ACKNOWLEDGEMENT

I would like to express my gratitude to Prof. Saritha, Department of Computer Science and Engineering, PES University, for her continuous guidance, assistance, and encouragement throughout the development of this UE20CS390A - Capstone Project Phase – 1.

I am grateful to the Capstone Project Coordinators, Dr.Sarasvathi V, Professor and Dr. Sudeepa Roy Dey, Associate Professor, for organizing, managing, and helping with the entire process.

I take this opportunity to thank Dr. Sandesh B J, Professor & Chairperson, Department of Computer Science and Engineering, PES University, for all the knowledge and support I have received from the department. I would like to thank Dr. B.K. Keshavan, Dean of Faculty, PES University for his help.

I am deeply grateful to Dr. M. R. Doreswamy, Chancellor, PES University, Prof. Jawahar Doreswamy, Pro Chancellor – PES University, Dr. Suryaprasad J, Vice-Chancellor, PES University for providing to me various opportunities and enlightenment every step of the way. Finally, this Capstone Project could not have been completed without the continual support and encouragement I have received from my family and friends.

ABSTRACT

Unorganized Agri Practices can be seen leading to the emission of greenhouse gases. Unpredictable weather patterns, droughts and extreme weather events are the primary effects of GHG emission and Global Warming. Main objective comes with people usually focus on all the other Factors affecting the Emission of GHG. But, the Agricultural sector remains Unchecked. Agricultural practices should be practiced in the way that controls the emission of GHG, changes in climate, disease outbreaks, nutritional deficiency in crops and inefficient resource usage. The Proper Management and change in the Agricultural Activities can bring down the Emission of GHG. Projection of GHG emission will guide the government and policy makers to plan accordingly.

TABLE OF CONTENTS

TABLE OF CONTENTS

LIST OF FIGURES

1. INTRODUCTION

Human activities such as burning fossil fuels, deforestation, and agriculture result in the discharge of various chemical gases like carbon dioxide, methane, and nitrous oxide into the atmosphere, which are referred to as greenhouse gas emissions. These emissions have the potential to trap heat in the Earth's atmosphere, leading to the well-known global warming phenomenon.

Unorganized Agri Practices can be seen leading to the emission of greenhouse gases. Unpredictable weather patterns, droughts and extreme weather events are the primary effects of GHG emission and Global Warming. Main objective comes with people usually focus on all the other Factors affecting the Emission of GHG. But, the Agricultural sector remains Unchecked. Agricultural practices should be practiced in the way that controls the emission of GHG, changes in climate, disease outbreaks, nutritional deficiency in crops and inefficient resource usage. The Proper Management and change in the Agricultural Activities can Bring down the Emission of GHG. Projection of greenhouse gas emission will guide the policy makers and the government to plan accordingly.

Analyzing and predicting the control of greenhouse gases (GHG) emission from agricultural activities is a critical area of research in sustainable agriculture. Agricultural activities are one of the major contributors to global GHG emissions, with livestock production, rice cultivation, and synthetic fertilizers being the primary sources of emissions. As such, there is a need for effective approaches to minimize the environmental impact of agricultural activities, especially with respect to GHG emissions.

To tackle the problem of greenhouse gas emissions, one possible solution is to employ Machine Learning (ML) models. ML is a branch of Artificial Intelligence (AI) that centers on the creation of algorithms and models capable of learning from data and making predictions. The utilization of ML models has become more widespread across different areas, including agriculture, with the aim of forecasting and improving outcomes.

Some examples of ML models used in the analysis and prediction of GHG emissions from agricultural activities include decision trees, classification models, regression models and stacking ensemble models. These models can be trained on datasets containing information on farming practices, environmental variables, and GHG emissions to make accurate predictions and provide insights into the drivers of GHG emissions.

Expected outcomes this project, predicting the emissions of Greenhouse gases from the agricultural sector will include the development of accurate predictive models, identification of key factors influencing Greenhouse gas emissions in different agricultural practices, and the assessment of the effectiveness of potential mitigation strategies. These outcomes will help to inform policy decisions and promote sustainable agricultural practices, leading to reduced emissions and improved environmental outcomes.

This project's significance is rooted in the urgent need to decrease GHG emissions as a means of mitigating the adverse effects of climate change, which poses a substantial challenge. Given that the agricultural industry is a major contributor to GHG emissions, it plays a vital role in curbing these emissions. By precisely predicting and comprehending GHG emissions in agriculture, policymakers and farmers can make informed choices on how to decrease emissions, implement effective mitigation approaches, and encourage sustainable practices. As a result, the findings of this project, which involve predicting GHG emissions in agriculture, are of considerable consequence for climate change policy and agricultural operations.

MOTIVATION

There are several potential motivations for a project analyzing the control of GreenHouse Gas (GHG) emissions in agriculture using ARMA and other Machine Learning (ML) models.

Firstly, Agriculture is a significant source of GHG emissions, with estimates suggesting it accounts for around a quarter of global emissions. As such, reducing emissions from the agricultural sector is critical to achieving global emissions reduction targets and mitigating the impacts of Climate change.

Secondly, there is a growing body of research suggesting that ML models can be effective tools for predicting GHG emissions from agricultural activities. ARMA models, in particular, have been used to model Time series data related to the emissions of Greenhouse gases in agriculture.

Thirdly, understanding the factors that influence GHG emissions in agriculture and developing effective strategies for reducing emissions can have significant economic benefits for farmers and other stakeholders. For example, reducing fertilizer use or implementing more efficient irrigation practices can not only reduce emissions but also lower costs for farmers.

Overall, a project analyzing the control of GHG emissions in agriculture using ARMA and other ML models has the potential to make a significant contribution to efforts to mitigate climate change while also providing practical benefits for farmers and other stakeholders in the agricultural sector.

2. PROBLEM DEFINITION

The goal of this project is to analyze and predict the control of greenhouse gas (GHG) emissions from agricultural activities using machine learning models. Agricultural activities, including livestock production and crop cultivation, are a significant source of GHG emissions, which contribute to climate change. Thus, it is crucial to develop effective strategies to reduce these emissions and mitigate their impact on the environment.

This project aims to leverage machine learning models to analyze large datasets of agricultural activities and GHG emissions data to identify patterns, correlations, and potential factors influencing GHG emissions. The ultimate goal is to develop predictive models that can help policymakers and stakeholders make informed decisions and implement effective strategies to reduce GHG emissions from agricultural activities.

3. LITERATURE SURVEY

3.1 PROVABLY CORRECT PEEPHOLE OPTIMIZATIONS WITH ALIVE

3.1.1 INTRODUCTION

Greenhouse gas (GHG) emissions from agricultural activities have become a significant contributor to global warming, and efforts to mitigate them have gained much attention in recent years. In this literature survey, we explore the different applications of ML models in analyzing and predicting the control of GHG emissions from agricultural activities. We begin by discussing the challenges associated with GHG emissions in the agricultural sector and their potential environmental impact. Then, we review existing studies that have applied ML models to predict and analyze GHG emissions from all the different agricultural activities. Finally, we discuss the limitations of the existing studies and propose future research directions to address these limitations.

