Table of Contents

[Introduction 1](#_Toc203685711)

[Descriptive 2](#_Toc203685712)

[Spatial Distribution of Biomass Across the Farm 2](#_Toc203685713)

[Biomass and central tendency 2](#_Toc203685714)

[Spatial Distribution of Biomass 3](#_Toc203685715)

[Best environment for forage 4](#_Toc203685716)

[Data Exploration and Analysis Approach 4](#_Toc203685717)

[Insights from the boxplots 5](#_Toc203685718)

[Statistical Observations and Ranges 5](#_Toc203685719)

[Analysis of Temperature and Solar Radiation 6](#_Toc203685720)

[Comparison between measures I used in descriptive 7](#_Toc203685721)

[importance of data analytical techniques to the decision-making 8](#_Toc203685722)

[Predictive Analytics 8](#_Toc203685723)

[Models 9](#_Toc203685724)

[Exponential Smoothing 10](#_Toc203685725)

[LSTM 11](#_Toc203685726)

[Compare table 11](#_Toc203685727)

# Introduction

The increasing challenges in agriculture and livestock raise the need for data analytics for informed decision-making to enhance sustainable farming. This project has been designed with the objective of conducting biomass production data analysis on pasture through descriptive, predictive, and prescriptive analytics for better management of pasture to achieve sustainability.  
  
These data, sourced from a research paper, include satellite-derived vegetation indices, weather data, and forage quality metrics. This diverse range of sources provides a robust foundation for analyzing current trends in biomass production, forecasting future variations considering environmental factors, and developing prescriptive models to optimize productivity while maintaining ecological balance.  
  
This report will discuss various stages of data analysis, starting with descriptive analytics to understand the key insights into the distribution and characteristics of data; predictive analytics will be used next to forecast the amount of biomass produced, and finally, prescriptive analytics to determine the best course of action to realize the desired outcomes. All analyses will be performed in Python using modern data analysis tools and techniques to make accurate, evidence-based recommendations.  
  
The evidence-based solutions that we will provide through this project will inform strategic decisions in the agricultural sector toward improved pasture productivity and environmental sustainability.

# Descriptive

The project on biomass production, environmental factors, and satellite-derived vegetation indices will be executed using descriptive analytics to explore the general characteristics of the dataset. In this section I will provide proper insight into the patterns and trend composition in the data, hence giving valuable insight in decision-making for pasture management. All the possible statistical measures that will be prepared involve central tendency, dispersion, and frequency distribution to arrive at completeness of clarity about variability in the main distribution of features.

This analysis will be further enhanced through visualization by applying techniques such as the histogram, box plot, and scatter plot in establishing trends and relationships that may exist within the data. Further, I will apply methods of outlier detection to reveal unusual data points that may impact the analysis.

## Spatial Distribution of Biomass Across the Farm

### Biomass and central tendency

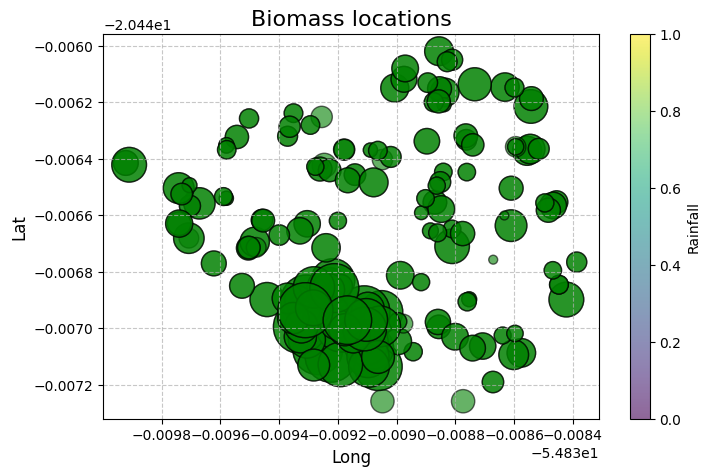
The calculated values are as follows:

* **Mean Biomass:** 4302.39 kg/ha
* **Median Biomass:** 3189.83 kg/ha
* **Mode Biomass:** 414.33 kg/ha

The big difference between the mean and median is an indication of a **right-skewed distribution**, hence showing that there could be higher biomass values pulling the mean upwards. This could be due to specific areas within the farm that exhibit exceptionally high biomass production, probably due to environmental factors such as soil fertility, rainfall distribution, or management practices.

