A black rectangle with a black background

Description automatically generated with low confidenceLogo, company name

Description automatically generated

**Data Dynamos**

Rani Gadiraju(Product Owner)

Kiran Kumar Reddy Yerrabathini(Scrum Master)

Akash Sathyamurthy(Developer)

Sreeja Badri(Developer)

Kinnera Veerapaneni(Developer)

Fall 2023

**GHG Emissions DASHBOARD AND PREDICTIVE ANALYTICS FOR GMU CAMPUS**

**About the Cover**

The cover photo is an aerial view of Pearmund Cellars Winery. Located in Broad Run, Virginia it has been one of Virginias top wineries for over 20 years and has gathered over 200 gold medals in international competitions and national acclaim for its production of 100% Virginia grown wine. The winery's vineyard, Meriwether, is the oldest Chardonnay vineyard in Virginia, producing wine from its 30 acres for over 40 years.

Pearmund Cellars produces a variety of wines from Virginia grown fruit such as Viognier, Petit Manseng, Cabernet Franc, Petit Verdot and more. Their signature Ameritage, a blend of 5 different Bordeaux varietals, has won Best in Class at the Tasters Guild International.

Chris Pearmund, owner and founder has consulted on over 20 winery openings in the state; most recently Effingham Manor located in Nokesville, Virginia. A sister winery to Pearmund Cellars, it marries the history of Virginia and the history of Virginia winemaking. Using varietals found in Virginia such as Traminette, Viognier, Petit Verdot and Tannat, Effingham Manor tells the story of Virginia’s wine history.

Together Effingham and Pearmund provides guests with education, history, quality and service unrivaled in the Virginia Wine business, and carry a reputation as the leaders in the Virginia Wine Industry.

Always an innovator, Pearmund Cellars uses geothermal heating and cooling in their winery to maintain a constant temperature for their wine production process. This helps them produce high-quality wines while also being environmentally friendly.

Pearmund Cellars is among the first Virginia wineries to incorporate statistical data analysis to aid in wine production. Partnering with a Fall 2022 DAEN capstone student project team resulted in a statistical data analysis tool which assisted in the creation of Pearmund’s “12 Pearls Chardonnay” — a new product that outsold other wines 2-to-1 during the summer.

Contents

Table of Contents

[Abstract 1](#_Toc135642685)

[Section 1: Problem Definition 3](#_Toc135642686)

[1.1 Background 3](#_Toc135642687)

[1.2 Problem Space 4](#_Toc135642688)

[1.3 Research 6](#_Toc135642689)

[1.4 Solution Space 7](#_Toc135642690)

[1.5 Project Objectives 7](#_Toc135642691)

[1.6 Primary User Stories 7](#_Toc135642692)

[1.7 Product Vision 8](#_Toc135642693)

[1.7.1 Scenario #1 8](#_Toc135642694)

[1.7.2 Scenario #2 8](#_Toc135642695)

[Section 2: Datasets 8](#_Toc135642696)

[2.1 Overview 8](#_Toc135642697)

[2.2 Field Descriptions 8](#_Toc135642698)

[2.3 Data Context 9](#_Toc135642699)

[2.4 Data Conditioning 10](#_Toc135642700)

[2.5 Data Quality Assessment 10](#_Toc135642701)

[2.6 Other Data Sources 10](#_Toc135642702)

[2.7 Storage Medium 10](#_Toc135642703)

[2.8 Storage Security 10](#_Toc135642704)

[2.9 Storage Costs 11](#_Toc135642705)

[Section 3: Algorithms & Analysis / ML Model Exploration & Selection 11](#_Toc135642706)

[3.1 Solution Approach 11](#_Toc135642707)

[3.1.1 Systems Architecture 11](#_Toc135642708)

[3.1.2 Systems Security 11](#_Toc135642709)

[3.1.3 Systems Data Flows 11](#_Toc135642710)

[3.1.4 Algorithms & Analysis 11](#_Toc135642711)

[3.2 Machine Learning 11](#_Toc135642712)

[3.2.1 Model Exploration 12](#_Toc135642713)

[3.2.2 Model Selection 12](#_Toc135642714)

[Section 4: Visualizations / ML Model Training, Evaluation, & Validation 12](#_Toc135642715)

[4.1 Overview 12](#_Toc135642716)

[4.2 Visualizations 12](#_Toc135642717)

[4.3 Machine Learning 12](#_Toc135642718)

[4.3.1 Model Training 12](#_Toc135642719)

[4.3.2 Model Evaluation 13](#_Toc135642720)

[4.3.3 Model Validation 13](#_Toc135642721)

[Section 5: Findings 13](#_Toc135642722)

[Section 6: Summary 13](#_Toc135642723)

[Section 7: Future Work 13](#_Toc135642724)

[Appendix A: Glossary 15](#_Toc135642725)

[Appendix B: GitHub Repository 16](#_Toc135642726)

[Overview 16](#_Toc135642727)

[GitHub Repository Link 16](#_Toc135642728)

[GitHub Repository Contents 16](#_Toc135642729)

[Appendix C: Risks 17](#_Toc135642730)

[Sprint 1 Risks 17](#_Toc135642731)

[Sprint 2 Risks 17](#_Toc135642732)

[Sprint 3 Risks 18](#_Toc135642733)

[Sprint 4 Risks 18](#_Toc135642734)

[Sprint 5 Risks 19](#_Toc135642735)

[Appendix D: Agile Development 21](#_Toc135642736)

[Scrum Methodology 21](#_Toc135642737)

[Sprint 1 Analysis 21](#_Toc135642738)

[Sprint 2 Analyis 22](#_Toc135642739)

[Sprint 3 Analysis 22](#_Toc135642740)

[Sprint 4 Analysis 22](#_Toc135642741)

[Sprint 5 Analysis 23](#_Toc135642742)

[Works Cited 25](#_Toc135642743)

Table of Figures

[Figure 1: ChatGPT is the latest Natural Language Processing tool for Data Analytics. 5](#_Toc135643858)

[Figure 2: Data Analytics Engineering has the power to be transformative for an organization. 6](#_Toc135643859)

[Table 1: Glossary Table 15](#_Toc135643860)

[Table 2: Sprint 1 Risks 17](#_Toc135643861)

[Table 3: Sprint 2 Risks 17](#_Toc135643862)

[Figure 3: Sprint project dates 21](#_Toc135643866)

**This page intentionally left blank**

Abstract

Abstract

By absorbing solar energy, greenhouse gases (GHGs) warm the atmosphere and cause global warming. Carbon dioxide (CO2), nitrous oxide (N2O), and methane (CH4) are the most prevalent GHGs. Unfortunately, one of the factors cited for the rise in these gases in the environment is agriculture. This study, born out of the urgent concern regarding GHG emissions and their environmental impact, delves into various facets of these emissions, exploring their sources, quantification methods, and potential mitigation strategies. Given that almost all our environmental effects can be linked to emissions generation, addressing this climate change is Mason's top environmental concern. To operationalize its commitment, Mason implemented various initiatives outlined in its 2010 Climate Action Plan. Notably, the university engaged in purchasing wind-powered Renewable Energy Certificates (RECs) to offset a percentage of its annual electricity consumption. Over the years, Mason progressively increased its REC purchases, offsetting 5% of electricity use in 2011, 10% in 2012, 2013, and 2014, and 15% in 2015. These purchases not only contributed to the growth of the U.S. renewable energy sector but also effectively offset millions of pounds of annual carbon dioxide emissions. Recognizing these substantial efforts, Mason earned recognition as a Green Power Partner by the U.S. Environmental Protection Agency. In tandem with these initiatives, Mason developed its inaugural Climate Action Plan (CAP) as a strategic roadmap towards achieving climate neutrality. The university has diligently conducted several Greenhouse Gas (GHG) Emissions Inventories to measure and analyze its emissions, showcasing significant progress since the initial climate commitment in 2007. Mason consistently reports its sustainability advancements through a comprehensive Sustainability Tracking, Assessment, and Rating System (STARS) report.

)

**This page intentionally left blank**

Report

# Problem Definition

## Background

The global movement towards sustainable practices has acquired significant momentum in recent years. In conjunction with these efforts, a significant event has occurred in which 880 universities, including GMU, have declared a firm commitment to achieve carbon neutrality by 2050. This agreement represents a significant step forward in tackling climate change and lowering higher education institutions' carbon footprint. A full understanding of carbon emissions, their sources, and their environmental implications is required to achieve carbon neutrality. Traditional carbon emission estimates have been widely utilized, but they have constraints that prevent them from fully capturing the full breadth of emissions.

Traditional carbon emission estimates, normally completed annually, have served as the foundation for calculating an organization's carbon footprint. Methods such as Simap have been used to estimate emissions from activities such as on-site combustion of fossil fuels, manufacturing processes, and transportation fleets. Simap gives the university a real-time perspective of emissions, allowing it to spot irregular effects and peak emissions that are otherwise hidden by annual averages. Second Nature, which informs GMU's decision-making processes, provides a comprehensive viewpoint on the complex relationship between emissions and sustainability. Simap and Second Nature were adopted by GMU because of their dedication to gaining a deeper understanding of its carbon emissions.

However, the capacity of these annual estimates to account for transient impacts and fluctuations in emissions over time is limited. They do not catch short-term spikes or swings, sometimes known as "peaks" or "tails," which can account for many emissions.

**Limitations of Traditional Emission Estimates:**

The inability of standard annual emission estimates to account for transitory effects is a significant constraint. Seasonal shifts, economic swings, and unforeseen catastrophes can all cause changes in emission patterns. Inadequate accounting for such fluctuations might lead to an inadequate knowledge of emissions and impede effective mitigation methods. Furthermore, the emphasis on annual estimates frequently overlooks the impact of short-lived emission spikes, which may account for a significant fraction of overall emissions.

To address emissions comprehensively, organizations often categorize them into three scopes:

**Scope 1**: Direct Emissions: This encompasses emissions from sources directly owned or controlled by the organization. It includes emissions from on-site fossil fuel combustion, manufacturing processes, and transportation fleets.

**Scope 2**: Indirect Emissions: These result from the consumption of purchased electricity, heat, or steam. They stem from off-site sources like electric power plants.

**Scope 3**: Other Indirect Emissions: This category covers all other emissions indirectly linked to the organization's activities, but not under its direct control. It involves activities across the value chain, such as purchasing goods and services, employee commuting, business travel, and waste disposal.

**Importance of Addressing Peaks and Tails:**

By focusing exclusively on average annual emissions, chances for effective mitigation may be ignored. Peaks and tails, despite their rarity, can contribute significantly to total emissions output. Identifying and focusing on these situations for reduction efforts can have a significant impact on achieving carbon neutrality. Universities and organizations can make significant progress toward their emission reduction targets by addressing these infrequent but high-impact events.

