CLASSIFICATION OF DRY BEANS BASED ON PHYSICAL FEATURES USING MACHINE LEARNING

NEELESH VASNANI

15 AUGUST 2021

FINAL PROJECT

161.777 PRACTICAL DATA MINING

DR. MATTHEW PAWLEY

MASTER OF ANALYTICS

MASSEY UNIVERSITY

TABLE OF CONTENTS

EXECUTIVE SUMMARY…………………………………………………………... 1

1. INTRODUCTION AND MOTIVATION…………………………………………..... 2-3

2. METHODS…………………………………………………………………… 3-9

    2.1 DATA OVERVIEW…………………………………………………..….. 3-7

    2.2 STATISTICAL METHODOLOGY………………………………………….. 7-9

3. RESULTS AND ANALYSIS…………………………………………………… 10-14

4. DISCUSSION AND CONCLUSIONS…………………………………………… 14-15

REFERENCES………………………………………………………………….. 16

APPENDICES…………………………………………………………………... 17-20

EXECUTIVE SUMMARY

Dry beans are the most produced and most consumed pulse in the world. Given the high degree of diversity among dry beans, there is a strong motivation to classify and sort dry beans according to variety in order to increase market value and avoid economic losses. Manual classification of dry beans is expensive, time-consuming, and inefficient. To this end, automated classification can provide operational and cost benefits. This analysis employed ten different machine learning techniques with the goal of automatically classifying dry bean varieties based on their physical features. The data used for the analysis was captured through image processing and contains 16 physical geometric attributes describing 13,611 grains of 7 dry bean varieties.

Prior to modeling, the data was explored to get a better feel of the variation in the variables. The feature variables were transformed for maximum normality to ensure optimal results. The analysis made use of 10 machine learning classification models, namely. standard logistic regression, polynomial logistic regression, decision tree, random forest, support vector machine (SVM), linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), neural network (NN), gradient boosting, and ensemble. 70% of the data was used to train the models while 30% was reserved exclusively for validation.

The results revealed that it is possible to automatically classify dry beans with high accuracy using physical features. The ten models had promising accuracy rates ranging from 90.2% to 93.24%. Polynomial logistic regression was the best performing model with the lowest misclassification rate of 6.76% and the highest accuracy of 93.24%. Based on the best model, the accuracy rates of the seven dry bean varieties - Sira, Barbunya, Dermason, Seker, Horoz, Cali, and Bombay - were 87.28%, 91.76%, 94.14%, 94.47%, 95.21%,95.56%, and 100% respectively. Moreover, the most important variables for differentiating the dry bean varieties were found to be ShapeFactor1, EquivDiameter, MinorAxisLength, Area, and Perimeter. With these results, it makes sense for dry bean producers to invest in automated systems for dry bean classification.

1. INTRODUCTION AND MOTIVATION

In many developing nations, agriculture is one of the largest drivers of economic growth. While these countries may be abundant in natural resources and crops, agricultural activities are often carried out with the lack of modern technology. Traditionally done based on human judgment and knowledge, the classification of seeds, beans, and pulses is one such activity with a great potential for automation (Agrawal & Dahiya, 2018). Accordingly, an increasing number of studies have explored the automation of seed classification using machine learning techniques. Machine learning techniques have been used for the classification of a variety of seeds including dry beans, chickpeas, coffee beans, corn seeds, and grain legumes, to name a few. For example, Pinto et al. (2017) developed a neural network (NN) model to classify green coffee beans with up to 98.7% accuracy. Similarly, Ayele and Tamiru (2020) employed machine learning techniques such as neural networks, support vector machine (SVM), and decision tree (DT) to classify Ethiopian chickpeas with up to 97.5% accuracy. In their nuanced work, Perez-Rodriguez (2019) employed linear discriminant analysis (LDA), random forest (RF), and support vector machine (SVM) models to classify cowpea beans with 93% accuracy. In this analysis, the focus will solely be on the classification of dry beans as a case study for automated seed classification.

Dry bean (scientific name: Phaseolus vulgaris L.) is the most edible and most produced pulse in the world. The production, supply, and consumption of dry beans play a crucial role in the agriculture sector of countries where it is cultivated. Dry beans come in several different varieties and there is a high degree of genetic diversity among these varieties. It is not uncommon for farmers and producers to deal with mixed, unsorted bags of dry beans. This is concerning because the simultaneous cultivation of different species of dry beans results in economic losses (Kokul & Ozkan, 2020). It follows that there is a strong need to sort and classify dry beans by variety. However, manual sorting and classification is not a feasible and practical solution in the long run for two main reasons. First, manual sorting of dry beans is a time-consuming process that requires heavy human labor resources which makes it expensive and inefficient at high production volumes. Second, dry bean varieties typically have similar color hues which makes visual inspection difficult and error-prone. These two reasons underpin a strong motivation for the automated classification of dry beans. Using machine learning techniques, dry beans can be automatically and accurately classified based on physical geometric measurements obtained through image processing. Accurate classification of dry beans leads to uniform seed varieties during cultivation. This enables farmers to adhere to better standards for planting and marketing, resulting in higher seed quality (Ajaz & Hussain, 2015). Better seed quality results in a higher production yield of dry beans per batch as well as a reduced likelihood of seed disease. These advantages lead to increased market value and value-based pricing of the dry beans which not only benefits the farmers but also the economy at large. In this way, the accurate classification of dry beans benefits both production and marketing activities involved in the end-to-end sale of dry beans.

This analysis extends upon the data developed by Koklu and Ozkan (2020) who used image processing to capture physical features of dry beans and summarized them into a single dataset. In their study, they also performed multiclass classification of the dry bean samples using computer vision and machine learning techniques and obtained 93% accuracy. This analysis in part aims to match or improve the accuracy obtained by the original authors using a variety of machine learning techniques. Specifically, the following research aims are addressed:

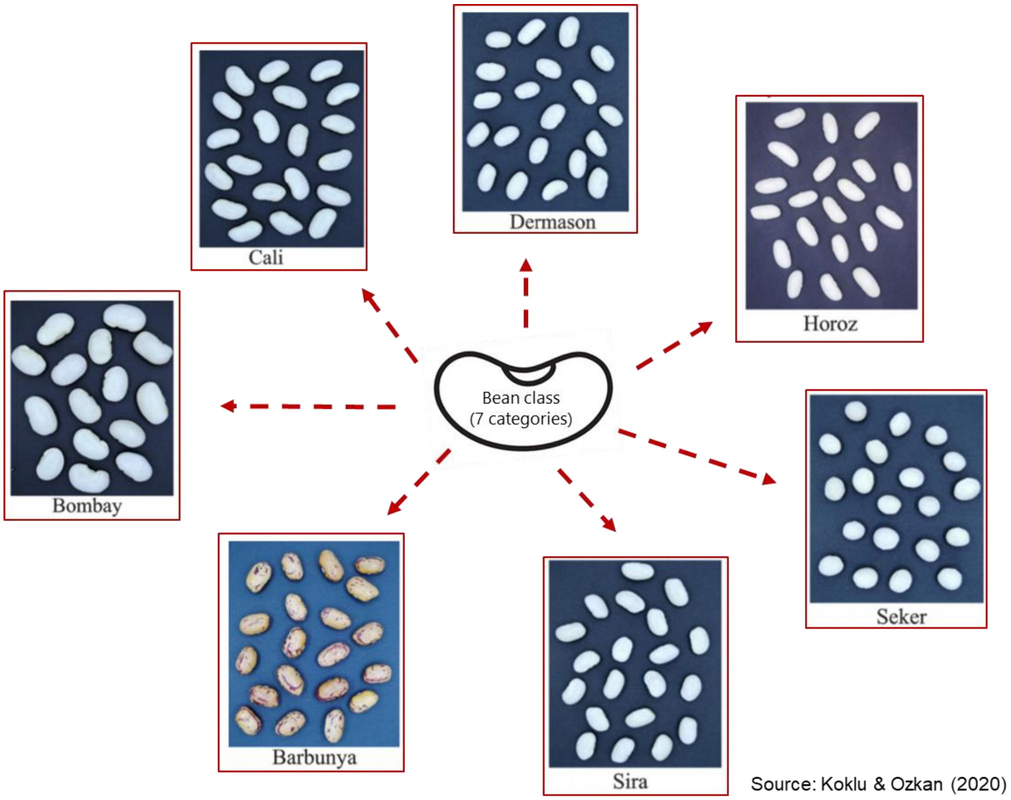
- Build and compare different machine learning models to accurately classify dry bean varieties using physical geometric attributes captured through image processing
- Evaluate whether the machine learning models proposed by the original authors of the dataset, Koklu and Ozkan (2020), can be improved in terms of predictive power
- Identify the machine learning technique that best discriminates among different dry bean varieties based on misclassification rates and relevant performance metrics
- Compare the different dry bean varieties in terms of ease of classification
- Determine which physical attributes are most important for classifying dry beans

2. METHODS

2.1. DATA OVERVIEW

The dataset is sourced from Koklu and Ozkan (2020) and consists of 13,611 observations which correspond to 13,611 samples of dry bean grains that were obtained from 236 images captured through image processing. For this analysis, the 13,611 observations are randomly split into 70% (9,521 observations) for training and 30% (4,090 observations) for validation. To obtain the images of dry beans, Koklu and Ozkan (2020) sourced 1 kilogram worth of the seven dry bean varieties from certified seed producers based in Turkey. They used a computer vision system consisting of a high-resolution 2.2 megapixel Prosilica GT2000C camera, a lens mount, and an illumination box to prevent shadow formation to capture the images. These images were then processed using MATLAB to extract geometric features of the dry beans. There are 16 numerical feature variables in the data which describe the physical geometric attributes of the dry beans. Out of the 16 features, 12 features are related to core geometric dimensions from which 4 additional features are derived to describe the shape form of the beans. The target variable is a multiclass nominal variable containing seven classes of dry beans: Barbunya, Bombay, Cali, Horoz, Seker, Sira, and Dermason. The figure below shows a sample of the images for each of the seven dry bean varieties.

