II Trimester MSc (AI & ML)

Advanced Machine Learning

Department of Computer Science

FORECASTING OF AGRICULTURAL CO2 EMISSIONS TO MITIGATE CLIMATE CHANGE

by

SaiDarshan (2348548)
Guhan K S (2348519)
Suhas S (2348563)

January 2024

CERTIFICATE

This is to certify that the report titled ***Forecasting of Agricultural CO2 Emissions to Mitigate Climate Change*** *is a bona fide record of work done by* ***SaiDarshan (2348548) Guhan K S (2348519) Suhas S (234563)*** *of CHRIST (Deemed to be University), Bangalore, in partial fulfilment of the requirements of II Trimester of MSc Artificial Intelligence and Machine Learning during the year 2023-24.*

Course Teacher
Dr. Nizar Banu PK

Valued-by: (Evaluator Name & Signature)
1.

2.

Date of Exam: 25-01-2024

Table of Contents

Team Details

Reg. no	Name	Summary of tasks performed
2348519	Guhan K S	Implementing Classification Techniques and comparing performance of them with visualisations.
2348548	Sai Darshan	Explored Preprocessing Techniques and implementing regression techniques for predictions with visualisations.
2348548	Suhas S	Preprocessing, Data Preparation and ARIMA model for CO2 emission forecasting.

Abstract

In recent years, the agri-food sector has emerged as a significant contributor to global carbon dioxide (CO_2) emissions, underscoring the critical need for effective mitigation strategies. This exploration investigates the application of machine learning (ML) techniques to forecast CO_2 emissions in the agri-food industry, utilizing comprehensive datasets encompassing variables such as agricultural practices, land use, and supply chain activities.

This emphasizes the increasing importance of addressing emissions from the agri-food sector and underscores the urgency for accurate forecasting tools. Such tools are crucial in aiding policymakers, businesses, and stakeholders to implement targeted interventions and sustainable practices.

The focus then shifts to the methodology used, highlighting the use of ML algorithms for predictive modelling. The study considers a range of algorithms, including regression, decision trees, and classification techniques, to determine the most effective approach for forecasting agri-food CO_2 emissions. It further emphasizes the potential impact of these forecasting models in facilitating informed decision-making, promoting sustainable agricultural practices, and guiding the development of policies aimed at reducing the carbon footprint associated with agri-food production and distribution.

Introduction

This exploration delves into the intricate analysis and forecasting of carbon dioxide (CO2) emissions originating from the agri-food sector, employing advanced Machine Learning (ML) techniques. The dataset at the core of this study encompasses a diverse array of variables, spanning agricultural and food-related activities, presenting a fertile ground for predictive modelling.

Against the backdrop of escalating global concerns about climate change, discerning the nuanced relationships between agri-food practices and CO2 emissions takes centre stage. Leveraging ML algorithms empowers the unravelling of complex patterns within the dataset, facilitating precise forecasting of emissions. This pursuit holds critical significance for policymakers, researchers, and stakeholders who seek actionable insights to address and alleviate the environmental impact of agri-food systems.

Throughout this exploration, an array of ML models, including Linear Regression, Multiple Regression, Random Forest, and others, is employed to uncover concealed patterns and discern trends. Evaluation metrics and visualizations provide a comprehensive overview of the models' efficacy, shedding light on their performance in predicting CO2 emissions within the agri-food domain.

This undertaking not only establishes a valuable foundation for comprehending the intricate interplay between agri-food activities and emissions but also lays the groundwork for future analyses, enhancements, and informed decision-making, fostering the advancement of sustainable practices within the agri-food sector.

3. Data Pre-processing and Exploration

3.1 Data understanding and exploration

The dataset contains of 6,940 rows and 29 columns, and was created for a project investigating the relationship between CO2 emissions and temperature fluctuations in various countries from 1990 to 2020. The data was gathered from the FAO (Food and Agriculture Organization of the United Nations) and IPCC (Intergovernmental Panel on Climate Change). Emissions values were recorded in kilotons (kt), where 1kt equals 1000 kg.