**Impact of Power Generation Sources:**

The distribution of GMU's electrification programs directly affects their environmental impact. The university's commitment to decreasing its carbon footprint and the effectiveness of electrifying Scope 1 emissions are both supported by its move to renewable energy sources.

**Problem Space**

The problem domain being discussed revolves around the evaluation and reduction of greenhouse gas (GHG) emissions at George Mason University (GMU) and other similar institutions, with the overarching objective of achieving carbon neutrality by 2050. Here is an analysis of the key challenges and objectives within this undertaking:

* **Precise Measurement of GHG Emissions:** Traditional methods, like Simmap and Second Nature, provide yearly approximations of emissions but might overlook temporary fluctuations or sudden emission spikes. Implementing a system for weekly emissions tracking could help identify such variations and offer a more accurate depiction of the university's carbon impact.
* **Incomplete Data Compilation:** Gathering precise data from diverse sources across the campus poses difficulties. Distinct departments, buildings, and facilities may lack consistent data collection practices or standardized reporting methods.
* **Scope 1 and Scope 2 Emissions:** Scope 1 emissions pertain to direct emissions from owned sources, while Scope 2 emissions include indirect emissions from purchased energy. Shifting Scope 1 emissions to Scope 2 could decrease immediate on-campus emissions, but the accountability for emissions would then be placed on external sources, potentially impacting genuine emission reduction.
* **Analysis of Power Generation Origins:** The origin of electricity significantly influences GHG emissions. Altering the power generation mix, such as incorporating more renewable sources, can lead to fluctuations in the carbon intensity of energy consumption.
* **Strategies for Mitigation**: Identifying effective methods for emission reduction is crucial. Electrifying Scope 1 emissions might cut direct emissions but could lead to increased indirect emissions if the electricity originates from carbon-intensive sources.
* **Timeline for Carbon Neutrality**: Meeting the carbon neutrality goal by 2050 necessitates meticulous planning and execution of emission reduction strategies. Swift adjustments might be required to stay on course with the set timeline.
* **Financial and Technical Feasibility**: Enforcing emission reduction strategies could demand investments in technology, infrastructure, and staff training. Assessing the financial and technical viability of these measures is pivotal.
* **Engagement of Stakeholders**: Achieving carbon neutrality involves various stakeholders, including university leadership, faculty, students, and local communities. Ensuring unanimous alignment and support for the objectives can be challenging.
* **Progress Tracking and Reporting**: Developing a mechanism for continuous emission monitoring, progress assessment, and transparent reporting is essential to demonstrate the university's commitment to carbon neutrality.
* **Adapting to Changing Regulations**: The regulatory framework regarding emissions could evolve over time. Remaining updated and adaptable to changing regulations is vital for sustained environmental responsibility.
* **Cultivating Behavioral Changes**: Attaining carbon neutrality demands alterations not only in technology but also in behaviors and routines among students, faculty, and staff. Encouraging eco-friendly practices within the campus community is a multifaceted endeavor.

## Research

Understanding, monitoring, and controlling the release of these gases into the atmosphere are all part of the large and ongoing subject of research on greenhouse gas emissions. Greenhouse gases (GHGs) are gases that cause the greenhouse effect, which in turn causes global warming and climate change. They do this by trapping heat in the Earth's atmosphere. Carbon dioxide (CO2), methane (CH4), nitrous oxide (N2O), and fluorinated gases are the main greenhouse gases.

The following are some important topics of study in relation to greenhouse gas emissions:

* **Monitoring and Measuring:**

The measurement of atmospheric amounts of greenhouse gases is the topic of research. This calls for the use of mobile sensors, satellite observations, and ground-based monitoring stations. To better understand how greenhouse gases are released and absorbed by various natural and human activities, scientists also investigate the sources and sinks of greenhouse gases. The main sources of greenhouse gas emissions, including energy production, transportation, industrial operations, agriculture, and deforestation, are the subject of research. Studies estimate the various industries and activities contributions to total emissions, which aids in efficiently focusing mitigation efforts.

* **Strategies for Mitigation:**

The goal of this research is to create and find methods for reducing greenhouse gas emissions. Some of the mitigation strategies under consideration include the use of renewable energy sources, increased energy efficiency, carbon capture and storage (CCS), reforestation, and adjustments to agricultural practices.

* **Impact of Greenhouse Gas Emissions:**

Scientists evaluate how higher greenhouse gas concentrations may affect ecosystems, extreme weather events, sea level rise, and climate trends. This study contributes to the public's understanding of the urgency of controlling emissions to reduce potential effects.

* **Policy and Regulations:** Research in this field assesses how well current policies and regulations for lowering greenhouse gas emissions are working. In order to encourage emission reductions, it also examines potential new legal frameworks, international agreements, such as the Paris Agreement, and financial incentives.
* **Technological Innovation:** Studies examine cutting-edge renewable energy technology, carbon collection and utilization techniques, and environmentally friendly transportation options as potential emissions-reducing technologies.
* **Socioeconomic Implications:** Researchers look at how climate change and greenhouse gas emissions impact vulnerable populations, economies, and local communities. This study sheds light on the advantages of reducing emissions as well as the social and economic costs of inaction.
* **Climate modeling:** Using simulations, climate models attempt to capture the intricate relationships between the many variables that influence climate change, such as greenhouse gas emissions, ocean currents, air circulation, and more.

These models are used by researchers to estimate possible future climate scenarios depending on various emission scenarios.

In general, research on greenhouse gas emissions is essential for leading mitigation efforts, informing legislation, and increasing awareness of the urgency of addressing climate change. To accomplish the transition to a future that is more sustainable and low carbon, it requires cooperation between scientists, governments, industries, and the public.

**SCOPE 1:**

**N-P-K (Nitrogen, Phosphorus, Potassium) in Fertilizers and Emission Factors:**

In modern agriculture, fertilizers containing the elements nitrogen (N), phosphorus (P), and potassium (K), also known as N-P-K fertilizers, are essential. Enhancing crop yields and ensuring food security depend on these nutrients. N-P-K fertilizer use and application in agriculture, however, can primarily be blamed for greenhouse gas emissions because of the nitrogen component.

* Nitrogen (N): Nitrogen is a vital fertilizer ingredient that is necessary for crop nutrition. The emission of nitrous oxide (N2O), a strong greenhouse gas, occurs because of nitrogen in fertilizers at various application phases and afterward through soil processes. The amount of N2O released per applied unit of nitrogen is measured by the emission factor for nitrogen in fertilizers.
* Phosphorus(P): Phosphorus (P) does not immediately release greenhouse gasses when applied. But when phosphorus runoff from agricultural areas reaches aquatic bodies, it raises worries since it can cause environmental problems. Due to the phosphorus content of fertilizers, this may cause emissions of methane (CH4) and carbon dioxide (CO2) in aquatic habitats.
* Potassium (K): The application of potassium-containing fertilizers does not directly contribute to the release of greenhouse gasses. However, improving its use may indirectly affect emissions. By reducing the requirement for excessive nitrogen fertilization, balanced potassium treatment can reduce N2O emissions.

Understanding the relationship between N-P-K fertilizers and emission factors during application is essential for emissions reduction strategies in agriculture. It highlights the need for sustainable agricultural practices that optimize nutrient use efficiency while minimizing greenhouse gas emissions. This understanding informs the development of strategies aimed at reducing the environmental impact of fertilizers, contributing to both food security and emissions reduction goals.

Discussions about measuring carbon dioxide (CO2) emissions from various vehicle types utilized within an organization can be found in the areas of transportation and public safety.

* Internal Combustion Engine (ICE) Cars (Greg Farley): An ICE automobile fleet is the subject of this analysis. To estimate CO2 emissions, it considers variables like the number of cars, the typical weekly mileage, and fuel economy.
* Facilities/Maintenance Vehicles (Greg Farley): This section of the study pertains to a fleet of ICE automobiles, considering their weekly mileage and fuel consumption, in a manner similar to that of passenger vehicles.
* Shuttles (Josh Cantor): In this section, the focus is now on shuttle buses, which are still viewed as ICE vehicles. The number of buses and their weekly mileage have a role in estimating CO2 emissions.
* List of Routes: The context mentions the need to gather data on the distance for each route, assuming one bus per route. Additionally, the frequency of routes is considered.

**SCOPE 2:**

Electricity Usage for Lighting: The first part of the analysis focuses on estimating weekly electricity usage for lighting in each building. This calculation considers factors such as the building's square footage, the assumed number of lights per square foot, daily light usage, and the wattage of the lights. These calculations help determine the weekly electricity consumption for lighting in each building.

* CO2 Emissions from Electricity: To convert electricity usage into CO2 emissions, a weighted average of CO2 emissions per kilowatt-hour (KWHrs) is considered, factoring in the percentage of power generated from different sources (e.g., coal, LNG, oil). This calculation allows us to estimate the weekly GHG emissions associated with the electricity consumption in each building.
* Electricity Usage for Appliances: Like lighting, the analysis extends to estimating electricity usage for appliances in each building. Factors include square footage, assumed appliances per square foot, daily appliance usage, and appliance wattage. These calculations help determine the weekly electricity consumption for appliances.
* CO2 Emissions from Appliances: Just as with lighting, electricity usage for appliances is converted into CO2 emissions using the weighted average CO2 emissions per KWHrs, considering the mix of power generation sources.
* Heating and Cooling: Additionally, the context introduces considerations for temperature control in buildings. It involves estimating weekly electricity usage for heating or cooling, considering factors like temperature differences, and building insulation coefficients.

**SCOPE 3:**

* Transportation of Commuters (Parking Spots): The context briefly touches on transportation-related emissions from commuters. It mentions calculating weekly fuel burn and CO2 emissions for commuters based on the number of cars, miles traveled, and fuel burn rates.

## Solution Space

The top-level block shows the area for decreasing Scope 1's greenhouse gas emissions.

