Master Degree Project

UNIVERSITY OF SKÖVDE

Study of evaluation metrics while predicting the yield of lettuce plants in indoor farms using machine learning models

Master Thesis for Ljusgårda

Divya Chedayan
Harry Geo Fernandez

Supervisor: Jonas Karlsson
Examiner: Juhee Bae

Master Degree Project (15 ECTS) in Informatics
with a specialization in Data Science
Spring term 2023

ABSTRACT

A key challenge for maximizing the world's food supply is crop yield prediction. In this study, three machine models are used to predict the fresh weight (yield) of lettuce plants that are grown inside indoor farms hydroponically using the vertical farming infrastructure, namely, support vector regressor (SVR), random forest regressor (RFR), and deep neural network (DNN).

The climate data, nutrient data, and plant growth data are passed as input to train the models to understand the growth pattern based on the available features. The study of evaluation metrics majorly covers Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), R-squared, and Adjusted R-squared values.

The results of the project have shown that the Random Forest with all the features is the best model having the best results with the least cross-validated MAE score and good cross-validated Adjusted R-squared value considering that the error of the prediction is minimal. This is followed by the DNN model with minor differences in the resulting values. The Support Vector Regressor (SVR) model gave a very poor performance with a huge error value that cannot be afforded in this scenario. In this study, we have also compared various evaluating metrics mentioned above and considered the cross-validated MAE and cross-validated Adjusted R-squared metrics. According to our study, MAE had the lowest error value, which is less sensitive to the outliers and adjusted R-squared value helps to understand the variance of the target variable with the predictor variable and adjust the metric to prevent the issues of overfitting.

Keywords: machine learning, lettuce yield prediction, Regression, SVR, RF, DNN, MAE, MSE, RMSE, R-squared, Adjusted R-squared

Table of Contents

1. Introduction

Food is one of the important necessities for survival and importing food products cannot be a sustainable way of living. Sweden due to its harsh climatic conditions, some parts of the country's environment are not suitable for traditional agriculture on farmlands. According to (Taghizadeh, 2021), the conventional way of agriculture has a high chance of failing in some places. Furthermore, relying heavily on imported food makes the food supply systems of importer nations vulnerable to unexpected shocks. In fact, the COVID-19 pandemic caused difficulties in the food trade, since transport vehicles were not allowed to enter other countries. Feeding a growing global population is one of the biggest difficulties, as it is predicted by 2050, global food production will need to expand by 50% (Alexandratos & Bruinsma, 2012). Another challenge is that food production must be increased at the same time reducing agriculture's negative impacts on the environment (Rhodes, 2019). As a solution to this and to become a self-sustainable nation in food production, the agricultural firms in Sweden have introduced indoor farms. However, it is hard to grow lettuce plants in Sweden with the severe climatic conditions and the farmers are facing difficulties in predicting the crop yield accurately (Shan, 2021). In order to overcome this issue, we have indoor farms where the cultivation is happening in a controllable environment with the best-growing conditions created artificially in an indoor setup on a large industrial scale. When growing lettuce on a giant scale indoors, it is important to generate maximum results and to take measures to make sure that the plant yields are improved by analysing all the possible parameters contributing to plant growth.

In a controlled environment agriculture (CEA), the data are collected and exchanged using the Internet of Things (IoT) such as sensors and actuators that are embedded with electronics, an interconnected network of physical devices that helps in the periodic monitoring of the environment for efficient growth of plants (Srivani & Manjula, 2019). These Internet of Things continuously help to monitor and optimize the growing conditions of the lettuce plants such as temperature, CO2 level, humidity, water pH and EC (Electrical Conductivity), light intensities, and nutrient levels. Thus, improving the health and the quality of lettuce plants resulting in more efficient production throughout the year.

Vertical farming is the process of growing plants in cylindrical tubes/pipes in layers that are stacked vertically to maximize the use of available space. The controlled environment agriculture (CEA) approach, in which the total environment can be controlled and managed is the major foundation of vertical farming (Ullah et al., 2023). This technique makes it possible to produce food during the whole year. The irrigation system, room temperature, and plant conditions are also considered in vertical farming for a good yield. The greatest advantage of vertical farms is that they provide organic minerals and enzymes to promote healthy plant growth and this technique allows a high mineral intake throughout the crop life cycle (Farooq et al., 2019). The basic essential ingredients for plant growth are proper sunlight, moisture, and heat. Thus, vertical farming techniques should take care of these parameters for the better growth of plants.

The aim of this thesis is to predict the yield of lettuce plants by using a large set of data recorded by sensors and manually by humans. This data consists of a combination of environmental data and plant growth data. Predicting the yield can be helpful in understanding how each factor can affect the growth of the lettuce plant and deciding its quality to categorize it for sale.

1.1 Problem definition

During the past years, changes in the climate and weather patterns led to the current land and water shortages, which in turn affected the agricultural sector (Majid et al., 2021). The water consumption is very high in the agricultural farms, and it is estimated that 70% of water consumer is from this specific area (Kloas et al., 2015). The application of modern technologies has improved the human's capacity to overcome the limited resources problem. Hydroponic systems are considered to be the best way to overcome the issues faced with traditional agricultural systems (Majid et al., 2021). Installing a hydroponic system can be very expensive, but it will help the agriculture sectors to gather the environmental and the plant growth data to predict the crop yield using machine learning models. The objective of our study is to apply different machine learning models to predict the various feature affecting the lettuce plant yield based on two scenarios made up of different combinations of input variables and then finding the optimal model scenario. A study conducted by (Hong et al., 2022) helped us to understand various water nutrients promoting the growth of lettuce plants. The ability of lettuce plants to grow, produce stems, and increase quality are all considerably improved by nutrients like N, P, and K. The ideal amount of N, P, and K applied lettuce grown in Wuhan, China was 315 kg N, 210 kg P2O5, and 285 kg K2O respectively. In their experiment, the lettuce leaves had the highest N and K accumulations, whereas the stem contained the highest P accumulations.

1.2 Aim of this study

Based on the requirements provided by the company, we have arrived at the below problem definitions to describe the problem. These research questions will help to conclude the research work and provide the results demanded by the company. This study focuses on answering the following research questions:

1.2.1 What are the major features that affect the improvement in the yield of lettuce and predict the yield based on these features?

1.2.2 Previous studies show that various models perform well when predicting crop yield. Which one is the best in our case?

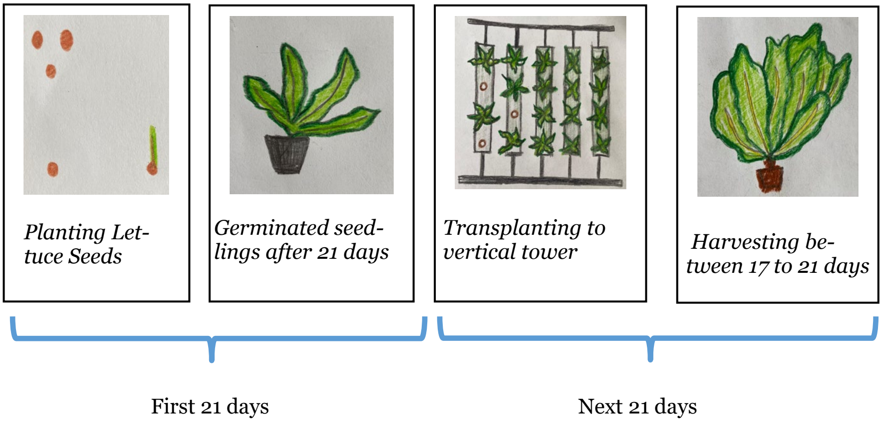
1.2.3 How different evaluation metrics are used to evaluate the performance of the models that predicts the yield of lettuce plants and which metrics are preferred in this study.

2. Background

In indoor farms, lettuce plants are cultivated in artificial growing conditions, with all the required environment setup. Several advanced sensors and equipment are used to control this environment and also to collect data about the environment and plant details at particular intervals of time for future analysis. The whole plants are divided into 15 different batches, and each batch is tested with varying growth conditions. These data put together can be analysed by using data mining techniques and machine learning algorithms to understand what are the optimum environmental conditions required for maximum plant yields. Many different studies are conducted in the same area across the world. Here, in this study, we are trying to understand the effects of variables which are a combination of climatic features, water nutrient features, and the physical and genetical features related to the growth of the lettuce plants.

2.1 Understanding the business process

In this section, we will be discussing the process of growing lettuce plants in Ljusgårda. The whole process is explained as shown in Figure 1. The company provided a tour of its production section to explain the process from seedling to harvesting. The seeds are planted in different trays and are kept in dark rooms for 21 days and no nutrients are provided at this time. Watering consistently helps the seeds germinate. There are no details collected during the germination process. Once germinated after 21 days, the plants are transplanted to vertical towers or tubes with different combinations of nutrient recipes provided for the growth of plants. During this stage onwards, they start to collect the plant data and environmental data from day 1 to day 22 after transplanting (DAT) to predict the yield of lettuce plants. The tower features such as the number of plants per tower, the separation distance of towers, and the distance between each plant are fixed. The maintenance of nutrient recipe levels in the hydroponic towers and the light recipes are maintained through a constantly monitored system. Between 17 to 21 days of transplantation, the plants are harvested without any delay, packed, and shipped on the same day to save shelf life, because the shelf life of the lettuce plant is only 14 days.



Planting Lettuce Seeds

Germinated seedlings after 21 days

Transplanting to vertical tower

Harvesting between 17 to 21 days

First 21 days

Next 21 days

Figure 1: Process of growing the lettuce plants

2.2 Research background

In this section, a few background research papers are mentioned which helped to understand the methods needed when predicting the crop yield. The first two research papers are elaborated briefly; we took the main reference from these papers.

(Mokhtar et al., 2022) conducted a study to predict the lettuce yield grown hydroponically. They monitored the lettuce yield prediction using four machine learning (ML) models. The models used are support vector regressor (SVR), random forest (RF), extreme gradient boosting (XGB), and deep neural network (DNN). They used three different hydroponic technique systems to cultivate the lettuce plants such as a pyramidal aeroponic system, a suspended nutrient film technique system, and an aeroponic tower system. These lettuces were grown in a greenhouse where the environment is controlled to promote the growth of lettuces. The data was collected during the years 2018 and 2019. In their experiment, they set three scenarios with different combinations of input variables such as leaf counts, water intake, stem measurement, and dry waste. The result showed the least RMSE value for the XGB model, with scenario 3 having all the input variables, followed by SVR with the same scenario. The RF model with scenario 1, having inputs like leaf counts and water intake, scored the highest yield. They calculated the Scatter Index (SI) values of all the models by dividing the RMSE value by the average values of the observed yield. The SI values are used to evaluate the performances of the models (Samadi et al., 2021). The SI values for all the model scenarios were less than 0.1, which indicates that all the models performed well. According to the researchers based on their experiment, they concluded the best two models were SVR having all input variables with scenario 3, and the DNN model with scenario 2 that has taken leaf counts, water intake, and dry weight as input. A DNN is an artificial neural network with multiple layers between the input, hidden, and output layers, which helps to examine more complex non-linear relationships between the input and output. According to them, among the two best models, DNN with fewer inputs is preferred, and the capacity of the DNN model to predict the yield of fresh lettuce is promising, and it is evaluated that the DNN can also be a rapid tool for the decision makers to manage crop yield on large scale.

