Individual Diploma Thesis

MACHINE LEARNING BASED MANAGEMENT AND STATISTICAL ANALYSIS OF DATA COLLECTED FROM A PLANT PHENOTYPING PLATFORM

Grigoris Michael

University of Cyprus

Department of Computer Science

May 2024

University of Cyprus

Department of Computer Science

Machine learning Based Management and Statistical Analysis of Data Collected from a Plant Phenotyping Platform

Grigoris Michael

Supervisor

Chryssis Georgiou

The Thesis was submitted in partial fulfillment of the requirements for obtaining the Computer Science degree of the Department of Computer Science of the University of Cyprus

May 2024

Acknowledgement

I want to express my deepest appreciation to Dr. Chryssis Georgiou, a Professor in the Department of Computer Science at the University of Cyprus, who not only entrusted me with this topic but also provided consistent guidance and support throughout my Diploma Thesis journey.

Additionally, I am profoundly thankful to the Plant Molecular Physiology Lab and Dr. Elias Bassil, Assistant Professor in the of Department of Biological Sciences at the University of Cyprus for generously providing the essential experimental data samples that formed the backbone of our research.

Lastly, I am deeply grateful to my parents and friends for their unwavering encouragement and support throughout this journey.

Abstract

The rise of machine learning, particularly with advancements in deep neural networks for tasks like image and language processing, has indeed ushered in a new era across various aspects of our lives. These developments have enabled remarkable progress in areas such as computer vision, natural language understanding, and speech recognition, among others.

While the emergence of deep neural networks has led to significant breakthroughs and superior performance in many tasks, it's important to acknowledge that simpler methods still have their place. Especially in the field of plant biology most of the analysis done using machine learning focuses on image classification.

In this Thesis, we believe that in this field there is room for other techniques to be useful in the analysis of data. We aim to establish a machine learning base methodology that makes use of clustering algorithms to give insides in the behavior of plants and the effect of the environment on them. To accomplish this, we used data taken from an experiment contacted by the Plant Molecular Physiology Lab of the Department of Biological Science of the University of Cyprus with objective to compare the growth and physiological effects of four seedling rootstocks of Avocado (Persea americana) to salinity.

The methodology we have concluded on consists of three main steps, *data cleaning*, *clustering*, and *cluster comparison*. The main idea, once the data gets cleaned and can be used, is to choose two parameters that we want to find more about their relationship and cluster the plants based on those two parameters separately. Once the clustering is done, we compare the resulting clusters from the two parameters to find the similarity they have. The advantage of this method is that the two parameters do not have to be homogenous, and we don't need to know the feature of the parameter that differentiates the plants.

This analysis on the experiment gave very promising results as it was evident from previous knowledge on the topic, and we are positive that it can be applied in other plant-based experiments that follow the same structure.

Contents

Chapter 1

Introduction

1.1 Motivation

The motivation behind this study is the fact that even though the machine learning is widely spread and used in plant biology, it is limited for the most part to image classification especially for disease identification [1][2], with no real use on data analysis. In this study the focus is on using other simpler machine learning methods, like clustering [3], to help us answer questions that are not obvious how to tackle with regular statistical methods but to also uncover new questions based on the results. Furthermore, clustering may generalize better for bigger and more complex datasets where statistical analysis cannot be efficiently used. The limits of statistical analysis in plant biology have driven us to search for better, more efficient, and more effective ways to analyze the data that result from this type of experiments and the rise of machine learning in the last years suggest that this is the correct first step forward.

1.2 Objective of the Study

The objective of this study is to establish a machine learning based methodology to primarily extract the relationship between the attributes of a plant with focus on the growth and how this relationship changes with respect to controlled outside parameters. As a first step we aim establish a simple methodology that makes use of clustering

algorithms which are by their nature fast, not very demanding in terms of computational resources and most importantly can be generalized very well and handle huge amounts of data. The data used in this study were taken from an experiment contacted by the Plant Molecular Physiology Lab. The goal of the experiment was to compare the growth and physiological effects of four seedling rootstocks of Avocado (Persea americana) to salinity. Differential responses to salt would then help examine the underlying molecular basis (gene networks that could be involved in salt stress acclimation). Eventually this information could be used to breed more salt tolerant rootstocks for areas where salinity limits avocado production.

1.3 Methodology

The methodology of this study can be seen in Figure 1.1

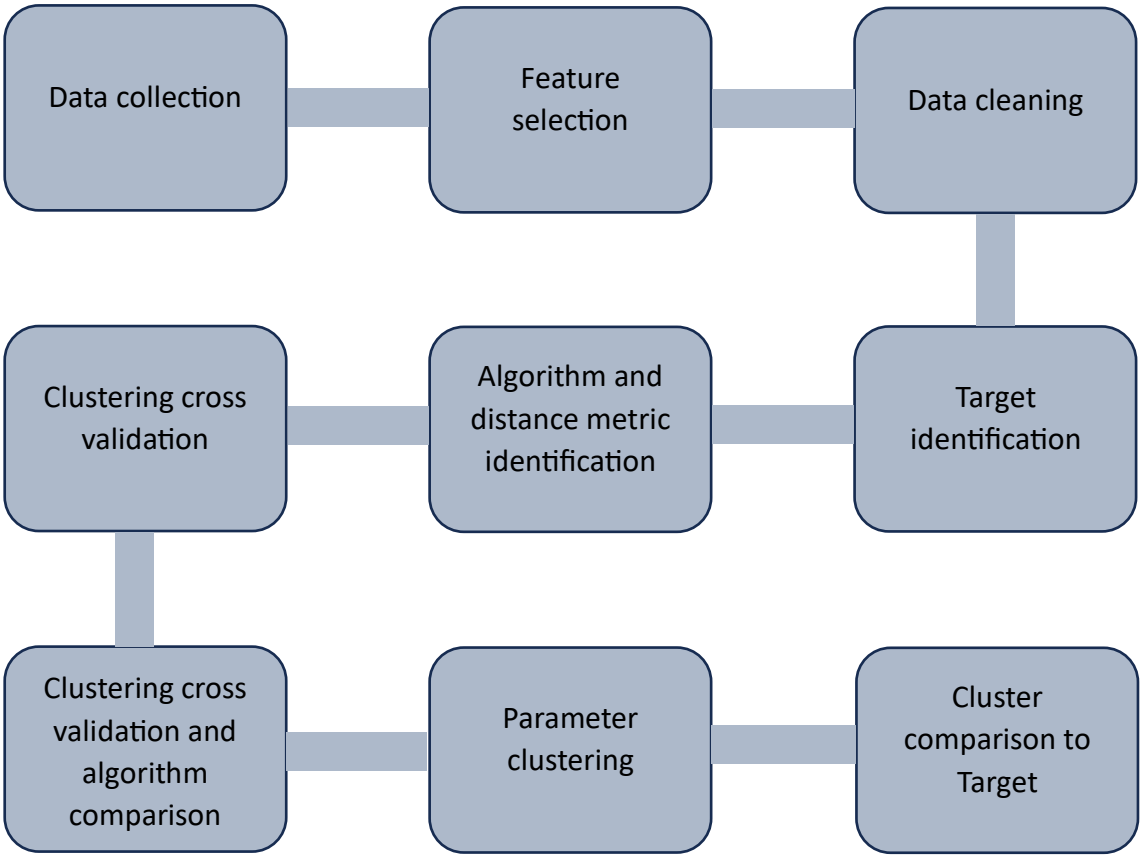

Figure 1.1: Methodology of Current Work

Firstly, the Plant Molecular Physiology Lab of the Department of Biological Science of the University of Cyprus contacted an experiment and gave us access to the data that were stored in an online platform named SPAC Analytics. Through the use of an API (Application Programming Interface) we were able to download the data and using by SQL lite to store it in a database for better handling of the data.

Afterwords it was necessary to identify the useful data because in many cases a large number of records were missing due to sensor failure. Anywhere the "damage" was small, the mean imputation [4] was used (the missing value was replaced with the average of the previous and the next value). In addition to that, many parameters appear to have the same behavior concerning the timeseries and they were identified and excluded using statistical correlation.

Once the data were cleaned, a target variable had to be chosen based on the questions that the experiment was trying to answer. In this particular case, the question was related to the growth of the plants, so the chosen target was a variable that shows the growth. Then we clustered the target variable with four different clustering algorithms (Section2.2) with two desirable clusters and compare them to get high confidence on the clustering. Furthermore, we investigated different types of distance metrics for the algorithms to find the best suited for this particular dataset. The resulting clusters separate the plants into two categories, with the first being those with high rate of growth and the second with low rate of growth. To the remaining parameters we followed the same procedure as the target variable in regard to clustering. In the end, after creating the clusters for every parameter we compared those clusters to the clusters of the target. This procedure gave us the information of which parameters affect the growth of the plants and how much they do so.

We also needed to compare the clustering of the target to the natural "clustering" we get from the experiment regarding the combination of the two groups (control and treatment) and the four cultivars. More specifically we need to compare the growth of every cultivar with respect to the others and with respect to the same cultivar with salinity applied. Because the growth can be approximated linearly and the feature of the

timeseries of the target that shows the growth is the rate of change (slope of the timeseries) we can calculate it easily with the use of statistical analysis. We used four different methods to approximate the rate of change of the growth to gain more confidence on the results and subsequently we compared the results of every cultivar to the rest in order to gain information on the behavior of every cultivar.

Finaly, we used the clustering method to compare the clustering of the target to the combination of the two groups (control and treatment) and the four cultivar and we compared the result from this method to the statistical analysis.

1.4 Thesis Organization

The rest of this thesis is split into four chapters. Table 1.1 reports the content of each chapter.