* **Scope 1 Categories:** production of fertilizer, transportation operations, and boilers and furnaces. There are several solution categories that organizations might investigate to cut emissions within each Scope 1 area.
* **Solutions for Fertilizer Production**: Included in this area are methods for using less fertilizer and practicing sustainable agriculture. Emission reduction may be aided by precision agriculture, which optimizes resource use.
* **Transportation Operations Solutions**: In this section, options including implementing clean energy measures, switching to alternate modes of transportation (such electric cars), and increasing fuel efficiency are offered.
* **Boilers and Furnaces Solutions**: Strategies for reducing emissions from heating systems include switching to cleaner fuels, improving combustion processes for increased effectiveness, and moving towards renewable energy sources.
* **Solution Subcategories**: Within each solution category, these subcategories indicate certain steps and strategies that organizations can follow.
* **Reduced Fertilizer Use:** Reducing the amount of fertilizer used in agricultural practices while preserving crop yields are referred to as "reduced fertilizer use" strategies.
* **Alternative Transportation Modes**: The use of alternate modes of transportation to offset the emissions produced by conventional fuel-powered vehicles, such as electric cars, public transportation, and cycling.
* **Renewable Energy Sources**: Using renewable energy sources, such as solar, wind, and hydropower, instead of fossil fuels for industrial and heating purposes.
* **Precision Agriculture:** It involves utilizing data and technology to maximize the use of resources in agriculture, such as water, fertilizer, and pesticides, while lowering emissions.
* **Electric Vehicles**: Converting to electric vehicles, which emit no exhaust emissions, for the transportation fleet.
* **Improved Combustion Processes:** Enhancing the performance of combustion processes in boilers and furnaces to increase efficiency and emissions.

A diagram of a diagram

Description automatically generated

* University can explore the installation of Combined Heat and Power (CHP) systems within Boilers & Furnaces Solutions. These systems make it possible to produce heat and electricity simultaneously from the same energy source, improving energy efficiency and lowering emissions. By capturing and reusing heat from industrial operations, waste heat recovery devices can also reduce the amount of energy required for heating and emissions.
* It can finance carbon offset initiatives, such reforestation and renewable energy projects, to make up for any residual emissions across all categories. In addition, assessing and lowering emissions across the supply chain, including when sourcing supplies and moving goods, can reduce emissions both directly and indirectly.
* It can adopt a more thorough strategy customized to their unique circumstances, industry, and location by incorporating these extra options into their emissions reduction initiatives. Such a comprehensive strategy enables well-informed choices and further development of a business that is more environmentally friendly and sustainable.

**Scope-2:**

**A diagram of a scientific experiment

Description automatically generated**

**Campus Buildings:**

Details of how much power is used in various campus buildings for lighting, heating, cooling, and running equipment.

Renewable Energy: Indicate the extent to which any structures that use renewable energy sources (such as solar panels on rooftops) contribute.

* **Heating and cooling systems:**

Identifying any energy efficiency measures in place, such as programmable thermostats or high-efficiency HVAC systems, as well as the primary energy source that heating and cooling systems primarily use, such as electricity, natural gas, or another fuel source. The use of renewable energy in heating and cooling systems is referred to as "renewable energy integration."

* **Laboratories:**

Describing in detail the electricity required by the energy-intensive equipment used in laboratories. Describe the exhaust and ventilation systems as they may use a lot of energy.

Research: If laboratories spend time on energy-related research, describe how this supports efforts to reduce emissions.

* **Data Coolers:**

Describe the size of the data center, including the number of servers and the energy requirements of those servers. Describe the cooling techniques used to keep server temperatures stable. Efficiency standards Talk about energy-efficient data center designs, including virtualization or hot/cold aisle confinement.

**Scope-3:**

A diagram of a company

Description automatically generated

* **Purchased Goods and Services Emissions:** Emissions linked to items bought from external suppliers. Includes emissions from production, transportation, and disposal of office supplies, equipment, textbooks, and contracted services.
* **Employee Commuting Emissions:** Emissions from faculty and staff commuting to/from the institution. Covers emissions from employee transportation methods, like cars, public transport, or biking.
* **Business Travel Emissions**: Emissions from employee work-related trips (conferences, meetings, etc.). Includes emissions from flights, hotels, rental cars, and related activities.
* **Waste Disposal Emissions:** Emissions tied to waste disposal generated by the institution. Includes emissions from waste incineration, landfilling, and recycling.
* **Student International Travel Emissions:** Emissions from students traveling to/from the institution for international programs. Includes emissions from flights, ground transport, and related travel.
* **Student Commuting Emissions:** Emissions from students commuting to campus for classes and activities. Like employee commuting, covering various student transportation modes.
* **Study Abroad Travel Emissions:** Emissions from student and faculty travel for study abroad programs. Includes emissions from international flights, accommodations, and study abroad travel.
* **Admin/Faculty Travel Emissions:** Emissions resulting from administrative staff and faculty work-related travel. Covers emissions from business trips, conferences, and professional travel.

A diagram of a company's strategy

Description automatically generated

**Mitigation Strategies:**

* **Purchased Goods and Services Emissions**: Choose eco-friendly suppliers. Reduce unnecessary purchases. Promote reuse and recycling.
* **Employee Commuting Emissions:** Encourage telecommuting. Promote carpooling and public transport. Support biking and walking.
* **Business Travel Emissions:** Favor virtual meetings. Offset essential travel emissions. Establish sustainable travel policies.
* **Waste Disposal Emissions:** Implement waste reduction and recycling. Explore waste-to-energy solutions. Pursue zero-waste initiatives. Student International
* **Travel Emissions:** Promote virtual international programs. Encourage carbon offsetting. Advocate for sustainable travel choices.
* **Student Commuting Emissions:** Offer campus shuttle services. Invest in biking and walking infrastructure. Promote carpooling.
* **Study Abroad Travel Emissions:** Encourage sustainable travel practices. Offset study abroad program emissions. Explore alternative travel modes.
* **Admin/Faculty Travel Emissions:** Establish sustainable travel guidelines. Consider virtual conference attendance. Implement carbon offset programs.

## Project Objectives

The primary objective related to greenhouse gases is to mitigate their effects to avoid the negative consequences associated with rising global temperatures:

* Develop a comprehensive model or tool that accurately assesses campus-wide greenhouse gas emissions, enabling us to track and manage our environmental impact more effectively.
* By analyzing weekly emissions data, we aim to identify transients and peaks in our carbon footprint, which will help us pinpoint specific timeframes or activities contributing significantly to emissions.
* An important aspect of our analysis involves understanding the potential impact of transitioning emissions from Scope 1 to Scope 2. This shift will allow us to evaluate how changes in energy sourcing influence our carbon profile.
* We are investigating shifts in the breakdown of power generation sources to gain insights into the changing dynamics of our energy consumption. This understanding will inform our sustainability strategies and resource allocation.
* Our goal is to identify specific opportunities for mitigation within our operations. Through careful analysis of our emissions data, we intend to pinpoint areas where interventions can have the greatest positive impact on reducing our carbon footprint.

## Primary User Stories

## President of George Mason University: In my role as President of George Mason University, I am resolute in our mission to transform our campus into a carbon-neutral environment by the year 2050. This ambitious commitment reflects our dedication to sustainability, innovation, and fostering a greener future for our students, community, and the world.

## Product Vision

The product vision is to develop a comprehensive tool for universities to achieve their pledge of becoming Carbon Zero by 2050. This tool will provide a detailed understanding of the university's greenhouse gas (GHG) emissions, including transient effects, "tails," and peak emissions. It will also help universities identify specific opportunities for mitigation and assess the impact of moving emissions from Scope 1 to Scope 2. Additionally, the tool will enable universities to monitor changes in the breakdown of power generation sources over time. By offering insights into weekly emissions patterns, this tool will empower universities to take targeted actions to reduce their carbon footprint and make informed decisions for a sustainable future.

### Scenario #1

The Carbon Zero University Tool is developed using a well-structured Agile methodology. It starts with defining key features and user stories, focusing on basic emissions data collection. Subsequent sprints build on this foundation, adding emissions analysis, Scope 1 and Scope 2 insights, and data monitoring capabilities. User feedback leads to continuous refinement, improved user interface, scalability, and documentation. The MVP is then tested by select universities, guiding future iterations with features like automated data collection, advanced analytics, and system integration. This iterative approach ensures a robust tool for universities to achieve their carbon neutrality goals.

### Scenario #2

In this iterative feature development scenario, the Agile approach drives continuous improvement of the Carbon Zero University Tool. It starts with core modules for data collection and emissions analysis, forming the base. Subsequent sprints add critical features like Scope 1 and 2 insights, data monitoring, weekly emissions analysis, and enhanced reporting. User experience is a priority with UI/UX enhancements. Automation streamlines data input, and reporting gets better visuals. Integration with common sustainability systems is sought for compatibility.

User feedback fuels iterative improvements, aligning the tool with evolving sustainability trends. This ensures the tool remains adaptable and responsive to universities striving for carbon neutrality.

# Datasets

## Overview

A GHG emissions dashboard and predictive analysis for GMU project requires the collection and analysis of a wide range of data across three main scopes. Scope 1 focuses on direct emissions generated by an organization's operations. This dataset includes critical data like fuel usage, fugitive emissions, and process emissions. It necessitates records of fuel consumption in heating, cooling, and transportation, as well as information on emissions from sources such as leaks and venting in industrial processes. For reliable emissions measurement within Scope 1, detailed records of energy use, specifying types of energy sources and quantities consumed, are also required.

The primary sources of Scope 2 emissions are purchased power and heat. Organizations will require energy bills or invoices indicating electricity and heat consumption to produce this dataset. It also entails utilizing grid emission factors to determine emissions associated with purchased electricity. Data on the purchase or generation of renewable energy is critical for organizations investing in renewable energy sources in order to account for emissions reductions. Finally, Scope 3 emissions, which are frequently the most complex, require a diverse set of data sources. These include supply chain statistics regarding emissions from the manufacture and delivery of purchased goods and services. Employee commuting data, business travel data, waste data, and other categories related to an organization's activities all help to provide a complete picture of Scope 3 emissions.These datasets must be collected, managed, and analyzed properly in order to set reduction objectives, track progress, and make informed choices to reduce an organization's carbon footprint across all three scopes.

## Field Descriptions

### Fleet Fuel Usage Dataset

* **Year (Type: Integer)** – This field describes the year of the fleet fuel usage
* **E85 (Type: Integer)** – This field describes the amount of fuel used with ethanol-gasoline blend (contains 51% to 83% ethanol) used by flex-fuel vehicles.
* **Diesel (Type: Integer)** – This field describes the amount of fuel used by diesel engines.
* **Gas - 10% Unlead Blend (Type: integer)** – This field describes the amount of fuel used in gasoline-powered vehicles, which is composed of 10% ethanol and 90% gasoline.
* **Gas - Mid Grade (Type: Integer**) – This field describes the amount of fuel that has octane levels 88-90.
* **Gas – Premium (Type: Integer)** – This field describes the amount of fuel with octane level 91-93, used by high-performance or turbocharged engines.

