Optimizing Plant Growth

John Jarvis

Old Dominion University

jjarv001@odu.edu

https://github.com/jjarv001/CS422termProject

Abstract

Millions of people across the planet perform gardening tasks in a variety of climates and conditions. Varying conditions can play critical roles in the success or failure of a plant. The average home gardener won't have the ability to control or mitigate potential conditions. In this study I seek to identify some of the critical features necessary for plant success and try to determine an optimal set of growing conditions. Using a supervised machine learning approach to identify important features, I seek to build a model that can predict whether a plant will succeed or fail given a set of input conditions.

1 Introduction

In this project, I test how different machine learning models can be utilized to determine features that are important to plant success. I also attempt to develop a model that can predict whether a plant will succeed or fail given a certain set of conditions. The conditions I examine are soil type, amount of sunlight, watering frequency, fertilizer type, temperature, and humidity. Some of these variables are categorical data and some are quantitative.

2 Data Description

As mentioned in Section 1, I am analyzing the effect of different independent variables on plant success/failure (the target variable). Ideal data regarding plant success/failure was not easy to find, and ideal data might be difficult to collect (especially in uncontrolled environments). The data that was settled on consists of 194 samples, and a subset of the data is shown below in Figure 1. The data that is used in this project is certainly less than ideal, so in this section we'll look at the different variables and their potential pitfalls.

plant_growth_data

Soil_Type	Sunlight_Hours	Water_Frequency	Fertilizer_Type	Temperature	Humidity	Growth_Milestone
loam	5.192294089205040	bi-weekly	chemical	31.719602410244100	61.59186060849000	0
sandy	4.033132702741610	weekly	organic	28.91948412187400	52.42227609891600	1
loam	8.892768570729000	bi-weekly	none	23.179058888285400	44.66053858490320	0

Figure 1: Dataset Visual

2.1 Categorical Variables: Soil Type, Watering Frequency, Fertilizer Type

The first categorical variable is Soil Type, which consists of three possibilities: loam, sandy, and clay. Soil Type can play a major role in plant development. The type of soil can affect the amount and type of nutrients present in the soil. It can also influence root development and water retention. These sub-variables would likely to be difficult to track, but it's important to keep these things in mind.

The second categorical variable is Watering Frequency, which also consists of three possibilities: daily, weekly, and bi-weekly (from the data it appears that bi-weekly is meant in the terms of "every other week" though it's not stated outright). Some additional data that might have been useful would include things such as volume, time of day, and watering system (sprinkler, drip, soaker hose, etc.).

The final categorical variable is fertilizer type with its three possibilities: chemical, organic, and none. No specific fertilizer brand is listed for the chemical or organic fertilizer, nor is the amount and deliver method mentioned. These sub-variables could certainly play a role in plant health.

All these variables were relatively evenly distributed amongst the samples, as shown in Figure 2 below.

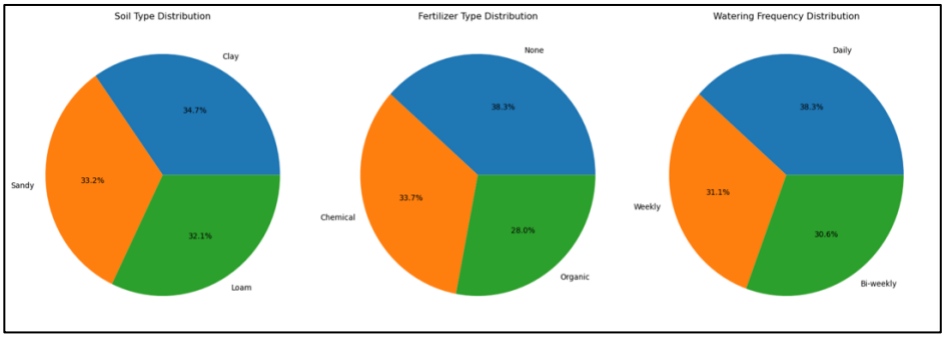

Figure 2: Pie Charts of Categorical Variables

2.2 Quantitative Variables: Sunlight Hours, Temperature, Humidity

The quantitative variables need less explanation, as it can be surmised that these are averages over the course of the plant sample's life cycle. It would likely be useful to have these measurements performed on a daily, or even hourly basis. This could probably be achieved through a system of sensors in a weather resistant environment, such as a greenhouse, but the initial setup would likely be quite expensive. It also would've been useful to know what the hardiness zone the samples were collected in, as this can have an impact on all three of these variables

The distributions of each of these variables were relatively normal across the samples set, with humidity being a bit bimodal. This can be viewed in Figure 3 below.

