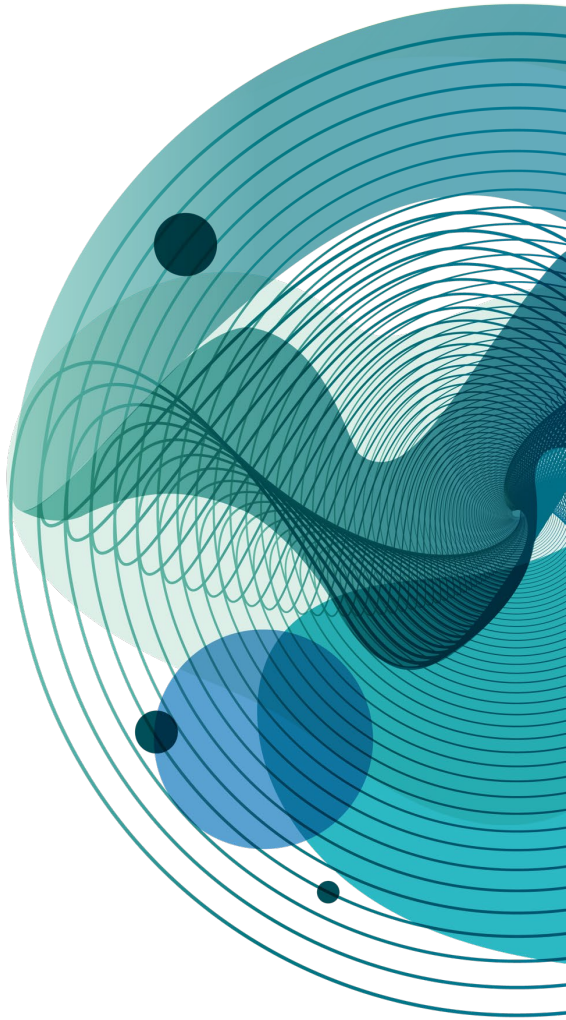
Deloitte.

Custom Crop Guidance with AI

Creating AI Models to Enhance Crop Selection

Agenda



Our Team

Alex Wroble
CEO

Joe Matthews
Analyst

Nicole Salas
Analyst

Drew Clutterbuck
Analyst

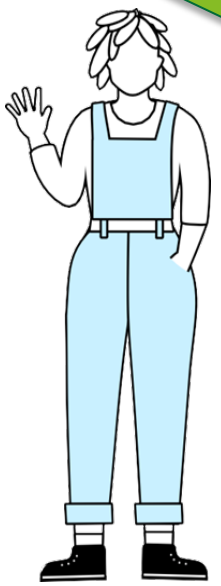
Dorothy Akpovwa
Analyst

Overview

Business Problem

A brief overview of the business problem and why it is important to solve it.

We aim to maximize our crop yields by understanding each crop's specific needs and nutrient requirements.



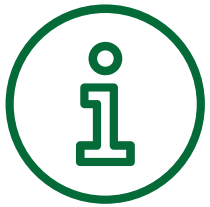
Farmer (Client)

- A lack of understanding about environmental conditions and nutrients can be detrimental to farmers' success.
- Different crops require different nutrients
- Poor crop selection can lead to negative losses for farmers
- Crops not suitable for local conditions can harm crop growth
- Incorrect Soil pH can prevent crops from absorbing nutrients

*Addressing the question, **"What crops are most suitable to grow given the soil and environmental conditions?"** can lead to better resource management, better crop health, amongst many other areas of agriculture that are important to farmers.*

Data Understanding

The dataset utilized gave information on main factors needed for crop health, and below details important information the team needed to understand before moving on to model creation.



Data Overview

- Entirely composed of numerical data
- All 2,200 rows of data are complete with no null values

Key Factors

- Nitrogen, Phosphorus and Potassium
- The dataset includes the ratio of each nutrient present in the soil

Environmental Variables

- Temperature – in degrees Celsius
- Humidity – relative in %
- pH_Value – numerical value of the soil
- Rainfall – in millimeters

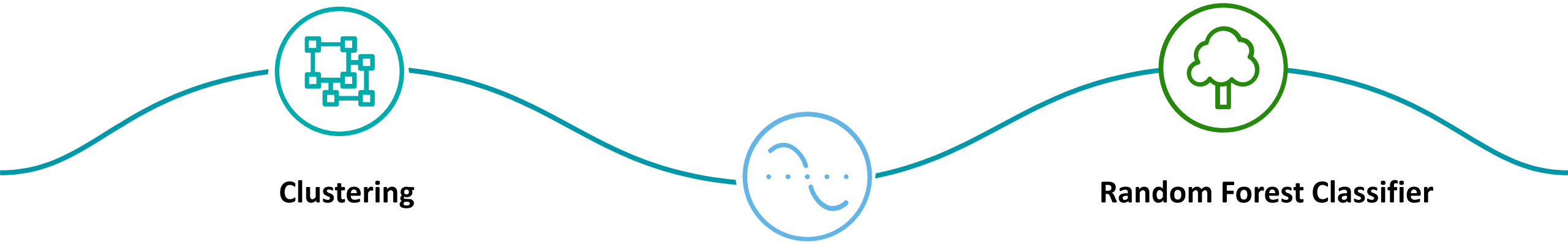
Target Variable

- “Crop”, representing the type of crop
- 22 unique crop types included

Methodology

Methodology

The methodology used for the duration of the project consisted of three main sections.



Clustering

- Models without clustering had very poor results when trying to predict certain classes
- By providing multiple crop recommendations based on environmental factors, the client can utilize economic factors to make a final decision
- An AI model was used to group together similar crops into clusters

Logistic Regression

- Machine learning technique that directly addresses the business classification problem.
- Allows for various combinations of hyperparameters to be tested and for k-fold cross validation.
- Enables for the final model to be tested on a separate unused test set.