### Energy Summary Dataset

* **Name of Building (Type: String)**- Identifies the name of the building for which energy summary data is recorded.
* **Usable Area of Building (Type: Numeric)-** Represents the total usable area within the building, measured in square feet.
* **Monthly Electricity Usage (Type: Numeric)** - Specifies the amount of electricity consumed by the building monthly, measured in kilowatt-hours.
* **Number of CO2 Emitted per Month (Type: Numeric)** - Represents the total amount of carbon dioxide emitted due to electricity consumption in a month, measured in kilograms.
* **Pounds of CO2 Emitted per Month (Type: Numeric)** - Specifies the total amount of carbon dioxide emitted due to electricity consumption in a month, measured in pounds.
* **Pound of CO2 Emitted per Usable Area (Type: Numeric)** - This represents the amount of carbon dioxide emitted per square foot of usable area in the building each month, measured in pounds per square foot.

### Fertilizer Dataset

* **Year (Type: Number)** – Represents the year fertilizer has been used
* **Date of Application (Type: Date**) - The date the fertilizer was applied
* **Fertilizer Type (Type: Float)** - The amount of fertilizer used on the specific date with different N-P-K ratios.
* **Co2 Emitted (Type: Float)** - The amount of CO2 emitted per fertilizer.

## Data Context

**Fleet fuel usage Dataset:**

George Mason University's fleet fuel usage dataset, which covers the fiscal years 2012 through 2022, provides details on the university's trends of fuel consumption. The dataset demonstrates a steady 22% decrease in gasoline consumption throughout this ten-year period, from 86,892.9 gallons in 2012 to 66,894.6 gallons in 2022. This decrease is linked to elements like the adoption of hybrid, electric, and fuel-efficient automobiles as well as changes in driving patterns.

In 2022, 69% of fuel usage will be from gasoline, with 11% coming from diesel. Notably, the use of gasoline mixes has decreased from 38% in 2012 to 20% in 2022, perhaps as a result of the accessibility of alternative fuels like E85. Additionally, the use of diesel dropped from 12% to 11%.

The most typical fuel is "gas-unleaded," which probably refers to regular gasoline-powered cars. Diesel and midgrade gas are the next two most favoured choices. E85, a blend of 85% ethanol and 15% gasoline, and gas-10%, probably 10% ethanol and 90% gasoline, are being used more frequently, which is an interesting sign of a shift towards cleaner fuels.The dataset also shows the effects of COVID-19, with fuel consumption in FY2020 significantly declining as a result of less activities connected to higher education and remote work.

Overall, the data indicates that George Mason University is making an effort to lessen its carbon footprint and greenhouse gas emissions, supporting sustainability programs by switching to alternative and environmentally friendly fuels like E85 and gas-10% blend.

**Energy Summary Dataset**

The "Energy Summary Analysis Dataset" is an extensive collection of information that was carefully assembled by a specialized division at George Mason University. Its main objective is to carefully assess how much energy is used in various campus buildings. This dataset is an essential tool for doing in-depth analysis of energy use, with an emphasis on assisting knowledgeable choices on resource allocation and energy efficiency upgrades.

The "Begin Date" and "End Date" columns in this dataset use a defined time period to indicate when energy consumption data for each building was captured. It enables examining how lockdowns, remote work arrangements, and changes in building occupancy may have altered energy consumption patterns given the backdrop of the COVID-19 pandemic. There are many different methods for gathering data, such as metering, utility bills, and monitoring systems.

Building locations, key energy sources, distinctive building identifiers, and specialized energy consumption purposes are among the important factors and fields. Additionally, it offers details on energy measurement systems and related financial expenses.

The organized format of this dataset, frequently in the form of spreadsheets, comprises information about the energy usage of each building. It is applicable in a variety of situations, such as benchmarking energy efficiency, locating potential for cost savings, and evaluating environmental effects. Given privacy and security issues, access to the dataset may be restricted in some ways, especially during a public health emergency.

**Fertilizer Application Records**

The utilization of numerous items at various places from fiscal years FY18 through FY23 is covered by this dataset. Fairfax, POV, Arlington, Scitech, and PSC are notable sites. The dataset includes columns that detail the weights (in pounds) of various items applied on particular dates, with each row matching to a given location and fiscal year. The collection offers information on how fertilizers are generally used for soil treatment and fertilization. The "Date of Application" column lists the usage dates for each product.

Analysis of the dataset might disclose important details about how products are distributed throughout different locations and consumption trends. By comparing the differences in product quantities over several fiscal years, researchers can identify patterns in agricultural and soil treatment operations. Comparing places like Fairfax and POV may also shed light on the preferences and demands of the community.

This dataset is used for a variety of tasks, including product performance evaluations, environmental impact analyses, and agricultural planning. In order to help decision-makers improve agricultural and environmental approaches, it provides a historical record of evolving product application techniques.

In conclusion, this dataset provides a thorough history of product applications over multiple geographies and fiscal years. Numerous possible uses for it exist, such as agricultural planning, environmental impact analysis, and product efficacy assessments. Researchers and decision-makers can learn a lot about product usage trends and improve agricultural and environmental practices by analyzing the data provided throughout.

## Data Conditioning

**Data collection and inspection:**

* Obtained the dataset on greenhouse gases from the sustainability office.
* Examined the dataset to comprehend its structure, considering the different types of variables, data formats, and any missing or incorrect values.

**Handling Missing Data:**

* Recognized and appropriately handled the missing data.
* If a row or column has a large percentage of missing values but is not essential to the analysis, it was removed. Using techniques like mean, median, or regression, impute missing values.

**Data cleaning:**

* Corrected data points that were anomalous or inaccurate. Considered eliminating or altering outliers because they can have a major impact on analytical outcomes.

**Feature Engineering:**

* Produced innovative features and insightful data for analysis. Data aggregation using emission rate calculations is part of this data.

**Integration of Data:**

* Worked with various datasets, appropriately integrating them using shared keys or IDs. During integration, consistency and data integrity were guaranteed.

## Data Quality Assessment

Several substantial data quality difficulties exist during the dataset compilation process, which might impair its reliability and usability for analysis. For starters, missing values are visible in some columns, which might lead to knowledge gaps and influence calculations across multiple datasets. To ensure that the datasets are complete and available for analysis, missing values must be corrected using data imputation or handling procedures.

Inconsistencies in date formats throughout the "Begin Date" and "End Date" columns, on the other hand, can impede reliable time-based analysis. These inconsistencies make it difficult to do temporal analysis or align data from diverse sources. It is critical to organize date formats and ensure that dates are correct and consistent across all datasets in order to do effective time-series analysis.

Furthermore, the time intervals vary throughout the datasets. Some databases provide monthly data, while others provide yearly data. Because disparate datasets may need to be aligned or aggregated appropriately to create a cohesive timeline, this mismatch might hamper data integration and analysis. It is critical to account for these variances and think about the best time resolution for the given analytical goals.

Outliers in electricity usage and cost values, as well as disparities in the emissions information, may also suggest data mistakes or anomalies. Outliers can distort statistical analysis and must be handled through data cleansing or validation procedures to verify the data's trustworthiness.

The fertilizer dataset has significant data quality issues that limit its utility for agricultural studies. These problems include differences in fertilizer mix, different measuring units, and uneven reporting of nutrient values. A complete study and harmonization of fertilizer formulas, standardization of measurement units, and validation of nutrient content are required to assure the credibility of this dataset.

Furthermore, the existence of unexpected zero values for electricity usage in some entries raises questions about the accuracy and completeness of the data. It is critical to investigate these zero numbers in order to discover whether they are valid readings or whether there are recording problems that need to be corrected.

Finally, gathering and analyzing these disparate datasets with various time periods and data quality issues might be difficult. A rigorous data quality evaluation, data cleansing, and harmonization process are required to assure the data's dependability and usability for analysis and decision-making. Furthermore, domain-specific knowledge and context are critical for effectively addressing dataset-specific difficulties and drawing relevant insights from the assembled data.

## Other Data Sources

For estimating its carbon footprint, George Mason University (GMU) took the strategic choice to utilize the SIMAP of the University of New Hampshire. This decision demonstrates GMU's dedication to a thorough and open assessment of greenhouse gas (GHG) emissions. SIMAP is a well-known and respected technique made especially for evaluating GHG emissions within the higher education sector.

GMU benefited in various ways by implementing SIMAP:

Reliable GHG Assessment: SIMAP offers a standardized framework for gathering and processing data on GHG emissions. This guarantees the accuracy and thoroughness of the GHG inventory process.

Advice and Expertise: SIMAP provides GMU, and other organizations seeking to measure and lessen their carbon footprint with helpful advice and resources. It helps with boundary selection, emission factor choice, and data accuracy. Independent evaluation: The use of SIMAP made it possible for the GMU GHG inventory to undergo an impartial evaluation. The university's emissions data are validated and made accurate by this external evaluation.

**A Second Nature review:**

GMU went one step further by sending its data to Second Nature after utilizing SIMAP to complete its GHG inventory. The Presidents' Climate Leadership Commitments, originally known as the American College & University Presidents' Climate Commitment, are supported by Second Nature, a well-known organization. The institution is in line with a broader commitment to addressing climate change within the higher education sector by providing Second Nature with GMU's GHG inventory. It displays GMU's commitment to openness and responsibility in its sustainability initiatives. As an impartial organization, Second Nature is essential in validating and confirming the sustainability accomplishments and promises made by higher education institutions.

GMU demonstrates its commitment to being held responsible for its carbon reduction efforts and promotes its image as an organization that is actively addressing climate change by having Second Nature examine its GHG inventory. Overall, GMU's engagement with Second Nature and use of the SIMAP platform demonstrate a thorough and rigorous approach to sustainability and carbon neutrality, ensuring that the university's actions are in line with acknowledged norms and obligations in the higher education sector.

## Storage Medium

The project includes datasets from diverse areas, such as energy consumption, utility data, fleet emissions, and fertilizer data, all of which have unique characteristics and requirements. A multi-tiered storage approach is suggested to ensure smooth data handling, accessibility, security, and scalability.

For datasets with well-structured schemas and time-series properties, such as energy consumption and utility data, a robust RDBMS such as PostgreSQL or MySQL is recommended. This option maintains data integrity, allows for complicated searches, and can handle the periodic nature of the data quickly.

A NoSQL database system would benefit datasets with variable properties and semi-structured or unstructured forms, such as fleet emissions and fertilizer data. Options like MongoDB and Cassandra provide the flexibility needed to properly manage varied data kinds.

For huge datasets that require advanced querying and reporting capabilities, cloud-based data warehousing systems such as Amazon Redshift or Google BigQuery are recommended. These platforms are especially well-suited for historical and aggregated data analysis, as they offer scalability and performance.

Creating a data lake using services like Amazon S3 or Azure Data Lake Storage is recommended to keep the raw, unfiltered version of the datasets before changes. A data lake saves original datasets, allowing for future analytics and exploration while also providing data transformation and processing pipelines.