Also, the mode is well below both the mean and median, indicating that low values of biomass are more frequent, while a few high values may act as outliers. These findings emphasize the importance of further analysis to pinpoint factors contributing to the variability in biomass production.

### Spatial Distribution of Biomass

To gain a general overview of the geographical distribution of biomass across the farm, I created a scatter plot using geographical coordinates (latitude and longitude) along with biomass values. This visualization effectively displays the biomass distribution across different locations, making it easy to identify any potential anomalies.

What I noticed from my analysis was that there are certain regions with much greater values of biomass relative to other regions. Such uneven distribution may signify that there are certain environmental conditions, like quality of soil, moisture levels, or vegetation density, which might be affecting biomass growth. Also, the possibility or otherwise of grazing animals might be at play, which we will examine in more detail.

#### Analysis of Biomass in Relation to Environmental Factors and Grazing Impact

To better understand the factors influencing biomass production, I made a heatmap to explore the correlations between biomass and environmental features. The heatmap helped me to identify the features that might have a significant impact. From the analysis, I determined that the features most likely to influence biomass include Animals, Radiative Direct Average, SAVI, and EVAPOT.

#### Analysis of Biomass and Radiative Direct Average

To explore the connection between biomass and radiative direct average, I performed a regression plot analysis to visualize their correlation. The scatter plot showed no significant relationship between these two variables. The data points seemed to be randomly scattered without any distinct trend, indicating that radiative direct average does not directly affect biomass production. This suggests that other factors, like soil moisture and vegetation indices, may have a more substantial influence on biomass than solar radiation.

#### Analysis of biomass and evapotranspiration

Next, I looked into how biomass relates to SAVI, which is a measure of vegetation health and density. I used boxplot analysis to divide the SAVI values into three categories: low, medium, and high. The findings revealed a strong positive correlation between SAVI and biomass, suggesting that areas with higher SAVI values generally produce more biomass. The biomass values were notably higher in regions with elevated SAVI levels, reinforcing the idea that healthier and denser vegetation contributes to increased biomass production. This observation underscores the significance of utilizing vegetation indices like SAVI as essential tools for monitoring pasture health and forecasting biomass yields.

#### Analysis of Biomass and Evapotranspiration (EVAPOT)

Lastly, I looked into how biomass relates to evapotranspiration (EVAPOT), which reflects the overall impact of evaporation and plant transpiration. The analysis of the scatter plot revealed a negative correlation, indicating that as evapotranspiration rises, biomass generally declines. This finding suggests that the loss of water through evaporation and transpiration may adversely affect pasture productivity, likely due to water stress hindering plant growth. To sustain biomass levels in dry conditions, it could be crucial to manage evapotranspiration through methods like irrigation or improving soil moisture retention.

#### Impact of Animals on Biomass Production

To understand how grazing affects biomass levels, I started by examining the distribution of sample types and their relationship with animal presence. The bar chart analysis showed that samples S1 and S2 were among the most used, even though no animals were present. This indicates that these samples may act as baseline or control sites, facilitating comparisons with paddocks that have animals. The spatial distribution analysis further revealed that areas with higher animal density had lower biomass levels, suggesting that grazing may negatively impact biomass accumulation. This trend indicates that grazing activities play a significant role in influencing biomass availability, likely due to the ongoing consumption of forage by livestock.

To delve deeper into the effects of grazing, I performed a comparative boxplot analysis of biomass levels in areas with and without animals. The results indicated that non-grazed areas had notably higher biomass values, while grazed areas showed lower and more consistent biomass levels, reflecting continuous forage use. These findings highlight the need for effective grazing management strategies, such as rotational grazing and controlled stocking density, to ensure pasture sustainability. Monitoring non-grazed areas like S1 and S2 can offer valuable insights into the natural growth potential of biomass, aiding in the optimization of pasture use and maintaining a balance between forage availability and animal nutrition.

