Shanay Tolat

Data Bootcamp (ECON-UB 232)

Prof. Jacob Koehler

12th May 2025

Morphological and physiological features impacting vegetative plant biomass growth

Introduction

Modeling plant growth is increasingly becoming important as greenhouses and hydroponics become more relevant today. And a great way to track the growth is through the biomass of the plant – which refers to the total mass of organic matter produced by plants in a given area. This organic matter is usually produced through the process of photosynthesis, where carbon dioxide and water is transformed into sugars and oxygen in the presence of sunlight ("Plant Biomass - an Overview | ScienceDirect Topics," n.d.).

During the process of plant growth, the plant's tissues congregate in large amounts of water. To accurately measure the growth of the total biomass of the plant, it is thus necessary to remove this water weight from the plant entirely. The biomass, or dry matter, of the plant consists of what remains after water has evaporated from the plant's architecture. This dry mass includes the stems, leaves, roots, cells etc (Calf Health Basics 2018).

Vegetative growth is growth that occurs in the non-reproductive organs of the plants, which don't include fruits and flowers (Poorter et al. 2011). These biological concepts are important before introducing the dataset since the target variable of the dataset is the percentage of dry matter in vegetative growth (PDMVG). While the exact definition of the vegetative part of the plant varies, this dataset explicitly states that it only includes above ground features (not the roots). The higher the PDMVG, the more efficient the plant is at converting nutrients and carbon dioxide into growth tissue. To measure the PDMVG of a plant, the vegetative part of the plant is harvested and weighed. The plant is then dried in an oven to make the water evaporate before weighing it again. Then use the formula:

$$PDMVG\,(\%) = \left(\frac{Dry\ weight}{Fresh\ weight}\right) \times 100$$

Dataset overview

Since many researchers have started to develop algorithms to predict plant growth rates based on certain conditions, I am to create my very own models to predict the biomass growth in plants using the 'Greenhouse Plant Growth' dataset from Kaggle. This dataset has 14 columns that explore the morphological and physiological characteristics that impact the plant biomass growth rate. The dataset includes 30,000 samples of different plants and these are all the feature variables included:

- Random - random identifier representing sample batches
- ACHP - average chlorophyll content per plant, an indicator of photosynthetic ability
- PHR - plant height rate - the vertical growth of the plant over the period of data collection
- AWWGV - average wet weight of vegetative growth - total fresh weight of the above ground parts
- ALAP - average leaf area per plant
- ANLP - average number of leaves per plant
- ARD - average root diameter
- ARL - average root length
- ADWR - average dry weight of roots
- AWWR - average wet weight of roots
- PDMRG - percentage of dry matter in roots
- Class - categorical label for experimental purposes

All these columns were grouped into certain sub-categories that allows for a more streamlined problem statement. The categories are: leaf architecture (ACHP, ANLP, ALAP), root structure

(ARD, ADWR, ARL, ADWR, AWWR, PDMRG), and the physiological feature (PHR). Even though the roots are not considered part of the vegetative segment of the plant in the study, they are essential for nutrient uptake and transportation. The models should show how the structure of the roots not only increases root biomass but also the rest of the plant. The groupings are relevant because, as theory will explain later, some features aren't relevant in isolation, but when they are used together with other features they increase their predictive property.

This brings us to the problem statement of our project: *Which structural features of the plant are the most relevant in predicting the PDMVG? And which model is the best at capturing relationships within the dataset to produce the most accurate results?*