3.1.2 CHARACTERISTICS AND IMPLEMENTATION

Characteristics:

Interdisciplinary: Requires knowledge from both the agricultural and computer science fields to build project on such topics based on GHG emissions using ML models.

Data-driven: Machine learning models require large amount of data with loaded features to train and make accurate predictions. Therefore, this literature survey focuses on studies that use comprehensive data sets.

Comparative: As there are multiple machine learning models available, we are comparing and contrasting different models and their effectiveness in predicting GHG emissions from agricultural activities using decision trees, classification models, regression models & ensemble models.

Innovative: Machine learning models are constantly evolving, so the most important part is that finding the techniques used like LSTM which deals with problem of gradient descent, K-means, bagging and boosting etc.

Implementation:

Scope & Objectives: The main scope and objectives of this project involve utilizing an intelligent, data-driven approach based on time-series prediction techniques, specifically the ARIMA model, to forecast GHG emissions. Understanding the key drivers of GHG emissions is crucial for policymakers and governments to combat them effectively, and forecasting emissions over the next decade can inform policy decisions. The ARIMA (0, 2, 1) model was used to forecast GHG emissions, revealing an upward trend.

Non-parametric models like KNN and SVR were used, with the number of parameters depending on the volume of training data, to further increase the predictions' accuracy. Boosting methods, which improve the regressor's performance and lower error rates, were also applied. The AdaBoost method, specifically, was used.

Because the dataset used for this project contains time-series data, a statistical model called SARIMA that takes seasonality into account was also taken into account.

Establishing a two layer ensemble model for predicting fossil energy consumption based on stacking ensemble learning.

K-means can be used to group the sectors according to their GHG emissions with the regions of the country to assess ability to identify and map similarities and differences in agricultural emission profiles, driving hyper local GHG emission reduction strategies.

3.1.3 FEATURES

Types of GHG Emissions from Agricultural Activities: Agricultural activities generate different types of Greenhouse gas (GHG) emissions, including: Nitrous Oxide (N2O), Carbon Dioxide (CO2), Methane (CH4), Direct Emissions etc.

Impact of Agricultural Activities on GHG Emissions: Many agricultural activities come into picture which affect GHG emissions like Land Use Change(LUC), Livestock Production, Use of Nitrogen Fertilizers, Energy Use, Crop Burning etc.

Machine Learning Models: There are several Machine Learning models to adopt to get a best prediction performance i.e decision trees – can be used for classification between the models and choose the efficient one. Classification Models – which predicts the class labels/categories based on input values given, Regression Models – used for prediction of continuous values in the given dataset and Ensemble Models to merge the two best accurate models to get best performance.

3.1.4 EVALUATION

1. Cross-validation: To assess the accuracy and generalizability of many machine learning models, validation is necessary. A common technique for this is cross-validation, in which the dataset is split into subsets for the training set and the test set, and the model is trained on the training set and assessed on the test set.

2. Root Mean Square Error (RMSE): The average of the discrepancies between the predicted values and the actual values is measured using the commonly used assessment metric for regression models, known as the RMSE.
3. F1-score: The F1-score, which calculates the harmonic mean of precision and recall, is a frequently used evaluation metric for classification models.
4. Model Evaluation: The software programme will be able to judge how well the machine learning models are working. The evaluation criteria are F1 score, recall, accuracy, and precision.
5. Mean Absolute Error (MAE): Another often employed metric for assessing regression models is the MAE, which calculates the mean absolute difference between the predicted and actual values.
6. Mean Squared Error (MSE): Another often employed metric for measuring the effectiveness of regression models is the MSE, which computes the average of the squared differences between the anticipated and actual values.

4. DATA

4.1 OVERVIEW

A database called FAOSTAT that gathers statistics on many facets of agriculture and food systems is maintained by the Food and Agriculture Organisation (FAO) of the United Nations. The Intergovernmental Panel on Climate Change (IPCC) has designated the FAOSTAT domain Emissions Totals as a comprehensive collection of data on greenhouse gas (GHG) emissions from agrifood systems as well as emissions from other economic sectors. The Tier 1 procedures of the IPCC Guidelines for National Greenhouse Gas Inventories are used to gather and process the data.

The domain covers emissions of Methane gas (CH4), Nitrous Oxide (N2O), Carbon Dioxide (CO2), and Fluorinated gases (F-gases) used in industrial processes. The emissions are categorized based on the activities that generate them, including agriculture, forestry, fisheries, and land use changes. The data is available by country, with global coverage for the period 1961-2020 and projections for 2030 and 2050 for some categories of emissions, while others cover the period 1990-2020.

The database is updated annually to ensure that the information is current and relevant. The data is used by researchers, policymakers, and other stakeholders to monitor and analyze trends in GHG emissions and to develop strategies to mitigate their impact on climate change. The FAOSTAT Emissions Totals domain is an essential tool for understanding the role of agrifood systems and other economic sectors in contributing to global GHG emissions, and for informing policy decisions aimed at reducing those emissions.

4.2 DATASET

Dataset with the attributes and their values :

	ï»¿Domain Code	Domain	Area Code (M49)	Area	Element Code	Element	Item Code	Item	Year Code	Year	Source Code	Source	Unit	Value	Flag	Flag Description
0	GT	Emissions Totals	356	India	7234	Direct emissions (N2O)	5064	Crop Residues	1961	1961	3050	FAO TIER 1	kilotonnes	27.8120	E	Estimated value
1	GT	Emissions Totals	356	India	7236	Indirect emissions (N2O)	5064	Crop Residues	1961	1961	3050	FAO TIER 1	kilotonnes	6.2577	E	Estimated value
2	GT	Emissions Totals	356	India	7230	Emissions (N2O)	5064	Crop Residues	1961	1961	3050	FAO TIER 1	kilotonnes	34.0697	E	Estimated value
3	GT	Emissions Totals	356	India	724313	Emissions (CO2eq) from N2O (AR5)	5064	Crop Residues	1961	1961	3050	FAO TIER 1	kilotonnes	9028.4589	E	Estimated value
4	GT	Emissions Totals	356	India	723113	Emissions (CO2eq) (AR5)	5064	Crop Residues	1961	1961	3050	FAO TIER 1	kilotonnes	9028.4589	E	Estimated value

Fig – 1, Overview of the Dataset

The dataset from Food and Agriculture Organization provides information on greenhouse gas (GHG) emissions from India for the period 1961 to 2020. The emissions are classified by different area codes, which could refer to regions or states within India.

The dataset includes emissions of three major GHG gases - Nitrogen Dioxide (N2O), Methane gas (CH4), and carbon Dioxide (CO2), as well as emissions of other GHG gases. These gases are known to contribute significantly to climate change and global warming.

The emissions are measured in kilo tons of CO2 equivalent (CO2e), which is a standard measure used to compare emissions of different Greenhouse gases. This enables comparison of emissions from different sources based on their contribution to climate change.

The dataset provides a comprehensive record of GHG emissions from India over several decades, enabling analysis of trends and patterns in emissions over time. This information is crucial for policymakers and researchers to develop effective strategies for reducing GHG emissions and mitigating the impacts of climate change.

