
# Biomass Estimation by Plant Phenotype Data Analysis

Implementation and comparison of machine learning models for biomass estimation using plant phenotype data, aiming to find the most effective and accurate approach for non-destructive biomass assessment in plants.

Faculty of Information, Media and Electrical Engineering
Technische Hochschule Köln

| | |
|---|---|
| Author: | Varun Sringeri Lakshmikanth |
| Matriculation Number: | 11145408 |

| | |
|---|---|
| Internal Examiner: | Prof. Uwe Dettmar |
| External Examiner: | Prof. Marcel Bucher |

December x, 2022

# Declaration

I certify that I have independently written the thesis I have submitted. All passages, I have taken all passages, whether verbatim or in spirit, from published or unpublished works of other authors or of the author himself/herself. of others or of the author himself/herself, I have marked them as such. marked as taken. All sources and aids that I have used for this work are indicated, are indicated. The thesis has not been submitted with the same content or in essential parts to another been submitted to any other examination authority.

Place, Date

Signature

# Abstract

Estimating biomass is a crucial challenge in plant science and agriculture because it offers important information on the development, productivity, and use of resources of plants. ecologists view predicting plant biomass as a vital objective. Finding a predictive biomass model across tests is, however, quite difficult. Destructive sampling, which is time-consuming and labor-intensive, is frequently used in traditional approaches. Therefore, quick, accurate, precise, and non-destructive phenotyping methods for biomass yield are required. Non-destructive biomass measuring techniques have made an effort to solve this issue with the introduction of image-based high-throughput plant phenotyping facilities. This study examines the use of machine learning models for biomass estimate using plant phenotypic data, including Random Forest (RF), Support Vector Machine (SVM), Multilayer Perceptron (MLP), k-Nearest Neighbours (KNN), and the required characteristics/features are extracted using an automated script and utilized to predict biomass using trained models. By utilizing the wealth of data included in plant phenotypic features, the study seeks to develop precise and effective methods for biomass assessment. Due to their aptitude for collecting complicated correlations and managing high-dimensional data, RF, SVM, MLP, and KNN were chosen. A broad dataset of plant phenotypic traits and accompanying biomass measurements were used to train the models. The outcomes reveal that these machine learning models perform effectively in estimating biomass. By capturing non-linear correlations and classifying various classes, respectively, RF and SVM are able to provide precise predictions. While KNN efficiently uses nearby samples for the estimate, MLP excels at learning complicated patterns. The strengths, weaknesses, and applicability of various models for biomass estimating tasks are shown through comparative analysis and evaluation. The research advances biomass estimating methodologies by providing non-destructive, scalable, and trustworthy ways to gauge plant growth and output. By enabling more effective monitoring and management of plant populations in various agricultural and ecological settings, the adoption of these techniques has the potential to influence plant science research and agricultural practices. This study promotes non-destructive and precise biomass estimation techniques in plant science and agriculture by offering a script-based solution for biomass estimation utilizing machine learning models.

**Keywords**:

# Contents

# Chapter 1

# Introduction

In plant science and agriculture, biomass estimation is essential because it offers important information on the growth, productivity, and resource use of plants. Biomass is a potent indicator for agricultural applications because of its direct relationship to the health and growth status of the crop. Predicting the sequential biomass of plants may be useful in correlating crop biomass with environmental growth. Resource management, crop prediction, and plant growth can all benefit from accurate and non-destructive biomass assessment. However, obtaining accurate and effective biomass estimations is extremely difficult. Destructive sampling is a common component of traditional approaches for estimating biomass, which is time- and labor-intensive. Thus, there is a rising need for non-destructive, precise, and scalable biomass estimating methods. The use of machine learning models for biomass estimation using plant phenotypic data is explored in this thesis study. This study focuses on using machine learning models for biomass prediction based on plant phenotypic traits, such as Random Forest (RF), Support Vector Machine (SVM), Multilayer Perceptron (MLP), and k-Nearest Neighbors (KNN). These models offer a more comprehensive and precise way to represent the complex interactions between phenotypic features and biomass data. These models are good at capturing complicated relationships and managing high-dimensional data. These models make use of the wealth of data included in plant phenotypic features in an effort to create precise and effective methods for biomass estimation.

An ensemble learning technique called Random Forest (RF) mixes different decision trees to produce predictions. It excels in handling high-dimensional data and capturing non-linear relationships. Contrarily, SVM is a supervised learning technique that divides many classes using a hyperplane. It's a good option for biomass estimation because it can handle both linear and non-linear data. A form of feedforward neural network called a multilayer perceptron (MLP) has many layers of interconnected nodes. MLP is an effective approach for estimating biomass because it can learn intricate patterns and correlations. A non-parametric technique called k-Nearest Neighbors (KNN) classifies samples according to how close they are to training data points. It has been used effectively in biomass estimates and can be modified for regression problems. This research examines various machine learning models in an effort to create precise and effective biomass estimation techniques that take advantage of the wealth of information in plant phenotypic data. For the purpose of estimating biomass, RF, SVM, MLP, and KNN will be compared and evaluated in order to shed light on their capabilities, weaknesses, and performance. By offering non-destructive, scalable, and reliable approaches for evaluating plant growth and production, the use of these models in biomass estimate has the potential to change plant science research and agricultural operations.