Random Forest Classifier

- Machine learning technique that evaluates feature importance in selecting a crop.
- Allows for higher accuracy as it is less prone to overfitting.
- Flexible and easy to interpret to ensure accurate and specialized results

Final Clusters

- Utilized K-Means clustering to find most similar observations
- Allocated each crop to the cluster where most of its observations reside (at least 50%)
- Analyzed options with different number of clusters using elbow and silhouette plots
- Crops in the same cluster can be substituted for each other when choosing a crop to plant
- Used the optimized clustering (shown right) to create new target labels

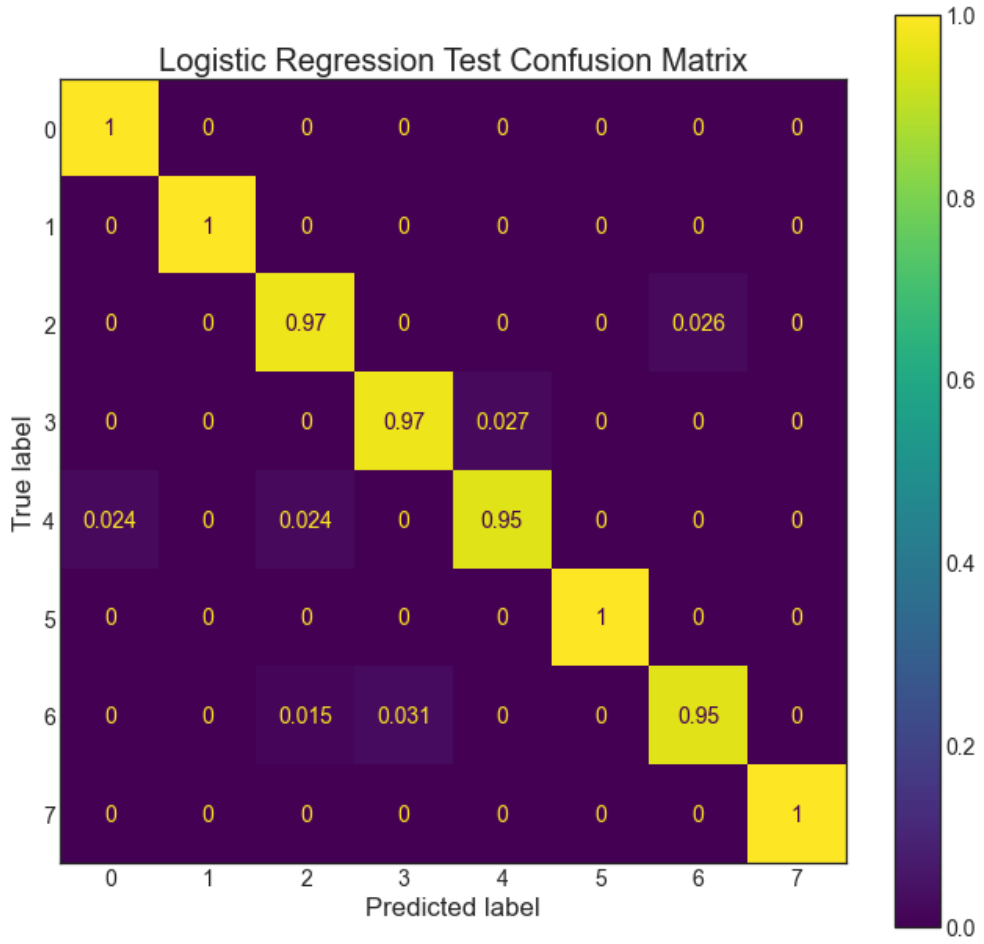
*See Appendix B

Findings

Logistic Regression

A machine learning model that calculates the probability that a given observation belongs to each class (crop cluster) and assigns the observation to the class with the highest probability.

Confusion Matrix



Key Takeaways

98%
Accuracy on Dedicated Testing Data

100%
Accuracy on 4 Clusters

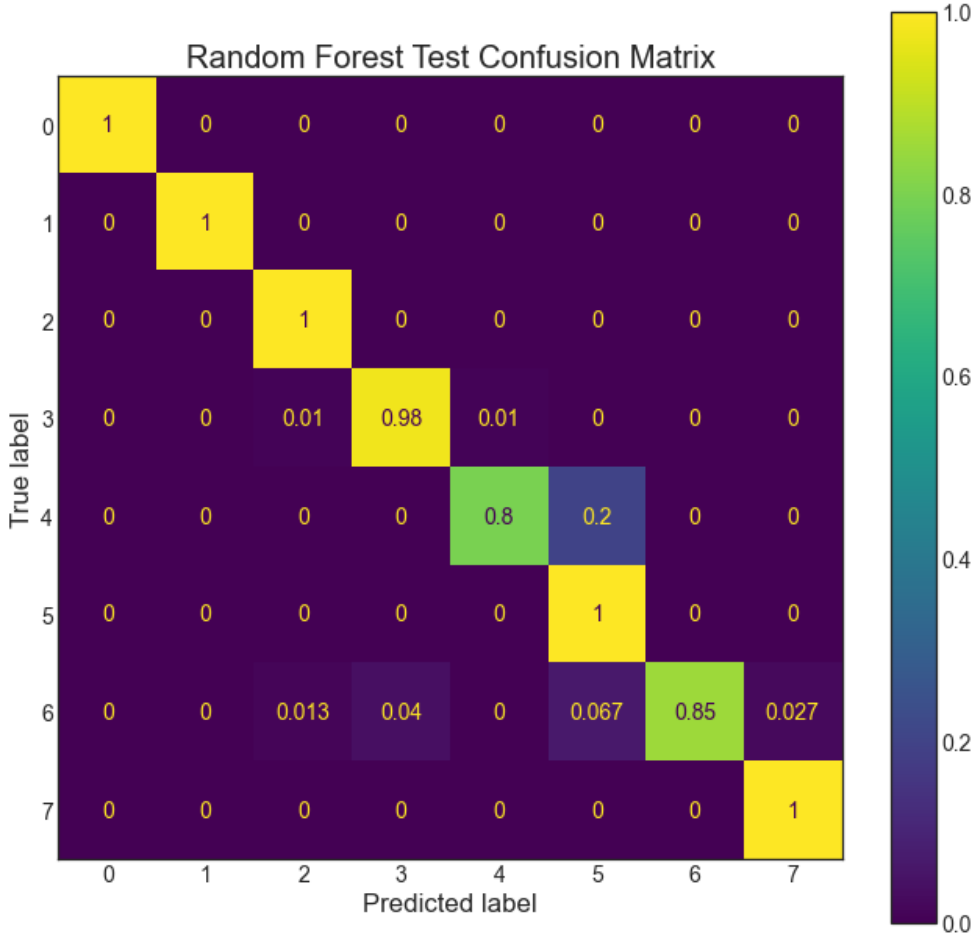
12
Different Model Variations

*See Appendix C

Random Forest Classifier

A machine learning model that combines the predictions of multiple decision trees to improve accuracy and reduce the risk of overfitting

Confusion Matrix



Key Takeaways

96%

Accuracy on Dedicated Testing Data

7x

More Accurate than Baseline Model

62%

More Accurate than other AI Models

*See Appendix D

Conclusion

Logistic Regression with 8 Clusters

We recommend the logistic regression over the random forest classifier.