Chapter Number	Chapter Description
2	We give an overview of the parameters used in the experiment and a brief description of the algorithms and distance metrics used in the study.
3	We explain the structure of the experiment and the methods used for data cleaning and feature extraction
4	Explains the similarities and difference of the algorithms used with respect to the data used in this study through examples of the results. Furthermore, in the chapter we discuss some results taken from the work done on the primary objective of the study and subsequently, we compare a statistical method and the clustering method on the second objective of the study
5	We give a summary of the study and some key points for future research

Table 1.1: Document Organization

Chapter 2

Background Knowledge

2.1 About the Biological Experiment

This thesis makes use of data taken from an experiment contacted by the Plant Molecular Physiology Lab of the Department of Biological Sciences at the University of Cyprus. The goal of the experiment was to answer some questions regarding the effect of salinity in the growth of different cultivars of avocado. The four cultivars used are Mexicola, Hass, Donnie and Weldin and for each cultivar eight (8) plants were used. Four (4) plants from each cultivar were placed in the control group and the other four in the treatment group resulting in sixteen (16) plants in each group for a total of thirty-two (32) plants. In Figure 2.1 we can see a diagram of the structure of the experiment in relation to the groups and cultivars. The experiment is divided into three phases. In phase 1, the plants were free to grow in a controlled environment. In phase 2, salinity was added to the treatment group, and nothing changed for the control group. In phase 3, the plants in the treatment group were cleaned from the salinity. While the experiment was taking place, some measurement devices tracked many key parameters and based on those the system derived more data. The dataset we were given consists

of a time series for every parameter for every plant. Most of the parameters were measured every three minutes and the rest once a day.

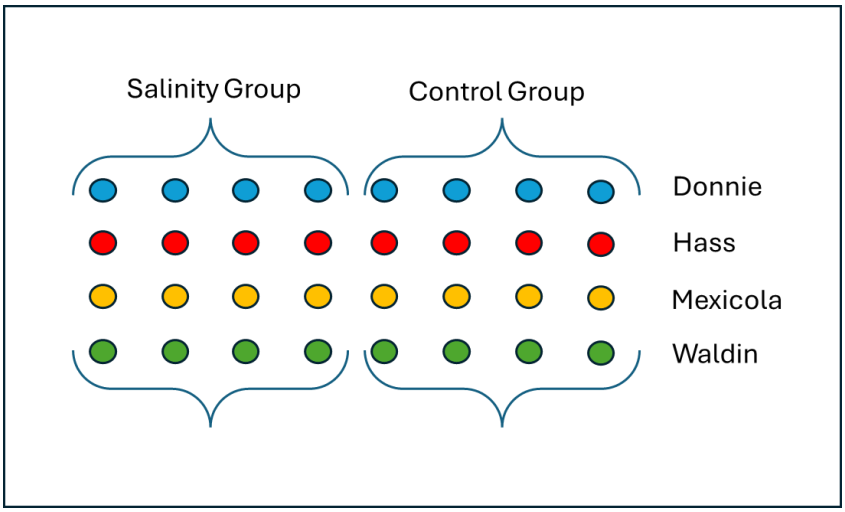

Figure 2.1: Experiment Diagram of Cultivars and Groups

In Table 2.1 we can find information of the parameters measured during the experiment. The description column briefly describes the parameter, the Value column gives the reference value of the parameter in the database (identification key) and the Units column gives the physical units in which the parameter is expressed in.

Name	Description	Value	Units
Weight	The raw plant's gross weight	s4	g
Weight Smooth	The plant's smoothed gross weight using Savizky-Golay algorithm	ws	g
Weight Normalized	The raw plant's gross weight divided by the Plant's gross weight on the first day	wn	g/g
Transpiration Rate	The rate at which the plant loses water through transpiration	tr	g/min
E	The transpiration normalized to plant weight	e	$g_{water}/g_{plant}/min$

Gs canopy (Watchdog)	Canopy conductance calculated as transpiration rate divided by vapor pressure deficit (VPD)	g2s	g_{water}/min
Gs canopy per weight (Watchdog)	Canopy conductance normalized to the plant weight, calculated as transpiration rate divided by VPD	gs	g_{water}/g_{plant}/min
Calculated VWC	Soil volumetric water content calculated from soil sensor measurements	cvwc	cm^3/cm^3
RDT Morning	Time reference of the first measurement for daily transpiration in the morning	rdtm	g
RDT Evening	Time reference of the second measurement for daily transpiration in the evening	rdte	g
Daily Transpiration	Daily mass difference between two time points, indicating the total water loss through transpiration during the day	dt	g
Normalized Daily Transpiration	Daily transpiration divided by the calculated plant weight, providing a normalized measure of water loss.	ndt	g_{water} /g_{plant}
Plant Growth	The difference in mass between consecutive days after applying water into soil capacity, indicating the increase in plant mass over time.	pg	g
Plant net Weight	The plant's weight after subtracting fixed pre-experiment measurements, giving the net	pnw	g

	change in plant weight during the experiment.		
Calculated Plant Weight	The plant net weight calculated by the water use efficiency (WUE) method, providing an estimate of the plant's mass	cpw	g
Plant Water Recharge	The daily mass difference between two time points, indicating the amount of water absorbed by the plant from the soil.	pwr	g
Analog P	-	analogp	-
Weather station PARLight	Measurement of Photosynthetically Active Radiation (PAR) from a weather station, indicating the intensity of light available for photosynthesis	wspar	µmol/m^2/s
Weather station RH	Relative humidity measurement from a weather station, indicating the amount of moisture in the air.	wsrh	%
Weather station Temp	Temperature measurement from a weather station, indicating the ambient temperature.	wstemp	°C
Weather Station VPD	Vapor Pressure Deficit (VPD) measurement from a weather station, indicating the difference between the amount of moisture in the air and its saturation point	vpd	kPa

Weather Station VPD smooth	Smoothed curve of Vapor Pressure Deficit (VPD) measured by a weather station.	vpd|sg	kPa
DIELEC/0/5TE	the raw output of the soil sensor, is used to calculate the volumetric soil water content	s10	Unitless
VWC (DIELEC/0/5TE)	Volumetric water content measured by a DIELEC/0/5TE sensor, indicating the amount of water present in the soil	vwc-10	cm^3/cm^3
VWC Smooth (DIELEC/0/5TE)	Smoothed curve of volumetric water content measured by a DIELEC/0/5TE sensor	vwcs-10	cm^3/cm^3
Influx (DIELEC/0/5TE)	Time derivative of change of volumetric water content from the soil into the plant.	inf-10	g/min
Plant water balance (DIELEC/0/5TE)	Difference between the water influx and outflux measured by a DIELEC/0/5TE sensor, indicating the overall water balance	pwc-10	g/min
Daily Influx (DIELEC/0/5TE)	Daily mass difference indicating the total water influx into the plant-soil system measured by a DIELEC/0/5TE sensor	dinf-10	g
EC/0/5TE	Electrical conductivity indicating the concentration of ions in the soil solution	s6	dS/m
Temperature/0/5TE	Temperature of the surrounding environment.	s141	°c

Table 2.1: Explanation of Parameters

2.2 Machine Learning Algorithms

Machine learning has a very broad range of different techniques and algorithms [5], and each one is suited best for a particular type of data. In general Machine Learning can be broken down into three main types. The first type is known as *supervised machine learning* [6], where each record in the data has its own label (which represent the class or the expected outcome of that data point) and the end goal of a supervised machine learning model is to predict the label of new unseen data. *Unsupervised machine learning* is the second type. In this type the data are not associated with labels and the goal is to find similarities or differences between them and that is usually done by clustering [3]. The third type of machine learning is *reinforcement learning* [7], in which the model is treated as an agent that performs an action and gets rewarded or punished depending on if the action brings the agent closer to the end goal or further away. In this thesis we are using the second type of machine learning, the unsupervised learning and more specifically we focus on four different clustering algorithms. A clustering algorithm [3][8], tries to group every data point in different sets based on the distance they have from other data points. These sets are called clusters and some algorithms let the user define the number of clusters and some do not. In this thesis we focus only on algorithms of the first kind. In Figure 2.2 we can see the different types of machine learning in a hierarchical tree-like structure.

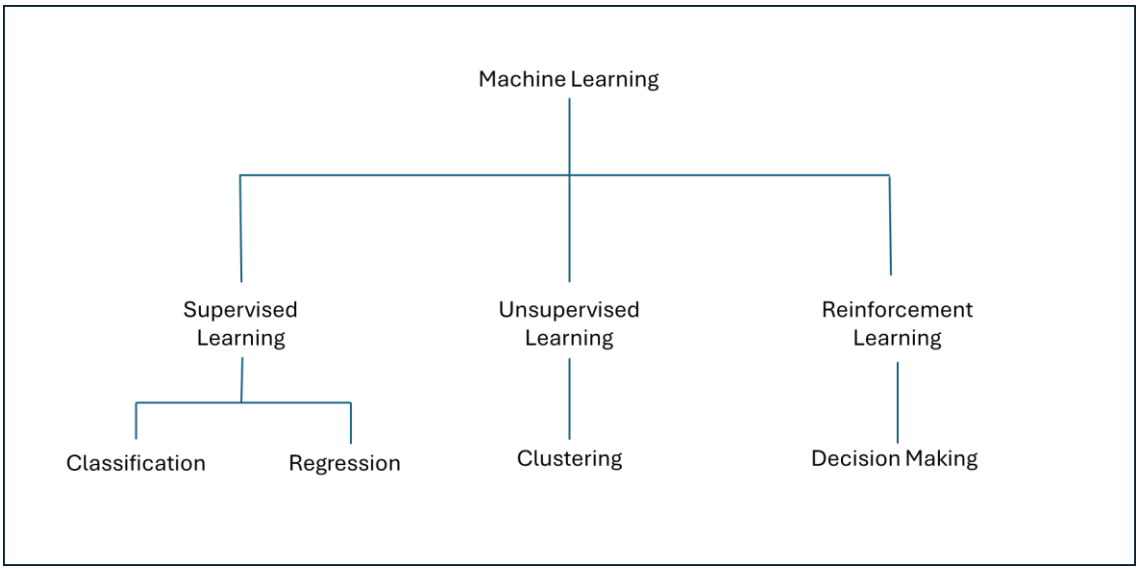

Figure 2.2: Machine Learning Diagram

2.2.1 Agglomerative Hierarchical Clustering

The agglomerative hierarchical clustering [9], is a bottom-up algorithm that creates a tree-like hierarchical structure of clusters. It starts by assigning a cluster to each datapoint in the datasets and calculates the distance of every pair of points. The choice of the distance metric depends on the nature of the data and the problem. Once all the distances are known the algorithm combines the two clusters with the minimum distance into a new cluster. This procedure is repeated until all the data points belong to a single cluster or a condition is met. When the clusters have more than two points to determine the distance between clusters the algorithm uses a linkage criterion. An example of such a criterion is defined as the minimum distance between any single point in the first cluster and any single point in the second cluster. In Figure 2.3, we can see the formation of the clusters with a bottom-up approach and the tree-like structure that emerges from the clustering process. The last step after the creation of the tree-like structure is the choice of the number of clusters (in this case two).

