A logo for university computing

Auto-generated description

**Evaluation cover page**

|  |  |
| --- | --- |
| *Student's Full Name:* | Taylan Ozgur Ozkan |
| *Student Number:* | 2024140 |
| *Module Title:* | Machine Learning |
| *Evaluation Title:* | CA1. |
| *Assessment Due Date:* | 27/10/2024. |
| *Presentation Date* | 01/11/2024. |

**Declaration**

By submitting this review, I confirm that I have read CCT's policy on academic misconduct and understand the implications of submitting work that is not mine or that does not appropriately reference material taken from a third party or other source.

I declare that it is my own work and that all third-party material has been properly referenced.

I further confirm that this work has not previously been submitted for evaluation by me or anyone else at CCT College Dublin or any other higher education institution.

Contents

Climate Change Impact on Agriculture 2024 1

[Introduction 1](#_Toc181377331)

[1. Data Preprocessing and Exploratory Data Analysis (EDA) 1](#_Toc181377332)

[Data Loading and Review 1](#_Toc181377333)

[Summary Statistics of Continuous Variables: 2](#_Toc181377334)

[Correlation Analysis 5](#_Toc181377335)

[2. Cluster Analysis 5](#_Toc181377336)

[Selection of Clustering Algorithm and Determination of Number of Clusters 5](#_Toc181377337)

[Running the Algorithm and Visualizing the Results 7](#_Toc181377338)

[3.Time Series Analysis and ARIMA Model 11](#_Toc181377339)

[Model Performance Evaluation and Interpretation 12](#_Toc181377340)

[Visualization of Forecast Trends 13](#_Toc181377341)

[General Assessment 13](#_Toc181377342)

[General Assessment and Recommendations 14](#_Toc181377343)

Climate Change Impact on Agriculture 2024

# Introduction

**The effects of climate change are increasingly putting pressure on the agricultural sector, posing serious risks to food security, economic sustainability, and environmental balance. This study examines the impacts of climate variables such as temperature, precipitation, and CO2 emissions on agricultural yields using data analytics and modeling methods. Short-term forecasts were employed to predict future trends.**

# 1. Data Preprocessing and Exploratory Data Analysis (EDA)

## Data Loading and Review

Column Names and Data Types:

* **The dataset includes a total of 15 columns, encompassing both categorical (e.g., Country, Region, Crop\_Type, Adaptation\_Strategies) and numerical (e.g., Average\_Temperature\_C, Total\_Precipitation\_mm) data types.**

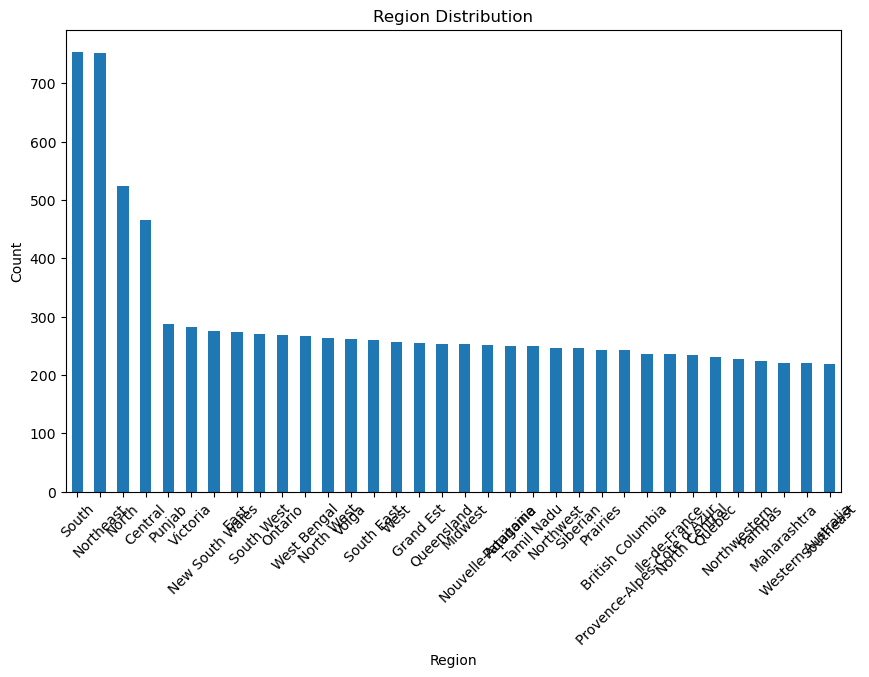
Dataset Dimensions:

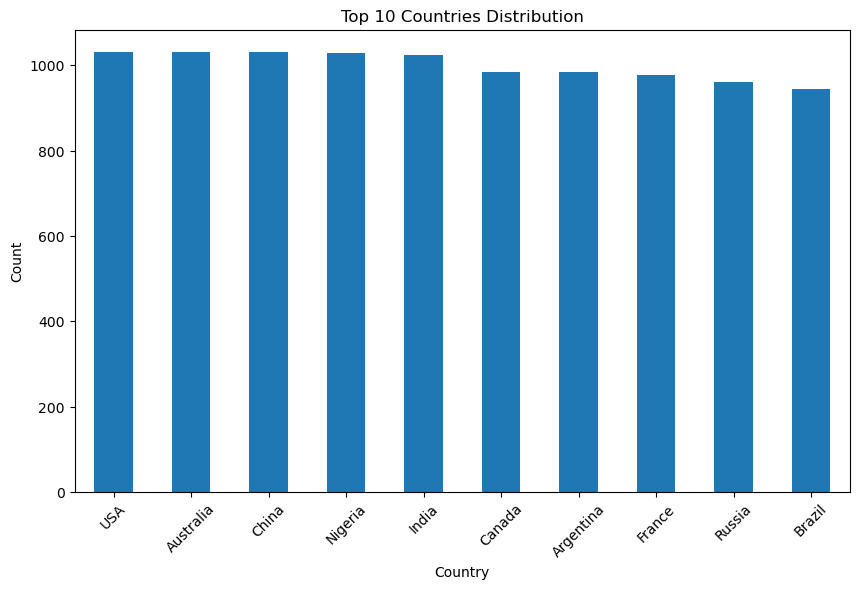
* **The dataset consists of 10,000 rows and 15 columns.**

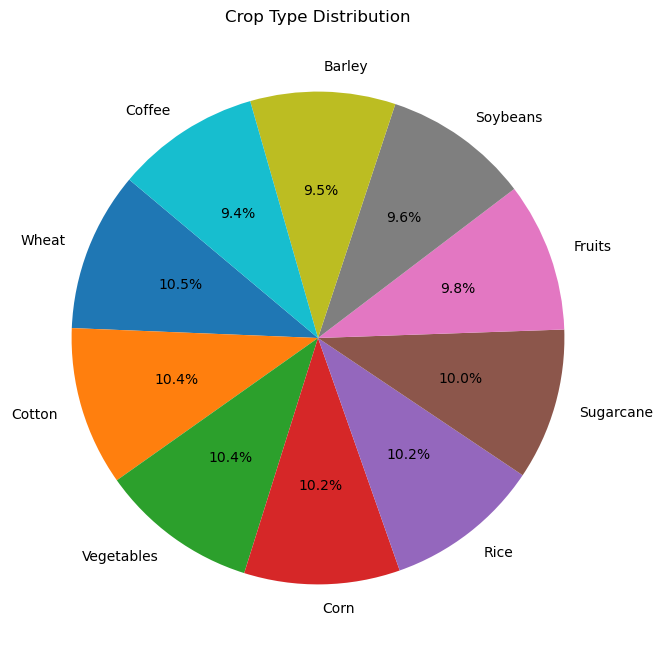
Missing Data Analysis:

* **No missing data were found in the columns.**

## Summary Statistics of Continuous Variables:

****

****

****

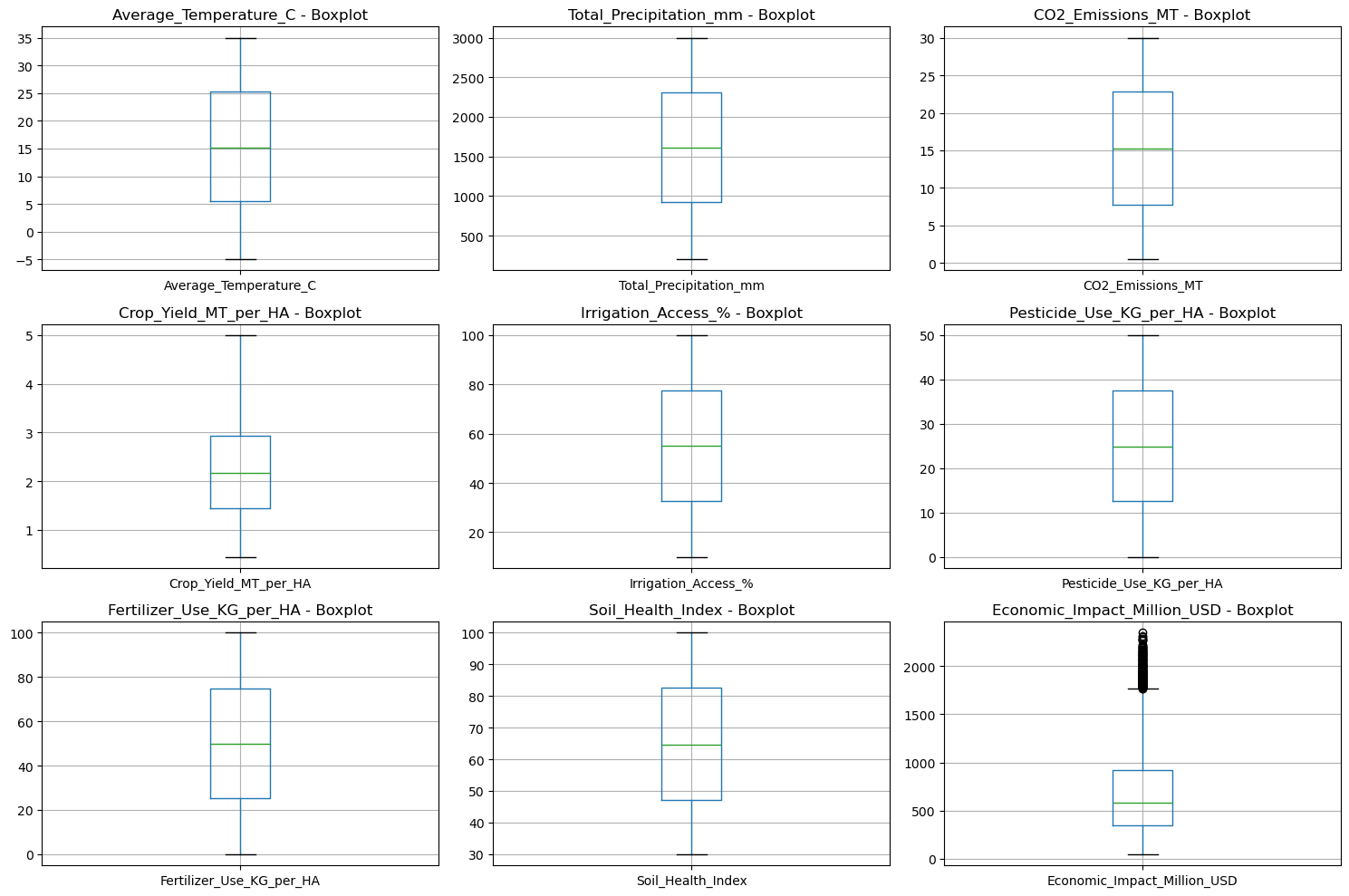
* **Basic statistics such as average values, minimums, maximums, and standard deviations were calculated for continuous variables.**

**Some key highlights:**

* **Temperature (Average\_Temperature\_C): Mean 15.24°C, minimum -4.99°C, maximum 35.00°C.**
* **Total Precipitation (Total\_Precipitation\_mm): Mean 1611 mm, with considerable variability observed.**
* **CO2 Emissions (CO2\_Emissions\_MT): Mean 15.25 MT, showing a broad distribution in emission levels.**
* **Crop Yield (Crop\_Yield\_MT\_per\_HA): Mean 2.24 MT/HA, displaying relatively low variability.**
* **Economic Impact (Economic\_Impact\_Million\_USD): Mean 674 million USD; economic impact is notably higher in certain countries.**