- ***Accuracy***

  The logistic regression model is advised due to its higher accuracy (98%) on the testing data and its robustness in classifying clusters

- ***Clustering***

  Allowed improved accuracy in the model while giving the farmer a variety of options to choose from based on their external knowledge

- ***Computational Efficiency***

  Logistic regression models are quicker to train and make predictions than random forest classifiers

Next Steps

Next Steps

Below are potential next steps the Deloitte team believes can help scale the current project to continue to expand upon findings and help solve the business problem.

01

Enhance the dataset with **real-time data** on weather, soil and environmental triggers by utilizing Application Programming Interface (APIs) and Internet of Things (IoT) devices.

02

Enhance the dataset with **economic factors** such as production costs (seeds, fertilizers, pesticides), market demand, and price

03

Incorporate additional datasets to grow the project to perform **deep learning models** for more complex pattern recognition.

04

Collaborate with **additional research organizations** such as the International Food Policy Research Institute (IFPRI) and Food and Agriculture Organization (FAO) to further scale the project by leveraging their data and best practices.

Thank You

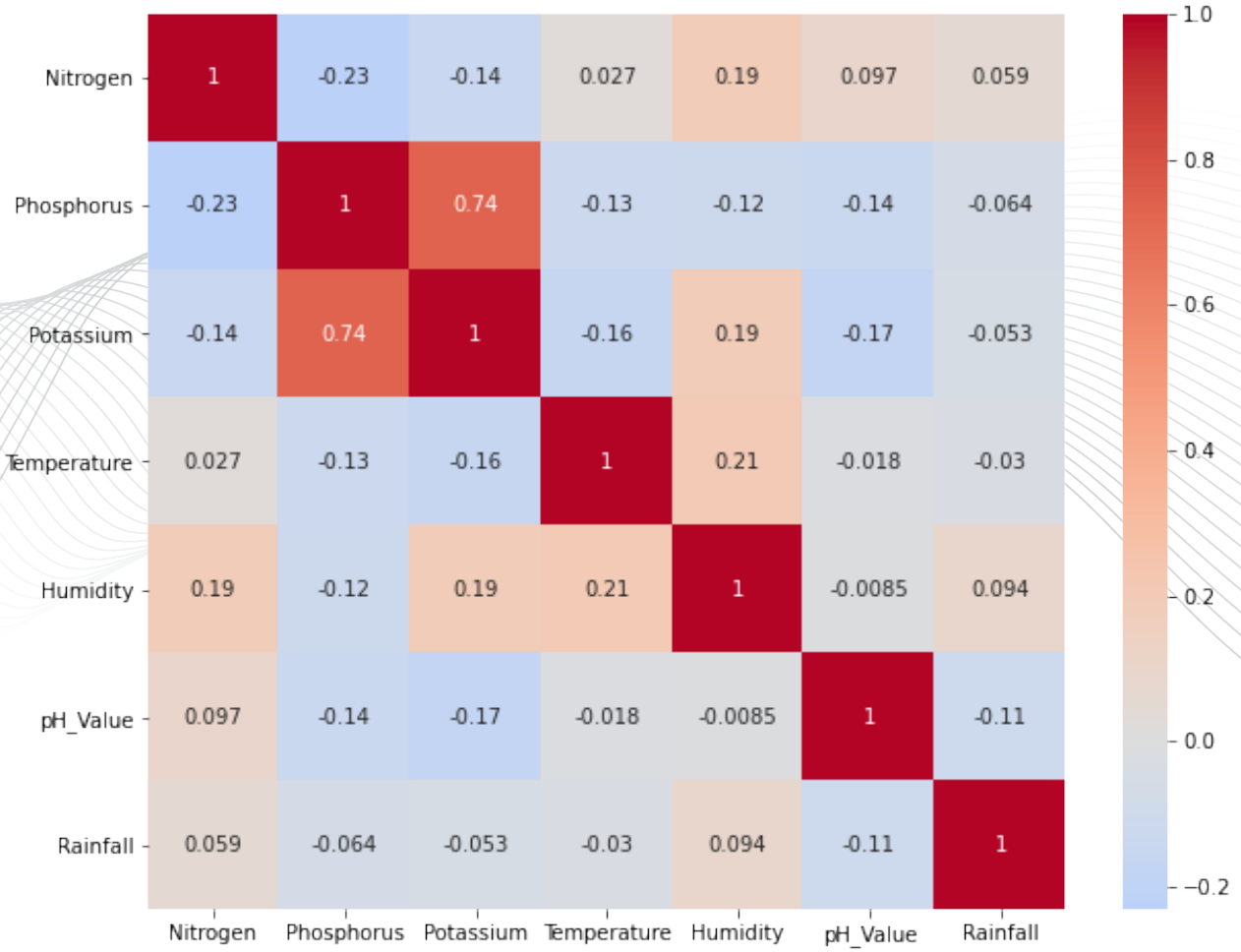
Any Questions?

Appendix

Appendix A – Data Understanding

As part of the initial Exploratory Data Analysis, a correlation matrix was created.

