Modeling Plant Health
Alyssa Yang (awyang)
December 18, 2024

1 Introduction

Understanding plant health is crucial for optimizing agricultural practices, improving crop yields, and promoting sustainable growth. Monitoring and predicting plant health, whether in large-scale agriculture or home gardening, can enhance care, growth outcomes, and sustainability.

Recent advancements in agricultural technology, particularly smart farming sensors, have transformed plant health management by enabling real-time environmental monitoring and precision agriculture [2]. These advancements stem from research on factors influencing plant health such as the vital role of balanced macronutrients in supporting essential plant functions [3].

However, previous research has often examined isolated factors, overlooking the variability from joint interactions. Thus, this project seeks to bridge this gap by leveraging Bayesian modeling techniques to predict plant health while accounting for these complex interactions with the aim to deepen our understanding of plant nutrition and contribute to future advancements in agricultural optimization.

2 Dataset

The Plant Health Monitoring dataset [1] includes data from 1,000 plants, capturing environmental and biological factors that influence plant health. It includes measurements for temperature (°C), humidity (%), soil moisture (%), soil pH, nutrient level, light intensity (lux), health score (0-100), and health status. The health score reflects the plant's overall health based on these factors, and health status is a binary variable (0 = Unhealthy, 1 = Healthy) based on the health score. Table 1 below presents the data for the first plant.

Plant ID	Temperature	Humidity	Soil Moisture	Soil pH	Nutrient Level	Light Intensity	Health Score	Health Status
Plant_1	26.4901	73.9936	34.8723	5.5461	41.3651	18728.7210	68.8592	0

Table 1: Data for Plant 1

All numeric variables (variables excluding plant ID and health status) exhibit a near-normal distribution with minimal to no skew and are centered around their respective means. Summary statistics and histograms for each variable are shown in Figures 1 and 2. Additionally, the dataset reveals an imbalance in health status with 826 plants classified as healthy and 174 as unhealthy as visualized in Figure 3.

To analyze trends between healthy and unhealthy plants, I grouped the data based on health status. The numeric variables, with the exception of health score (directly influenced by health status), appear to be normally distributed across both groups, exhibiting the same distribution as before. Additionally, the mean and median values for most variables are very similar between healthy and

temp	humidity	soil_moisture	soil_ph	nutrient_level	light_intensity	health_score
Min. :15.28	Min. :30.60	Min. : -0.2927	Min. :5.035	Min. :18.23	Min. :11301	Min. : 52.87
1st Qu.:23.06	1st Qu.:53.94	1st Qu.: 35.2800	1st Qu.:6.131	1st Qu.:43.17	1st Qu.:17919	1st Qu.: 72.45
Median :25.08	Median :60.63	Median : 44.9962	Median :6.500	Median :49.82	Median :19872	Median : 79.45
Mean :25.06	Mean :60.71	Mean : 45.0875	Mean :6.491	Mean :49.51	Mean :19860	Mean : 79.72
3rd Qu.:26.94	3rd Qu.:67.29	3rd Qu.: 54.9137	3rd Qu.:6.833	3rd Qu.:56.39	3rd Qu.:21837	3rd Qu.: 87.00
Max. :36.56	Max. :91.93	Max. :103.8936	Max. :8.122	Max. :81.13	Max. :29295	Max. :115.29