Figure 1. Dry Bean Class (Target Variable)



Source: Koklu & Ozkan (2020)

As can be observed in Figure 1 above, the seven dry bean varieties appear to be quite similar in terms of color. Therefore, color would not be an effective differentiator for classifying and sorting the dry beans. Discrimination based on geometric features which encapsulate the size, form, shape, and structure of the dry beans is more appropriate. Table 1 below provides a short description of the seven dry bean varieties (Koklu & Ozkan, 2020).

Table 1. Description of Dry Bean Varieties

Dry Bean Variety	Qualitative description
Cali	White color, kidney-shaped, slightly plump
Horoz	White color, long and cylindrical, medium-sized
Dermason	White color, fuller flat, round-shaped
Seker	White-color, round shape, large-sized
Bombay	White color, oval and bulging shape, very large size
Barbunya	Beige and red speckled color, oval-shaped, large size
Sira	White color, flat and round shape, small size

Table 1 below summarizes all the variables in the dataset along with the description and role for each variable.

Table 2. Definition of Variables

Variable	Description	Role
Class	Dry bean classification (Seker, Sira, Horoz, Cali, Bombay, Dermason, or Barbunya)	Target
Area (A)	Area of bean zone in pixels	Input/Feature
Perimeter (P)	Circumference of bean	Input/Feature
Major Axis Length (L)	Length of longest line that can be drawn from a bean	Input/Feature
Minor Axis Length (l)	Length of longest line drawn from a bean, perpendicular to main axis	Input/Feature
Aspect Ratio (K)	Ratio of major and minor axis lengths (L / l)	Input/Feature
Eccentricity (Ec)	Eccentricity of ellipse with same moments as bean region	Input/Feature
Convex Area (C)	Number of pixels in smallest convex polygon containing bean area	Input/Feature
Equivalent Diameter (Ed)	Diameter of the circle whose area equals bean area	Input/Feature
Extent (Ex)	Ratio of pixels of the bean-bounding rectangle to the bean area	Input/Feature
Solidity (S)	Ratio of pixels in convex shell to area of the bean in pixels	Input/Feature
Roundness (R)	Calculated as: (4πA)/(P^2)	Input/Feature
Compactness (CO)	Calculated as: Ed / L	Input/Feature
Shape Factor 1 (SF1)	Calculated as: L / A	Input/Feature
Shape Factor 2 (SF2)	Calculated as: l / A	Input/Feature
Shape Factor 3 (SF3)	Calculated as: A / (L*L*π/4)	Input/Feature
Shape Factor 4 (SF4)	Calculated as: A / (L*l*π/4)	Input/Feature

Figure 2 below shows the number of records in the dataset for each dry bean variety. The Dermason variety has the highest number of observations (3546, 26% of total) while the Bombay variety has the lowest number of observations (522, 3.8% of total). For seven classes, the ideal number of observations in a perfect sample for each variety would be 1,944 or 14% per variety.

Figure 2. Number of Records per Bean Class Type

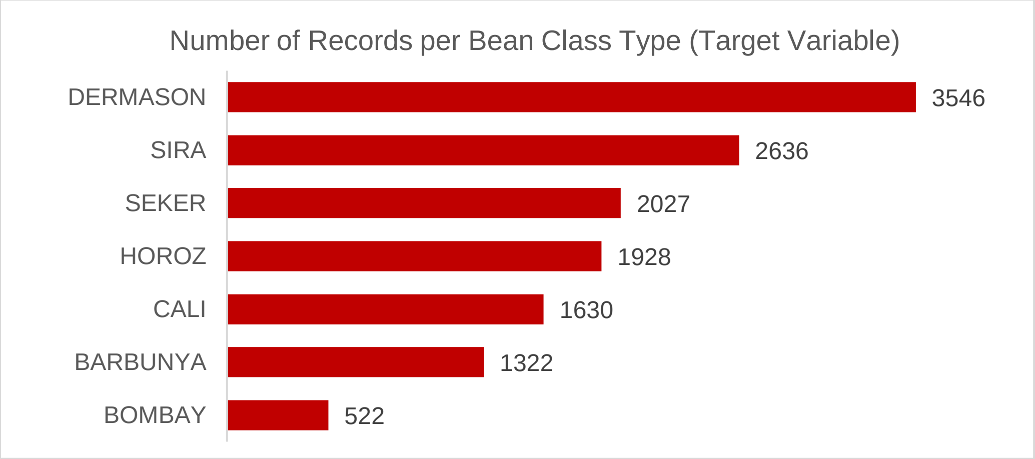


Figure 3. Distribution of Feature Variables

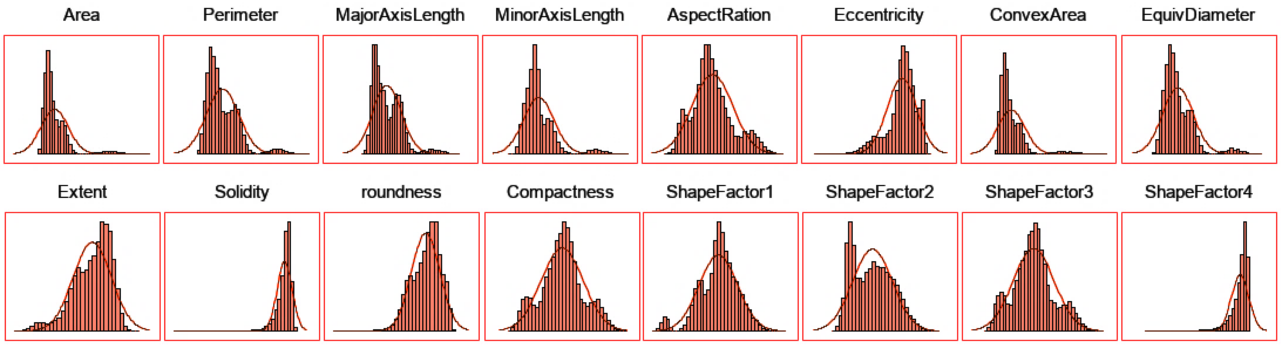


Figure 3 above shows the distribution of the 16 feature variables in the dataset. Generally, most of the variables appear to have an approximately symmetric, normal distribution. However, the variables area, perimeter, and convex area appear to be right-skewed. Moreover, solidity, shape factor 4, and roundness appear to be left-skewed. Some bimodal patterns are present as well; for instance, in perimeter, major axis length, and eccentricity. Moreover, compactness and shape factor 3 appear to have trimodal distributions. These bimodal and trimodal patterns could represent the inherent variation in physical attributes between the different dry bean varieties.

Table 3. Mean Values of Features by Bean Variety

Bean Variety	Area	Perimeter	MajorAxisL	MinorAxisL	AspectRatio	Eccentricity	ConvexArea	EqDiameter	Extent	Solidity	Roundness	Compactness	SF1	SF2	SF3	SF4
BARBUNYA	69804	1046.11	370.04	240.31	1.54	0.75	71026	297.31	0.749	0.983	0.800	0.805	0.005	0.001	0.649	0.996
BOMBAY	173485	1585.62	593.15	374.35	1.59	0.77	175813	468.94	0.777	0.987	0.864	0.793	0.003	0.001	0.629	0.992
CALI	75538	1057.63	409.50	236.37	1.73	0.81	76689	309.54	0.759	0.985	0.846	0.757	0.005	0.001	0.573	0.991
DERMASON	32119	665.21	246.56	165.66	1.49	0.74	32498	201.68	0.753	0.988	0.908	0.819	0.008	0.002	0.672	0.997
HOROZ	53649	919.86	372.57	184.17	2.03	0.87	54440	260.73	0.706	0.985	0.794	0.701	0.007	0.001	0.492	0.992
SEKER	39881	727.67	251.29	201.91	1.25	0.58	40270	224.95	0.772	0.990	0.945	0.897	0.006	0.003	0.805	0.998
SIRA	44729	796.42	299.38	190.80	1.57	0.77	45273	238.34	0.749	0.988	0.885	0.797	0.007	0.002	0.636	0.995
Overall	53048	855.28	320.14	202.27	1.58	0.75	53768	253.06	0.750	0.987	0.873	0.800	0.007	0.002	0.644	0.995

Table 3 above shows the descriptive mean values (measured in pixels) for each feature broken down by bean variety. This helps to provide a visual hypothesis regarding which features can be expected to be most useful in discriminating among the bean varieties. For example, the mean values for area and perimeter seem to vary significantly across the seven varieties. Accordingly, these variables can be expected to be important features. On the other hand, the mean values for variables such as extent, solidity, roundness, and shape factor 4 appear to be relatively similar across bean varieties. From the table, it can also be inferred that the Bombay variety is the largest in size since it has the highest area and perimeter. The complete descriptive statistics including standard deviations, maximum, and minimum values can be found in Appendix A.

With optimized model performance in mind, the 16 feature variables were transformed to achieve maximum normality. Table 4 below summarizes the data transformations used.