By advancing biomass estimation tools, this research makes it possible to monitor and manage plant populations in various agricultural and ecological situations more successfully.
Automated image-based phenotyping approaches for field crops have become popular with plant breeders [**?**]. Specifically, a script that uses an image as input, extracts the necessary features, and then produces biomass estimations using the trained models has been created to simplify the estimation procedure. The goal of the script is to automate the estimation of biomass while minimizing manual labor and increasing the effectiveness of biomass assessment. The script enables rapid and non-destructive assessments by utilizing image-based high-throughput plant phenotyping facilities. It takes the input image and extracts pertinent properties that the machine learning models created in this study can use as inputs. Based on the features retrieved from the input image, the script uses the trained models to produce estimations of the biomass. This method offers a non-destructive, scalable, and reliable alternative to destructive sampling for biomass estimation. The script offers effective monitoring and evaluation of plant development and productivity by automating the estimating procedure. The creation and application of a script that automates the biomass estimation procedure using machine learning models is one of the achievements of this thesis report. The accuracy and effectiveness of biomass estimation are improved by including image-based plant phenotypic data. In conclusion, the goal of this thesis paper is to use plant phenotypic data to overcome the difficulties in biomass calculation. The findings of this study have important ramifications for agricultural practices and plant science research. The accuracy, effectiveness, and non-destructive nature of biomass estimation are improved by the introduction of a script that accepts photos as input, extracts the necessary features, and outputs biomass estimates using machine learning models. The methods, findings, and discussions that help enhance biomass estimating methodologies in plant science and agriculture are thoroughly examined in the following chapters.

# Chapter 2

# Fundamentals

## 2.1 Machine Learning

Artificial Intelligence (AI) and Machine Learning (ML) have had a significant impact on almost every aspect of contemporary civilization, resulting in a period of extraordinary technical developments and revolutionizing a number of industries. AI and machine learning (ML) are now essential in multiple fields. In order to recognize patterns and connections that could escape the human eye, ML algorithms can analyze large and heterogeneous datasets, including genomes, medical imaging, and patient records. This not only speeds up the diagnosis procedure but also makes it possible to create individualized treatment programs that improve patient results. Additionally, AI-driven robotic devices are transforming surgical procedures by enabling minimally invasive interventions and increasing surgeons' abilities with increased precision and real-time data feedback. Diverse industries have been transformed by AI and ML, giving professionals the tools they need to solve complex issues, improve decision-making, and spur creativity. The combination of AI and ML has increased production and efficiency while also creating new opportunities for social and scientific advancement. A future where AI and ML constructively contribute to human well-being and sustainable development must be encouraged as these technologies continue to advance. To do this, it is crucial to strike a balance between innovation and ethical considerations.

## 2.2 Neural Network

In the fields of artificial intelligence and machine learning, a neural network is a fundamental idea that aims to replicate the extraordinary capabilities of the human brain. A neural network is made up of a number of nodes, or artificial neurons, stacked in layers and is inspired by the complex network of neurons and their connectivity in the brain. Each neuron receives input data, processes it by adding the inputs' weights together and adding a bias term, and then applies an activation function to the processed data to produce an output. By introducing non-linearity, the activation function enables neural networks to model complicated relationships and identify subtle patterns in the data. An input layer, one or more hidden layers, and an output layer typically make up a neural network's architecture. As the input data spreads across the network, the hidden layers operate as intermediate representations, gradually extracting higher-level properties from the data. On the basis of the learnt representations from the hidden layers, the output layer generates the final prediction or decision.

The ability of neural networks to learn complex and abstract representations from massive amounts of input is one of their primary benefits. They excel at tasks like picture identification, audio recognition, natural language processing, and more because of their ability to automatically find significant patterns in the data. Numerous sectors have been transformed by neural networks, which have helped advance applications for autonomous vehicles, healthcare diagnostics, recommendation systems, and a plethora of other uses. In order to advance artificial intelligence and address challenging real-world challenges, neural networks are expected to play an increasingly important role as they develop and change the machine learning environment.

## 2.3 Data analysis

Data analysis is a critical step in research, business, and decision-making. It involves looking at and interpreting raw data to find important patterns, insights, and trends. It includes a broad range of methodologies, including fundamental descriptive statistics as well as more sophisticated approaches like data mining, machine learning, and statistical modeling. Data is prepared for analysis by being cleansed, arranged, and translated into a structured format during data analysis. Charts and graphs are popular visualization tools used to show facts visually and promote greater understanding. The ultimate purpose of data analysis is to unearth relevant information that might support research hypotheses, assist decision-making, highlight opportunities or problems, and help organizations or enterprises develop strategies. Through data analysis, researchers and analysts may make sense of complicated datasets, identify patterns, and arrive at relevant conclusions that aid in the growth of numerous disciplines and support systems for making decisions using the best available evidence.

## 2.4 Machine learning branches

Machine learning can be divided into several subsets or branches, each with its specific focus and applications. The main subsets of machine learning are:

### 2.4.1 Supervised Learning

The primary machine learning paradigm is supervised learning, where the algorithm is trained on a labeled dataset where each input data point has a corresponding target or label. A mapping from input features to output labels must be learned during the process of supervised learning for the algorithm to generate precise predictions on brand-new, untainted data. To reduce the difference between its predictions and the actual labels in the training data, the algorithm iteratively modifies its internal parameters, such as weights and biases, throughout the training phase. Usually, this optimization procedure uses backpropagation and gradient descent algorithms.

### 2.4.2 Unsupervised Learning

Unsupervised learning is a branch of machine learning that deals with unlabeled data and seeks to identify underlying structures, relationships, or patterns without the use of explicit training data or predefined output labels. It is a more difficult and exploratory procedure than supervised learning

because there is no "ground truth" to compare the algorithm's predictions to. Unsupervised learning algorithms instead concentrate on locating groups of related data points using methods like clustering or locating condensed versions of the data using dimensionality reduction. In contrast to dimensionality reduction approaches, which try to reduce the number of features while keeping important information, clustering algorithms combine data points based on their similarity. For data exploration, anomaly identification, and understanding the fundamental properties of the data, unsupervised learning is essential. It is essential for tasks where the true labels are unknown and where the algorithm needs to find hidden structures in order to extract meaningful information from the input.

