Ben-Gurion University of the Negev
The Faculty of Engineering Science
The Department of Industrial Engineering and Management

M.Sc. Seminar

Analyzing the Primary Factors Influencing the Severity of Macrophomina Phaseolina in Israel to Enhance Cotton Cultivation Yield

Submitted by:
Matan Spiro, M.Sc. student
The Department of Industrial Engineering and Management
Ben-Gurion University of the Negev
E-mail: spiro@post.bgu.ac.il

Supervisor:
Prof. Gilad Ravid
The Department of Industrial Engineering and Management
Ben-Gurion University of the Negev
E-mail: rgilad@bgu.ac.il

Table of Contents

Tables' Table of Contents

Figures' Table of Contents

1. Abstract

Cotton crop yield in Israel is significantly affected by various environmental factors, including the prevalence of diseases such as Macrophomina phaseolina (MP). MP is a common pathogen in cotton crops in Israel and has a substantial negative impact on yields globally. This study analyzed key factors influencing the severity of MP, which ultimately affect cotton crop yield - the more severe the MP infection, the lower the expected yield. Data on rainfall and temperature was collected for the weeks preceding and following the sowing period, and numerous machine learning models were employed to predict yield categories (low, medium, and high). While the accuracy of these models was modest, with a maximum accuracy of 56.8% using a random forest model, the analysis highlighted the most informative features that contribute to cotton crop yields in MP-infected soil, and their relationships with the target variable. Through logistic regression equations and random forest's most influential factors, we identified key environmental factors that significantly influence the probability of each yield category. These findings provide valuable insights into the complex dynamics of cotton yield in Israel and can inform agricultural practices and disease management strategies to enhance crop productivity.

2. Introduction

Cotton cultivation is an important part of agricultural activity in Israel, mostly as a contributor to economic prosperity in terms of gross domestic product. In regions like Israel, where agricultural endeavors play a significant role in shaping the local economy, understanding the dynamics of a growing's yield becomes paramount. Farmers try to maximize agricultural output, and their effort is hindered by different challenges, for example the presence of pathogens like MP. This fungal disease, known for its detrimental effects on cotton crops, poses an obstacle to farmers. This fungal pathogen targets cotton plants, causing a range of fatal effects: by infecting the roots and vascular tissues of cotton plants, MP disrupts the uptake and transport of water and nutrients. This impairs the plant's ability to access essential resources from the soil, leading to nutrient deficiencies and stunted growth [1], that ultimately result in reduced yield and economic losses for growers.

In order to examine the disease's effect on cotton crops' yield in Israel and to ultimately mitigate its disastrous effect, we conducted this study based on data of sowing and yield of cotton crops, that was collected in farm areas that are inflicted by MP.

We examined some of the parameters that have been associated to influence the severity of MP, with a goal to find the most influential ones and their relationship with each other, with respect to the time dimension. We relied on past studies that concluded that MP suppresses yield in cotton crops grown in Israel [2] [3]. By understanding that and considering that environmental factors such as rainfall and temperature affect the severity of MP [1] [4], we managed to construct supervised machine learning models that can predict the cotton crops' yield to some extent.

We found it useful to include rainfall and temperature data of the weeks that preceded the sowing date and the weeks that followed it. The initial soil and environmental conditions can of course influence the presence and severity of soilborne diseases such as MP. In addition, data collected in dates that followed the sowing date is also of vital importance, as it holds ongoing information of the presence and severity of the disease (which is directly linked to the environmental conditions that were used as data). Integrating data from before and after sowing should be done to create a much fuller picture of the disease's severity, and it also allows for a comprehensive assessment of disease risk and management strategies. By identifying risk factors present in the soil before planting, and monitoring disease development throughout the growing season, it is possible to deduce management practices to mitigate the impact of MP. This may include measures such as crop rotation, soil amendments, and fungicidal treatments aimed at reducing disease incidence and preserving crop yield.

3. Literature Review

3.1. Macrophomina Phaseolina

MP is a generalist soil-borne fungus that is present all over the world. It infects more than 500 plant hosts and attacks any part of the plant at any stage of growth from germination to harvest, thus causing damping off (a disease of seedlings), seedling blight, leaf blight, color rot, stem rot and root rot [5]. Under certain conditions this fungus can cause substantial yield losses in crops such as soybean, strawberry, and cotton. An average yield loss of 10% has been reported, but in case of severe infection it may increase up to [5]. In soybean alone researchers reported yield losses of more than 2 million tons during 2003ï 2012 [6]. MP favors high temperatures (30Á35Á) and low soil moisture [6]. The pathogen becomes severe in hot (35ÁC or more) weather [5]. MP can still cause great harm with a more humid weather, even though it generally favors environments with low moisture. Low soil

moisture can occur due to insufficient rainfall, high temperatures causing evaporation, low soil water retention capacity, or excessive water usage by plants.

In some studies, the fungus' genetic diversity has been associated with the host plant's origin and/or geographical locations, while in others the researchers could not clearly differentiate cases of presence of the organism based on its pathogenicity, host, or geographical origins. Furthermore, in numerous studies the distribution of MP genotype has been found to be independent of sampling location and host. The disease's wide host range and heterogeneous nature make it difficult to control [7]. In addition to these difficulties, another one can be found in the field of fungicides. fungicides are chemicals used to control or kill fungi that cause diseases in plants, animals, or humans. These chemicals are specifically designed to inhibit the growth of fungal pathogens or to destroy them altogether. The chemical control of MP is difficult, since there are no systemic fungicides that move towards the root, and more specifically - no fungicides have been registered to control this pathogen [7]. In [4] the researchers tested 16 different fungicides to suppress the wilt caused by MP, and out of those 16 only 3 were concluded to be effective. The plants studied after being exposed to the fungicides cannot be defined as resistant to MP since the pathogen can penetrate the roots, but disease incidence has been significantly reduced compared to the untreated control [4]. That is a valuable conclusion for the fight against MP, but it is also important to remember that agro-environmental policies and the increasing negative perception of the public on the agrochemicals may lead to the evaluation and comparison of chemicals agents with more sustainable alternatives to control plant diseases caused by MP.

The pathogen causes charcoal rot disease (CRD) in cotton, with symptoms that develop mainly in the late stages of growth. These include drying leaves and stems, plant wilting, and death. In recent years, the disease has increased in cotton fields in Israel, possibly due to the transition to susceptible extra-long staple (ELS) Pima-type (a type of cotton known for its long and fine fibers, which make exceptionally strong and durable cotton) and the climatic conditions changes. It is now widely agreed upon that CRD significantly impacts Israeli cotton agriculture [3].

The fungus survives in the soil primarily as multicellular sclerotia, which are formed in the host tissues and are released into the soil as tissues decay. Multicellular

sclerotia are specialized structures formed by certain fungi, such as certain species of molds and mushrooms. Sclerotia serve various purposes for these fungi, including survival during unfavorable conditions such as drought or cold temperatures.

The inoculum densities (concentration of material containing microorganisms that are introduced into an environment for the purpose of initiating growth or causing an infection) of sclerotia in the soil are directly related to the incidence of the disease in the field [6], and those densities can potentially increase in plant disease through mechanisms involving mutations and hyphal fusion, which tend to occur in fungi such as MP [1]. Fungi without a known sexual phase, such as MP, must rely on mutations and hyphal fusion to generate genetic variation. Hyphae are thread-like structures that make up the body of a fungus; those can sometimes be seen in rotten vegetables even at home. Hyphal fusion is a process in fungi where the hyphae from two different fungal individuals come into contact and fuse together. The fusion of hyphae allows for the exchange of genetic material, cytoplasm, and organelles between the two fungi [1].