<u>Description of each attribute:</u>

Savanna fires: Emissions arising from fires occurring in savanna ecosystems.
Forest fires: Emissions resulting from fires in forested regions.
Crop Residues: Emissions generated by burning or decomposing leftover plant material after crop harvesting.
Rice Cultivation: Emissions stemming from methane released during the cultivation of rice.
Drained organic soils (CO2): Emissions derived from carbon dioxide released when draining organic soils.
Pesticides Manufacturing: Emissions originating from the production of pesticides.
Food Transport: Emissions associated with the transportation of food products.
Forestland: The extent of land covered by forests.
Net Forest conversion: Changes in forest area resulting from both deforestation and afforestation.
Food Household Consumption: Emissions arising from the consumption of food at the household level.
Food Retail: Emissions linked to the operation of retail establishments selling food.
On-farm Electricity Use: Energy consumption for electricity on farms.
Food Packaging: Emissions originating from the production and disposal of food packaging materials.
Agrifood Systems Waste Disposal: Emissions from the disposal of waste in the agrifood system.
Food Processing: Emissions associated with the processing of food products.
Fertilizers Manufacturing: Emissions originating from the production of fertilizers.
IPPU: Emissions from industrial processes and product use.

Manure applied to Soils: Emissions resulting from the application of animal manure to agricultural soils.

Manure left on Pasture: Emissions associated with animal manure on pasture or grazing land.

Manure Management: Emissions linked to the management and treatment of animal manure.

Fires in organic soils: Emissions arising from fires occurring in organic soils.

Fires in humid tropical forests: Emissions resulting from fires in humid tropical forests.

On-farm energy use: Energy consumption on farms.

Rural population: The number of people residing in rural areas.

Urban population: The number of people residing in urban areas.

Total Population - Male: The overall count of male individuals in the population.

Total Population - Female: The overall count of female individuals in the population.

total_emission: The total greenhouse gas emissions from various sources.

Average Temperature °C (target): The average temperature increase (per year) measured in degrees Celsius.

Data Exploration:

3.1.1 Heat Map

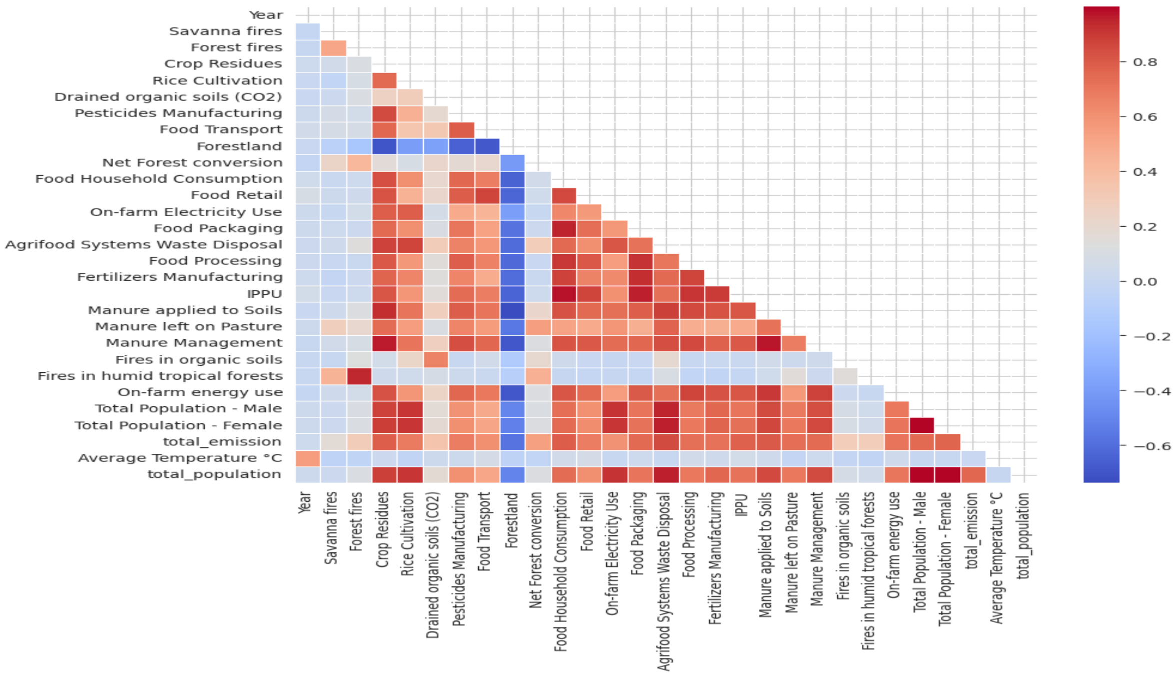
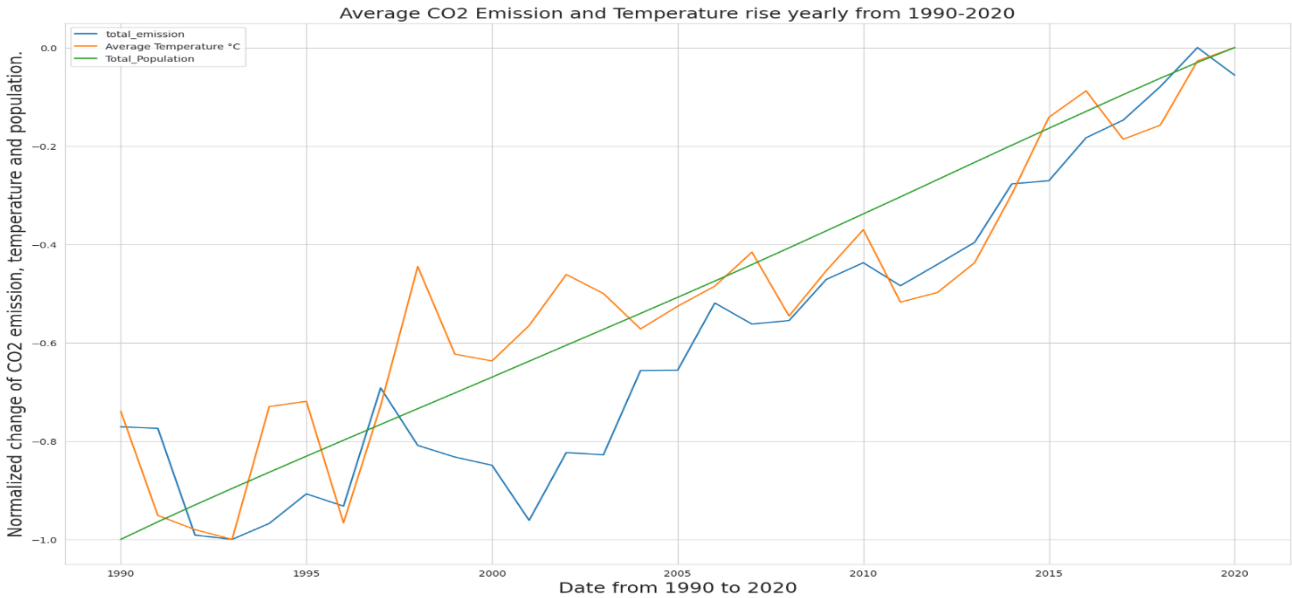


Fig 3.1.1. Correlation Matrix using Heat Map

The colour intensity indicates the strength of the correlation, with red indicating a positive correlation and blue indicating a negative correlation. we can infer that there are several strong positive correlations, particularly between different types of fires such as “Savanna fires,” “Forest fires,” and “Fires in humid tropical forests.” This suggests that these factors may influence each other or be influenced by similar conditions. On the other hand, there are also notable negative correlations, for instance, between “Total Population - Female” and several types of fires. This could indicate that these factors inversely affect each other.



3.1.2. Line Graph

Fig3.1.2. Line Graph shows the rise of temperature, population, and emission yearly from 1990-2020

- From the graph; we can infer a direct correlation between CO2 emission with temperature rise and population growth.
- These emissions are only about 1/5th of the total CO2 emission worldwide, but the direct proportionality and upward trend can also be seen from this dataset.