(Gertphol et al., 2018) developed predictive models for predicting crop yield. According to the researchers, yield forecasting is significant for farmers because that affects the quantity and the quality of the crop due to climatic changes happening. A large dataset on the environmental conditions in which the lettuce is grown was generated using the Internet of Things by deploying a smart hydroponic lettuce farm. In this study, regression models using machine learning (ML) techniques were created. They constructed different models such as Support Vector Regression (SVR), Multiple Linear Regression (MLR), and Artificial Neural Networks (ANN) and integrated linear models to predict the yield on a weekly basis. They considered the features such as the intensity of light, humidity, and temperature, along with the measurement of plant growth. The target variables used were the total fresh weight, nitrate content, leaf counts, and the area of the leaf. Two scenarios were introduced to perform predictions. The first scenario contains a set of data for the predictive models that consider crop yield during farming. As a result, the farmer will need to take measurements of the plant growth every week which can be a time-consuming task. The plant growth data and the environmental data from the first to fifth weeks of farming are the features utilized in this scenario. These features can be adjusted to weekly environmental data for 168 features and the growth data for 25 features, in total 194 features are considered. In the second scenario, they used the dataset which only considers the environmental factors collected weekly which is 168 features and the farmer does not have to make the extra effort to measure the plant growth every week. The data from the sixth week is the harvesting week and is used as the target for the model to learn both scenarios. According to the study, the prediction using scenario 1 is better than scenario 2, which the researchers already expected. Still, they were excited to see that the error rates for the two scenarios were mostly identical during the first three weeks. The researchers implemented integrated models to predict crop yield, and the result shows that their model outperformed SVR. Finally, they suggest not sticking to a single model to produce an accurate crop yield prediction. The later weeks have improved accuracy when comparing weekly forecasts. The author mentioned that this study did not implement the Feature Selection technique, so the result was not good enough. Taking measurements of plant growth may encounter some error levels as it is performed manually. The result could have been better if they had a sensor-based system to collect the data without any error and apply the feature selection technique. As per the guidance provided, the SHAP method can also be used to identify the most influential features of the target variable because of the various benefits it has over other methods. Various other studies have been performed in the field of crop yield predictions. (Jeong et al., 2016) used Random Forest (RF) and Multiple linear Regression (MLR) models to predict the yield of various crops like wheat, maize, and potato at the local and global scale. According to them, RF outperformed with higher accuracy and precision than MLR's desired outcome and was determined to be extremely proficient at predicting crop yield. A study on comparison between the Artificial intelligence models for predicting crop yields such as corn and soybean was conducted by (Kim et al., 2019). They build an optimized DNN model to predict the crop yield and according to them, this model could be used in the new region only if the parameter optimization is conducted for the new region. Another study was conducted by (Kamir et al., 2020) to predict the yield of wheat across Australia using various machine learning models to reduce the risk of food shortages due to the growing global demand. The machine learning regression models were constructed, in which climate data such as rainfall and the maximum temperature and the satellite image time series were passed. These data helped them to find the yield gap hotspots.

SHAP Feature Selection is performed using the models to understand the features that really impact the target feature. A SHapley Additive exPlanations, in short SHAP (Lundberg & Lee, 2017), is a visualization tool that is used for explaining the impact of features on the machine learning model outcomes. There are several machine learning models that make predictions with great accuracy and performance. There is one limitation with these models as we are unable to explain how well they create results. It is always necessary to explain the results generated by the models. SHAP is one among many popular methods that can be used to identify the contribution of individual features to the model's prediction. SHAP works to decompose the model's output by the sum of the impact of each feature. An extended study conducted by (Scavuzzo et al., 2022) based on (Lundberg & Lee, 2017) used SHAP to find out the influence and the interactions between the variables in the generated models. The SHAP analysis and visualization helped to understand how all the variables interact with one another and how this relationship is reflected in the model. SHAP showed the variables having the greatest influence on the target feature. In our case, we have decided SHAP because of some of the benefits it has over other models. It is based on cooperative game theory which ensures consistency in identifying the contribution of individual features, which is more reliable. Shapley values are easier to understand at the same time can handle complex models. It can rank the features based on the importance of the features and also can identify the relation within the features itself.

2.3 Machine learning models

2.3.1 Support vector regressor (SVR)

A support vector regressor (SVR) is a type of machine learning algorithm which predicts a continuous output variable that is used for regression problems. This is a supervised learning algorithm that helps to predict target values (Shi et al., 2022). Support Vector Machine (SVM) is a supervised algorithm for classification, proposed by (Vapnik et al., 1992). Support Vector Machine (SVM) are a class of algorithms for classification, regression and other applications that indicate the current state of art in the field. Later in 1996, Drucker et al., introduced the regression technique which is based on Vapnik's concept on support vectors. SVR is based on the same principles as the support vector machine (SVM) which indeed is used for classification tasks (Mokhtar et al., 2022).

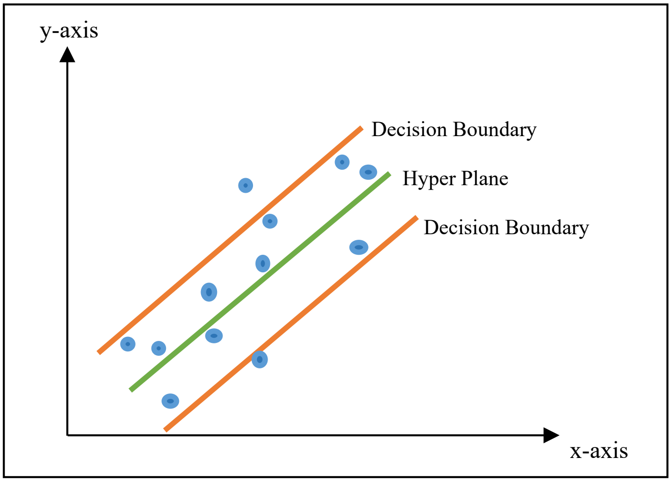


Figure 2: Support Vector Regressor

An SVR seeks to minimize the differences between the actual and the predicted output values by locating the possible line or hyperplane that may divide the data points in feature space as shown in Figure 2. The input data are mapped by the SVR algorithm to a high-dimensional feature space where an optimization problem is used to determine the decision boundary.

The performance of the model is determined using the following hyperparameters.

- Kernel: The process through which the lower-dimensional data are mapped to higher dimensional data.
- Hyper Plane: A hyperplane is a line that separate data classes, and used to predict the continuous or target value.
- Decision boundary: A decision boundary is represented by a hyperplane that maximizes the distance between the support vectors and the boundary and thereby, minimizes the prediction error.
- Support vectors: The data points nearest to the decision boundary.

SVR is a powerful and flexible algorithm that can be used for a wide range of regression problems. There are different types of SVR algorithms such as linear SVR, non-linear SVR, and epsilon-insensitive SVR. According to each algorithm, each one has a different approach to determine the best possible decision boundary for the set of given data.

The function for SVR is based on the concept of finding a hyperplane or decision boundary that separates the input data into two classes. In SVR, the goal is to find a hyperplane that separates the input data in such a way that the predicted output values are as close as possible to the actual output values.

The function of SVR can be represented as follows:

$$y = f(x) + \varepsilon$$

Where x is the input variable, y is the output variable, and ε is the error term or noise in the data. The objective of the support vector regressor algorithm is to calculate the value of f(x) that minimizes the error term ε, subject to a set of constraints. These constraints are based on the distance between the predicted output values and the actual output values, along with the margin or distance between the decision boundary and the support vectors. The function f(x) can take different forms depending on the type of SVR algorithm being used. In linear SVR, the function f(x) is a linear function of the input variables and in non-linear SVR, the function f(x) can be a polynomial function, radial basis function (RBF), or another non-linear function.

Advantages of SVR

- Non-Linear Relationships: SVR can capture the non-linear relationships between the features and the target variable.
- Robustness to Outliers: SVR is less sensitive to outliers in the data compared to traditional linear regression methods.
- Effective in High-Dimensional Spaces: SVR is ideal for complex datasets with numerous features because SVR effectively performs even in high-dimensional feature space.
- Global Optimization: SVR aims to find the best-fitting line that has the maximum margin from the data points, leading to a more robust model.
- Flexibility: The various kernel functions allow SVR to model complex relationships in the data accommodating various data patterns.

Disadvantages of SVR

- Data Scaling: The performance of SVR can be affected by the scale of the features. So, it is frequently required to preprocess the data by scaling or normalizing the features before training the model.
- Memory Consumption: SVR can use a lot of memory in addition to the computational intensities, especially when working with huge datasets.
- Computationally Expensive: Training an SVR model is computationally expensive when dealing with huge datasets. As the dataset size increases, the time complexity can become a concern.

- Overfitting with complex Kernels: SVR is capable of modeling complex relationships, however highly complex kernels can result in overfitting if not well controlled.
- Limited to Large Datasets: SVR may perform poorly when working with very large datasets, as it might struggle to handle the computational demands.
- Interpretability: While SVR can produce accurate predictions, interpreting the model's result, especially using non-linear kernels.

2.3.2 Random Forest regressor (RFR)

Random Forest is another popular machine learning algorithm that is used for both regression and classification problems. In regression, the Random Forest algorithm works by building an ensemble of decision trees, where each tree is trained on a random subset of features and a random subset of the training data. To make a prediction, each decision tree in the Random Forest independently produces a prediction, and the final prediction is the average of the individual tree predictions. The Random Forest model was developed by (Breiman, 2001). This model uses the bagging method in order to ensemble a group of decision trees with controlled variances. Even though Random Forest can be used for both regression and classification, it is mostly used for regression problems and for making predictions. A particular kind of bootstrap ensemble is RF regression. In bootstrapping, a random subset of the training dataset is selected from the raw dataset and used to develop the model. It works with the random binary trees that use the subset of the observations. The RF algorithm is explained step by step below (Maheswari & Ramani, 2023) :

Step 1: K occurrences are selected from the training sample randomly.

Step 2: The decision trees are generated for the selected occurrences.

Step 3: The number of estimation techniques to be produced is mentioned by N

Step 4: Iterate Steps 1 and 2.

