highway, fuel consumption comb and our target variable CO2 emissions.

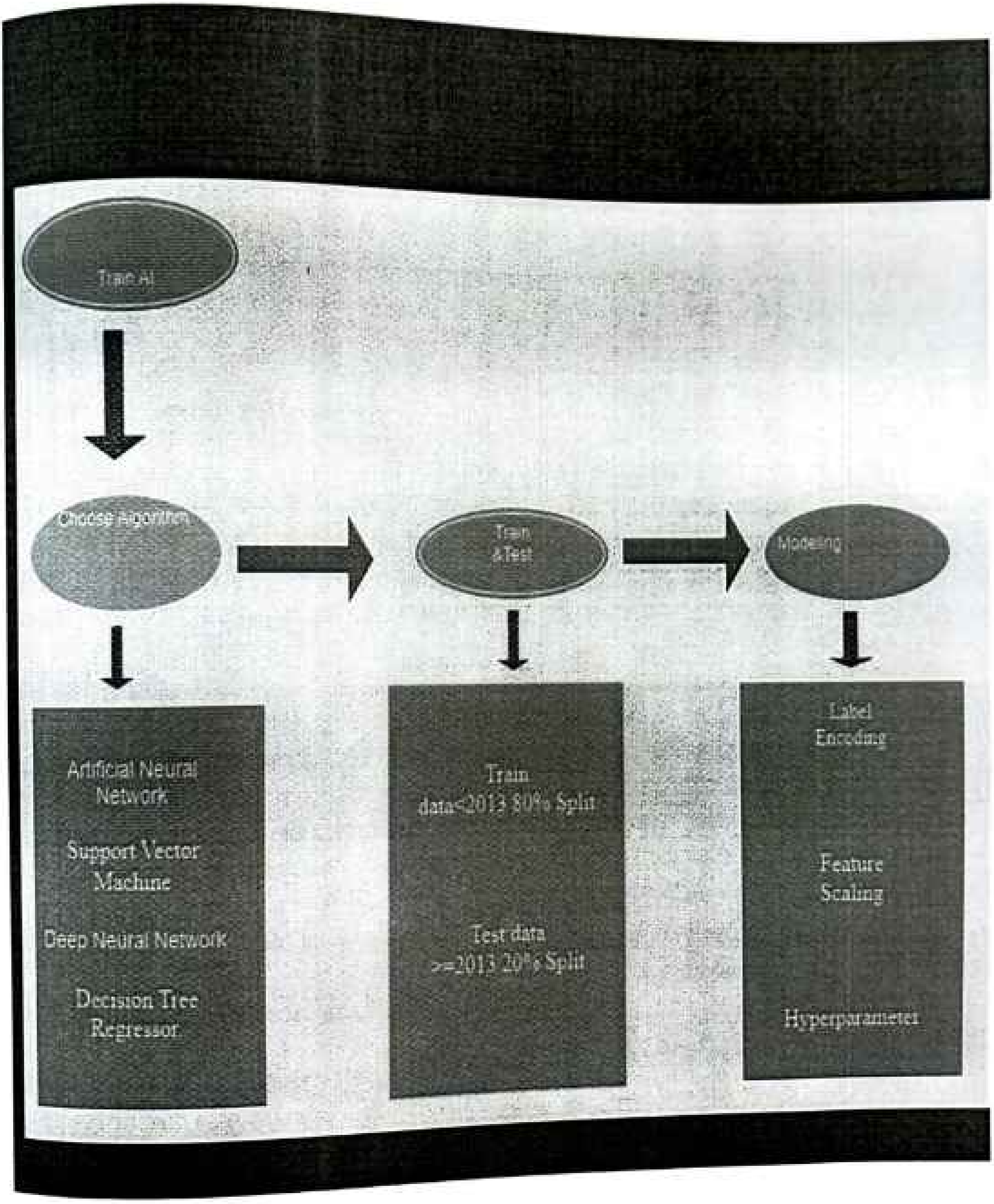
Table 2: Fuel consumption rating dataset features

Columns	Details
Transmission	The transmission type with the number of gears
Fuel Type	The type of fuel used
Fuel Consumption City	The fuel consumption on city roads (L/100 km)
Fuel Consumption HWY	The fuel consumption on highways (L/100 km)

Fuel Consumption Comb	The combined fuel consumption (55% city, 45% highway) is shown in L/100 km
CO2 Emission	The amount of carbon dioxide emissions rate

Figure 3: ML Algorithm flowchart

8.2. Flow chart:



Three new columns were produced via feature engineering the dataset. The first new column named MODEL FREQ, encodes and counts the frequency of occurrences of the vehicle model. The second new column, FUEL_CONS_HWY_CITY sums up the fuel consumption of both city and highway and finally the last new column, ENG_POWER computes the product of 'Number of Cylinders' and the' engine size' to determine the engine power and fuel consumption. TABLE 4 shows how the data set looks after feature engineering.

Label encoding on the categorical features transformed the following categorical variables in the dataset ['MAKE', 'MODEL', 'VEHICLE CLASS', 'TRANSMISSION', 'FUEL TYPE'] into nu merical values. The final table output can be seen in TABLE 5. Label encoding makes the data ready for scaling and modelling.

To examine the performance success of the machine learning algorithms in terms of CO2 emissions forecasting, the dataset with model years between 1995 to 2012 is used to train the algorithms, and then model years within the last 9 years (2013–2021) are forecasted with different algorithms.

8.3 Machine Learning Algorithms

The train test split method comes next after deciding the X independent variables which in this data set are ['MODEL YEAR', 'MAKE', 'MODEL', 'VEHICLE CLASS', 'ENGINE SIZE(L)', 'CYLINDERS','TRANSMISSION','FUELTYPE','FUELCONSUMPTIONCITY(L/100km)', 'FUELCONSUMPTIONHWY(L/100km)','FUELCONSUMPTIONCOMB(L/100km)','FUEL NSUMPTIONCOMB (mpg)','MODELFREQ','FUEL_CONS_HWY_CITY','ENG_POWER'] and ['CO2 EMISSIONS,' the target variable, also known as the Y variable. TABLE3showstheproportion of the train test split on the dataset. The modelling process includes Label encoding, feature scaling and hyperparameter tuning after which the best-performing algorithm is used to forecast the CO2 emission from 2022– 2050 on retail sales in Canada.

In this research study, different machine learning algorithms were used– Artificial Neural Network (ANN), Support Vector Machine (SVM), Deep Neural Network (DNN), Decision Tree Regressor and Gradient Boost Regressor (GBR) are studied by using the tool Google Colab in the prediction of transportation-related CO2 emissions in Canada.

8.3.1 Artificial Neural Network (ANN)

In our study, we employ an artificial neural network (ANN) as a computational model for tasks such as prediction, classification, and decision-making. The ANN consists of artificial neurons, like those found in the human brain. Neurons within the network are connected, and the strength of these connections, known as weights, modifies the inputs. Each neuron processes inputs and produces an output using a non-linear activation function.

For our supervised learning study, both inputs and outputs are provided to the network. In order to modify the weights, the network compares its outputs to the intended outputs and propagates faults back through the system. This iterative process continues as the weights are continuously adjusted. The structure of the ANN model algorithm is depicted in FIGURE 4.

8.3.2 Support Vector Machine (SVM)

One of the most often used supervised learning methods for machine learning's classification or regression is Support Vector Machine (SVM). The goal of the SVM algorithm is to create the best line or decision boundary called a hyperplane, that can segregate n-dimensional space into classes so that we can Easily assign the new data point to the appropriate category going forward. SVM chooses the extreme points/vectors that help in creating the hyperplane called support vectors.

8.3.3 Deep Neural Network (DNN)

A deep neural network (DNN) is an ANN with multiple layers of interconnected nodes between the input and output layers. In supervised learning and reinforcement learning challenges, neural networks are extensively utilized. These networks are built using interconnected layers. A neural network's primary function is to take in a collection of inputs, process them through more sophisticated calculations, and then produce an output in order to solve issues in the real world.

8.3.4 Decision Tree Regressor

A decision tree is a non-parametric supervised learning algorithm used for classification and re aggression. It consists of a hierarchical tree structure with nodes and branches FIGURE 5. shows a hierarchical, tree structure, which consists of a root node, branches, internal nodes, and leaf nodes. The algorithm selects the best feature to split the data into different classes or categories and the process continues until the tree is fully grown or a stopping criterion is met. The decision tree can then be used for predictions on new data by following the tree's path based on input feature values.