In conclusion, MP poses a significant challenge in the cultivation of cotton due to its resilient nature, widespread prevalence, and detrimental impact on crop growth. As a soilborne fungus, its survival structures, such as sclerotia, allow it to persist in adverse conditions and serve as a continuous threat to the production of many types of plants, specifically cotton, across diverse agroecological regions. The difficulty in controlling MP stems from its ability to infect various plant species, including cotton, as well as its capacity to resist chemical fungicides. Addressing the challenges posed by MP demands a multifaceted approach, integrating more informed practices, genetic resistance, and sustainable management strategies to mitigate its effects and safeguard cotton yields and general agricultural sustainability.

3.2. Big Data in Agriculture

The 2023 State of Food Security and Nutrition in the World report, that was published by the United Nations, reveals that 900 million people are facing severe food insecurity. According to World Food and Agriculture Organization (FAO), every

third child in the world dies from hunger, and another 200 million children under the age of five go hungry regularly [8].

In addition to these troubling facts, FAO stated that food production must increase by 70% by 2050 to meet growing global populationôs demands. Agriculture remains a primary source of food for the population and raw material for many industries. Population growth and climate changes are increasing the importance of using science to improve agriculture, as humanity need to produce more food using fewer inputs, and so agriculture is seeking new products, practices, and technologies [9]. Big data in agriculture refers to the use of large and complex datasets generated from various sources. The data can be related to ground, weather, consumption, energy use, prices, economic information and so on, and it can be gathered using sensors, machines, drones, satellites and more. The data obtained is used to make more timely or accurate decisions, through constant monitoring or specific big data science enquiries [9]. Utilizing technologies that rely on big data, food could be produced more efficiently, of higher nutritional quality, in more stable supplies, with less environmental damage, and likely with additional economic and ecological benefits [10].

Agricultural data can be used in terms of precision agriculture, which mean optimization of inputs such as water, fertilizers, and pesticides by providing information about soil conditions, weather patterns, crop health, and pest infestations. This helps in reducing waste, increasing productivity, and minimizing environmental impact. Satellite imagery and drones equipped with sensors can capture detailed information about crop health, growth patterns, and field conditions. This data can be used to identify areas of concern such as nutrient deficiencies, pest infestations, or irrigation issues. Thus, farmers can intervene swiftly which can enhance production efficiency. Big data analytics can also streamline the agricultural supply chain by improving logistics, inventory management, and distribution processes. By tracking the movement of goods from farm to market, stakeholders can identify inefficiencies, reduce spoilage, and ensure timely delivery of products to consumers.

Regarding aspects that are not directly related to farmed fields, it is also important to mention that analyzing market trends and consumer preferences using big data can help farmers and agribusinesses make strategic decisions about crop selection, pricing, and marketing strategies. This ensures that agricultural products meet the demands of consumers while maximizing profitability for producers. In the realm of

risk management, big data analytics can assist in assessing and managing risks that arise from diseases that attack the plants.

In [8], the researchers concluded that big data analysis had not yet been widely applied in agriculture, meaning that it was still at an early development stage. But while big data analysis had not yet been widely applied in agriculture, there are cases where it had been used, and in [9] several examples of such applications are given: GPS-based applications in precision farming are being used for farm planning, field mapping, soil sampling, tractor guidance, crop scouting, and yield mapping. Drones with infrared cameras and GPS technology led to better decision making and better risk management. In livestock farming, smart dairy farms are replacing labor with robots in activities like feeding cows, cleaning the barn, and milking the cows. Chemical sensors are also used to register the presence and amount of specific chemicals or chemical groups.

It is also important to mention that along with the incredible advantages that can be gained using big data in agriculture, there are some downfalls to it. For example, the data needs to be protected both in transit and at rest. This requires constant monitoring and constant automated response [11]. In addition, farmers may have concerns about who owns and controls the data generated by their agricultural operations. If they will use technologies such as sensors, drones, or satellite imagery provided by third-party vendors, they may worry about losing control over their sensitive information and how it is being used or shared.

4. Methodology

We were given data of cotton crops that were sowed in different farm areas in Israel, that had their soil infected with MP. We then used the daily rainfall and maximum & minimum temperature that were measured in those areas, to construct a dataset that could serve a machine learning model. We aggregated the daily data into weekly intervals to create a more manageable dataset. We used data of 18 weeks in total: 8 weeks that preceded the sowing (t_{-8} to t_{-1}), the week of the sowing itself (t_0) and 9 weeks that followed the sowing week (t_1 to t_9). We also examined the possibility of transforming the continuous target variable, yield in kilogram (KG), into three distinct categories, ensuring that each category contained a roughly equal number of records,

thereby facilitating the model's ability to discern patterns and relationships. We then employed several machine learning models, as well as a deep learning model.

4.1. Data Exploration

The work commenced with a data file that its records contained information about each sowing of cotton crops. The file contained 1,058 records of cotton crops, that were sowed in 38 different farm areas, in dates ranging from 6th of March and 26th of April, in each of the years between 2015 and 2023. We used the fields: farm area, sowing date, and yield. By examining the distribution of the yield, we noticed that most values are centered at around 600 KG, and the distribution resembles a normal distribution with left skewness which mean there were a few records with relatively low yield.

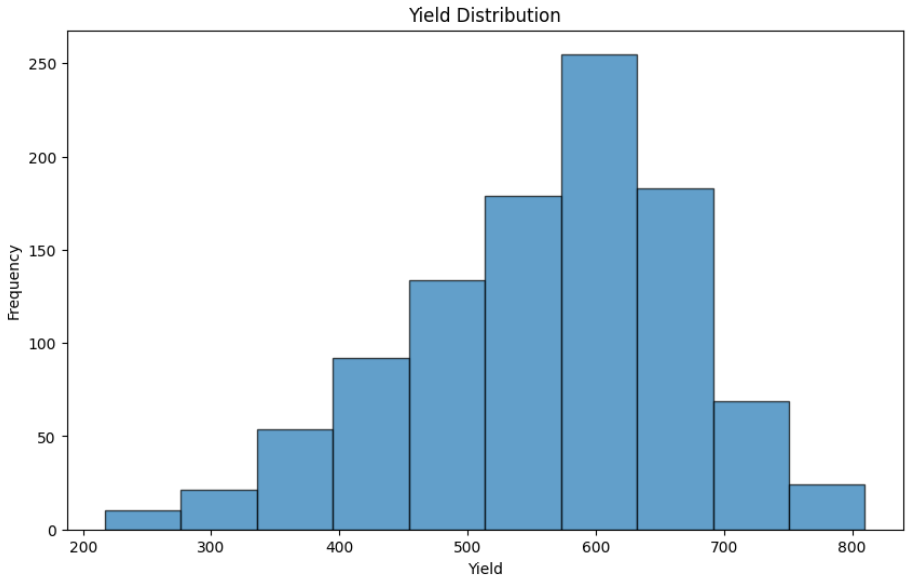

Figure 1 - yield distribution

In addition to the fields that the data file contained, we manually found the farm area's coordinates using Google Maps. We also collected rainfall and temperature data from Israel's Meteorological Service (IMS). IMS has many stations scattered around the country, and they differ in rate of sampling and in the elements they sample. We constructed a list of relevant stations per each of the two elements we examined (rainfall and temperature), and each list contained the name of the station, its coordinates and the station's commencement date and cessation date, which had a null value in case the station is still active. We found that there are 1325 unique relevant rain stations, and 204 unique relevant temperature stations. We also noted that the longitude and latitude fields have a redundant tailing string.