Step 5: The class with the most votes is chosen after the estimates of each prediction model are calculated for the new sample.

The Random Forest algorithm has several advantages compared to other regression algorithms. It is robust to noise and outliers in the data. It has the capacity to handle a large number of in-out features and is computationally efficient. Another main advantage of RF is it will not accept null values (Geetha et al., 2020). Other than these advantages, it can also be used to estimate the importance of each input feature in the prediction, which can be useful for the feature selection and interpretation of the model. In our case, the best score is obtained using the Random Forest model which was trained using 44 trees.

Advantages of Random Forest Regressor

- Robust to Outliers and Noise: Random Forest are less sensitive to outliers and noisy data compared to a single decision tree due to the algorithm's ensemble structure.
- Non-Linear Relationships: Random Forest are excellent for a variety of regression issues because they can capture complex non-linear correlations between features and the target variable.
- Feature Importance: Random Forest can provide insights into feature importance helping to identifying the features that are most important and contributing the most to predictions. Thus, feature selection may benefit from this information and understanding the problems.
- High Predictive Accuracy: Random Forest often provide high predictive accuracy due to their ability to reduce overfitting by averaging predictions from multiple decision trees.
- Handling Missing Values: Random Forest are known to handle missing values in the dataset without the requirement for imputation.
- Scalability: Random Forests can be parallelized to speed up the training on multiple-core processors and can handle a wide range of dataset sizes.

- Reduced Risk of Overfitting: The ensemble approach of combining multiple decision trees reduces the risk of overfitting.

Disadvantages of Random Forest Regressor

- Memory and storage requirements: Random Forests can use a lot of memory, particularly if the ensemble includes a lot of decision trees.
- Training Time: Since several decision trees are constructed, training a random forest might take longer than simpler algorithms like linear regression.
- Complexity and Interpretability: The predictions made by Random Forest can be challenging to interpret particularly for large, complicated datasets with multiple decision trees.
- Bias and Variance: Random Forests typically exhibit little bias, but they can still exhibit variance especially if there are too many trees. Balancing bias and variance are important for optimal model performance.
- Hyperparameter Tuning: Various hyperparameters should be tuned to get optimal performance. Tuning then could be time-consuming and require cross-validation.

2.3.3 Deep neural network (DNN)

The DNN is a powerful Deep Learning model, which stands for Deep Neural Network. (Fukushima, 1980) was the first researcher to create a mathematical model of a neural network known as Neocognitron, which was using (Hubel & Wiesel, 1962) discoveries. DNN is a type of artificial neural network that has multiple hidden layers between the input and the output layers (Mokhtar et al., 2022). The working of DNN architecture is shown in Figure 3. As the name suggests, deep refers to the fact that these networks are composed of several layers, which help them to learn more complex and abstract features of data. The deep neural network is composed of multiple layers of interconnected artificial neurons that work together to learn and extract patterns from input data. The figure below, explains the DNN architecture where X input are initial layer passed into the

DNN. These nodes accept the input and then process it mathematically and then transfer the outcome to the next layer.

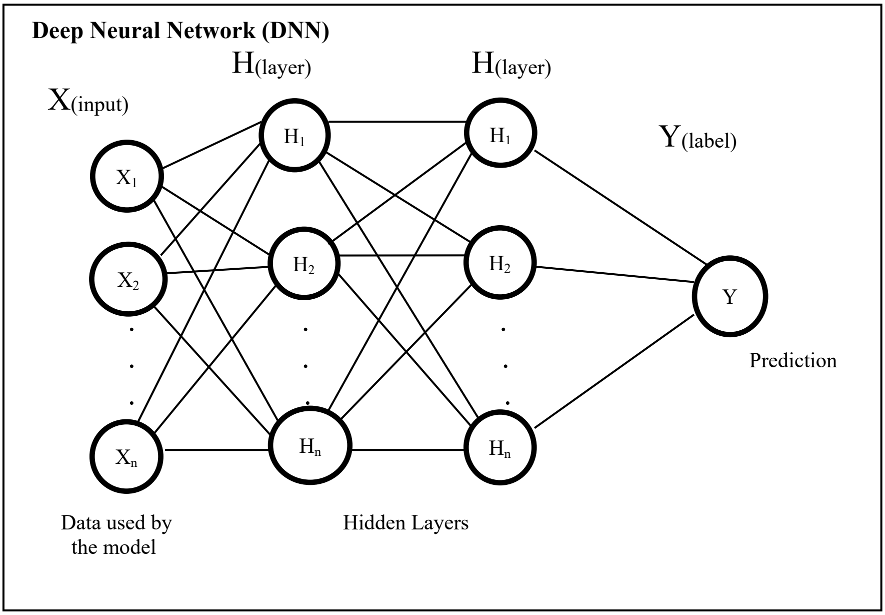


Figure 3: DNN Architecture

Later hidden layer can extract this data information through backpropagation. This helps the hidden layer to learn the characteristics of the features. In the final stage, another layer uses an activation function based on our problem, and finally returns a prediction output $Y_{(label)}$. Each layer in DNN is made up of several nodes which are known as neurons. The rectified linear unit (ReLU) is frequently used activation function to build input-output relationships (Bing Xu et al., 2015) which is defined as follow:

$$\boxed{ReLu(s) = \begin{Bmatrix} x(x > 0 \\ 0(x \leq 0 \end{Bmatrix}}$$

is implemented in our study also. The loss function in DNN is defined as,

$$\boxed{loss = \frac{1}{2n}\sum_{i=1}^{n}\left(T_i - T_i'\right)^2}$$

where, n is the number of observations T_i and T_i' is the expected value by the DNN model. The activation function ReLU is:

$$\boxed{T' = ReLu\varpi_4(\varpi_3(ReLu(\varpi_2(ReLu)(\varpi_1 + b_1)) + b_3)) + b_4}$$

where ω_1, ω_2, ω_3, and ω_4 are the weights, and b_1, b_2, b_3, and b_4 are the bias terms.

Each connection between neurons in adjacent layers is associated with a weight value. This weight represents the strength or importance of the connection. The weights are applied to the connections between neurons during training to improve the network performance on a specific task. The bias term that is linked to each neuron also serves as an offset, and enables the network to learn non-linear correlations. The difference between the network's predicted output and the true value is measured using the loss function. Depending on the problem, a loss function may be selected. For instance, MSE for regression problems and categorical cross-entropy for classification problems.

The DNN model learns the more abstract and the high-level features, thereby performing more complex tasks. One of the challenges faced by deep neural network, as the gradient used to update the weights can become very small or very big, thus making it hard to optimize the network. Backpropagation is a key algorithm for training DNNs. It calculates the gradient of the loss function with respect to the weights and biases of the network. This gradient is then utilized to update the weights and biases in the opposite direction of the gradient to minimize the loss function. For the effective and efficient training of deep neural networks, the gradient is essential. It directs the optimization procedures, enabling the network to find its way to a combination of weights and biases that minimize the loss function and produce accurate predictions on new data.

Advantages of DNN

- Feature Learning: DNN can automatically learn complicated features and representations from raw data, thereby reducing the need for manual feature engineering.
- High Performance: DNNs have attained state-of-art performance in several domains, exceeding conventional machine learning techniques.
- Non-Linear Relationships: DNNs can model complex and non-linear relationships in data, making them suitable for tasks with complex patterns.
- Managing Different Data Types: DNNs can handle different kinds of data such as image, text, and audio, making them excel in various tasks like image recognition, natural language processing, etc.
- Scalability: By adding more layers and neurons, DNN can handle large datasets and complex problems, even though it depends on the computational resources.
- Parallel Processing: Faster computations can be achieved by using DNN that can make use of parallel processing on hardware with multiple cores and specialized accelerators.

Disadvantages of DNN

- DNN may struggle with insufficient data and frequently need huge datasets to achieve good performance.
- For highly deep architectures, training deep neural network models can be computationally demanding and time-consuming. In many cases, specialized hardware like GPU is needed.
- Heavy resources needed: Significant computing resources including powerful hardware and energy is needed for building and training DNN.
- Biased Data: DNNs are capable of picking up biases from the training data, which might lead to biased predictions.

2.4 Evaluation metrics

The most common metrics used to evaluate the models are mean absolute error (MAE), mean squared error (MSE), root mean squared error (RMSE), R-squared, and adjusted R-squared.

Mean Squared Error (MSE): This is one of the commonly used metrics used to evaluate regression models. The concept of MSE was introduced by (Gauss & Bertrand, 1957). It measures the average of the squared differences between the predicted and the true values. The lower the MSE, the better the model.

$$\mathrm{MSE} = 1/n * \sum (y-\hat{y})^2$$

Where y is the true value, ŷ is the predicted value, and n is the number of samples.

Root Mean Squared error (RMSE): This is the square root of MSE and is also commonly used metrics for evaluating regression models. The advantage is that it has the same units as the target variable, making it more interpretable.

$$\mathrm{RMSE} = \mathrm{sqrt}\ (\ 1/n * \sum (y-\hat{y})^2\ )$$

Mean Absolute Square (MAE): This is the average of the absolute differences between the predicted and true values. This was introduced by (Hyndman & Koehler, 2006). The main advantage of MAE is it is less sensitive to outliers than the MSE.

$$\mathrm{MAE} = 1/n * \sum |\ y-\hat{y}|$$

R-Squared (R^2): The concept of R-Squared was initially proposed by (Wright, 1921). This is a measure of how well the model fits the data compared to a baseline model that simply predicts the mean value of the target variable. R^2 ranges from 0 to 1, with higher value indicating better model performance.

$$R^2 = 1 - \sum (y-\hat{y})^2 / \sum (y-\bar{y})^2$$

Where ȳ is the mean value of the target variable.

3. Method

In this section, we describe the process of the data science life cycle to predict the yield of the lettuce grown in the agricultural farm. The method includes various processes to ultimately predict the crop yield. The various methods include data collection, data pre-processing, creating different models, evaluating the models, the results are analysed for all the models, and finally selecting the best models that predict the crop yield (Kumar et al., 2023).