5. SYSTEM REQUIREMENTS SPECIFICATION

5.1 INTRODUCTION

The document will guarantee that the system is created to fulfil the unique needs of the stakeholders and that it lives up to their expectations. The SRS will also serve as a platform for interaction and coordination between the project team and stakeholders.

The SRS will outline the system's functional and non-functional requirements, such as its capacity to preprocess and analyse sizable datasets, offer interactive data visualisations, train and assess various machine learning models, and produce predictions and recommendations based on the analysis. The SRS will also include the user needs, including the type of user interface needed and any user training or documentation requirements, as well as the hardware and software requirements necessary to operate the system effectively.

5.2 FUNCTIONAL REQUIREMENTS

A simple User Interface for the user to give their queries describing the symptoms they are experiencing in simple language. Prediction of the GHG emission based on the past data values. Predicting for Achieving NetZero Emission of GHG.

The interface of Analyzing and Predicting the control of GHG Emissions from Agricultural Activities using Machine Learning Models should have clear and consistent screen formats with GUI standards for styles, providing a user-friendly experience. Simple navigation and standard features like help should be included on the screen layout. An appropriate relative timing for the user's workflow should be used for inputs and outputs. The presence of a programmable function key of some kind could improve user efficiency and experience. In order to help the user fix the problem, error messages

should be understandable and offer useful feedback. The interface should also be able to present large amounts of data in an orderly and understandable format, enabling the effective.

5.2.1 INTERFACES

A simple interface for the user interaction to take the inputs from the user and give their prediction results and the graphical outputs. An Interface for Predictive system that uses Machine learning models and Displays the Prediction of the Emission rates for the user. An Interface for Analytical model to analyze and display the results for achieving the NetZero emission of GHG.

5.2.2 HARDWARE REQUIREMENTS

- Pentium IV and above
- RAM 512 mb and above
- Minimum 20 Gb of Disk space

5.2.3 SOFTWARE REQUIREMENTS

- Windows, Linux, MacOS or any other OS
- Any Web browser like, Google Chrome, Safari, etc.
- Libraries like Pandas, Numpy, Matplot library etc.
- Other ML libraries like Natural Language toolkit, scikit – learn etc.
- Code editor and IDE's to build the code like Jupyter Notebook, V S code and others.

5.2.4 COMMUNICATION INTERFACE

HDMI Cable : for Connecting to External Display

Ethernet Cable : If the Project will be connected to cloud, and needs internet access.

5.3 NON – FUNCTIONAL REQUIREMENTS

5.3.1 PERFORMANCE REQUIREMENTS

Accuracy: The machine learning models must be highly accurate in predicting GHG emissions from agricultural activities. The accuracy of the models must be tested against actual data to ensure that the models provide accurate predictions.

Reliability: The system must be reliable and consistent in its performance. The machine learning models must be tested under various scenarios to ensure that they are reliable and produce consistent results.

Robustness: The system must be robust enough to handle large amounts of data and various input parameters. The machine learning models must be able to handle changes in input data and still provide accurate predictions

5.3.2 SAFETY REQUIREMENTS

Ethical Considerations: The machine learning models will be developed and used in an ethical manner. The models will not be used to discriminate against individuals or groups based on race, gender, age, or other protected characteristics.

5.4 ETHICAL CONSIDERATIONS:

In the development and use of machine learning models, it is important to consider ethical considerations to ensure that the models are used in an ethical manner. Discrimination against individuals or groups based on race, gender, age, or other protected characteristics is not acceptable and should be avoided.

In this context, it is important to ensure that the machine learning models are designed and trained in a way that is fair and unbiased. This can be achieved by ensuring that the training data is diverse and representative of the population being studied, and by using appropriate algorithms and techniques to mitigate any biases that may be present in the data.

In addition, it is important to ensure that the machine learning models are transparent and explainable, so that the decisions made by the models can be easily understood and justified. This can be achieved by using techniques such as decision trees or rule-based systems that provide clear and understandable explanations for the decisions made by the models.

Finally, it is important to ensure that the machine learning models are used in a way that respects the privacy and autonomy of individuals. This can be achieved by ensuring that the data used to train the models is collected and stored in a secure and ethical manner, and by providing individuals with the option to opt-out of data collection and analysis if they so choose.

In summary, ethical considerations are an important safety requirement for the development and use of machine learning models. It is important to ensure that the models are designed and trained in a fair and unbiased manner, that they are transparent and explainable, and that they are used in a way that respects the privacy and autonomy of individuals. By considering these ethical considerations, we can ensure that the machine learning models are used in a way that benefits society as a whole.

6. SYSTEM DESIGN

The above diagram represents a master class diagram for a system that analyzes and predicts GHG emissions from agricultural activities using machine learning models. The system consists of five major components: Data Collection and Preprocessing, Machine Learning Models, Model Evaluation and Selection, GHG Emissions Analysis and Prediction, and User Interface. Each component is represented as a class in the diagram.