**Distribution of Categorical Variables:**

* **Country: The distribution for the first 10 countries is visualized with a bar chart.**
* **Region: All regions are represented with a bar chart, illustrating data distribution across regions.**
* **Crop Type: Displayed with a pie chart, showing that certain crop types are more prevalent.**

****

**As a result of the outlier analysis we conducted for continuous variables, it is observed that there are outliers in some variables in particular:**

**• Average\_Temperature\_C and Total\_Precipitation\_mm: There are some outliers at the upper limit in temperature and precipitation variables.**

**• CO2\_Emissions\_MT and Economic\_Impact\_Million\_USD: Some upper limit outliers are also observed in CO2 emissions and economic impact variables.**

**• Irrigation\_Access\_% and Pesticide\_Use\_KG\_per\_HA: There are also extreme values ​​in these variables.**

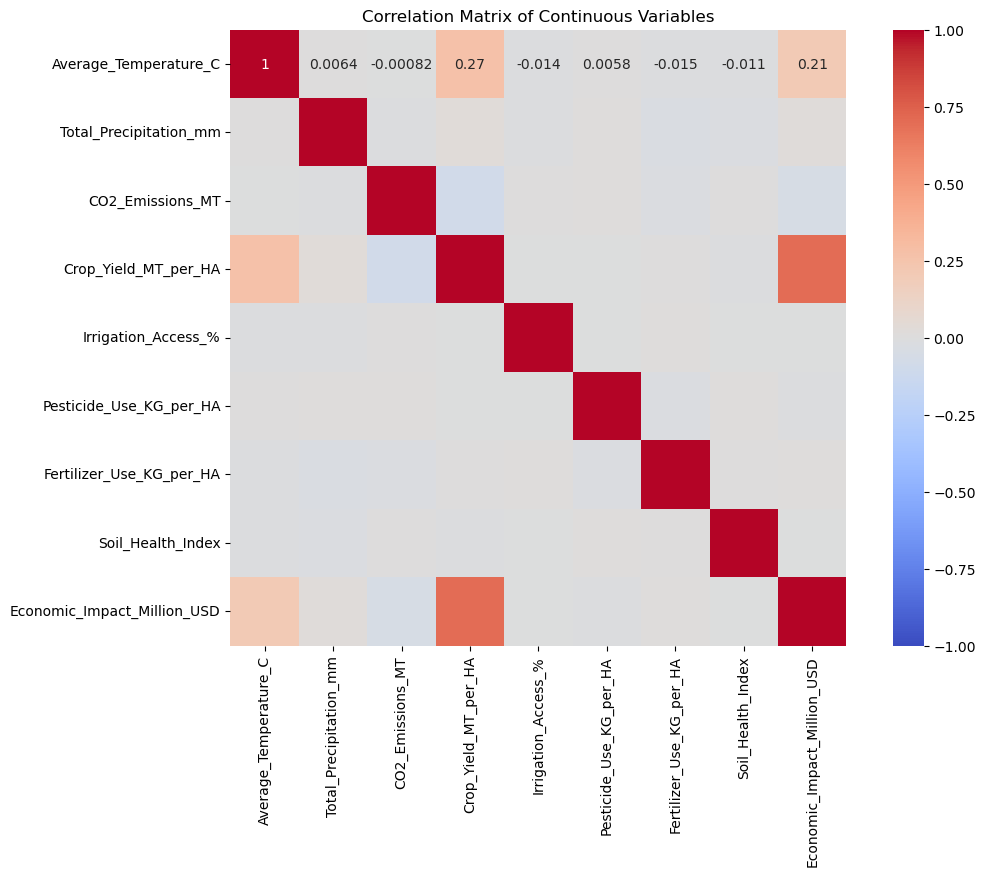
**The IQR method (Interquartile Range) may be a good choice to clean outliers.**

**IQR is Calculated: The first quartile (Q1) and third quartile (Q3) of each continuous variable are found, then IQR = Q3 - Q1 is calculated.**