## Best environment for forage

### Data Exploration and Analysis Approach

To determine the best conditions for forage, I adopted a systematic exploratory data analysis strategy that combined both visual and statistical techniques. The main steps in this process involved examining the distribution of key forage quality indicators—Neutral Detergent Fiber (NDF), Acid Detergent Fiber (ADF), and Crude Protein (CP)—and exploring their connections with environmental factors. Initially, I used boxplots to analyze the distribution of NDF, ADF, and CP across various sample locations. These visual tools were instrumental in spotting outliers, which play a crucial role in identifying areas with either high or low forage quality. Lower NDF and ADF values typically signify high-quality forage that is more digestible, while elevated CP levels indicate nutrient-rich forage that is advantageous for livestock. By assessing the spread and central tendency of these indicators, I was able to pinpoint key regions that show promising conditions for forage production. Subsequently, I performed correlation analysis with a heatmap to investigate the potential relationships between forage quality indicators and environmental variables such as temperature, wind speed, rainfall, and atmospheric pressure. This analysis shed light on how these factors affect forage quality, allowing me to identify the strongest and weakest correlations. Grasping these relationships is vital for making informed decisions about pasture management and optimizing environmental conditions to boost forage productivity.

### Insights from the boxplots

**NDF (Neutral Detergent Fiber)**

The analysis of NDF values showed that lower NDF levels are a sign of high-quality forage, as they relate to better digestibility and nutrient availability for livestock. The dataset revealed a significant variation in NDF values, with the lowest levels found in specific environmental conditions. Low NDF values were noted in areas with low rainfall (around 2.5-3.0 mm), moderate wind speeds (about 12.4 m/s), and stable atmospheric pressure. These conditions indicate that regions with less moisture and consistent airflow may produce forage with enhanced digestibility characteristics

**ADF (Acid Detergent Fiber)**

Higher values of ADF are known to reduce digestibility, hence making it difficult for livestock to obtain nutrients from such forage. On the other hand, locations with low ADF values indicated good quality and easy-to-digest forage. The analysis of data indicated that low values of ADF occurred at a temperature range of 30-32°C, rainfall about 0-2.5 mm, and wind speed at a moderate level. These findings indicate that areas with an even temperature, with lower moisture availability, can yield better digestibility of forage.

**CP (Crude Protein)**

Crude Protein (CP) is a key factor in assessing the quality of forages, with elevated CP values that add value to the nutritional status of livestock. Findings of the study indicated that high CP value areas coincided with high levels of biomass production where CP ranges were in excess of 12%, which is ideal. Meteorological conditions coinciding with high CP value areas are mean precipitation levels of 2.3-2.9 mm, lower temperature ranges of 21-28°C, and relatively homogeneous wind speeds. These conditions favor the production of proteinaceous forage where improved pasture yield and animal health are realized.

### Statistical Observations and Ranges

Forage quality measurement comparisons, specifically Neutral Detergent Fiber (NDF), Acid Detergent Fiber (ADF), and Crude Protein (CP), provided valuable insight into how each compares between various environments. Referring to filtered data in optimum environmental conditions, each measure had a specific range that was applied to determine the optimal condition for quality forage.

**NDF (Neutral Detergent Fiber) Range**

NDF values for the high-quality forage zones fluctuated from 36.27% to 50%, with a difference of 14.25%. These were recorded during periods of low precipitation (2.5-3.3 mm), moderate temperatures (31-34°C), and constant wind speeds (9.5-13.9 m/s). Lower NDF values were related to increased biomass production, meaning that forage in these zones is more digestible and of better quality.

**ADF (Acid Detergent Fiber) Range**

ADF percentages ranged from 37.93% to 54.12%, and total differences were 16.18%. Minimum ADF values showing quality forages were found in areas with lower rainfall (0-2.5 mm), moderate wind speed (12.4-21.8 m/s), and 30-32°C temperature. Above, we see that low ADF is associated with greater biomass yield and it is best to plant forage in those areas.

**CP (Crude Protein) Range**

Crude Protein (CP) content plays a significant role in feeding cattle and other animals suitably. Its range varied between 12.84% to 16.77% and had an aggregate difference of 3.93%. Where greater CP percentages occurred, locations of moderate precipitation levels (2.3-2.9 mm), approximately 21-28°C temperature ranges, and 7-15 m/s wind speeds existed. Such meteorological conditions assisted plants with greater protein content in development, giving elevated biomass percentages.

### Analysis of Temperature and Solar Radiation

In order to know how temperature and solar radiation are related, we did a special analysis founded on data groups in the dataset. We divided the biomass, temperature, and solar radiation values into different levels to allow us to discern patterns and relationships.