Table 4. Summary of Variable Transformations

Logarithm	Power	Square root	Exponential
Area	Eccentricity	AspectRation	Extent
Perimeter	Solidity	ShapeFactor2	Roundness
MajorAxisLength	ShapeFactor4	ShapeFactor3	
MinorAxisLength			
ConvexArea			
EquivDiameter			
Compactness			
ShapeFactor1			

2.2. STATISTICAL METHODOLOGY

In machine learning terminology, predictive modeling is known as supervised learning. If the variable to be predicted is categorical, the prediction task is a classification problem. If the variable to be predicted is continuous, then it is a regression model (Ye, 2013). Since the goal of this analysis is to predict or classify the category of dry bean among seven different varieties, classification models are appropriate. Several machine learning algorithms exist for classification tasks. Moreover, different algorithms can produce different results depending on the dataset. Accordingly, it is important to compare several machine learning techniques to uncover which algorithm performs best for the task at hand. For this analysis, 10 machine learning models are employed as described below:

- **Logistic regression** – Logistic regression is similar to linear regression but is used when the target variable is binary or multi-class as opposed to continuous. The resulting predicted values surface as the probability of the levels of the target variable with respect to a certain set of inputs. In a similar study, Ajaz and Hussain (2015) used logistic regression to classify Chickpea beans. Since Chickpea beans are the second most consumed pulse after dry beans, it would be meaningful to see how logistic regression performs for dry beans.
- **Polynomial logistic regression** – This model is simply a more complex and advanced version of the standard logistic regression in that it includes polynomial terms, two-way interactions, and a 100-max-step stepwise method. Although more computationally intensive, this model can produce better results for nonlinear data.
- **Decision tree** – A decision tree is a visual classification method that follows a tree-like path to show classification decisions for different ranges of input variables at different hierarchies of partitions (Ye, 2013). Decision trees are widely used for classification tasks because of their easy interpretation, visual appeal, and simple implementation. Ayele and

Tamiru (2020) used decision trees to classify chickpeas while Koklu and Ozkan (2020) used the method to classify dry beans.

- **Neural Network** – Hinged on the biological concept of how neurons operate, Neural Networks (NN) - also known as Multilayer Perceptrons (MLP) - are powerful algorithms that can be understood as advanced extensions of regression models in that they include an additional hidden layer between the input variables layer and target variable layer. The advantage of Neural Networks is that they can model virtually any relationship between an input and target variable with high accuracy. The drawback, however, is poor interpretability of the model. Among seed classification studies, Neural Network is the most prevalent and widely used algorithm due to its high performance (Koklu & Ozkan, 2020).
- **Linear discriminant analysis (LDA)**- LDA is one of the simpler methods used for classification and is based on dimensionality reduction. LDA is included as an alternate model primarily because it has shown promising results for seed classification when combined with image processing techniques (Medeiros et al., 2020). For this analysis, an LDA model with proportional (as opposed to equal) prior probabilities is used.
- **Quadratic discriminant analysis (QDA)** – QDA is similar to LDA except in that it can learn and use quadratic boundaries which therefore makes it more flexible. If the null hypothesis that the covariance matrixes are homogeneous is rejected, QDA is more appropriate than LDA. Moreover, QDA may work better for nonlinear data compared to LDA. As such, it is important to test the performance of QDA in addition to LDA.
- **Support vector machine (SVM)** – SVM is a kernel-based algorithm with high computational power for classification and regression problems. It is known to have better generalization, a stronger theoretical base, and more accurate results, especially for nonlinear data. The motivation for including this algorithm is that Koklu and Ozkan (2020) found SVM to be the best performing model for dry bean classification.
- **Gradient Boosting** – Gradient boosting is a recursive algorithm that creates a series of decision trees by fitting the error of the prediction from the earlier tree in the series. It essentially creates a strong model from several weak models. The model was run in SAS Enterprise Miner with a shrinkage or learning rate of 0.1 using 100 iterations.
- **Random Forest** – The random forest algorithm works by constructing several decision trees and returns the output based on the class selected by most trees (for classification models). One advantage of random forests is that they correct for decision trees' tendency of overfitting. The HP Forest node in SAS Enterprise Miner with default settings is used for this model.
- **Ensemble** – An ensemble model is treated as a single model that is comprised of several constituent models, making it powerful in terms of predictive accuracy and model stability. For categorical target variables such as in this case, the majority "vote" across the

constituent models is taken. A manual ensemble model is used in this analysis by combining the Neural Network, Logistic Regression, Decision Tree, and Gradient Boosting models. Ensemble models are relatively new, so they are not as widely used in the literature.

Table 5. Comparison of Machine Learning Technique Usage with Related Seed Classification Studies

Model	**This Analysis**	Koklu & Ozkan (2020)	Pinto et al. (2017)	Medeiros et al. (2020)	Ayele & Tamiru (2020)	Basol & Toklu (2021)	Perez et al. (2019)	Ali et al. (2020)	Ajaz & Hussain (2015)	Agrawal & Dahiya (2020)	Number of Studies (excluding this analysis')
Seed type	**Dry Bean**	Dry Bean	Coffee bean	Curcas	Chickpea	Various	Cowpea	Corn seed	Chickpea	Various	
Decision Tree	■	■			■				■	■	4
Standard Logistic Regression	■							■	■		2
Polynomial Logistic Regression	■										
Random Forest	■						■	■			2
Neural Network	■	■	■		■	■		■	■		6
Gradient Boosting	■										
Support Vector Machine (SVM)	■	■			■		■			■	4
Linear Discriminant Analysis	■			■			■			■	3
Quadratic Discriminant Analysis	■										
K-Nearest-Neighbors (kNN)		■					■			■	3
Ensemble	■										

Table 5 above shows a comparison of models used in this analysis relative to nine other seed classification studies. This analysis compares 10 different models for dry bean classification, most of which have been used by other seed classification studies as well. However, four models, namely: polynomial logistic regression, gradient boosting, quadratic discriminant analysis, and manual ensemble, have not been used by any of the other nine studies. Neural Network is the most prevalent model as it was used by six out of the nine reviewed studies, followed by SVM and decision trees which were employed by four studies. Overall, the 10 models selected for analysis represent a good combination of linear (e.g. Logistic regression, LDA, etc) and nonlinear (e.g. SVM, QDA, Decision trees, Neural networks, etc.) machine learning algorithms.

During the fitting process, all models used misclassification rates as the selection criterion for model assessment. Moreover, a 70% train – 30% validation split was used wherein all models were assessed based on misclassification rates on the validation data. In addition to misclassification rates (i.e. 100% – accuracy), the 10 models were also evaluated using model performance metrics such as precision, specificity, recall, and F1-Score. The models were ran using SAS Enterprise Miner software, SAS Base, and Python on a Windows 10 machine with 16GB ram and an Intel i7 processor. Specifically, SAS Base was used for the linear and quadratic discriminant analysis models (See Appendix B) while Python was used for the SVM model. The other seven models were all ran using SAS Enterprise Miner. Some results were reformatted using MS Excel for ease of reporting.

3. RESULTS AND ANALYSIS

Table 6. Comparison of Models Ranked by Misclassification Rate on Validation Data

Model	Misclassification Rate	ASE	ROC Index
1 - Polynomial Logistic Regression	6.76%	0.0145	0.986
2 - Support Vector Machine (SVM)	6.85%	0.0149	0.985
3 - Ensemble	7.02%	0.0152	0.985
4 - Quadratic Discriminant Analysis (QDA)	7.20%	0.0153	0.985
5 - Standard Logistic Regression	7.48%	0.0155	0.985
6 - HP Random Forest	7.75%	0.0164	0.984
7 - Neural Network	7.87%	0.0163	0.985
8 - Gradient Boosting	7.97%	0.0168	0.984
9 - Linear Discriminant Analysis (LDA)	8.25%	0.0182	0.975
10 - Decision Tree	9.80%	0.0247	0.945

As can be seen in Table 6, all models perform fairly well in classifying the dry beans with misclassification rates below 10%, ranging from 6.76% for the best model to 9.80% for the worst model. The ROC Index scores (i.e. area under the ROC curve) for the models which range from 0.945 to 0.986, coupled with low ASEs, further indicate a healthy model fit. The best performing model is polynomial logistic regression with a misclassification rate of 6.76% or an accuracy of 93.24%. This is marginally better than the best model obtained by the original authors, Koklu and Ozkan (2020), who got an accuracy of 93.19% using SVM. As shown in the table, SVM is the second-best model with a misclassification rate of 6.85% or an accuracy of 93.15%. This result coincides with that obtained by the original authors. In their model for classifying cowpea beans, Perez-Rodriguez et al. (2019) got an accuracy of 93% using SVM, which was also their best model outperforming LDA, k-Nearest-Neighbors, and random forest. This potentially suggests that SVM tends to perform well for the classification of not only dry beans but seeds in general. While the ensemble model does not have the best misclassification rate as one might expect, it performs better than three of its constituent models including neural network, decision tree, and gradient boosting. The neural network model had a misclassification rate of 7.87% or an accuracy of 92.13% which is higher than that obtained by Koklu and Ozkan (2020) who got an accuracy of 91.73% for their MLP model. On the flip side, the misclassification rate of 9.8% or accuracy of 90.2% for the decision tree model is worse compared to the 92.36% obtained by the original authors. Among the 10 models, the decision tree has the highest (i.e. poorest) misclassification rate of 9.8%. This deviates greatly from the results of Ayele and Tamiru (2020) who found that decision trees work best in classifying chickpea beans with an accuracy rate of 97.5%. Also noteworthy is that the superior performance of polynomial logistic regression compared to standard logistic regression, and quadratic discriminant analysis compared to linear discriminant analysis, suggests that non-linear decision boundaries work better for classifying dry beans based on physical features.