**Outlier Boundaries Are Determined: Outlier boundaries are calculated as Q1 - 1.5 \* IQR (lower bound) and Q3 + 1.5 \* IQR (upper bound).**

**Filtering Outliers: Data falling outside these boundaries are considered outliers and are removed from the data set.**

**After filtering outliers, the size of our data set decreased from 10,000 rows to 9,823 rows. This process helped prevent extreme outliers from negatively affecting the analysis results.**

Correlation Analysis****

# 2. Cluster Analysis

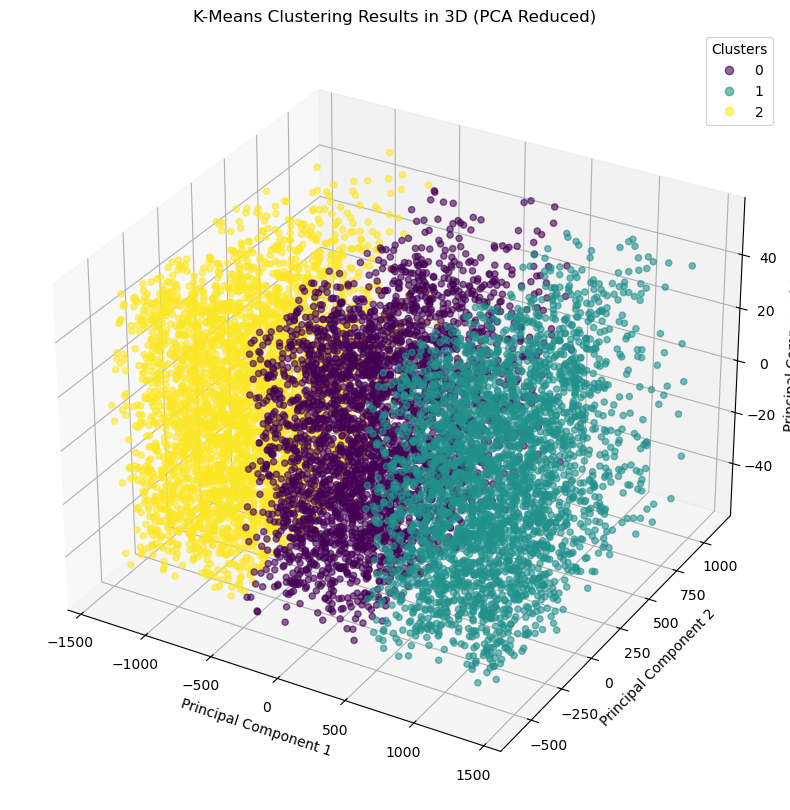
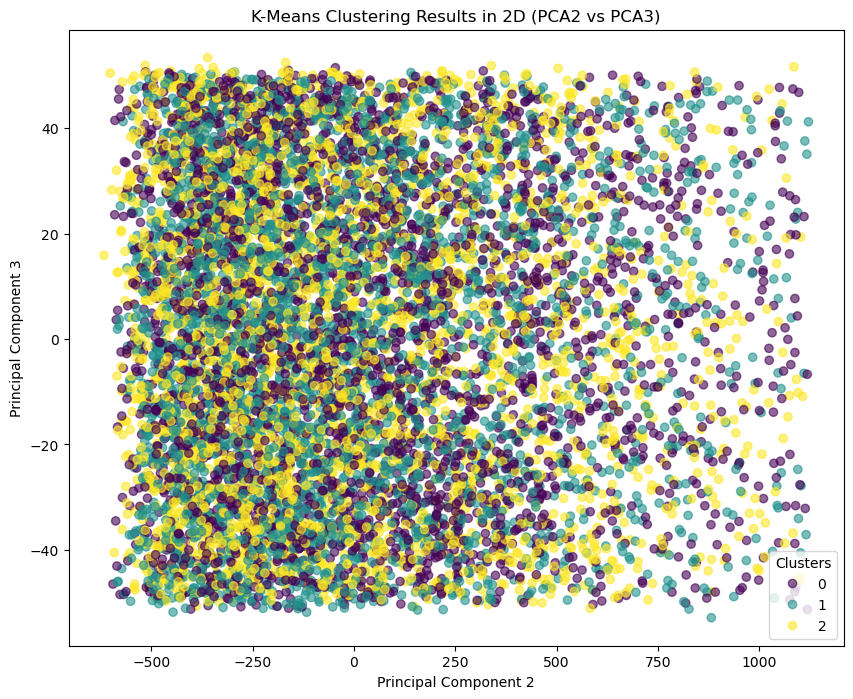
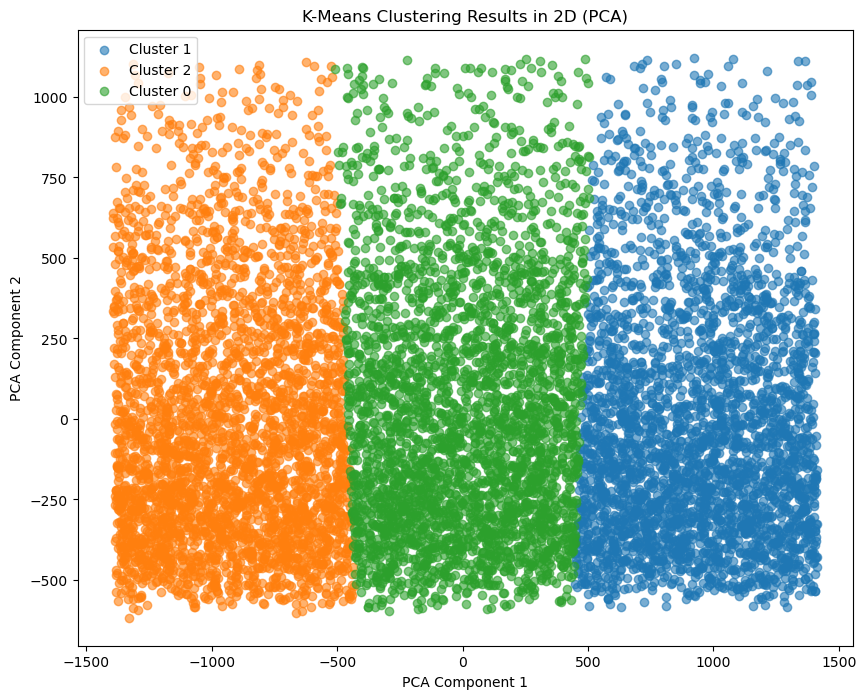
## Selection of Clustering Algorithm and Determination of Number of Clusters