4.2. Preprocessing

We removed stations that were not active during the time span of the data, and those that actually did not collect any data even though were listed in the IMS database. We also removed the redundant string at the end of the longitude and latitude numeric strings. The next step was finding the IMS stations that were closest to the farm areas, so that the rainfall and temperature data that will be used will resemble the actual rainfall and temperatures that occurred in the farm area itself. For each farm area, we found the closest rain station and temperature station using haversine formula, which calculates the shortest distance between two points on the surface of a sphere. We ended up with two data frames, each for rainfall and temperature, and each contains the fields: farm area, closest station and distance (in kilometers).

The next step was to fetch the data of the relevant stations. However we needed to fetch a certain subset of the data, that was relevant for combinations of station type (rainfall or temperature), station name and time span. For each year included in the study, we aimed to retrieve data spanning from January 1st to June 30th. As described earlier, we structured this timeframe to encompass data both preceding and following the sowing period, which occurred annually in March or April. The farm areas that were studied differ in the timing of the data collection. For 2015 for example, the dataset contained 25 relevant farm areas (out of the total of 38), meaning that only a certain 25 farm areas were being experimented and thus yielded data for this study, and for 2016 there were 15 relevant farm areas (that some of them were also examined in 2015). Because of the vast number of IMS station and the fact that only some were relevant to the study, as well as the fact that the relevant stations are different per each year in the study, we grouped the results and found which rainfall and temperature station were relevant for each year of the study. The results are as follows:

operational_years	rain_station
2015-2015	אושה, כפר מסריק, נחשון מאוישת, צובה, צרעה
2015-2016	בית ניר, גת מאוישת, חדרה תחנת הכח
2015-2017	משואות יצחק
2015-2018	כפר ביל"ו, עין החורש מאוישת
2015-2020	גבעת ברנר, כפר הרואה, כפר מנחם, קדמה
2015-2022	חולדה, ניצנים קבוץ
2015-2023	גן שלמה, רבדים
2016-2016	נתיב הלה מאוישת
2016-2018	אילון מאוישת

2016-2020	עין צורים
2016-2022	בני דרום
2016-2023	בית חלקיה, יסודות, שעלבים
2018-2018	נגבה מאוישת
2020-2020	גן השומרון

Table 1 - relevant rain stations by year

operational_years	temp_station
2015-2015	אפק, חיפה בתי זיקוק, צובה
2015-2020	גת, חדרה נמל, נגבה
2015-2022	ניצן
2015-2023	חפץ חיים, נחשון, קבוצת יבנה
2016-2016	נתיב הלה
2016-2018	אילון

Table 2 - relevant temperature stations by year

We then fetched daily data from IMS by the criterions of station and year, spanning from January 1st to June 30th. The rainfall values are measured in millimeters, and the minimum/maximum temperatures values are measured in Celsius. The dataset lacked dates in which the station did not measure rain, so we incorporated the missing dates with the value of "0". The dataset also occasionally lacked minimum and maximum temperatures, and so we filled the missing dates and their measured values by forward fill, or backward fill in case forward fill was not possible.

The next step involved aggregating the data into weekly intervals instead of daily ones. In this dataset, each record corresponds to a distinct sowing event, with individual sowing dates varying across records. To standardize the temporal alignment of the features with respect to each sowing event, we used a dynamic time window approach. We used a set of features denoted as week t_{-8} to week t_9, where week t_0 represents the week of sowing for each respective record. Consequently, the value in column week t_1 for the first record for example, corresponds to the week immediately following the specific sowing event that is given in the first record, and the value of column week t_1 for the second record corresponds to the week immediately following the specific sowing event that is given in the second record. Thus, column week t_1 does not depict any specific week counted since the beginning of a certain year, it merely depicts the information of the week that followed each sowing. This dynamic framework allows for the consistent characterization of temporal patterns relative to the unique timing of each sowing event. Week t_1 for a given record may represent a period in early March, whereas for another record, it could denote early April, depending on the individual

sowing dates. This approach accommodates the inherent variability in sowing schedules across records, enabling meaningful temporal analysis tailored to the specific context of each sowing event.

After examining different aggregation methods with the goal of increasing the accuracy of the machine learning models we used later on, the values of the dynamic framework included summation of daily rainfall and averaging the daily values of each of the features maximum temperature and minimum temperature. We can inspect the weekly mean values' trend of each category of features:

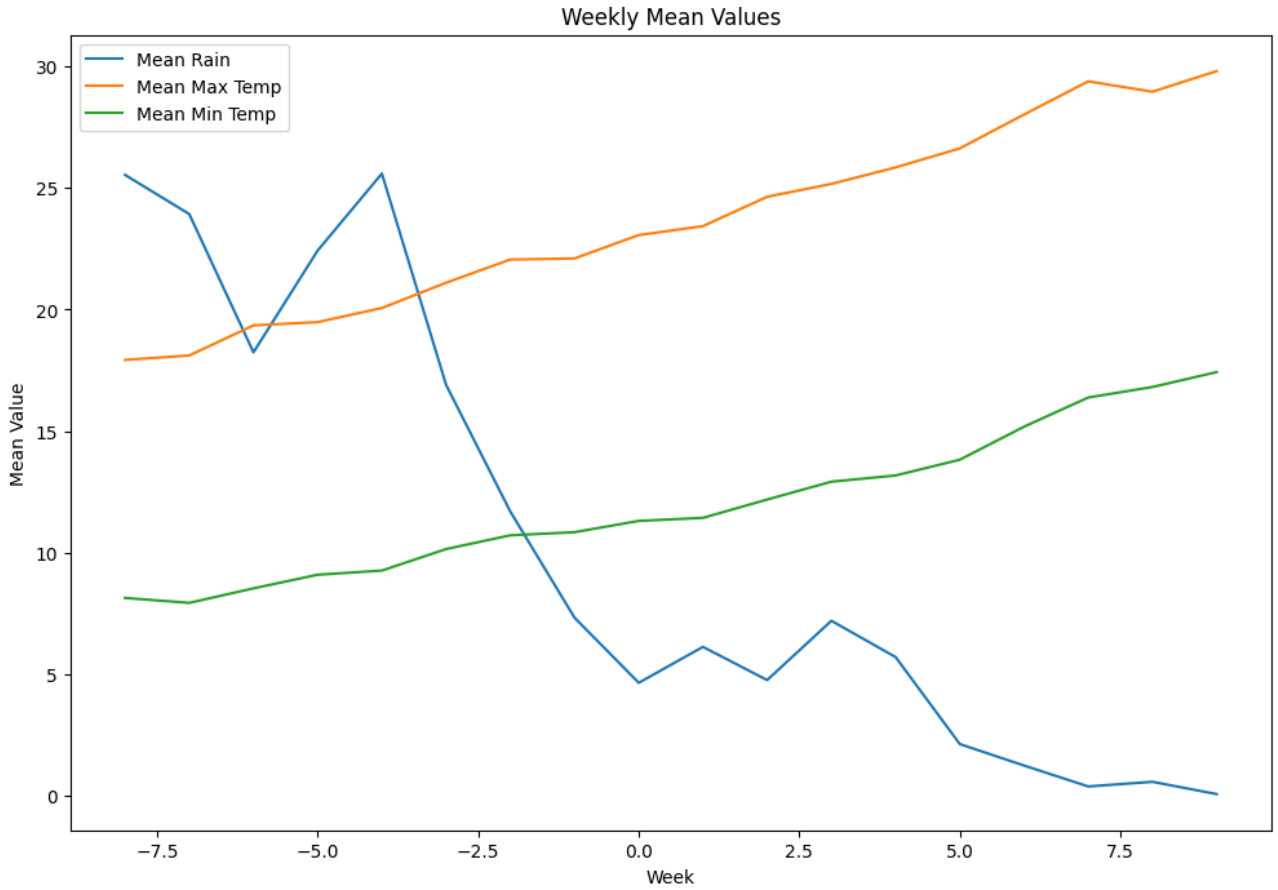

Figure 2 - weekly mean values

The rainfall values are generally decreasing as the weeks pass by, which can be expected. Note that each week represents a dynamic timeframe, that is relevant for each record separately, and so the x axis does not mimic a timeline. We can see that there is a constant gap between the maximum and minimum temperature graphs, that shows the difference between the values of each of these feature

domains. Both lines are increasing, which can we expected as the temperature should increase during this time of year.