Implementing a version control system like Git is critical for maintaining the integrity and repeatability of code scripts, data conversions, and analysis pipelines. This assures change traceability and enables communication among project stakeholders.

In conclusion, the diversity of the datasets in the project needs a careful approach to data storage. The recommended multi-tiered storage solution handles each dataset's specific characteristics, improving accessibility, security, and scalability while maintaining data integrity throughout the project's lifecycle. The project can ensure efficient data management, data-driven insights, and informed decision-making by using the proper storage solutions for each dataset type.

## Storage Security

To ensure the security of the project's diverse datasets, a comprehensive approach is essential. Multi-tiered storage solutions, coupled with robust access controls and encryption mechanisms, play a pivotal role in safeguarding sensitive data.

The use of RBAC and encryption at rest and in transit should be implemented to restrict unauthorized access and protect data both during storage and transmission. Regular security audits and monitoring help detect and respond to potential threats promptly, ensuring data remains protected throughout its lifecycle.

Additionally, maintaining data integrity and privacy compliance, such as GDPR or HIPAA where applicable, is crucial. This includes data anonymization and de-identification techniques for sensitive information.

By adopting a holistic security strategy encompassing data storage, access controls, encryption, and compliance measures, the project can mitigate risks, uphold confidentiality, and maintain the trust of stakeholders while leveraging its diverse datasets for valuable insights and decision-making.

## Storage Costs

Storage costs for this project can vary significantly based on the chosen storage solutions and the volume of data. For datasets with well-structured schemas and time-series properties like energy consumption and utility data, using traditional RDBMS such as PostgreSQL or MySQL typically incurs moderate costs.

For datasets with variable properties and semi-structured or unstructured forms, such as fleet emissions and fertilizer data, NoSQL database systems like MongoDB or Cassandra offer flexibility at a reasonable cost.

Cloud-based data warehousing systems such as Amazon Redshift or Google BigQuery are suitable for large datasets requiring advanced querying and reporting capabilities, but they may involve higher costs, particularly as data volume increases.

Creating a data lake using services like Amazon S3 or Azure Data Lake Storage can be cost-effective for storing raw, unfiltered datasets, but operational costs may arise from data transformation and processing pipelines.

In summary, precise storage costs can fluctuate based on dataset characteristics and the scalability of chosen storage solutions. It's essential to strike a balance between data accessibility, security, and cost-effectiveness to ensure efficient data management throughout the project's lifecycle.

# Algorithms & Analysis / ML Model Exploration & Selection

## Solution Approach

### Systems Architecture

Gather data from various sources, such as databases, APIs, or CSV files. Ingest and save the data in a model- and analysis-friendly manner.

Engineering features and performing data preprocessing tasks including resolving missing values, scaling features, and encoding categorical variables are examples of tasks that fall under this category.

If necessary, incorporate features to enhance the model's functionality. Create training and test datasets from the preprocessed data. Utilize the training set of data to create a model.

Utilize the testing dataset to verify the model's effectiveness. If necessary, adjust hyperparameters for better outcomes. Install the learned model in a real-world setting. Make a service or API to expose the model for in-the-moment predictions. Make sure that the deployed model can handle incoming requests and is scalable.

A diagram of data processing

Description automatically generated

### Systems Security

The security of the GHG Emissions Monitoring Project's data flow remains paramount to its effectiveness and integrity. To safeguard the system, a multi-layered security strategy is in place. Firstly, robust access controls and authentication mechanisms ensure that only authorized personnel can access and modify the data. Role-based access control (RBAC) is implemented, granting specific permissions based on user roles and responsibilities.

Secondly, data encryption is applied both in transit and at rest, ensuring that sensitive user inputs and calculated emissions data are protected from unauthorized access during transmission and while stored within the system.

Regular security audits and monitoring processes are conducted to detect and respond to potential threats promptly. Intrusion detection systems and log analysis are employed to identify any unusual activities or breaches in real-time, maintaining data integrity and confidentiality.

Additionally, compliance with relevant data privacy regulations, such as GDPR or HIPAA, is upheld to ensure that user data is handled in a lawful and ethical manner, reinforcing transparency and accountability.

### Systems Data Flows

**A diagram of a data processing process

Description automatically generated**

The GHG Emissions Monitoring Project includes a carefully planned data flow that forms the framework of extensive environmental sustainability measures. The project's central focus is the effective monitoring and control of greenhouse gas (GHG) emissions. User inputs and computed data are the two primary forms of data inputs that it uses. User inputs, which are manually submitted by users and serve as the foundation for emissions estimates, include essential data including vehicle specifications, fuel consumption, and mileage. On the other hand, calculated data shows the results of complex calculations and provides useful metrics like overall GHG emissions, scope-specific allocations, and precise source identifications.

This project's data flow is a well-organized procedure with three crucial phases: data management and storage, predictive analysis, and data presentation. In order to guarantee data integrity, data storage ensures the safe and orderly storing of user inputs and calculated emissions data. In order to forecast future trends in emissions, predictive analysis makes use of historical data and existing patterns, enabling proactive decision-making. The project's strength is seen in the data presentation, where complex emissions data is translated into formats that users can easily understand and apply to gain useful insights.

This data-driven project offers numerous advantages, including enhanced visibility into emissions data, data-driven decision-making capabilities, and increased transparency and accountability. By leveraging the data flow, universities can identify areas for emissions reduction and make informed choices aligned with sustainability goals. The project's holistic approach empowers universities to effectively reduce their environmental footprint, contributing to a more sustainable future.

The GHG Emissions Monitoring Project's data flow is, in essence, more than just a technical infrastructure; it is an essential weapon in the global fight against climate change. It gives businesses the tools they need to get a clear, data-driven understanding of emissions, enabling them to make wise, sustainable choices. In order to promote a more sustainable and environmentally conscious future, this project's complete approach to data gathering, analysis, and presentation positions companies to successfully navigate the challenging emissions reduction landscape.

### Algorithms & Analysis

The greenhouse gas emissions prediction project uses a machine learning algorithm to predict greenhouse gas emissions based on historical data. The specific algorithm used will depend on the specific data set and the desired outcomes. However, some common types of models for greenhouse gas emissions prediction include:

* Regression models: Regression models are a type of machine learning algorithm that is used to predict a continuous output variable based on one or more input variables. Regression models are often used to predict greenhouse gas emissions because they can be trained on historical data to identify the relationships between the input variables and the output variable.
* Time series models: Time series models are a type of machine learning algorithm that is used to predict future values of a time series based on its past values. Time series models can be used to predict greenhouse gas emissions because they can account for the temporal relationships between the input and output variables.
* Decision tree models: Decision tree models are a type of machine learning algorithm that is used to predict a categorical output variable based on one or more input variables. Decision tree models can be used to predict greenhouse gas emissions because they can be trained on historical data to identify the decision rules that are most predictive of different levels of emissions.

Analysis

Once a machine learning model has been trained and evaluated, it can be used to analyze greenhouse gas emissions data and identify trends and patterns. For example, the model can be used to:

* Identify the sectors and activities that are contributing the most to greenhouse gas emissions
* Assess the impact of different policies and initiatives on greenhouse gas emissions
* Forecast future greenhouse gas emissions under different scenarios

The analysis of greenhouse gas emissions data can be used to inform decision-making about how to reduce emissions and mitigate the effects of climate change.

Additional considerations

When developing and deploying a greenhouse gas emissions prediction model, it is important to consider the following:

* Data quality: The quality of the data used to train the model has a significant impact on the accuracy of the model's predictions. It is important to carefully clean and prepare the data before using it to train the model.
* Model selection: The best model for a particular project will depend on the specific data set and the desired outcomes. It is important to carefully consider different types of models and select the one that is most likely to produce accurate and useful predictions.
* Model evaluation: It is important to evaluate the performance of the model on a held-out dataset before deploying it to production. This helps to ensure that the model is able to generalize to new data and is not simply overfitting to the training data.
* Model monitoring: Once the model has been deployed, it is important to monitor its performance and make adjustments as needed. This is because the relationships between the input and output variables can change over time.

By carefully considering these factors, it is possible to develop and deploy a greenhouse gas emissions prediction model that can produce accurate and useful predictions.

## Machine Learning

* Multi-Scope Challenge: GHG emissions are divided into three scopes, each of which has its own set of parameters and data sources. It takes a sophisticated, multi-modal machine learning model to create a CRS for all scopes.
* Training Data: The model is trained using historical emissions data, including direct emissions (Scope 1), indirect emissions linked to energy (Scope 2), and indirect emissions connected to value chains (Scope 3). The dataset contains details on the sources of emissions, how much energy is consumed, who the suppliers are, and other pertinent information.
* Model Evaluation: Metrics like Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) are used to evaluate the performance of the model.
* Specific Recommendations for Each Scope: Scope 1: Based on factors including output, energy consumption, and operating parameters, direct emissions are projected. To cut emissions, operational processes can be optimized.
  + Scope 2: Energy-related emissions are anticipated, and suggestions include predicting energy trends and encouraging the use of cleaner energy sources. Scope 3: Forecasted emissions from value chain vendors. These emissions can be decreased with better logistics.
  + The CRS is incorporated into an environmental management system and offers projections and recommendations for emission reductions to decision-makers. Additionally, it automates GHG reporting and measurement, making regulatory compliance and sustainability objectives simpler.
  + Model Selection: The dataset and goals influence which machine learning algorithm is used. Regression models, time series models, and decision tree models are examples of common models.
* Model Analysis: Data on emissions are analyzed using the model after training and evaluation. It identifies industries that contribute significantly to emissions, evaluates the effects of policies, and projects future emissions based on various scenarios.
* Implementation Considerations: A successful implementation depends on factors including model choice, data quality, rigorous evaluation, and ongoing monitoring.

### Model Exploration

The K-Nearest Neighbors (KNN) method is a machine learning technique used in the context of a carbon emissions reduction project to provide insights and make predictions based on previous emissions data. The method used to forecast future CO2 emissions is based on the similarity of data points. In this instance, the month and the sources of energy use are examples of input variables. The quantity of CO2 emissions is the output variable.

KNN makes use of historical data that includes details on the energy consumption sources and associated CO2 emissions for every month. By examining the historical patterns and associations between these input variables and the output variable (CO2 emissions), KNN is capable of predicting the expected level of CO2 emissions for future months. It does so by identifying the K nearest data points in the historical dataset that closely resemble the current scenario in terms of energy sources and month, and then computes the emission prediction based on their emissions data.

Moreover, KNN is proficient in identifying anomalies or unusual emissions events. It accomplishes this by comparing the current data point, representing the energy sources and month, to those in the historical dataset. If the current data point significantly deviates from the typical patterns observed in historical data, it can signal an anomaly, such as equipment malfunctions or irregularities in energy consumption, warranting immediate attention and corrective actions. Thus, KNN serves as a valuable tool for both emissions prediction and anomaly detection in the quest for effective carbon emissions reduction.