6.1 MASTER – CLASS DIAGRAM

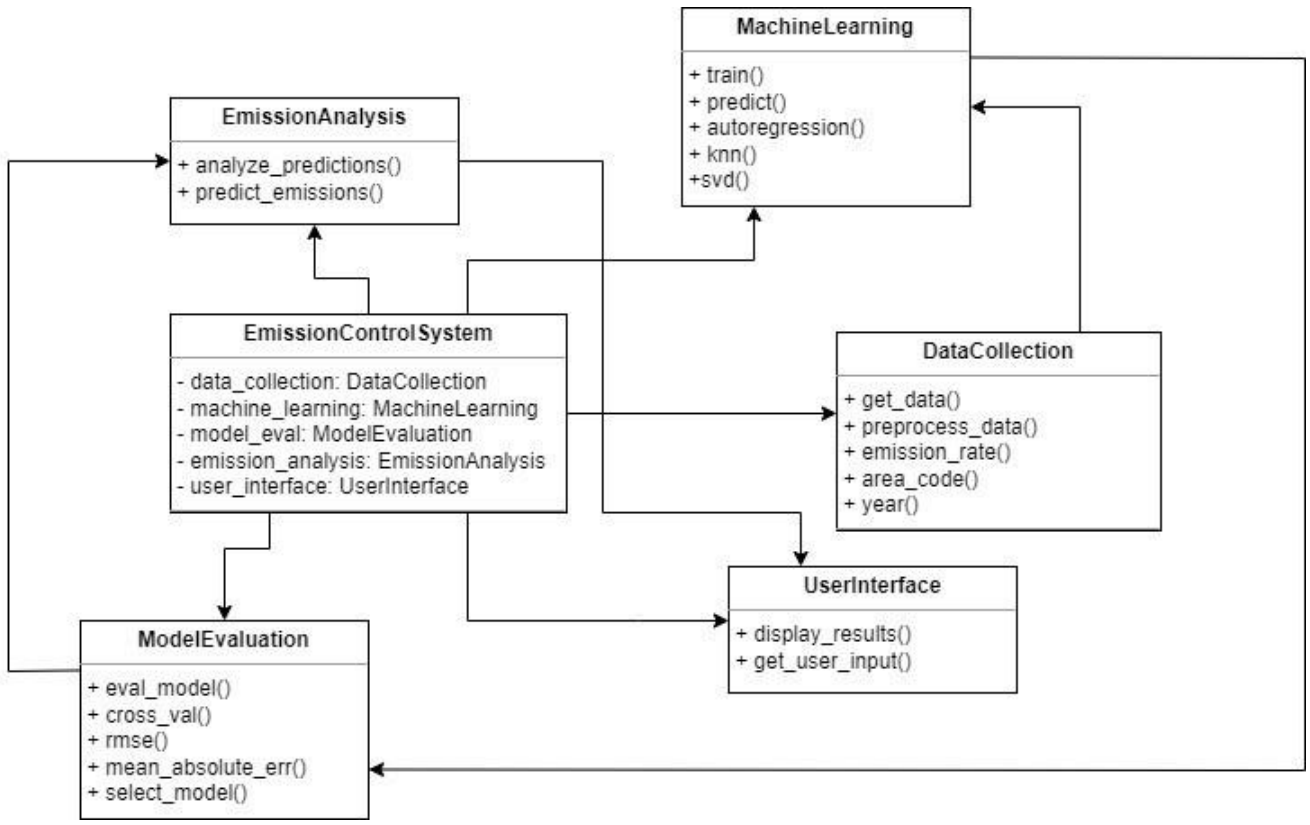

Fig – 2, Master Class Diagram

The above diagram represents a master class diagram for a system that analyzes and predicts GHG emissions from agricultural activities using machine learning models.
The system consists of five major components: Data Collection and Preprocessing, Machine Learning Models, Model Evaluation and Selection, GHG Emissions Analysis and Prediction, and User Interface. Each component is represented as a class in the diagram.

The Emission Control System class is the main class that aggregates all the other classes. It contains references to each of the other four major components, i.e., Data Collection and Preprocessing, Machine Learning Models, Model Evaluation and Selection, GHG Emissions Analysis and Prediction, and User Interface.
The Data Collection and Preprocessing class is responsible for collecting the necessary data for analysis and preprocessing it to make it suitable for machine learning. It contains two methods: get_data(), emission_rate(), area_code() and preprocess_data()

The Machine Learning Models class is responsible for training machine learning models and making predictions. It contains two methods: train(), auto_regression(), svd(), knn() and predict().

The Model Evaluation and Selection class is responsible for evaluating and selecting the best machine learning model. It contains two methods: evaluate_model(), cross_val(), rmse(), mean_absolute_err() and select_model().

The GHG Emissions Analysis and Prediction class is responsible for analyzing and predicting GHG emissions from agricultural activities using the selected machine learning model. It contains two methods: analyze_emissions() and predict_emissions().

Finally, the User Interface class is responsible for displaying the results and collecting user input. It contains two methods: display_results() and get_user_input().Overall, this master class diagram provides a high-level view of the system's major components and their relationships, which can be used as a starting point for further analysis and design.

6.2 SYSTEM ARCHITECTURE DIAGRAM

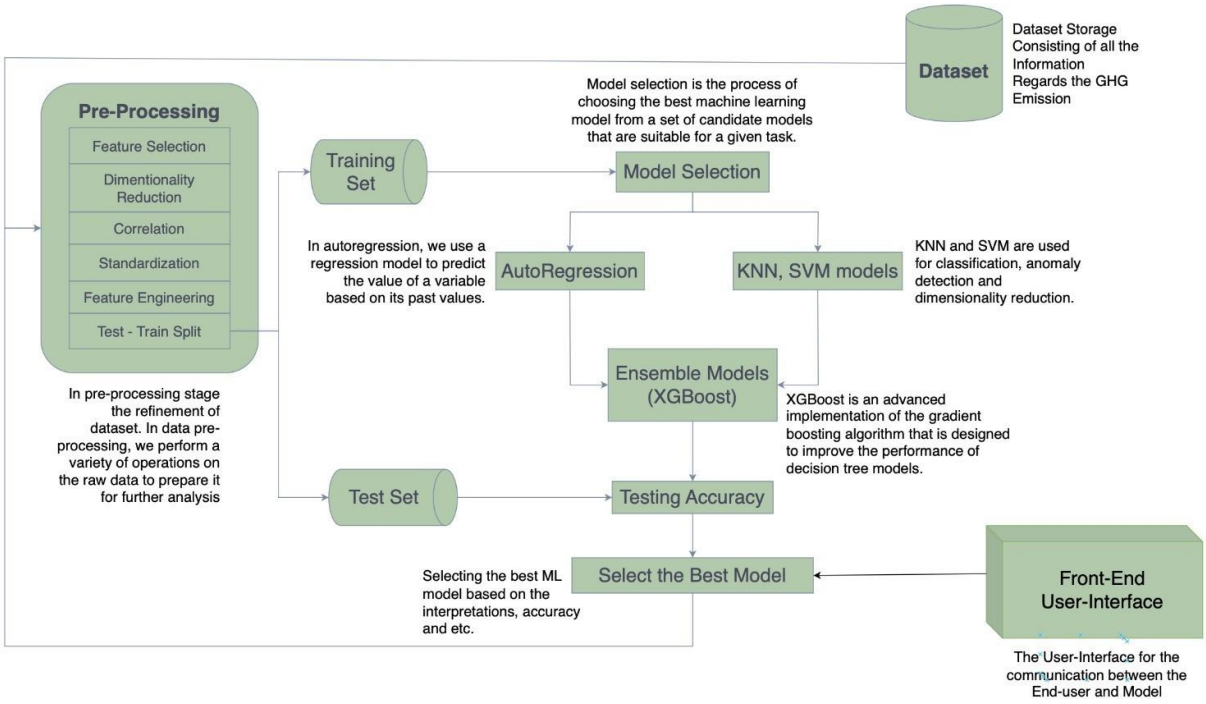

Fig – 3, System Architecture Diagram

The architecture diagram explains about all the modules and phases built in the project and their flow of working to generate predictive results. It starts with the data loading, and pre-processing. Here the Data will be loaded into the model and few cleaning operations will be done. This may include cleaning the data values, removing missing values, and transforming it into a usable format.

The next step is to select the most relevant features or variables that contribute to GHG emissions. This may involve using statistical methods or domain knowledge to identify the most important factors. Correlation matrix is used to find the correlation between attributes, to understand the dependency of target attribute with other attributes. The selected features are used to train machine learning models,

such as regression or decision trees, to predict GHG emissions. The models are trained on a subset of the data known as the training set. Once the models have been trained, they are evaluated to determine

how well they perform. This is done using a separate subset of the data, known as the test set, to test the model's accuracy and identify any potential issues. The models may be further improved by

engineering new features or variables that better capture the relationships between agricultural activities and GHG emissions. Ensemble models, such as random forests or boosting algorithms, can be used to combine the predictions of multiple machine learning models and improve overall accuracy.

Once the models have been trained and evaluated, they can be deployed in a production environment. This may involve integrating the models into an existing software system or creating a new application that uses the models to predict GHG emissions from agricultural activities. It is essential to monitor the models' performance regularly to ensure they continue to provide accurate predictions. This may involve monitoring the input data for any changes or monitoring the output of the models to identify any unexpected results.