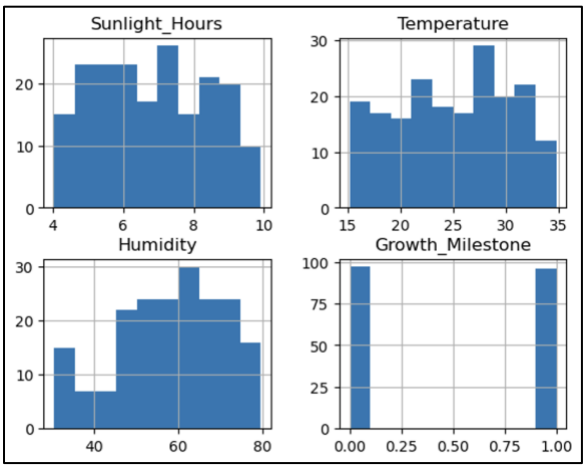

Figure 3: Quantitative Data Distribution

Additionally, there were no significant outliers, as shown in Figure 4 below.

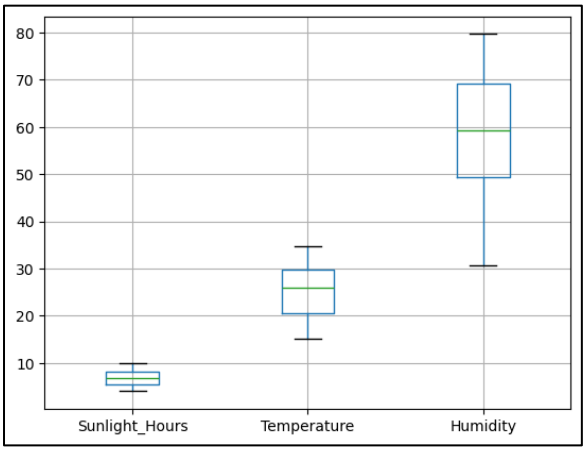

Figure 4: Quantitative Data Boxplot

3 Machine Learning Methods and Tools

This was a supervised learning classification problem, as the target variable was known and the dataset was clearly labeled.

Being that this was a classification problem and that the dataset was relatively small, multiple machine learning models were analyzed to determine which was the most accurate in predicting whether a plant would succeed or fail.

The data was split into training and testing sets using scikit-learns train_test_split method with an 80/20 ratio, respectively. The categorical data was encoded using one-hot encoding to work appropriately with the models.

Finally, the models used were all from scikit-learn and included: Random Forest Classifier, Logistic Regression Classifier, Gaussian Naïve Bayes, Decision Tree Classifier, Support Vector Classifier, Gradient Boosting Classifier, and KNN Classifier.

4 Implementation of Experiment

This experiment was carried out using the Python programming language. All of the code for this experiment was carried out in a Jupyter Notebook, using Visual Studio Code as the primary code editor. The code is publicly accessible at the following link:

https://github.com/jjarv001/CS422termProject/blob/main/termProjFinal.ipynb

The basic structure of the program included the following steps:

1. Data visualization (as shown in the figures above, with more available at the link listed above)
2. Statistical analysis on data (ANOVA , Chi square, interactions)
3. Preprocessing data (encoding and transformations)
4. Analysis of models and results
5. Summarizing observations

5 Results, Discussion, and Conclusion

In this section we'll look at the results obtained from this experiment, both good and bad. We'll also discuss the importance/implications of these results and what could've been better.

5.1 Results

From experimenting with different models, the best models were concluded to be the Decision Tree Classifier, Random Forest Classifier, and KNN Classifier (though this is subject to slight change depending on the run). None of these models were able to reliably predict whether a plant would succeed or fail given a set of input variables. The most accurate was the Decision Tree Classifier at an accuracy of 67%, with the remaining two models close behind. The comparison of the top three models can be seen in Figure 5 below.

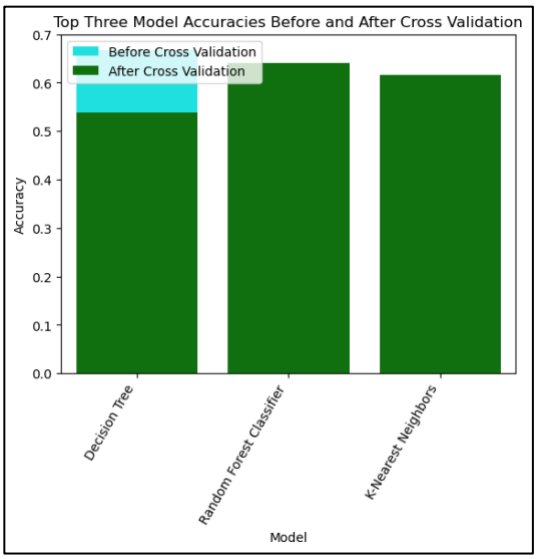

Figure 5: Comparison of Model Accuracies

While none of the models were particularly accurate with regards to prediction, there were some potential statistical takeaways. Prior to modeling, statistical analysis was performed on all independent variables and their interactions. Not all of this will displayed here for brevity sake, but all data is available at the previously mentioned link in Section 4.