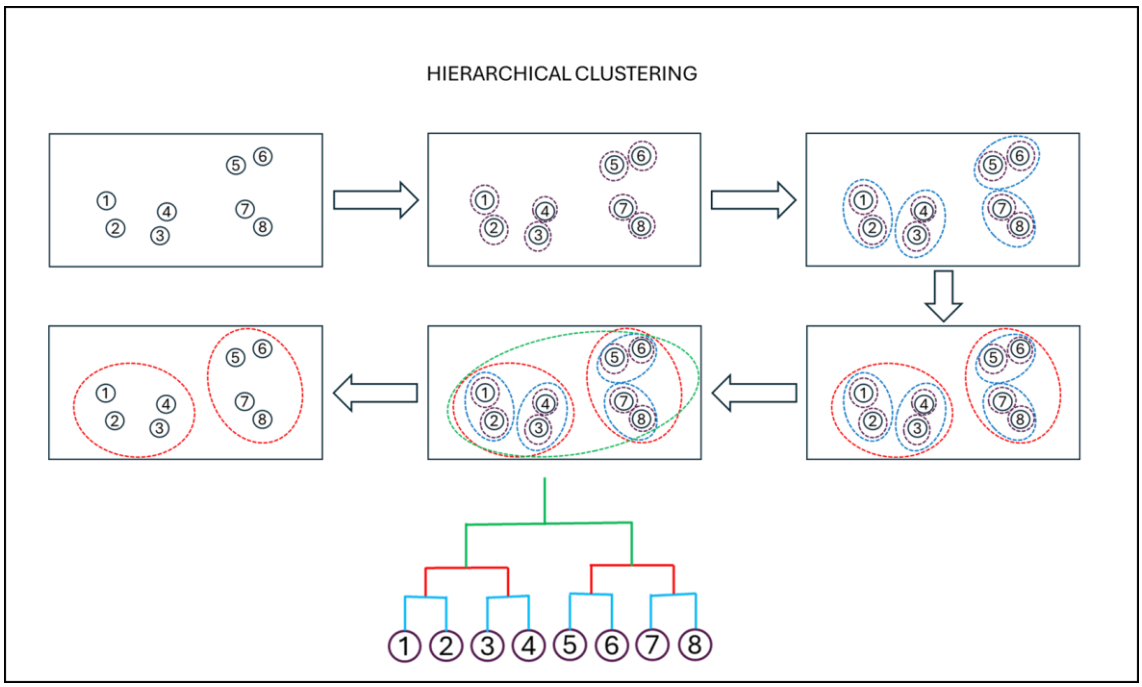

Figure 2.3: Hierarchical Clustering Explanation

2.2.2 K-Means

The k-means algorithm [10], tries to separate the datapoints into k clusters. Initially the algorithm chooses k centroids randomly selected from the data point. These centroids represent the initial centers of the k clusters. Once the centroids have been established the rest of the data points are assigned to the nearest cluster based on the distance from the point to the centroid of the cluster. After all the points are assigned to a cluster the new centroid of each cluster is calculated as the mean of all the point in the cluster. This process of assigning the data to the nearest cluster and calculating the new centroid ends when the new centroid is the same or very close to the previous one for every cluster, at this point is said that the algorithm has converged. In Figure 2.4 we get a visual representation of the clustering process of K-means algorithm through a simple example. We can see that the randomly chosen centers in the first step created two clusters which changed in the second step from the calculation of the new centers. After the third step there are no more changes in the clusters which means that the clustering process has converged, and the algorithm terminates. The initial choice of clusters was two.

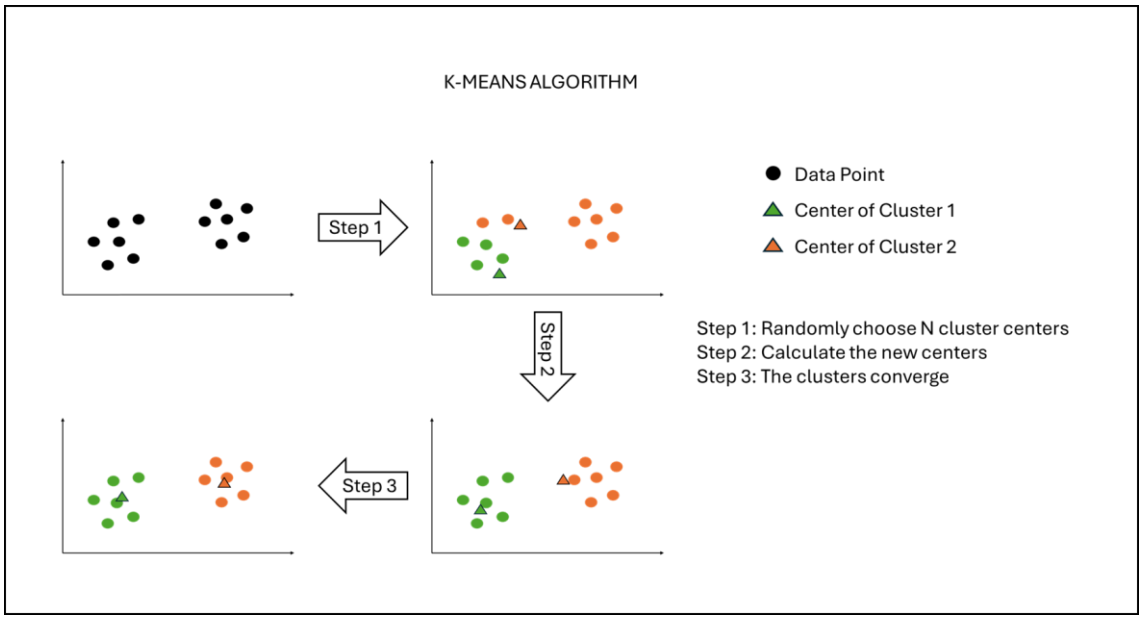

Figure 2.4: K-Means Algorithm Explanation

2.2.3 K-Means Time Series

The K-means Time Series algorithm is a type of k-means clustering suitable for timeseries data. It works in the same way that the k-means algorithm works with the only difference being the preprocessing of the data where each timeseries in the dataset is calculated by computing the difference between consecutive points. Effectively performing the k-means algorithm on the derivative of the original timeseries. In Figure 2.5 we have two examples of timeseries clustering. In the first example we have six timeseries with a sin wave like form and three of those have similar frequencies and the other three have different frequencies from the first ones and similar to each other. The result of the algorithm is the separation of the timeseries into two clusters with the feature in question being the frequency. In the second example we have two groups of linear equations with the main difference being the slope and the algorithm manages to create two clusters based on the different slopes

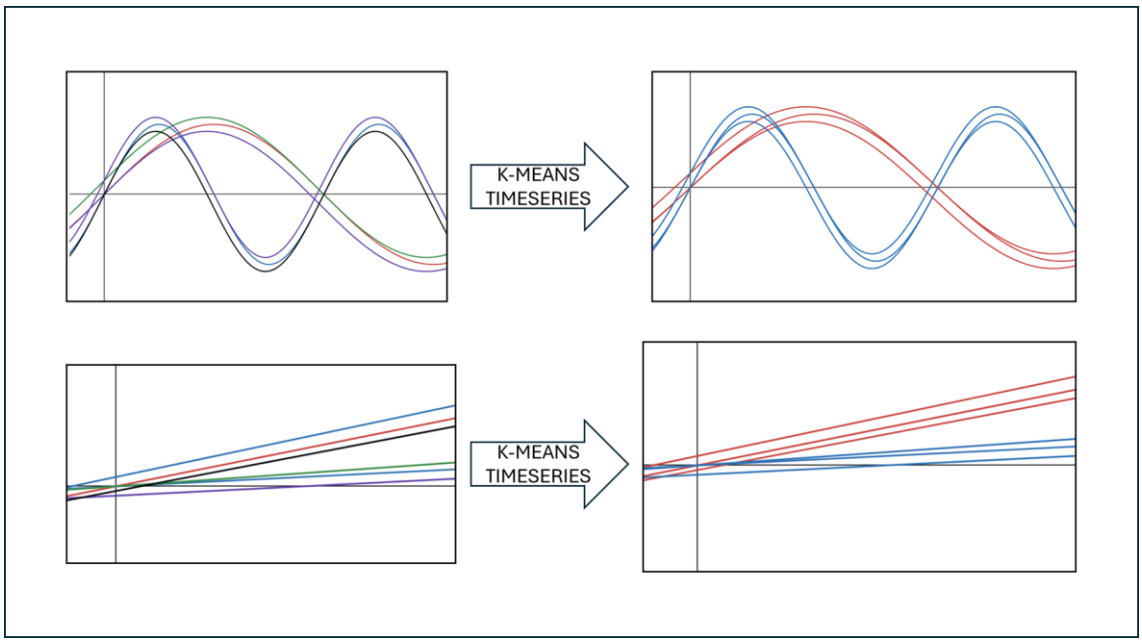

Figure 2.5: Timeseries Clustering Examples

2.2.4 K-Medoids

The k-medoids algorithm [12], is a variant of the k-means algorithm designed to be more effective than k-means in datasets that have noise. It has the two basic steps that k-means has where it assigns all points to the cluster with the closes center and then it calculates a new center for each cluster. The difference lays in the way the new center is calculated. In k-medoids the centers of the clusters are called medoids and they are calculated in the following way, for each point in the cluster the distance of that and every other point is calculated and is summed. The new medoid is the point with the smaller sum. This algorithm works in a similar way to k-means but is said to have better initial conditions due to the choice of medoids from the dataset [13][14]. In Figure 2.6 is shown the difference between the medoids in the k-medoids algorithm and the means (centers) in k-means algorithm. We can see that the medoids must be part of the dataset but the means in k-means have more freedom which makes them more sensitive outliers.