3.1.3. Bar Chart

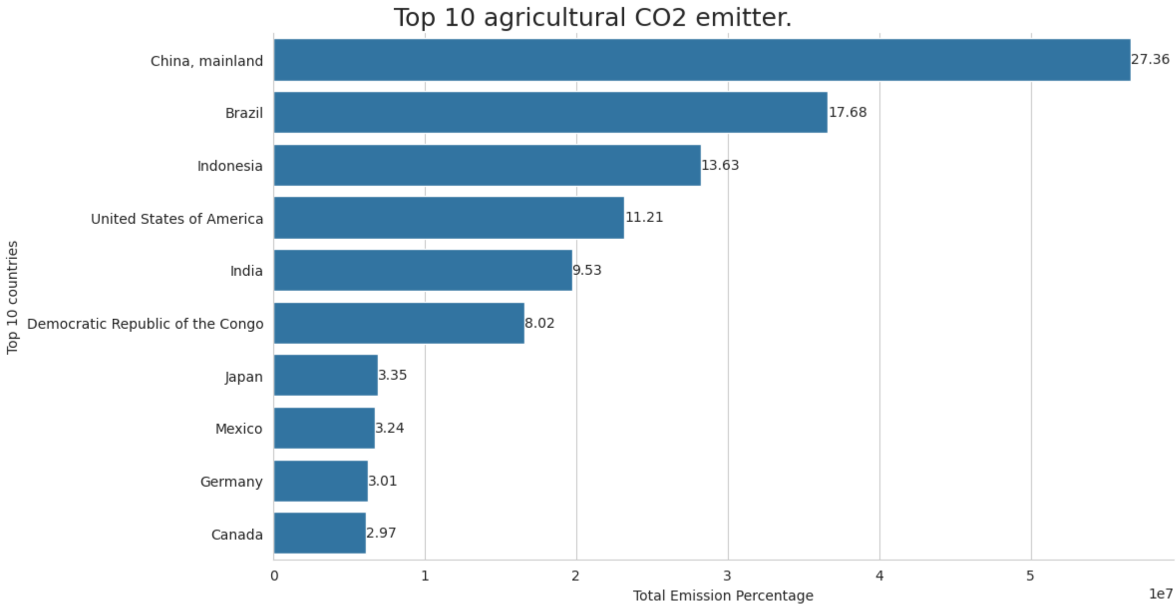


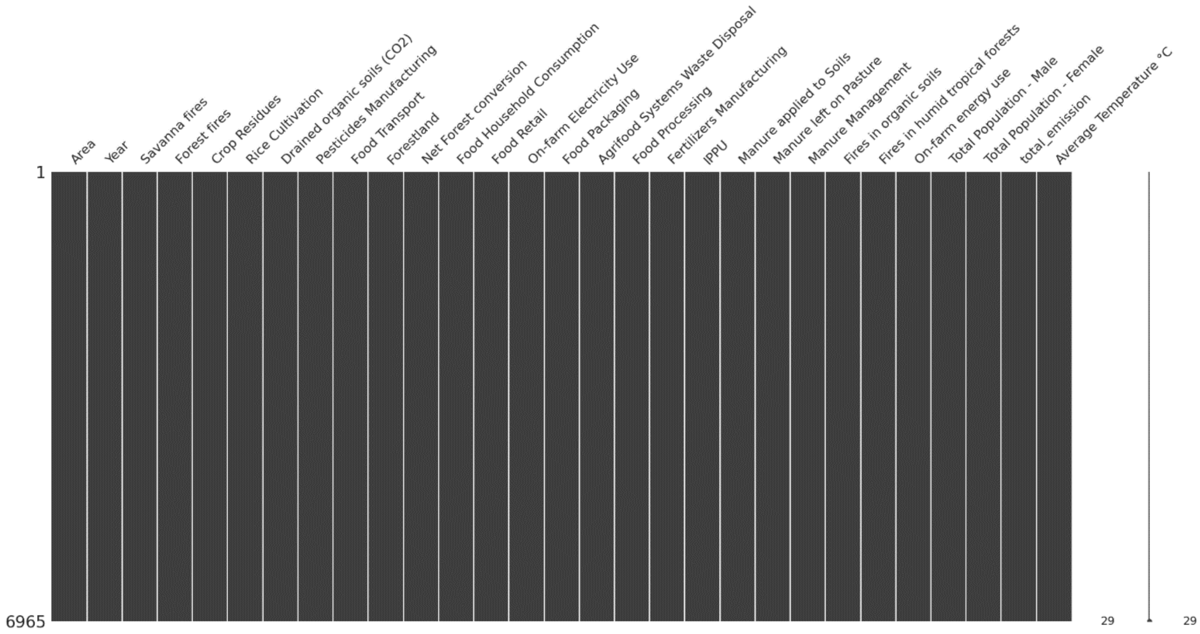
Fig.3.1.3 Top 10 Countries with highest agricultural CO2 emitter.