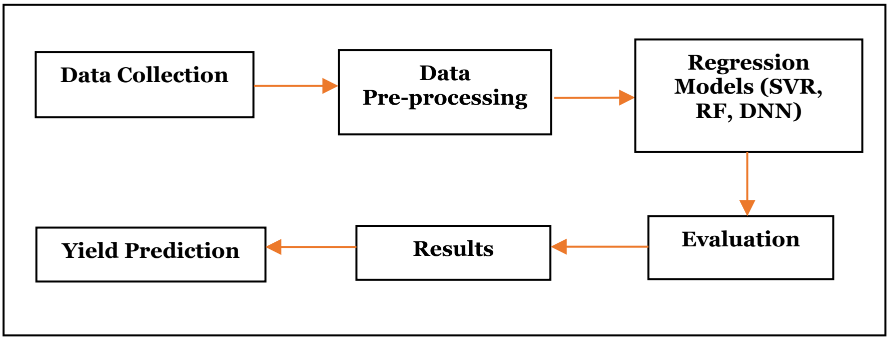


Figure 4: Lettuce Plant Yield Prediction Using Machine Learning

Here we have followed a hybrid model making use of a traditional data science life cycle project and CRISP-DM. The process starts with data collection, where the raw data is collected from indoor farms. We have done a rough analysis of the data to understand the data at a higher level. This gives a brief idea of what immediate cleaning is required, and how the data can be processed. Next, we move on to the data pre-processing step in which we work on the data for cleaning, transformation, and aggregation to prepare the data for the machine learning models. Based on the literature review performed, we have arrived at three machine learning models for the prediction of the yield of lettuce plants. We build the models, and experiment with various combinations of parameters for the models. Once the models are built, and fit, we proceeded to evaluate them. This is to understand how the models are performing on different data, and how good are their performance. These results are noted and analysed to go back to the model building stage. Again, we experiment with different model parameters, and changes of features in input data and check the results. Each model with the best results is noted and a yield prediction is made with these models. Finally, each model is compared to others, to decide the best model for yield prediction.

For this case, we have followed a systematic literature review and interviews with the domain experts. At first, we conducted an interview with the plant expert to understand the characteristics of the lettuce plant, how the features affect the growth, what is the growth process and regarding each feature. We also interviewed the data analyst at the company to understand how we can handle the plant data in a meaningful way so that the results are in real-time valuable. To understand the business process, the environmental and ethical aspects, and the impacts of the research work was explained by the CEO of the company. Based on this understanding, we conducted a systematic literature review of previous academic papers, and other scientific articles and also giving importance to ideas published from Israel, since they are a pioneer in the farming revolution. Choosing these methods over other methods is based on the requirement of the research work. Since the data is already collected and given, for us to understand the needs, approaches, and data – the only way we came across is to conduct an interview with the experts and understand from them. Studying the previous work and research conducted on this topic will be greatly helpful and guiding for our research and hence a systematic literature review is conducted.

3.1 Dataset

As part of the requirement, there are four different datasets provided to implement the prediction of yield for lettuce plants. Below is the dataset explained.

The climate dataset (10312256 rows × 4 columns) includes columns such as GrowRoom, TimeStamp, Attribute, and Values. The attribute column consists of temperature, humidity, relative humidity, and CO2 values which will be transformed into individual columns during data cleaning. The climate data records the climate conditions with the help of sensors.

The second dataset provided was the manual data (36103 rows × 10 columns) which collects features such as GrowRoom, TimeStamp, DAT (Days After Transplant), BatchID, Seed_type, Attribute, Values, Wagon, and Plant_position. The attributes include air_temperature, canopy_area, Weight_packable, Weight_waste, Weight, Leaf_temperature, Humidity, Roots_EC, Roots_temperature, leaf_vpd, Plant_height and Leaf_count.

The third data provided was the water data (75 rows × 26 columns) which records all the nutrient features provided to the lettuce plant for growth. The water data contains features such as DateReceived, GrowZone, GrowRoom, pH, EC (Electrical conductivity), Alk (Alkaline level), K, Cl, Zn, Mg, Mn, Mo, NO3-N, S, B, Al, NH4-N, Ca, Cu, Si, Na, Fe, Ni, and P. Most of these are chemical nutrients provided to the lettuce plant to aid the growth.

The last data provided was the light data (49 rows × 16 columns) which contains the characteristics of light recipe provided to the lettuce plant. Since only 1 recipe was used throughout the year, we were instructed by the data analyst in the company to eliminate this light data which can lead to wrong predictions.

3.1.1 Data preparation

The climate data was provided in 15 files having data collected for the features of 15 grow rooms each from 01 to 15. Using the concat function and the same column names, the data is combined into a single dataset. To convert the attributes in the Attribute column to individual columns, the pivot_table function was used with Date and GrowRoom as the index. Also, the temperature attribute was converted into mean, minimum, and maximum values for each day, so that multiple readings of each day can be consolidated into a single day's reading to make ready for the final dataset. Finally, the index was reset using reset_index and the temperature column is dropped.

The manual dataset has 36,103 records which is the actual reading that can help for building the model. Out of this, ID, Wagon and Plant_position was initially dropped as per the discussion with plant expert since there are more null values and they cannot be interpolated as they are specific and categorical variables. To convert the attributes into individual columns a pivot_table was used with index as GrowRoom, Date, DAT, BatchID, and Seed_type, columns as Attribute and value as Values. The index was reset using reset_index function. Out of the individual attributes Air_temperature and humidity were immediately dropped since they are duplicate data from the climate data. Leaf count, Leaf temperature, Plant height, Weight, Leaf_temperature, leaf_vpd (Leaf Vapour Pressure Deficit), Roots_EC, Roots_temperature, and canopy area was also dropped since out of 799 records majority of them were null values due to the manual error while data collection. Plant expert advised that it is better to drop these attributes since the reading has plant from DAT 1 to DAT 22 and the data interpolation technique used can corrupt the prediction of the yield. Features Weight_Waste, Weight_Packable, humidity and CO2 have few null values which was populated using the interpolate function.

Out of water data, the features with the most contribution for the growth of the lettuce plant was selected which are GrowRoom, GrowZone, Date, pH, EC (Electrical Conductivity), K, NO3-N, NH4-N, Ca, Fe, and P. We could select these features by finding the correlation with the target variable and after having inputs from the plant specialist in the company. All the other features were dropped. The data records each are for one week's water supply provided to the lettuce plants, and we utilised this data to be prepared in a daily manner to prepare for the final data. Initially to populate for all the days, using a for loop each row was taken and appended into a new and blank dataset. The Date was accessed using the iloc method and iterated by using timedelta (days=1). Applying this for the whole dataset, the data for all the dates individually was populated. Next the GrowRoom should be populated based on the GrowZone since there are only three GrowZones and each Growzone has 5 GrowRooms and only one row data for 5 GrowRooms. Again, a for-loop was used to iterate through each row in the dataset and three inner for-loops were used to split each GrowZone into 5 GrowRooms based on the conditions. The rows were copied using copy() and by using append() inserted into a new and empty dataset. Since the index values are copied from the same, a manual index is set using set_index() with the size of the dataset passed as series using range(). The light data consists of only one light recipe which is now being tested for the plants and it was advised by experts that we ignore the whole data.

To join the three datasets is difficult now as the data in climate data is in a very frequent timestamp (several thousands of readings per day) and manual and water data in a daily timeframe. To overcome this issue, we have converted the climate data entries into a daily time frame. The entries are grouped by the date and the mathematical mean of the features (temperature, humidity, relative humidity, and CO2) are calculated using the pivot function. Finally, when all the datasets are in a common daily timeframe, they are merged into a single dataset by using the merge() function with Date and GrowRoom as a common column to combine manual data, climate data, and water data.

To handle the outliers in the 'Weight_Waste' feature having wastage of weight within 6 days, this is replaced by 0. For the feature 'Weight_Packable', which is the target variable it is noted that there are outliers based on the 'DAT' values. For example, a lettuce plant with just 5 days of growth will not be having 100 grams of weight. We divided the days of growth into group of 5 days to calculate the average 'Weight_Packable' during this period and this value was assigned to corresponding 'DAT' feature for which the 'Weight_Packable' is out of the limits. Finally, there was one single 'DAT' value which is 23 and according to our interview with the Company it was said that the lettuce plants will be harvested within 22 days maximum. Based on this information, we have deleted that row considering it an outlier.

Features	Description
GrowRoom	Grow Rooms from 1-15
DAT	Date of Transplant after 21 days from the growing tray to the vertical tower
Weight_Packable	Yield – fresh head weight (target variable)
Weight_Waste	The weight of the lettuce plant is not consumable
Humidity	Concentration of water vapors in the air (grams per cubic meter of air)
Relative Humidity	Measure of water vapor in water air mixture compared to max amount possible (%)
CO2	Carbon dioxide is measure in parts per millions(ppm)
Max Temperature	Maximum recorded temperature for the specific date
Min Temperature	Minimum recorded temperature for the specific date

Mean Temperature	Average recorded temperature for the specific date
pH	Acidity level of water (5.5 to 8 is the optimal pH level)
EC	Electrical Conductivity (nutrient level in water, normal range is from 2 to 3.5
K	Potassium level (mg/L)
NO3-N	Nitrate Content
NH4-N	Nitrate Content
Ca	Calcium level
Fe	Iron Level
P	Phosphorus Level
SeedType_Exanimo	Type of seed planted
SeedType_Zac	Type of seed planted
Day	Day of the reading taken
Month	Month of the reading taken
Year	Year of the reading taken

The cleaned final dataset consists of 708 rows and 23 columns.

Seedtype feature has two categories (Crystal lettuce and Incised lettuce) with only six records each which were later dropped since those types of seeds are not in consideration. With other seed types Zac and Exanimo, One hot encoding was performed using the get_dummies() function. The GrowRoom containing 15 categories was label encoded using LabelEncoder() instead of One hot encoding in order to keep the number of features minimal. The Date feature was split into Day, Month, and Year features by using dt.date, dt.month and dt.year respectively so that the model is able to input them.

We split the dataset into train and test with a ratio of 70:30 respectively. The test set should be smaller than the training set. 70:30 is a common split ratio used by the data professionals when working on small to medium-sized datasets. The choice of splitting the data such as 70:30 is not a strict rule but rather a common guideline that can be adjusted based on the size of the dataset, the nature of the problem, and the requirements. As our final dataset was small, we proceeded to take a 70:30 split as an initial point for small to medium-sized datasets. This 30% is a significant amount of test data to evaluate the performance of the models. According to statistical analysis performed by (Nguyen et al., 2021), the training and testing data is divided as a ratio of 70-30 and they are considered the best ratio for training and validating the models. (Fashoto et al., 2021) experimented with different ranges of splitting training and testing data while predicting the yield of maize crops, and he concluded that an 80-20 ratio is often considered but he does not support that an 80-20 ratio should be used in all the scenarios.

4. Implementation

In this section, the steps followed to implement the solutions, what all functions are used, what are the reason behind them, the thought process behind each step is explained.