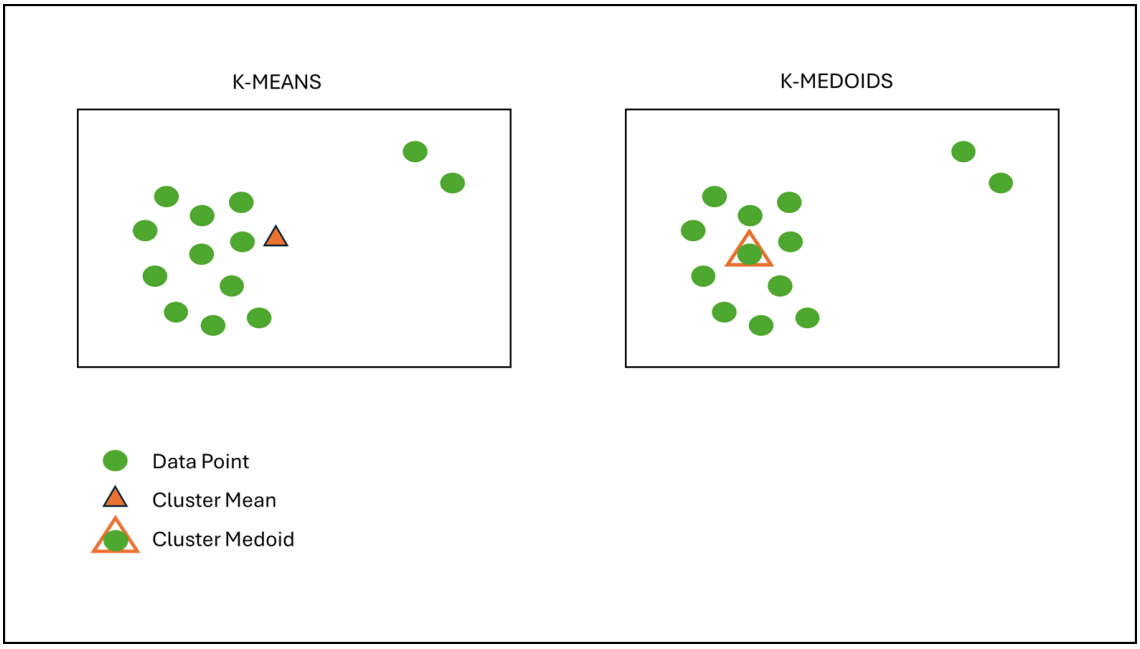

Figure 2.6: Cluster Mean and Cluster Medoid Difference

2.2.5 Distance Metrics for Time Series

Distance metrics [17], is the formula for the distance used to determine the similarity or difference of two points in the dataset or in this case of two time series. The three main distance metrics are Euclidian Distance, Dynamic Time Warping [15], Shape-based Distance [18]. The Euclidian distance calculates the Euclidian distance of every point of the timeseries to the corresponding point of the second timeseries and adding them all. The Dynamic time warping measures the similarity of the two timeseries in nonlinear making it suitable for timeseries with different duration or timing. the Shape-based Distance measure the similarity of the timeseries based on the similarity in the shape of the graphs.

After testing the clustering that results from the three different distances, we concluded that they all give the same result, unlike [16], for this particular set of data so in order to keep things simple we use the Euclidian Distance. In Figure 2.7, we can see two similar timeseries and how their similarity is measured using Euclidian distance and Dynamic Time Warping, notice that Euclidian distance is stricter in the comparison of individual moments and Dynamic Time Warping is more flexible making it better for finding shape-based similarities in uneven timeseries.

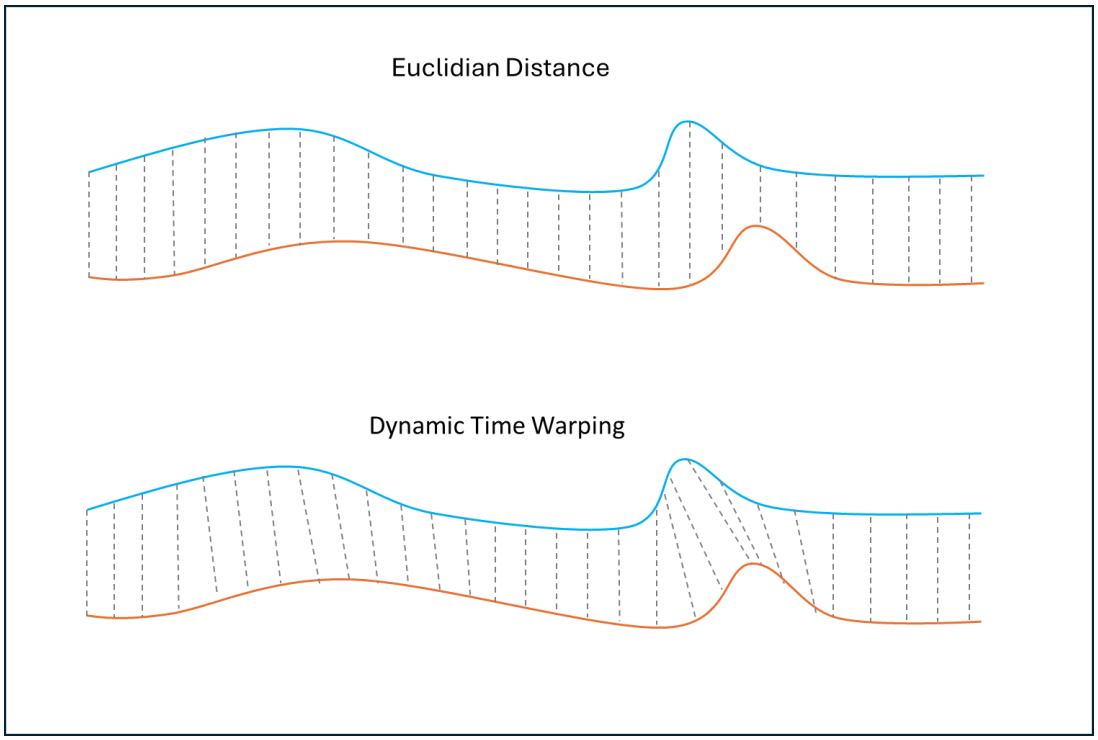

Figure 2.7: Euclidian and DTW Difference

Chapter 3

Data Collection

3.1 Obtaining and Managing the Data

The data the Plant Molecular Physiology Lab collected from this experiment and other experiments of the same nature are managed through an online application called SPAC analytics [22] and stored in their server. They were kind enough to give us access to the data through an API. In order to fetch the data, we developed a simple program that made use of the “request” library in Python.

The procedure was very straightforward. At first, we had to create an identification token to verify the validity of our request. Once we had the token, it was a matter of choosing the desirable experiment that the Plant Molecular Physiology Lab requested. To be consistent with the tool the Plant Molecular Physiology Lab uses we had to use the same identification codes of the parameters and names of the plants. We had to create a local database using SQL lite through Python to better manage and store the data of that experiment. This gave us flexibility on the combinations of plants and parameters we wanted to investigate with the use of machine learning and made the cleaning of the data easier.

3.2 Target Identification and Objective

The objective we are focusing on in this study is understanding the effect of salinity on the growth of plants with respect to the way the growth is affected by or affects other parameters. A secondary objective is understanding how different cultivars of avocado behave in an environment with increased salinity. As target or Target variable in the context of the methodology we will refer to the main parameter that we want to see the relation with the other parameters. Based on the objective of the experiment we concluded that the appropriate target variable is the net weight of the plant as it is a clear indication of the growth.

3.3 Feature Selection

Feature selection in machine learning is the process of selecting a subset of relevant features for use in model construction. This helps improve model performance by reducing overfitting, enhancing model interpretability, and decreasing training time. The Plant Molecular Physiology Lab, while contacting the experiment kept track of thirty (30) parameters including the ones that were derived. Between the thirty (30) parameters some of them have very high similarity to each other due to their derivation. With the use of statistical correlations, we identified which have more than 90% correlation and put them aside (as these parameters would give the same results with each other in the clustering). In Table 2.2 we can find the parameters and the relationship they have with other parameters in the same or in different plant.

Parameter name	Correlation type
"Weight" "Weight_Smooth" "Weight_Normalized"	High correlation in the same plant
"Transpiration_Rate" "E"	High correlation in the same plant

"Gs_canopy_per_weight_Watchdog"	
"Gs_canopy_Watchdog"	Problem with fetching from the server
"Analog_P"	Does not exist for most plants
"Weather_Station_PARLight" "Weather_Station_RH" "Weather_Station_Temp" "Weather_Station_VPD" "Weather_Station_VPD_smooth" "Temperature_0_5TE"	High correlation between plants
"DIELEC_0_5TE" "VWC_DIELEC_0_5TE" "VWC_Smooth_DIELEC_0_5TE"	High correlation in the same plant
"Calculated_VWC" "Influx_DIELEC_0_5TE" "Plant_water_balanc_DIELEC_0_5TE" "EC_0_5TE"	No correlation in the same plant or between plants
"RDT_Morning" "Plant_net_Weight" "Calculated_Plant_Weight"	High correlation in the same plant
"Daily_Transpiration" "Normalized_Daily_Transpiration"	High correlation in the same plant
"Plant_Growth", "Plant_Water_Recharge", "Daily_Influx_DIELEC_0_5TE" "RDT_Evening",	No correlation in the same plant or between plants

Table 2.2: Correlation of Parameters

Parameters exhibiting high correlation between plants were excluded from the clustering process as there were no discernible differences between plants for these

parameters. This exclusion was necessary as clustering would not provide meaningful insights when there is little variability across plants.

Similarly, for parameters showing high correlation within the same plant, only one of them was utilized in the clustering process. Using both parameters would essentially yield identical results, as they capture the same underlying information. By selecting only one of these highly correlated parameters, we avoid redundancy in the clustering analysis. Unfortunately, we lost an additional two parameters due to problems with the fetching from the server or because the values did not exist for the majority of the plants. The parameters that were eventually in the clustering can be seen in Table 2.3

Parameter group	Frequency of measurement	Parameter Name
Weight	3 minutes	"Weight" "Transpiration_Rate" "Gs_canopy_per_weight_Watchdog" “Calculated_VWC”
Weight	1 Day	"RDT_Morning" "RDT_Evening" "Daily_Transpiration" "Plant_Growth" "Plant_Water_Recharge"
Weather	3 minutes	None
Weather	1 Day	None
0/5TE	3 minutes	"DIELEC_0_5TE","Influx_DIELEC_0_5TE" "Plant_water_balanc_DIELEC_0_5TE" "EC_0_5TE"
0/5TE	1 Day	"Daily_Influx_DIELEC_0_5TE"

Table 2.3: Parameters used for Clustering.

3.4 Data Cleaning

Data cleaning in machine learning is the process of identifying and correcting or removing errors and inconsistencies in data to improve its quality and make it suitable for analysis. The raw data we obtained could not be utilized by the algorithms due to the presence of missing values. Many timestamps lacked values due to sensor failures, and some parameters exhibited illogical values, such as negative weights, necessitating data cleaning.