### 2.4.3 Semi-Supervised Learning

Between supervised and unsupervised learning is semi-supervised learning. To train the model, it makes use of both labeled and unlabeled data. The algorithm's performance is improved because of the small amount of labeled data which assists it in learning from the unlabeled data. When gathering labeled data is expensive or time-consuming, semi-supervised learning is advantageous since it makes better use of the data that is already available.

### 2.4.4 Reinforcement Learning

A branch of machine learning called reinforcement learning is concerned with training agents ways to interact with the environment and gain knowledge from positive or negative feedback. Over time, the agent learns to operate in a way that maximizes cumulative rewards; this training depends on the concept of trial and error. This method is especially well suited for sequential decision-making problems seen in games, robotics, and autonomous systems.

### 2.4.5 Deep Learning

Deep learning is a specialized and sophisticated branch of machine learning that focuses on developing deep neural networks, which are multi-layered neural networks. In difficult problems like speech recognition, computer vision, and natural language processing, these networks have demonstrated remarkably good results. By gradually extracting more abstract information at each layer, the depth of the neural networks enables them to automatically develop hierarchical representations from unstructured data. The network's numerous variables are altered throughout the training phase, which is often carried out using large-scale datasets, in order to minimize the prediction error. Due to its ability for handling enormous volumes of data and comprehending intricate patterns, deep learning has revolutionized a number of industries and allowed for the creation of cutting-edge models with previously unheard-of capabilities and accuracy.

### 2.4.6 Transfer Learning

Transfer learning is the process of using the skills developed when training a model for one job to enhance performance on a different but related task. Transfer learning involves refining a previously trained model on a fresh dataset in order to apply the learnt features to the fresh task. When the target dataset is small, transfer learning is helpful because it enables the use of information from a larger dataset or domain.

Each machine learning subset has unique abilities and tackles various issues. Depending on the unique properties of the data, the issue at hand, and the desired output, the right subset or combination of subsets should be chosen. The divisions between these subsets are becoming more flexible as machine learning develops, opening up new opportunities for study and applications in the area of artificial intelligence.

## 2.5 Arabidopsis thaliana

Thale cress or Arabidopsis thaliana is a small flowering plant that is a member of the Brassicaceae family. For a number of compelling reasons, it has become a popular model organism in plant biology research. The selection of Arabidopsis thaliana as the study plant in the context of this thesis report on biomass estimation using plant phenotype and machine learning models is well-justified, taking into account its numerous positive traits that support thorough and insightful research.
The exceptionally quick life cycle of Arabidopsis thaliana, which normally lasts six to eight weeks from seed germination to seed production, is one of the plant's major advantages. This quick life cycle makes it possible for researchers to run tests and see several generations develop in a short amount of time, which makes it easier to gather a wealth of data and quickens the speed of research. The research of various growth phases and responses to environmental changes is also made possible by the rapid return time, giving information on the dynamic features of plant growth and development. Furthermore, there are several wild accessions of Arabidopsis thaliana that can be studied, demonstrating a high degree of genetic variety. The identification of important genetic elements controlling biomass accumulation and growth patterns is made possible by the genetic variety that allows researchers to examine a wide range of phenotypic features and reactions to various environmental situations. The diversity also improves the research's robustness and generalizability because conclusions drawn from several accessions are more likely to be relevant generally. The modest size and simplicity of cultivation of Arabidopsis thaliana make it a useful plant for studies in greenhouses or other controlled environments. Its modest size makes data gathering and picture analysis much simpler, especially when using sophisticated phenotyping platforms like the ones discussed in this thesis report. The simplicity of culture also guarantees constant and repeatable experimental conditions, which minimizes potential confounding variables and improves the validity of research findings.
In conclusion, due to its short life cycle, small genome, genetic variety, genetic tractability, and ease of cultivation, Arabidopsis thaliana stands out as an essential model organism in this research project. Together, these beneficial characteristics offer researchers an effective set of tools that they can use to investigate the nuances of plant phenotype, understand the nuanced mechanisms governing biomass estimation, and investigate the exciting potential of machine learning models in the fields of agriculture and plant biology. This thesis report uses the study of the plant Arabidopsis thaliana to advance biomass estimation procedures and deepen our comprehension of the dynamics of plant growth in response to environmental factors.

## 2.6 Plant phenotype

The term "plant phenotype" describes the observable physical and physiological characteristics of a plant that are the outcome of how its genotype interacts with its environment. Plant height, leaf form, flower color, the architecture of the root system, and reactions to biotic and abiotic stressors are only a few examples of the qualities that these attributes cover. The complex interplay between a plant's DNA and the environmental conditions it experiences throughout its life cycle result in its phenotypic. It is essential in determining the ability of plants to respond to various environmental cues, tolerate fluctuations, and engage in ecosystem interactions. In order to improve agricultural productivity, resilience, and nutritional value, breeders must be able to choose and create plants with desirable qualities, which requires a thorough understanding of plant phenotypes. Studying plant phenotypes also aids in understanding how plants react to disease, climate change, and other ecological factors, providing insights into ecosystem dynamics and biodiversity conservation. A deeper knowledge of plant phenotypes has been made possible by improvements in high-throughput phenotyping tools and imaging methods, advancing plant biology and agriculture.