I divided them like this:

* Biomass into four categories: Very Low, Low, High, and Very High.
* Temperature (TEMP\_MAX) into three categories: Cold, Moderate, and Hot.
* Solar Radiation (RAD\_SOL) is classified into three types: Low, Moderate, and High.

And after I used a contingency table, this is the results I found:

**Dominance of Low and Moderate Solar Radiation:**

* Low and moderate solar radiation levels happened most often, 106 and 116 times, respectively. High solar radiation was seen the least, with only 90 times.
* Low solar radiation: Mostly linked to cold (52 times) and moderate (54 times) temperatures, and none in hot conditions.
* Moderate Solar Radiation: It occurs in Cold (40), Moderate (52), and Hot (24) temperatures, implying a healthy equilibrium.
* High Solar Radiation: This occurs primarily in Hot temperatures (74 times), rarely in Cold (16 times), and never in Moderate conditions.

**Distribution of Temperature Across Solar Radiation:**

* Cold occurred most frequently among the temperature categories, occurring 108 times at all solar radiation levels.
* Hot temperatures were the least common, appearing in 98 cases in total but dominating the High solar radiation category.

**Insights**

The comparison indicates solar radiation intensities to be closely correlated with temperature classes, and high solar radiation occurs most commonly under hot conditions. The relationship can be utilized in predicting forage growing conditions, enhancing pasture management, and planning grazing programs according to the anticipated temperature and radiation intensities.

## Comparison between measures I used in descriptive

|  |  |  |  |
| --- | --- | --- | --- |
| Category | Measures Used | Purpose | Key Insights |
| Measures of Frequency | Count, Frequency Distribution | Understand how often specific values occur. | Identified the most commonly occurring forage conditions across different sample types. |
| Measures of Central Tendency | Mean, Median | Determine the central value of key forage indicators (NDF, ADF, CP). | Provided a benchmark for typical forage quality and highlighted differences across conditions. |
| Measures of Dispersion | Range | Assess the spread of forage quality indicators. | Identified the variability in NDF, ADF, and CP values to determine forage consistency. |
| Outlier Detection | Box Plot | Identify and analyze extreme values in forage quality indicators. | Highlighted areas with significantly high or low forage quality based on key parameters. |
| Relationships Between Features | Contingency Table | Analyze relationships between categorical features like temperature and solar radiation. | Found strong associations between temperature categories and solar radiation levels. |
| Visualizations | Box Plot, Scatter Plot, Heatmap | Visualize variations and relationships between environmental factors and forage quality. | Helped identify trends and dependencies, such as the impact of environmental factors on biomass. |

## importance of data analytical techniques to the decision-making

Data analytics techniques are core in decision-making, where raw data is transformed into actionable insights. These techniques will enable the stakeholders to comprehend patterns, trends, and relationships within the data for more informed and strategic decisions. For instance, measures of central tendency such as mean and median give a clear indication of the typical values that the dataset contains, thus enabling the decision-maker to set benchmarks for forage quality. Measures of dispersion, such as range and variance, provide insight into the variability of data, which is important to understand the consistency of forage conditions across different environmental settings. The identification of outliers through box plots helps in the detection of anomalies that may indicate areas that need further investigation or intervention to ensure that resources are allocated effectively.

Besides, the analysis of the relationship through contingency tables, for temperature and solar radiation for example, allows the identification of environmental factors that influence biomass production. This information is useful in pasture management, such as determining the timing of grazing or adjusting fertilization plans to increase yields of forage. Visualizations, such as scatter plots and heatmaps, add much to making complicated data more understandable and provide an intuitive feel for the underlying patterns and correlations that are not obvious from numerical raw data.

In the end, data analytical techniques enhance the process of decision-making by laying down a structured approach to problem-solving and strategic planning. Whether it be understanding forage quality indicators, evaluating environmental impacts, or predicting future outcomes, these techniques empower stakeholders with data-driven decisions leading to better productivity and resource management. These analytical methods help the organizations minimize uncertainty, optimize operations, and achieve objectives with better efficiency.

# Predictive Analytics

In the initial half of this section, I spent my time on data preparation for modeling and analysis. I initially converted the date-time columns to proper datetime objects. This enabled me to work and analyze time-based data easily. This conversion helped in time-series analysis and determining trends in the data set.