### Model Selection

For precise forecasts and anomaly detection in a project to reduce carbon emissions, choosing the right machine learning model is crucial, especially when working with binary emissions data where the objective is to distinguish between emissions and non-emissions.

For many other reasons, K-Nearest Neighbors (KNN) turns out to be the better option. The KNN model is non-parametric, which gives it flexibility and adaptability to different data distributions. This is especially helpful when dealing with emissions data, a typical feature of environmental datasets that may not follow linear or specified patterns. KNN is excellent at classification tasks like predicting binary outcomes; it is not just good at regression. It locates the 'K' nearest data points that closely resemble the current circumstance, making accurate projections of future emissions possible. The main competency of KNN is its similarity-based methodology. Analyzing how closely the current circumstance resembles previous patterns is crucial for an emissions forecast.

By comparing the present data point to comparable historical data, KNN successfully satisfies the need for an emissions forecast based on historical data. KNN is also good at detecting anomalies. By comparing the present data point to past data, it identifies departures from regular patterns, enabling the diagnosis of anomalies necessitating prompt attention.

This project could not be a good fit for linear regression, which is frequently used for continuous numerical forecasts. In binary emissions data, linear connections between variables may not hold, which could result in errors and inappropriate forecasts. For binary emissions data, the model's assumption of constant variance across various values of independent variables is not ideal.

K-Nearest Neighbors (KNN) is the method of choice for this project to reduce carbon emissions because of the characteristics of emissions data, the emphasis on similarity-based forecasts, and anomaly detection. It is the ideal model for emissions forecast and anomaly detection since it fits the project's goals and data properties effectively.

A screenshot of a graph

Description automatically generated

# Visualizations / ML Model Training, Evaluation, & Validation

## Overview

The GMU campus's GHG Emissions Dashboard is for tracking and overseeing environmental efforts. It uses Tableau for data visualization and K-Nearest Neighbors (KNN) machine learning for emissions analysis. Users can obtain data-driven insights regarding the energy and carbon footprint of the university through this dashboard. Users may easily analyze emissions data by using Tableau's advanced visualization features, which offer an engaging and informative interface. A range of charts, graphs, and maps are included in visualizations, which provide dynamic views of sustainability measures, energy use, and greenhouse gas emissions.

The data can be filtered and further examined by users to identify trends, sources of emissions, and development opportunities. The GMU community may stay involved in the university's sustainability, make informed decisions, and monitor progress toward sustainability goals by having online access to the dashboard. Through this program, GMU will be better equipped to mitigate its environmental impact while also increasing transparency and awareness.

## Visualizations

**A graph of emission

Description automatically generated with medium confidence**

The bar graph shows the total carbon dioxide (CO2) emissions of George Mason University (GMU) campus over a year, calculated monthly. The x-axis represents the month, and the y-axis represents the total CO2 emissions in pounds. The graph shows that the highest CO2 emissions occurred in September and the lowest CO2 emissions occurred in January. However, the difference between Highest and lowest is herding notable with a value of around 2M pounds.

A screenshot of a graph

Description automatically generated

The visualization shows the scope 3 user input monthly emissions of GMU campus. Scope 3 emissions are those that are produced indirectly by the university, such as the emissions from commuting students, faculty, and staff. The worksheet allows user input of “% of change in average commuter” for a particular month on the right side. This allows us to identify the changes in emissions if the average number of commuters is changed in that month.

A screenshot of a computer

Description automatically generated

The visualization shows the scope 3 commuter emissions of George Mason University (GMU) campus, broken down by student, faculty, and staff. The visualization shows that students produce the most scope 3 commuter emissions, followed by faculty and staff. This might be due to the vast number of student population compared to faculty or staff. Students might also have longer commuted than faculty and staff. Another possibility is that students are more likely to drive to campus than faculty and staff. It is also possible that the variation in emissions is due to a combination of factors.

A graph of a graph

Description automatically generated

The visualization shows the scope 2 user input monthly emissions of George Mason University (GMU) campus. Scope 2 emissions are those that are produced indirectly by the university, such as the emissions from the generation of electricity that is used to power the campus. The graph shows that the highest CO2 emissions occurred in September and the lowest CO2 emissions occurred in January. The worksheet allows user input of “% of change in Electricity usage” for a particular month on the right side. This allows us to identify the changes in emissions if the electricity usage is changed in that month.

A screenshot of a computer

Description automatically generated

The visualization shows the scope 2 emissions of George Mason University (GMU) campus building wise, in pounds. The visualization shows that the Johnson Center has the highest scope 2 emissions, followed by the Aquia building, mason global center, engineering building. The buildings with the lowest scope 2 emissions are the whitetop hall, washington hall and truman Hall. The worksheet allows user input of “% of change in Electricity usage” for a particular month on the right side. This allows us to identify the changes in emissions of that building if the electricity usage is changed in that month.

A screenshot of a computer

Description automatically generated

The visualization shows the scope 1 fuel/fleet emissions of George Mason University (GMU) campus, broken down by fleet type. Scope 1 emissions are those that are produced directly by the university, such as the emissions from the combustion of fuel in university vehicles. The visualization shows that the gasoline fleet has the highest scope 1 fuel/fleet emissions, followed by the diesel fleet, the CNG fleet, and the E85 fleet.

A screenshot of a graph

Description automatically generated

The visualization shows the scope 1 fuel/fleet emissions of George Mason University (GMU) campus, broken down by fleet type and month. The worksheet allows user input of “% of change in Gasoline usage” for a particular month on the right side. This allows us to identify the changes in emissions of if the gasoline usage is changed in that month.

A screenshot of a computer

Description automatically generated

The visualization shows the scope 1 fertilizer emissions of George Mason University (GMU) campus, broken down by fertilizer type. Scope 1 emissions are those that are produced directly by the university, such as the emissions from the application of fertilizer to campus lawns and gardens. The visualization shows that the most common fertilizer types used on GMU campus are 19-0-15, 17-0-5, and 0-45-0.

A graph of growth in a graph

Description automatically generated with medium confidence

The visualization shows the scope 1 fertilizer emissions of George Mason University (GMU) campus, broken down by month. Scope 1 emissions are those that are produced directly by the university, such as the emissions from the application of fertilizer to campus lawns and gardens. This visualization comes with an assumption that a fertilizer emits CO2 till 6 months after it has been applied/used.

A graph of a graph of a graph

Description automatically generated with medium confidence

The visualization shows the scope 1 LNG/natural gas emissions of George Mason University (GMU) campus, broken down by month. Scope 1 emissions are those that are produced directly by the university, such as the emissions from the combustion of LNG/natural gas in university boilers and furnaces.

The visualization shows that the highest scope 1 LNG/natural gas emissions occur in the winter (December, January, and February). The lowest scope 1 LNG/natural gas emissions occur in the summer (June, July, and August).The worksheet allows user input of “% of change in natural gas usage” for a particular month on the right side. This allows us to identify the changes in emissions of if the natural usage is changed in that month.

## Machine Learning

GMU's GHG Emissions Dashboard utilizes the K-Nearest Neighbors (KNN) machine learning algorithm to analyze emissions data and support sustainability objectives. KNN, a supervised learning method, leverages data point similarity to predict future emissions based on historical data, weather conditions, and energy consumption. This information is presented through charts and graphs to visualize emissions trends and progress. Moreover, KNN assists in identifying emission sources and reduction opportunities, facilitating the development of targeted strategies and recommended actions for emissions reduction.

Machine learning is highly effective in carbon emissions reduction due to its capacity to handle complex and voluminous data. It can process large datasets, discern intricate patterns, and make data-driven predictions, adapting to various data sources and types. KNN stands out for its simplicity and interpretability, allowing stakeholders to grasp the rationale behind specific predictions, fostering trust and transparency in emissions reduction efforts. It excels in anomaly detection, swiftly identifying deviations from expected emissions patterns, which is crucial for addressing equipment malfunctions and energy consumption irregularities. Additionally, KNN is proficient in emissions prediction, using historical data and relevant variables to anticipate future emissions, enhancing the ability to proactively develop emissions reduction strategies. This makes KNN the preferred choice for this project.

### Model Training

There are several essential steps involved in training a K-Nearest Neighbors (KNN) model for predicting greenhouse gas (GHG) emissions. First, training and testing data sets are separated, usually with a 70% training and 30% testing split. The fuel usage records, energy consumption statistics, emission factors, operational information, and labels identifying the scopes of the emissions (Scope 1, Scope 2, and Scope 3) are among the features included in the dataset.

KNN is a supervised learning algorithm that uses similarity as its foundation. To generate predictions, it evaluates how close together data points are. Instead of undergoing conventional training in the instance of GHG emissions, KNN retains the training data together with its labels for future use. It is therefore an instance-based or memory-based algorithm.

To make predictions, KNN uses a selected distance metric, such as Euclidean distance, to determine how similar a test data point is to its k nearest neighbors. One important hyperparameter that affects prediction quality is the number of neighbors or KNN can do classification (classifying emissions into the three scopes) or regression (predicting emissions levels), depending on the task. Model evaluation is essential after predictions are made. While classification tasks employ metrics like accuracy, precision, recall, and the F1-score to evaluate the model's classification performance, regression tasks commonly use MSE to estimate the accuracy of predictions.

It might be required to fine-tune hyperparameters, particularly k, to maximize the model's performance. KNN is an important tool for estimating GHG emissions, supporting sustainability initiatives, and compliance reporting due to its ease of use and capacity to identify regional patterns. It offers a simple yet efficient method for estimating emissions from different sources by comparing them to nearby data points, which promotes environmental responsibility and well-informed decision-making.

### Model Evaluation

In this project, the primary objective is to rigorously evaluate the effectiveness of a K-Nearest Neighbors (KNN) model in predicting CO2 emissions. The evaluation process commences with the partitioning of the dataset into distinct training and testing subsets, with 70% of the data designated for training and the remaining 30% reserved for testing. This partitioning is crucial because it allows us to assess how well the model can generalize its predictions to new, unseen data. Subsequently, a set of regression evaluation metrics, including MAE, MSE, and RMSE, are applied. These metrics provide quantitative measures of the disparities between the model's predicted CO2 emissions and the actual emissions, offering a basis for accurately assessing the model's predictive performance.