9.Code and Output

1. Import Libraries

```
!pip install mlxtend -qq
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
import statsmodels
from scipy import stats
import statsmodels.api as sm
from scipy.stats import shapiro
import statsmodels.stats.api as sms
from statsmodels.compat import lzip
from statsmodels.formula.api import ols
from statsmodels.stats.anova import anova_lm
from statsmodels.tools.eval_measures import rmse
from statsmodels.graphics.gofplots import qqplot
from statsmodels.stats.outliers_influence import variance_inflation_factor
from sklearn.decomposition import PCA
from sklearn.preprocessing import StandardScaler, MinMaxScaler
from sklearn.metrics import mean_squared_error, mean_absolute_error
from sklearn.feature_selection import RFE
from mlxtend.feature_selection import SequentialFeatureSelector as sfs
from sklearn.linear_model import Lasso, Ridge, ElasticNet, SGDRegressor, LinearRegression
```

from sklearn.model_selection import KFold, LeaveOneOut, GridSearchCV, cross_val_score, train_test_split

3. Read Data

data = pd.read_csv("../input/co2-emission-by-vehicles/CO2 Emissions_Canada.csv")

data.head()

data.info()

Table of Contents

1. **[Import Libraries](#import_lib)**

2. **[Set Options](#set_options)**

3. **[Read Data](#Read_Data)**

4. **[Exploratory Data Analysis](#data_preparation)**

- 4.1 - [Preparing the Dataset](#Data_Preparing)

  - 4.1.1 - [Data Dimension](#Data_Shape)

  - 4.1.2 - [Data Types](#Data_Types)

  - 4.1.3 - [Missing Values](#Missing_Values)

  - 4.1.4 - [Duplicate Data](#duplicate)

- 4.2 - [Understanding the Dataset](#Data_Understanding)

  - 4.2.1 - [Summary Statistics](#Summary_Statistics)

  - 4.2.2 - [Correlation](#correlation)

  - 4.2.3 - [Analyze Categorical Variables](#analyze_cat_var)

  - 4.2.4 - [Anaylze Target Variable](#analyze_tar_var)

  - 4.2.5 - [Analyze Relationship Between Target and Independent Variables]

  - 4.2.6 - [Feature Engineering] (#feature_eng)

5. **[Data Pre-Processing] (#data_pre)**

- 5.1 - [Outliers](#out)
  - 5.1.1 - [Discovery of Outliers](#dis_out)
  - 5.1.2 - [Removal of Outliers](#rem_out)
  - 5.1.3 - [Rechecking of Correlation](#rec_cor)
- 5.2 - [Categorical Encoding](#cat_enc)

6. **[Building Multiple Linear Regression Models](#bui_mlr_mod)**
- 6.1 - [Multiple Linear Regression - Basic Model](#bas_mod)
- 6.2 - [Feature Transformation](#fea_tra)
- 6.3 - [Feature Scaling](#fea_sca)
- 6.4 - [Multiple Linear Regression - Full Model - After Feature Scaling](#mod_aft_sca)
- 6.5 - [Assumptions Before Multiple Linear Regression Model](#ass_bef)
  - 6.5.1 - [Assumption #1: If Target Variable is Numeric](#tgt_num)
  - 6.5.2 - [Assumption #2: Presence of Multi-Collinearity](#pre_mul_col)
- 6.6 - [Multiple Linear Regression - Full Model - After PCA](#mod_pca)
- 6.7 - [Feature Selection](#fea_sel)
  - 6.7.1 - [Forward Selection](#for_sel)
  - 6.7.2 - [Backward Elimination](#bac_eli)
- 6.8 - [Multiple Linear Regression - Full Model - After Feature Selection](#mod_fea_sel)
- 6.9 - [Assumptions After Multiple Linear Regression Model](#ass_aft)
  - 6.9.1 - [Assumption #1: Linear Relationship Between Dependent and Independent Variable
  - 6.9.2 - [Assumption #2: Checking for Autocorrelation](#che_aut_cor)
  - 6.9.3 - [Assumption #3: Checking for Heterskedacity](#che_het)
  - 6.9.4 - [Assumption #4: Test for Normality](#tes_nor)
    - 6.9.4.1 - [Q-Q Plot](#qq_plt)
    - 6.9.4.2 - [Shapiro Wilk Test](#sha_wil_tes)

7. **[Model Evaluation](#mod_eva)**

- 7.1 - [Measures of Variation](#mea_var)

- 7.2 - [Inferences about Intercept and Slope](#inf_int_slo)

- 7.3 - [Confidence Interval for Intercept and Slope](#con_int_slo)

- 7.4 - [Compare Regression Results](#com_reg_res)

8. **[Model Performance](#mod_per)**

- 8.1 - [Mean Square Error(MSE)](#mse)

- 8.2 - [Root Mean Squared Error(RMSE)](#rmse)

- 8.3 - [Mean Absolute Error(MAE)](#mae)

- 8.4 - [Mean Absolute Percentage Error(MAPE)](#mape)

- 8.5 - [Resultant Table](#res_tab)

9. **[Model Optimization](#mod_opt)**

- 9.1 - [Bias](#bias)

- 9.2 - [Variance](#var)

- 9.3 - [Model Validation](#mod_val)

  - 9.3.1 - [Cross Validation](#cro_val)

  - 9.3.2 - [Leave One Out Cross Validation(LOOCV)](#loocv)

- 9.4 - [Gradient Descent](#gra_des)

- 9.5 - [Regularization](#reg)

  - 9.5.1 - [Ridge Regression Model](#ridge)

  - 9.5.2 - [Lasso Regression Model](#lasso)

  - 9.5.3 - [Elastic Net Regression Model](#ela_net)

  - 9.5.4 - [Grid Search CV](#gri_sea)

8. Model Performance


```
train_pred = MLR_full_model.predict(X_train)
test_pred = MLR_full_model.predict(X_test)
```

8.1 Mean Squared Error (MSE)


```
mse_train = round(mean_squared_error(y_train, train_pred),4)
mse_test = round(mean_squared_error(y_test, test_pred),4)
print("Mean Squared Error (MSE) on training set: ", mse_train)
print("Mean Squared Error (MSE) on test set: ", mse_test)
```

8.2 Root Mean Squared Error (RMSE)

```
mse_train = mean_squared_error(y_train, train_pred)
rmse_train = round(np.sqrt(mse_train), 4)
mse_test = mean_squared_error(y_test, test_pred)
rmse_test = round(np.sqrt(mse_test), 4)
print("Root Mean Squared Error (RMSE) on training set: ", rmse_train)
print("Root Mean Squared Error (RMSE) on test set: ", rmse_test)
```

8.3 Mean Absolute Error (MAE)

```
mae_train = round(mean_absolute_error(y_train, train_pred),4)
mae_test = round(mean_absolute_error(y_test, test_pred),4)
print("Mean Absolute Error (MAE) on training set: ", mae_train)
print("Mean Absolute Error (MAE) on test set: ", mae_test)
```