Next, I used one-hot encoding for the categorical variables to convert them into numerical values that machine learning algorithms can use. One-hot encoding helped ensure that categorical variables were represented without establishing any order, which avoided possible biases in the model.

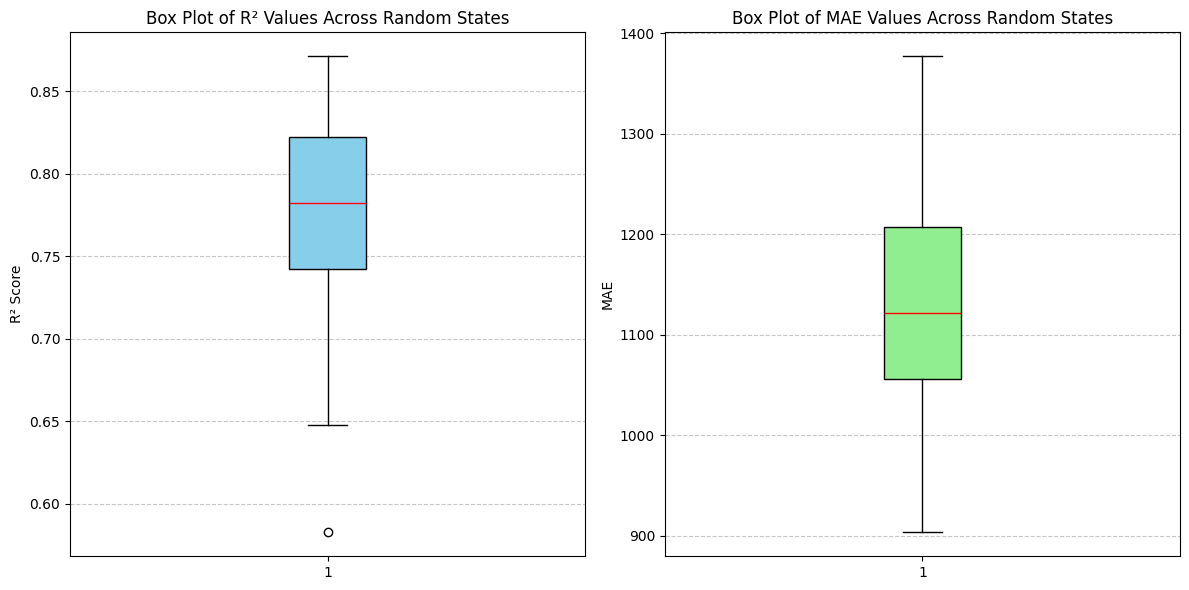
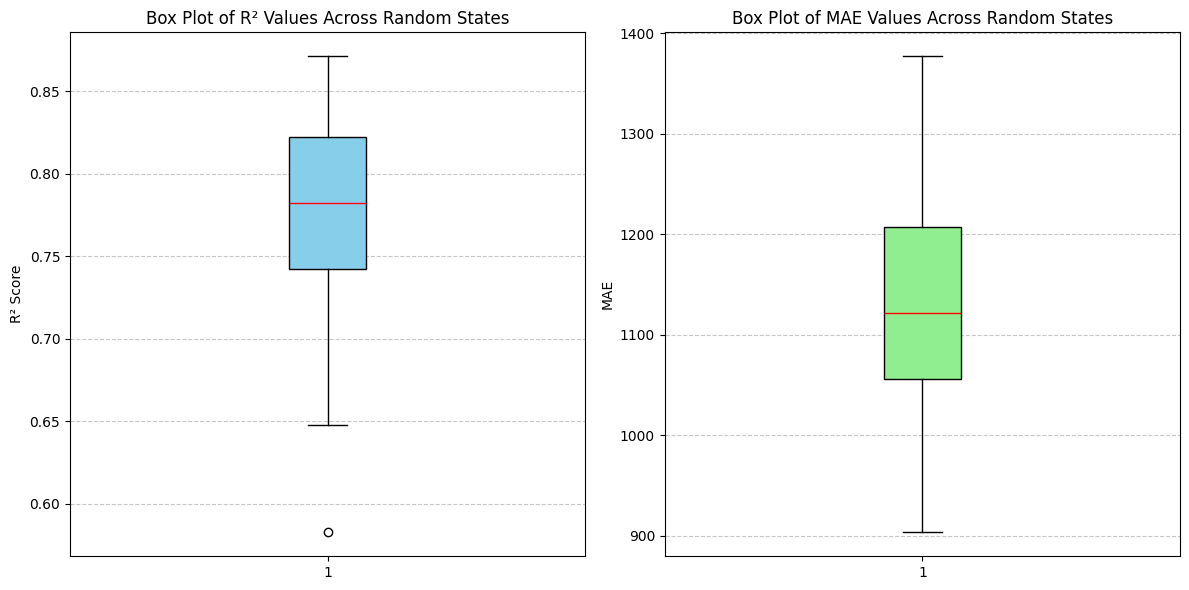
Once I had cleaned the data, I chose features to identify the most significant ones for the predictive model. I applied two primary approaches: Select K Best and Recursive Feature Elimination (RFE). These approaches allowed me to rank and choose the most significant features based on their statistical significance and model performance. Select K Best tests each feature separately with some statistical tests, whereas RFE removes less significant features one by one in an attempt to make the model more precise.