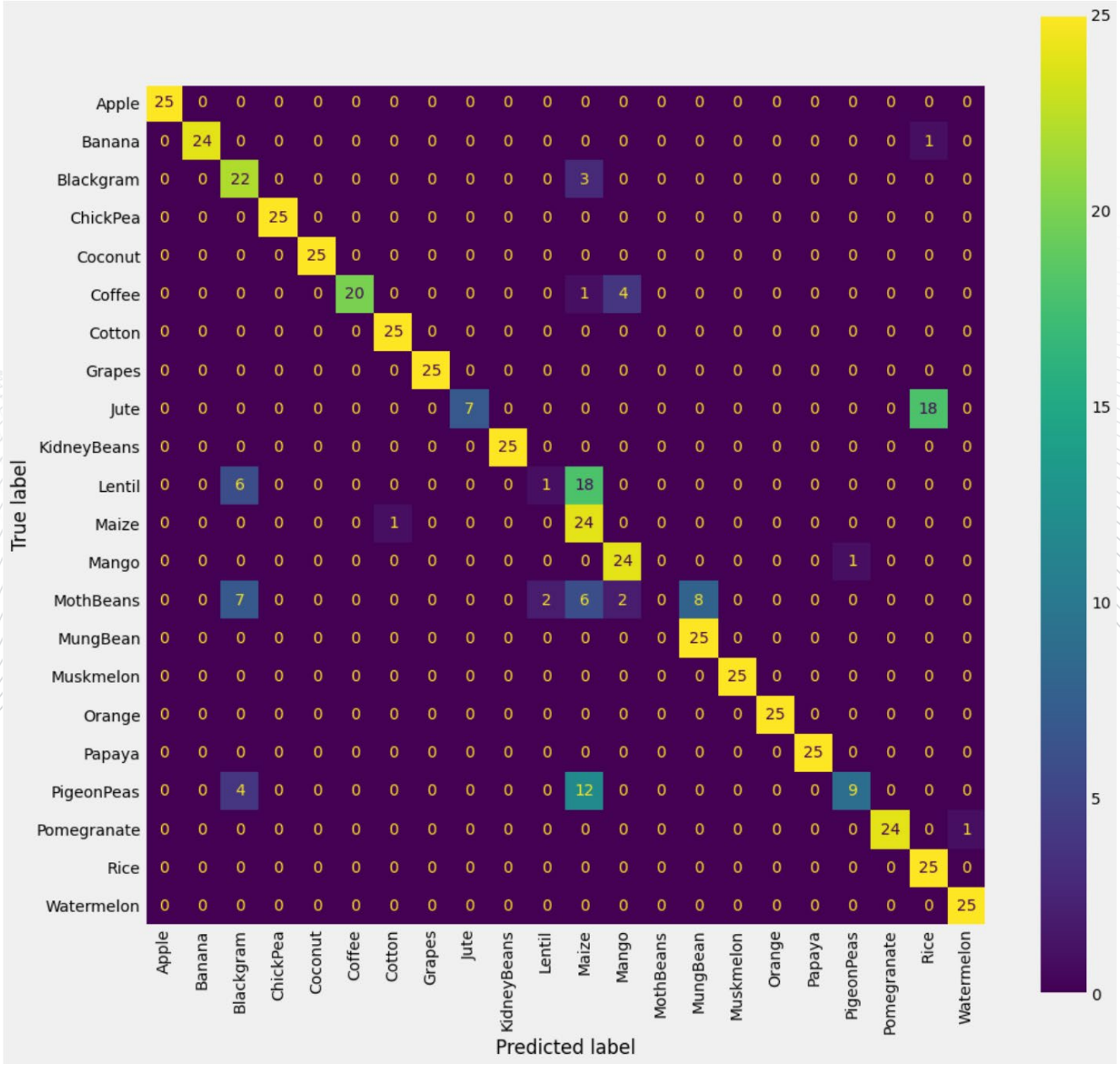
This showcases which data columns of the dataset are highly correlated.



Appendix B – Clustering

Confusion Matrix using random forest without Clustering

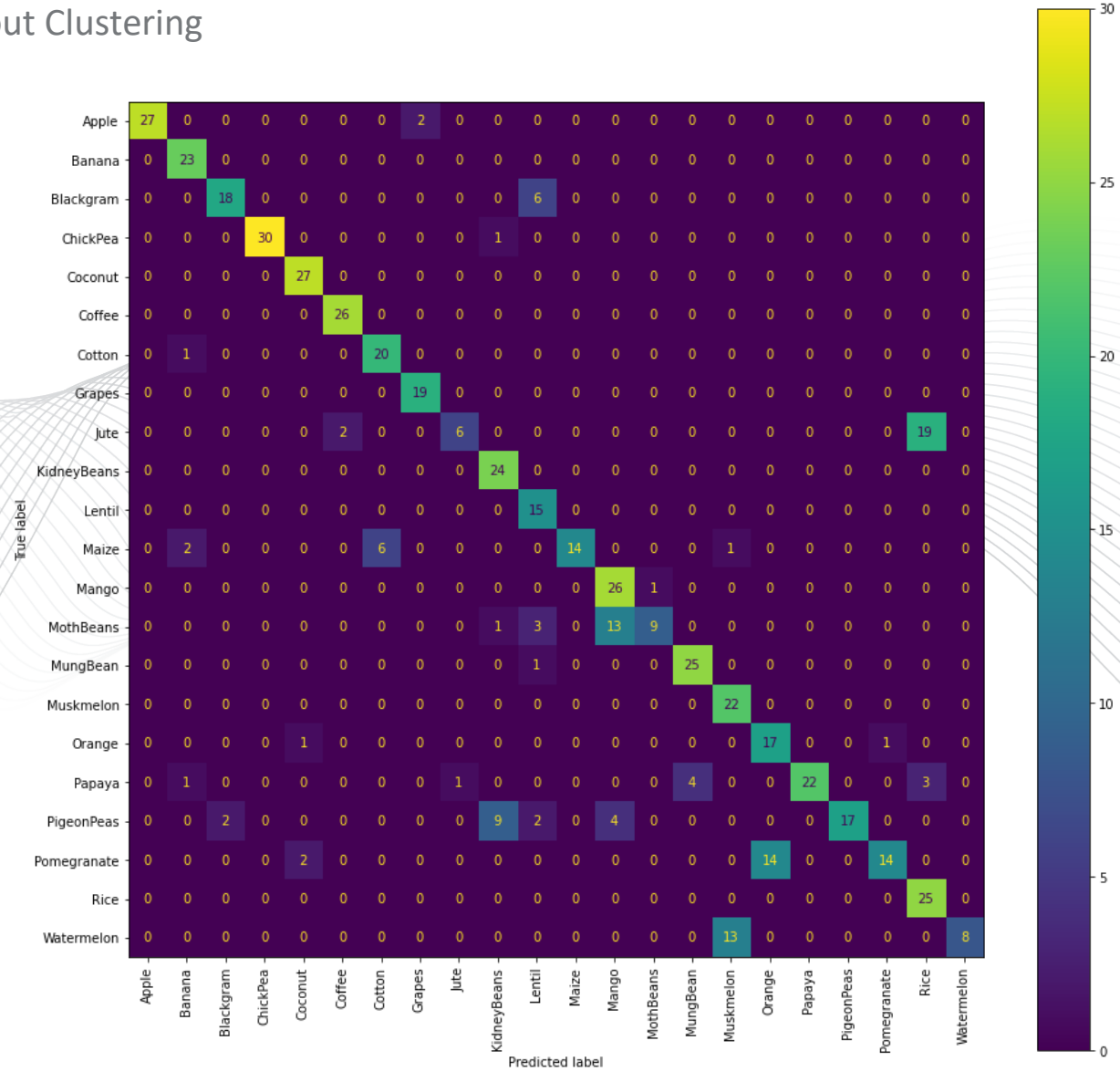
Confusion Matrix without Clustering: True vs Predicted Label