To ensure the model's robustness and reliability, k-fold cross-validation is employed. This technique involves dividing the dataset into k subsets and systematically training the model on k-1 subsets while testing it on the remaining subset. This process is repeated k times, with the testing subset rotating each time. Cross-validation provides a comprehensive view of the model's performance, allowing us to assess its ability to handle diverse data scenarios, reduce overfitting, and enhance our confidence in its predictive capabilities. Moreover, the evaluation process includes hyperparameter tuning, with a specific focus on optimizing the number of neighbors (k) in the KNN model. Grid search or randomized search techniques are employed to identify the hyperparameters that lead to the best model performance. This step ensures that the KNN model operates at its peak efficiency and aligns with the project's objectives. Lastly, the interpretability of the KNN model is assessed through visualizations like scatter plots comparing predicted and actual CO2 emissions. These visual aids help identify patterns, trends, and potential areas for model improvement, enhancing our understanding of how the model makes predictions.

In summary, this project systematically evaluates a K-Nearest Neighbors model's performance in predicting CO2 emissions. It involves data partitioning, the use of regression metrics, k-fold cross-validation for robustness, hyperparameter tuning for optimization, and visualizations for interpretability. These steps collectively ensure a comprehensive assessment of the model's capabilities, accuracy, and reliability in predicting CO2 emissions.

|  |  |  |
| --- | --- | --- |
| **Model** | **RMSE** | **Accuracy** |
| KNNBasic | 0.85 | 0.91 |
| KNNMeans | 0.89 | 0.90 |

### Model Validation

Developing a robust and effective KNN model for predicting greenhouse gas (GHG) emissions involves a systematic approach encompassing various critical factors. Firstly, hyperparameter tuning and validation are pivotal to optimize the model's performance. Techniques such as k-fold or stratified cross-validation should be employed to assess the model's generalization ability and mitigate overfitting concerns. Feature scaling is imperative due to KNN's reliance on distance measures. Standardizing or normalizing features ensures that no single feature disproportionately influences distance estimates. Addressing incomplete or missing data through techniques like imputation or judicious removal is crucial to maintain data integrity.

Secondly, the dimensionality of the dataset should be considered, and feature selection may be employed to reduce the risk of overfitting. Outliers, which can significantly impact nearest neighbor estimates, should be identified and handled diligently. Visualizations play a key role in understanding feature correlations, data distributions, and model predictions, providing insights into the model's behavior. While KNN is often considered a black-box model, efforts can be made to decipher and understand its predictions. Improved model interpretability is achieved by comprehending decision boundaries and the influence of specific features on predictions.

Lastly, ensuring the reliability of the KNN model involves deploying it in a manner that reflects its training and testing environments. Establishing a mechanism for continuous performance tracking is crucial, particularly with evolving datasets. This ongoing monitoring helps detect any drift in data distribution or model degradation over time. Iterative refinement through thorough validation and consistent evaluation ensures the creation of a trustworthy and valuable forecasting tool for GHG emissions prediction, aligning the model with real-world applications and enhancing its overall robustness and usefulness.

# Findings

The comprehensive greenhouse gas emissions analysis and prediction project conducted by George Mason University (GMU) encompassed a meticulous investigation across Scopes 1, 2, and 3. The initial phase involved a meticulous data collection process, spanning fuel consumption, energy statistics, emission details, and Scopes 1-3 classification, laying the groundwork for the project's subsequent phases.

A detailed examination through visualizations elucidated monthly emission trends, aiding in discerning emission peaks, variations among scopes, and contributions from distinct sources. Leveraging machine learning techniques like regression, time series, and decision trees, the project accurately estimated emissions patterns. The K-Nearest Neighbors (KNN) algorithm particularly excelled in forecasting emissions and detecting anomalies due to its historical data-driven accuracy.

The project's intricate visualizations offered vital insights into emissions sources, highlighting Scope 1 sources such as fleet, fertilizer, and LNG/natural gas, along with Scope 2 sources like electricity consumption and building-wise emissions, and Scope 3 emissions from commuters. These insights guided targeted mitigation efforts effectively.

Moreover, the project delved into diverse sustainability and emissions reduction methods, exploring renewable energy utilization, efficiency enhancements, carbon capture, reforestation, and agricultural practice modifications. It also contributed to the discourse on sustainability by offering insights into emissions assessment research, mitigation techniques, and policy implications.

The ultimate outcome was the development of the Carbon Zero University Tool through iterative Agile methodology. This tool continually enhanced monitoring capabilities, Scope insights, and emissions analysis, enabling GMU and other institutions to track weekly emissions more effectively, thereby facilitating monitoring and reduction of carbon footprints. GMU's systematic approach—from data collection and machine learning to visualizations, research insights, and tool development—positioned it at the forefront of comprehending and addressing emissions challenges, steering strategic initiatives towards sustainability goals.

# Summary

George Mason University's (GMU) greenhouse gas emissions analysis and prediction project set out on a thorough journey to comprehend, forecast, and reduce carbon emissions across Scopes 1, 2, and 3. To lay a solid platform for further study, the project began with a thorough collection of data covering a wide range of topics, including fuel consumption, energy statistics, emission details, and Scopes 1-3 classification.

A detailed analysis of direct and indirect emissions was provided by visualizations created to show monthly emission trends. Understanding emission peaks, variation among scopes, and the distinct contributions of different emission sources was made easier with the help of these intricate visual aids.

Machine learning methods were used to estimate emissions with accuracy. To predict emissions patterns, regression, time series, and decision tree models were applied, each of which made use of past data. Because of its historical data-driven accuracy, the K-Nearest Neighbors (KNN) algorithm stood out among these models for emissions forecasting and anomaly detection.

Important insights into the sources of emissions were made possible by the project's intricate visualizations. They emphasized sources of Scope 1 emissions (fleet, fertilizer, LNG/natural gas), as well as sources of Scope 2 emissions (electricity consumption, building-wise emissions) and Scope 3 emissions (commuters). Targeted mitigation efforts benefited from the complex insights into emissions profiles provided by these visualizations.

The project explored the complex issues of sustainability and emissions reduction programs in addition to their analytical components. It examined a wide range of mitigation techniques, including the use of renewable energy, improved efficiency, carbon capture, reforestation, and modifications to agricultural practices. The project also shaped the conversation around sustainability by offering insights into current research on emissions assessment, mitigation techniques, and policy consequences.

The project's ultimate result was the Carbon Zero University Tool's iterative development, made possible by Agile methodology. The monitoring capabilities, Scope insights, and emissions analysis were continuously enhanced by this instrument. Iterative upgrades included automation, interoperability with sustainability systems, and improvements to the user interface and user experience. The program made it possible for GMU and other colleges to track their weekly emissions and provide more detailed reporting, which helped them to monitor and reduce their carbon footprints.

GMU was positioned at the forefront of comprehending and addressing emissions challenges through this methodical approach, which included everything from data collection and machine learning applications to detailed visualizations, research insights, and tool development. This allowed for informed decision-making and propelled strategic initiatives toward the achievement of sustainability objectives.

# Future Work

* Real-Time Data Integration: A shift to real-time data integration will enable for the capturing of instantaneous changes in emissions, allowing for more dynamic and responsive mitigation techniques. Implementing Internet of Things (IoT) sensors or API connections with power grids can provide a constant stream of data for more accurate and timely emissions calculations.
* Policy Impact Predictive Analytics: Create predictive analytics to estimate the probable impact of various sustainability policies or projects. Simulate the results of various carbon reduction options prior to deployment to inform decision-makers and optimize resource allocation.
* Collaborative Platforms: Develop a collaborative web platform or mobile application that engages the university community in efforts to reduce emissions. Carbon footprint tracking, emission reduction challenges, and a venue for discussing sustainable practices and ideas might all be included on this platform.
* Gamification for Behavioral Change: Incorporate gamification aspects into the university community to foster sustainable behaviors. Stakeholders, including students and teachers, might be inspired to actively participate in reaching carbon reduction goals by turning emission reduction activities into a game or challenge.
* Data Visualization Innovations: Push the boundaries of data visualization by experimenting with immersive technologies like virtual or augmented reality. Creating interactive, three-dimensional visualizations of emissions data could offer a more engaging and insightful experience for stakeholders.

Appendix

Appendix A: Glossary

|  |  |
| --- | --- |
| **Term** | **Definition** |
| GHG | Greenhouse gases |
| N-P-K | Nitrogen Phosphorus Potassium |
| CCS | carbon capture and storage |
| ICE | Internal Combustion Engine |
| CHP | Combined heat and Power |
| SIMAP | Sustainability Indicator Management & Analysis Platform |
| RBAC | Role-Based Access Control |
| GDPR | General Data Protection Regulation |
| RDBMS | Relational Database Management Systems |
| KNN | K-Nearest Neighbors |
| RMSE | Root Mean Square Error |
| MAE | Mean Absolute Error |
| MSE | Mean Squared Error |

Table : Glossary Table

Appendix B: GitHub Repository

GitHub Repository Link

<https://github.com/DataDynamos/GHG-Emissions-Dashboard-and-Predictive-Analytics-for-GMU-Campus.git>

Appendix C: Risks

Sprint 1 Risks

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Risk | Description | Probability | Impact | Mitigation |
| Incomplete Data or Inaccuracies | Incomplete or inaccurate information may lead to incorrect emissions calculations and analysis, affecting the tool's reliability | Moderate | High | Implement data validation checks and quality control measures to identify and rectify data issues. Work closely with data providers to ensure data completeness and accuracy |
| Resistance to Behavioral Changes: | Achieving carbon neutrality may require changes in behavior and practices among students, faculty, and staff, which could face resistance. | Moderate | Moderate | Develop comprehensive awareness and engagement campaigns to educate and involve the campus community in sustainability efforts. Provide incentives and rewards for eco-friendly practices |
| Budget Constraints | Budget limitations may impede the development and implementation of the tool and related emission reduction | Moderate | High | Seek external funding sources, grants, or partnerships to support the project's financial needs. Prioritize initiatives with a high return on investment and cost-effective solutions |
| Data Privacy and Security | Managing sensitive emissions data and ensuring its privacy and security may pose risks | Low | High | Implement robust data encryption, access controls, and cybersecurity measures. Comply with relevant data privacy regulations and standards to protect data integrity |
| Changing Energy Market | Fluctuations in energy prices and availability of renewable energy sources may affect the feasibility of certain emission reduction strategies. | Moderate | Moderate | Diversify energy sourcing strategies to adapt to changing market conditions. Explore long-term contracts with renewable energy providers to stabilize energy costs. |