I also determined the optimal number of features required for optimal model performance. By experimenting with various sets of features individually, I was able to separate the most useful variables that improved the predictability of the model. This ensured that only the significant features were retained, which made the model easier and capable of performing its task in the optimum way.

## Models

**Linear regression:**

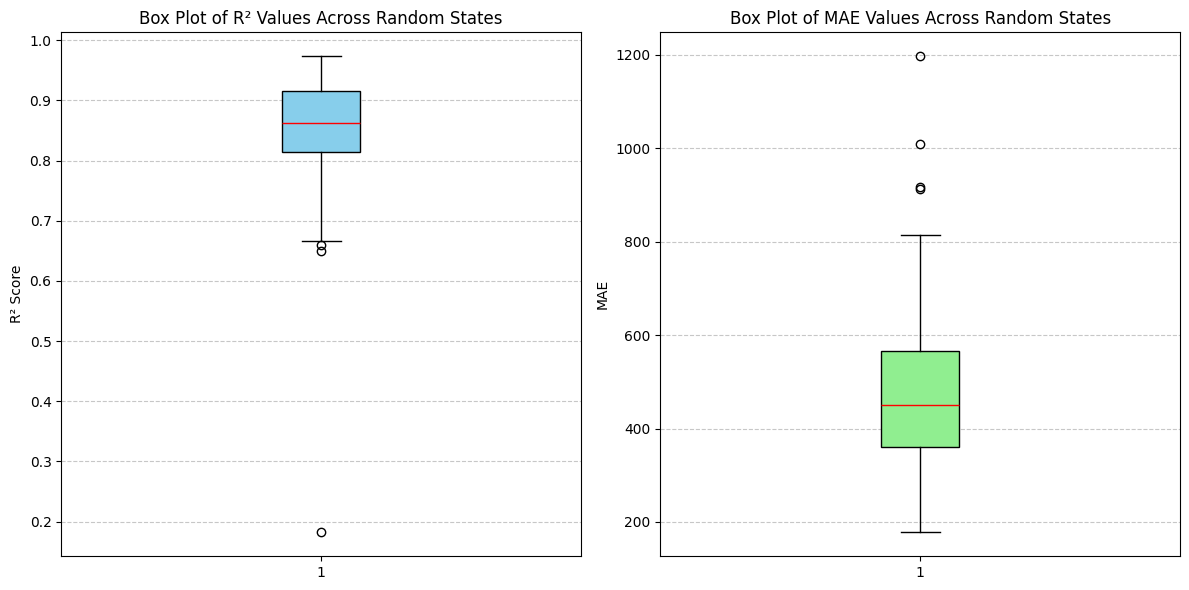
**With RFE:** **with SelectKBest:**

As we see here the results is two close to each other because linear regression I a simple model and both the features selection chose select a very close feature so that give us this result.