Sunlight hours and humidity appear to have some amount of influence on plant success/failure, with a moderate humidity appearing to have a positive impact and overexposure to sunlight having a negative impact. The results of analysis of variance (ANOVA) testing appear to show these two variables borderline on being statistically significant, as highlighted in Figure 6 below.

	Sunlight_Hours	Temperature	Humidity
F-statistic	2.895978	0.549833	3.300261
p-value	0.090428	0.459297	0.070836

Figure 6: ANOVA of Quantitative Variables

Concerning categorical variables, the fertilizer type (though not statistically significant with this data set) appeared to show some level of influence on plant success/failure. Figures 7 and 8 display Chi square results on these variables as well as a visualization of fertilizer type vs. success below, respectively.

	Soil_Type	Water_Frequency	Fertilizer_Type
Chi-squared	0.130766	0.005155	1.226920
p-value	0.717640	0.942764	0.268007

Figure 7: Chi Square of Categorical Variables

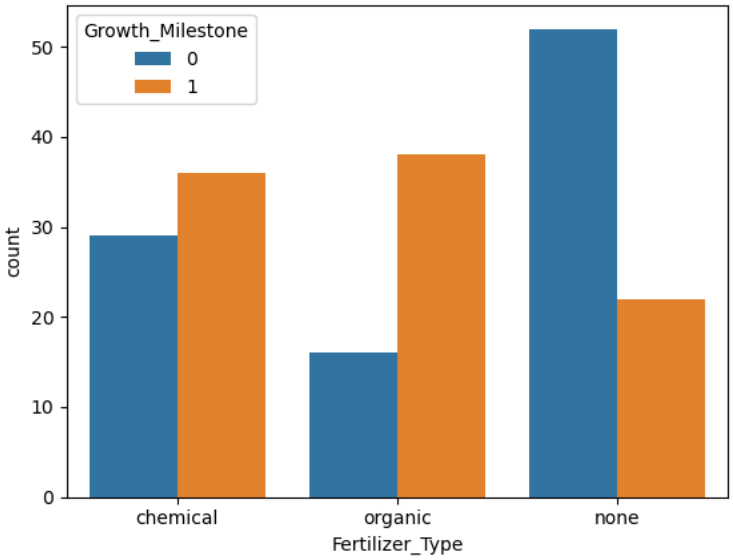

Figure 8: Fertilizer Type vs. Success (Orange)

5.2 Discussion

While the models observed and tested weren't particularly good for determining whether a plant would succeed or fail given a set of conditions, much of the inaccuracy was likely to do with the amount and quality of the data. The data utilized would likely have been far more useful had it been taken over a time series rather than an average. Additionally, its unclear what truly defined a success vs. failure.

That being said, this type of data is not easy to collect on a large scale, and there are many additional factors that play into growth in all forms of life.

5.2 Conclusion

I think it's reasonable to think that this project would be worth attempting again with a larger and more descriptive dataset. Plants play a vital role in not just human health, but the planet's health. Learning the factors that play important roles in plant health and how to optimize those factors can have many benefits, and I think this experiment shed some amount of light on some of those factors. But there's certainly much more that could be done.

There are many and much better projects going on around the world utilizing machine learning to optimize agricultural functions. Unfortunately, much of the data is not publicly available as the research being conducted is often for-profit and/or proprietary.

While this project wasn't a complete success, I wouldn't view it as a failure either. Some interesting interactions and their effects on plant success were observed, and I think this experiment also highlighted the importance of high-quality data.

I hope you enjoyed reading this and can take something useful/interesting away from it.

6 References/Acknowledgements

Dataset publicly available at: https://www.kaggle.com/datasets/gorororororo23/plant-growth-data-classification/code

[1] *An Introduction to Machine Learning* by Andreas C. Muller and Sarah Guido (O'Reilly). Copyright 2017 Sarah Guido and Andreas Muller, 978-1-449-36941-5.

[2] *Python for Data Analysis* by Wes McKinney (O'Reilly). Copyright 2022 Wes McKinney, 978-1-098-10403-0.