Appendix B – Clustering

Confusion Matrix using Logistic Regression without Clustering

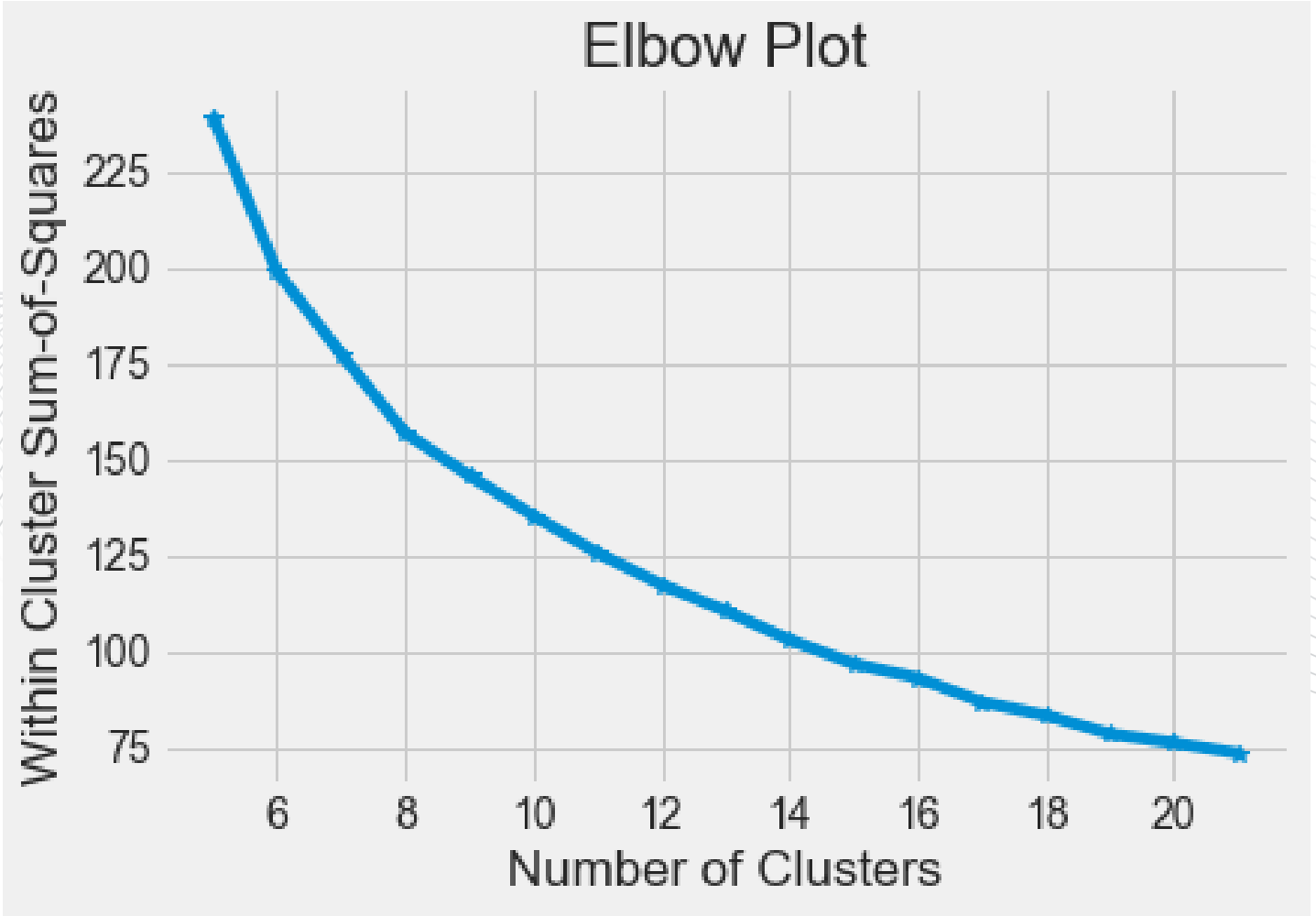
Logistic Regression Confusion Matrix: True vs Predicted Label



Appendix B – Clustering

Clustering Elbow Plot

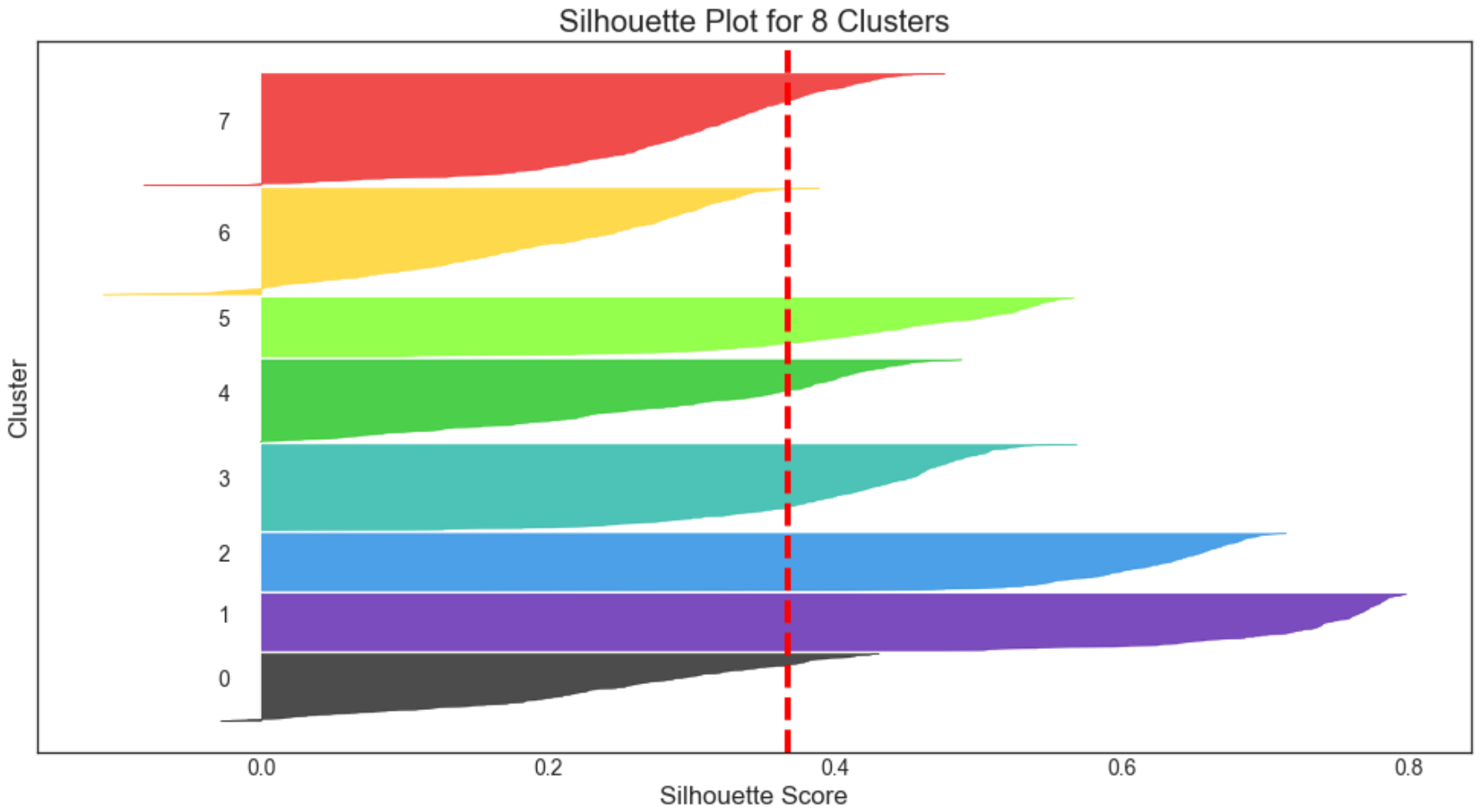
Elbow Plot: K-means Clustering Sum-of-Squares by Number of Clusters