**Decision Tree:**

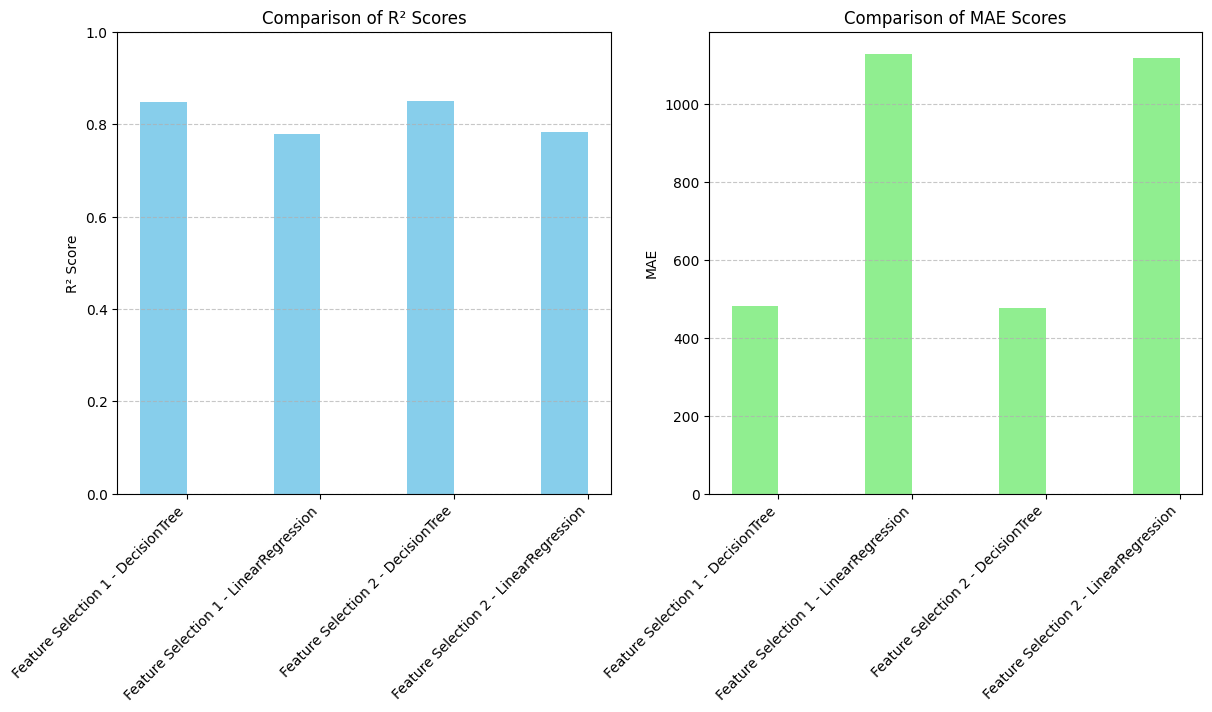
**With RFE: With Select k best:**

 A close-up of a graph

Description automatically generated

Here we can find defferent results with defferent random states because decision tree is more complicated than linear regression.