After aggregating the data in a weekly interval, we noticed that there were some identical records, that had similar values in all fields except the target variable. This happened because the dataset included some individual sowings that took place in the same farm area and in the same date, that naturally resulted in different yield. To increase the accuracy of the machine learning models we used later, we aggregated those records by averaging their yield values. To the same end, we then opted to remove the rainfall data for week t_9 due its sparse nature. We also examined if they were any correlated columns with a threshold of +/-0.75, but there were not.

This resulted in a dataset of 294 records and 54 features ï 17 rainfall fields, 18 maximum weekly temperature fields, 18 minimum weekly temperature fields and 1 target field. Clearly the number of records is not sufficiently larger than the number of fields, and so it can be expected that it will be difficult for the machine learning models to properly learn the complexity of the data.

The initial testing was conducted using the target variable in its natural form, namely with its continuous values. However, the results were not satisfactory. We then chose to change the target variable's form to a categorial field with 3 classes. After the conversion, we observed improvements in the results. At first, we divided the values to classes while making sure the size of each categorial group is similar. This method is used to create bins of equal frequency, while the width of each bin may not be uniform across all bins:

- category 0 - a group of 98 samples, with values ranging from 336 to 538.
- category 1 - a group of 99 samples, with values ranging from 539 to 600.
- category 2 - a group of 97 samples, with values ranging from 601 to 732.

However, we found that using a different method of binning that does not aim to balance the number of samples in each of the classes, had enhanced all the measures we examined. This method aims to create bins of equal width and similar distribution of data within each bin. This usually results in a different number of samples in each bin, as seen in the following imbalanced classification:

- category 0 - a group of 168 samples, with values ranging from 336 to 468.
- category 1 - a group of 97 samples, with values ranging from 469 to 600.
- category 2 - a group of 29 samples, with values ranging from 601 to 732.

Essentially, the cut-off value between category 0 and category 1 has decreased, and now the width of each bin is the same as the others. Samples are distributed unevenly among all bins, but each bin's samples yield a similar distribution.

4.3. Applying machine learning models

We employed several machine learning models: support vector regression [12]/classifier [13] (SVR/SVC), K-nearest neighbors (KNN) [14], decision tree [15], random forest [16], gradient boosting [17], and linear and logistic regression (each for each form of the target variable).

Logistic regression for multiple categories (or multinomial logistic regression) extends the binary logistic regression model to handle more than two categories in the target variable. Each category has a separate set of coefficients. The logistic regression model predicts the probability of each category using the softmax function, which is a generalization of the sigmoid function used in binary logistic regression. It converts a vector of values into a probability distribution, where the probabilities of each value are proportional to the exponential of that value. The output of the softmax function is a vector of values that sum to 1. To train the multinomial logistic regression model, we use maximum likelihood estimation (MLE) to find the coefficients that maximize the likelihood of the observed data given the model. We also employed a basic deep learning model after grid-searching for the best hyper parameters: Rectified Linear Unit (ReLU) activation, 256 neurons in the first layer, 128 neurons in the second layer, 64 neurons in the third layer and an adam optimizer. ReLU introduces non-linearity into the network by setting the output value to the maximum of the input value and 0, meaning that only if the input is positive the output will be the same as the input, which helps the model learn complex patterns in the data. The Adam optimizer is an adaptive learning rate optimization algorithm. It adjusts the learning rate for each parameter dynamically based on the estimated first and second moments of the gradients.

All tests (including the confusion matrix's) were conducted while cross-validating using 10 folds, and for the categorial form of the target variable we tested the

models both for X as it was driven after preprocessing and for X that was scaled using MinMaxScaler. MinMaxScaler transforms the features by scaling each one of them to a range between zero and one. This is done by computing the minimum and maximum values of each feature and then transforming the data accordingly.

5. Results

Using the target variable with its continuous values:

Model	Measurement	Score
Linear Regression	MSE	7574.204
	R^2	-0.661

Table 3 - linear regression model's results

Using the target variable after converting the values to classes, the resulting accuracy measure of each model is as follows:

Model	Score for Non-Scaled X	Score for Scaled X
SVC Linear Kernel	0.469	0.541
SVC RBF Kernel	0.555	**0.565**
SVC Polynomial Kernel	0.494	0.521
KNN	0.534	0.5
Decision Tree	0.432	0.436
Random Forest	**0.568**	**0.568**
Gradient Boosting	0.463	0.473
Logistic Regression	0.466	0.554
Deep Learning	**0.565**	0.521

Table 4 - models' results

The tested measurements are as follows:

Measurement	Score for Non-Scaled X
Accuracy	0.5748
Precision	0.5065
Recall	0.5061
F1	0.5043

Table 5 - resulting measurements' values

The confusion matrix is as follows:

Values for Non-Scaled X	Predicted		
	Class 1	Class 2	Class 3
Actual Class 1	12	12	5
Actual Class 2	14	118	36
Actual Class 3	5	53	39

Table 6 - confusion matrix

6. Discussion

A high MSE indicates that the predicted values are far from the actual values, implying that the model is making large errors in its predictions. The magnitude of the MSE is proportional to the scale of the target variable, and as the yield ranges from 200 to 800, a score of 7574 is high.

An R² score indicates how well the model explains the variance in the target variable. Negative R² scores suggest that the model is performing worse than a simple horizontal line (mean prediction) and is a strong indicator of poor fit. R² it is not the square of any term, thus can be a negative number. The R^2 score is defined as: $R^2 = 1 - \frac{SSres}{SStot}$, where SS_{res} is the residual sum of squares (the sum of squared differences between the observed and predicted values), and SS_{tot} is the total sum of squares (the sum of squared differences between the observed values and the mean of the observed values). A negative R^2 score occurs when SS_{res} is greater than SS_{tot}. This indicates that the model's predictions are worse than simply using the mean of the observed data as predictions. Thus, using the natural form of the target variable would be a bad choice.

The results for the categorial form of the target variable's values indicate that the scaling X yield different results for most of the models, but the maximum accuracy value remains the same: 56.8% accuracy using the random forest model, followed closely by the classic deep learning model when using non-scaled X, and by SVC with an RBF kernel when using scaled X. We chose to focus on the measure of accuracy because it is crucial for the chosen model to be able to predict the yield's class as accurately as possible, and this measure measures the proportion of correctly classified instances out of the total instances. The result is slightly worse for other examined measures, and lands at about 50% for all of them.