For timestamps with a small number of missing values, the gaps were filled using the average of the preceding and following values. When a parameter had a large number of missing values, the affected time series for that parameter was ignored for the duration of the problem. Similarly, time series with illogical values were also disregarded. Notably, the first nine to ten days of data were excluded for every parameter across all plants due to prevalent illogical values at the start of the experiment.

When comparing two parameters, only the intersecting valid data of both parameters was used, further reducing the usable dataset. As a result, from an initial dataset of thirty-two plants and thirty parameters with three-month time series, the effective data was significantly reduced. On average, we ended up with usable data from sixteen to twenty plants per parameter pair, thirteen useful parameters, and fifty-six days of time series data for each parameter. In Figure 3.1 we can see a simple diagram of the problems we encountered with the data and the solutions we used.

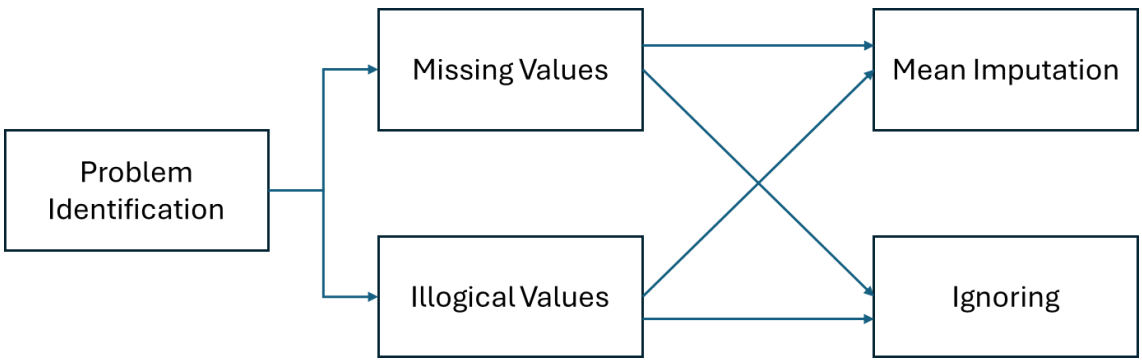

Figure 3.1: Data Problems and Solutions Diagram

Chapter 4

Clustering

Clustering algorithms are best suited for this type of analysis, as they inherently uncover patterns in the data that may not be easily detectable through traditional statistical methods. Additionally, the dataset is not homogeneous; some parameters have values recorded every three minutes, while others are recorded once a day, making direct comparisons difficult. Furthermore, the useful features of each time series for every parameter are unknown.

The main idea of the methodology is to cluster the two parameters we are interested in finding more about their relationship for every plant independently and then compare the resulting clusters. This comparison will yield a statistical representation of the relationship of the two parameters. The primary objective of the experiment requires only two clusters for the target variable because we only need to separate the plants based on the growth they had and to keep things simple we will to the same for the rest of the parameters and we will use the names high and low to distinguish the clusters.

4.1 Algorithm Comparison and Confidence

It is very important to be confident in the clusters that result from the algorithms, so we used four different algorithms. The algorithms we focus on are agglomerative, k-means, k-means Time Series and k-medoids, discussed in Section 2.2. The comparison of the algorithms in a practical sense was done by clustering four parameters with the four different algorithms and comparing the resulting clusters for each parameter. The four parameters consist of the target variable for the primary objective and three randomly chosen.

For the results we will focus on the target variable. The resulting clusters from the four different algorithms were mostly the same with 100% agreement on one of the two clusters, typically the low cluster. The agglomerative was the strictest regarding the formation of the high cluster, the rest were giving very similar results for both low and high clusters.

Tables 4.1, 4.2, 4.3, each show the similarity in the clustering for every pair of the four algorithms in that group. For the creation of these tables, we used the timeseries of all the plant. Specifically, the H in the cells represent the high cluster and the L the low cluster. Every cell gives the similarity of the algorithm on the row to the algorithm on the column. For example, the second cell in the first row in Table 4.1 gives the following information. 100% of the plants in the high cluster (H) of the agglomerative algorithm are also in the high cluster (H) of the k-means algorithm and 84% of the plants in the low cluster (L) of the agglomerative algorithm are also in the low cluster (L) of the k-means algorithm. Table 4.1 shows the similarity for the group that includes every plant, Table 4.2 shows the similarity for the control group and Table 4.3 shows the similarity for the treatment group (salinity group). The all-plants group contains twenty-seven (27) plants, the salinity group contains fourteen (14) plants and the control group thirteen (13) plants.

	Agglomerative	K-Means	K-Means Time Series	K-Medoids
Agglomerative	H:100% L:100%	H:100% L:84%	H:100% L:92%	H:100% L:60%
K-Means	H:63.64% L:100%	H:100% L:100%	H:81.82% L:100%	H:100% L:71.43%
K-Means Time Series	H:77.78% L:100%	H:100% L:91.30%	H:100% L:100%	H:100% L:65.22%
K-Medoids	H:41.18% L:100%	H:64.71% L:100%	H:52.94% L:100%	H:100% L:100%

Table 4.1: Similarity of Clusters for All Plants

	Agglomerative	K-Means	K-Means Time Series	K-Medoids
Agglomerative	H:100% L:100%	H:100% L:83.33%	H:100% L:83.33%	H:100% L:66.67%
K-Means	H:66.67% L:100%	H:100% L:100%	H:100% L:100%	H:100% L:80%
K-Means Time Series	H:66.67% L:100%	H:100% L:100%	H:100% L:100%	H:100% L:80%
K-Medoids	H:50% L:100%	H:75% L:100%	H:100% L:75%	H:100% L:100%

Table 4.2: Similarity of Clusters for Control Group

	Agglomerative	K-Means	K-Means Time Series	K-Medoids
Agglomerative	H:100% L:100%	H:100% L:84.62%	H:100% L:84.62%	H:100% L:53.85%
K-Means	H:60% L:100%	H:100% L:100%	H:100% L:100%	H:100% L:63.64%
K-Means Time Series	H:60% L:100%	H:100% L:100%	H:100% L:100%	H:100% L:63.64%
K-Medoids	H:66.67% L:100%	H:55.56% L:100%	H:55.56% L:100%	H:100% L:100%

Table 4.3: Similarity of Clusters for Salinity Group

	All plants	Control	Salinity
Agglomerative	H:100% L:92%	H:100% L:77.77%	H:100% L:74.36%
K-Means	H:81.82% L:90.47%	H:88.89% L:93.33%	H:86.6% L:87.88%
K-Means Time Series	H:92.59% L:97.1%	H:88.89% L:93.33%	H:86.6% L:87.88%
K-Medoids	H:52.91% L:100%	H:75% L:91.6%	H:59.26% L:100%

Table 4.4: Average Similarity for Each Group

From Table 4.1 we can see that the agglomerative algorithm has 100% similarity on the high cluster with the rest of the algorithms meaning that the agglomerative is the strictest regarding the high cluster. Similarly, the k-medoids algorithm has 100% similarity on the low cluster with the rest meaning that this algorithm is the strictest regarding the low cluster. The k-means and k-means Time Series seem to be somewhere in the middle in terms of similarity. As we can see from Table 4.4 on average the k-means Time Series is more similar with the rest of the algorithms than the k-means in both high and low clusters. When comparing Tables 4.2 and 4.3 we can see that in almost all pairs in the control tables (Table 4.2) the similarity is greater than the same pair in the salinity table (Table 4.3). In Sections 4.2 and 4.3 the clustering was done by all four algorithms, but we will focus on the results of k-means Time Series because the resulting clusters are more similar to the rest of the algorithms. In other words, k-means Times series is in the middle of the spectrum that defines the strictness of an algorithm.

4.2 Clustering with Respect to Features

The primary objective of the experiment was to understand the effect that the measured parameters have on the growth of the plants. To achieve this, we used the four algorithms mentioned in Section 2.2 to cluster the relevant parameters that resulted from the data selection at Section 3.3. The number of clusters was chosen to be two because the necessary clusters for the primary objective required two clusters for the target (high and low). One more reason for the choice of only two clusters is the fact that the small volume of data we had did not allow for more clusters. Furthermore, because the experiment had three different phases and the plants can be divided into three groups, *control*, *treatment,* and *all plants*, we clustered the data for every combination separately for a total of six combinations. The same procedure was followed for the target variable. The last step was to compare the clustering of every combination of parameters, phase, and group to the appropriate clustering of the target variable.

Because of the enormous number of possible combinations in this study we are focusing on three parameters and the target variable. The parameters are (a) Daily transpiration,

which is the total water loss through transpiration during the day, (b) Plant water recharge, which is the amount of water absorbed by the plant from the soil, (c) Daily_Influx_DIELEC_0_5TE which is the total water influx into the plant-soil system measured by the DIELEC/0/5TE sensor and for validation the EC_0_5TE which express the concentration of ions in the soil solution indicating the level of salinity in the plant. The clustering in the results mentioned in the study was made with k-means Time Series witch was the most balanced algorithm of the four mentioned in Section 2.2. Every row of Table 4.5.1 shows the similarity of the high or low cluster (shown in the last column with the letters H for High and L for Low) of the parameter 1 for a specific phase of the experiment and a specific group to the high or low cluster respectively of the parameter 2 for the same phase and group.

Num.	Parameter 1	Parameter 2	Phase	Group	Similarity
1	Daily Transpiration	Plant net Weight	Phase 1	All plants	H: 77.78% L: 90.91%
2	Daily Transpiration	Plant net Weight	Phase 2	All Plants	H: 63.64% L: 88.89%
3	Daily Transpiration	Plant net Weight	Phase 3	All Plants	H: 77.78% L:75%
4	Daily Transpiration	Plant net Weight	Phase 1	Control	H: 80% L:88.89%
5	Daily Transpiration	Plant net Weight	Phase 2	Control	H:71.43% L:85.71%
6	Daily Transpiration	Plant net Weight	Phase 3	Control	H: 71.43% L:71.43%
7	Daily Transpiration	Plant net Weight	Phase 1	Salinity	H: 85.71% L:75%
8	Daily Transpiration	Plant net Weight	Phase 2	Salinity	H: 71.43% L:100%
9	Daily Transpiration	Plant net Weight	Phase 3	Salinity	H: 80% L:90%

Table 4.5.1: Similarity using K-Means Timeseries of Daily Transpiration.