Figure 1: Summaries of numeric variables

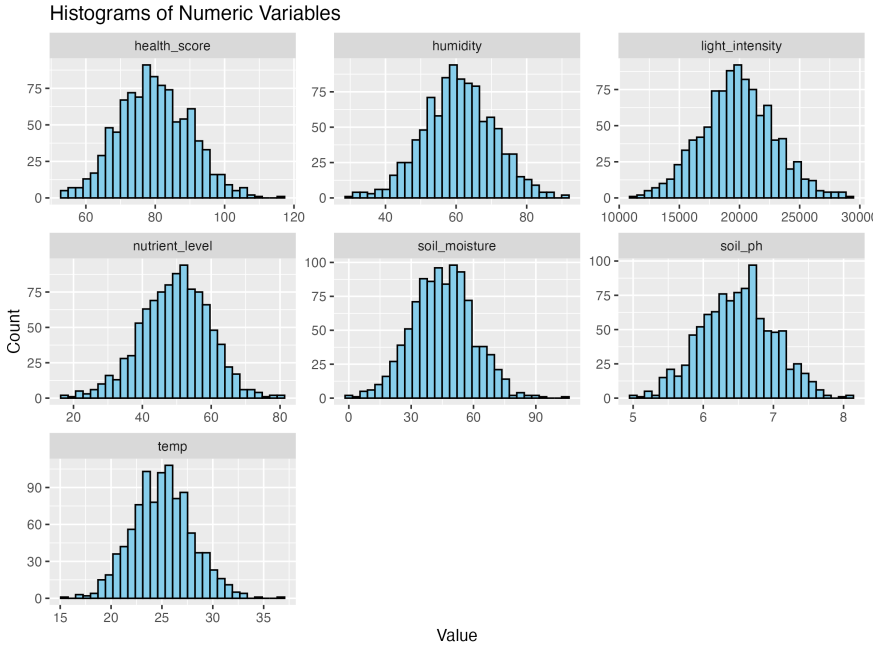

Figure 2: Histograms of numeric variables

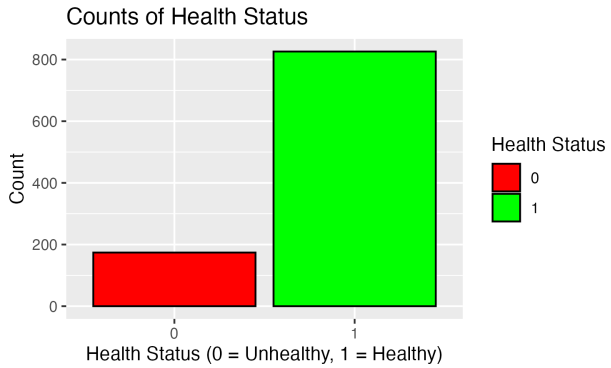

Figure 3: Counts of health status

unhealthy plants, with the only notable differences being slightly lower soil moisture and slightly higher light intensity in healthy plants compared to unhealthy ones. This suggests that while these predictors may have some role in plant health, the differences across them are relatively minor, indicating that other factors may play a more significant role. These trends are illustrated in Figures 4 and 5.

```
$`0`
      temp          humidity     soil_moisture       soil_ph      nutrient_level  light_intensity  health_score
 Min.   :18.80   Min.   :30.79   Min.   : 5.937   Min.   :5.035   Min.   :23.75   Min.   :12694   Min.   :52.87
 1st Qu.:23.33   1st Qu.:53.46   1st Qu.:37.318   1st Qu.:6.123   1st Qu.:43.52   1st Qu.:17863   1st Qu.:62.72
 Median :25.29   Median :60.85   Median :47.214   Median :6.528   Median :49.97   Median :19852   Median :65.95
 Mean   :25.43   Mean   :60.25   Mean   :46.301   Mean   :6.480   Mean   :49.33   Mean   :19792   Mean   :64.90
 3rd Qu.:27.30   3rd Qu.:66.35   3rd Qu.:56.048   3rd Qu.:6.801   3rd Qu.:55.75   3rd Qu.:21504   3rd Qu.:68.01
 Max.   :32.90   Max.   :85.80   Max.   :75.184   Max.   :7.592   Max.   :69.45   Max.   :27618   Max.   :69.99

$`1`
      temp          humidity     soil_moisture         soil_ph      nutrient_level  light_intensity  health_score
 Min.   :15.28   Min.   :30.60   Min.   : -0.2927   Min.   :5.047   Min.   :18.23   Min.   :11301   Min.   : 70.02
 1st Qu.:22.97   1st Qu.:54.00   1st Qu.: 35.0050   1st Qu.:6.133   1st Qu.:43.15   1st Qu.:17943   1st Qu.: 76.42
 Median :24.96   Median :60.48   Median : 44.6637   Median :6.498   Median :49.77   Median :19874   Median : 81.43
 Mean   :24.98   Mean   :60.80   Mean   : 44.8319   Mean   :6.493   Mean   :49.55   Mean   :19874   Mean   : 82.84
 3rd Qu.:26.90   3rd Qu.:67.36   3rd Qu.: 54.3840   3rd Qu.:6.839   3rd Qu.:56.55   3rd Qu.:21905   3rd Qu.: 88.72
 Max.   :36.56   Max.   :91.93   Max.   :103.8936   Max.   :8.122   Max.   :81.13   Max.   :29295   Max.   :115.29
```