8.4 Mean Absolute Percentage Error (MAPE)

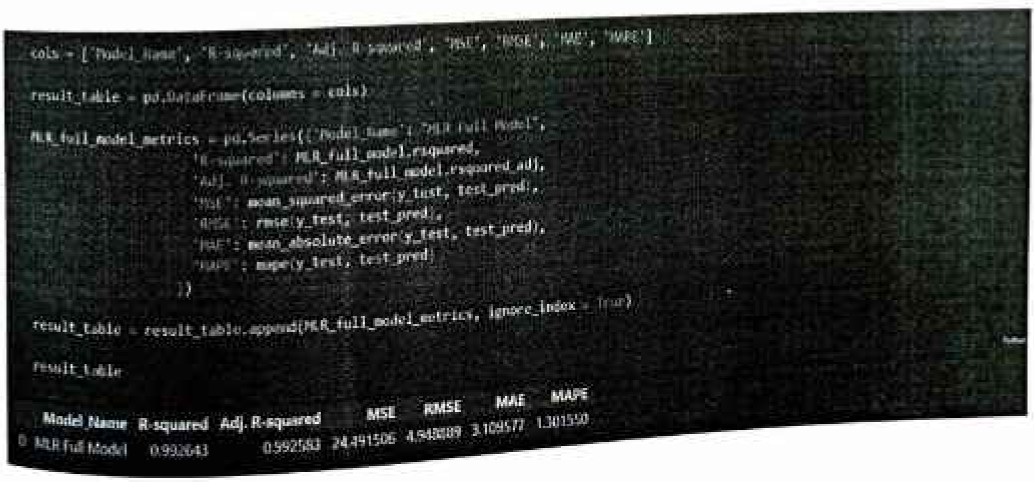
```
def mape(actual, predicted):
return (np.mean(np.abs((actual - predicted) / actual)) * 100)
mape_train = round(mape(y_train, train_pred),4)
mape_test = round(mape(y_test, test_pred),4)
```

```
print("Mean Absolute Percentage Error (MAPE) on training set: ", mape_train)
print("Mean Absolute Percentage Error (MAPE) on test set: ", mape_test)
# 9. Model Optimization<a id="mod_opt"></a>
# 9.5 Regularization<a id="reg"></a>
## 9.5.1 Ridge Regression Model<a id="ridge"></a>
ridge = Ridge(alpha = 0.1, max_iter = 500)
ridge.fit(X_train, y_train)
update_score_card(algorithm_name='Ridge Regression(with alpha=.1)', model=ridge, alpha = .1)
print('RMSE on test set:', get_test_rmse(ridge))
ridge = Ridge(alpha = 1, max_iter = 500)
ridge.fit(X_train, y_train)
update_score_card(algorithm_name='Ridge Regressio (with alpha=1)', model = ridge, alpha = 1)
print('RMSE on test set:', np.round(get_test_rmse(ridge),2))
ridge = Ridge(alpha = 2, max_iter = 500)
ridge.fit(X_train, y_train)
update_scorecard(algorithm_name='Ridge Regression (with alpha = 2)', model = ridge, alpha =2)
print('RMSE on test set:', get_test_rmse(ridge))
ridge = Ridge(alpha = 0.5, max_iter = 500)
ridge.fit(X_train, y_train)
update_score_card(algorithm_name='Ridge Regression(with alpha=.5)', model=ridge, alpha = .5)
print('RMSE on test set:', get_test_rmse(ridge))
plt.subplot(1,2,1)
plot_coefficients(MLR_model, 'Linear Regression (OLS)')
plt.subplot(1,2,2)
plot_coefficients(ridge, 'Ridge Regression (alpha = 0.5)')
```

```
plt.tight_layout()
## 9.5.2 Lasso Regression Model<a id="lasso"></a>
lasso = Lasso(alpha = 0.01, max_iter = 500)
lasso.fit(X_train, y_train)
print('RMSE on test set:', get_test_rmse(lasso))
plt.subplot(1,2,1)
plot_coefficients(MLR_model, 'Linear Regression (OLS)')
plt.subplot(1,2,2)
plot_coefficients(lasso, 'Lasso Regression (alpha = 0.01)')
plt.tight_layout()
lasso = Lasso(alpha = 0.05, max_iter = 500)
lasso.fit(X_train, y_train)
print('RMSE on test set:', get_test_rmse(lasso))
plt.subplot(1,2,1)
plot_coefficients(MLR_model, 'Linear Regression (OLS)')
plt.subplot(1,2,2)
plot_coefficients(lasso, 'Lasso Regression (alpha = 0.05)')
plt.tight_layout()
df_lasso_coeff = pd.DataFrame({'Variable': X.columns, 'Coefficient': lasso.coef_})
print('Insignificant variables obtained from Lasso Regression when alpha is 0.05')
df_lasso_coeff.Variable[df_lasso_coeff.Coefficient == 0].to_list()
update_score_card(algorithm_name = 'Lasso Regression', model = lasso, alpha = '0.05')
## 9.5.3 Elastic-Net Regression Model<a id="ela_net"></a>
enet = ElasticNet(alpha = 0.1, l1_ratio = 0.55, max_iter = 500)
enet.fit(X_train, y_train)
```

```
update_score_card(algorithm_name = 'Elastic Net Regression', model = enet, alpha = '0.1',
l1_ratio = '0.55')

print('RMSE on test set:', get_test_rmse(enet))

enet = ElasticNet(alpha = 0.1, l1_ratio = 0.1, max_iter = 500)

enet.fit(X_train, y_train)

update_score_card(algorithm_name = 'Elastic Net Regression', model = enet, alpha = '0.1',
l1_ratio = '0.1')

print('RMSE on test set:', get_test_rmse(enet))

enet = ElasticNet(alpha = 0.1, l1_ratio = 0.01, max_iter = 500)

enet.fit(X_train, y_train)

update_score_card(algorithm_name = 'Elastic Net Regression', model = enet, alpha = '0.1',
l1_ratio = '0.01')

print ('RMSE on test set:', get_test_rmse(enet))

plt.subplot(1,2,1)

plot_coefficients(MLR_model, 'Linear Regression (OLS)')

plt.subplot(1,2,2)

plot_coefficients(enet, 'Elastic Net Regression')

plt.tight_layout()
```

Screenshots



```
cols = ['Model_Name', 'R-squared', 'Adj. R-squared', 'MSE', 'RMSE', 'MAE', 'MAPE']

result_table = pd.DataFrame(columns = cols)

MLR_full_model_metrics = pd.Series({'Model_Name': "MLR Full Model",
                 'R-squared': MLR_full_model.rsquared,
                 'Adj. R-squared': MLR_full_model.rsquared_adj,
                 'MSE': mean_squared_error(y_test, test_pred),
                 'RMSE': rmse(y_test, test_pred),
                 'MAE': mean_absolute_error(y_test, test_pred),
                 'MAPE': mape(y_test, test_pred)
                 })

result_table = result_table.append(MLR_full_model_metrics, ignore_index = True)

result_table
```

	Model_Name	R-squared	Adj. R-squared	MSE	RMSE	MAE	MAPE
0	MLR Full Model	0.992643	0.992583	24.491506	4.948889	3.109577	1.301550

```
sgd = SGDRegressor(random_state = 10)
linreg_with_SGD = sgd.fit(X_train, y_train)

print('RMSE on train set:', get_train_rmse(linreg_with_SGD))
print('RMSE on test set:', get_test_rmse(linreg_with_SGD))

RMSE on train set: 4.3716
RMSE on test set: 4.9856
```

Linear regression

```
ridge = Ridge(alpha = 0.1, max_iter = 500)
ridge.fit(X_train, y_train)

update_score_card(algorithm_name='Ridge Regression (with alpha = 0.1)', model = ridge, alpha = 0.1)

print('RMSE on test set:', get_test_rmse(ridge))

RMSE on test set: 4.9461

ridge = Ridge(alpha = 1, max_iter = 500)
ridge.fit(X_train, y_train)

update_score_card(algorithm_name='Ridge Regression (with alpha = 1)', model = ridge, alpha = 1)

print('RMSE on test set:', np.round(get_test_rmse(ridge),2))

RMSE on test set: 4.95
```

Ridge

```
ridge = Ridge(alpha = 2, max_iter = 500)
ridge.fit(X_train, y_train)

update_score_card(algorithm_name='Ridge Regression (with alpha = 2)', model = ridge, alpha = 2)

print('RMSE on test set:', get_test_rmse(ridge))

RMSE on test set: 5.0109

ridge = Ridge(alpha = 0.5, max_iter = 500)
ridge.fit(X_train, y_train)

update_score_card(algorithm_name='Ridge Regression (with alpha = 0.5)', model = ridge, alpha = 0.5)

print('RMSE on test set:', get_test_rmse(ridge))

RMSE on test set: 4.9424
```

Lasso

```
enet = ElasticNet(alpha = 0.1, l1_ratio = 0.55, max_iter = 500)
enet.fit(X_train, y_train)

update_score_card(algorithm_name = 'Elastic Net Regression', model = enet, alpha = '0.1', l1_ratio = '0.55')

print('RMSE on test set:', get_test_rmse(enet))

RMSE on test set: 15.5696


enet = ElasticNet(alpha = 0.1, l1_ratio = 0.1, max_iter = 500)
enet.fit(X_train, y_train)

update_score_card(algorithm_name = 'Elastic Net Regression', model = enet, alpha = '0.1', l1_ratio = '0.1')

print('RMSE on test set:', get_test_rmse(enet))

RMSE on test set: 19.1626
```