From Table 4.5.1 we can see that the Daily Transpiration is a good predictor of Plant Net Weight because in every case the majority of plants that belong to a cluster of daily transpiration also belong to the same cluster of plant net weight. The all-plants group contains twenty-seven (27) plants, the salinity group contains fourteen (14) plants and the control group thirteen (13) plants. The average similarity for the high cluster is 82.5% and for the low is 85%, so we can be confident that high Daily transpiration is necessary but not sufficient for increase in Plant net weight.

As we have discussed in Section 3.1 the experiment is divided into three phases. The first phase (phase 1) where there is no salinity, the second (phase 2) where salinity was added and the third (phase 3) where the plants got cleaned from salinity. With the structure of the experiment in mind we can compare the change in similarity in rows four and five to the change in similarity in rows seven and eight to see the effect that the addition of salinity has in the confidence of Daily transpiration as a predictor for plant net weight. The change in similarity for the control group from phase 1 to phase 2 (rows four and five) is at most 10% but the corresponding (rows seven and eight) change for the salinity group is 25%.In addition to that the similarity for the salinity group for phase 2 for the low cluster is 100%, this might suggest that the Daily Transpiration is a better predictor of plant net weight in an environment with increased salinity. In other words, low Daily Transpiration in such an environment almost always results in reduced growth but high transpiration does not guarantee high growth, but more research needs to be done.

We can also compare the phases 1 and 3 of the salinity group to see if the effect of the salinity in the predictability of Plant Net Weight from Daily Transpiration remains after the removal of Salinity. As we can see from rows seven and nine there is a 5% decrease in the similarity of the high cluster and 15% increase for the low cluster which might mean that the plants need more time to reach the same state as before the addition of salinity.

Num.	Parameter 1	Parameter 2	Phase	Group	Similarity
1	Daily_Influx_DIELEC_0_5TE	Plant net Weight	Phase 1	All plants	H: 73.33% L: 80%
2	Daily_Influx_DIELEC_0_5TE	Plant net Weight	Phase 2	All Plants	H: 55.56% L: 81.25%
3	Daily_Influx_DIELEC_0_5TE	Plant net Weight	Phase 3	All Plants	H: 62.50% L:70.59%
4	Daily_Influx_DIELEC_0_5TE	Plant net Weight	Phase 1	Control	H: 62% L:100%
5	Daily_Influx_DIELEC_0_5TE	Plant net Weight	Phase 2	Control	H:66.67% L:83.33%
6	Daily_Influx_DIELEC_0_5TE	Plant net Weight	Phase 3	Control	H: 80% L:71.43%
7	Daily_Influx_DIELEC_0_5TE	Plant net Weight	Phase 1	Salinity	H: 83.33% L:71.43%
8	Daily_Influx_DIELEC_0_5TE	Plant net Weight	Phase 2	Salinity	H: 40% L:75%
9	Daily_Influx_DIELEC_0_5TE	Plant net Weight	Phase 3	Salinity	H: 50% L:72.73%

Table 4.5.2: Similarity using K-Means Timeseries of Daily_Influx_DIELEC_0_5TE.

Table 4.5.2 shows the similarity in the clusters created by k-means timeseries of Daily_Influx_DIELEC_0_5TE to Plant net weight similarly to Table 4.5.1. The all-plants group contains twenty-five (25) plants, the salinity group contains thirteen (13) plants and the control group twelve (12) plants. The average similarity of the two parameters for the high cluster is 63.71% and for the low cluster is 78.41% which is a lot higher than the high cluster. Overall, it is clear that like Daily transpiration the Daily Influx is an important parameter that can work as predictor of the plant net weight even though the similarity percentages are not as high as in the daily transpiration from Table 4.5.1.

We can see from the rows four, five and six that for the first two phases of the control group the similarity of the high cluster is lower than the similarity of the low cluster with the opposite being true for the last phase. This phenomenon may be the result of the increase in the importance of this parameter with respect to the growth, but further research needs to be done with a larger dataset.

We can also see that from the comparison of the rows seven, eight and nine we can derive how does the effect of Daily Influx on growth changes in an environment with increased salinity. In the salinity group from the first phase to the second the similarity of the low clusters seems to not be affected by salinity, but the similarity of the high clusters is 43% lower. We can conclude that salinity has a negative effect in the predictability of plant net weight from Daily influx, in other words it seems that high Daily influx is not enough for a plant to have, in increased salinity, in order to also have a high growth rate.

Lastly the similarity in both clusters from phase 2 to phase 3 of the salinity group of the experiments (rows eight and nine) does not change significantly with the implications that even after the salinity is removed the plant continues to have the same behavior as before.

Num.	Parameter 1	Parameter 2	Phase	Group	Similarity
1	Plant_Water_Recharge	Plant net Weight	Phase 1	All plants	H: 77.78% L: 91.67%
2	Plant_Water_Recharge	Plant net Weight	Phase 2	All Plants	H: 66.67% L: 85.71%
3	Plant_Water_Recharge	Plant net Weight	Phase 3	All Plants	H: 77.78% L:76.19%
4	Plant_Water_Recharge	Plant net Weight	Phase 1	Control	H: 80% L:88.89%
5	Plant_Water_Recharge	Plant net Weight	Phase 2	Control	H:71.43% L:85.71%
6	Plant_Water_Recharge	Plant net Weight	Phase 3	Control	H: 71.43% L:71.43%
7	Plant_Water_Recharge	Plant net Weight	Phase 1	Salinity	H: 77.78% L:85.71%
8	Plant_Water_Recharge	Plant net Weight	Phase 2	Salinity	H: 62.50% L:100%
9	Plant_Water_Recharge	Plant net Weight	Phase 3	Salinity	H: 80% L:90.91%

Table 4.5.3: Similarity using K-Means Timeseries of Plant_Water_Recharge.

In Table 4.5.3 we can find the similarity of both high and low clusters of the parameter plant water recharge to the clusters of plant net weight in all phases of the experiment in the same way that is shown in Tables 4.5.1 and 4.5.2. Once again, the algorithm used to derive the clusters is the k-means timeseries because it gave the most balanced results. The all-plants group contains twenty-five (25) plants, the salinity group contains thirteen (13) plants and the control group twelve (12) plants.

From the rows four, five and six we can compute the average similarity for the high cluster, which is 74.28% and for the low cluster, which is 82.01%, in other words high plant water recharge is an indication for high growth rate and low plant water recharge is an indication for low growth rate. This phenomenon seems to change when salinity is added to the plants. The row seven represents the salinity group for the first phase and the row eight represents the same group for the second phase. When comparing these rows, we can see that the similarity in the high cluster is 15% lower in the second phase than in the first and the similarity in the low cluster is 15% higher in the second phase than in the first phase.

The parameter plant_water_recharge has the same behavior as the parameter daily transpiration when salinity is added to the plants and when it is removed from the plants. This behavior of the plants with respect to these two parameters when salinity is added might indicate that for high rate of growth some other parameter becomes more important, but this needs further research with a larger dataset.

Num.	Parameter 1	Parameter 2	Phase	Group	Similarity
1	EC_0_5TE	Treatment	Phase 1	All plants	H: 50% L: 57.14%
2	EC_0_5TE	Treatment	Phase 2	All Plants	H: 100% L: 85.71%
3	EC_0_5TE	Treatment	Phase 3	All Plants	H: 100% L:75%
4	EC_0_5TE	Treatment	Second half of phase 3	All Plants	H: 72.73% L:75%

Table 4.5.4: Similarity using K-Means Timeseries of EC_0_5TE to Salinity.

Table 4.5.4 shows the similarity of the parameter EC_0_5TE, which is a measure of the existence of salinity in the plants, to the treatment groups (salinity and control groups) for every phase of the experiment. All thirty-two (32) plants were used for this table.

The salinity group is treated as if it was the high cluster, and the control group is treated as if it was the low cluster because they are true representations of the existence or no of salinity based on the experiment. In other words, Table 4.5.4 shows the similarity of the created clusters from a measurement to the real clusters according to the groups. We included this table to validate the methodology with known clusters because this is the only parameter that has a measurement and an objective clustering base on the way the experiment was contacted.

From the first row it is evident that in the first phase the plants have no salinity as expected and we can conclude that because the similarity for the low and high clusters is around 50% which suggests that the distribution of the plants from the high and low clusters of EC_0_5TE in the treatment groups is random. This fact changes in the second phase (second row of Table 4.5.4) where there is 100% similarity in the high cluster and 85.71% in the low cluster which means that the EC_0_5TE is an accurate representation of salinity as expected. The similarity in the low cluster is not 100% because it seems that the salinity in some of the plants was not great enough for the algorithm classify them as plants with high salinity even though they had some salinity.

The Figure 4.1 shows the timeseries of the parameter EC_0_5TE for three plants. With blue color is shown the timeseries of a plant that is wrongly classified by the algorithm as not having salinity even though it has, with orange color is shown a timeseries of a plant that belong in the salinity group and with green is shown the timeseries of a plant that belongs to the control group, as we can see there is no clear difference from the blue and green timeseries which may suggest that the salinity was not applied as intended in the experiment.

From the third row of the table were we have the similarity for the third phase of the experiment we can see that even after the salinity is removed the parameter EC_0_5TE gives an accurate representation of the control and salinity groups and even in the second half of the third phase (fourth row of Table 4.5.4) the clusters from the parameter EC_0_5TE is still a good representation of the control and salinity groups with more than 70% similarity.