Even though the best model's performance is not satisfactory, we have managed to witness interesting characteristics by studying the most meaningful features and the relationships between them. By delving into the random forest classifier's results, we can witness the importance of the features. These scores reflect how much each feature contributes to reducing the Gini impurity in the forest's decision trees. The Gini impurity is a measure of how often a randomly chosen element from the set would be incorrectly labeled if it was randomly labeled according to the distribution of labels in the set. The top 20 features:

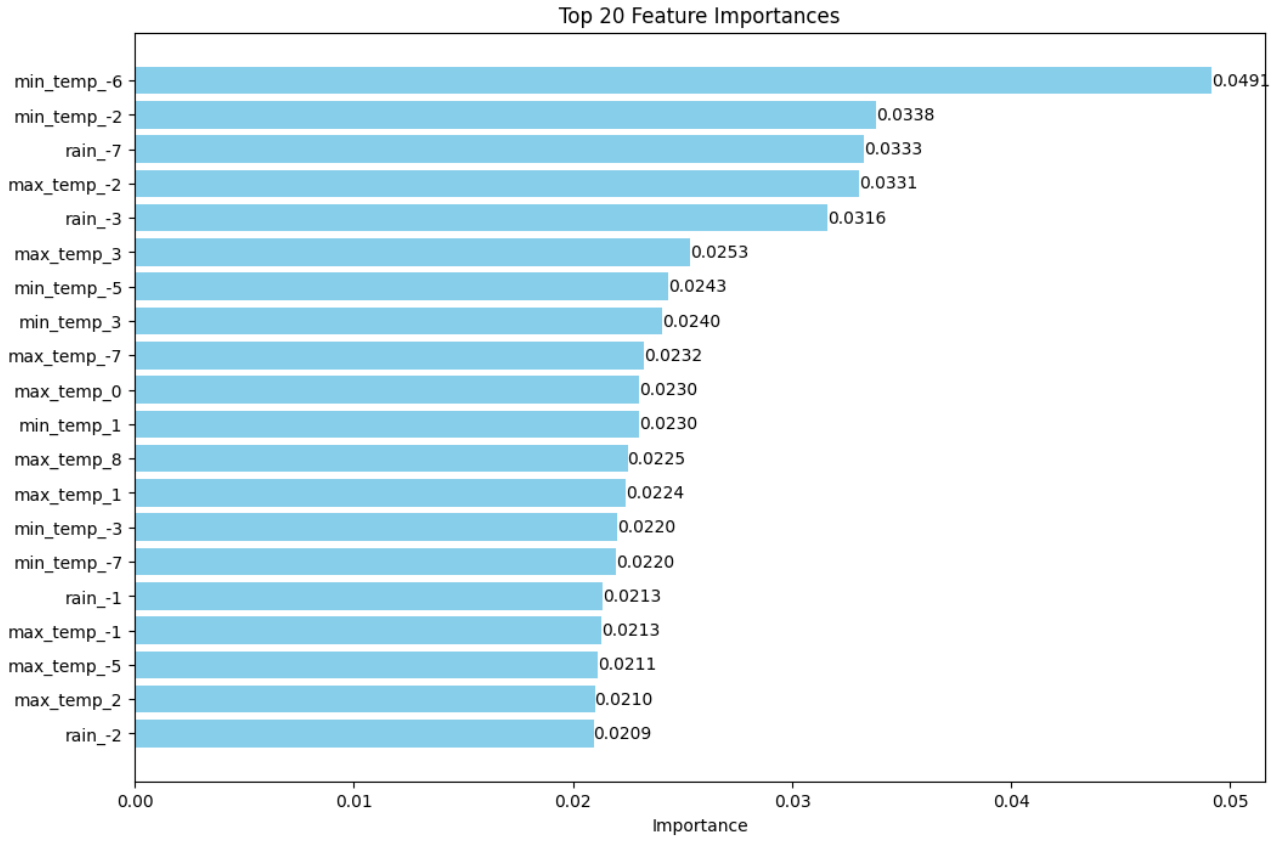

Figure 3 - random forest, top 20 important features

It seems like the rain features have lesser effect on the predictions of the random forest model, meaning that most of them are not top influential factors used by the model to make predictions about the yield's class. The rest of the fields have modest effect. Similarly, we investigated the top 20 most influential factors by magnitude, that affected the logistic regression model's classification:

Class 0 (low yield):

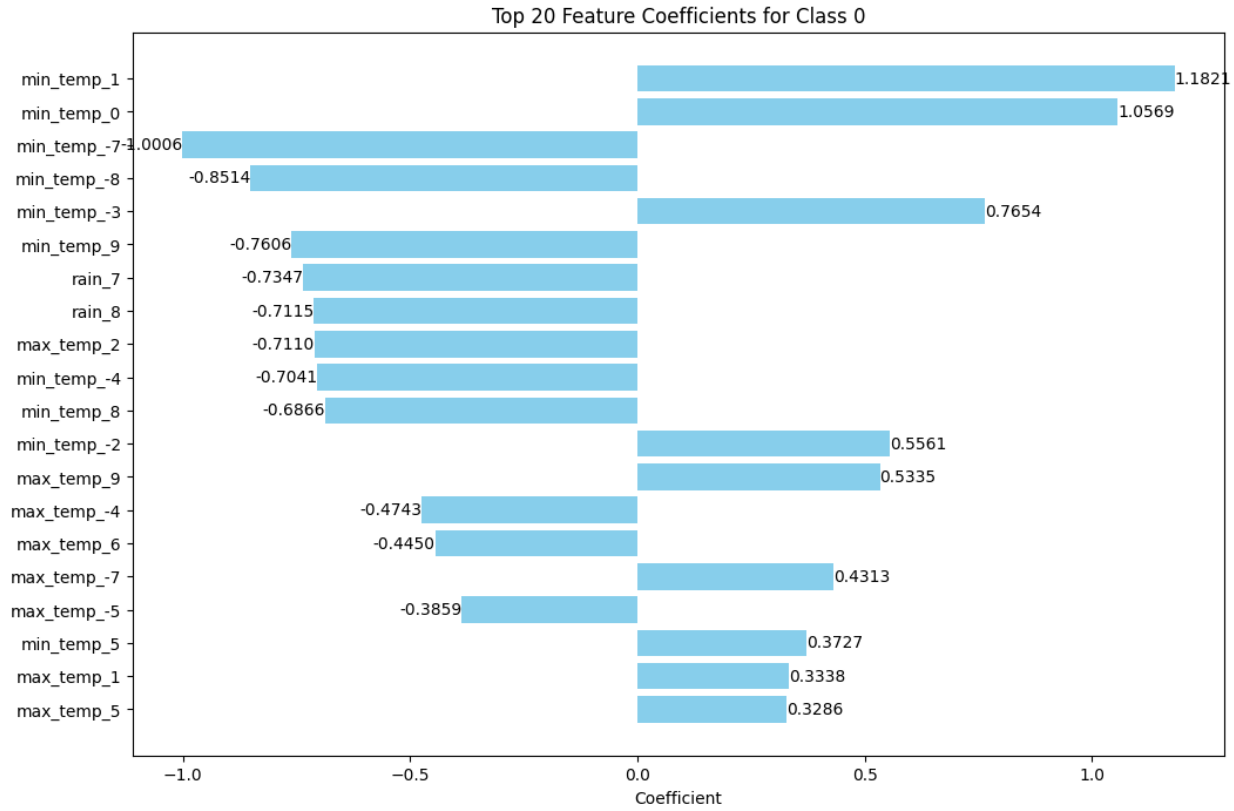

Figure 4 - logistic regression, class 0, top contributing features

Class 1 (medium yield):