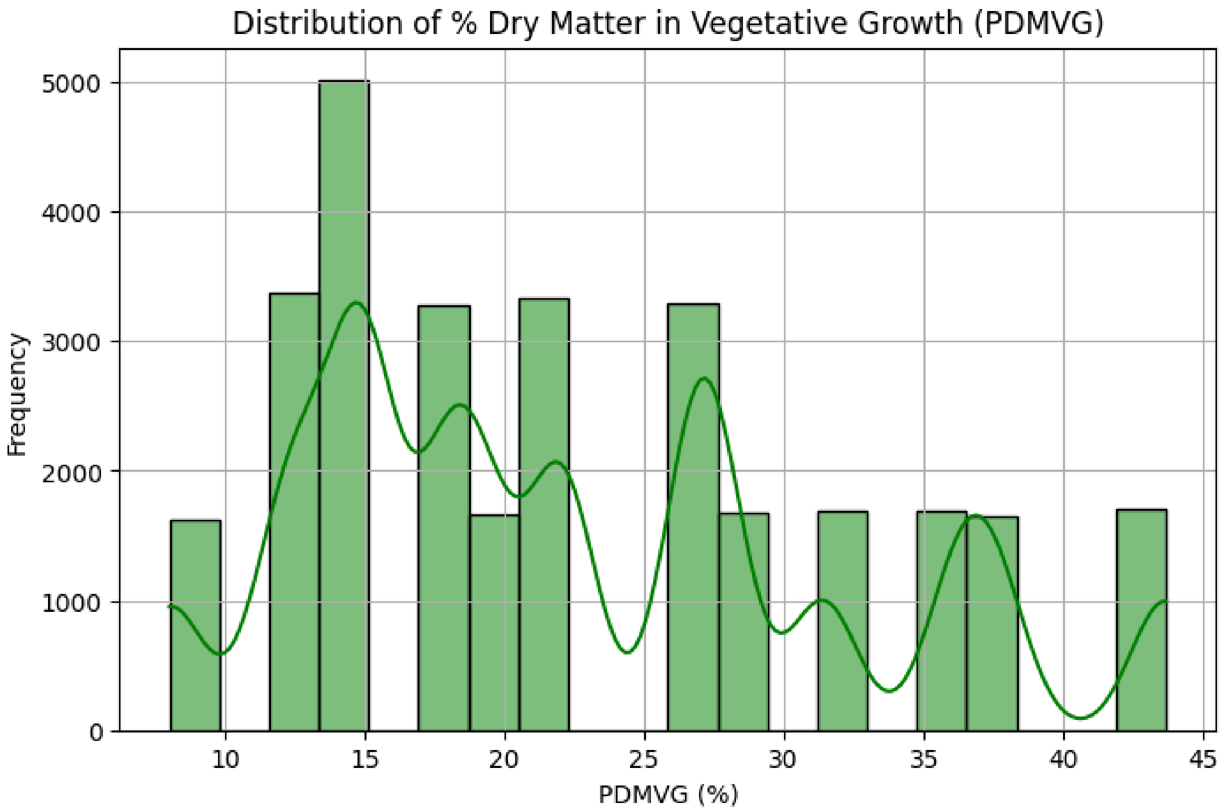
The models that are going to be used are a Baseline model (for statistical benchmarking), Multiple Regression, K-Nearest Neighbours, Random Forest, Neural Networks, and Decision Trees. Before that, however, we must preprocess the data and perform an Exploratory Data Analysis (EDA) to understand trends in the dataset.

Data Preprocessing

Before using the dataset for EDA and Regression Analysis, some data preprocessing was conducted. Initially a mean imputation was performed since there were some missing values in physiological variables, possibly due to measurement gaps or sensor dropouts during the trials. The mean imputation for continuous features preserves the central tendency for data. As discussed in the Introduction, since the PDMVG is calculated using the Average wet weight of vegetative parts and the average dry weight of vegetative parts, those two columns were dropped. Along with this, the columns 'Random' and 'Class' were also dropped since they both were only present for experimental purposes and had no predictive value.