ElasticNet

```
tuned_paramaters = [{'alpha':[1e-15, 1e-10, 1e-8, 1e-4,1e-3, 1e-2, 0.1, 1, 5, 10, 20, 40, 60, 80, 100]}]

ridge = Ridge()
ridge_grid = GridSearchCV(estimator = ridge,
                          param_grid = tuned_paramaters,
                          cv = 10)

ridge_grid.fit(X_train, y_train)

print('Best parameters for Ridge Regression: ', ridge_grid.best_params_, '\n')
print('RMSE on test set:', get_test_rmse(ridge_grid))

Best parameters for Ridge Regression:  {'alpha': 0.01}

RMSE on test set: 4.9486
```

Ridge regression using grid cv

```
tuned_paramaters = [{'alpha':[1e-15, 1e-10, 1e-8, 0.0001, 0.001, 0.01, 0.1, 1, 5, 10, 20]}]

lasso = Lasso()
lasso_grid = GridSearchCV(estimator = lasso,
                          param_grid = tuned_paramaters,
                          cv = 10)

lasso_grid.fit(X_train, y_train)

print('Best parameters for Lasso Regression: ', lasso_grid.best_params_, '\n')
print('RMSE on test set:', get_test_rmse(lasso_grid))

Best parameters for Lasso Regression:  {'alpha': 0.001}

RMSE on test set: 4.9495


update_score_card(algorithm_name = 'Lasso Regression (using GridSearchCV)',
                  model = lasso_grid,
                  alpha = lasso_grid.best_params_.get('alpha'))
```

Lasso reg. using grid cv

```python
tuned_paramaters = [{'alpha':[0.0001, 0.001, 0.01, 0.1, 1, 5, 10, 20, 40, 60],
                     'l1_ratio':[0.0001, 0.0002, 0.001, 0.01, 0.1, 0.2, 0.4, 0.55]}]

enet = ElasticNet()
enet_grid = GridSearchCV(estimator = enet,
                         param_grid = tuned_paramaters,
                         cv = 10)

enet_grid.fit(X_train, y_train)

print('Best parameters for Elastic Net Regression: ', enet_grid.best_params_, '\n')
print('RMSE on test set:', get_test_rmse(enet_grid))

Best parameters for Elastic Net Regression:  {'alpha': 0.0001, 'l1_ratio': 0.55}

RMSE on test set: 4.9445

update_score_card(algorithm_name = 'Elastic Net Regression (using GridSearchCV)',
                  model = enet_grid,
                  alpha = enet_grid.best_params_.get('alpha'),
                  l1_ratio = enet_grid.best_params_.get('l1_ratio'))
```

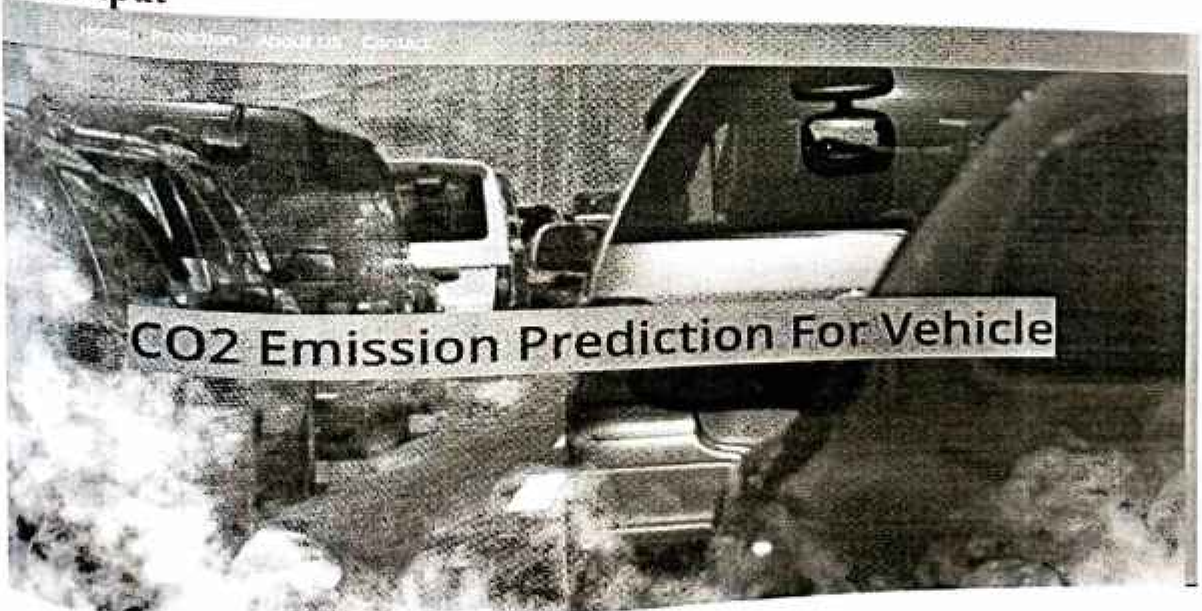
Net elastic using grid cv

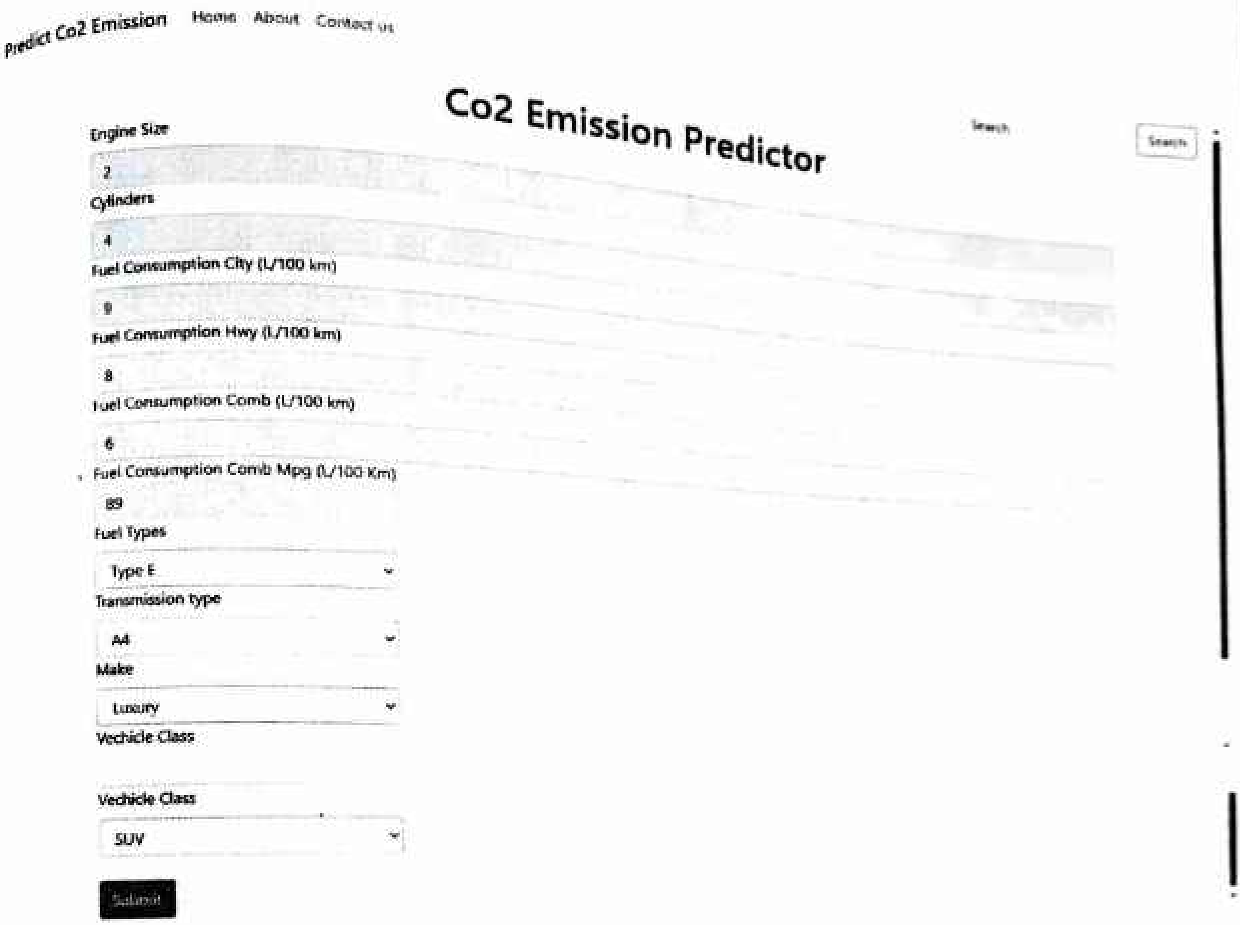
```python
score_card = score_card.sort_values('Test_RMSE').reset_index(drop = True)
score_card.style.highlight_min(color = 'lightblue', subset = 'Test_RMSE')
```