## 2.7 Image processing

Within the broader field of computer vision, image processing is a specialized and comprehensive science devoted to modifying, analyzing, and enhancing digital images in order to gain insightful knowledge, enhance visual quality, and facilitate the extraction of useful information. Image processing is essential to many applications across a wide range of sectors as the world grows more visually oriented. Images are represented as two-dimensional arrays of pixels in digital image processing, where each pixel carries information about the color or grayscale intensity of an image. Spatial domain processing and frequency domain processing are the two basic categories under which image processing techniques can be divided. Spatial domain processing entails carrying out operations directly on the image's pixel values. Image filtering, which blurs or sharpens the image, edge detection, which draws attention to the borders between various sections, and image enhancement, which seeks to enhance visual quality by modifying brightness, contrast, and color balance, are common procedures.

The rapid development of computing power, which has resulted in the creation of more intricate and sophisticated algorithms, is largely responsible for the continuing advancement of image processing. Machine learning's branch of deep learning, which enables cutting-edge approaches to image recognition, segmentation, and synthesis, has also made a significant contribution to image processing. The way we view and interact with the visual environment has been completely transformed by the important field of image processing. Our daily lives have been made better by its numerous uses across numerous industries, which have also aided in scientific developments, medicinal advances, and technology advances. The future will surely be shaped by image processing as it develops, pushing the frontiers of what is feasible in the field of computer vision and beyond.

### 2.7.1 Computer vision

The fields of computer vision and image processing are closely related and collaborate to provide machines with the ability to comprehend and interpret the visual environment. Image processing, which uses a variety of methods to enhance and preprocess digital images to make them more acceptable for further analysis, serves as the foundation for computer vision. To enhance image quality, reduce noise, and isolate regions of interest, techniques including filtering, denoising, and image segmentation are essential. Following the preprocessing of the images, computer vision algorithms assume control and use the data obtained from image processing to carry out complex tasks like item recognition, scene interpretation, and even human position estimates. Machines can identify their surroundings, recognize objects and patterns, and make intelligent decisions based on visual data thanks to the synergy between computer vision and image processing. The development of autonomous vehicles, medical diagnostics, surveillance systems, augmented reality applications, and other fields have all been made possible by this potent combination, which has changed many different industries. We can anticipate far more noteworthy advances as the fields evolve and advance, which will improve our comprehension of the visual environment and spur innovation in a variety of fields.

### 2.7.2 Threshold

### 2.7.3 contors

## 2.8 Important terms and concepts

### 2.8.1 Scikit-learn (sklearn)

Popular machine learning library Scikit-learn offers a full range of tools for various machine learning tasks, including as classification, regression, clustering, dimensionality reduction, and more. It is meant to be user-friendly and effective and is built on top of other scientific Python libraries like NumPy and SciPy. Both beginners and seasoned machine learning practitioners can use Scikit-learn because of its consistent API and variety of algorithms. It offers modules for model training, evaluation, hyperparameter adjustment, feature selection, and data preprocessing. Furthermore, Scikit-learn easily interfaces with other Python tools, allowing programmers to quickly create end-to-end machine learning pipelines.

### 2.8.2 Pandas

Python has a robust data analysis and manipulation module called Pandas. It provides simple data structures, principally the DataFrame, that enable users to efficiently work with labeled and structured data. Users can execute necessary data cleaning, transformation, and manipulation activities with Pandas and import data from a variety of file formats, including CSV, Excel, and SQL databases. It offers features for data wrangling and preparation, including filtering, sorting, grouping, aggregating, and merging. Additionally, Pandas is made to integrate easily with other NumPy and Matplotlib-compatible scientific Python tools, allowing users to swiftly complete challenging data analysis and visualization jobs.

### 2.8.3 R-squared value

A statistical metric used to assess the goodness of fit of a regression model is the R-squared value, sometimes referred to as the coefficient of determination. It measures the percentage of the variance in the dependent variable's output that can be accounted for by the model's independent variables' inputs. The R-squared value ranges from 0 to 1, where a value of 0 means that the model does not explain any variance in the dependent variable and a value of 1 means that the model perfectly fits the data and fully accounts for all variance. Higher R-squared values, in general, show that the model fits the data better and that the predictions of the model closely match the actual data points. To prevent overfitting and ensure the model's validity for making correct predictions on fresh, unforeseen data, R-squared must be interpreted in conjunction with other evaluation metrics and domain expertise.

### 2.8.4 Mean squared value (MSE)

The average squared difference between projected values and actual values in a regression or estimation problem is measured by the mean squared value (MSE), a commonly used statistical metric. It is a crucial performance indicator for assessing how accurate a prediction model is. The difference between each predicted value and its associated actual value is squared in order to calculate MSE, which is then calculated as the average of these squared disparities. The outcome value offers an indicator of how well the model's predictions reflect the actual values. A lower MSE suggests a more accurate and precise model because it shows that the model's predictions are more in line with the actual values. Overall, the MSE is a useful tool for evaluating the effectiveness of regression models, assisting in the choice of models, and directing efforts to improve models.

### 2.8.5 Hyperparameter tuning

Determining the ideal settings for hyperparameters, which are parameters set prior to the model's training that influence its performance and behavior, is a crucial step in the machine learning process of tuning hyperparameters. Hyperparameters, which are determined by the developer or data scientist before to training and have a substantial impact on the model's generalization and effectiveness, contrast with model parameters, which are learned from the training data (such as weights in a neural network). Learning rates, regularization strengths, the number of hidden layers in a neural network, and the number of trees in a random forest are examples of common hyperparameters. In order to determine the configuration that yields the highest performance on a validation set, hyperparameter tuning systematically examines various combinations of hyperparameter values. This often involves the use of methods like grid search, random search, and Bayesian optimization. As a result, more robust and trustworthy machine learning models may be produced. Appropriate hyperparameter tuning can greatly increase the model's accuracy and minimize problems like overfitting or underfitting.