6.3 DATA FLOW DIAGRAM

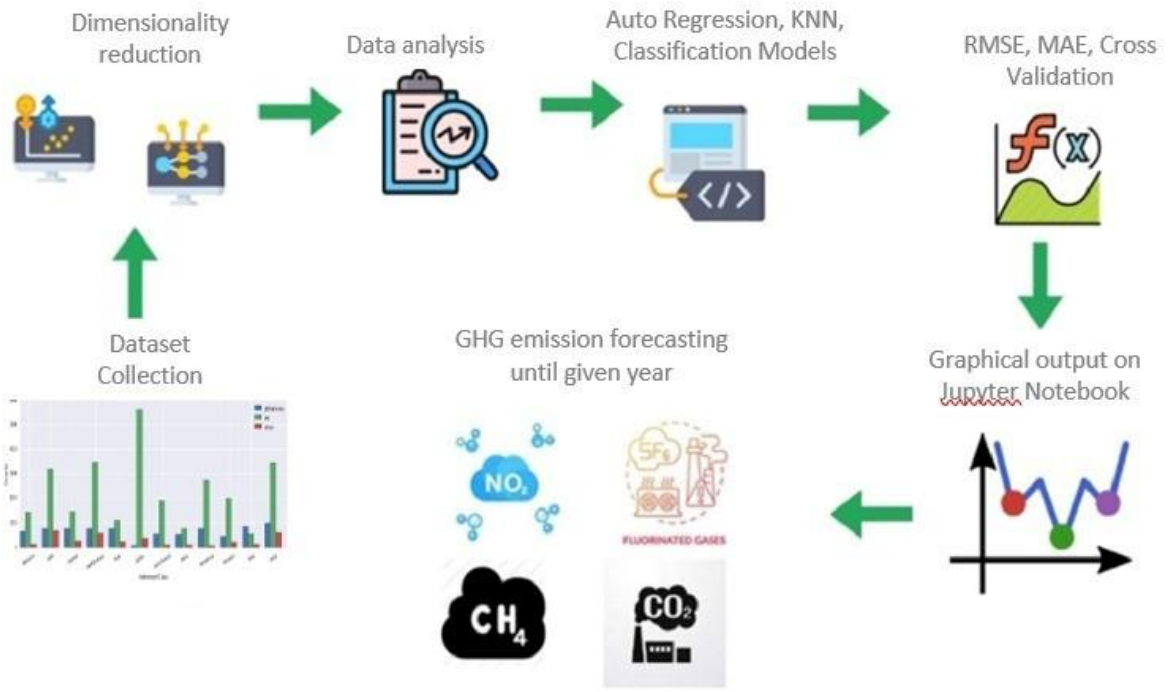

Fig – 4, Data flow Diagram

The Above Diagram explains the data flow throughout the model. This is a representation of data flowing procedure in the predictive model to obtain the emission graph results. Dataset is the main source of collection, through which the data values will be passed into the model. To make understand the model with all the data values correctly, the dataset needs to be pre-processed so removing all the garbage values and cleaning the dataset. This will also helps to increase the accuracy and precision of the model. Then the data is going through many operations to make it perfect for the Analysis. Then the data is analysed using many visualization tools and techniques. This helps to understand the dataset characteristics and other features that helps to decide which machine learning models will be the right fit for that dataset.

Once the dataset is completely analyzed, now it is ready to train the model, a part of dataset is separated to test the model, and is known as test dataset. While training the model multiple best machine learning model will be used to train with the training set. Few of the best selected model for this project are AR, ARMA, ARIMA, KNN, XGBoost. Then some of the best performing machine learning models will be merged together to get even more accurate results and achieve higher level of perfection.

These best models will be selected through accuracy testing which will calculate the accuracy results for all these models. Few of the accuracy tests that are commonly used includes MAE, RMSE, RAE etc. Once the models are completely built with checking accuracy, now the model is ready to test with external end-user with providing their own parameters as input values, which is also known as last step of models is Deployment and Meintenance.

6.4 USE CASE DIAGRAM

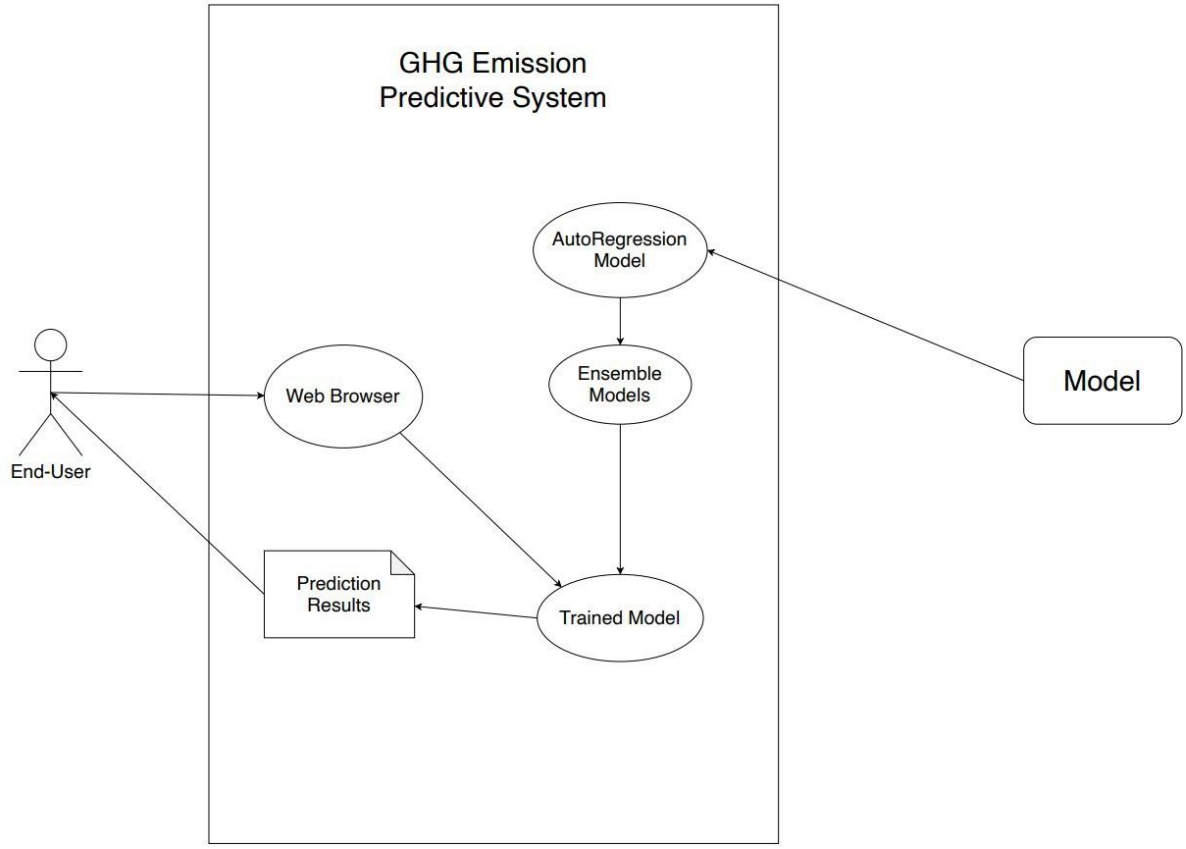

Fig – 5, Use – Case Diagram

The above Use case diagram explains the behavioural interaction of machine learning model with the end-user. This use case diagram has 2 actors mainly, the model and the end-user. It shows the way of communication between both the actors, model and the end-user. In this case, the model processes the

dataset loaded, and mainly performs regression models and classification models, then by using ensemble models selects best based on their accuracy tests. Once the model is trained on the training data, it is ready to test the model using test data, For the communication between the end-user and the model an interface with webpages will be available, that helps the end-user to give the input parameters and to display the output results received from the model. The graphical models are used to display the graphical predictions of the future emission rates and changes in emissions of different greenhouse gases and the other pattern after receiving some required parameters from the end-user helps to predict the chances for controlling the emission rates of greenhouse gases.