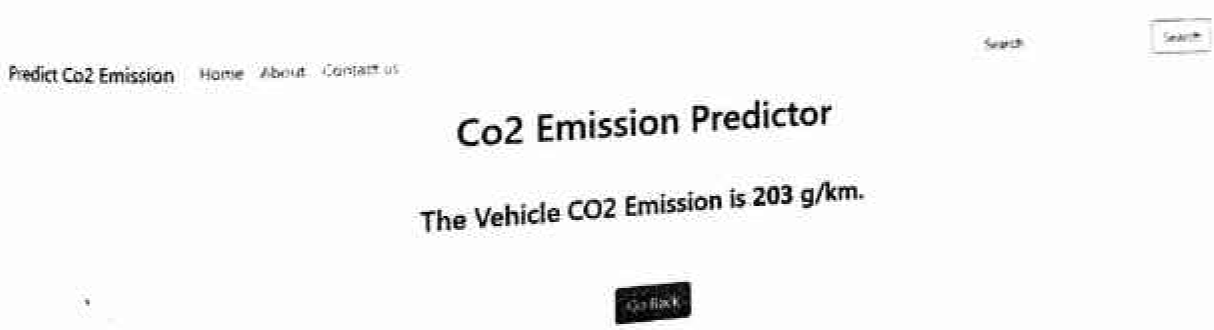
	Model_Name	Alpha (Wherever Required)	l1-ratio	R-Squared	Adj. R-Squared	Train_RMSE	Test_RMSE	Test_MAPE
0	Ridge Regression (with alpha = 0.5)	0.500000	-	0.992608	0.992546	4.320100	[illegible]	1.295323
1	Elastic Net Regression (using GridSearchCV)	0.000100	0.550000	0.992638	0.992576	4.311400	4.944500	1.298470
2	Ridge Regression (with alpha = 0.1)	0.100000	-	0.992642	0.992580	4.310400	4.946100	1.299823
3	Ridge Regression (using GridSearchCV)	0.010000	-	0.992643	0.992581	4.309900	4.940600	1.301376
4	Lasso Regression (using GridSearchCV)	0.001000	-	0.992642	0.992580	4.310200	4.949500	1.300953
5	Ridge Regression (with alpha = 1)	1	-	0.992512	0.992449	4.348300	4.952200	1.294861
6	Linear Regression (using SGD)	-	-	0.992431	0.992367	4.371600	4.985600	1.316064
7	Ridge Regression (with alpha = 2)	2	-	0.992171	0.992106	4.446000	5.010900	1.310282
8	Lasso Regression	0.05	-	0.990470	0.990390	4.905500	5.514400	1.458071
9	Elastic Net Regression	0.1	0.55	0.903067	0.902251	15.644600	15.569600	4.825320
10	Elastic Net Regression	0.1	0.1	0.852578	0.851336	19.293500	19.162600	6.153716
11	Elastic Net Regression	0.1	0.01	0.843006	0.841683	19.910000	19.774300	6.382421

Display overall score of the models

9.1 Output




Predict Co2 Emission
Home
About
Contact us
Co2 Emission Predictor
Search
Engine Size
2
Cylinders
4
Fuel Consumption City (L/100 km)
9
Fuel Consumption Hwy (L/100 km)
8
Fuel Consumption Comb (L/100 km)
6
Fuel Consumption Comb Mpg (L/100 Km)
89
Fuel Types
Type E
Transmission type
A4
Make
Luxury
Vechicle Class
Vechicle Class
SUV


Predict Co2 Emission
Home
About
Contact us
Search
Co2 Emission Predictor
The Vehicle CO2 Emission is 203 g/km.

9.2 Development Life Cycle

- **Data Collection and Pre-processing:** Gather relevant data sources, which may include historical CO_2 emissions data, economic indicators, demographic data, and environmental factors. Pre-process the data to handle missing values, outliers, and inconsistencies. Data pre-processing may also involve normalization, transformation, and feature engineering to prepare the data for modelling.

- **Feature Selection and Engineering**: Identify the most relevant features (variables) for predicting CO_2 emissions. This may involve domain knowledge, statistical analysis, and machine learning techniques to select informative features and create new features that capture important patterns in the data.

- **Model Selection and Training**: Choose appropriate machine learning models for CO_2 emission prediction, considering factors such as interpretability, accuracy, and scalability. Commonly used models include regression models, time series analysis, and machine learning algorithms like random forests, support vector machines, or neural networks. Train the selected models using historical data, and evaluate their performance using appropriate metrics and validation techniques.

- **Model Evaluation and Validation:** Assess the performance of the trained models on unseen data to ensure generalization and reliability. Use techniques such as cross-validation, holdout validation, or time series split validation to evaluate model performance. Compare the performance of different models and select the best-performing one(s) for deployment.

- **Deployment and Integration:** Deploy the CO_2 emission prediction system in a production environment, making it accessible to end-users through an application programming interface (API), web interface, or integration with existing systems. Ensure that the system meets performance, scalability, and security requirements.

- **Monitoring and Maintenance**: Continuously monitor the performance of the deployed system and update it as needed to accommodate changes in data patterns, regulations, or user requirements. Perform regular maintenance tasks such as retraining models with new data, updating software dependencies, and addressing any issues or bugs that arise.

- **Feedback and Iteration**: Gather feedback from users and stakeholders to identify areas for improvement and iteratively enhance the CO_2 emission prediction system. This may involve refining models, adding new features, or incorporating additional data sources to improve prediction accuracy and usability.

Chapter 4
10: IMPLEMENTATION & RESULTS

In densely populated cities, especially along toll road corridors, vehicle traffic and associated pollutants are mostly to blame for the harmful levels of air pollution, including CO2. While some models are intended to illustrate the spatial prediction of these emissions, others are used to assess the effects of transportation-related CO2 emissions on both people and the environment. For the experiment, we used the MATLAB 2020 software. The machine's physical configuration included an Intel i7 Quad-Core CPU running at 2.8 GHz and 8 GB of RAM. This sample size remains tiny when compared to earlier studies. On the other hand, it necessitates a successful plan for handling overfitting problems. In the first model, class of vehicle, and size of engine (L), The input variables used in the model are cylinders transmission, fuel type, fuel consumption city (L/100 km), fuel consumption highway (L/100 km), fuel consumption comb (L/100 km), fuel consumption comb (mpg), and co2 emissions (g/km). To normalize the data, the max-min method was applied. The built-in model's ability to generate precise predictions was assessed using four statistical measures: determination coefficient (R2), R%, mean squared error (MSE), and root mean square error (RMSE). Deep learning models are capable of simulating complex non-linear problems, such CO2 emissions from vehicles. Recent improvements in processing power have enabled this capability. The training process of LSTM and biLSTM methods is shown in Figures 12 and 13.