Appendix B – Clustering

Clustering Silhouette Plot for 8 Clusters

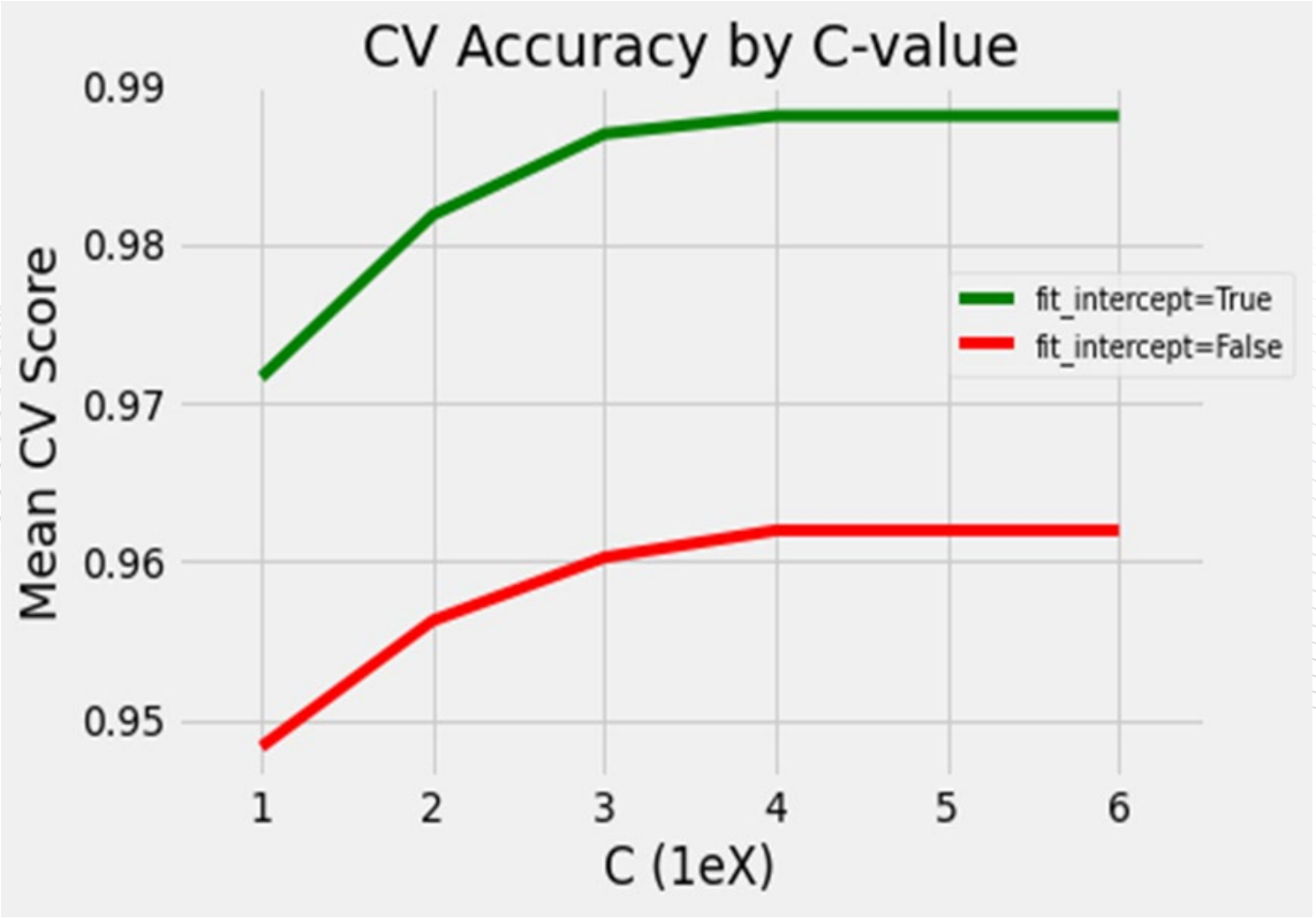
Silhouette Plot: Cluster vs Silhouette Score