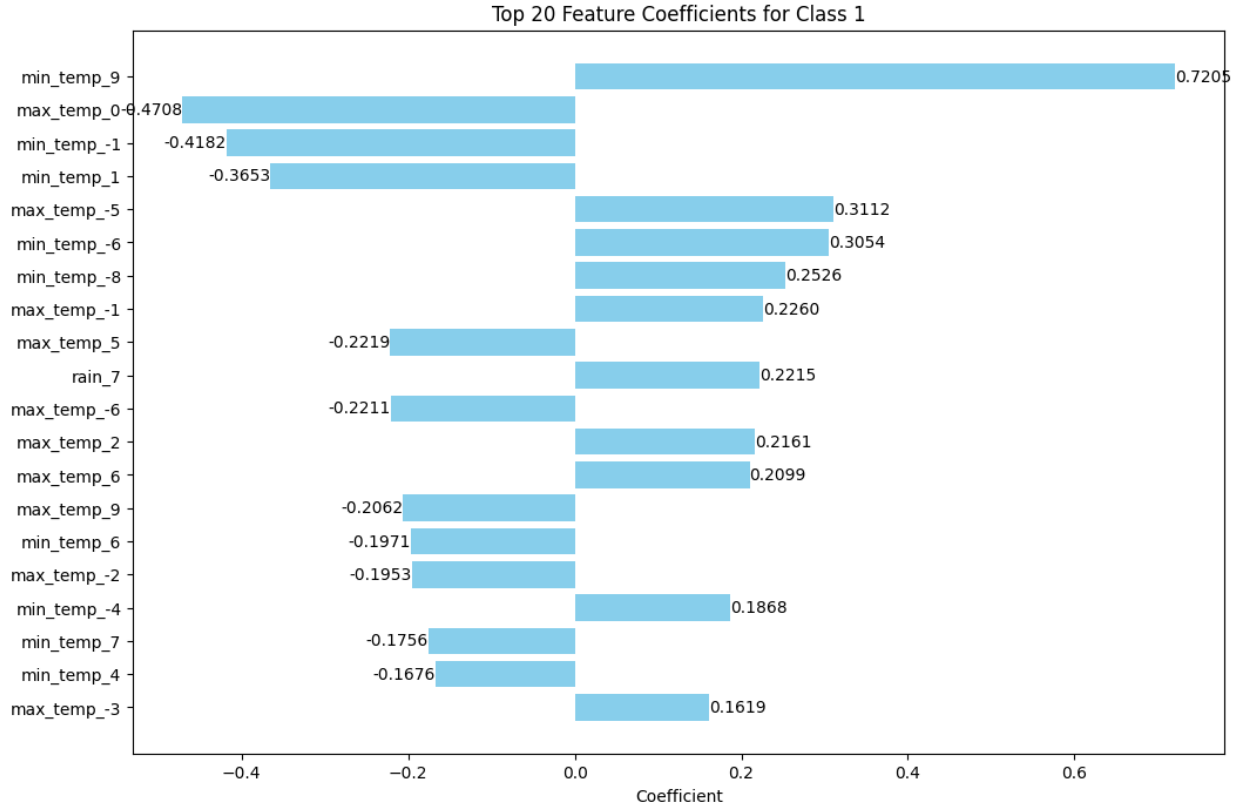

Figure 5 - logistic regression, class 1, top contributing features

Class 2 (high yield):

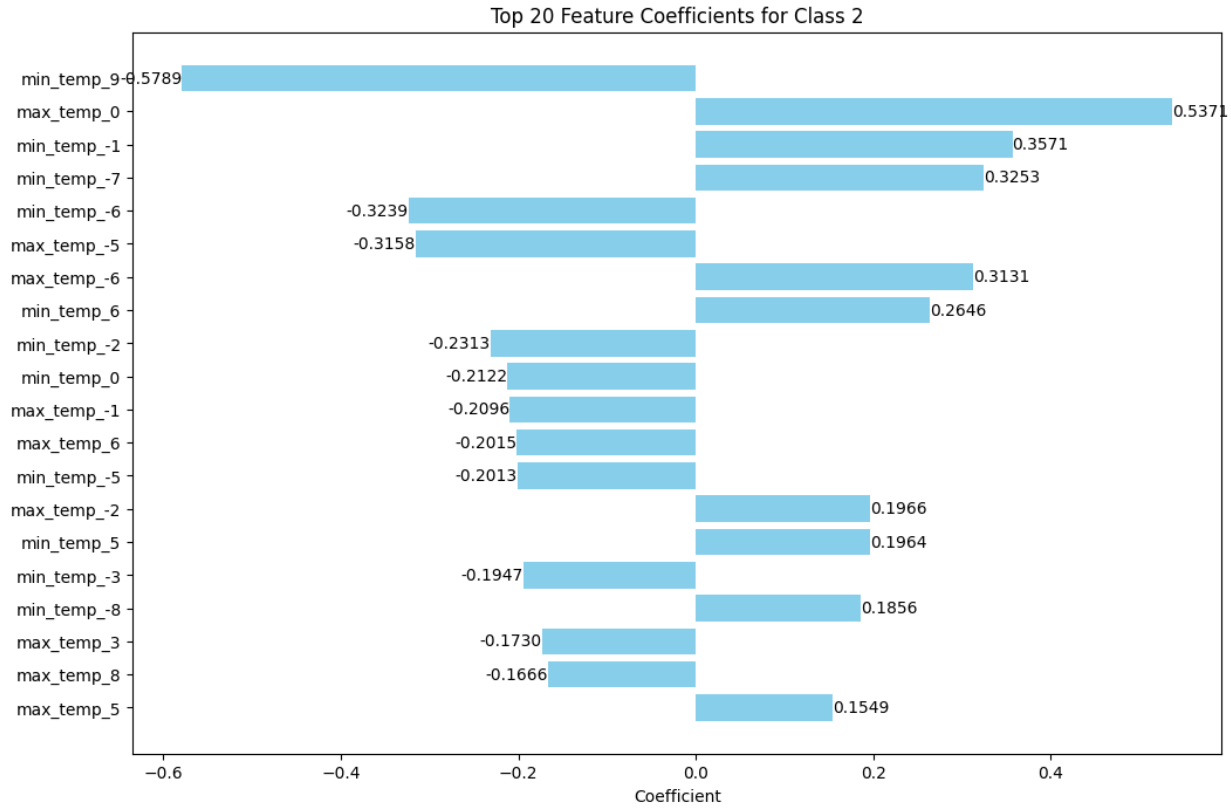

Figure 6 - logistic regression, class 2, top contributing features

We can break down the key features and their impact (positive or negative) on each class:

Class 0 (low yield)

- Positive influence: higher minimum temperatures at weeks -3, 0, 1, 5, and higher maximum temperatures at lags -7, 1, 5, 9.
- Negative influence: lower minimum temperatures at weeks -8, -7, -4, -2, 8, 9, and lower maximum temperatures at weeks -5, -4, 2, 6.

Class 1 (medium yield)

- Positive influence: higher minimum temperatures at weeks -8, -6, 9, and higher maximum temperatures at weeks -5, -3, -1, 2, 6; also, more rain at week 7.
- Negative influence: lower minimum temperatures at weeks -1, 1, 4, 6, 7, and lower maximum temperatures at weeks -6, -2, 0, 5, 9.

Class 2 (high yield)

- Positive influence: higher minimum temperatures at weeks -8, -7, -1, 5, 6, and higher maximum temperatures at weeks -6, -2, 0 ,5.

- Negative influence: lower minimum temperatures at weeks -6, -5, -3, -2, 0, 9, and lower maximum temperatures at weeks -5, -1, 3, 6, 8.

Similarly to the random forest's most influential factors, here both minimum and maximum temperatures are the most influential feature across all classes, but the impact direction (positive or negative) varies by week number and class.