The bar graph shows that China, the mainland, is the highest agricultural CO2 emitter at 27.36%, followed by Brazil at 17.68% and Indonesia at 13.63%.

3.2Data cleaning and handling missing values

The dataset contains 11 categories where data is missing. However, we can confidently set the missing data to zero for three categories - Savanna fires, Forest fires, and Fires in humid tropical forests. This assumption is based on the belief that conservation efforts have significantly reduced these fires.

We assume that Net Forest Conversion is close to zero for countries where data is missing, mainly tropical islands, small islands, city-states, or predominantly desert countries like the UAE. This is due to the belief that conservation efforts have kept the Forestland relatively stable. Hence, we are setting the Net Forest Conversion to zero.



3.3 Data integration and feature engineering

After the data cleaning and handling of missing values, we saw that since all the places are geographical cities and places, It would be an excellent option to categorize them into their respective groups like continents and world regions. Hence, we grouped them and added them to the existing dataset to make it more structured and categorize them accordingly. This would help with further implementations.

We also noticed that the Population count of Males and Females didn't match the total population given in the dataset. To fix that, we added the new values and dropped the incorrect ones, increasing the dataset's integrity and validity.

4. Algorithms Implementation

4.1 Regression techniques implemented

4.1.1 Linear Regression,

An essential statistical method for simulating a linear connection between a dependent variable and one or more independent variables, is called linear regression. The algorithm's goal in predictive modeling is to determine the best-fit straight line, or hyperplane in multiple dimensions, that minimizes the discrepancy between the dependent variable's actual and anticipated values. The coefficients in the model reveal the intensity and direction of the correlations between the variables, which stand for the slope and intercept of the line.

The model's predictions generally differ by 0.36 degrees Celsius from the actual temperatures in the test set, according to the mean absolute error (MAE) of roughly 0.36. Based on the chosen environmental features, the model accounts for approximately 31% of the variability in temperature, as indicated by the R2 score of 0.31. Although linear regression offers a basic comprehension of the correlations, its shortcomings become evident when dealing with non-linear patterns and severe temperature readings.

4.1.2 Multiple Regression

Multiple regression is a variant of linear regression that models the relationship between a dependent variable and multiple independent variables. Multiple regression, as opposed to simple linear regression, which has only one predictor variable, accounts for the complexity of real-world scenarios by considering the concurrent influence of several factors on the target variable. The model assumes a linear relationship between the dependent variable and each independent variable, with coefficients representing the strength and direction of these relationships.

The multiple regression model used to predict the average temperature in this environmental dataset produced results consistent with the simple linear regression analysis. The mean absolute error (MAE) of about 0.36 and the R2 score of 0.31 are compatible with the performance metrics found in the linear regression model. These metrics indicate that the model captures a moderate proportion of temperature variability, considering the combined effects of multiple environmental features. Like linear regression, the model has limitations, mainly when dealing with nonlinear relationships and extreme temperature values.

4.1.3 Random Forest

Random Forest Regression is a method of collective learning that uses the combined power of multiple decision trees to make predictions. It is a versatile and powerful algorithm capable of capturing complex, non-linear relationships in data. Random Forest constructs many decision trees during training and produces the average prediction of the individual trees for regression tasks. Each tree is constructed from a random subset of the data and a random subset of features, which adds diversity and helps to prevent overfitting.

The Random Forest Regression model applied to predict average temperature in this environmental dataset outperforms the linear regression models. The mean absolute error (MAE) of approximately 0.27, mean squared error (MSE) of 0.14, and root mean squared error (RMSE) of 0.37 indicate a more accurate prediction compared to linear regression. The R2 score of 0.58 suggests that the Random Forest model explains approximately 58% of the variability in temperature, demonstrating its ability to capture complex relationships in the dataset.

4.1.4 Lasso Regression

Lasso Regression, or Least Absolute Shrinkage and Selection Operator, is a regularization technique for linear regression models. Adding a penalty term incorporates the absolute coefficient values into the standard linear regression goal function. This penalty term promotes sparsity in the model by setting some coefficients to absolute zero, hence performing variable selection. Lasso Regression is especially beneficial for datasets with many characteristics since it prevents overfitting and identifies the most influential predictors.