This model is mainly built with the interface that helps for the easier communication and understanding the importance of emission control of greenhouse gases. As easier it is understood by the farmers and the policy makers, one will come up with the optimal solutions to control the emission rates.

7. IMPLEMENTATION AND PSEUDOCODE

Model Selection process will start once the features are engineered, a suitable predictive model needs to be selected. This will involve choosing between regression models, decision trees, neural networks, or other types of models. Few of the selected models for this project include Autoregression model, classification model like KNN, and ensemble models like XGBoost etc. These models will be trained to give the better output results.

In the realm of statistical modeling for predicting future behavior in stationary time series data lies Autoregression (AR). By leveraging historical behavior from past instances and applying self-regression techniques, AR models can successfully extrapolate these results further into the future while accounting for inherent randomness in systems via error terms. Enhancing such precision further comes from utilizing both moving average and autoregressive approaches within our forecasting model – creating what's known as the autoregressive moving-average process or ARMA for short. Data analysis has broad applications across industries because it enables professionals to identify patterns and trends within datasets. With this information they can estimate the impact of past values on present variables and make insightful predictions about future values within a given time series.

7.1 PSEUDO CODE FOR AUTO REGRESSION MODEL.

1. Set the parameters for the model, p (autoregressive order) and q (moving average order).
2. Determine the series' data's mean and standard deviation.
3. By removing the mean, centre the data.
4. Calculate the centred data's autocorrelation function (ACF) and partial auto correlation function (PACF).
5. Use the ACF and PACF charts to determine the AR and MA processes' relative priority.
6. Set the AR and MA coefficients to zero.
7. Use the maximum likelihood estimation approach to fit the ARMA model.
8. Use the Augmented Dickey-Fuller (ADF) test to determine whether or not the residuals are stationary.
9. Change the data and re-fit the ARMA model if the residuals are non-stationary.
10. Make predictions for next time points using the fitted model.
11. Calculate the confidence intervals for the predictions.
12. In case of multivariate time series, perform a VARMA model which extends the ARMA model to the multivariate case.
13. Evaluate the model’s performance using metrices such as Mean Squared Error (MSE) and Mean Absolute Error (MAE).
14. Visualize the results of the model predictions and compare them to the actual data.

15. Refine the model as necessary by adjusting the model parameters or using different techniques such as seasonal ARMA models.

CLASSIFICATION MODEL.

Classification is a supervised learning technique in machine learning that involves predicting the class or category of a given input data point. In classification, the goal is to learn a model that can assign a set of input features to one of several predefined classes.

Classification is typically used when the target variable is categorical or qualitative, such as predicting whether an email is spam or not, classifying images of animals into different species, or determining whether a customer is likely to buy a product or not.

Classifying data begins by segregating it into categorized 'training' and 'test' sets. During this stage's main aim—namely: 'training,' or teaching of sorts—we focus on building a strong link between our input features and relevant output classes via an accrued understanding of their interactions within datasets at-large. Once deciphered successfully, we apply such learnings onto our dataset's other half—the 'test' set—and judge them based on predictive accuracy relative to given outcomes from similar datasets before us.

Classification is integral in a variety of fields today - but what algorithm should be used remains up for debate depending on context. For instance: when faced with limited amounts of training data - neural networks may not perform as well as well-established techniques like logistic regression models or decision trees. Context matters here: an understanding of datasets and available tools allow us ultimately decide on the best methodological approach based on performance requirements such as precision recall trade-offs etc.,

K-NN Algorithm

Say you're working on a complex problem that requires classification or regression analysis - what do you use? Well if accuracy and reliability are important to you (as they should be) we'd recommend considering the k nearest neighbors (k NN) algorithm.

This powerful tool can handle both types of analyses with ease. Essentially when faced with new data to classify or predict against existing datasets this technique will search for its nearest k neighbors in those datasets - using those neighbors labels/values as guidance to come up with an accurate prediction for your new piece of information.

If its regarding a classification problem - pinpointing which category your piece belongs in - then your new piece will be classified based on which label came up most frequently amongst the k nearest neighbors. Meanwhile for regression problems - where you're predicting an outcome's value - the predicted value of your new piece will simply be an average of its k nearest neighbors.

7.2 PSEUDO CODE FOR K – NEAREST NEIGHBOUR MODEL.

K-nearest neighbors (KNN) algorithm involves the following steps:

Step 1: Loading the training and test data is the first step in implementing any algorithm.

Step 2: The value of K, which represents the number of nearest data points, needs to be chosen.

Step 3: When determining classification for points in test data, this algorithm measures their distance from every row of training data with techniques such as Euclidean, Manhattan or Hamming distance. Generally speaking, Euclidean measurements tend to be more commonly used than others. After difference scores are obtained they get ranked from smallest to largest and then K number of top results get chosen for closest matches before it assigns a classification according to majority grouping in those matched rows.

ENSEMBLE MODELS.

Ensemble modeling is a powerful technique used in machine learning that improves prediction accuracy by combining outputs from various individual machine learning algorithms known as "base learners." Not only are ensemble methods applicable to both classification and regression modeling tasks but they're also highly versatile offering several types such as bagging, boosting, stacking among others.

Bagging trains base learners independently on different randomly selected subsets of the training dataset then combines all the predicted outcomes to obtain the final result while boosting trains base learners sequentially where later iterations focus on instances where earlier iterations misclassified data points.

Stacking uses an approach whereby diverse base learners output predictions which are fed into another model called a meta learner that learns how best to combine these disparate predictions into one single output.

Random forests combine two powerful approaches bagging with decision trees where observation labels make decisions based on voting rather than consensus method employed in traditional decision trees.

Although Ensemble modeling requires larger datasets and more computational resources compared to individual machine learning algorithms it remains a valuable tool for solving complex machine learning problems. Target Readability: Academic professionals/Researchers.

7.3 PSEUDO CODE FOR XG BOOST MODEL.

1. Initialize the model parameters, including the learning rate, maximum depth of trees, number of trees, regularization parameters, etc.