4.1 Implementation of support vector regressor (SVR)

The Support Vector Regressor model is selected based on the previous research that it can handle non-linear relationships between target variable and the features. It is also less sensitive to outliers in the data. The SVR model is imported from svm in sklearn package as it is a regression model based on Support Vector Machine. The kernel is provided as 'linear' which provided the best score. The data is split into the target variable (Weight_packable) and predictor features. The StandardScaler is used to transform the data from the preprocessing package. Once the model is trained using fit(), prediction is done with the help of model.predict() and the predicted values (y_pred) are compared with the true values (y_test) to check the evaluation metrics of mean squared error, root mean squared error, cross validated mean absolute error, r-squared, and cross validated adjusted r-squared.

4.2 Implementation of deep neural networks (DNN)

The Deep Neural Networks is selected based on its ability to handle non-linear relationships. It can also automatically learn the hierarchical importance of features in different levels of data. The DNN is implemented with the help of keras and tensorflow packages in Python. All required modules such as sequential, dense, layers, and activation which build the DNN model are imported.

A sequential model is built by defining the required layers which are the input, hidden, and output layer using Dense from layers in keras. The number of neurons in the input layer is decided by the number of input features. We have decided on two hidden layers with 11 and 6 units of neurons in each layer. The activation function used is relu which gives a stable evaluation score for each model execution. Input shape is provided as the number of features in the input dataset. The DNN model is compiled with the model.compile() function setting the loss as 'mean_squared_error', and the optimizer as Adam with a learning rate of 0.03. Then the model is trained using fit() provided 50 epochs, a batch_size of 100, and verbose = 1. Then the model is evaluated with the evaluate() function and prediction is done. The parameters for hyper-tuning are decided by several iterations of manual testing with various values and fixed the value that provides the best results. Finally, the predicted results and the true value are compared to populate the of mean squared error, root mean squared error, mean absolute error, and adjusted r-squared. A KFold method also is used to generalize the model and its evaluation metrics with n_splits as 10. The n_splits value is tested with various values and is always careful in order to avoid overfitting issues.

4.3 Implementation of random forest regressor

Random Forest Regressor is selected based on previous research papers that it is highly robust to outliers and noisy data. It can handle missing data automatically while constructing trees. The chances of overfitting while using Random Forest are very unlikely. Randomforestregressor() is imported from the sklearn package to implement the model. The model is defined with an n_estimator value of 44 which sets the number of decision trees to include in the random forest ensemble. In our case, n_estimator is set to 44 means the model will consist of 44 individual decision trees. The maximum depth of the individual decision tree is set to 11. These values are decided with different manual iterations and fixed when the model is giving the best results. Initially, the model is created with n_estimator as 10 and the depth of the tree as the number of features. In the next test iterations, these figures are changed until they are fixed at 44 and 11 which is convincing enough. The model is trained using fit(), a prediction is made using model.predict() and the evaluation metrics cross-validated MAE, MSE, RMSE, R-squared, and cross-validated adjusted r-squared are calculated. Among the three models, it is to be noted that the Random Forest Regressor gives the best results.

Why not LSTM?

Initially, it was fixed that an LSTM model would also be built with the final data. However, after careful analysis, we have concluded that even though the data includes a timestamp, this cannot be considered a time series as a whole. This is due to the reason that the lettuce plant has a life cycle of 21 days after transplanting to the vertical tower and this keeps repeating. Also, it is to be noted that the four datasets provided are in different timeframes each and recorded in irregular intervals.

5. Results

The models explained in Section 4 are implemented, and the evaluation metrics are calculated. Results are analysed depending on the model selected. Below are the scores for the evaluation metrics observed in the models.

Scenario – 1 (all variables as input)

Model	MAE	RMSE	MSE	R-squared	Adjusted R-squared
DNN	9.67	2.24	14.25	0.83	0.80
SVR	15.63	12.47	158.62	0.65	0.61
RF	7.43	12.05	148.15	0.88	0.86

Scenario – 2 (only considering GrowRoom, DAT, Weight_Waste, Relative humidity, Mean Temperature, NH4-N, NO3-N)

Model	MAE	RMSE	MSE	R-squared	Adjusted R-squared
DNN	12.44	3.11	17.45	0.73	0.70
SVR	15.62	14.26	205.98	0.65	0.61
RF	8.77	14.08	201.04	0.83	0.83

For analysing the DNN model, we can observe the loss and the accuracy score of the model for the train and test data. This will help to understand if the model may have any issues like overfitting or underfitting.

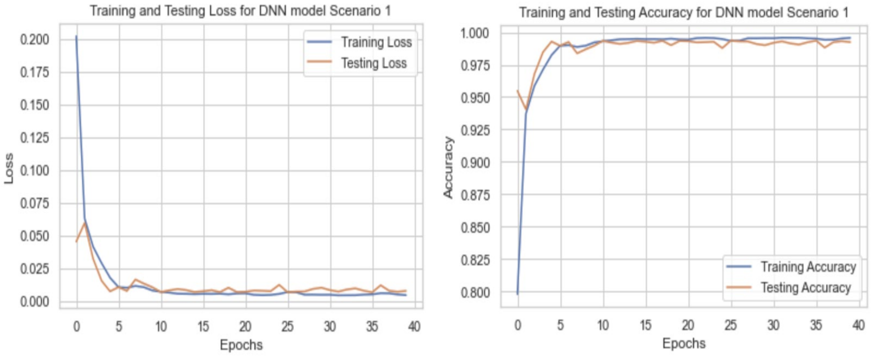


Figure 5. Loss and accuracy plots for DNN model's training and testing data (Scenario 1)

In the Figure 5 results for DNN, the blue line represents the training data and the red line represents the testing data. It is noted that the loss and accuracy for the both testing and training data are maintained mostly in a same trend. Even though there is a slight fluctuation of test data for loss and accuracy, we believe this to be because of the lack of sufficient data points.

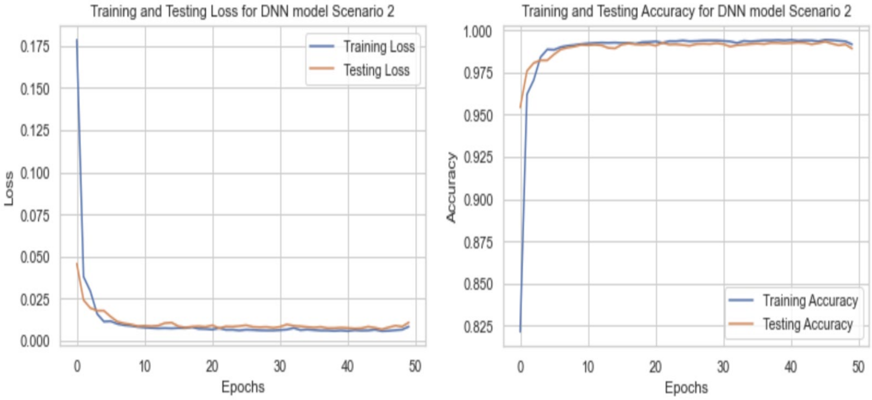


Figure 6. Loss and accuracy plots for DNN model's training and testing data (Scenario 2)

In the above Figure 6, we can note that the fluctuations in the testing as well as the training data are much subtle and not immediate. This can be because the input was only selected features which avoided confusion to the model.

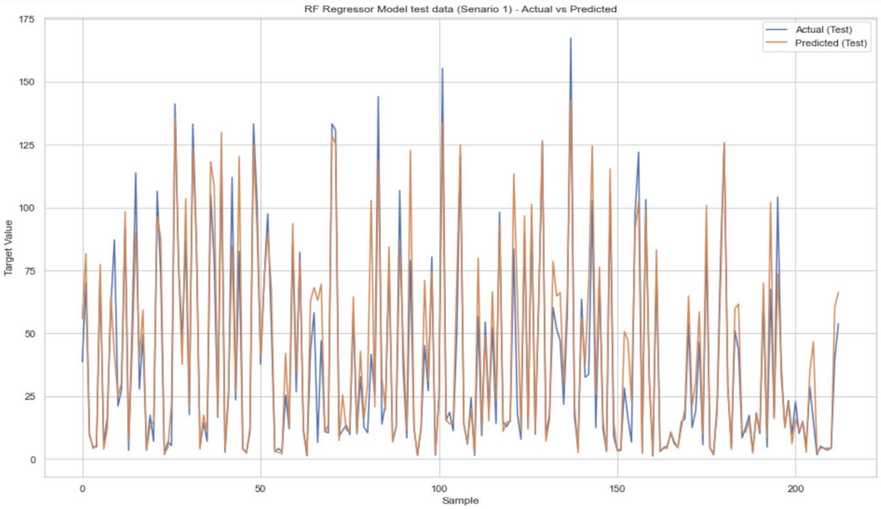


Figure 7. Actual vs predicted value for Random Forest model with test data – Scenario 1.

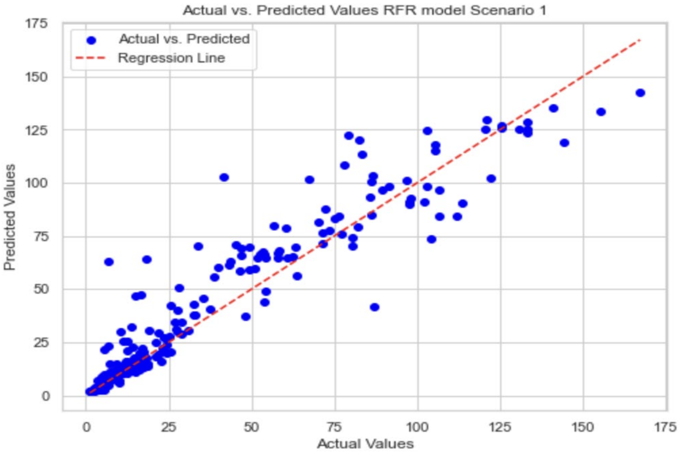


Figure 8. Actual vs predicted value for Random Forest model against the regression line – Scenario 1.

In Figure 7, we can see the actual vs predicted target values for the random forest model with all the features. It has returned 0.86 R-squared value. Also, in the above test data plot, we can understand that there are very few variations from actual to predicted. This clearly means that Random Forest Regressor has predicted most of the time correctly. It is to be noticed the few errors between actual vs predicted where the top peaks of few areas are coloured blue because the predicted value is less than the actual value. Based in figure 8, we can understand that the actual vs predicted values are very close to the regression line which means that the model has good prediction capacity.