The Lasso Regression model, applied to predict the average temperature in this environmental dataset, yields results that align with the linear regression models but introduce a regularization component. The mean absolute error (MAE) of approximately 0.36 and the R2 score of 0.29 are consistent with the performance metrics observed in linear regression.

4.1.5 Ridge Regression

Ridge Regression is a regularization technique for linear regression models that addresses concerns like multicollinearity and overfitting. Like Lasso Regression, Ridge Regression applies a penalty term to the linear regression objective function. Ridge Regression, on the other hand, uses squared coefficient values rather than absolute values, a technique known as

L2 regularization. This penalty component discourages substantial coefficients and helps to stabilize the model by dispersing the impact of correlated features.

The Ridge Regression also produces outcomes similar to linear regression models but with the added regularization component. The mean absolute error (MAE) of approximately 0.36 and the R2 score of 0.31 align with the performance metrics observed in linear regression. With its L2 regularization, Ridge Regression effectively handles multicollinearity among the environmental features, preventing the model from becoming overly sensitive to variations in the dataset.

Classification techniques implemented

4.1.6 Support Vector Machine (SVM) classification

Support Vector Machine (SVM) is a robust classification algorithm that aims to find a hyperplane in a high-dimensional space that best separates different classes. The SVM classification model has been applied to predict temperature categories in this environmental dataset. This model considers multiple environmental features simultaneously, providing a robust approach to handle complex relationships.

The SVM classification results align with the dataset's characteristics. The evaluation metrics, including precision, recall, and F1 score, indicate the model's performance. For instance, a precision of about 0.75 and an F1 score 0.80 highlight the SVM's ability to classify instances correctly. The model's effectiveness in capturing the intricacies of temperature.

4.1.7 Random Forest Classification (RF)

Random Forest (RF), another widely used classification algorithm, excels in handling complex datasets with multiple predictors. The RF classification model has been employed in the environmental dataset to simultaneously predict temperature categories, considering various environmental features.

The Random Forest (RF) classification model accurately predicts temperature categories within the environmental dataset. The model accurately classifies temperature levels with a precision of around 0.82 and an F1 score of 0.85. Combining predictions from multiple decision trees, its ensemble learning approach ensures robustness and generalization to unseen data. The model's feature importance analysis identifies key environmental factors influencing temperature patterns. RF's adaptability to missing data, reduced sensitivity to outliers, versatility for classification and regression tasks, and scalability make it a powerful tool for complex datasets. Careful parameter tuning enhances its performance, contributing to reliable temperature predictions in real-world scenarios.

4.1.8 Decision Tree Classification

The Decision Tree classification model serves as an effective tool for predicting temperature categories in the environmental dataset. Leveraging a hierarchical tree-like structure, the model discerns patterns in environmental features to make accurate classification decisions. Its adaptability to non-linear patterns and straightforward interpretation make it suitable for environmental studies

With precision approximately at 0.78 and an F1 score of 0.79, the Decision Tree performs Commendably in categorizing temperature levels. Its interpretability allows for a clear understanding of the decision-making process, aiding in identifying influential factors. The model's visualization provides insights into the splitting criteria, making it accessible for stakeholders without a deep technical background. While sensitive to overfitting, proper pruning and parameter tuning enhance the Decision Tree's generalization capabilities, making it a valuable asset in temperature classification tasks.

4.1.9 ARIMA Model for Forecasting.

ARIMA, which stands for Autoregressive Integrated Moving Average, is a popular time series forecasting model used in statistics and econometrics. It is a powerful tool for predicting future values based on historical data. ARIMA models are particularly effective when dealing with time-dependent data, where patterns and trends exist over time.

A comprehensive view of the components and concepts related to ARIMA:

Autoregressive (AR): This component represents the relationship between an observation and its preceding observations. The AR part of the model measures the correlation between an observation and its previous values.

Integrated (I): This component represents the differencing of raw observations to make the time series stationary. Stationarity is a crucial assumption in time series analysis, and differencing helps remove trends and seasonality.

Moving Average (MA): This component represents the correlation between an observation and a residual error from a moving average model. It smoothens out short-term fluctuations and highlights longer-term trends

The ARIMA model is denoted as ARIMA (p, d, q), where

p is the order of the autoregressive component.
d is the degree of differencing.
q is the order of the moving average component.