2. Split the data into training and validation sets.

3. For each tree in the model:

   a. Compute the gradients and hessians of the loss function for each training instance.

   b. Build a decision tree to fit the negative gradients using the training data.

   c. Prune the tree using regularization to prevent overfitting.

   d. Add the tree to the model.

   e. Evaluate the model on the validation set and record the performance.

   f. If the performance on the validation set stops improving, stop training and use the current model as the final model.

4. Use the final model to make predictions on new data.

XGBoost (Extreme Gradient Boosting) is a highly popular machine learning algorithm deployed in regression as well as classification applications. It belongs to the category of boosting algorithms designed with ensemble learning methodology incorporating several weak models (e.g. decision trees) thereby producing a robust model with high accuracy over other conventional tools available in the market.

Its known for its exceptional performance in terms of speed and flexibility across various domains including winning several Machine Learning competitions worldwide. The procedure involves adding new decision trees iteratively into the existing model correcting errors committed previously using Gradient Descent techniques thereby optimizing loss function minimizing differences between predicted versus actual values.

Note: The above pseudocode is just a high-level overview of the XGBoost algorithm, and there are many details and variations in the implementation of XGBoost that can affect the performance and accuracy of the model.

7.4 PSEUDO CODE FOR DECISION TREE

When it comes to creating models that handle both regression and classification tasks efficiently look no further than decision trees! These algorithms use recursive partitioning of input space based on binary decisions until certain criteria have been met.

In a classification setting, decision trees partition the input space into regions that correspond to different classes, and assign a label to a new input based on which region it falls into. In a regression setting, decision trees partition the input space into regions that correspond to different values of the target variable, and predict a continuous value for a new input based on the region it falls into.

Here's a stepwise process of pseudo code for building decision trees:

1. Define a function that takes as input a dataset and a list of possible features to split on.
2. Within the function, compute the impurity of the dataset using a chosen impurity metric (e.g. Gini index or entropy).
3. Splitting data into subsets doesn't have to be complicated: simply iterate over each feature possibility and calculate their corresponding information gain values.
4. Choose the highest value for your split point; from there create two new subsets of data.
5. Keep iterating through these subsets recursively until you hit your cutoff criteria -- perhaps when reaching a certain depth or minimum number of samples in a leaf node.
6. By following these steps you can build a decision tree thats both effective and efficient.

MEAN ABSOLUTE ERROR (MAE)

Mean Absolute Error (MAE) is one of several methods used within regression analysis time series analysis and machine learning models to evaluate overall predictive performance — measuring the average absolute difference between predicted vs actual values in a given dataset.

To calculate MAE you'll need to find the difference between predicted vs actual values that define each sample in your data set and then take their absolute value — avoiding any potential cancellations caused by both positive or negative errors. The resulting scores are then averaged over all samples using this formula:

The formula for MAE is:

MAE = 1/n * sum(|y_pred - y_actual|)

Lower scores tend to indicate better overall predictive model performance when looking at MAE evaluations but its worth noting that this metric focuses more on the practical accuracy of predictions rather than just their square differences (as with RMSE).

ROOT MEAN SQUARE ERROR (RMSE)

Measuring prediction accuracy is vital in several domains from finance to healthcare to smart city planning. One useful metric used in this regard is Root Mean Squared Error or simply RMSE. A prerequisite for calculating RMSE is obtaining Mean Squared Error or MSE which conveys average squared differences between forecasted and factual figures.

This gets followed by taking square root of MSE that results in computing RMSE conveying standard deviation of residuals or the distance between predictions and actual values. Typically higher RMSE values suggest a greater degree of error in model estimation. The formula for computing RMSE is:

The formula for Root Mean Square Error is:

Rmse = Sqrt (1/n * sum ((y_predicted - y_actual) ^2))

In this equation 'y_predicted' stands for predicted value 'y_actual' represents true value and 'n' denotes total number of samples under analysis.

7.5 MODEL TRAINING

Model training is an important step in the process of developing Machine Learning algorithms. It involves feeding the pre-processed data into the algorithm, which learns patterns and relationships from the data. The goal of this training process is optimizing the algorithm's parameters to make accurate predictions for new data.

One important aspect of model training is the splitting the dataset into training set and test datasets. The test set is used to assess the performance of these machine learning algorithms, while the training set is used to teach them how to make predictions. This lessens the likelihood of overfitting, which occurs when an algorithm memorises the training set rather than discovering patterns that apply to new data.

The performance of the model must be assessed on the validation set after it has been trained. Metrics including accuracy, recall, precision, and F1 score are computed to achieve this. These indicators enable us to assess the model's effectiveness and pinpoint its weak points.

After training the model and evaluation, it is ready for the deployment in a production environment. This involves integrating the model into a software application or web service that can be accessed by end-users. In order to ensure that the model remains accurate over time, it is important to regularly update the dataset to reflect changes in emission levels. This allows the model to continue making accurate predictions even as the environment changes.

It is important to monitor the model's performance over time to ensure its continued accuracy. This includes monitoring metrics such as model drift, which pertains to changes in the data distribution over time that could affect the model's precision. If the model drifts, it may require retraining with updated data or adjustments to align with the evolving environment.

In conclusion, model training is a critical step in developing machine learning algorithms for predicting emissions levels. It Involves the splitting of dataset into training set and test sets, optimizing the algorithm's parameters, and evaluating its performance. Once the model is trained completely, it is ready for the deployment in a production environment and updated over time to ensure accuracy. Regular monitoring is essential to ensure that the model remains accurate and up-to-date with changing environmental conditions. By following these steps, machine learning can provide valuable insights into emissions levels and contribute to a more sustainable future.

8. CONCLUSION OF CAPSTONE PROJECT PHASE – 1

The capstone project phase I has demonstrated the potential of machine learning in monitoring pollution levels, providing essential data to identify sources of pollution and enabling effective remediation measures. The use of machine learning models for pollution monitoring can significantly impact sustainable agriculture practices, benefiting farmers, policymakers, and governments. This research has significant implications for the agricultural sector, as it can help in reducing environmental impacts and promoting sustainability.

The literature review clarifies the general machine learning models to be used for predicting greenhouse gas emissions, such as autoregression, classification models, and ensemble models like XGBoost. It outlines the procedure for training the model, including collection of dataset, pre-processing of data, feature selection process, and model development. The use of machine learning models enables accurate predictions, allowing for early warning systems that could prevent long-term environmental damage.

The findings of this research provides valuable insights onto the potential of precision techniques and machine learning in monitoring pollution levels. By identifying the sources of pollution, remedial measures can be taken, contributing to a cleaner and safer environment. Additionally, these techniques could enable farmers to reduce their environmental impact, helping to promote sustainable agricultural practices.

However, it is important to acknowledge the limitations of this research, and further studies are required to validate the efficiency of precision techniques and machine learning models for pollution monitoring. The effectiveness of these techniques may vary depending on the environmental conditions, and the results may not be applicable in all scenarios. Additionally, it is essential to consider the ethical implications of the data collection process, ensuring that privacy and data protection regulations are respected.

In conclusion, this research provides a solid foundation for future studies and practical applications of precision techniques and machine learning in agriculture and environmental monitoring. By adopting these techniques, farmers can potentially contribute to reducing their environmental impact, paving the way for a more eco-friendly and efficient agricultural sector. Furthermore, the use of machine learning models for pollution monitoring could have significant implications for environmental protection, leading to cleaner and safer environments for future generations.