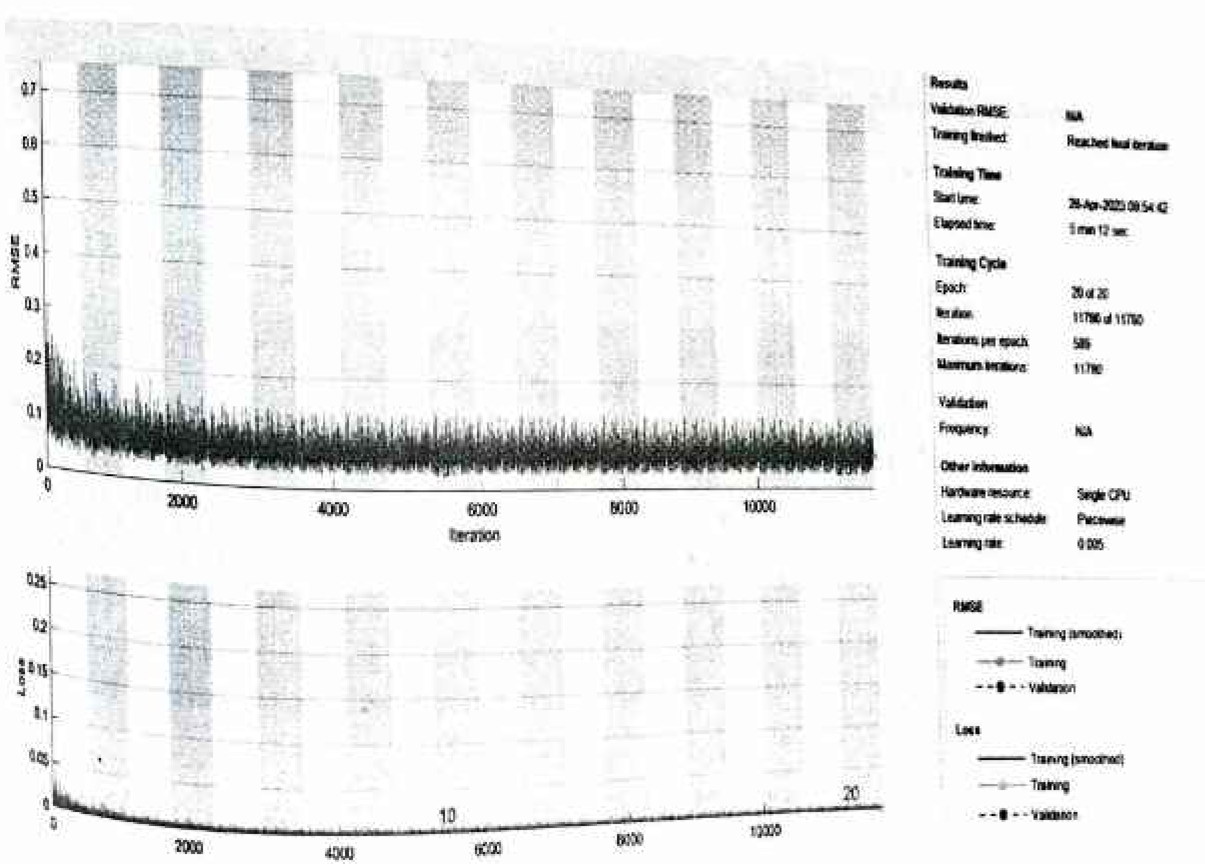


Figure 12. Training process of the long short-term memory network model.

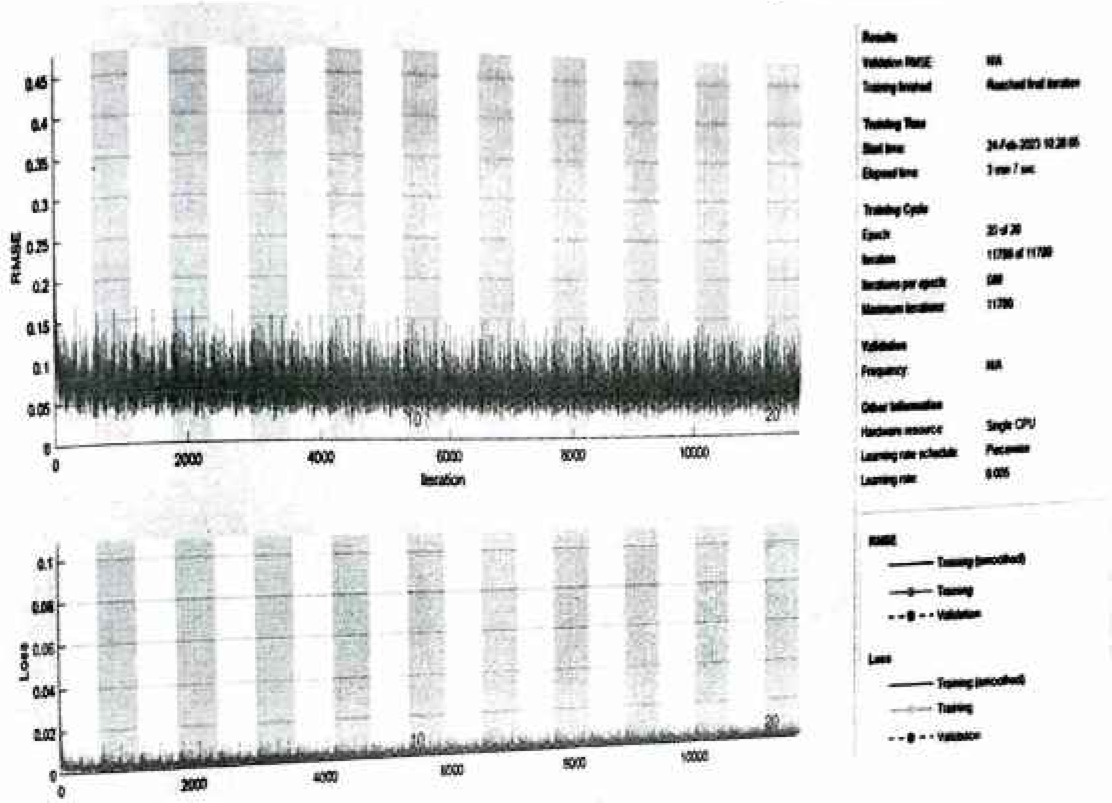


Figure 13. Training process of the bidirectional long short-term memory network model.

10.1 Training Results

Using specific experimental data to create a highly successful model involves a crucial step called training. Throughout this phase of the procedure, almost 80% of the datasets were used for this purpose. The performance of the designed LSTM and BiLSTM models is shown in Figure 14, and the evaluation metrics' values are included in Table 3. The anticipated CO2% (Y-axis) values and the experimental values (X-axis) are entirely consistent between datasets, as Table 4 and Figure 14 demonstrate. If the built LSTM and BiLSTM models exhibit low MSE and RMSE values, high R% (97.07%) and R2 (93.78) values, respectively, they are suitable for testing. These numbers show that the system is capable of achieving the desired objectives [18].