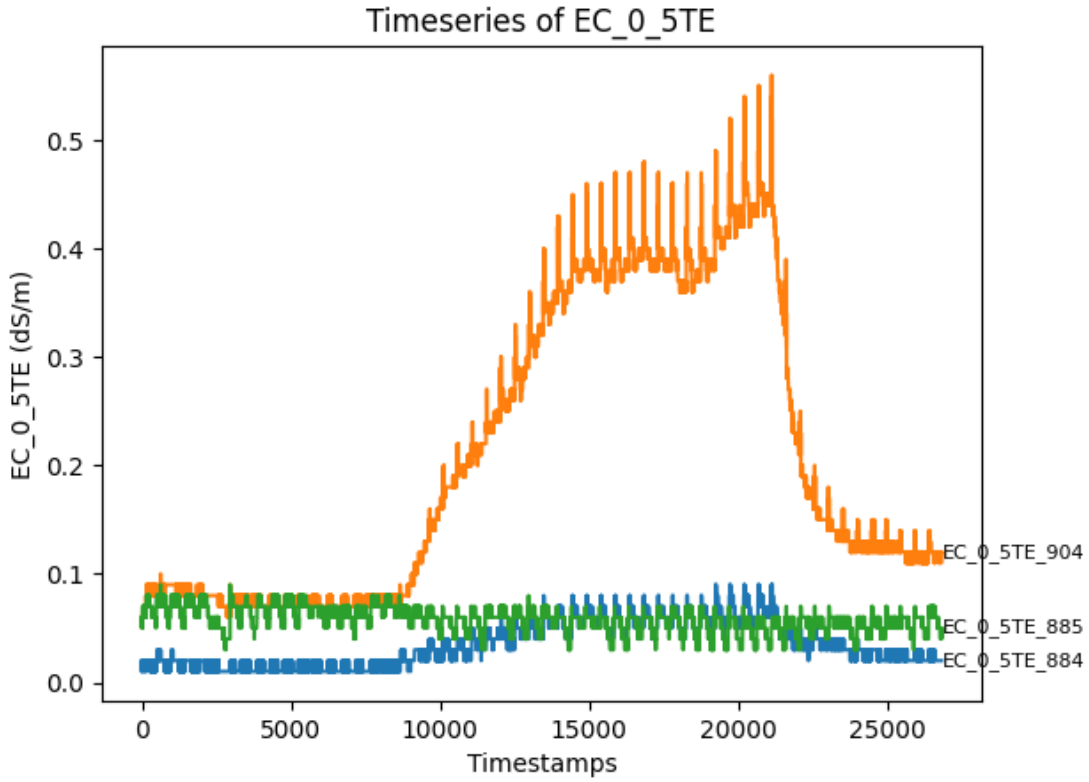

Figure 4.1: Timeseries of EC_0_5TE

4.3 Clustering with Respect to Cultivars

The same methodology used in Section 4.2 can be applied to the second objective of this study, which is understanding the effect that salinity has on the growth of different cultivars of avocado. There are some major differences in the two objectives that we need to take into consideration. For the second objective the parameter that affects the growth is a combination of the existence or not of Salinity and the cultivar every plant belongs too. We can treat these as parameters but there is no need to use clustering for them because we already have the plants separated bases on the group (Salinity and Control) and their cultivar. The target variable is the plant net weight because it gives the best picture of the growth of the plants. For this objective, we know the property of the time series we would like to measure, namely the rate of change, we are free to use

some statistical tools instead of clustering to have more accurate results. We will use both methods and compare the results.

4.3.1 Clustering Method

We used the same four clustering algorithms discussed in Section 2.2 and used in Section 4.1 for clustering the target variable (plant net weight) and similarly to Section 4.2 we will discuss the results of the k-means Time Series as the most representative algorithm. Because there are eight different combinations of group and cultivar, we are focusing on the comparison of two cultivars for both groups at a time. Furthermore, we are focusing on phase 2 of the experiment where the addition of salinity happened. In addition to that we excluded the cultivar named Donnie for simplicity because it was getting placed only in the low cluster.

	High Cluster of Plant Net Weight	Low Cluster of Plant Net Weight
Hass - Control	75%	25%
Hass - Salinity	50%	50%
Waldin - Control	25%	75%
Waldin- Salinity	0%	100%

Table 4.6: Results of Clustering for Hass-Waldin

	High Cluster of Plant Net Weight	Low Cluster of Plant Net Weight
Hass - Control	75%	25%
Hass - Salinity	50%	50%
Mexicola - Control	25%	75%
Mexicola- Salinity	25%	75%

Table 4.7: Results of Clustering for Hass-Mexicola

	High Cluster of Plant Net Weight	Low Cluster of Plant Net Weight
Waldin - Control	75%	25%
Waldin - Salinity	50%	50%
Mexicola - Control	100%	0%
Mexicola- Salinity	50%	50%

Table 4.8: Results of Clustering for Mexicola-Waldin

	High cluster phase 1	High cluster phase 2	High cluster phase 3
Hass - Control	100%	100%	100%
Hass - Salinity	100%	75%	75%
Waldin - Control	25%	25%	25%
Waldin - Salinity	50%	25%	25%
Mexicola - Control	0%	25%	50%
Mexicola - Salinity	50%	25%	25%
Donnie - Control	0%	0%	0%
Donnie- Salinity	0%	0%	0%

Table 4.9: Results of Clustering for Every Group and Every Phase

Tables 4.6, 4.7, 4.8 show the similarity in the clustering of Plant Net Weight using only the plants that belong in Hass - Waldin, Hass - Mexicola, Waldin - Mexicola respectively and the clusters that result from the combination of the cultivars used and the two groups. We node that for the creation of these tables all plants were used due to the large number of combinations (8).

From Tables 4.6 and 4.7 we can conclude that the cultivar Hass has better growth than the cultivar Waldin and Mexicola both with and without Salinity. Furthermore, the cluster of Hass that had salinity has better growth than the cluster of Waldin and Mexicola that didn't have salinity. From the Table 4.8 we can see that the cultivars Mexicola and Waldin are very close in the growth rate for both control and salinity groups.

Table 4.9 summarizes the clustering for all cultivars in all phases of the experiment for both groups independently. Only the similarity to the high cluster is shown for simplicity. We can see that the Hass cultivar has consistently the highest growth and the Donnie the lowest. The Mexicola control group from 0% of plant in the high cluster in the first phase it ended up with 50% of the plants in the high cluster while all other control groups have the same percentage of plants in the high cluster in all three phases. That might indicate that the second order rate of change of the Mexicola's growth is higher between phases than the rest of the cultivars.

4.3.2 Statistical Method

For the second objective we can use some statistical methods to understand better how salinity and cultivar affect the growth of a plant. This is possible and even better than clustering because we already have clusters for the cultivars and the groups (*Control* and *Salinity*). In addition, we know exactly which is the property of the timeseries of the Plant Net Weight we need to measure namely the rate of change which happened to be possible to approximate with the slope of a line because of the small timescale of the experiment. This statistical method is used to validate the clustering method, to give more accurate results regarding the rate of growth of every cultivar with respect to salinity but also to measure how salinity affects the rate of growth.
To calculate the rate of change of the timeseries we used four different methods. The first method works by finding the average of the first and last N (N=3) data points and find the slope between them. The second method works by using linear regression to approximate the data with a line from which we get the slope. The third method works

by finding the slope of every possible pair of points and averaging them. The fourth method works by finding the slope of every pair of consecutive points and calculating the average of those slopes.

	Phase 1	Phase 2	Phase 3
Donnie-Control	4.09	0.27	1.39
Donnie-Salinity	3.7	1.47	1.45
Hass-Control	7.28	5.01	4.48
Hass-Salinity	8.4	4.66	6.38
Mexicola-Control	6.27	3.15	5.93
Mexicola-Salinity	5.91	2.74	2.81
Waldin-Control	5.26	3.39	5.05
Waldin-Salinity	5.93	2.96	3.17

Table 4.10: Rate of Change Calculated with the “Average” Method.

	Phase 1	Phase 2	Phase 3
Donnie-Control	4.1	0.06	1.02
Donnie-Salinity	3.6	1.6	1.29
Hass-Control	7.4	5.12	4.46
Hass-Salinity	8.56	4.66	6.29
Mexicola-Control	6.23	3.29	5.73
Mexicola-Salinity	6.06	2.88	2.72
Waldin-Control	5.29	3.4	5
Waldin-Salinity	6.19	3.07	3.06

Table 4.11: Rate of Change Calculated with the “Linear Regression” Method.

	Phase 1	Phase 2	Phase 3
Donnie-Control	4.09	0.12	1.16
Donnie-Salinity	3.6	1.56	1.4
Hass-Control	8.08	5.91	4.03
Hass-Salinity	8.65	4.65	6.47
Mexicola-Control	6.18	3.24	5.71
Mexicola-Salinity	6.11	2.84	2.7
Waldin-Control	5.3	3.38	5.03
Waldin-Salinity	6.32	3.03	3.05

Table 4.12: Rate of Change Calculated with the “All Pairs” Method.

	Phase 1	Phase 2	Phase 3
Donnie-Control	3.98	0.5	1.65
Donnie-Salinity	3.62	1.29	1.35
Hass-Control	8.27	5.78	4.5
Hass-Salinity	9.06	4.64	7.66
Mexicola-Control	5.86	3.01	5.59
Mexicola-Salinity	6.35	2.58	2.52
Waldin-Control	5.31	3.25	5.29
Waldin-Salinity	6.92	2.79	2.97

Table 4.13: Rate of Change Calculated with the “Consecutive Pairs” Method.

	Average	Linear regression	All pairs	Con pairs
Hass	19%	21%	27%	27%
Mexicola	8%	10%	11%	21%
Waldin	22%	23%	25%	34%

Table 4.14: Estimated Loss in Rate of Change of Growth for each Method.

Tables 4.10, 4.11, 4.12, 4.13 show the average slope of every combination of cultivar and group for every phase of the experiment calculated by the methods discussed earlier. In

order to create these tables, we used only the plants that appear to have logical values except for the Donnie-control which have only one plant with logical values, so we kept the rest that didn't have missing values in the calculation. The total number of plants used is twenty-seven. From Tables 4.10, 4.11, 4.12 we can see that the methods "average", "linear regression", and "All pairs" give very similar results with the "con pairs" method giving slightly different results which is expected because it is not as tolerant to noise. For simplicity we are going to use the "linear regression" method to discuss the results, but the same results can be derived from the other methods and in the context of this study they were used to increase the confidence in the methods and validate the results.

The cultivars in descending order based on the rate of growth are Hass, Mexicola, Waldin, Donnie with the Mexicola and Waldin being very close. When salinity is added the list becomes Hass, Waldin, Mexicola, Donnie but once again the Waldin and Mexicola are very close in terms of growth. That might suggest that the difference in growth because of salinity is similar or at least is not significantly larger for one cultivar in comparison to the others. Furthermore, if we focus on the salinity groups, we can see that in the case of Hass the growth rate in the third phase is bigger than the rate of growth in the second phase. This is not true for the other three cultivars.