### 2.8.6 Cross-validation

A reliable and popular method in machine learning for evaluating a model's performance and reducing the danger of overfitting is cross-validation. It entails partitioning the dataset into nu-

merous folds or subsets, referred known as "k-folds." The model is tested on the remaining fold after being trained on folds k-1. Each fold is used as the validation set exactly once during the course of the next k iterations of this operation. The average of the evaluation outcomes from each of the k folds serves as the final performance statistic. Since the model may be evaluated on other data points that it hasn't encountered during training, cross-validation yields a more accurate estimate of the model's generalization capabilities. As the model is assessed on various data partitions, decreasing reliance on a single train-test split, it also helps in spotting potential overfitting concerns. K-fold cross-validation, stratified k-fold cross-validation (used for imbalanced datasets), and leave-one-out cross-validation (used when the dataset is small) are three commonly used cross-validation techniques. Cross-validation is an essential phase in the development of a model since it directs hyperparameter tuning, influences model choice, and aids in the creation of more reliable and reliable machine learning models.

### 2.8.7 Overfitting and underfitting

The performance and generalizability of a model are both impacted by two prominent machine learning problems: overfitting and underfitting.

When a model becomes overly complicated, it learns to recognize noise or random oscillations in the training data rather than the underlying patterns, which is known as overfitting. As a result, although the model excels on the training data, it is unable to generalize to brand-new, untried data. Because the model has grown too dependent on the training data and is unable to manage variances in various datasets, overfitting can result in subpar performance on real-world tasks. Underfitting, on the other hand, occurs when a model is simple and fails to recognize the underlying patterns in the training data. Due to its inability to recognize the true relationships between the input features and the target variable, an underfit model performs badly on both the training data and the fresh data. Underfitting frequently means that the model is unable to adequately capture the complexity of the data. Techniques like regularization, early halting, and model complexity reduction can be utilized to alleviate overfitting. These techniques penalize excessively complicated models and stop them from remembering noise. More complicated models, feature engineering, or more data can allow the model to better catch significant trends in order to counteract underfitting.

Building reliable and accurate machine learning models that can perform well on new, untested data requires striking a balance between overfitting and underfitting. It is crucial to regularly evaluate the model using methods like cross-validation to make sure it generalizes well and can handle real-world scenarios successfully.

# Chapter 3

# Methodology

A systematic strategy is used in the proposed methodology for the thesis report on biomass estimation utilizing plant phenotypic and machine learning models to address the challenges with conducting accurate and non-destructive biomass assessment. A comprehensive and representative dataset of plant phenotypic data, including characteristics like leaf area, and leaf length, will first be gathered. To guarantee the data's quality, it will go through an extensive preprocessing process that takes into account things like image resolution inconsistencies, shadows, and imaging artifacts. The preprocessed data will next be used to extract relevant and discriminative features using the appropriate methods. These features will capture crucial details pertaining to biomass estimation. Then, four effective machine learning algorithms will be chosen for the task: Random Forest (RF), Support Vector Machine (SVM), Multilayer Perceptron (MLP), and k-Nearest Neighbors (k-NN). Cross-validation techniques will be used to train the models on the preprocessed data, ensuring generalizability and reducing the risk of overfitting. To gauge their precision and robustness, their performance will be examined using appropriate metrics like mean squared error and accuracy.
To comprehend the connections between the phenotypic characteristics and biomass estimation offered by the models, interpretability approaches will also be used. The best machine learning (ML) model(s) will be chosen through comparative analysis based on performance, interpretability, and applicability to biomass estimation. In the end, a script will be created to automate the estimation of biomass, allowing for non-destructive, scalable, and effective evaluation. By employing this methodology, the thesis seeks to advance biomass estimation and offer fresh perspectives, with possible applications in forestry, agriculture, and environmental studies.

## 3.1 Data Collection

Information was gathered for the data collection process from two different sources. First, utilizing the PlantScreen phenotyping technology created by Photon Systems Instruments (PSI), a collection of top-down, time-lapse visible spectrum images of Arabidopsis thaliana were taken. These images were captured at the National Plant Phenomics Centre in Aberystwyth, UK, at Plas Gogerddan. The dataset provides important details on the development and growth of the Arabidopsis thaliana plants under controlled conditions. [1]

Additional data-gathering efforts included the planting of about 350 Arabidopsis thaliana plants in 192 pots split into 16 trays with each tray hosting 12 pots in a glasshouse environ-

ment at the University of Cologne. The plants were carefully weighed, and measured during a five-week period. A variety of samples reflecting a wide range of shapes, sizes, and weights in Arabidopsis thaliana were obtained using this manual data-gathering approach. In order to create a comprehensive and well-rounded dataset for the purpose of training machine learning models for biomass estimation in Arabidopsis thaliana, data from the PlantScreen phenotyping platform and the glasshouse studies were combined.