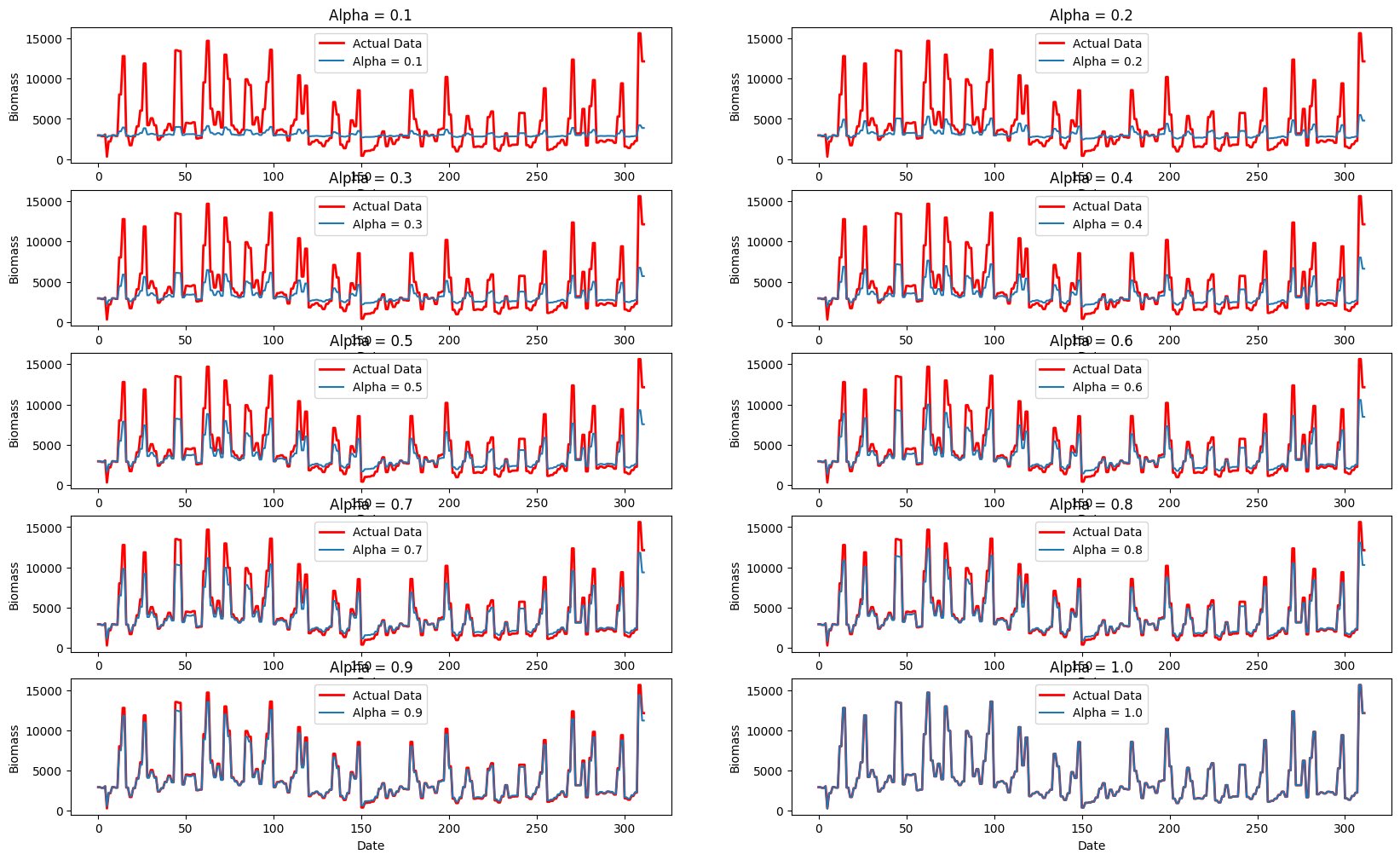
The analysis of model performance across different random states reveals significant variability in the results, highlighting the instability of the dataset. The box plots of R2 and Mean Absolute Error (MAE) measures show that while the median values show a good performance of the model, the large range of observed values along with the existence of outliers shows inconsistencies in the data distribution. It is an indication that there are most probably inherent differences in the dataset. differences in feature distribution or existence of noise, which affect the generalizability of the model. The variations in performance for different random states indicate that model predictions are extremely sensitive to how the data is split into training and test sets, and it is difficult to arrive at reliable conclusions. To mitigate this and improve the stability of the data, one should attempt methods like stratified sampling, feature scaling, and cautious feature selection. Besides, applying strict evaluation techniques, such as cross-validation, is able to mitigate the effect of data instability and provide a more stable model performance.



And According to the results the best results comes when I used select K Best with DcisionTree because it gives us lowest MAE and Highier R2

## Exponential Smoothing

I also used Exponential Smoothing for forecasting. and I try many alpha values to find the best one between them:



And the result is the best one is (Alpha = 1) and its give me (MAE=0 , MSE=0, R2=1).

## LSTM

Also I used LSTM and to give me the best results I add more layers and put the (epochs=1000)

Which give me (MAE=883, R2=0.77).

## Compare table

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model | Feature Selection Method | R² Score (Performance) | MAE (Error) | Strengths | Weaknesses |
| Linear Regression | RFE | Moderate | Moderate | Simple, interpretable, fast training | Sensitive to multicollinearity, assumes linearity |
| SelectKBest | Moderate | Moderate | Feature selection improves interpretability | Limited capability for complex relationships |
| Decision Tree | RFE | Varies across states | Varies | Handles complex relationships, interpretable | Prone to overfitting, sensitive to data splits |
| SelectKBest | Higher R², lower MAE | Better | Selects optimal features, improves performance | May still overfit if not pruned properly |
| Exponential Smoothing | Not Applicable | 1.00 | 0 | Perfect fit for trend forecasting, smooth predictions | Not suitable for non-trend data, may lag behind sudden changes |
| LSTM (Neural Network) | Not Applicable | 0.77 | 883 | Handles sequential data well, captures dependencies over time | Computationally expensive, requires tuning |