Exploratory Data Analysis

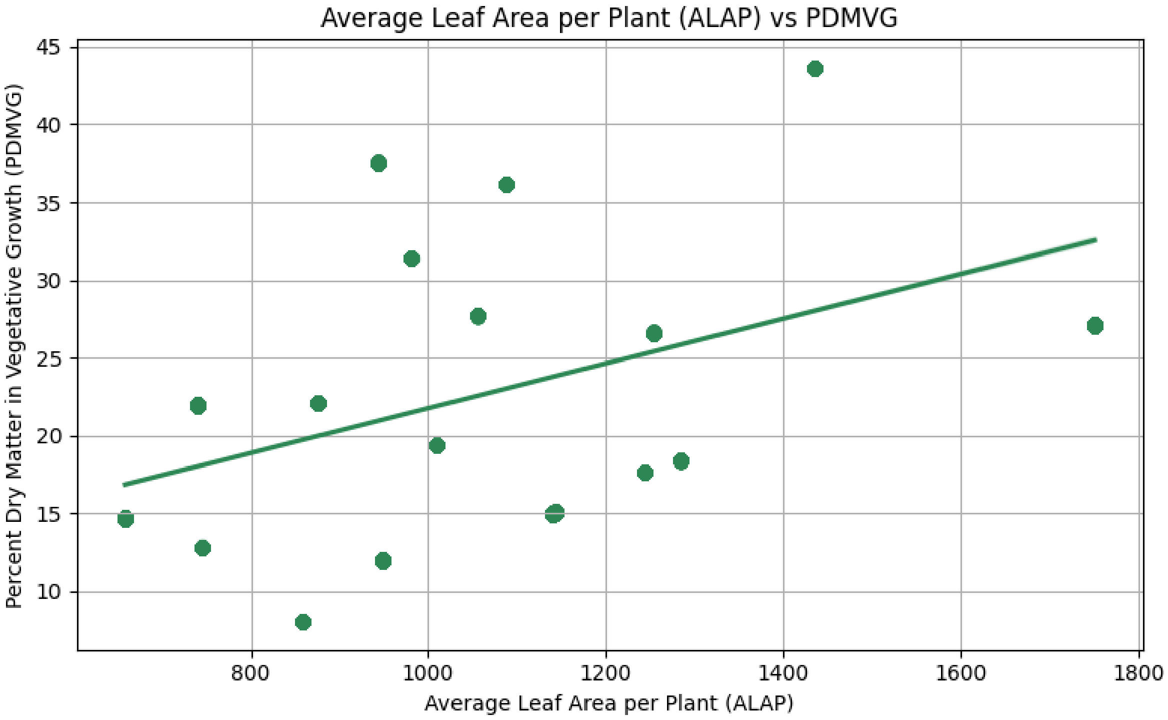
Distribution of PDMVG



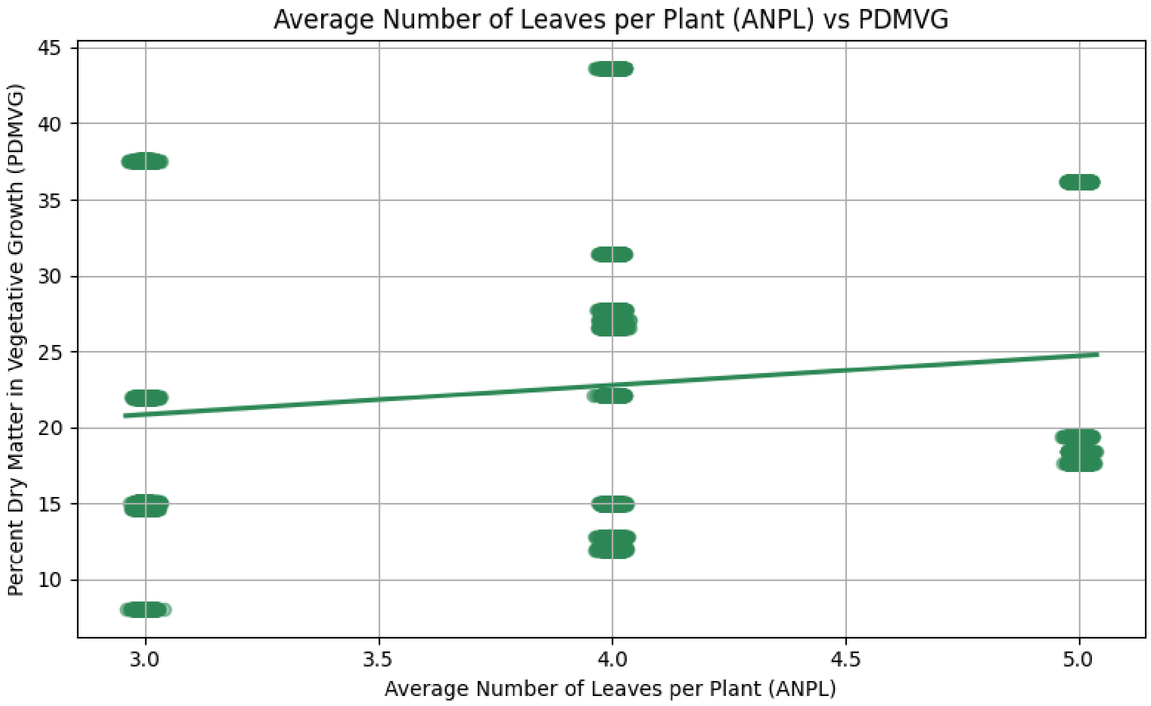
The distribution shows its highest peak at 12 to 15%, which is considered ideal for most plants. It is multimodal, meaning that there are heterogeneous plant responses in the dataset. This means that the dataset is made up of different subgroups – perhaps implying that various species of plants were used in the sample. These subgroups correspond with the ‘Class’ column in the dataset, which I dropped since it didn’t have any predictive significance.

Leaf Architecture as a determinant of Vegetative Biomass Accumulation

The accumulation of dry matter in the vegetative tissues of the plant, which is what is quantified by PDMVG, is a result of a process called net primary production (NPP). By theory, NPP is directly proportional to the total green surface area of the plant ("Net Primary Production - an Overview | ScienceDirect Topics" 2018). Since the total green surface area of the plant would be equal to the average number of leaves multiplied by the average leaf area it would make sense for both of them to show positive correlations with the target variable. To view the relationship, a scatterplot with a regression line was plotted between ALAP and PDMVG.