Table 7. Confusion Matrix for All Models (Validation Data)

	Predicted Positive		Predicted Negative	
Actual Positive	**True Positives (TP)**		**False Negatives (FN)**	
	1 - Polynomial Logistic Regression	691	1 - Polynomial Logistic Regression	100
	2 - Support Vector Machine (SVM)	687	2 - Support Vector Machine (SVM)	104
	3 - Ensemble	683	3 - Ensemble	108
	4 - Quadratic Discriminant Analysis (QDA)	686	4 - Quadratic Discriminant Analysis (QDA)	105
	5 - Standard Logistic Regression	689	5 - Standard Logistic Regression	102
	6 - HP Random Forest	679	6 - HP Random Forest	112
	7 - Neural Network	688	7 - Neural Network	103
	8 - Gradient Boosting	679	8 - Gradient Boosting	112
	9 - Linear Discriminant Analysis (LDA)	675	9 - Linear Discriminant Analysis (LDA)	115
	10 - Decision Tree	672	10 - Decision Tree	119
Actual Negative	**False Positives (FP)**		**True Negatives (TN)**	
	1 - Polynomial Logistic Regression	94	1 - Polynomial Logistic Regression	3205
	2 - Support Vector Machine (SVM)	95	2 - Support Vector Machine (SVM)	3204
	3 - Ensemble	96	3 - Ensemble	3203
	4 - Quadratic Discriminant Analysis (QDA)	103	4 - Quadratic Discriminant Analysis (QDA)	3196
	5 - Standard Logistic Regression	113	5 - Standard Logistic Regression	3186
	6 - HP Random Forest	101	6 - HP Random Forest	3198
	7 - Neural Network	113	7 - Neural Network	3186
	8 - Gradient Boosting	110	8 - Gradient Boosting	3189
	9 - Linear Discriminant Analysis (LDA)	119	9 - Linear Discriminant Analysis (LDA)	3181
	10 - Decision Tree	127	10 - Decision Tree	3172

To evaluate whether a model is good enough to scale for practical use, the classification accuracy is not sufficient (Ye, 2013). For an exhaustive assessment, close attention must be paid to the true positives, true negatives, false positives, and false negatives, as well as additional performance metrics associated with these four parameters such as precision, sensitivity, specificity, and F1-Score. The confusion matrix in Table 7 above shows that polynomial logistic regression (the best model based on misclassification rates) indeed has the highest number of true positives and true negatives and the lowest number of false negatives and false positives. From the actual and predicted values, the performance metrics are computed as shown in Table 8 below.

Table 8. Model Performance Metrics

Model	Sensitivity	Specificity	Precision	F1-Score	Misc. Rate	Accuracy
1 - Polynomial Logistic Regression	87.4%	95.2%	88.0%	87.7%	6.8%	93.24%
2 - Support Vector Machine (SVM)	86.9%	95.2%	87.9%	87.3%	6.9%	93.15%
3 - Ensemble	86.3%	95.2%	87.7%	87.0%	7.0%	92.98%
4 - Quadratic Discriminant Analysis	86.7%	95.0%	86.9%	86.8%	7.2%	92.80%
5 - Standard Logistic Regression	87.1%	94.7%	85.9%	86.5%	7.5%	92.52%
6 - HP Random Forest	85.8%	95.0%	87.1%	86.4%	7.8%	92.25%
7 - Neural Network	87.0%	94.7%	85.9%	86.4%	7.9%	92.13%
8 - Gradient Boosting	85.8%	94.8%	86.1%	85.9%	8.0%	92.03%
9 - Linear Discriminant Analysis (LDA)	85.4%	94.5%	85.0%	85.2%	8.3%	91.75%
10 - Decision Tree	85.0%	94.3%	84.1%	84.5%	9.8%	90.20%

In the context of dry bean production and distribution, the presence of other dry bean varieties in a batch of a certain dry bean variety is an undesirable scenario as it results in lower market value (Koklu & Ozkan, 2020; Ali et al., 2020). Therefore, if a classification model is used to automatically classify dry beans, the number of false positives within a batch should be minimum. This can be assessed using the precision and specificity of the model. The best model has a healthy precision of 88%. This means that if a batch of 100 dry beans of a particular variety is prepared based on the predictions of the best model, 88 are expected to actually be of that variety while 12 are expected to be of other varieties (i.e. false positives). The models are also highly specific, as evidenced by the high specificity (i.e. true negative %) rates of 94.3% to 95.2%. High specificity suggests a small number of false positives, which is desirable. Taking the Seker dry bean variety as an example, a specificity of 95.2% for the best model suggests a high probability that other dry bean varieties would indeed be classified as non-Seker and would not end up being falsely classified as the Seker variety.

Also crucial in dry bean cultivation is that the yield is maximized and the waste from incorrect classifications is minimized (Pinto et al., 2017). This can be assessed to a large extent using the sensitivity (i.e. true positive rate) of the model. The best model has a sensitivity of 87%. This means that out of 100 dry beans of a particular variety, the best model would be able to classify 87 of them correctly into batches designated for that variety, but incorrectly classify 13 as other varieties. This means that 13% of the yield of a particular variety would not end up being classified into the batches for that variety. That said, a precision of 87% across the board provided by an automated system is still much better than manual classification. The models are also compared and evaluated based on F1-Scores as shown in Table 8. The F1-Score is the harmonic mean of the precision and sensitivity of a model meaning it considers both false positives and false negatives. Moreover, the F1-Score is a useful performance metric especially for irregular class distributions such as in this case (Ye, 2013). The F1-Score of 87.7% indicates a good balance between precision and sensitivity. Figure 4 below shows the key model performance metrics for the best model. Overall, the model has high values for all metrics but the strength of the model particularly lies in its high specificity and high accuracy rates.

Figure 4. Model Performance Metrics for Polynomial Regression (Best Model)

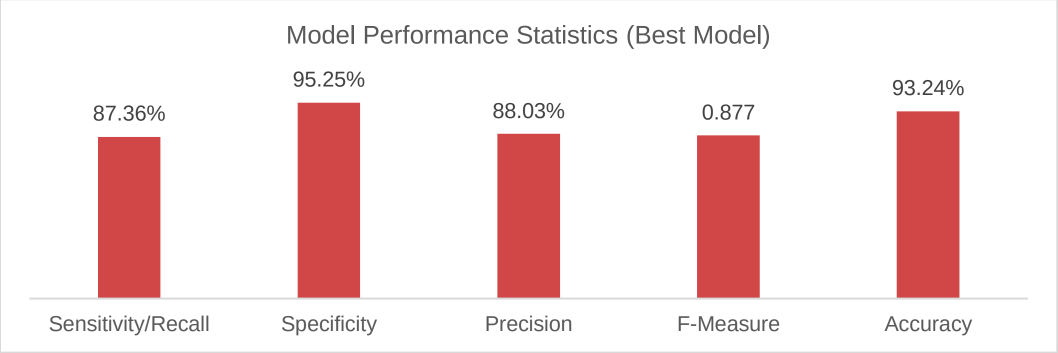


Figure 5. Accuracy by Bean Variety Based on Polynomial Regression (Best Model)

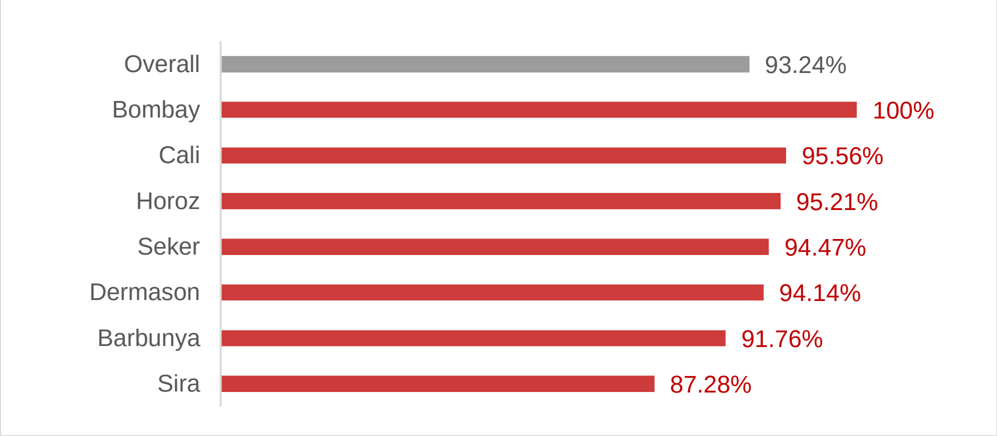
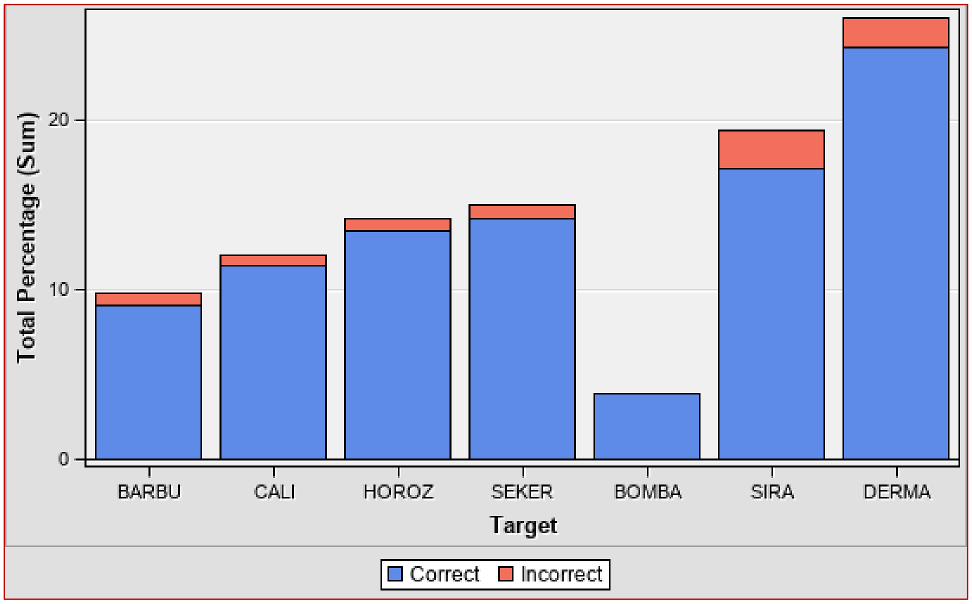