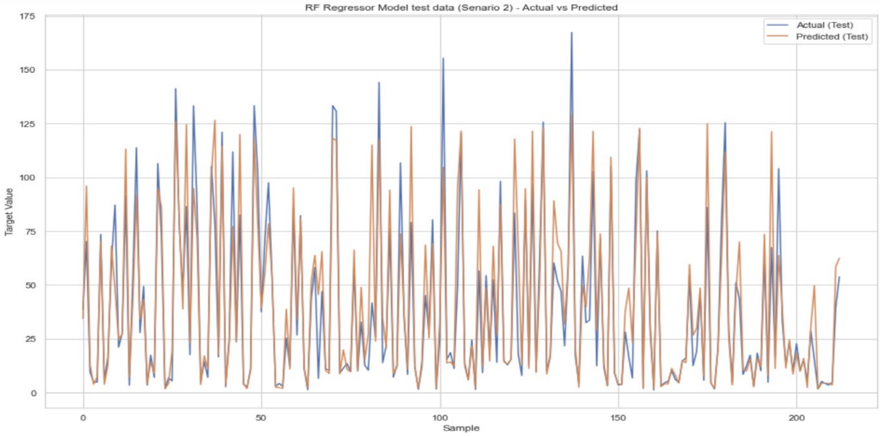


Figure 9. Actual vs predicted value for Random Forest model with test data – Scenario 2.

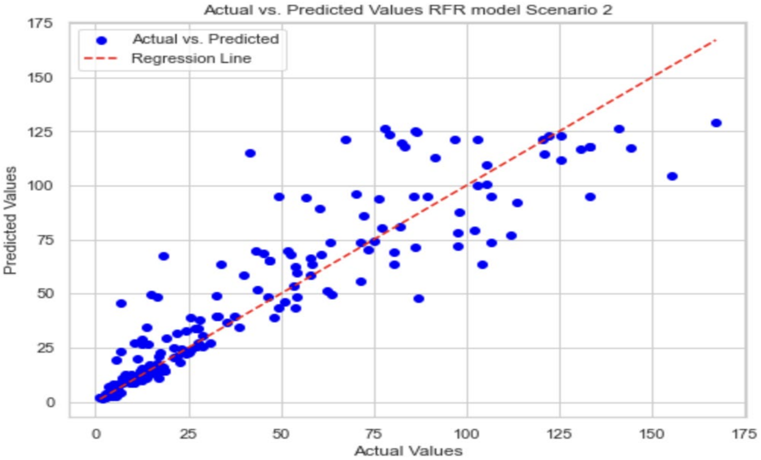


Figure 10. Actual vs predicted value for Random Forest model against the regression line – Scenario 2.

Figure 9 shows the actual vs predicted for scenario 2. Even though the random forest model is provided with selected features, the performance is poor compared to the previous random forest iteration. In figure 10 we can see that when compared to figure 8, the points are really far off from the regression line.

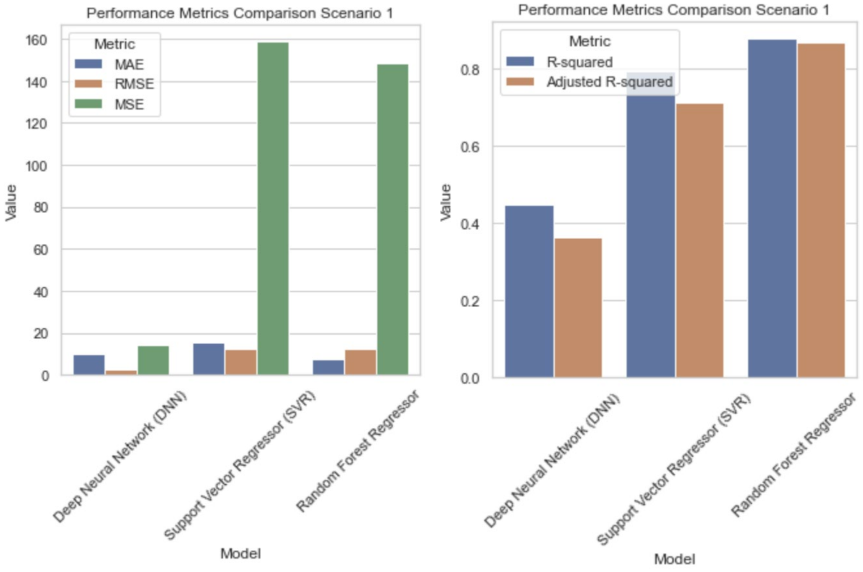
Support Vector Regressor models provided poor results compared to RF and DNN models. But at the same time, other research groups working on the same research area, noted that with the inclusion of light data SVR showed improved performance.

5.1 Methods, implementation, and results

Different scenarios were used by selecting the important features that affect the growth of lettuce plants, and these scenarios were used as the input for the implementation of our models. The first scenario (Scenario 1) is that we consider all the input variables for the models. In the second scenario (Scenario 2), we consider only variables that are most correlated with the target value and the most important features which can impact the target variable as suggested by the experts. We have built different machine learning models such as SVR, RF and DNN to predict the yield of lettuce. The climate data, water nutrient data, and plant growth data are passed as input to train the models to understand the growth pattern based on the features. Our study also included various evaluation metrics majorly covers Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), R-squared, and Adjusted R-squared value. We had considered the MAE and Adjusted R-squared metrics because MAE is less sensitive to the outliers and adjusted R-squared value helps to understand the variance of the target variable with the predictor variable and adjust the metric to prevent the issues of overfitting. The results of our study shows that Random Forest with all the features is the best model having the best results with the least cross-validated MAE score and good cross-validated Adjusted R-squared value considering that the error of the prediction is minimal. This is followed by the DNN model with minor differences in the resulting values. The Support Vector Regressor (SVR) model gave a very poor performance with a huge error value that cannot be afforded in this scenario.

5.2 Analysis of the joint results

The above results are analysed as follows. Since we have the target variable for the prediction as the yield of the lettuce plant, i.e., “Weight_packable”, the models cannot afford to have any large errors. Hence, we are considering the RMSE score as it measures the model’s accuracy by calculating the square root of the average of the square of errors. In this case, the possibility of high errors is less since the errors are squared. MSE, the mean squared error provided us with the square of the average errors as visible in Fig 11 and 12. A lower MSE score indicates better model performance. Another better option along with the RMSE score here is the MAE score. This is the mean absolute error, which calculates the mean of the absolute difference between predicted and true values. The MAE score is able to handle the unnoticed outliers present in the final dataset. R2 score helps to understand the variance of the target variable depending on the predictor variables and it has a value from 0 to 1. The RF model has an R2 score of 0.86, followed by DNN with a 0.80 with the scenario 1. Using both scenarios, the RF model have an R2 score above 0.80 and is considered to be the best model in our case.


Performance Metrics Comparison Scenario 1

Performance Metrics Comparison Scenario 1

Figure 11: Scenario 1 model scores

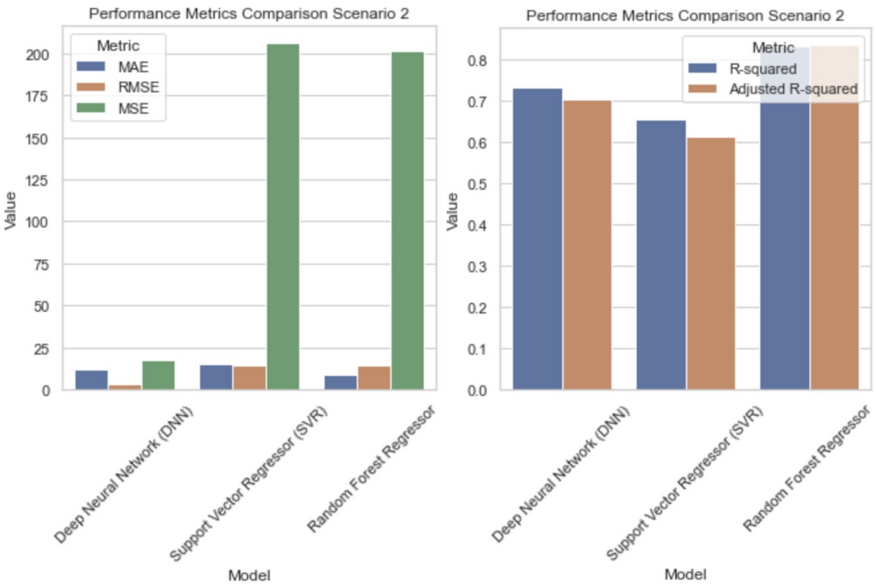
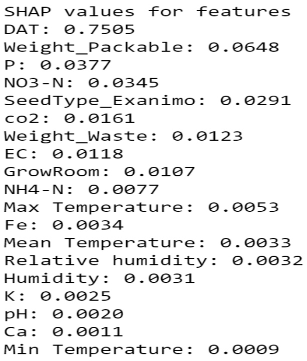


Figure 12: Scenario 2 model scores

5.3 SHAP analysis

The SHAP model can be applied only on an existing model. So, it is normal that the target variable is also visible in the features list. Here, in this experiment, the feature selection method is used to understand the impact of features on the model's output based on the Random Forest regressor model and the SHAP score for each feature is printed. The result showed that DAT has the highest impact on yield, followed by P, NO3-N, SeedType_Exanimo, CO2, and Weight_Waste. Also, it can be noted that all top features has positive impact on the prediction of the model.



```
SHAP values for features
DAT: 0.7505
Weight_Packable: 0.0648
P: 0.0377
NO3-N: 0.0345
SeedType_Exanimo: 0.0291
co2: 0.0161
Weight_Waste: 0.0123
EC: 0.0118
GrowRoom: 0.0107
NH4-N: 0.0077
Max Temperature: 0.0053
Fe: 0.0034
Mean Temperature: 0.0033
Relative humidity: 0.0032
Humidity: 0.0031
K: 0.0025
pH: 0.0020
Ca: 0.0011
Min Temperature: 0.0009
```

Figure 13: SHAP scores for features from Random Forest (Scenario 1)

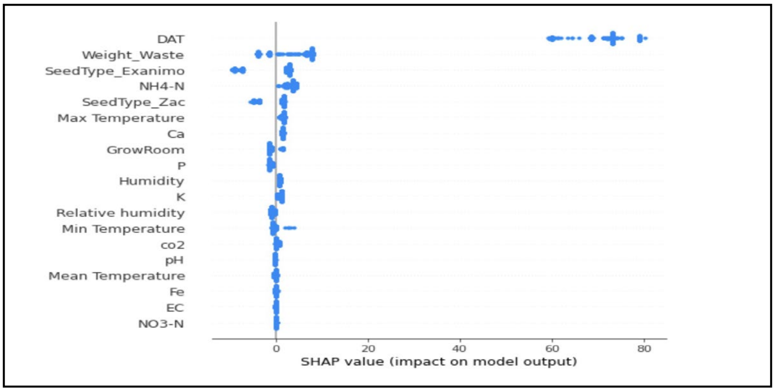


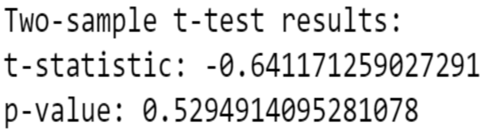
Figure 14: SHAP visualization summary plot

In figure 14, it is clearly visible that the feature DAT has the most impact on the prediction in RF model, followed by weight_waste. In other words, most of the growth of the lettuce plants can be explained by how old the plant gets up to 21 days. After that the lettuce stops growth and with further time it starts decaying.