Appendix C – Logistic Regression

Logistic Regression GridSearch Results

Logistic Regression Model Comparison: Accuracy by C-Value and fit_intercept

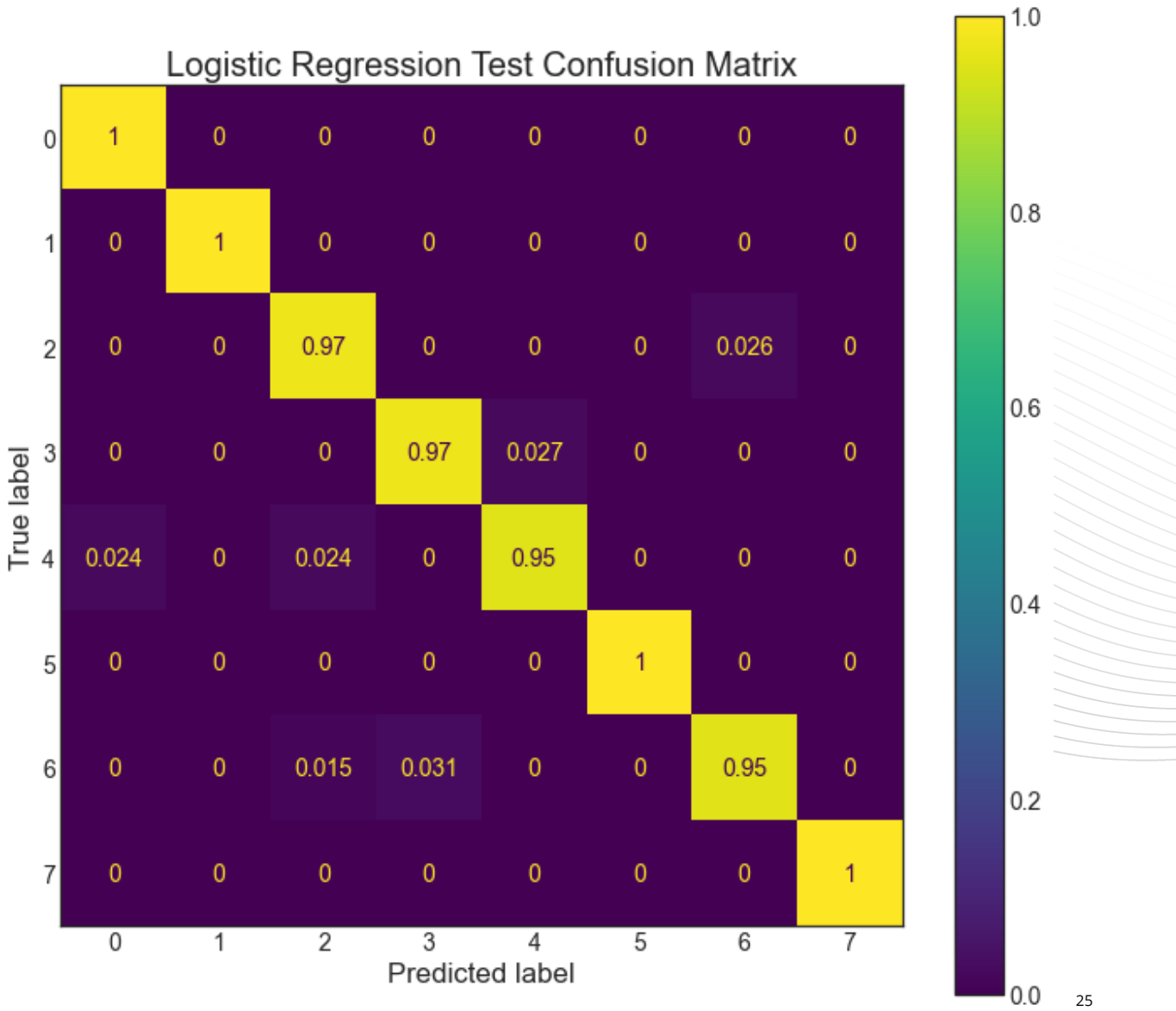


Appendix C – Logistic Regression

Logistic Regression Confusion Matrix on Testing Data with Legend

Confusion Matrix with Legend: True vs Predicted Label

Cluster	Crops
0	Chickpea, Kidney Benas
1	Grapes, Apples
2	Maize, Banana, Cotton, Coffee
3	Moth Beans, Mung Beans, Blackgram, Lentil
4	Pigeon Peas, Mango
5	Pomegranate, Orange, Coconut
6	Rice, Papaya, Jute
7	Watermelon, Muskmelon



Appendix D – Random Forest Classifier

Random Classifier Comparison Statistics

Random Selection: 12.5% Success rate

Decision tree comparable performance: 58.91% Accuracy

- Confusion Matrix & Stats Below

Random Forest Classifier:

- Max Depth: 3
- Estimators: 30
- Train Score: 96.42%
- Test Score: 96.00%

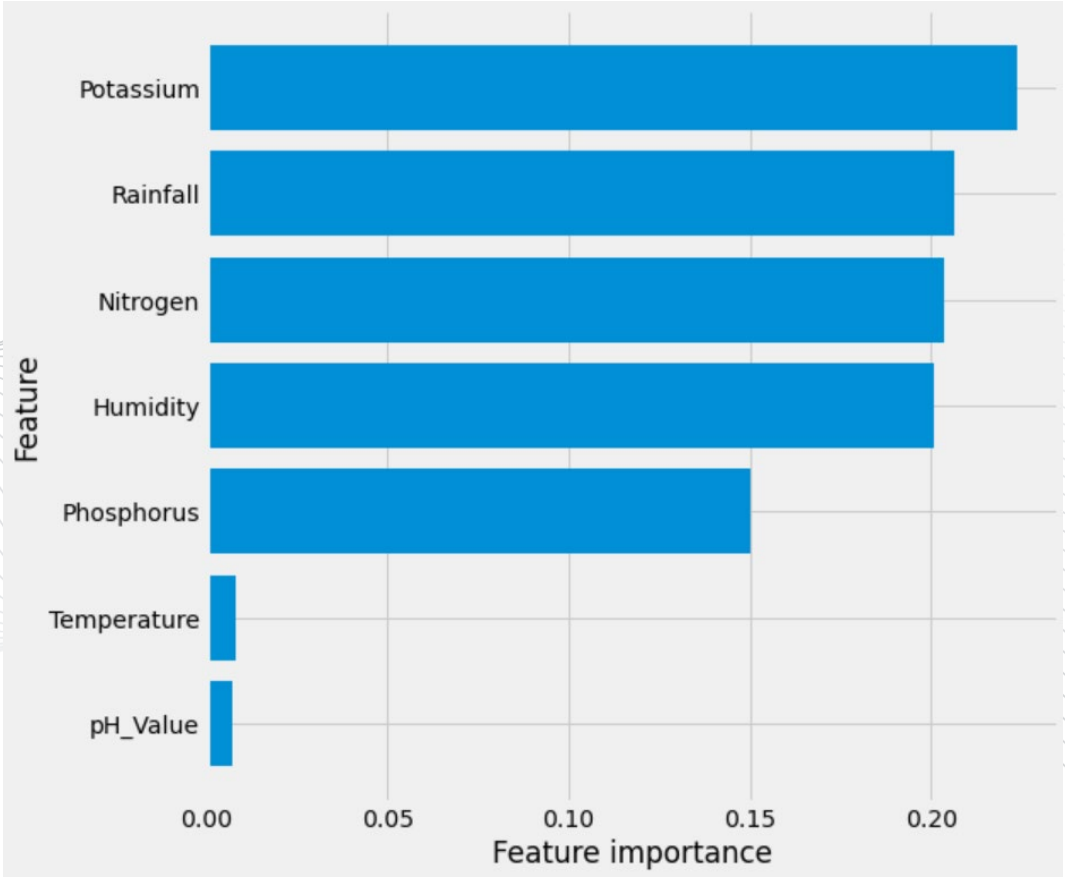
```
[[  0   0   0  50   0   0   0   0]
 [  0   0   0  50   0   0   0   0]
 [  0   0 100   0   0   0   0   0]
 [  0   0   1  99   0   0   0   0]
 [  0   0   0  33   0  17   0   0]
 [  0   0   0   0   0  75   0   0]
 [  0   0  53  20   0   0   0   2]
 [  0   0   0   0   0   0   0  50]]
```