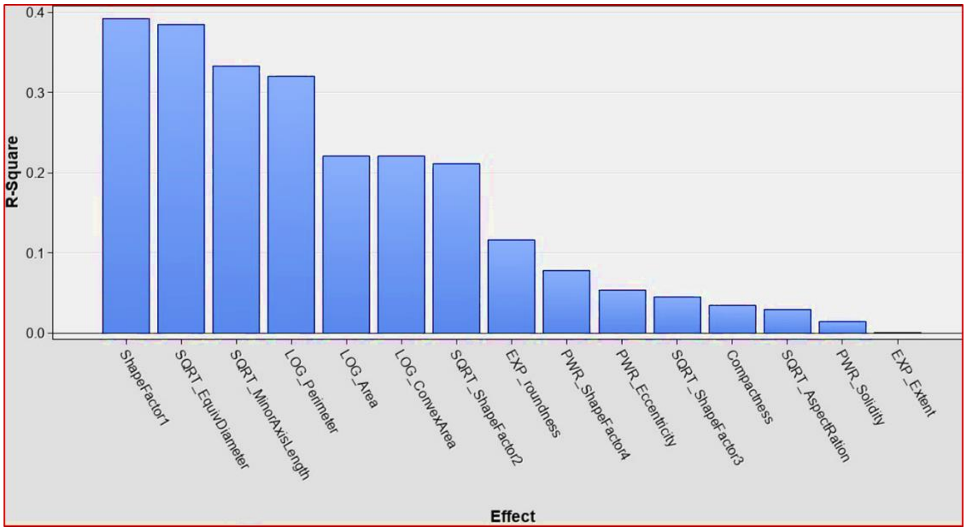
Figure 5 above shows the accuracy of classification for each dry bean variety. Bombay beans are the easiest to fully classify as evidenced by the accuracy of 100%. On the other hand, Sira beans are the hardest to classify with an accuracy rate of 87.28%, making them the only variety with a sub-90% accuracy rate. As noted in Table 1 earlier, Sira beans are the smallest in size among the seven beans whereas Bombay beans are very large in size. Given that physical features based on processed images are used as inputs to classify the beans, the relative size of the dry bean varieties could be a potential factor that affects classification accuracy.

Figure 6. Classification Chart for Polynomial Regression (Best Model)



Based on the classification chart in Figure 6, the best model has no issues in classifying all the dry bean varieties with roughly equal misclassification rates despite the data having slightly disproportional classes. For Bombay beans, there were no incorrectly classified samples.

Figure 7. Variable Importance Plot for Polynomial Regression (Best Model)



The figure above shows which physical features are most important in differentiating the dry bean varieties from each other. The top five most important features include ShapeFactor1, EquivDiameter, MinorAxisLength, Perimeter, and Area for the polynomial logistic regression model (best model). Three of these variables, namely: MinorAxisLength, ShapeFactor1, and Perimeter, were also observed to be the most important variables in splitting the Decision Tree (See Appendix D for the complete decision tree). Therefore, these variables should be prioritized in situations where decisions must be made based on limited data availability.

4. DISCUSSION AND CONCLUSIONS

The results have shown that it is possible to accurately classify the seven dry bean varieties based on their physical features using automated machine learning techniques. Ten different machine learning models were employed and compared, namely: standard logistic regression, polynomial logistic regression, decision tree, random forest, SVM, LDA, QDA, neural network, gradient boosting, and ensemble. The results revealed high accuracy rates for all models - ranging from 90.2% to 93.24%. In addition to classification accuracy, all the models performed well based on performance metrics such as ASE, ROC Index, sensitivity, specificity, precision, and F1-score. Polynomial logistic regression was the best model with the highest accuracy rate of 93.24% and lowest misclassification rate of 6.76% based on a 70/30 train-validation split. This best model was able to marginally improve the accuracy of 93.19% obtained by the original authors, Koklu and Ozkan (2020). On the other hand, the decision tree which has the easiest interpretability unfortunately had the poorest accuracy rate. It was also noted that nonlinear techniques such as polynomial logistic regression and quadratic discriminant analysis had higher accuracy rates.

Generally, the models were able to classify the fairly imbalanced and disproportional bean classes with high accuracy. The accuracy rates of the seven dry bean varieties: Sira, Barbunya, Dermason, Seker, Horoz, Cali, and Bombay were 87.28%, 91.76%, 94.14%, 94.47%, 95.21%, 95.56%, and 100% respectively. Based on these results, the Bombay dry bean variety can be fully classified with a 100% accuracy whereas the Sira variety is most difficult to classify (87.28% accuracy). Moreover, the results suggest that Physical 2D attributes alone can be used to accurately classify dry beans with relatively high accuracy. Specifically, the most important variables for differentiating the dry bean varieties were found to be ShapeFactor1, EquivDiameter, MinorAxisLength, Area, and Perimeter for the best performing model. Overall, given the high classification accuracy of 93.24% for the best model, it therefore makes sense for dry bean producers to invest in automated systems for classifying dry beans as opposed to manual sorting.

While the results have provided several useful insights, the analysis can be enhanced in many ways. First, the current data is based on physical attributes derived from 2D images. If additional variables that measure the third dimension (i.e. depth) of the beans are included, the classification accuracy can be potentially increased. Moreover, the possibility of adding variables describing the texture of the beans should also be explored. Second, alternative validation methods such as 10-fold cross-validation can be used to compare and verify the results obtained with the 70%-train, 30%-validation split used in this analysis. Finally, modeling the tangible lost dollars from incorrect classifications of dry beans using profit/loss matrices would serve as a key input for producers and farmers to assess the financial viability of investing in automated systems. Future enhancements to the analysis should be hinged on the premise that the degree to which the automated classification of dry beans increases the seed quality, market value, yield, and profitability of dry beans within the agriculture industry ultimately determines model success.

REFERENCES

Agrawal, D., & Dahiya, P. (2018). Comparisons of classification algorithms on seeds dataset using machine learning algorithm. Compusoft, 7(5), 2760–2765. https://doi.org/10.6084/ijact.v7i5.720

Ajaz, R. H., Hussain, L., Jammu, A., & Muzaffarabad, K. (2015). Seed Classification using Machine Learning Techniques. Journal of Multidisciplinary Engineering Science and Technology, 2(5), 3159–40. Retrieved from www.jmest.org

Ali, A., Qadri, S., Mashwani, W. K., Brahim Belhaouari, S., Naeem, S., Rafique, S., … Anam, S. (2020). Machine learning approach for the classification of corn seed using hybrid features. International Journal of Food Properties, 23(1), 1097–1111.

Ayele, N. A., & Tamiru, H. K. (2020). Developing Classification Model for Chickpea Types using Machine Learning Algorithms. International Journal of Innovative Technology and Exploring Engineering, 10(1), 5–11. https://doi.org/10.35940/ijitee.a8057.1110120

Basol, Y., & Toklu, S. (2021). A Deep Learning-Based Seed Classification with Mobile Application. Turkish Journal of Mathematics and Computer Science, 13(1), 192–203. https://doi.org/10.47000/tjmcs.897631

Fukai, H., Furukawa, J., Katsuragawa, H., & Pinto, C. (2018). Classification of Green Coffee Beans by Convolutional Neural Network and its Implementation on Raspberry Pi and Camera Module. Timorese Academic Journal of Science and Technology, 1, 10.

Koklu, M., & Ozkan, I. A. (2020). Multiclass classification of dry beans using computer vision and machine learning techniques. Computers and Electronics in Agriculture, 174, 105507

Medeiros, A. D. de, Pinheiro, D. T., Xavier, W. A., Silva, L. J. da, & Dias, D. C. F. dos S. (2020). Quality classification of Jatropha curcas seeds using radiographic images and machine learning. Industrial Crops and Products, 146(September 2019), 112162. https://doi.org/10.1016/j.indcrop.2020.112162

Pérez-Rodríguez, M., Gaiad, J. E., Hidalgo, M. J., Avanza, M. V., & Pellerano, R. G. (2019). Classification of cowpea beans using multielemental fingerprinting combined with supervised learning. Food Control, 95, 232–241.