**Selection of Clustering Algorithm: We chose K-Means algorithm to divide the data into meaningful segments. K-Means is an algorithm that provides fast and effective results in large data sets and groups the data according to a certain number of clusters. In this analysis, we numerically converted categorical variables such as country, region and product type and applied clustering with K-Means.**

**A graph with a line

Description automatically generated**

**Determining the Number of Clusters: We used the Elbow Method to determine the optimal number of clusters. The graph above shows a curve showing the relationship between the number of clusters and WCSS (within-cluster total squared deviation). The "elbow" point in the graph indicates the number of clusters where the WCSS decrease slows down. In this analysis, we determined that the optimal result can be achieved with 3 clusters.**

Running the Algorithm and Visualizing the Results****

**First 2D Graph (PCA1 vs PCA2):**

**In this graph, we see the separation between the clusters through the first two components of PCA.**

**Second 2D Graph (PCA2 vs PCA3):**

**It can be observed that there is a more complex distribution between the clusters. There is more overlap between some clusters through these components.**

**3D Graph (PCA1, PCA2, PCA3):**

**There is a certain boundary between the clusters, but there are transitions in some areas. This may indicate that there is small overlap in the nature of the data.**

**In order to analyze the characteristic differences between the clusters, we learned the summary statistics of each cluster with the code df\_encoded.groupby("Cluster").mean().**

**Average Temperature and Precipitation:**

**• While Cluster 0 and Cluster 2 have similar values ​​in terms of average temperature, Cluster 1 shows a slightly lower average temperature.**

**• The highest rainfall is found in Cluster 2 (average 2536 mm), which may indicate that the region receives more rainfall. Cluster 1 has significantly lower rainfall (average 676 mm).**