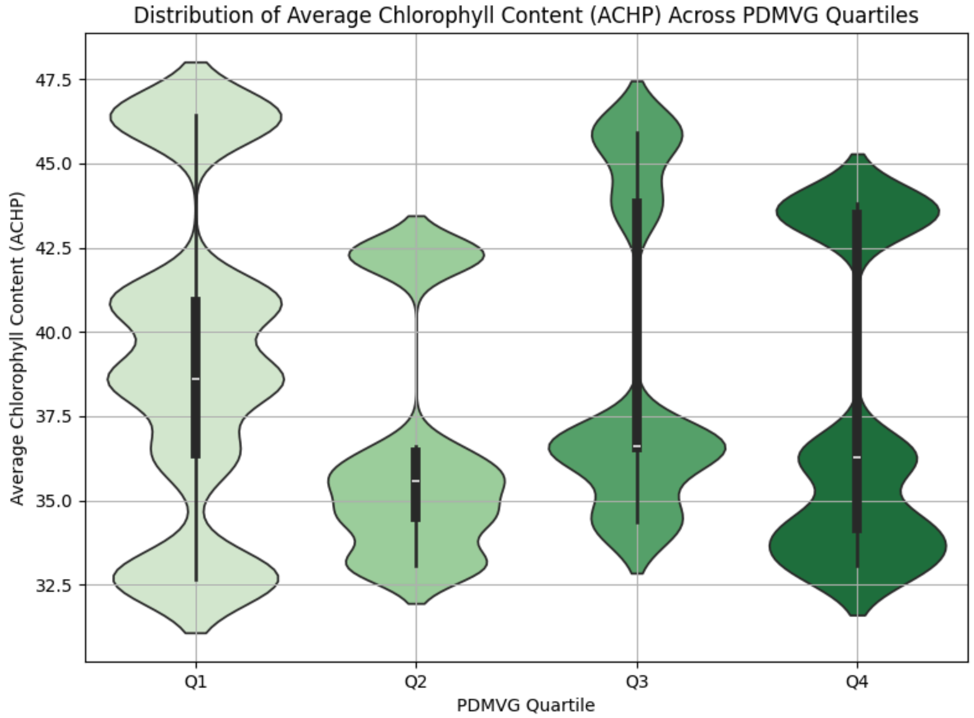
The graph obtained shows a clear positive trend between the dry matter and leaf area per plant. Studies in the past using tomato and maize have shown that higher ALAP correlates with more photosynthesis, which is a driver of metabolic activity in plants. A similar scatter plot was drawn for ANPL.



The ANLP shows a weakly positive relationship with PDMVG. This can likely be explained by the fact that excessive number of leaves may also cause internal shading within the plant. This means that the more leaves there are, the more likely it is that the leaves' shadows would prevent other leaves from getting sunlight in the greenhouse. Thus, simply having more leaves wouldn't always result in a more effective photosynthetic environment. In all the plants used in the experiment, the average number of leaves also varied from only 3 to 5, which is not a sufficiently wide range to make any judgements regarding the relationship between the two variables. While scientifically there would be a positive correlation, it is not fully supported by the dataset.

Chlorophyll levels

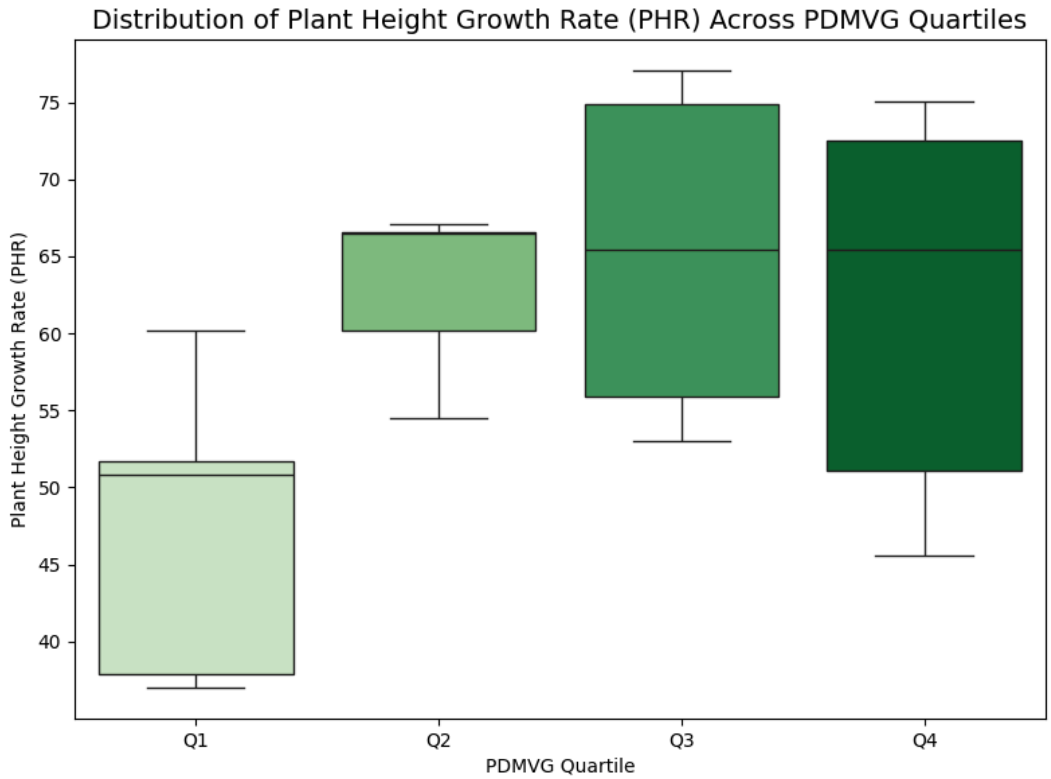
Chlorophyll is the primary pigment that drives photosynthesis in plants, it helps with the absorption of light to convert carbon dioxide and water to produce sugar and oxygen. The presence of chlorophyll is probably one of the strongest features that would predict organic biomass for this very reason. In the latter prediction models that will be built as well, it would be expected for this to come as one of the factors with most importance. To analyse the relationship between Average chlorophyll content (ACHP) and PDMVG, I first created a new variable called 'PDMVG_q' to split the PDMVG into quartiles. The violin plot would allow us to see not only the variation but also the central tendencies of each quartile.