5.4 Statistical testing of evaluation metrics

Based on the study on previous research and a discussion with the supervisor, we decided to conduct a statistical comparison on the results (Cross validated Adjusted R-squared score) of DNN and RF model with scenario 1. We have decided to proceed with a two-sample t-test. This helped us to understand if there is any significant difference between the cross validated adjusted R-squared values of two models.

The test provides two values which are the p-value and the t-statistic value. The t-statistic values simply provide the difference in the means of two models. But the p-value provides the confirmation on the basis of a null hypothesis.



```
Two-sample t-test results:
t-statistic: -0.641171259027291
p-value: 0.5294914095281078
```

In our test we found the p-value to be greater than 0.05 which is a significant value. So based on this, we accept the null hypothesis that there is no significant difference between the cross validated adjusted R-squared values of both the models. Also, with the confirmation that having lesser MAE value of 7.43 and better cross validated R-squared score of 0.86, RF model is the best performing model. Also, we are aware that the RF model can handle noisy and outlier data.

It is also to be noted that even though we had chosen MAE, there is a significant difference between RMSE and MSE value of the two models which contradicts our conclusion. This is justified with the facts that the input data to the models are transformed using MinMaxScaler which had reduced the magnitude of outliers from the actual. Hence, we have given importance to the MAE value over RMSE and MSE considering the ability of MAE value to be less influenced by outliers.

5.5 Results for research questions

RQ1. We have two results based on two approaches for this research question. One is by the instructions provided by the plant experts who confirmed that DAT, Relative_Humidity, Temperature, NO3-N, and NH4-N features are of high importance. The second insight is generated from the model results and the SHAP method implemented – this showed us that DAT is the most impacting feature for the prediction of RF model. Followed by a slight impact by Weight_waste, NH4-N, and Min_temperature. All other features according to the SHAP method have a Shapley score of less than 0.01.

RQ2. The best performing model in this case with the data we have is RF model with an MAE value of 7.43 and cross validated R-squared score of 0.86.

RQ3. We have studied five metrics as part of our regression problem – MAE, MSE, RMSE, R-squared, and Adjusted R-squared. With the analysis we have performed, the conclusion arrived is that the best metrics that can be concentrated are MAE (Mean Absolute Error) and Adjusted R-squared score. MAE is chosen over MSE and RMSE because in MSE value, in order to overcome the negative value issue, the square of the error is calculated which increases the magnitude of the error. Even though RMSE is the root of the MSE value, the effect of the magnitude gets carried over to the final value which cannot be reliable. The R-squared score is analysed to understand the fit of the model, but it is identified that the better one is the Adjusted R-squared score which also takes into consideration the number of predictors and in turn gives a better assessment of the model.

6. Discussion

In this section, we will be discussing the work that we have done as a part of the thesis work, what motivated us to select certain methods. We will also be analysing the results which we have generated to find a solution to the research questions. Different models such as support vector regressor (SVR), random forest regressor (RF) and deep neural network are constructed to predict the yield of the lettuce plant. In this study, we have also compared the different evaluation metrics and took into account the cross-validated MAE and cross-validated Adjusted R-squared metrics because MAE is less sensitive to outliers. The adjusted R-squared metric helps to understand the variance of the target variable with the predictor variable and adjust the metric to prevent overfitting issues.

SHAP feature selection is used for a better understanding of the models. This helps us to understand the importance of each feature, and to explain the result of the model to a human. We have also tried to implement statistical tests to study how there is a significant difference between the scores of different models used in our study.

According to the data, we found out there are many important features that affect the growth of lettuce plants. According to past experiments conducted to increase the yield of the lettuce plants, light intensities are a very important feature to increase the yield of the crop. In our study, they lack light data as the company provided only one light recipe throughout the year in every season. We were advised not to incorporate the light data as it may lead to the wrong prediction. In the coming years, the company will be collecting the data with sensors as some of the data provided was taken manually. Finally, with the available data provided to us, and within our limited capacity, we have concluded that the RF model with all features provided as input, works the best in this scenario.

As part of this research work, we have a few suggestions that can help the company to improve the process in the future:

1. Assuming light data is important for lettuce growth, it should be collected and included for the prediction in the future. Currently, there is no light intensity data available for enough samples.
2. Most influenced manual data should be collected in an automated and regular way since it has a direct impact on recording yield and several other physical aspects of the lettuce plant.
3. It is advised that if the data collection is integrated into one single database for each plant/batch would help to improve the analysis process.

6.1 Theoretical framework (or previous research)

Many studies have been performed previously to predict the yield of different kinds of plants such as lettuce, tomato, soybean, etc. Predicting the crop yield helps the farmers to optimize the production of the crop. More effective resource allocation is made possible by accurate yield forecasts for farmers and agricultural businesses. Based on the output, they can optimize the use of water, fertilizers, pesticides, and other inputs, cutting down the plant waste, and increasing cost-effectiveness. This also helps to plan their harvest and post-harvest operations more successfully. (Mokhtar et al., 2022) conducted a study to predict the lettuce yield grown hydroponically. They monitored the lettuce yield prediction using different machine learning (ML) models such as support vector regressor (SVR), random forest (RF), extreme gradient boosting (XGB), and deep neural network (DNN). The data was collected between 2018 and 2019. In their experiment, they created three scenarios with various combinations of input factors, including dry waste, stem measurement, water intake, and leaf count. The XGB model with scenario 3 having all the input variables has the lowest RMSE value, followed by SVR with the same scenario. The researchers came to the conclusion that the two best models were SVR, which had all input variables with scenario 3, and DNN, which had leaf counts, water intake, and dry weight as inputs with scenario 2. They claim that among the two best models, DNN with fewer inputs is more favoured as the ability of the DNN model to predict the yield of fresh lettuce is promising, and it is estimated that the DNN may also be a rapid tool for decision-makers to control crop yield on a wide scale. Our study also included

SVR, RF, and DNN models with two scenarios. Scenario 1 takes all the input variables and scenario 2 with only a few selected features that are more correlated to the target variable and the essential features recommended by the plant expert in the company. Ultimately our final results show that RF is the best model with scenario 1 with less MAE values of 7.43 and an R-squared value of 0.88.

Many previous research studies have given importance to light-intensity data as they are essential for the proper and healthy growth of the plant. Here, we lack light data as we don't have enough light intensity data collected in the previous years by the company. They are planning to collect all the plant and environment data in regular intervals with IoT in the coming years to help in better prediction of the lettuce yield in the future. With the help of light data, the predictions showed good results in previous research using the SVR model (Gertphol et al., 2018). But since we are lacking light data and with the limitations of our data the RF model is providing the best results in our case.

6.2 Ethical and societal aspects

The ethical and societal aspects of predicting the yield of lettuce plants can be understood from different perspectives, including the impact on society, the environment, and the economy.

The social impact of predicting the yield of lettuce plants can help to increase food security by ensuring that enough food is generated to feed the growing population. This helps to reduce the risk of food shortages and price spikes which can have a significant impact on social communities. Predicting lettuce plant yields might encourage the consolidation of the agricultural sector because larger farms might benefit from having an advantage in deploying new technology. Small-scale farmers and rural communities may suffer as a result of this.

The environmental impact of predicting the yield can help the producers to increase agricultural productivity, reduce waste, and minimize the use of harmful pesticides and fertilizers. This may result in an agriculture sector that is more environmentally responsible and sustainable. In turn, the more usage of agrochemicals could lead to environmental damage if not used responsibly.

The ability to predict lettuce plant yields can assist farmers in planning their planting and harvesting operations and in streamlining their manufacturing procedures. This could lead to higher crop productivity and gaining more profits, which can contribute to the economic development of rural communities. On the other hand, predictive models may benefit large-scale commercial farmers who have the resources to invest in technology and data analysis. However, this could lead to widening the gap between large and small-scale farming.

Overall, there may be ethical considerations about the use of predictive models in agriculture. It is really important to carefully consider and address any ethical concerns that might arise. The predicting process requires data that are collected and analysed data from multiple sources. This includes weather patterns, quality of soil, and plant growth data. Farmers may be concerned about the privacy and security of the data their data. The other issue is that the predictive models may be biased if the data is incomplete or inaccurate. This could lead to unfair treatment of farmers or contribute to environmental harm. Predicting the yield enables farmers or agricultural firms to allocate resources more efficiently. So, they can optimize the use of water, fertilizers, pesticides, and other inputs based on the prediction, thus helping reduce waste and cost-effectiveness. Businesses are able to plan their harvest and post-harvest operations more effectively with accurate yield predictions. This also includes scheduling labour, transportation, and storage resources to match the expected yield. Also, helps to coordinate their supply chains better and can ensure that the correct number of lettuces reaches the consumers controlling shortages or excess supply. It is really important that these models are developed and used in a responsible and transparent manner.

7. Conclusion

Predicting the yield of lettuce plants is an essential task for farmers to optimize their crop production and increase their overall yield. There are different parameters affecting the growth of lettuce plants, such as climatic conditions, water nutrients, and other environmental factors. In our project, we have used different scenarios with different input variables. These data are fed into the models to understand the features affecting the growth of lettuce. To evaluate the performance of different models, various metrics were used, such as MSE, RMSE, cross-validated MAE, and cross-validated adjusted R-squared. These metrics have their own strength and weaknesses. Based on the research scenario and the inputs, we have evaluated the results using the metrics available. RF model is identified as the best model with a cross-validated MAE score of 7.43 and a cross-validated adjusted R-squared score of 0.86 with scenario 1 that included all the features.

7.1 Future work

A future study could test the models on different varieties of lettuces. The current models are trained only on two specific varieties of lettuce. So, the study can be conducted to explore the generalizability of these models of other varieties of lettuce, which might have different optimal growing conditions and yield potential. These models take a limited set of input features to predict the yield of lettuce plants. The data was collected in two ways, climate data using sensors, and other data was collected manually. Future work can explore using more advanced sensing technologies to gather data on additional features such as light intensities, plant growth rates, disease prevalence, and insect infestations. This may increase the accuracy of the models and provide more comprehensive insights into the factors that affect lettuce yield.

Our project is predicting the yield of the lettuce plant, which will help to increase their business by increasing productivity, thereby increasing the profit for the company. In the future, the development of mobile applications will help the company or farmers to input their data and receive real-time predictions of lettuce yield. This helps the farmers make informed decisions about their lettuce production and optimize the yield potential.