**CO2 Emissions and Economic Impact:**

**• CO2 emissions are similar in all three clusters, indicating that there is no significant difference between the clusters.**

**The highest mean economic impact is seen in Cluster 0, while the lowest mean is seen in Cluster 1.**

**Agriculture and Soil Health Indicators:**

**• However, Pesticide\_Use\_KG\_per\_HA and Fertilizer\_Use\_KG\_per\_HA differ slightly. In particular, pesticide use is higher in Cluster 2, while it is lower in Cluster 1.**

**PCA Components:**

**• According to the first three PCA components, the clusters are spread over certain areas**

3.Time Series Analysis and ARIMA Model**A graph showing the temperature of years

Description automatically generated**

According to the graph of the average temperature time series by year, there are significant fluctuations in the data and sudden increases and decreases in some periods. This situation shows that there are sudden changes in the data as well as trend and seasonal characteristics.

We can say the following by analyzing the graph:

1. Trend: Although there is no general upward or downward trend, there are increases and decreases from time to time.

2. Seasonality: No specific cycle or periodic structure is observed. However, there are sudden changes in some years.

3. Stationarity: According to the ADF test results,

ADF Statistic: -4.81, this value is generally associated with stationary series with lower negative values.

p-value: 5.20e-05, i.e. much smaller than 0.05. This shows that the series is stationary.

Conclusion:

Our time series can be considered stationary at a 95% confidence level. In this case, we can apply the series to the ARIMA model without further transformation.

**A comparison of a graph

Description automatically generated with medium confidence**

According to ACF and PACF graphs:

In the ACF (Autocorrelation Function) graph, there is a high correlation at the first lag (lag 1), but the correlation decreases rapidly at subsequent lags. This shows that a value of 1 may be appropriate for the q parameter.

A similar situation is seen in the PACF (Partial Autocorrelation Function) graph; here, the first lag is prominent, and then the correlation decreases rapidly. In this case, a value of 1 seems appropriate for the p parameter.

## Model Performance Evaluation and Interpretation

MAE, MSE and RMSE error metrics were used to evaluate the performance of the ARIMA model and the results were obtained as follows:

• Mean Absolute Error (MAE): 0.437

• Mean Squared Error (MSE): 0.201

• Root Mean Squared Error (RMSE): 0.448

These error metrics are important indicators of how well the model makes predictions:

MAE (Mean Absolute Error): The MAE value was obtained as 0.437, which means that the model's predictions deviate from the true values ​​by 0.437 units on average. The lower the MAE value, the higher the model's prediction accuracy.

MSE (Mean Squared Error): The MSE value is 0.201 and this metric is more sensitive to large errors because it takes the square of the errors. This low value shows that the model avoids large errors and keeps its predictions close to the true values.

RMSE (Root Mean Squared Error): The RMSE value is 0.448, which indicates that the estimates deviate from the actual values ​​by approximately 0.448 units on average.

Visualization of Forecast Trends**A graph showing a graph of value

Description automatically generated with medium confidence**

**In the graph, the fluctuating structure of the real values ​​in recent years and the temperature values ​​predicted by the model are shown side by side. The ARIMA model's predictions are consistent with the fluctuations in past years by remaining stable around a certain average after 2020. The model predicts that temperatures after 2020 will be around 15.2 °C by capturing past trends.**

## General Assessment

**• Accuracy: The low error metrics obtained indicate that the model has good accuracy. Especially in a climatic variable such as temperature, an error rate at this level is considered reasonable.**

**• Trend Predictability: The temperature values ​​predicted by the model showed slight fluctuations in line with the trends in past years and provided a stable prediction.**

**• Future Forecast Ability: Since the model offers a prediction close to the average with current data, it offers a reliable prediction in short-term predictions. However, more complex and detailed models may be needed for long-term projections.**

# General Assessment and Recommendations

**The results of this study contribute to a better understanding of the negative effects of climate change on the agricultural sector and shed light on future measures to be taken. Model performance analyses revealed that the ARIMA model used provides reliable results in short-term forecasts and accurately reflects the trends of the available data. However, for long-term projections and more complex scenarios, different modeling techniques and more comprehensive data sets need to be evaluated.**