Plants that are in the lowest quartile of PDMVG showed a very wide distribution of average chlorophyll content ranging from 32.5 to 47.5. This suggests more inconsistencies in the efficiency of plants that had the lowest growth rates. It could mean that other factors such as nutrient deficiencies and other environmental conditions were hampering the growth of the plant, not just the chlorophyll content. It was surprising that the peak chlorophyll content was seen in a plant that was in the lowest quartile of PDMVG. Quartile 2 showed a more compact but relatively low chlorophyll content compared to the other groups. The last two quartiles have higher median ACHP values as predicted. However, the graph doesn't give us enough confidence alone to believe that ACHP should be a strong indicator as the theoretical backing suggests.

Plant height growth rate (PHR)

A plant’s height growth rate is theoretically supposed to have a positive relationship with the PDMVG. This is because taller plants generally have a better photosynthetic canopy and can also receive more sunlight since they are less likely to be covered by other plants. To explore this relationship, a box plot was drawn, with the same quartiles of PDMVG.



The graph showed a relationship that does exhibit positive correlation between the two variables but this relationship plateaus after a certain point. It means that the plant's height growth rate only influences the PDMVG up to some height after which a taller plant doesn’t show higher biomass. The first quartile of the PHR showed a median of 51 after which the median increased to 66 and remained the same in the third and fourth quartiles as well.

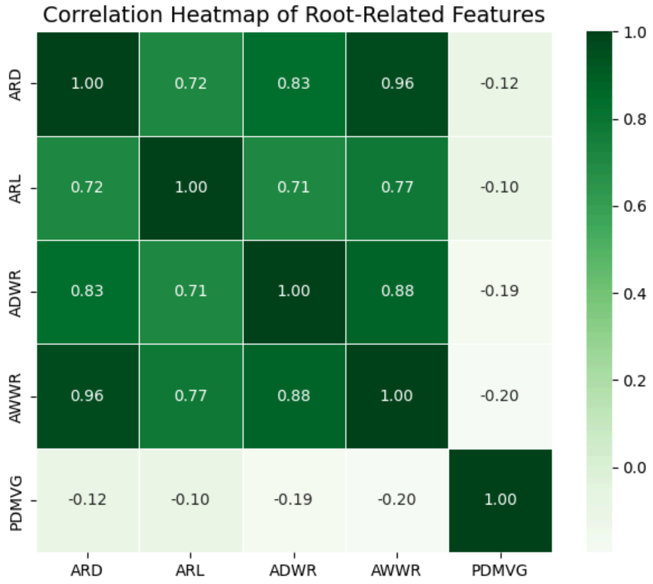
Root-related physiological features

Since the dataset contained multiple root-related features a correlation heatmap can be drawn. It would be expected that the ADWR and AWWR would show strong correlations with the PDMVG since they also represent biomass accumulation. Theoretically, plants with larger roots can have more expansive access to nutrients and water, permitting a higher metabolism and growth rate. Thus there should be a positive correlation.