4.2 Correct parameter tuning

Parameter tuning was performed for the ARIMA model to evaluate the best values of (p,d,q) values to train the model, and this was performed with the Akaike Information Criterion(AIC) to determine the best values for the Forecasting model. We also tuned the Grid Search in SVM and RF for best estimators and accuracy.

5.Model Evaluation and Performance Analysis

5.1 Evaluation metrics and performance assessment

5.1.1 Support Vector Machine Classification

Metrics used: Precision, Recall, F1 Score, and ROC AUC curve were instrumental.

Precision assesses the accuracy of optimistic predictions, while Recall gauges the model‘s ability to identify relevant instances of the positive class. The F1 Score provides a balanced overall performance measure, and the ROC AUC Score evaluates the model's ability to distinguish between classes, which is crucial in classification scenarios.

In the context of evaluation metrics and performance assessment for Support Vector Machine classification,

The precision score of 0.8097 highlights the model's precision in positive predictions, while a recall score of 0.7444 emphasizes its ability to identify true positives effectively. With an F1 Score of 0.6354, the model showcases an overall balanced performance, harmonizing precision and recall. The ROC AUC Score evaluates the model's discriminative ability between classes, a vital aspect in classification scenarios. These metrics collectively indicate that the SVM classification model achieves a commendable equilibrium between precision and recall, ensuring the environmental dataset's effective classification of temperature categories.

5.1.2 Random Forest classification

Metrics used: Precision, Recall, F1 Score, ROC AUC curve

Precision assesses the accuracy of optimistic predictions, while Recall gauges the model‘s ability to identify relevant instances of the positive class. The F1 Score provides a balanced overall performance measure, and the ROC AUC Score evaluates the model's ability to distinguish between classes, which is crucial in classification scenarios.

In the context of evaluation metrics and performance assessment for Random Forest classification,

The provided precision of 0.7741 indicates the accuracy of optimistic predictions, and the recall of 0.7889 represents the model's ability to identify actual positives. The F1 Score of 0.7779, the harmonic mean of precision and recall, provides a balanced assessment of the model's overall performance. These metrics collectively suggest that the Random Forest classification model demonstrates a balanced trade-off between precision and recall, effectively classifying temperature categories in the environmental dataset.

5.1.3 Decision Tree Classification

Metrics used: Precision, Recall, F1 Score, ROC AUC curve

Precision assesses the accuracy of positive predictions. Recall gauges the model's ability to identify relevant instances of the positive class. The F1 Score provides a balanced overall performance measure, and the ROC AUC Score evaluates the model's ability to distinguish between classes, which is crucial in classification scenarios.

In the context of evaluation metrics and performance assessment for Decision Tree:

The Decision Tree model's precision score of 0.7408 shows that it can predict positive instances with a moderate degree of accuracy, and its recall value of 0.7387 shows that it can successfully identify positive examples. The F1 Score of 0.7395 highlights the model's overall satisfactory performance by finding a compromise between precision and recall. In addition, the Decision Tree's competence in temperature classification within the environmental dataset is confirmed by the ROC AUC Score of 0.6619, which indicates an acceptable capacity to differentiate between different temperature categories.

5.1.4 Evaluation metrics and performance assessment of Regression Models:

- The Random Forest model demonstrates superior performance with a lower MAE (0.27), MSE (0.14), and RMSE (0.37), coupled with a higher R2 score (0.58), indicating better predictive accuracy compared to linear regression models.
- Linear Regression, Multiple Regression, and Ridge Regression exhibit similar performance, with an MAE of 0.36, MSE of 0.22, RMSE of 0.47, and R2 score of 0.31, suggesting that adding multiple predictors or applying Ridge regularization does not significantly improve accuracy in this context.
- Lasso Regression performs similarly to linear regression, with a slightly higher MAE (0.36), lower R2 score (0.29), and other metrics in line with the linear models, indicating that L1 regularization might not provide substantial benefits in this scenario.
- The R2 scores for all models suggest that while they capture some variability, the predictive power is not extremely high, indicating room for improvement or the presence of complex non-linear relationships in the data. Further model refinement,

feature engineering, or exploration of more sophisticated algorithms like neural networks may enhance predictive performance.