### 3.1.1 Plant material

Arabidopsis thaliana was the plant utilized in the study at the National Plant Phenomics Centre. It was grown in PSI 6 cm square pots, each of which contained 180 ml of compost (Levington F2 + 205 grit sand). They used the PlantScreen phenotyping platform to promote the growth of the plants. The plants were kept in a glasshouse under a 10-hour day, 20°C/15°C regime, with gravimetric watering to keep the soil moisture at about 75% of its maximum capacity. These plants were seeded, and after ten days, they were properly replanted into the weighed pots. The plants were observed and harvested for data collection at predetermined intervals. Following the final harvest, which took place 56 days following seeding, the Arabidopsis thaliana plants had completed their whole growth cycle.
Additionally, before being planted in the glasshouse, the seeds were kept at a temperature of 4 degrees Celsius to maintain their viability and health. The homogeneous germination and subsequent growth of the plants were encouraged by the regulated storage environment, resulting in a stable and dependable dataset for biomass estimation. The investigation of various phenotypic traits was made possible by the use of Arabidopsis thaliana plants and the careful cultivation process, which increased the precision and efficiency of the proposed biomass estimation methodology using plant phenotype and machine learning models.

### 3.1.2 Image acquisition

The images from National Plant Phenomics Centre, it lasted for 35 days and ended with the final harvest. During the daylight hours, the plants were top-view photographed using the visible spectrum every 15 minutes (the actual interval was closer to 13 minutes). Typically, the first photograph was captured shortly after 9 am, and the final one was captured after 8 pm.
Images in the glasshouse were captured with A USB Webcam from Sandberg with a 1080p resolution. The webcam is made to record precise image quality (2MP). Its USB interface enables simple connectivity to a wide range of devices having USB ports, including laptops, desktop computers, and tablets. Because of this attribute, it is a practical option for video conferencing, online meetings, live streaming, and other applications that require video communication. The images are captured and stored in JPG/JPEG format

## 3.2 Preprocessing and Feature Extraction

The main objective of this study is to construct biomass estimation models using a simplified method that considers two essential traits which are, the number of leaves and the leaf area of the plant. With these two characteristics acting as the primary input elements, the goal is to forecast the plant's biomass, which serves as the output variable. A wide and representative collection

of plant images as well as measurements of the quantity, size, and fresh weight of the leaves are gathered in the first step of the procedure. The dataset must include a wide range of plant species, growth phases, and environmental circumstances to guarantee the models' capacity to generalize. The collected features, in particular, leaf area and count, will then be used as input variables in machine learning models. A number of models are taken into account, including Random Forest, Support Vector Machine, Multilayer Perceptron, and k-Nearest Neighbors. To enable the calculation of biomass from plant phenotypic data, it is intended to train each model using fresh weight (FW) as the desired output. Analysis of the models' performance and hyperparameter optimization are critical steps in the model construction process. Cross-validation techniques are used to assess the models' accuracy, robustness, and generalizability. To further evaluate the predictive effectiveness of the models, computations of assessment metrics like root mean square error (RMSE) will be made. These comprehensive evaluations will provide valuable insight into the performance of machine learning models for biomass estimation based on plant phenotypic data.

Precision, computational efficiency, and scalability are assessed as part of the evaluation of the machine learning models' efficiency. This study also intends to show the benefits and possibilities of combining machine learning with plant phenotypic features for biomass estimation. This thesis mainly focuses on developing models for machine learning using the parameters extracted from plant images, particularly the leaf count and area of leaf coverage. These models' training will focus on using fresh weight (FW) as the output variable. The primary objective is to create models for biomass estimation that are precise and easy to understand. This approach involves a number of steps, including data collection, preprocessing, feature extraction, model development, and thorough evaluation.

## 3.3 Model Development and Training

I developed predictive models based on four different machine-learning techniques: Random Forest (RF), Multilayer Perceptron (MLP), Support Vector Machine (SVM), and k-Nearest Neighbors (KNN) to comprehend the underlying relationship between image-derived parameters and the accumulated biomass (FW). These models used the observed FW as the response variable and the normalized phenotypic profile matrices for a representative range of phenotypic features as predictors (explanatory variables).

### 3.3.1 Random Forest

Random Forest (RF) regression is a machine learning algorithm used for biomass estimation based on the attributes of the number of leaves and leaf area extracted from plant images. RF regression is a powerful and versatile ensemble learning technique that combines multiple decision trees to create a robust predictive model. The RF algorithm builds several decision trees, each of which is trained on a different subset of the dataset. These decision trees are created using a random subset of the training samples and a random selection of characteristics. Based on the input attributes (number of leaves and leaf area), each tree independently predicts the output variable (in this case, fresh weight) during the training phase. By averaging or collecting the majority vote of all the forest's trees' estimates, the final prediction is obtained. The ability of RF regression to handle

intricate interactions between input characteristics and the output variable is one of its primary advantages. The algorithm's ability to capture non-linear and interaction effects is particularly helpful in situations where there may not be a linear relationship between the number of leaves, leaf area, and fresh weight. RF regression can reveal complex connections and patterns that may not be seen using conventional statistical techniques.

In this instance, the dataset containing plant images together with the related measures of the number of leaves, leaf area, and fresh weight is used to train the RF regression model. To develop a prediction model, the RF algorithm examines the link between the input attributes (number of leaves and leaf area) and the goal variable (fresh weight). Then, based on the number of leaves and leaf area in the new, unseen plant images, the trained RF model is used to estimate the fresh weight of the plants. The accuracy and dependability of the biomass estimation are assessed using a variety of measures, including mean squared error, mean absolute error, and R-squared value. In order to verify the model's performance over several subsets of the dataset and reduce overfitting, cross-validation techniques are frequently used. Overall, by utilizing the variables of leaf area and number of leaves retrieved from plant photos, the RF regression approach plays a significant role in this instance by estimating biomass. It is a useful method for biomass estimation in the context of this study since it can handle non-linear connections, handle high-dimensional data, and give interpretable feature importance analysis.