As for the average root diameter and average root length, if we assume that roots are perfect cylinders for simplicity, we know that ("Volume of a Cylinder Calculator," n.d.):

$$Volume \propto \left(\frac{D}{2}\right)^2 \times L$$

If the volume of roots increases it could be assumed that the water and nutrient uptake would also increase, making us expect a weakly positive correlation between these factors and PDMVG.



There are strong interrelationships between the root related factors that would be expected. Thicker and longer roots are likely to have a higher biomass. However, all the features showed a weakly negative correlation with PDMVG, meaning that there is almost no correlation. It would have been assumed that plants with higher root biomass would also correlate with higher biomass in the above ground parts. It is likely that these relationships are either non-linear or that these features work better together to predict the biomass accumulation, so they will be evaluated after building the models.

Modelling

Baseline model

To set the initial benchmark for comparison with other models, a baseline regression model was created and implemented. This would predict the mean of the training dataset for all test observations. This model yielded a Mean Squared Error (MSE) of 92.68, which serves as a statistical floor. All the other models that are created should reduce this MSE value since the baseline model doesn't use any predictor variables and assumes homogeneity.

Multiple Regression model

A multiple regression will examine the linear relationship between the PDMVG and the independent variables in the dataset unlike a simple regression that only analyses a single features impact on the dependent variable. This regression model will allow for the leaf architecture, chlorophyll levels, and root-related physiology to be accounted for in predictions. The MSE of the test set was 27.27, which compared to the baseline MSE of 92.67 is a great improvement. It means that the model is indeed capturing the predictive value of the features.

The fact that the difference between the test MSE and training MSE was only about 1 unit means that the model is not overfitting and has learned to generalize well to unseen data.

The R^2 value is 70.6 which indicates that 70.6% of the variation in the dry biomass growth is attributed to changes in the independent variables. This R^2 value is not very high for a good statistical model, but it would have been expected given that there was mutli-colinearity between some of the features, which other models account for. After looking at the coefficients in the given model, it was noted that the plant height rate growth, average dry root weight, and average root diameter are the strongest positive contributors to vegetative dry matter. In the correlation heatmap that I created in my EDA, there was actually a negative relationship between plant diameter and PDMVG so it is surprising that it is one of the strongest predictors. However, as the volume of the plant roots is directly correlated with the square of the diameter, it is a reasonable assumption to make in terms of biological backing.

Apart from the coefficients, the importance of each feature was also looked at through a method that estimates how much shuffling the feature worsens the performance of the model. This permutation method is used to tell how each feature contributes to predictive quality of the model. By this analysis, it was noted that the average root diameter drastically reduces the performance of the multiple regression, meaning that it is important for the accumulation of plant biomass. After this, the average leaf area and the plant height growth are the next two physiological features that had highest importance. This makes sense as our theoretical backing and EDA suggested that they would serve as important features in the models.

In biology, the resource allocation hypothesis suggests that plants that have strong root systems are important to absorb water and nutrients, contributing to higher dry mass growth (Keller

2010). Even though our EDA correlation heatmap showed weak and negative correlations between PDMVG and the root features, they were found to be the most significant features in the multiple model.

K-Nearest Neighbours

The K-Nearest Neighbours model is used in this analysis since it helps capture non-linear relationships that are likely to exist in greenhouse plant growth. While the multiple linear regression assumes a global linear structure between the independent and dependent variables, KNN approximates predictions based on ‘neighbouring’ plants with similar traits. After using a grid search the optimal number of neighbours that was obtained was 30. The training MSE was 0.00009583 and the test MSE was 0.0001, which was a huge improvement from both my baseline and multiple regression model. The difference between the performance of the model in my training and testing set is also lesser compared to previous models. The R^2 value was also extremely high: 0.99999892, which means it was near perfect in capturing the patterns of the dataset.