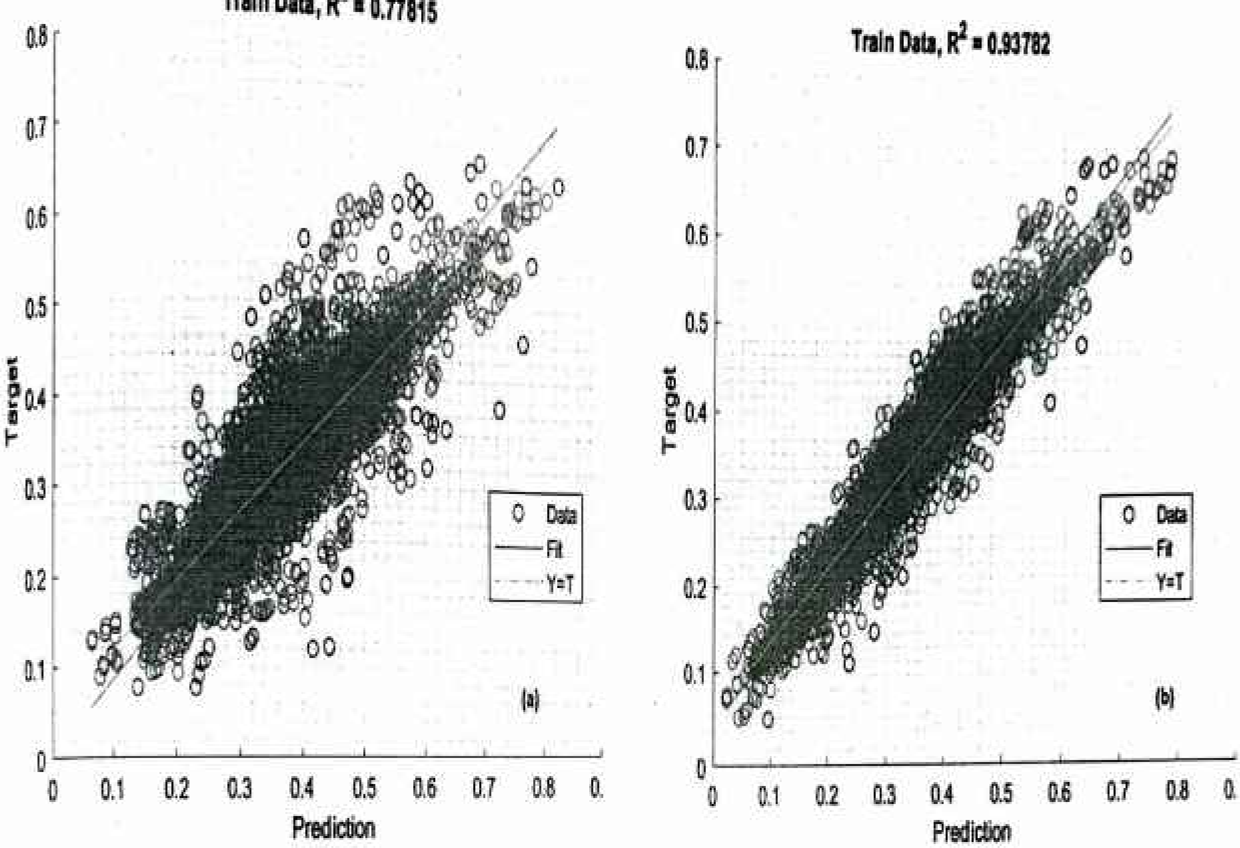


Figure 14. Regression plot of training carbon dioxide (a) LSTM (b) BiLSTM models.

Table 4. Results of LSTM and BiLSTM models in predicting carbon dioxide emissions at the training phase.

Models	#MSE	#RMSE	#R (%)	R^2 (%)
LSTM model	0.004980	0.07057	90.47	77.81
BiLSTM model	0.001177	0.0343	97.07	93.78

10.2 Testing Results

During testing, the LSTM and BiLSTM models were validated using unseen data from 20% of the datasets. The testing results for the LSTM and BiLSTM models are displayed in Figure 12 and Table 5, respectively. A robust connection is observed in Figure 15 between the experimental and projected values. Furthermore, the R% and R2 values were exceptionally high at 97.07% and 0.93.78, respectively, while the MSE and RMSE values were abnormally low at 0.0012 and 0.0035, respectively. The government's ability to prevent pollution from increasing can be aided by the accuracy with which the BiLSTM model predicts CO_2 emissions. According to our findings, the BiLSTM model predicts with the highest level of accuracy [17].

Table 5. Results of LSTM and BiLSTM models in predicting carbon dioxide emissions at the test- ing phase.

Models	#MSE	#RMSE	#R (%)	#R^2 (%)
LSTM model	0.005075	0.07125	90.14	75.73
BiLSTM Model	0.0012678	0.03560	96.95	93.55

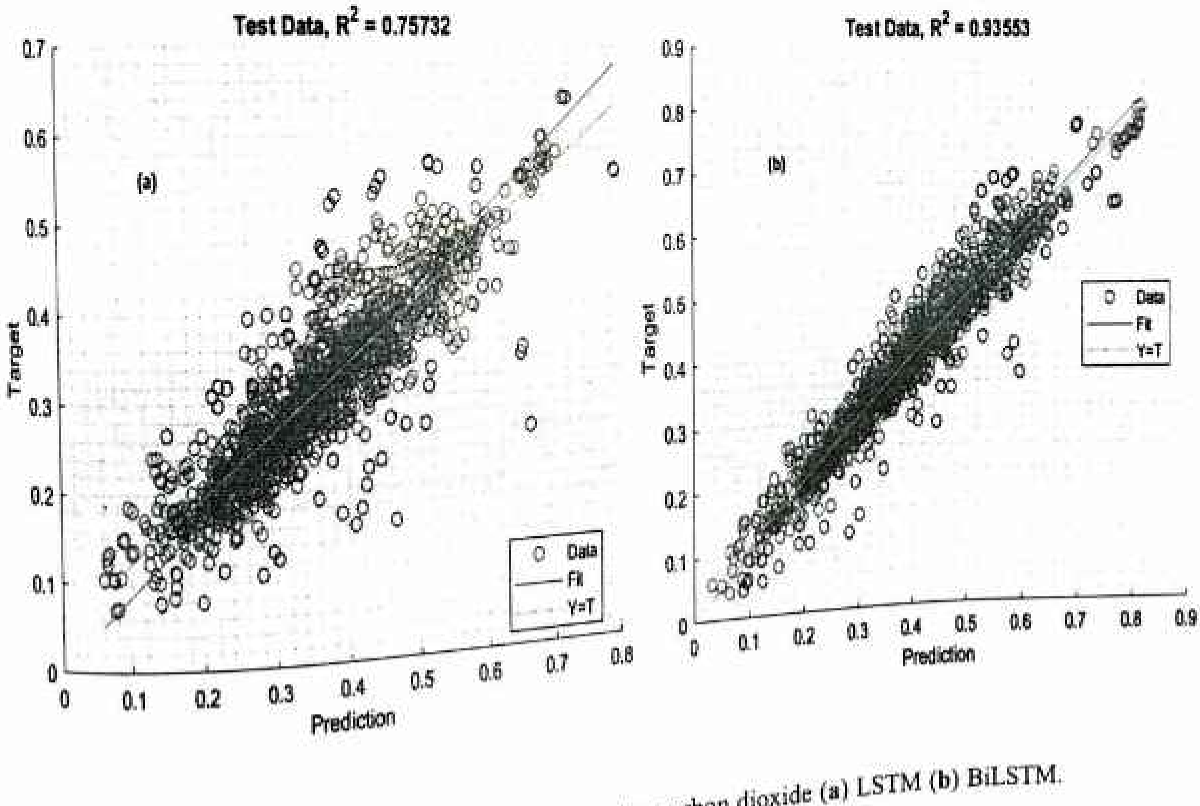


Figure 15. Regression plot of testing carbon dioxide (a) LSTM (b) BiLSTM.

Conclusion

The entire planet is being impacted by the effects of climate change, which include extreme weather events like storms and droughts. The main substances causing climate change are greenhouse gases, of which carbon dioxide (CO_2) makes up a greater percentage of the overall amount released into the atmosphere. Since simulation and data mining methods rely on previous data, it is expected that the concentration of CO_2 will continue to rise. Eighty percent of global carbon dioxide emissions are attributed to the burning of fossil fuels, primarily by the manufacturing and transportation sectors. Governments in both rich and developing nations have passed laws to regulate CO_2 emissions in an effort to transfer accountability for doing so from consumers to manufacturers or vice versa. Dripping data from the car file was used to evaluate a number of methodologies, and with an accuracy of 0.90, it was found to be the most effective approach for determining the quantity of carbon monoxide emission in ADAboost.

The conclusion about CO_2 emissions is that they are a major cause of global warming and climate change. Human activities, like burning fossil fuels and deforestation, release excessive CO_2 into the atmosphere, trapping heat and leading to temperature rise. To combat this, reducing CO_2 emissions is crucial. Solutions involve transitioning to renewable energy sources, enhancing energy efficiency, and preserving forests to absorb CO_2. Acting to curb CO_2 emissions is essential to safeguard our planet and future generations.

As we have seen the feature importance: Cause Analysis,

• The Energy use per kg oil per capita

• Renewable energy usage (% Total)

• Urban Population (% total population)

• Population (total) • Electricity access (% total population)

These are the features got most importance and therefore have more impact on the Target: „ CO2 Emission".

In the above given features as we see that the energy use per kg oil percapita and renewable energy usage are the most dependent features on carbon dioxide emissions. But if we see the energy use per kg oil is the negative factor so we should stop using it and renewable energy use is the positive factor so we have to increase its usage to reduce the carbon dioxide emissions in future. So, we should increase the usage of renewable energy and decrease usage of energy per kg oil.

As we see after the feature extraction the above-mentioned factors are more dependent on carbon dioxide emissions so we have to take necessary actions to reduce the negatively impacted factors so as to reduce the carbon dioxide emissions.