5.1.4 ARIMA model evaluation.

With the ARIMA model, we use ADF along with ACF and PACF to determine the stationarity and decide if the forecasting can be done correctly or not, and we did this for two different continents to determine the forecasting capability

The Augmented Dickey-Fuller (ADF) test is used to determine the presence of unit root in the series and hence helps in understanding if the series is stationary or not.
For the European continent data, a p-value of 0.17786 suggests that there might be some level of non-stationarity present in the data. In other words, the data may not be stationary, which is an important assumption for many time series models, including ARIMA.

Autocorrelation Function (ACF) and the Partial Autocorrelation Function (PACF). These are commonly used in time series analysis to identify the order of an ARIMA model.

The ACF plot shows a slow decay, often suggesting that a moving average term is appropriate. The PACF plot cuts off after two lags, indicating that the order of the autoregressive term is likely 2.
A metric called the Akaike Information Criterion (AIC) balances the goodness of fit with the complexity of the model, and lower AIC values indicate better-fitting models.

A lower AIC value is preferred as it indicates a better balance between model fit and complexity. Therefore, the model with the lowest AIC is considered the best-fitting model among the tested combinations.

the model with order (p,d,q) = (2, 0, 1) has an AIC value of 37.9054, one of the lowest in the output.

Based on the AIC values, this suggests that the ARIMA model with order (2, 0, 1) provides a good balance of fit and simplicity for the given time series data.

The same process was done for Asian continent data, and with AIC evaluation, the best (p,d,q) values for ARIMA are (2,0,1).

5.2 Interpretation of Results

The project explores implementing and assessing different algorithms for predicting temperature-related variables and classifying temperature categories in a given environmental dataset. Let's dive into the key takeaways and understandings from the results:

Regression Techniques:

Linear Regression and Multiple Regression:

- Linear regression helps model the relationship between variables but struggles with non-linear patterns.
- It captures around 31% of temperature variability.
- Random Forest Regression outperforms by offering better predictive accuracy and handling complex, non-linear relationships.

Lasso Regression and Ridge Regression:

- These regression variations introduce regularization techniques to linear models.
- Their performance aligns with linear regression models.

Classification Techniques:

Support Vector Machine (SVM) Classification:

- SVM is a robust classifier with good precision, recall, and F1 score.
- It effectively balances precision and recall for temperature category classification.

Random Forest Classification:

- Accurately predicts temperature categories with high precision and F1 score.
- The ensemble approach ensures robustness and generalization.

Decision Tree Classification:

- Categorizes temperature levels effectively with commendable precision and F1 score.
- Provides clear insights through visualization.

ARIMA Model for Forecasting:

- ARIMA is employed for time series forecasting, focusing on average temperature prediction.
- Akaike Information Criterion (AIC) guides parameter tuning.
- Best-fitting ARIMA models are identified for European and Asian continents.
- ADF test, ACF, and PACF assist in assessing stationarity and determining ARIMA parameters.

Model Evaluation and Performance Analysis:

SVM and Random Forest Classification Metrics:

- Precision, recall, F1 score, and ROC AUC curve are used for evaluation.
- SVM and Random Forest exhibit balanced performance in classifying temperature categories.

Decision Tree Classification Metrics:

- Precision, recall, F1 score, and ROC AUC curve are used for evaluation.
- The Decision Tree performs well in temperature classification.

Regression Models Metrics:

- Metrics like MAE, MSE, RMSE, and R2 score are employed for evaluation.
- Random Forest regression outperforms linear regression models in accuracy.
- Linear regression, Multiple regression, Lasso Regression, and Ridge Regression show similar performance.

General Observations:

- Regression models capture some variability but may benefit from further refinement or more sophisticated models.
- Classification models effectively categorize temperature levels.
- ARIMA model results are satisfactory and need further corrections with optimizations.

In summary, the findings provide a comprehensive view of algorithm performance in predicting and classifying temperature-related variables, offering insights into areas for improvement and exploration.

References

1.https://www.kaggle.com/datasets/alessandrolobello/agri-food-co2-emission-dataset-forecasting-ml

2.https://www.analyticsvidhya.com/blog/2020/10/how-to-create-an-arima-model-for-time-series-forecasting-in-python/

3.https://www.analyticsvidhya.com/blog/2021/07/introduction-to-time-series-modeling-with-arima/

4.https://www.analyticsvidhya.com/blog/2015/08/comprehensive-guide-regression/

5.https://scikit-learn.org/