Class 0 is more negatively impacted by lower minimum temperatures at the weeks that preceded the sowing and at the later weeks that following it, which can be expected as class 0 is where the yield was minimal, and lower minimum temperatures encourage MP, which causes lower yield, so it is expected that lower minimum temperature will not result in class 0 cases. However, it is not as negatively affected by lower minimum temperatures at the weeks that followed the sowing date. This is an interesting result, because it may indicate that the **minimum temperate at the weeks that followed the sowing do not have much impact of the yield** in an MP infected soil.
It also negatively impacted by lower maximum temperatures at the weeks that are adjacent to the sowing. These can be expected as low yield is related to strong presence of the pathogen, but it does not seem like it is negatively impacted by lower maximum temperatures at the edges of the time span, which could indicate that the **maximum temperature at the weeks that followed the sowing do not have much impact of the yield** in an MP infected soil. Also, it is positively impacted by higher minimum and maximum temperatures at the weeks around the sowing date. As class 0 is not the desired outcome for a farmer, this conclusion is coherent to the information found in the literature review ï MP prefers higher temperatures, and when MP is more prevalent, the less the yield of the cotton crops.

Class 1, like class 0, is negatively impacted by lower minimum and maximum temperatures throughout most of the weeks that precede and follow the sowing. However, it is positively impacted by higher minimum temperatures at the weeks at the edges of the studied time span, but less so by the higher minimum temperatures at the weeks that are adjacent to the sowing date. On the flip side, higher maximum temperatures at weeks that are adjacent to the sowing date are positively impacting the yield. This could be a strong indication that the **minimum and maximum temperatures at weeks around the sowing date act as a tuning factor to the yield**, as it may indicate the severity of the pathogen in the soil.

Class 2 is the desirable outcome in the eyes of the farmer, as it indicates high yield. We should expect to find values that indicate of conditions that are less favored by MP, however, we were surprised to find one odd conclusion - we noticed that **higher minimum temperatures at several of the studied weeks had positive influence**. However higher maximum temperatures had a positive influence at only a handful of weeks (as expected due to the traits and favoritism of MP). Furthermore, lower minimum and maximum temperatures were found to be of negative influence, which can be expected for the same reasons mentioned in this paragraph. For this class, the main conclusion was also derived in the previous classes' examination ï the weekly minimum temperature can serve as a tuning factor to the yield of cotton crops sowed in MP infected soil.

The goal of this study was to identify optimal rainfall and temperature conditions for growing cotton in soil infected with MP. The machine learning models we examined did not perform too well, which can be expected due to the lack of data. The poor performance also indicates a need for further research to develop a reliable model capable of accurately classifying cotton crop yields. Despite this, the study revealed significant findings regarding temperature conditions. Specifically, **weekly average minimum and maximum temperatures around the sowing date appear to play a critical role in influencing cotton yields**. These temperatures may act as tuning factors, affecting the yield outcomes.
Further research should focus on a more detailed analysis of temperature conditions on a per-week basis throughout the growing season. This approach could provide deeper insights into the specific temperature requirements at different stages of cotton growth, leading to more precise and actionable recommendations for optimizing cotton yields in MP-infected soils.

7. Acknowledgements

I sincerely thank Prof. Gilad Ravid from Ben-Gurion University for his valuable discussions and continuous support. I also extend my gratitude to Dr. Hagai Raanan from the Agricultural Research Organization ï Volcani Institute for his insightful contributions and support.

8. Bibliography

[1] S. Kaur, G. S. Dhillon, S. K. Brar, G. E. Vallad, R. Chand, and V. B. Chauhan, ñEmerging phytopathogen Macrophomina phaseolina: Biology, economic importance and current diagnostic trends,ò *Critical Reviews in Microbiology*, vol. 38, no. 2. pp. 136ï 151, May 2012. doi: 10.3109/1040841X.2011.640977.

[2] O. Degani, P. Becher, and A. Gordani, ñPathogenic Interactions between Macrophomina phaseolina and Magnaporthiopsis maydis in Mutually Infected Cotton Sprouts,ò *Agriculture (Switzerland)*, vol. 12, no. 2, Feb. 2022, doi: 10.3390/agriculture12020255.

[3] O. Degani, A. Gordani, E. Dimant, A. Chen, and O. Rabinovitz, ñThe cotton charcoal rot causal agent, Macrophomina phaseolina, biological and chemical control,ò *Front Plant Sci*, vol. 14, 2023, doi: 10.3389/fpls.2023.1272335.

[4] R. Cohen, N. Omari, A. Porat, and M. Edelstein, ñManagement of Macrophomina wilt in melons using grafting or fungicide soil application: Pathological, horticultural and economical aspects,ò *Crop Protection*, vol. 35, pp. 58ï 63, May 2012, doi: 10.1016/j.cropro.2011.12.015.

[5] M. K. Biswas and M. Ali, ñA review on characterization, therapeutic approaches and pathogenesis of Macrophomina phaseolina,ò2018. [Online]. Available: https://www.researchgate.net/publication/324113715

[6] S. Lodha and R. Mawar, ñPopulation dynamics of Macrophomina phaseolina in relation to disease management: A review,ò *Journal of Phytopathology*, vol. 168, no. 1. Blackwell Publishing Ltd, pp. 1ï 17, Jan. 01, 2020. doi: 10.1111/jph.12854.

[7] N. Marquez, M. L. Giachero, S. Declerck, and D. A. Ducasse, ñMacrophomina phaseolina: General Characteristics of Pathogenicity and Methods of Control,ò *Frontiers in Plant Science*, vol. 12. Frontiers Media S.A., Apr. 22, 2021. doi: 10.3389/fpls.2021.634397.

[8] A. Kamilaris, A. Kartakoullis, and F. X. Prenafeta-Bold¼ ñA review on the practice of big data analysis in agriculture,ò *Computers and Electronics in Agriculture*, vol. 143. Elsevier B.V., pp. 23ï 37, Dec. 01, 2017. doi: 10.1016/j.compag.2017.09.037.

[9] L. Klerkx, E. Jakku, and P. Labarthe, ñA review of social science on digital agriculture, smart farming and agriculture 4.0: New contributions and a future

research agenda,ò *NJAS - Wageningen Journal of Life Sciences*, vol. 90ï 91. Elsevier B.V., Dec. 01, 2019. doi: 10.1016/j.njas.2019.100315.

[10] S. A. Osinga, D. Paudel, S. A. Mouzakitis, and I. N. Athanasiadis, ñBig data in agriculture: Between opportunity and solution,ò *Agric Syst*, vol. 195, Jan. 2022, doi: 10.1016/j.agsy.2021.103298.

[11] C. S. Nandyala and H. K. Kim, ñFrom cloud to fog and IoT-based real-time U-healthcare monitoring for smart homes and hospitals,ò *International Journal of Smart Home*, vol. 10, no. 2, pp. 187ï 196, 2016, doi: 10.14257/ijsh.2016.10.2.18.

[12] H. DruckerŁ, C. J. C. Burges, L. Kaufman, A. SmolaŁ, and V. Vapoik, ñSupport Vector Regression Machinesò

[13] V.Vapnik, "The Nature of Statistical Learning Theory".

[14] W. G. Cochran, ñThe Combination of Estimates from Different Experiments,ò1954. [Online]. Available: https://www.jstor.org/stable/3001666?seq=1&cid=pdf-

[15] J. R. Quinlan, ñInduction of Decision Trees,ò1986.

[16] L. Breiman, ñRandom Forests,ò2001.

[17] J. H. Friedman, ñGreedy Function Approximation: A Gradient Boosting Machine,ò 2001.