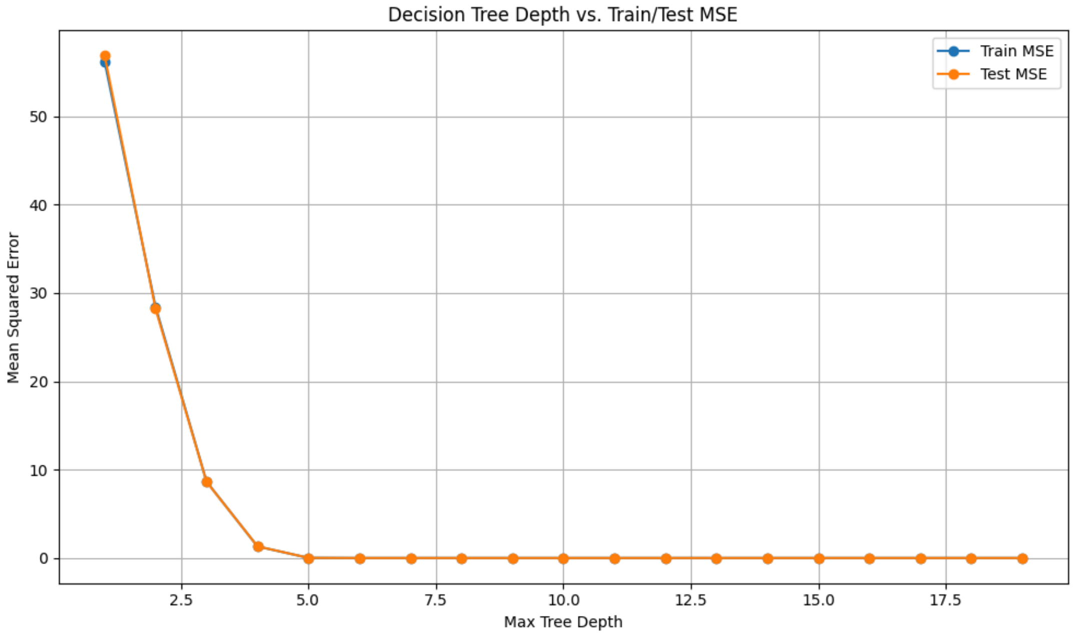
After the permutation based feature importance was carried out, the average leaf area per plant and the plant height growth ratio were significantly the most important variables. Again, this is backed by the biological theory discussed in the EDA. As plants grow taller, they are likely to be able to access more sunlight compared to other plants in the greenhouse and thus increase metabolic activity and biomass accumulation. Along with this, more leaf area corresponds to a greater photosynthetic surface on the plant. The average root length and diameter were the next most important features, since the volume of the root and hence weight is reliant on the diameter and the length of the roots. It was surprising that this model also didn’t show a very high importance for chlorophyll content. This must be because having a higher chlorophyll content doesn't compensate for having a lower leaf area, but having a larger leaf area can proportionally increase efficiency and compensate for a lower chlorophyll content.

The high performance of the KNN model suggests that the relationship between PDMVG and the plant features isn't linear and highly depends on the combination of features that a plant has. In this scenario, especially with the dataset having different subgroups of plants, the model can compare each plant with others that have similar characteristics rather than merely finding an overarching pattern in the entire dataset. This makes it easier to actually model greenhouse conditions where the micro-environment of the plant varies, and these small adjustments can be accounted for in this model.

Decision tree model

A decision tree model helps with the identification of hierarchical relationships in the dataset. Rather than using a linear or distance based approach, the model trains itself to create a flow-chart of sorts that splits the data into homogeneous subgroups based on the features of the plants. In the decision tree, the most important features that were used in splitting the dataset into subgroups of plants included average leaf area per plant, plant height growth rate, and average root diameter. Yet again, these features seem to consistently be the most significant in predictions. The visualized tree allows for ease in following the rules that are created.

One strength of this model is that the train and test MSE are the closest out of any model (0.5043 and 0.5078 respectively) we have used till now. It means that the model is good at generalizing to newer data. The model shows that after a decision tree depth of 5, the model would have started overfitting to particular segments of plants that would reduce its performance.



As shown from the graph, the lines of the train MSE and test MSE almost perfectly align with one another, showing that the model trained itself to the new data very well. Along with this, each layer of the decision tree exponentially reduced the MSE until a tree depth of 5, after which the MSE plateaued.

Random Forest model

Random forest works by aggregating multiple decision tree outcomes to stabilize the predictions that it makes and seeks nonlinear relationships across the dataset. It would be useful in modelling plant biomass growth since the process includes the interactions of various features in a nuanced manner. The optimal model that was found used 200 estimators and a maximum tree depth of 6 to get a training MSE of 0.00009657 and a test MSE of 0.00009721. This has by far been the best performing model. While the test MSE of the KNN model was almost the same, the KNN model's training MSE was slightly higher (however the difference is trivial). Again, as with the other models, the permutations showed that the plant height growth rate was an important factor.

However, along with this, it used the average root length as well. The biological basis of this model essentially suggests that a plant that is taller and better anchored to the ground will be able to maximize its access to both above ground and below ground resources. Finally, as predicted in our EDA, this model used the average chlorophyll content as one of the more important predictors in plant biomass growth.

Neural network regressor

Neural networks are machine learning models that learn certain patterns from data, essentially layers of neurons to do calculations each step of the way. The train MSE was 0.00501 and the test MSE was 0.00483. While this model still worked extremely well, KNN and Random Forest worked better. Neural networks are good at finding non-linear relationships between the features. The R-squared value from the neural network was extremely high at 0.99995, suggesting that 99.995% of the variation in PDMVG was explained by the neural network regressor.