Figure 4: Summaries of numeric variables across health status

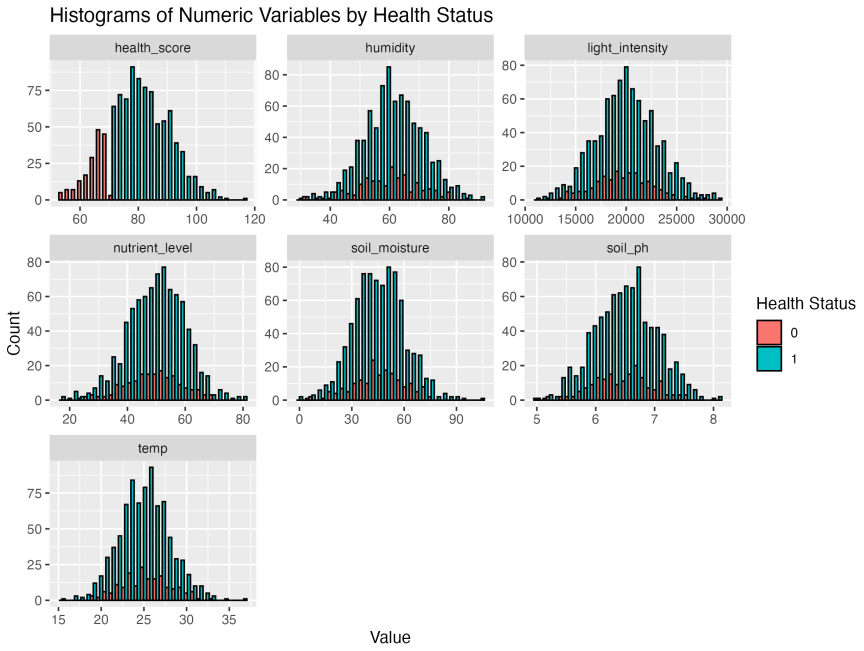

Figure 5: Histograms of numeric variables across health status

3 Method

To analyze general plant health trends, I developed a Beta-Binomial conjugate model to predict the probability p that a plant is healthy (health status = 1), given the health statuses of all plants in the dataset, $y_i \in \{0, 1\}$. I used an uninformative prior, $p \sim \text{Beta}(\alpha = 1, \beta = 1)$, as I did not have any prior beliefs about p, and a Binomial likelihood, $y|p \sim \text{Binomial}(n = 1000, p = p)$, as health status is binary. Thus, the conjugacy of these distributions yields the posterior distribution:

$$p|y \sim \text{Beta}(1 + \sum_{i=1}^{1000} y_i, 1001 - \sum_{i=1}^{1000} y_i)$$

I performed Markov Chain Monte Carlo (MCMC) to assess the model fit, using convergence diagnostics such as trace and autocorrelation plots and evaluating effective sample size. A posterior

predictive check was done by extracting posterior samples to compare the predicted number of healthy plants with the observed data.

However, the Beta-Binomial model provides only aggregate-level predictions and assumes a constant probability across all observations, failing to account for variability due to plant characteristics or external factors. To address this, I created a hierarchical Bayesian logistic regression model to incorporate group-level variation and additional predictors.

Plants were clustered into four groups using k-means based on environmental and soil attributes (temperature, humidity, soil moisture, pH, nutrient level, and light intensity) with each scaled to ensure equal contribution. I then performed PCA using 2 principal components to visualize the clusters, shown in Figure 6. In the STAN model, I modeled each predictor's coefficient as $\beta \sim \text{Normal}(0, 5)$ and group-specific intercepts as $\alpha_j \sim \text{Normal}(0, \sigma_\alpha^2)$. The standard deviation of cluster intercepts, $\sigma \sim \text{Cauchy}(0, 2)$, quantifies the variability between clusters. The likelihood for each plant i in cluster j was modeled as: $y_i \sim \text{Bernoulli}\left(\text{logit}^{-1}(\alpha_j + X_i \cdot \beta)\right)$

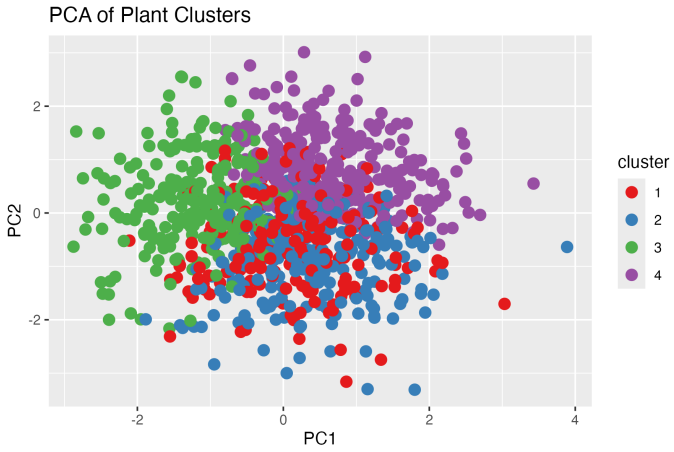

Figure 6: PCA of plant clusters

4 Results

Diagnostic checks for the MCMC model confirmed convergence, with trace plots showing overlapping chains, an effective sample size of 1,473, and autocorrelation quickly dropping to zero. The posterior predictive mean number of healthy plants was 825.248 (95% CI: 791-857), closely matching the observed value of 826. The simulated values are closely and normally distributed around the true number as shown in Figure 7, suggesting that the Beta-Binomial model fits the data well.