Pinto, C., Furukawa, J., Fukai, H., & Tamura, S. (2017). Classification of Green coffee bean images based on defect types using convolutional neural network (CNN). 2017 International Conference on Advanced Informatics, Concepts, Theory, and Applications (ICAICTA). doi:10.1109/icaicta.2017.8090980

Ye, N. (2013). The Handbook of Data Mining. Routledge.

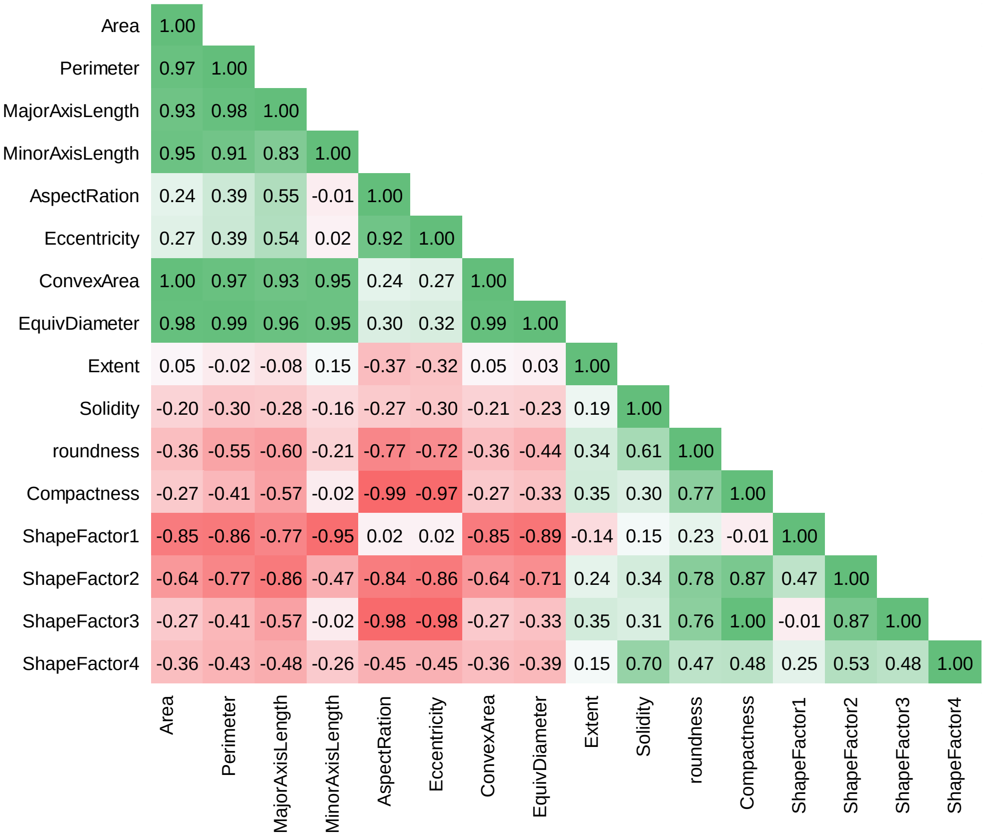
APPENDICES

Appendix A – Descriptive Statistics and Correlation Matrix of Features

Table 9. Descriptive Statistics of Feature Variables (Measured in Pixels)

Variable	Mean	Std. Deviation	Minimum	Maximum	N
Area	53048.3	29324.1	20420.0	254616.0	13611
Perimeter	855.28	214.29	524.74	1985.37	13611
MajorAxisLength	320.14	85.69	183.60	738.86	13611
MinorAxisLength	202.27	44.97	122.51	460.20	13611
AspectRation	1.58	0.25	1.02	2.43	13611
Eccentricity	0.75	0.09	0.22	0.91	13611
ConvexArea	53768	29775	20684	263261	13611
EquivDiameter	253.06	59.18	161.24	569.37	13611
Extent	0.75	0.05	0.56	0.87	13611
Solidity	0.99	0.00	0.92	0.99	13611
Roundness	0.87	0.06	0.49	0.99	13611
Compactness	0.80	0.06	0.64	0.99	13611
ShapeFactor1	0.01	0.00	0.00	0.01	13611
ShapeFactor2	0.00	0.00	0.00	0.00	13611
ShapeFactor3	0.64	0.10	0.41	0.97	13611
ShapeFactor4	1.00	0.00	0.95	1.00	13611

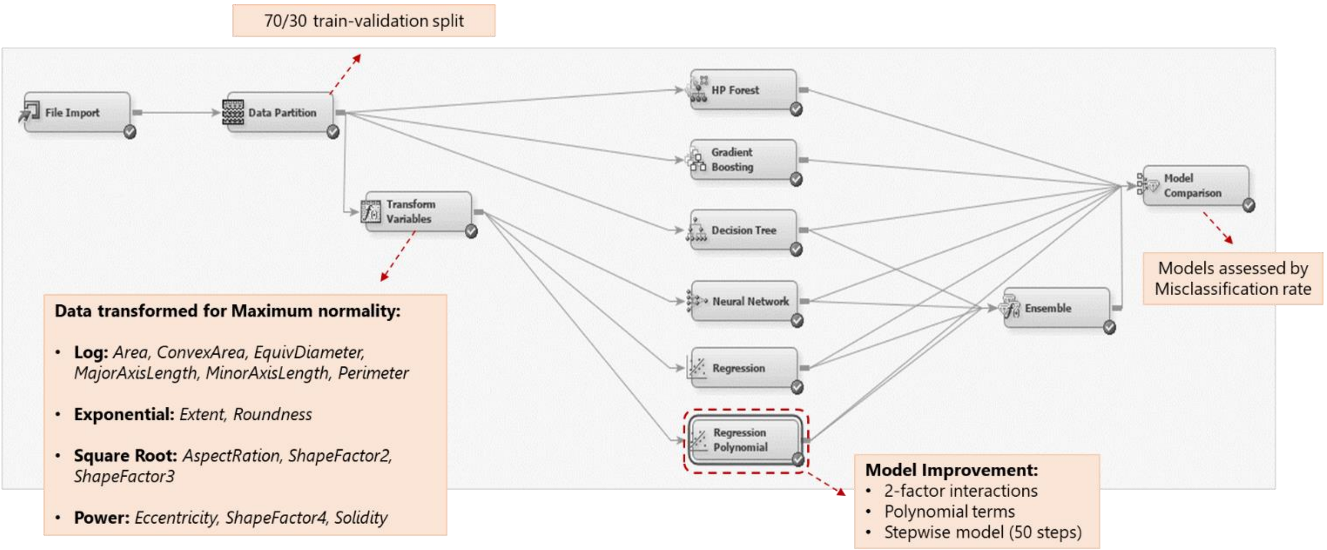
Figure 8. Correlation Matrix of Feature Variables



	Area	Perimeter	MajorAxisLength	MinorAxisLength	AspectRation	Eccentricity	ConvexArea	EquivDiameter	Extent	Solidity	roundness	Compactness	ShapeFactor1	ShapeFactor2	ShapeFactor3	ShapeFactor4
Area	1.00															
Perimeter	0.97	1.00														
MajorAxisLength	0.93	0.98	1.00													
MinorAxisLength	0.95	0.91	0.83	1.00												
AspectRation	0.24	0.39	0.55	-0.01	1.00											
Eccentricity	0.27	0.39	0.54	0.02	0.92	1.00										
ConvexArea	1.00	0.97	0.93	0.95	0.24	0.27	1.00									
EquivDiameter	0.98	0.99	0.96	0.95	0.30	0.32	0.99	1.00								
Extent	0.05	-0.02	-0.08	0.15	-0.37	-0.32	0.05	0.03	1.00							
Solidity	-0.20	-0.30	-0.28	-0.16	-0.27	-0.30	-0.21	-0.23	0.19	1.00						
roundness	-0.36	-0.55	-0.60	-0.21	-0.77	-0.72	-0.36	-0.44	0.34	0.61	1.00					
Compactness	-0.27	-0.41	-0.57	-0.02	-0.99	-0.97	-0.27	-0.33	0.35	0.30	0.77	1.00				
ShapeFactor1	-0.85	-0.86	-0.77	-0.95	0.02	0.02	-0.85	-0.89	-0.14	0.15	0.23	-0.01	1.00			
ShapeFactor2	-0.64	-0.77	-0.86	-0.47	-0.84	-0.86	-0.64	-0.71	0.24	0.34	0.78	0.87	0.47	1.00		
ShapeFactor3	-0.27	-0.41	-0.57	-0.02	-0.98	-0.98	-0.27	-0.33	0.35	0.31	0.76	1.00	-0.01	0.87	1.00	
ShapeFactor4	-0.36	-0.43	-0.48	-0.26	-0.45	-0.45	-0.36	-0.39	0.15	0.70	0.47	0.48	0.25	0.53	0.48	1.00

Appendix B – SAS EM Process Flow and SAS Base Code for Models

SAS EM Process flow:



70/30 train-validation split

Data transformed for Maximum normality:

- **Log:** *Area, ConvexArea, EquivDiameter, MajorAxisLength, MinorAxisLength, Perimeter*
- **Exponential:** *Extent, Roundness*
- **Square Root:** *AspectRation, ShapeFactor2, ShapeFactor3*
- **Power:** *Eccentricity, ShapeFactor4, Solidity*

Models assessed by Misclassification rate

Model Improvement:

- 2-factor interactions
- Polynomial terms
- Stepwise model (50 steps)