Key Findings

1. *Random Forest was the most effective model*

   Among all the regression models tested, Random Forest provided the lowest mean squared error in the testing data. It also had the closest testing and training MSE. The fact that it aggregates the outcomes of various decision trees permits it to really model the interplay between the leaf structure, height, and root structure of the plants. Since both the KNN model and the Random Forest model are good at capturing non-linear relationships, it was fitting that both of them yielded the best results. This means that the interaction between the plant features are dependent on interrelationships.

2. *Height and length proved to be the most important*

   Across all the models, the plant height growth rate and the average root length were repeatedly among the most important features in predicting PDMVG. Their predictive power is also grounded in agronomic principles discussed earlier.

3. *Inconsistencies in root related features across models*

   The EDA showed that root related features individually had no predictive power for the PDMVG. This was due to the fact that they work together as a system to impact the biomass growth in the vegetative part of the plant. This is supported by the fact that the diameter and length hold no value in isolation, but together they can be used to calculate the volume of roots. The volume of roots then can be used as a basis for the biomass of the root. This collinearity between the features was only accounted for in the models other than the multiple linear regression.

Next steps/Improvements

1. *Integration of environmental conditions of the greenhouse*

   Some very imperative features that biomass growth is reliant on is humidity, temperature, light intensity etc. While it may be hard to obtain for micro-environments (the environment for each plant within the greenhouse), there are new sensors and technologies that could be placed in every meter square of the greenhouse and record data for those specific plants. The dataset currently contained data on the physiological features in different parts of the plant, however, it was assumed that they all received the same amount of water, carbon dioxide, light etc, which may or may not be true. Especially when the MSE of the model reaches a point where it is in the thousandths, then understanding the micro-climate would be the only thing that could improve the model.

2. *Create a classification model*

   The dataset had a ‘Class’ column which indicated the experimental group that the data was a part of, but this was dropped since we wanted to make a regression model rather than a classification model. Thus, in the future, the author of the dataset themselves suggested using machine learning tools to create classification models for the plants.

3. *Switching to a multi-output regression model*

   Even though the target variable was the PDMVG, there was a column for the plants’ biomass growth in its root. If we conduct a multi-output regression then we can analyse not only the predictive efficiency of the models, but also understand the conditions under which there are tradeoffs between the dry mass accumulation in the root and in the vegetative part of the plant.

4. *SHAP (Shapley Additive Explanations)*

   This is a tool that is used to interpret machine learning models, similar to the permutations method that we currently used. It helps identify the importance of various features within the regression model. SHAP provides values given on the basis of game theory: how much each feature contributes to the final prediction made by the model. It is almost as though all the features together form a team, whose leader you are trying to find through these values (Lundberg 2018). If we only look at coefficients then it only says the magnitude and the direction of the change (and this is limited to linear relationships). The permutation importance tells us how much the performance of the model drops if each feature is shuffled. The disadvantage of this is that it can sometimes be unsuitable if the features are correlated. While the scope of this project was to limit analysis to 3500 words, in the future I hope to calculate these values and further analyse them.

Bibliography

Calf Health Basics. 2018. "Measuring the Dry Matter Content of Feeds." Uga.edu. August 2018. https://extension.uga.edu/publications/detail.html?number=SB58&title=measuring-the-dry-matter-content-of-feeds.

Keller, Markus. 2010. "Environmental Constraints and Stress Physiology." *Elsevier EBooks*, January, 227–310. https://doi.org/10.1016/b978-0-12-374881-2.00007-6.

Lundberg, Scott. 2018. "An Introduction to Explainable AI with Shapley Values — SHAP Latest Documentation." Readthedocs.io. 2018. https://shap.readthedocs.io/en/latest/example_notebooks/overviews/An%20introduction%20to%20explainable%20AI%20with%20Shapley%20values.html.

"Net Primary Production - an Overview | ScienceDirect Topics." 2018. Sciencedirect.com. 2018. https://www.sciencedirect.com/topics/earth-and-planetary-sciences/net-primary-production.

"Plant Biomass - an Overview | ScienceDirect Topics." n.d. Www.sciencedirect.com. https://www.sciencedirect.com/topics/engineering/plant-biomass.

Poorter, Hendrik, Karl J. Niklas, Peter B. Reich, Jacek Oleksyn, Pieter Poot, and Liesje Mommer. 2011. "Biomass Allocation to Leaves, Stems and Roots: Meta-Analyses of Interspecific Variation and Environmental Control." *New Phytologist* 193 (1): 30–50. https://doi.org/10.1111/j.1469-8137.2011.03952.x.

“Volume of a Cylinder Calculator.” n.d. Www.omnicalculator.com. https://www.omnicalculator.com/math/cylinder-volume.