Table 4.14 shows the estimated loss in growth of every cultivar except Donnie calculated based on the four methods. The loss is calculated with the following formula.

$$Estimated\ Slope = Slope_{(control,phase2)} / Slope_{(control,phase1)} * Slope_{(Salinity,phase1)}$$

$$Estimated\ loss = (Estimated\ Slope - Slope_{(Salinity,phase2)}) / Estimated\ Slope$$

Donnie was excluded from the calculation because the Donnie control group does not give a reasonable measure for the rate of growth in phase 2 and that might be the result of illogical values which were included only in the case of the second phase for completion purposes. The formula calculates the average Estimated slope that the cultivar in question would have if it was not in an environment with salinity based on the experimental results from the control group. Subsequently, the loss is calculated as the difference between the Estimated slope and the real slope normalized by the Estimated slope.

From Table 4.14 we can conclude that Mexicola has the lowest loss in growth rate due to salinity followed by Hass and Waldin, and the first three methods agree that the loss is about 10% which is approximately half of the loss in the growth of Hass. As we can see in absolute numbers Hass has the best behavior in an environment with increased salinity but in terms of the minimum loss in the growth rate the best cultivar is Mexicola.

The two methods we used for the second objective of the study agree on the results they yield and even though the clustering methods is not best suited for this objective, the statistical method validates the results it has given. Of course, the statistical method for this objective can be used to generate more results like the loss in the rate of growth due to salinity but the same method cannot be used in the first objective of the study. Firstly, because we do not know what feature of the timeseries of the parameters we are interested in and, secondly a direct comparison of the timeseries themselves is not possible because they are not homogeneous.

Chapter 5

Discussion

5.1 Summary

This thesis focuses on creating a machine learning based methodology that helps with the analysis of the relationship between the attributes of a plant. To accomplish this, we made use of data kindly given to us by the Plant Molecular Physiology Lab. Specifically, the Plant Molecular Physiology Lab contacted an experiment with objective to investigate the effects of Salinity to different cultivars of avocado.

Four cultivars were used with eight plants in each cultivar for a total of thirty-two (32) plants and four of the eight plants of every cultivar were treated as control group and the rest as treatment group. The experiment consists of three phases, phase one, two, and three. The environment for phase 1 was normal in terms of salinity and for phase 2 salinity was added to the treatment group but it was removed for phase 3.

During the experiment, sophisticated sensors took measurements of some important features of the plants like the weight and transpiration rate. Some features where measured every three minutes and some once a day. The data were then saved in an online platform for which we were given access through an API. With the use of python and SQL lite we were able to fetch the data and store it in a small database for better and easier management.

Once we had the data locally, we had to clean it since the sensors were not always working properly resulting in missing and illogical values. Because many of the parameters were expressing the same feature of the plant it was necessary to reduce them to a few distinct parameters and we did that with the use of statistical correlation.

Once we had the data ready to be use, we chose four different clustering algorithms to gain confidence regarding the results. Our primary objective was to establish a methodology for revealing the relationship between the different parameters we had data on making clustering the best option because it was not clear what was the relevant feature of every timeseries of the parameters and the data was not homogeneous.

The idea behind the methodology is use the clustering algorithms on the timeseries of the two parameters for all plants and compare the resulting clusters. The result would be a statistical representation of the effect of one parameter on the other. Because there are many combinations of parameters we chose to focus on the growth of the plants as the most important one and in the thesis, we discuss its relationship with Daily transpiration.

A secondary objective was used to compare the results of this methodology and statistical analysis. The objective was to analyze how Salinity affects the average growth of the plants in each cultivar. Even though the objective is not best suited for the methodology because the cultivars are already in clusters and the treatment plants are known in advance, it still gives the correct results, and the statistical analysis validates those results giving us confidence in methodology.

For the statistical analysis four methods were used to calculate the rate of the growth of the plants to give us confidence in the results and based on those we were also able to calculate the average loss in rate of growth of every cultivar.

Finally, this thesis accomplished to show that it is possible to analyze similar types of data and get results using the methodology we discussed.

5.2 Problems Encountered

During the course of this study, we encountered many problems related to the fetching and analysis of the data which impacted the resolution of the results. The first problem we encounter was related to the fetching of the parameter “Gs_canopy_Watchdog”. Specifically, when we tried to fetch the data given the identification code of the parameter we got an error that suggested that the data did not exist. The problem we

encountered was unresolved, leading to a decrease in the number of available parameters by one.

Furthermore, because of the structure of the data it was not possible to find an intuitive schema for the database. Specifically, the data consisted of thirty-two different plants and every plant had thirty parameters and each parameter was a timeseries of three months with timesteps every three minutes or once a day. This is a 3D structure that cannot be directly mapped on a 2D structure in an efficient way. Thankfully because of the small number of data we realized that the mapping did not need to be efficient. With that in mind we created two tables for each plant, one table for the parameters with timestamps every 3 minutes and one for the parameters that had timestamps once a day.

The worst problem we encountered was the fact that the data were damaged and had many missing or illogical values making some parameters and some parts of the timeseries unusable. This resulted in a cleaned dataset smaller than what we would like to work with to have reliable results.

5.3 Future Work

The current study, although showing encouraging outcomes, also highlights areas for future enhancement and improvement, specifically in data collection and algorithm application.

Firstly, the dataset used in this study did not have the size required for the algorithms applied to it and after the data cleaning the resulting dataset was even smaller because of a large number of missing values. Further research could involve the use of a larger dataset with more plants of the same cultivar with the same time window. With a larger dataset is would be possible to make reliable clustering with more than two clusters to gain more resolution on the relationship of the parameters. There is also the possibility to make different number of clusters based of the nature of the parameters. In addition, one further improvement on the methodology may be the clustering and comparison of

more than two parameters at a time to capture more complex relationships between more than two parameters.

Furthermore, future research could investigate a larger variety of algorithms and distance metrics to find the best combination. As mentioned before, a good point of research is to try clustering algorithms like DBSCAN [20] and OPTICS [21], which do not require in advance the number of clusters, but they choose the best number of clusters based on the data.

Bibliography

[1] C. Jackulin and S. Murugavalli, "A comprehensive review on detection of plant disease using machine learning and deep learning approaches," Measurement: Sensors, vol. 24, p. 100441, 2022. [Online]. Accessed on 10 October 2023

[2] Li, L., Zhang, S., & Wang, B. (2021). Plant disease detection and classification by Deep Learning—A Review. *IEEE Access*, *9*, 56683–56698.

[3] Namratha M 1 , Prajwala T R. (2012). A comprehensive overview of clustering algorithms in pattern recognition. *IOSR Journal of Computer Engineering*, *4*(6), 23–30. https://doi.org/10.9790/0661-0462330

[4] Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, *7*(2), 147–177. https://doi.org/10.1037//1082-989x.7.2.147

[5] Mahesh, Batta. "Machine Learning Algorithms - A Review." International Journal of Science and Research (IJSR), vol. 9, no. 1, January 2020, ISSN: 2319-7064, pp. [381-386].

[6] Pádraig Cunningham, Matthieu Cord, and Sarah Jane Delany. 2008. Supervised learning. In Machine Learning Techniques for Multimedia. Springer, 21–49.

[7] M. E. Harmon and S. S. Harmon, “Reinforcement Learning: A Tutorial.,” WRIGHT LAB WRIGHT-PATTERSON AFB OH, 1997

[8] G. Fung, “A comprehensive overview of basic clustering algorithms,” IEEE Trans. Inf. Theory, vol. 27, no. 1, pp. 49–60, Jan. 2001

[9] Tokuda, E. K., Comin, C. H., & Costa, L. da. (2022). Revisiting agglomerative clustering. *Physica A: Statistical Mechanics and Its Applications*, *585*, 126433. https://doi.org/10.1016/j.physa.2021.126433

[10] Ryan P. Adams. “K-Means Clustering and Related Algorithms”. 226 In: Elements of Machine Learning (2018). 227 Stanford Center for Academic Medicine, 453 Quarry Road, Palo Alto, 228 CA 94304

[11] V. Niennattrakul and C. A. Ratanamahatana, “On clustering multimedia time series data using k-means and dynamic time warping,” in Proc. Int. Conf. Multimedia Ubiquitous Eng. (MUE), 2007, Seoul, South Korea, pp. 733–738.

[12] Schubert, E., & Rousseeuw, P. J. (2019). Faster K-medoids clustering: Improving the PAM, Clara, and Clarans algorithms. *Similarity Search and Applications*, 171–187. https://doi.org/10.1007/978-3-030-32047-8_16

[13] Arbin, N., Suhaimi, N. S., Mokhtar, N. Z., & Othman, Z. (2015). Comparative analysis between K-means and K-medoids for statistical clustering. *2015 3rd International Conference on Artificial Intelligence, Modelling and Simulation (AIMS)*. https://doi.org/10.1109/aims.2015.82

[14] Arora, P., Deepali, & Varshney, S. (2016). Analysis of K-means and K-medoids algorithm for Big Data. *Procedia Computer Science*, *78*, 507–512. https://doi.org/10.1016/j.procs.2016.02.095

[15] Wang, W., Lyu, G., Shi, Y., & Liang, X. (2018). Time series clustering based on dynamic time warping. *2018 IEEE 9th International Conference on Software Engineering and Service Science (ICSESS)*. https://doi.org/10.1109/icsess.2018.8663857

[16] Bouhmala, N. (2016). How good is the Euclidean distance metric for the clustering problem. *2016 5th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI)*. https://doi.org/10.1109/iiai-aai.2016.26

[17] S. Aghabozorgi, A. S. Shirkhorshidi, and T. Y. Wah. Time-series clustering–a decade review. Information Systems, 53:16–38, 2015.

[18] W. Meesrikamolkul, V. Niennattrakul, and C. A. Ratanamahatana. Shape-based clustering for time series data. In PAKDD, pages 530–541. 2012.

[19] P. Berkhin, "Survey of clustering data mining techniques," Yahoo, Sunnyvale, CA, USA, Tech. Rep., 2002, https://doi.org/10.1007/3-540-28349-8_2

[20] M. Ester, H. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in Proc. 2nd Int. Conf. Knowl. Discov. Data Mining, 1996, pp. 226–231

[21] M. Ankerst, M. M. Breunig, H.-P. Kriegel, and J. Sander, "OPTICS: Ordering points to identify the clustering structure," ACM SIGMOD Rec., vol. 28, no. 2, pp. 49–60, 1999.

[22] Plant Phenotyping: Automated Phenotype Platform | Plant Ditech (plant-ditech.com)