9. Appendixes

9.1. Appendix A: logistic regression equations

$$logit(p) = ln(\frac{p}{1-p})$$

Class 0:

$$Logit(P(y = 0 \mid x))$$
$$= 0.0184 * rain\_-8 + 0.0219 * rain\_-7 + 0.0021 * rain\_-6 + -0.0023 * rain\_-5 + -0.0055 * rain\_-4 + -0.0149 * rain\_-3 + -0.0810 * rain\_-2 + -0.2104 * rain\_-1 + 0.0814 * rain\_0 + 0.0930 * rain\_1 + -0.0513 * rain\_2 + -0.0707 * rain\_3 + -0.1791 * rain\_4 + -0.0172 * rain\_5 + 0.0829 * rain\_6 + -0.7347 * rain\_7 + -0.7115 * rain\_8 + -0.2189 * max\_temp\_-8 + 0.4313 * max\_temp\_-7 + -0.1683 * max\_temp\_-6 + -0.3859 * max\_temp\_-5 + -0.4743 * max\_temp\_-4 + -0.2564 * max\_temp\_-3 + 0.1070 * max\_temp\_-2 + 0.2025 * max\_temp\_-1 + 0.3142 * max\_temp\_0 + 0.3338 * max\_temp\_1 + -0.7110 * max\_temp\_2 + 0.0945 * max\_temp\_3 + -0.2418 * max\_temp\_4 + 0.3286 * max\_temp\_5 + -0.4450 * max\_temp\_6 + 0.2553 * max\_temp\_7 + -0.1240 * max\_temp\_8 + 0.5335 * max\_temp\_9 + -0.8514 * min\_temp\_-8 + -1.0006 * min\_temp\_-7 + -0.2483 * min\_temp\_-6 + 0.0434 * min\_temp\_-5 + -0.7041 * min\_temp\_-4 + 0.7654 * min\_temp\_-3 + 0.5561 * min\_temp\_-2 + 0.1432 * min\_temp\_-1 + 1.0569 * min\_temp\_0 + 1.1821 * min\_temp\_1 + 0.2031 * min\_temp\_2 + -0.0094 * min\_temp\_3 + 0.1308 * min\_temp\_4 + 0.3727 * min\_temp\_5 + -0.0554 * min\_temp\_6 + 0.0790 * min\_temp\_7 + -0.6866 * min\_temp\_8 + -0.7606 * min\_temp\_9 + -0.0169$$

Class 1:

$$
\begin{aligned}
&Logit(P(y = 1 \mid x)) \\
&\quad = -0.0079 * rain\_-8 + -0.0122 * rain\_-7 + -0.0095 * rain\_ \\
&\quad -6 + 0.0079 * rain\_-5 + -0.0190 * rain\_-4 + -0.0427 \\
&\quad * rain\_-3 + -0.0158 * rain\_-2 + 0.0351 * rain\_-1 \\
&\quad + -0.0361 * rain\_0 + -0.0359 * rain\_1 + 0.0030 * rain\_2 \\
&\quad + -0.0152 * rain\_3 + 0.0339 * rain\_4 + 0.0566 * rain\_5 \\
&\quad + -0.0184 * rain\_6 + 0.2215 * rain\_7 + -0.0033 * rain\_8 \\
&\quad + 0.1175 * max\_temp\_-8 + -0.0646 * max\_temp\_-7 \\
&\quad + -0.2211 * max\_temp\_-6 + 0.3112 * max\_temp\_-5 \\
&\quad + -0.0322 * max\_temp\_-4 + 0.1619 * max\_temp\_-3 \\
&\quad + -0.1953 * max\_temp\_-2 + 0.2260 * max\_temp\_-1 \\
&\quad + -0.4708 * max\_temp\_0 + -0.0059 * max\_temp\_1 + 0.2161 \\
&\quad * max\_temp\_2 + 0.0864 * max\_temp\_3 + 0.0358 * max\_temp\_4 \\
&\quad + -0.2219 * max\_temp\_5 + 0.2099 * max\_temp\_6 + 0.0641 \\
&\quad * max\_temp\_7 + 0.1243 * max\_temp\_8 + -0.2062 * max\_temp\_9 \\
&\quad + 0.2526 * min\_temp\_-8 + -0.1139 * min\_temp\_-7 + 0.3054 \\
&\quad * min\_temp\_-6 + 0.1042 * min\_temp\_-5 + 0.1868 * min\_temp\_ \\
&\quad -4 + -0.0760 * min\_temp\_-3 + 0.1210 * min\_temp\_-2 \\
&\quad + -0.4182 * min\_temp\_-1 + -0.0340 * min\_temp\_0 + -0.3653 \\
&\quad * min\_temp\_1 + 0.0901 * min\_temp\_2 + 0.0152 * min\_temp\_3 \\
&\quad + -0.1676 * min\_temp\_4 + -0.1572 * min\_temp\_5 + -0.1971 \\
&\quad * min\_temp\_6 + -0.1756 * min\_temp\_7 + -0.1493 * min\_temp\_8 \\
&\quad + 0.7205 * min\_temp\_9 + 0.2799
\end{aligned}
$$

Class 2:

$$
\begin{aligned}
&Logit(P(y = 2 \mid x)) \\
&\quad = 0.0044 * rain\_-8 + 0.0134 * rain\_-7 + 0.0207 * rain\_-6 \\
&\quad + -0.0031 * rain\_-5 + 0.0255 * rain\_-4 + 0.0412 * rain\_-3 \\
&\quad + 0.0312 * rain\_-2 + 0.0053 * rain\_-1 + 0.0180 * rain\_0 \\
&\quad + 0.0072 * rain\_1 + 0.0048 * rain\_2 + 0.0353 * rain\_3 \\
&\quad + -0.0159 * rain\_4 + -0.0505 * rain\_5 + 0.0203 * rain\_6 \\
&\quad + -0.1353 * rain\_7 + 0.0693 * rain\_8 + -0.1366 * max\_temp\_ \\
&\quad -8 + -0.0430 * max\_temp\_-7 + 0.3131 * max\_temp\_-6 \\
&\quad + -0.3158 * max\_temp\_-5 + 0.1092 * max\_temp\_-4 + 0.0592 \\
&\quad * max\_temp\_-3 + 0.1966 * max\_temp\_-2 + -0.2096 \\
&\quad * max\_temp\_-1 + 0.5371 * max\_temp\_0 + -0.1429 \\
&\quad * max\_temp\_1 + -0.0360 * max\_temp\_2 + -0.1730 \\
&\quad * max\_temp\_3 + -0.0569 * max\_temp\_4 + 0.1549 * max\_temp\_5 \\
&\quad + -0.2015 * max\_temp\_6 + -0.0335 * max\_temp\_7 + -0.1666 \\
&\quad * max\_temp\_8 + 0.1131 * max\_temp\_9 + 0.1856 * min\_temp\_-8 \\
&\quad + 0.3253 * min\_temp\_-7 + -0.3239 * min\_temp\_-6 + -0.2013 \\
&\quad * min\_temp\_-5 + -0.0627 * min\_temp\_-4 + -0.1947 \\
&\quad * min\_temp\_-3 + -0.2313 * min\_temp\_-2 + 0.3571 \\
&\quad * min\_temp\_-1 + -0.2122 * min\_temp\_0 + 0.1034 * min\_temp\_1 \\
&\quad + 0.0714 * min\_temp\_2 + 0.1138 * min\_temp\_3 + 0.1365 \\
&\quad * min\_temp\_4 + 0.1964 * min\_temp\_5 + 0.2646 * min\_temp\_6 \\
&\quad + -0.0732 * min\_temp\_7 + 0.1009 * min\_temp\_8 + -0.5789 \\
&\quad * min\_temp\_9 + -0.2692
\end{aligned}
$$