### 3.3.2 Support Vector Regression

Support Vector Regression (SVR) is a machine learning algorithm used for biomass estimation based on the attributes of the number of leaves and leaf area extracted from plant images. SVR is a variant of Support Vector Machines (SVM) specifically designed for regression tasks. In a high-dimensional feature space, SVR seeks to identify a hyperplane that best reflects the relationship between the input attributes (number of leaves and leaf area) and the output variable (fresh weight). SVR focuses on establishing a hyperplane that achieves a particular degree of precision, known as the margin, around the training samples, as opposed to conventional regression techniques that seek to reduce the errors between the predicted and actual values. When dealing with non-linear correlations between the attributes and the target variable, SVR is especially helpful. The SVR algorithm converts the input attributes into a higher-dimensional space, where the hyperplane is generated, using a kernel function. SVR is capable to detect complicated correlations and non-linear patterns between leaf area, leaf number, and fresh weight as a result of this change. The qualities of the data and the issue at hand determine whether kernel function, such as linear, polynomial, or radial basis function (RBF), should be used.

The SVR model is used to estimate the biomass of new plant images based on their number of leaves and leaf area once it has been trained on the dataset comprising plant images and their related biomass. The SVR method predicts the appropriate biomass by applying the learned hyperplane to the input attributes. In order to evaluate the accuracy and reliability of the biomass estimation, the performance of the SVR model is measured using metrics like mean squared error and R-squared value. SVR also has the benefit of illuminating the significance of the input features in establishing biomass. Each attribute's contribution to the biomass estimation process can be evaluated by looking at the coefficients or weights corresponding to it. The relationship between the number of leaves, the size of the leaves, and the fresh weight can be better understood

with the use of this information. In summary, Support Vector Regression (SVR) is an effective method used in this case to estimate biomass based on the attributes of the number of leaves and leaf area derived from plant photos. A useful tool for biomass estimation, it can handle high-dimensional data, capture non-linear correlations, and give interpretability via support vectors. SVR aids in revealing intricate patterns and connections between input qualities and fresh weight, enhancing comprehension and decision-making in plant growth, yield prediction, and agricultural management.

### 3.3.3 Multilayer Perceptron

A popular artificial neural network design called a multilayer perceptron (MLP) is used to estimate the biomass of plants based on the properties of leaf area and leaf number that are derived from plant images. MLP is a feedforward neural network model made up of many interconnected artificial neurons or nodes arranged in layers. The output variable (biomass) and the input qualities (number of leaves and leaf area) are recognized as complicated non-linear relationships, and MLP is known for its ability to capture these interactions. It is particularly useful in situations where the qualities' intricate connections with the target variables make it difficult to represent them using conventional linear regression methods. An input layer, one or more hidden layers, and an output layer make up the MLP architecture. Multiple neurons make up each layer, which executes calculations on the incoming data and relay the results to the following layer. Weights are connected to the connections between the neurons, and they are changed throughout training to improve the effectiveness of the network. By minimizing a predetermined loss function, such as mean squared error or mean absolute error, which measures the difference between the anticipated and actual values throughout the training phase, the MLP model learns which weights are optimal.

This model is used to estimate the biomass of new plants based on their number of leaves and leaf area once it has been trained on the dataset comprising plant images and their related fresh weight. The forward pass of the MLP model produces the final output prediction as the input attributes spread through the network, activating the neurons. The accuracy and predictive capability of the biomass estimation are evaluated using a variety of measures, such as mean squared error, mean absolute error, and R-squared value. In order to evaluate the model's performance over several subsets of the dataset and guarantee its generalizability, cross-validation techniques is also used. It has a number of benefits for problems involving biomass estimation. Large and varied datasets can be used to generate learning from complex non-linear relationships. Due to MLP's great degree of flexibility, its modeling ability can be improved by adding more hidden layers, neurons, or activation functions. Additionally, MLP can handle high-dimensional data, making it appropriate for situations where the biomass estimating process depends on a number of factors. In conclusion, Multilayer Perceptron (MLP) is an effective neural network architecture that is employed in this instance to estimate biomass based on the attributes of the number of leaves and leaf area derived from plant images. It is a useful tool for precise biomass estimation because of its capacity to model intricate non-linear interactions and learn from various datasets. MLP assists in analyzing plant growth, forecasting yields, and making agricultural decisions by enhancing our knowledge of the correlations between input variables and fresh weight.

### 3.3.4 k-Nearest Neighbors

k-Nearest Neighbors (k-NN) is a popular non-parametric machine learning algorithm used for biomass estimation based on the attributes of the number of leaves and leaf area extracted from plant images. k-NN is a simple yet effective algorithm that operates on the principle of similarity and makes predictions based on the characteristics of the k nearest neighbors in the training dataset. The training dataset, which comprises of plant images together with the associated biomass, is first stored by the k-NN method. When a fresh plant image is given during the prediction phase, the algorithm looks for the k nearest neighbors in the training dataset based on how similar their attributes like leaf area and number of leaves are. Typically, distance metrics like Euclidean distance or Manhattan distance are used to calculate how similar two plant images are to one another. The characteristics of the new plant image and the attributes of the k nearest neighbors are measured using the k-NN technique. The predicted biomass of the new plant image is determined by aggregating the biomass of the k nearest neighbors, typically through averaging or weighted averaging.