SAS Base Code for Linear and Quadratic Discriminant Analysis Models:

```
/*Import datasets*/
proc import datafile="C:\Users\user\Google Drive\School\Master of Analytics\161.777 Practical Data
Mining\Final Project\Dry_bean.csv"
out=drybean dbms=csv replace; getnames=yes; run;

proc import datafile="C:\Users\user\Google Drive\School\Master of Analytics\161.777 Practical Data
Mining\Final Project\Dry_beantrain.csv"
out=drybeantrain dbms=csv replace; getnames=yes; run;

proc import datafile="C:\Users\user\Google Drive\School\Master of Analytics\161.777 Practical Data
Mining\Final Project\Dry_beantest.csv"
out=drybeantest dbms=csv replace; getnames=yes; run;

/*Stepwise Discriminant Analysis to determine important variables*/
proc stepdisc data=drybeantrain method=stepwise;
class class;
var Area--ShapeFactor4;
run;

/*Fisher LDA*/
proc discrim data=drybeantrain pool=yes
testdata=drybeantest testlisterr;
class class;
priors prop;
var &_STDVAR;
run;

/*Quadratic DA*/
proc discrim data=drybeantrain pool=test
testdata=drybeantest testlisterr slpool=.05;
title1 'Test for equality of covariance matrices';
title2 'and quadratic discriminant analysis';
class class;
priors prop;
var &_STDVAR;
run; title;
```

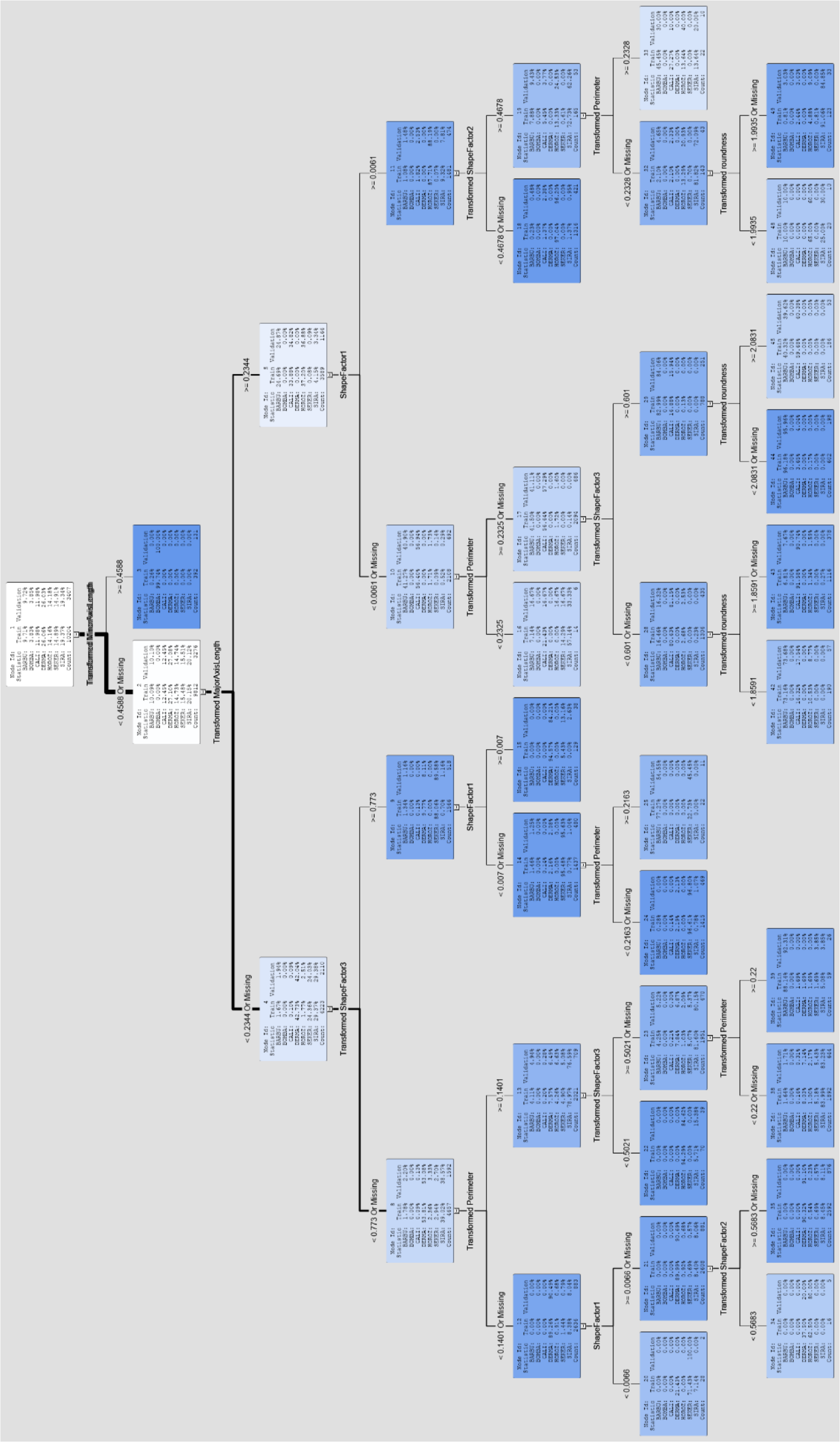
Appendix C – Results of Stepwise Polynomial Regression with Interactions (Best Model)

```
                                                 Analysis of Maximum Likelihood Estimates

                                                                            Standard       Wald                     Standardized
Parameter                                        Class    DF    Estimate      Error    Chi-Square    Pr > ChiSq      Estimate       Exp(Est)

Intercept                                        SIRA       1     76.4878     25.2245          9.19        0.0024                     999.000
Intercept                                        SEKER      1       151.2     28.0203         29.10        <.0001                     999.000
Intercept                                        HOROZ      1       102.3     23.0474         19.70        <.0001                     999.000
Intercept                                        DERMA      1       133.6     27.0527         24.38        <.0001                     999.000
Intercept                                        CALI       1    -22.5888     27.6257          0.67        0.4135                       0.000
Intercept                                        BOMBA      1     39.3077      1277.3          0.00        0.9754                     999.000
EXP_roundness                                    SIRA       1    -72.7067     21.8835         11.04        0.0009         -9.9643       0.000
EXP_roundness                                    SEKER      1      -114.5     23.9743         22.80        <.0001        -15.6894       0.000
EXP_roundness                                    HOROZ      1    -73.9384     20.9353         12.47        0.0004        -10.1331       0.000
EXP_roundness                                    DERMA      1    -90.6077     24.8614         13.28        0.0003        -12.4176       0.000
EXP_roundness                                    CALI       1    -50.2874     23.0719          4.75        0.0293         -6.8918       0.000
EXP_roundness                                    BOMBA      1    -71.8027      1100.8          0.00        0.9480         -9.8404       0.000
LOG_ConvexArea                                   SIRA       1      -184.1       123.6          2.22        0.1365         -9.6874       0.000
LOG_ConvexArea                                   SEKER      1    -88.7530       134.2          0.44        0.5083         -4.6713       0.000
LOG_ConvexArea                                   HOROZ      1    -90.8922       111.9          0.66        0.4166         -4.7839       0.000
LOG_ConvexArea                                   DERMA      1      -512.4       154.9         10.95        0.0009        -26.9702       0.000
LOG_ConvexArea                                   CALI       1       278.8     91.7414          9.24        0.0024         14.6757     999.000
LOG_ConvexArea                                   BOMBA      1      -322.7      1586.6          0.04        0.8388        -16.9855       0.000
EXP_Extent*EXP_Extent                            SIRA       1      1.6315      1.4508          1.26        0.2608                       5.111
EXP_Extent*EXP_Extent                            SEKER      1      3.7820      1.7508          4.67        0.0308                      43.904
EXP_Extent*EXP_Extent                            HOROZ      1      4.4987      1.4229         10.00        0.0016                      89.904
EXP_Extent*EXP_Extent                            DERMA      1      2.0761      1.5566          1.78        0.1823                       7.973
EXP_Extent*EXP_Extent                            CALI       1     -2.9580      1.0564          7.84        0.0051                       0.052
EXP_Extent*EXP_Extent                            BOMBA      1      0.5214     13.2008          0.00        0.9685                       1.684
EXP_Extent*ShapeFactor1                          SIRA       1     -1246.8       904.5          1.90        0.1681                       0.000
EXP_Extent*ShapeFactor1                          SEKER      1     -2747.8      1089.2          6.36        0.0116                       0.000
EXP_Extent*ShapeFactor1                          HOROZ      1     -2797.1       888.5          9.91        0.0016                       0.000
EXP_Extent*ShapeFactor1                          DERMA      1     -1725.0       952.5          3.28        0.0701                       0.000
EXP_Extent*ShapeFactor1                          CALI       1      2091.4       717.6          8.49        0.0036                     999.000
EXP_Extent*ShapeFactor1                          BOMBA      1      -527.4      9608.3          0.00        0.9562                       0.000
EXP_roundness*EXP_roundness                      SIRA       1     19.8204      5.7964         11.69        0.0006                     999.000
EXP_roundness*EXP_roundness                      SEKER      1     30.1965      6.5161         21.48        <.0001                     999.000
EXP_roundness*EXP_roundness                      HOROZ      1     21.1801      5.5594         14.51        0.0001                     999.000
EXP_roundness*EXP_roundness                      DERMA      1     22.6315      6.9288         10.67        0.0011                     999.000
EXP_roundness*EXP_roundness                      CALI       1     18.2931      5.6657         10.42        0.0012                     999.000
EXP_roundness*EXP_roundness                      BOMBA      1     22.9992       251.6          0.01        0.9272                     999.000
EXP_roundness*LOG_ConvexArea                     SIRA       1       294.2     59.3996         24.53        <.0001                     999.000
EXP_roundness*LOG_ConvexArea                     SEKER      1       255.6     69.6732         13.46        0.0002                     999.000
EXP_roundness*LOG_ConvexArea                     HOROZ      1       108.3     47.8602          5.12        0.0237                     999.000
EXP_roundness*LOG_ConvexArea                     DERMA      1       177.1     75.8664          5.45        0.0196                     999.000
EXP_roundness*LOG_ConvexArea                     CALI       1       252.6     43.3022         34.02        <.0001                     999.000
EXP_roundness*LOG_ConvexArea                     BOMBA      1       187.8       643.2          0.09        0.7703                     999.000
EXP_roundness*PWR_Eccentricity                   SIRA       1     -9.0946      5.7620          2.49        0.1145                       0.000
EXP_roundness*PWR_Eccentricity                   SEKER      1    -26.7967      6.6172         16.40        <.0001                       0.000
EXP_roundness*PWR_Eccentricity                   HOROZ      1      3.3551      5.7695          0.34        0.5609                      28.650
EXP_roundness*PWR_Eccentricity                   DERMA      1    -10.0718      6.0376          2.78        0.0953                       0.000
EXP_roundness*PWR_Eccentricity                   CALI       1     -7.7998      5.0008          2.43        0.1188                       0.000
EXP_roundness*PWR_Eccentricity                   BOMBA      1      8.3504       153.3          0.00        0.9566                     999.000
EXP_roundness*PWR_Solidity                       SIRA       1    -48.3416     13.9118         12.07        0.0005                       0.000
EXP_roundness*PWR_Solidity                       SEKER      1    -47.1114     17.0624          7.62        0.0058                       0.000
EXP_roundness*PWR_Solidity                       HOROZ      1    -41.8072     12.8501         10.58        0.0011                       0.000
EXP_roundness*PWR_Solidity                       DERMA      1     -9.5823     17.5418          0.30        0.5849                       0.000
EXP_roundness*PWR_Solidity                       CALI       1    -81.3006     12.9285         39.55        <.0001                       0.000
EXP_roundness*PWR_Solidity                       BOMBA      1    -79.1607       266.1          0.09        0.7661                       0.000
LOG_EquivDiameter*LOG_MinorAxisLength            SIRA       1      -921.1       229.3         16.14        <.0001                       0.000
LOG_EquivDiameter*LOG_MinorAxisLength            SEKER      1      -579.7       210.6          7.57        0.0059                       0.000
LOG_EquivDiameter*LOG_MinorAxisLength            HOROZ      1      -101.8       208.2          0.24        0.6249                       0.000
LOG_EquivDiameter*LOG_MinorAxisLength            DERMA      1      -141.2       248.3          0.32        0.5697                       0.000
LOG_EquivDiameter*LOG_MinorAxisLength            CALI       1      -854.2       151.4         31.83        <.0001                       0.000
LOG_EquivDiameter*LOG_MinorAxisLength            BOMBA      1       279.1      1888.3          0.02        0.8825                     999.000
LOG_MajorAxisLength*SQRT_ShapeFactor3            SIRA       1     29.5811     41.8408          0.50        0.4796                     999.000
LOG_MajorAxisLength*SQRT_ShapeFactor3            SEKER      1      -162.6     46.7291         12.11        0.0005                       0.000
LOG_MajorAxisLength*SQRT_ShapeFactor3            HOROZ      1      -198.1     38.1112         27.03        <.0001                       0.000
LOG_MajorAxisLength*SQRT_ShapeFactor3            DERMA      1     -5.8436     52.3125          0.01        0.9111                       0.003
LOG_MajorAxisLength*SQRT_ShapeFactor3            CALI       1      -102.7     36.9331          7.73        0.0054                       0.000
LOG_MajorAxisLength*SQRT_ShapeFactor3            BOMBA      1      -155.7       492.0          0.10        0.7516                       0.000
PWR_ShapeFactor4*PWR_ShapeFactor4                SIRA       1     -6.2376      3.3247          3.52        0.0606                       0.002
PWR_ShapeFactor4*PWR_ShapeFactor4                SEKER      1     11.7637      4.4574          6.97        0.0083                     999.000
PWR_ShapeFactor4*PWR_ShapeFactor4                HOROZ      1      2.6055      3.6560          0.51        0.4761                      13.538
PWR_ShapeFactor4*PWR_ShapeFactor4                DERMA      1     -4.6894      3.7689          1.55        0.2134                       0.009
PWR_ShapeFactor4*PWR_ShapeFactor4                CALI       1     -2.4343      3.5628          0.47        0.4944                       0.088
PWR_ShapeFactor4*PWR_ShapeFactor4                BOMBA      1     -1.4351     91.9484          0.00        0.9875                       0.238
PWR_ShapeFactor4*SQRT_AspectRation               SIRA       1     -7.3665      6.6684          1.22        0.2693                       0.001
PWR_ShapeFactor4*SQRT_AspectRation               SEKER      1    -28.9528     11.7475          6.07        0.0137                       0.000
PWR_ShapeFactor4*SQRT_AspectRation               HOROZ      1    -25.5158      6.6237         14.84        0.0001                       0.000
PWR_ShapeFactor4*SQRT_AspectRation               DERMA      1     -2.3681      8.1058          0.09        0.7702                       0.094
PWR_ShapeFactor4*SQRT_AspectRation               CALI       1    -19.7073      6.9748          7.98        0.0047                       0.000
PWR_ShapeFactor4*SQRT_AspectRation               BOMBA      1    -23.1565       191.5          0.01        0.9037                       0.000
PWR_Solidity*PWR_Solidity                        SIRA       1     26.6167      9.3817          8.05        0.0046                     999.000
PWR_Solidity*PWR_Solidity                        SEKER      1     57.0931     14.0450         16.52        <.0001                     999.000
PWR_Solidity*PWR_Solidity                        HOROZ      1     25.5479      8.5236          8.98        0.0027                     999.000
PWR_Solidity*PWR_Solidity                        DERMA      1     20.5704     12.5428          2.69        0.1010                     999.000
PWR_Solidity*PWR_Solidity                        CALI       1     31.9515      7.8501         16.57        <.0001                     999.000
PWR_Solidity*PWR_Solidity                        BOMBA      1     32.3482       223.7          0.02        0.8850                     999.000
PWR_Solidity*SQRT_ShapeFactor3                   SIRA       1      1.7567     22.0122          0.01        0.9364                       5.793
PWR_Solidity*SQRT_ShapeFactor3                   SEKER      1    -63.3384     25.0415          6.40        0.0114                       0.000
PWR_Solidity*SQRT_ShapeFactor3                   HOROZ      1    -44.8152     22.0526          4.13        0.0421                       0.000
PWR_Solidity*SQRT_ShapeFactor3                   DERMA      1    -77.4166     24.7596          9.78        0.0018                       0.000
PWR_Solidity*SQRT_ShapeFactor3                   CALI       1     79.5773     20.7785         14.67        0.0001                     999.000
PWR_Solidity*SQRT_ShapeFactor3                   BOMBA      1     63.3614       557.5          0.01        0.9095                     999.000
PWR_Solidity*ShapeFactor1                        SIRA       1     13156.9      2239.6         34.51        <.0001                     999.000
PWR_Solidity*ShapeFactor1                        SEKER      1     14831.6      3146.0         22.23        <.0001                     999.000
PWR_Solidity*ShapeFactor1                        HOROZ      1     14948.8      2102.0         50.58        <.0001                     999.000
PWR_Solidity*ShapeFactor1                        DERMA      1      8998.5      2511.5         12.84        0.0003                     999.000
PWR_Solidity*ShapeFactor1                        CALI       1     16888.8      2242.2         56.73        <.0001                     999.000
PWR_Solidity*ShapeFactor1                        BOMBA      1     18354.6     41116.8          0.20        0.6553                     999.000
```