Diagnostic checks for the hierarchical model also confirmed convergence with results identical to the MCMC model but with effective sample sizes of around 2,000 for each variable. The posterior predictive mean number of healthy plants was 823.207 (95% CI: 789-855), and this is reflected in Figure 8 where the simulated values are normally distributed but shifted slightly lower. Thus, since these predictions deviate more from the true number, the hierarchical model performed slightly worse than the Beta-Binomial model.

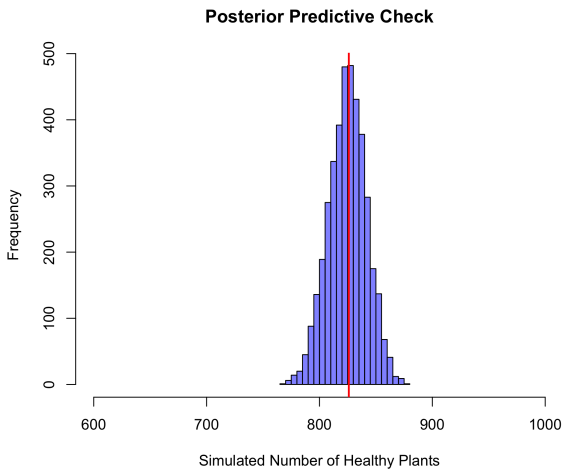

Figure 7: MCMC posterior samples

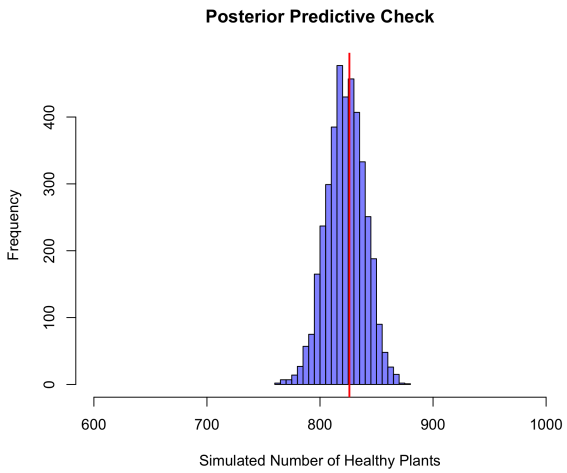

Figure 8: Hierarchical posterior samples

The PCA plot reveals overlapping clusters, indicating weak group separation. This suggests that plant health is not strongly influenced by shared environmental or soil conditions, and this is also supported by the posterior predictive check of the hierarchical model. These results imply that variability in plant health is likely driven more by individual predictors rather than group-level effects and that a simpler model, like the Beta-Binomial model, is able to capture variability effectively without the added complexity of hierarchical structure. We conclude that accounting for cluster-level variability is unnecessary for modeling plant health.

5 Conclusion

Through the analysis of this plant health monitoring dataset, this project suggests that plant health can be modeled at a population level without explicitly accounting for differences between plants under similar conditions. The Beta-Binomial model demonstrated strong performance due to its simplicity whereas the hierarchical model performed slightly worse, reflecting a lack of significant cluster effects.

However, it is important to note that the dataset used included a limited set of observations and predictors, and the hierarchical model clustered data using assumptions about group effects not directly observed. Clustering based on these environmental attributes may not capture meaningful patterns, and missing predictors may have a more significant role in predicting plant health. Thus, future work could involve validating these model assumptions, testing simpler approaches such as logistic regression, exploring alternative clustering methods, and incorporating more missing variables and data to improve predictions.

Without significant group effects, plant health management should focus on optimizing individual conditions rather than focusing on shared environmental or latent factors. Simpler models can effectively capture plant health variability, but further exploration of additional predictors and clustering methods is needed to enhance predictive power and application.

References

[1] Ziya, Plant Health Monitoring Dataset
https://www.kaggle.com/datasets/ziya07/plant-health-monitoring

[2] Sciforce, Smart Farming
https://www.iotforall.com/smart-farming-future-of-agriculture

[3] AGQLabs, Macronutrients in Plants
https://www.agqlabs.us.com/macronutrients-in-plants/

Link to the Github Repository: https://github.com/alyssawyang/stats551-final-project