The purpose of this study is to forecast Canada's CO2 emissions connected to transportation. Five machine learning algorithms—Artificial Neural Network (ANN), Support Vector Machine (SVM), Deep Neural Network (DNN), Decision Tree Regressor, and Gradient Boost Regressor—were used to successfully generate the prediction. (GBR)] are applied. The model was trained on 80% of the dataset that had years older than 2013, and it was tested on the remaining 20% of the dataset that contained years older than 2013. Six statistical criteria on the success of the algorithms in the forecast are explained in order to assess the model's performances.
This research is focused on predicting transportation-related CO2 emissions in Canada. To achieve accurate predictions, five machine learning algorithms (Artificial Neural Network (ANN), support vector machine (SVM), Deep Neural Network (DNN), Decision Tree Regressor, and Gradient Boost Regressor (GBR)) were utilized. The model was trained using 80% of the dataset with years up to 2013, while the remaining 20% was used for testing. Six statistical metrics were employed to evaluate the performance of the algorithms in forecasting CO2 emissions.

The results indicate that Artificial Neural Network (ANN), support vector machine (SVM), and Deep Neural Network (DNN) had the lowest performance, while Gradient Boost Regressor (GBR), Decision Tree Regressor, and Boosting Regressor showed the best forecasting results. Most algorithms in the study demonstrated high prediction accuracy in forecasting transportation-related CO2 emissions. The findings suggest that current policies and techniques, such as alternative fuel types, are effective in reducing CO2 emissions from the transportation sector.
The study's success implies that Canada is not likely to face significant threats from transportation-based CO2 emissions in the future. It is recommended to increase the use of Vehicle Alternative Fuel Types and introduce low carbon fuels to further reduce fossil fuel consumption in transportation.
Based on the analysis of the combustion process and the proximate analysis data, a method for predicting the fuel characteristic factor to estimate CO2 emissions from coal-fired power plants has been developed. This method serves as a guide for implementing online monitoring. Through the calculation of CO2 emissions from the Wangling thermal power plant unit, and subsequent comparison with results obtained using the 2006 IPCC Guidelines and the coal consumption rate method for power supply, the following conclusions can be drawn:

The results from both the IPCC method and the coal consumption rate method differ significantly from the actual results. The disparity in the former is attributed to variations in coal quality between China and other regions, while the latter is due to unburned carbon being released as ash, fly ash, and other forms. However, by utilizing Matlab software for multiple linear regression to establish the correlation between proximate analysis data and the fuel characteristic coefficient, this study has achieved results that align more closely with reality, making the calculations more accurate.

Future Work

Enhanced Modeling and Decision Support:

As the urgency to address climate change intensifies, the role of CO_2 emission prediction becomes increasingly pivotal in informing policy decisions, guiding mitigation strategies, and fostering sustainable development. Looking ahead, several avenues of future work can enhance the accuracy, applicability, and impact of CO_2 emission prediction models. This comprehensive approach encompasses data refinement, model development, uncertainty quantification, decision support integration, stakeholder engagement, and long-term research directions.

Data Refinement and Expansion:

Future work will focus on refining and expanding the scope of data used for CO_2 emission prediction. This involves enhancing the quality of existing data sources, such as emission inventories, satellite imagery, socioeconomic indicators, and climate variables. Rigorous data cleaning, validation, and preprocessing techniques will be employed to ensure data accuracy and consistency. Additionally, efforts will be made to integrate additional data sources and variables that may improve prediction accuracy and capture new insights into CO_2 emission dynamics. Real-time sensor data, remote sensing technologies, and emerging data streams will be explored to enrich predictive capabilities.

Model Development and Evaluation:

Advanced modeling techniques will be investigated to develop more sophisticated predictive models capable of capturing complex relationships and nonlinear dynamics inherent in CO_2 emissions. Ensemble modeling approaches, such as model averaging and stacking, will be employed to leverage the complementary strengths of multiple predictive models. Model evaluation metrics will be enhanced to comprehensively assess prediction performance, including mean absolute error, root mean squared error, and calibration plots. Rigorous validation exercises using cross-validation and sensitivity analysis will validate model robustness and generalization ability.

Uncertainty Quantification and Sensitivity Analysis:

Future work will focus on quantifying and communicating uncertainties associated with CO_2 emission predictions. Probabilistic modeling techniques, Monte Carlo simulations, and uncertainty propagation methods will be implemented to assess prediction reliability and provide decision-makers with probabilistic forecasts. Sensitivity analysis will be conducted to identify influential factors and assess the impact of input variables on emission predictions, guiding efforts to improve prediction accuracy and prioritize data collection and model refinement efforts.

References –

1. Bai, Femilda Josephin Joseph Shobana. "A Machine Learning Approach for Carbon di oxide and Other Emissions Characteristics Prediction in a Low Carbon Biofuel-Hydrogen Dual Fuel Engine." *Fuel* 341 (2023): 127578.
2. Peters, Glenn P., et al. "Carbon dioxide emissions continue to grow amidst slowly emerging climate policies." *Nature Climate Change* 10.1 (2020): 3-6.
3. Al-Nefaie, Abdullah H., and Theyazn HH Aldhyani. "Predicting CO2 emissions from traffic vehicles for sustainable and smart environment using a deep learning model." *Sustainability* 15.9 (2023): 7615.
4. Salem, Israa Ezzat, et al. "Introduction to The Data Mining Techniques in Cybersecurity." *Mesopotamian journal of cybersecurity* 2022 (2022): 28-37.
5. Azeez, Omer Saud, et al. "Modeling of CO emissions from traffic vehicles using artificial neural networks." *Applied Sciences* 9.2 (2019): 313.
6. Zhang, Rongshuo, et al. "A deep learning micro-scale model to estimate the CO2 emissions from light-duty diesel trucks based on real-world driving." *Atmosphere* 13.9 (2022): 1466.
7. Chadha, Akalbir Singh, et al. "Predicting CO2 Emissions by Vehicles Using Machine Learning." *International Conference on Data Management, Analytics & Innovation*. Singapore: Springer Nature Singapore, 2022.
8. Shah, Samveg, et al. "A Comparative Study of Machine Learning and Deep Learning Techniques for Prediction of Co2 Emission in Cars." *arXiv preprint arXiv:2211.08268* (2022).
9. Debone, Daniela, Tiago Dias Martins, and Simone Georges El Khouri Miraglia. "Modeling Carbon Release of Brazilian Highest Economic Pole and Major Urban Emitter: Comparing Classical Methods and Artificial Neural Networks." *Climate* 10.1 (2022):
10. Mitra, Sushmita, Sankar K. Pal, and Pabitra Mitra. "Data mining in soft computing framework: a survey." *IEEE transactions on neural networks* 13.1 (2002): 3-14.
11. Fayyad, Usama, Gregory Piatetsky-Shapiro, and Padhraic Smyth. "The KDD process for extracting useful knowledge from volumes of data." *Communications of the ACM* 39.11 (1996): 27-34.
12. Jun Lee, Sang, and Keng Siau. "A review of data mining techniques." *Industrial Management & Data Systems* 101.1 (2001): 41-46.
13. Khalid, B., Rachid, C., & Mounir, B. (2019). Data Mining: Analysis and Comparative Study of Supervised Techniques. *International Journal of Computer Applications*, *975*, 8887.
14. Bajpani, R. P., and Roopesh Kumar Dwivedi. "Data Mining Techniques For Dynamically Classifying And Analyzing Library Database." (2007).

Annexure I

Responsibility Chart

Roll No.	Name	Responsibilities
2201560024	Suraj Saurav	Fronted and Documentation
2201560039	Nikhil Kumar	Deployment and Research
2201560041	Abhishek Pathak	ML model and backend

Progress Reports duly signed by Mentors.

Annexure II:

K.R. MANGALAM UNIVERSITY
THE COMPLETE WORLD OF EDUCATION

School of Engineering & Technology
Final Year Project Progress Report

Project Title	**Co2 Emission Prediction**
Project Group Members	1.SURAJ SAURAV-2201560024 2.NIKHIL KUMAR-2201560039 3.ABHISHEK PATHAK-2201560043
Project Faculty Mentor	MS. RUCHIKA BAKHAR
Date of Meeting	28th March 2024
Tasks Assigned	Co2 Emission Prediction Model.
Tasks Completed	Co2 Emission Prediction Model.
Remarks by Project Mentor	Satisfactory

Overall Rating on Project Progress (Scale of 10)

No progress (0)	Average Progress (1-5)	Efforts are seen (6-8)	Satisfied with Progress (9-10)
		✓	

Signature of Faculty project mentor (with date): Ruchika 28/03/24

Signature (Coordinator/HoD/Dean)