9. PLAN OF WORK FOR CAPSTONE PROJECT PHASE - 2

The first process in building a predictive model is to conduct further research and analysis on the available data. This step is crucial because the efficiency of the model is largely dependent on the quality of dataset, and input values obtained through feature extraction. Feature extraction is a process that involves refining the dataset further to extract valuable insights. This process determines the appropriate machine learning models and techniques that should be used in the next phase of implementation.

Machine learning models play a critical role in predictive modeling. The selection of the right model for the available data is essential for achieving accurate results. During the initial phase, several models will be evaluated, including autoregression models like ARMA and classification models like KNN algorithms and ensemble models.

The next phase involves training the models using the available data sets. This process involves analyzing the training set of data using the selected models. Based on their learning, the accuracy of the models will be tested using the testing data. Once the accuracy tests have been conducted, the best models will be selected based on their performance.

Ensemble models will be used to combine the best models and obtain more accurate results. These models are designed to improve the overall performance of predictive models by combining multiple models.

The interface is an essential component that bridges the gap between the end-user and the predictive model. The interface plays an important role in helping the end-user understand the prediction results obtained by the model. Factors like visual appeal, ease of use, responsiveness, and other considerations will be taken into account while building the frontend interface.

Once the interface has been developed, the model will be ready for use by the end-users. The overall process of developing a predictive model involves several steps and requires a deep understanding of machine learning techniques, data analysis, and software development skills.

Accuracy testing is an important aspect of predictive modeling that involves measuring the accuracy of these models using the test dataset. The accuracy of a model is a measure of how well it predicts the outcomes based on the testing data set. The testing data set is a subset of the data that is used to evaluate the performance of the model.

Predictive modeling involves various accuracy testing techniques, such as the confusion matrix, receiver operating characteristic (ROC) curve, and precision-recall curve. A confusion matrix displays the number of true positives, true negatives, false positives, and false negatives in a table format. The ROC curve graphically plots the true positive rate against the false positive rate. The precision-recall curve illustrates the precision-versus-recall relationship in a graphical format.

Once the accuracy testing has been conducted, the best models will be selected based on their performance. Ensemble models will be used to combine the best models and obtain more accurate results.

The interface is an essential component that builds the communication between the predictive models and the End-user. The interface plays a role in making the end-user understand the prediction results obtained by the model. While building a frontend interface, various factors like visual appeal, ease of use, responsiveness, and others will be taken into consideration.

In conclusion, developing a predictive model involves several steps that require a deep understanding of machine learning techniques, data analysis, and software development skills. Accuracy testing is a crucial aspect of predictive modeling that involves measuring the accuracy of the model using the test dataset. The interface is an essential component that bridges the gap between the end-user and the predictive model.

REFERENCES/BIBLIOGRAPHY

[1] Forecasting the Emission of Greenhouse Gases from the Waste using SARIMA Model, Vaishnavi Jayaraman, Saravanan Parthasarathy, Arun Raj Lakshminarayanan, Department of Computer Science and Engineering, B.S.Abdur Rahman Crescent Institute of Science and Technology, Chennai, India. 2022 6th International Conference on Trends in Electronics and Informatics (ICOEI) | 978-1-6654-83285/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICOEI53556.2022.9777119

[2] Forecasting of GHG (greenhouse gas) Emission using (ARIMA) Data Driven Intelligent Time Series Predicting Approach, Dr. Somesh Sharma School of Management, Graphic Era Hill University Bhimtal, India. Dr. Ashish Kumar Saxena SOM, IFTM University Moradabad, UP, India. Mr. Manmohan Bansal Faculty of Management, Invertis University Bareilly, UP, India. 2022 7th International Conference on Communication and Electronics Systems (ICCES) | 978-1-6654-9634-6/22/$31.00 ©2022 IEEE | DOI:10.1109/ICCES54183.2022.9835888

[3] Carbon Emissions forecasting based on Stacking Ensemble Learning. Quanmao Zhang, Ying Wang, Dongliang Qin, Kailin Zhao, Wanying Xie, Economic Research Institute of state Grid Hebei Electric power company, Shijiazhuang, China. 2022 IEEE 5th International Electrical and Energy Conference (CIEEC) | 978-1-6654-1104-2/22/$31.00 ©2022 IEEE | DOI: 10.1109/CIEEC54735.2022.9845939.

[4] A k-Means-Based-Approach to Analyze the Emissions of GHG in the Municipalities of MATOPIBA Region, Brazil, Lucas Ferreira-Paiva , Attawan G. L. Suela Cardona-Casas , Domingos S. M. Valente, and Rodolpho V. A. Neves , Member, IEEE, IEEE LATIN AMERICA TRANSACTIONS, VOL. 20, NO. 11, NOVEMBER 2022. References

[5] Performance Monitoring Insight using Predictive Analytics: A Step towards IMO's GHG Emission Goals 2030. Nithish Balaji J, Himanshu Uppal, Pritam Patel and B.M. Shameem. OCEANS 2022 – Chennai | 978- 1-6654-1821-8/22/$31.00 ©2022 IEEE | DOI: 10.1109/OCEANSChennai45887.2022.9775128.

[6] The Assessment of Energy Consumption and Carbon Emission from Maize Production Process in Northern Thailand. Tanate Chaichana, Kunyaporn Chaiwoung, Saritporn Vittayapadung, Ukrit Samaksaman, ICUE 2020 on Energy, Environment, and Climate Change Asian Institute of Technology, Thailand. 20 – 22 October 2020

[7] Assessment of Emissions with Carbon-smart Farming Practices and Participatory Sensing in Rice, Rushikesh Kulat, Mariappan Sakkan, Prachin Jain, Sanat Sarangi, Srinivasu Pappula TCS Research and Innovation, Mumbai, India. 2022 IEEE Global Humanitarian Technology Conference (GHTC) | 978-1- 665450973/22/$31.00©2022IEEE|DOI:10.1109/GHTC55712.2022.9910612

[8] Assessing Impact of Carbon-smart Farming Practices in Rice with Mobile Crowdsensing, Rushikesh Kulat, Mariappan Sakkan, Prachin Jain, Sanat Sarangi, Srinivasu Pappula TCS Research and Innovation,
Mumbai, India. 2022 IEEE Region 10 Symposium (TENSYMP) | 978-1-665466585/22/$31.00©2022IEEE|DOI:10.1109/TENSYMP54529.2022.9864367

APPENDIX A DEFINITIONS, ACRONYMS AND ABBREVIATIONS

ARMA : AutoRegressive Moving Average.

ARIMA: AutoRegressive Moving Average + Trend Differencing.

GHG : Any gas, such as carbon dioxide and methane, that contributes to the greenhouse effect and results in climate change is referred to as a greenhouse gas.

KNN : K - Nearest Neighbor

LSTM : Long - Short Term Memory networks

MAE : Mean Absolute Error

LUC: Land Use Change

RAE : Relative Absolute Error

RMSE : Root Mean Square Error

SVM : Support Vector Machines

XGBoost : Extreme Gradient Boosting