	precision	recall	f1-score	support
ChickPea_KidneyBeans	0.00	0.00	0.00	50
Grapes_Apple	0.00	0.00	0.00	50
Maize_Banana_Cotton_Coffee	0.65	1.00	0.79	100
MothBeans_MungBean_Blackgram_Lentil	0.39	0.99	0.56	100
PigeonPeas_Mango	0.00	0.00	0.00	50
Pomegranate_Orange_Coconut	0.82	1.00	0.90	75
Rice_Papaya_Jute	0.00	0.00	0.00	75
Watermelon_Muskmelon	0.96	1.00	0.98	50
accuracy			0.59	550
macro avg	0.35	0.50	0.40	550
weighted avg	0.39	0.59	0.46	550

Appendix D – Random Forest Classifier

Random Classifier Feature Importance

Random Forest Feature Importance

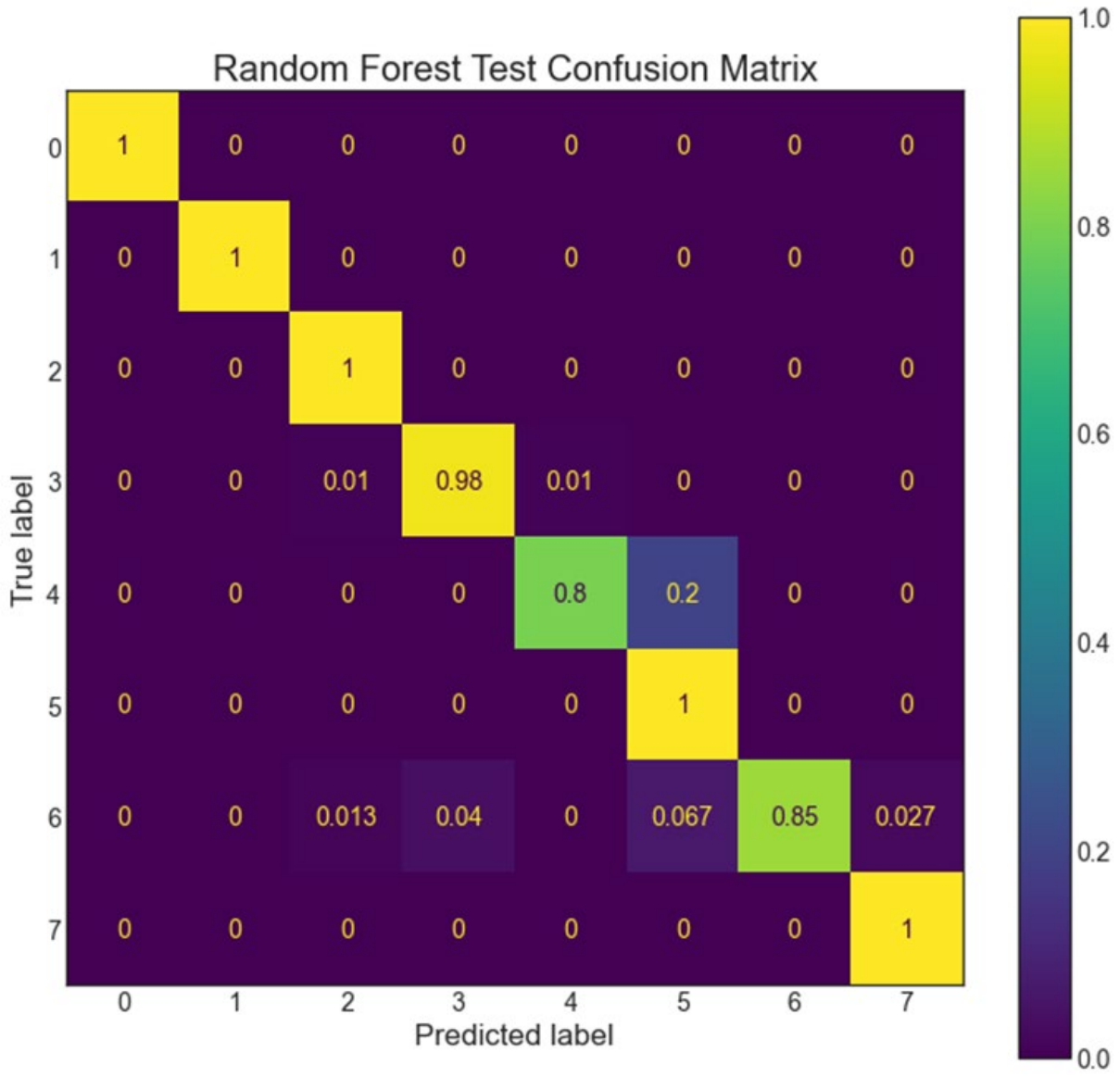


Appendix D – Random Forest Classifier

Random Forest Confusion Matrix on Testing Data with Legend

Confusion Matrix with Legend: True vs Predicted Label

Cluster	Crops
0	Chickpea, Kidney Benas
1	Grapes, Apples
2	Maize, Banana, Cotton, Coffee
3	Moth Beans, Mung Beans, Blackgram, Lentil
4	Pigeon Peas, Mango
5	Pomegranate, Orange, Coconut
6	Rice, Papaya, Jute
7	Watermelon, Muskmelon



Deloitte.