Appendix D – Classification Decision Tree


Transformed MinorAxisLength
< 0.4588 Or Missing
>= 0.4588
Transformed MajorAxisLength
< 0.2344 Or Missing
>= 0.2344
Transformed ShapeFactor3
ShapeFactor1
< 0.773 Or Missing
>= 0.773
< 0.0061 Or Missing
>= 0.0061
Transformed Perimeter
ShapeFactor1
Transformed Perimeter
Transformed ShapeFactor2
< 0.1401 Or Missing
>= 0.1401
< 0.007 Or Missing
>= 0.007
< 0.2325
>= 0.2325 Or Missing
< 0.4678 Or Missing
>= 0.4678
ShapeFactor1
Transformed ShapeFactor3
Transformed Perimeter
Transformed ShapeFactor3
Transformed Perimeter
< 0.0066
>= 0.0066 Or Missing
< 0.5021
>= 0.5021 Or Missing
< 0.2163 Or Missing
>= 0.2163
< 0.601 Or Missing
>= 0.601
< 0.2328 Or Missing
>= 0.2328
Transformed ShapeFactor2
Transformed Perimeter
Transformed roundness
Transformed roundness
Transformed roundness
< 0.5683
>= 0.5683 Or Missing
< 0.22 Or Missing
>= 0.22
< 1.8591
>= 1.8591 Or Missing
< 2.0831 Or Missing
>= 2.0831
< 1.9935
>= 1.9935 Or Missing