The k-NN algorithm's performance is influenced by the choice of the parameter k, which represents the number of neighbors to take into account. While a bigger number of k takes into account a wider range of neighbors, a smaller value of k tends to produce more focused predictions. simplicity and ease of use in implementation are the benefits of using this algorithm. It can adapt to various data types and makes no assumptions about the distribution of the underlying data. K-NN is also tolerant of outliers and can handle multi-modal data. The k-NN algorithm does, however, have a few drawbacks. The prediction time can drastically increase as the training dataset gets larger since finding the closest neighbors is necessary. The choice of distance measure and the scale of the input attributes may also have an impact on the way the k-NN algorithm performs. To address these problems, feature scaling is utilized. In conclusion, the k-Nearest Neighbors (k-NN) technique, which was employed in this instance to estimate biomass based on the attributes of the number of leaves and leaf area derived from plant images, is simple and efficient. K-NN offers predictions for the fresh weight of fresh plants by taking into consideration the traits of the closest neighbors in the training dataset. It is an effective tool for biomass estimation as to its simplicity, adaptability, and capability to handle multi-modal data.

## 3.4 Hyperparameter tuning

In order to maximize the performance of machine learning models, hyperparameter tuning is an essential phase in the modeling process. It entails looking for the optimal set of hyperparameters, or parameters that are chosen at the beginning of model training rather than ones that are learned from the data. GridSearchCV is a well-known method for hyperparameter tuning that cross-validates the model's performance while methodically examining a predetermined grid of hyperparameter values. In our case, we apply hyperparameter tuning using GridSearchCV to fine-tune the Random Forest (RF), Multilayer Perceptron (MLP), Support Vector Machine (SVM), and k-Nearest Neighbors (k-NN) models for biomass estimation based on the number of leaves and leaf area attributes extracted from plant images.

A collection of hyperparameters are specified for each model, and their associated values are to be investigated. Different combinations of values for each hyperparameter are specified to create the grid of hyperparameters. Following are the hyperparameters taken into account for each

model:

- **Random Forest**
  - n_estimators: the number of trees in the forest
  - max_depth: the maximum depth of each tree
  - min_samples_split: the minimum number of samples required to split an internal node
  - min_samples_leaf: the minimum number of samples required to be at a leaf node
- **Multilayer Perceptron**
  - hidden_layer_sizes: the number of neurons in each hidden layer
  - activation: the activation function for the hidden layers
  - solver: the optimization algorithm
  - learning_rate: the learning rate schedule
- **Support Vector Regression**
  - C: the regularization parameter
  - kernel: the kernel function used for mapping the input data to higher-dimensional space
  - gamma: the kernel coefficient (for non-linear kernels)
- **k-Nearest Neighbors**
  - n_neighbors: the number of neighbors to consider
  - weights: the weight function used in prediction
  - algorithm: the algorithm used to compute the nearest neighbors

Following the definition of the hyperparameters and their corresponding values, GridSearchCV conducts a thorough search by cross-validating the model while it is being trained on various combinations of the hyperparameters. Cross-validation reduces overfitting and aids in estimating the model's performance on hypothetical data. In order to determine the effectiveness of each set of hyperparameters, GridSearchCV applies a scoring metric, such as mean squared error or R-squared value. The optimal set is chosen as the set of hyperparameters that produces the best performance.

I intend to identify the hyperparameter configurations that provide the highest accuracy and best generalization capabilities for biomass estimation using the provided attributes by doing hyperparameter tweaking using GridSearchCV. We can increase the performance of the models by selecting the ones that work the best in our particular situation. After hyperparameter tuning, we assess the models with the ideal hyperparameter configurations using cross-validation or a different validation set to get a more accurate idea of how well they perform. This guarantees that the models are well-optimized and prepared for deployment for tasks involving biomass estimation.

In summary, the optimization of the Random Forest, Multilayer Perceptron, Support Vector Machine, and k-Nearest Neighbors models for biomass estimate requires hyperparameter tuning using

GridSearchCV. We can determine the optimal hyperparameter configurations that lead to accurate and generalizable models by methodically examining various combinations of hyperparameters and evaluating their performance. By using the specified variables retrieved from plant photos, this technique improves the prediction capabilities of the models and enables us to estimate biomass with greater accuracy.

bla bla bla

if you want a figure then this is the template ::


Figure 3.1: pie chart

## 3.5 second section

# Chapter 4

# Literature Review

write entire literature review here it doesnt consist of any sections

# Chapter 5

# Important Concepts

## 5.1 First section

## 5.2 second section

# Chapter 6

# Methodology

## 6.1 First section

## 6.2 second section

# Chapter 7

# Results and Discussion

explain your results here

## 7.1 sec 1

if you want further sections::

### 7.1.1 subsection 1

if you want bullet points :

- Hour of the day
- Day of the week
- Solar Radiation Level

if you want a table

| Model | Number of Lags | $R^2$ | RMSE | MAE |
|---|---|---|---|---|
| Random Forest Regressor | 24 lags | 0.993198 | 0.000455 | 0.000294 |
| Random Forest Regressor | 12 lags | 0.992970 | 0.000463 | 0.000291 |
| Random Forest Regressor | 6 lags | 0.992794 | 0.000469 | 0.000293 |
| Random Forest Regressor | 1 lag | 0.992776 | 0.000469 | 0.000299 |

Table 7.1: Model Training Results for Monthly Split of Data

# Chapter 8

# Conclusion

write summary of results along with future work etc

## 8.1 Summary of Results

## 8.2 Future Work

# List of Figures

# List of Tables

# Bibliography

[1] Bell, Jonathan, Dee, Hannah M. (2016). Aberystwyth Leaf Evaluation Dataset [Data set]. Zenodo. https://doi.org/10.5281/zenodo.168158