Table : Sprint 1 Risks

Sprint 2 Risks

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Risk | Description | Probability | Impact | Mitigation |
| Incomplete Data or Inaccuracies | Incomplete or inaccurate data may lead to incorrect emissions calculations and analysis, affecting the tool's reliability | Moderate | High | Implement data validation checks and quality control measures to identify and rectify data issues. Work closely with data providers to ensure data completeness and accuracy |
| Limitations of Developed Model/Tool | Due to Scope 3 inconsistent data, there is a limitation for calculating emissions for commuting | Medium | Low | Adopt iterative development with continuous testing and feedback loops. Engage emission experts in the model's creation and validation processes |
| Misinterpretation of Scope 1 and Scope 2 Emissions Impact | The risk of misinterpreting Scope 1 and Scope 2 emissions lies in failing to accurately identify and account for all sources within an organization's operational boundaries. | Low | Medium | Conduct regular training sessions and maintain clear and detailed documentation of definitions and differences between Scope 1 and Scope 2 emissions. Use automated checks to validate categorizations. |
| Changes in Power Generation Sources Over Time | The risk here lies in not accurately accounting for these shifts in the energy mix. | Medium | High | Design the model to be adaptable and modular. Schedule regular updates or reviews to incorporate the latest data on power generation sources. |

Sprint 3 Risks

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Risk | Description | Probability | Impact | Mitigation |
| **Data Quality Issues** | Incomplete, inaccurate, or biased data can lead to incorrect model predictions and analysis. | Medium | High | Conduct thorough data preprocessing, cleaning, and validation. Implement data quality checks and monitoring systems. |
| **Model Overfitting** | Models may be too complex and fit the training data perfectly but perform poorly on new data. | Medium | Medium | Use appropriate model complexity based on available data. Implement  cross-validation techniques. |
| **Bias and Fairness** | Biases in data or model training can lead to unfair or discriminatory analysis and predictions. | Medium | High | Regularly audit and assess models for bias. Use diverse and representative training data. |

Sprint 4 Risks

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model Overfitting** | Models may be too complex and fit the training data perfectly but perform poorly on new data. | Medium | Medium | Use appropriate model complexity based on available data. Implement  cross-validation techniques. |
| **Bias and Fairness** | Biases in data or model training can lead to unfair or discriminatory analysis and predictions. | Medium | High | Regularly audit and assess models for bias. Use diverse and representative training data. |

Sprint 5 Risks

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model Overfitting** | Models may be too complex and fit the training data perfectly but perform poorly on new data. | Medium | Medium | Use appropriate model complexity based on available data. Implement  cross-validation techniques. |
| **Bias and Fairness** | Biases in data or model training can lead to unfair or discriminatory analysis and predictions. | Medium | High | Regularly audit and assess models for bias. Use diverse and representative training data. |

Appendix D: Agile Development

Scrum Methodology

In order to address emissions reduction, the team will collaborate. We outlined precise and quantifiable objectives for lowering greenhouse gas emissions. These objectives must be clear, quantifiable, doable, timely, and relevant. Making a task for Youtrack encompasses all feasible steps and initiatives for emission reduction. These could consist of new energy-saving technologies, renewable energy sources, altered government regulations, and more. divided the emissions reduction projects into 2-4 weeks "sprints," or time-limited iterations. The team chooses which items from the product backlog to work on during the sprint during sprint planning.

The team works on the chosen initiatives and actions during the sprint. They work together, gather information, run tests, and put new policies into place with the goal of lowering emissions. Hold daily standup meetings to establish team alignment, update everyone on progress, and identify and address any barriers. Conduct a review at the conclusion of each sprint to evaluate the outcomes and determine whether the targets for emissions reduction were accomplished. Environmental impact studies, data analysis, and stakeholder feedback may all be a part of this review.

Figure : Sprint project dates

Sprint 1 Analysis

The team sought to comprehend the project's problem statement and goal. The analysis shows that the organization's use of fossil fuels for power generation, heating, and cooling is responsible for a sizable amount of its greenhouse gas emissions. Diesel fuel and natural gas are included in this. A significant amount of the carbon footprint is caused by the consumption of electricity, which shows a reliance on fossil fuel-based power sources. Because company cars and employee commutes consume petrol and diesel, the transportation industry has a considerable impact on emissions. A source of concern is recognized as emissions from industrial and heating furnaces. Alternative forms of transportation should be taken into consideration because employee commuting, particularly when done in personal vehicles, increases emissions.

We prepared the groundwork for a methodical strategy to address emissions from multiple sources. It entails gathering data and establishing baselines, taking early action to gain quick wins, and engaging stakeholders to ensure a cooperative effort.

Sprint 2 Analyis

Gathered pertinent emissions-related information from George Mason University (GMU), which may include old documents, utility bills, travel logs, and fertilizer data. The acquired dataset was cleaned, arranged, and organized for analysis. This could entail missing value filling, format conversion, and data cleansing. Preprocessing lowers the possibility of errors and inaccuracies while ensuring that the data is prepared for analysis. spotted and fixed any mistakes or inconsistencies in the dataset, like missing numbers or outliers. Data was transformed into a standardized format for analysis and integration with the data infrastructure of the organization. The second sprint establishes the groundwork for succeeding sprints' data-driven decision-making. For identifying emission patterns, establishing reduction goals, and assessing the effectiveness of emission reduction activities, accurate and well-structured data is crucial.

Sprint 3 Analysis

Evaluated the data flow from data sources to data processing and analysis. Identified data sources, data collection methods, and data processing techniques. Ensured data integrity and consistency throughout the flow. Assessed data storage and retrieval mechanisms. Reviewed the architecture of your system for managing greenhouse gas emissions data. Considered scalability, modularity, and flexibility to accommodate future data. Assessed the efficiency of data processing and storage components. Evaluated the integration of various system modules. Examined the security measures in place to protect sensitive greenhouse gas emissions data. Verified access controls, encryption, and data privacy. Ensured compliance with relevant data protection regulations. Analyzed the algorithms used for processing and analyzing emissions data. Considered the suitability of the algorithms for specific data patterns. Identified areas for algorithm optimization or improvement. Explored opportunities for incorporating additional ML techniques. Considered ensemble methods or hybrid models for better performance.

Sprint 4 Analysis

In Sprint 4, the analysis focused on evaluating the accuracy of the developed model and performing validation to ensure its reliability. The first step involved assessing the model's accuracy, which is a critical measure of its predictive performance. Metrics such as accuracy, precision, recall, and F1 score may have been calculated to quantify the model's performance on the validation set. Additionally, techniques like confusion matrices may have been employed to gain a deeper understanding of the model's ability to correctly classify instances.

Following the model evaluation, machine learning techniques were applied to train the model. This includes using algorithms such as K-Nearest Neighbors (KNN), which was mentioned earlier in the context of predicting greenhouse gas (GHG) emissions. The training process involves feeding the model with labeled data, allowing it to learn patterns and relationships within the dataset.

Visualizations played a crucial role in this analysis. Various visualizations may have been created to provide a comprehensive view of the model's behavior and predictions. For instance, visualizations could include graphs or charts depicting the relationships between input features and predicted outcomes. Additionally, user input values were utilized to generate visualizations that showcase how the model responds to specific scenarios, helping stakeholders better understand its functionality and decision-making process.

Furthermore, a hierarchical breakdown of all the scopes was performed. This likely involved breaking down the analysis into different scopes or categories, potentially based on the hierarchy of factors influencing GHG emissions. This breakdown aids in understanding the contributions of various factors to the overall emissions, providing a detailed and nuanced perspective.

Overall, Sprint 4 involved a comprehensive analysis that not only evaluated the model's accuracy and validated its performance but also utilized machine learning techniques, visualizations, and hierarchical breakdowns to provide a thorough understanding of the model's behavior and its implications for GHG emissions prediction.

Sprint 5 Analysis

Sprint 5 extended the examination of greenhouse gas (GHG) emissions by concentrating on improving the model and gaining deeper insights into specific facets. Drawing from the insights garnered in Sprint 4, adjustments and refinements were likely implemented to boost the model's accuracy and predictive capabilities. Similar to the preceding sprint, validation techniques were probably employed to ensure the dependability of the refined model. Building on the fundamental machine learning techniques introduced in Sprint 4, Sprint 5 may have delved into more advanced methods. This could involve experimenting with ensemble methods, deep learning architectures, or other sophisticated algorithms to further elevate the model's predictive accuracy. The role of visualizations remained pivotal in Sprint 5, serving to depict the refined model's behavior and how it responds to various input scenarios. Graphs, charts, and other visual aids were utilized to convey intricate relationships between input features and predicted outcomes, offering stakeholders a clearer understanding. Employing user input values, akin to Sprint 4, likely facilitated the creation of visualizations portraying how the refined model reacts to specific scenarios, aiding stakeholders in anticipating its behavior in diverse real-world situations.

Reference

Works Cited

[1] “Greenhouse gas inventory summary report FY2019,” University Sustainability, https://green.gmu.edu/ghg-inventory/ (accessed Sep. 12, 2023).

[2] “Greenhouse gas inventory summary report FY2019,” University Sustainability, https://green.gmu.edu/ghg-inventory/ (accessed Sep. 12, 2023).

[3] R. K. Pachauri et al., “Climate change 2014 synthesis report. Contribution of Working Groups I, II, and III to the Fifth Assessment Report of the intergovernmental panel on climate change,” the UWA Profiles and Research Repository, https://research-repository.uwa.edu.au/en/publications/climate-change-2014-synthesis-report-contribution-of-working-grou (accessed Sep. 12, 2023).

[4] “Environmental science and policy (EVPP),” Environmental Science and Policy (EVPP) < George Mason University, https://catalog.gmu.edu/courses/evpp/ (accessed Sep. 12, 2023).

[5] R. K. Pachauri et al., “Climate change 2014 synthesis report. Contribution of Working Groups I, II, and III to the Fifth Assessment Report of the intergovernmental panel on climate change,” the UWA Profiles and Research Repository, https://research-repository.uwa.edu.au/en/publications/climate-change-2014-synthesis-report-contribution-of-working-grou (accessed Sep. 12, 2023).

[6] A. Rogerson, “Bulk Fuel Transport | Foster Fuels,” *Foster Fuels Inc.* <https://fosterfuels.com/products-services/transportation-bulk-fuel/?msclkid=c977c287abc719247acfa78e12e85c1a>

[7] “Fleet Fuel Usage| Element,” *Element Fleet Management*. <https://www.elementfleet.com/resources/blog/how-to-manage-fleet-fuel-usage-most-efficiently>

(accessed Sep. 19, 2023).

[8] IPCC, 2022: Climate Change 2022: Mitigation of Climate Change. Contribution of Working Group III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change [P.R. Shukla, J. Skea, R. Slade, A. Al Khourdajie, R. van Diemen, D. McCollum, M. Pathak, S. Some, P. Vyas, R. Fradera, M. Belkacemi, A. Hasija, G. Lisboa, S. Luz, J. Malley, (eds.)]. Cambridge University Press, Cambridge, UK and New York, NY, USA. doi: 10.1017/9781009157926

**This page intentionally left blank**

**This page intentionally left blank**