With proper automated and organized data collection at regular intervals, the company will have records of its plant data for future studies. This will also help to implement the LSTM model in the future, which will provide better results for plant yield prediction. Also, it is observed that light is an important factor in promoting the growth of plants. In the future, the combination of water and light data can be used to develop an improved prediction model that can provide accurate predictions for business process planning.

References

Taghizadeh, Rouzbeh. (2021). Assessing the Potential of Hydroponic Farming to Reduce Food Imports: The Case of Lettuce Production in Sweden.

Alexandratos, N., & Bruinsma, J. (2012). World agriculture towards 2030/2050: the 2012 revision.

Rhodes, C. J. (2019). Only 12 years left to readjust for the 1.5-degree climate change option–Says International Panel on Climate Change report: Current commentary. Science progress, 102(1), 73-87.

Shan, Y. (2021). Opportunities and Challenges for Developing High-tech Urban Agriculture in Sweden: A case study in Stockholm.

Srivani, P., & Manjula, S. H. (2019, December). A controlled environment agriculture with hydroponics: variants, parameters, methodologies and challenges for smart farming. In 2019 Fifteenth In-ternational Conference on Information Processing (ICINPRO) (pp. 1-8). IEEE.

Ullah, R., Asghar, I., Griffiths, M. G., Stacey, C., Stiles, W., & Whitelaw, C. (2023, March). Internet of Things based Sensor System for Vertical Farming and Controlled Environment Agriculture. In 2023 6th Conference on Cloud and Internet of Things (CIoT) (pp. 136-140). IEEE.

Farooq, M. S., Riaz, S., Abid, A., Abid, K., & Naeem, M. A. (2019). A Survey on the Role of IoT in Agriculture for the Implementation of Smart Farming. Ieee Access, 7, 156237-156271.

Majid, M., Khan, J. N., Shah, Q. M. A., Masoodi, K. Z., Afroza, B., & Parvaze, S. (2021). Evaluation of hydroponic systems for the cultivation of Lettuce (Lactuca sativa L., var. Longifolia) and comparison with protected soil-based cultivation. Agricultural Water Management, 245, 106572.

Kloas, W., Groß, R., Baganz, D., Graupner, J., Monsees, H., Schmidt, U., Staaks, G., Suhl, J., Tschirner, M., Wittstock, B. and Wuertz, S. (2015). A new concept for aquaponic systems to improve sustainability, increase productivity, and reduce environmental impacts. Aquaculture environment interactions, 7(2), pp.179-192.

Hong, J., Xu, F., Chen, G., Huang, X., Wang, S., Du, L., & Ding, G. (2022). Evaluation of the Effects of Nitrogen, Phosphorus, and Potassium Applications on the Growth, Yield, and Quality of Lettuce (Lactuca sativa L.). Agronomy, 12(10), 2477.

Mokhtar, A., El-Ssawy, W., He, H., Al-Anasari, N., Sammen, S. S., Gyasi-Agyei, Y., & Abuarab, M. (2022). Using machine learning models to predict hydroponically grown lettuce yield. Frontiers in Plant Science, 13, 197.

Samadi, M., Sarkardeh, H., & Jabbari, E. (2021). Prediction of the dynamic pressure distribution in hydraulic structures using soft computing methods. Soft Computing, 25, 3873-3888.

Gertphol, S., Chulaka, P., & Changmai, T. (2018, November). Predictive models for lettuce quality from internet of things-based hydroponic farm. In 2018 22nd International Computer Science and Engineering Conference (ICSEC) (pp. 1-5). IEEE.

Jeong, J.H., Resop, J.P., Mueller, N.D., Fleisher, D.H., Yun, K., Butler, E.E., Timlin, D.J., Shim, K.M., Gerber, J.S., Reddy, V.R. and Kim, S.H. (2016). Random forests for global and regional crop yield predictions. PloS one, 11(6), p.e0156571.

Kim, N., Ha, K. J., Park, N. W., Cho, J., Hong, S., & Lee, Y. W. (2019). A comparison between major artificial intelligence models for crop yield prediction: Case study of the midwestern United States, 2006–2015. ISPRS International Journal of Geo-Information, 8(5), 240.

Kamir, E., Waldner, F., & Hochman, Z. (2020). Estimating wheat yields in Australia using climate records, satellite image time series and machine learning methods. ISPRS Journal of Photogrammetry and Remote Sensing, 160, 124-135.

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in neural information processing systems, 30.

Scavuzzo, C. M., Scavuzzo, J. M., Campero, M. N., Anegagrie, M., Aramendia, A. A., Benito, A., & Periago, V. (2022). Feature importance: Opening a soil-transmitted helminth machine learning model via SHAP. Infectious Disease Modelling, 7(1), 262-276.

Shi, B., Yuan, Y., Zhuang, T., Xu, X., Schmidhalter, U., Ata-UI-Karim, S.T., Zhao, B., Liu, X., Tian, Y., Zhu, Y. and Cao, W. (2022). Improving water status prediction of winter wheat using multi-source data with machine learning. European Journal of Agronomy, 139, p.126548.

Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992, July). A training algorithm for optimal margin classifiers. In Proceedings of the fifth annual workshop on Computational learning theory (pp. 144-152).

Drucker, H., Burges, C. J., Kaufman, L., Smola, A., & Vapnik, V. (1996). Support vector regression machines. Advances in neural information processing systems, 9.

Breiman, L. (2001). Random Forests. Statistics Depart-ment. University of California, Berkeley, CA, 4720.

Maheswari, M. U., & Ramani, R. (2023, March). A Comparative Study of Agricultural Crop Yield Prediction Using Machine Learning Techniques. In 2023 9th International Conference on Advanced Computing and Communication Systems (ICACCS) (Vol. 1, pp. 1428-1433). IEEE.

Geetha, V., Punitha, A., Abarna, M., Akshaya, M., Illakiya, S., & Janani, A. P. (2020, July). An effective crop prediction using ran-dom forest algorithm. In 2020 International Conference on System, Computation, Automation and Networking (ICSCAN) (pp. 1-5). IEEE.

Fukushima, K. (1980). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. Biological cybernetics, 36(4), 193-202.

Hubel, D. H., & Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. The Journal of physiology, 160(1), 106.

Xu, B., Wang, N., Chen, T., & Li, M. (2015). Empirical evaluation of rectified activations in convolutional network. arXiv preprint arXiv:1505.00853.

Gauss, C. F., & Bertrand, J. (1957). Gauss's Work (1803-1826) on the Theory of Least Squares.

Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. International journal of forecasting, 22(4), 679-688.

Wright, S. (1921). Correlation and causation.

Kumar, D., Kumar, Y., Gulati, A., & Kukreja, V. (2022, October). Wheat Crop Yield Prediction Using Machine Learning. In 2022 International Conference on Data Analytics for Business and Industry (ICDABI) (pp. 433-437). IEEE.

Nguyen, Q.H., Ly, H.B., Ho, L.S., Al-Ansari, N., Le, H.V., Tran, V.Q., Prakash, I. and Pham, B.T. (2021). Influence of data splitting on performance of machine learning models in prediction of shear strength of soil. Mathematical Problems in Engineering, 2021, pp.1-15.

Fashoto, S. G., Mbunge, E., Ogunleye, G., & den Burg, J. V. (2021). Implementation of machine learning for predicting maize crop yields using multiple linear regression and backward elimination. Malaysian Journal of Computing (MJoC), 6(1), 679-697.

Appendix

Climate dataset

Features		Description
Date		Date and time
GrowRoom		Grow Rooms from 1-15
Attributes	CO2	Carbon dioxide is measure in parts per millions(ppm)
	Temperature	Recorded temperature for specific dates
	Humidity	Concentration of water vapors in the air (grams per cubic meter of air)
	Relative Humidity	Measure of water vapor in water air mixture compared to max amount possible (%)

Manual dataset

Features	Description
ID	Serial number for the records
GrowRoom	Grow Room from 1 to 15
Date	Date and time the details were taken
DAT	Date of Transplant after 21 days from the growing tray to the vertical tower
BatchID	Unique batch ID for each plant
SeedType	The four different types of seeds (Exanimo, Zac, Crystal lettuce and Incised lettuce)
Wagon	wagon number
PlantPos	The position of each plant in the vertical tube in 1-15 growrooms.

Value	Values for the different attributes
Attribute	Different values taken during the whole plant life cycle
Weight	total weight of the lettuce plant
Weight_Packable	Yield – fresh head weight (target variable)
Weight_Waste	The weight of the lettuce plant is not consumable
Humidity	Concentration of water vapors in the air (grams per cubic meter of air)
Air_temperature	the ideal air temperature
Leaf_temperature	Temperature of the lettuce leaves
Leaf_count	Number of leaves recorded for each plant
Leaf_vpd	VPD stands for Vapor Pressure Deficit, refers to the difference in water vapor pressure between the amount of moisture the air could hold at a specific temperature
Plant_height	the height of the plant recorded
Canopy_area	Refers to the total surface area covered by the leaves of the plants in a given area
Roots_temperature	The recorded temperature of the roots of each lettuce plants
Roots_EC	Electrical conductivity of the nutrient solution in which the lettuce plant's roots are submerged

Water analysis dataset

Features	Description
DateReceived	Date and time the water nutrients were given

EndDate	The date until the same water nutrients are supplied
GrowZone	GrowRooms were merged in different GrowZones (1-5)
GrowRoom	Grow Rooms from 1 to 15
pH	Acidity level of water (5.5 to 8 is the optimal pH level)
EC	Electrical Conductivity (nutrient level in water, normal range is from 2 to 3.5
Alk	Alkaline level
K	Potassium level (mg/L)
Cl	Chlorine
Zn	Zinc
Mg	Magnesium is a component of chlorophyll
Mn	Manganese plays an important role in photosynthesis
Mo	Molybdenum (This is required for nitrogen metabolism)
NO3-N	Concentration of nitrates in nitrate ion termed as “Nitrate Nitrogen”
S	Sulphur
B	Boron
Al	Aluminium
NH4-N	Ammonium – an inorganic nitrogen compound
Ca	Calcium level
Cu	Copper level
Si	Silicone level
Na	Sodium level
Fe	Iron Level
Ni	Nickel level
P	Phosphorus Level

Declaration of student's efforts

- Section 1 and 1.1 written by Divya and Harry
- Section 2, 2.1 and 2.2 by Divya
- Section 2.3 and 2.4 written by both
- Section 3, and 3.1 written by both
- Section 4 and 4.1 by Harry
- Section 4.2, and 4.3 by both
- Section 5, 5.1, 5.2, 5.3, 5.4 and 5.5 written by both
- Section 6 and 6.1 by both
- Section 7 and 7.1 by both
- RQ1 by Harry
- RQ2 by Divya and Harry
- RQ3 by Divya
- Discussion with Ljusgårda team and supervisor was performed by both of us.