Dry Bean MultiClass Classification

Ryan Chui

Abstract

The aim of this project is to develop a supervised machine learning algorithm to perform a multi-classification of dry beans species harvested from population cultivation from a single farm. The dataset contains images of 13,611 grains of 7 different registered dry beans were taken by a high-resolution camera. A total of 16 features; 12 dimension and 4 shape forms, were obtained from the grains. Support Vector Machine (SVM), k-Nearest Neighbors (kNN), logistic regression, Decision Trees (DT) were performed and compared. Overall correct classification rates have been determined as 92.1% 92.4%, 91.4%, 92.7% for logistic regression, kNN, DT and SVM respectively. The SVM classification model, which has the highest accuracy results, has classified the **Barbunya, Bombay, Cali, Dermason, Horoz, Seker and Sira bean** varieties with 95.7%, 100%, 92.9%, 93.1%, 95.3%, 96.9%, 89.6%, respectively. With these results, the demands of the producers and the customers are largely met about obtaining uniform bean varieties.

Motivation

There is a wide range of genetic diversity of dry beans, which is the most produced among the edible legume crops globally. Seed quality is influential in crop production. Therefore, seed classification is essential for both marketing and production to provide the principles of sustainable agricultural systems. The primary objective of this study is to provide a method for obtaining uniform seed varieties from crop production, which is in the form of population, so the seeds are not certified as a sole variety.

Dataset(s) from UC Irvine ML Reprosity

16 Features (12 dimensions and 4 shape forms):

- Area (A): The area of a bean zone and the number of pixels within its boundaries.
- Perimeter (P): Bean circumference is defined as the length of its border.
- Major axis length (L): The distance between the ends of the longest line that can be drawn from a bean.
- Minor axis length (l): The longest line that can be drawn from the bean while standing perpendicular to the main axis.
- Aspect ratio (K): Defines the relationship between L and l.
- Eccentricity (Ec): Eccentricity of the ellipse having the same moments as the region.
- Convex area (C): Number of pixels in the smallest convex polygon that can contain the area of a bean seed.
- Equivalent diameter (Ed): The diameter of a circle having the same area as a bean seed area.
- Extent (Ex): The ratio of the pixels in the bounding box to the bean area.
- Solidity (S): Also known as convexity. The ratio of the pixels in the convex shell to those found in beans.
- Roundness (R): Calculated with the following formula: (4piA)/(P^2)
- Compactness (CO): Measures the roundness of an object: Ed/L
- ShapeFactor1 (SF1)
- ShapeFactor2 (SF2)
- ShapeFactor3 (SF3)
- ShapeFactor4 (SF4)
- Class (**Seker, Barbunya, Bombay, Cali, Dermosan, Horoz and Sira**)

Dataset: https://archive.ics.uci.edu/ml/datasets/Dry+Bean+Dataset

Data Preparation and Cleaning

This dataset is collected using a computer vision system that extracted shape features from beans images.

Steps include:

- Exploring the dataset, getting summary statistics and checking for null values and duplicates and there weren't any.
- Plot and check for class of imbalance, but plot of distribution suggest there is no affect and no need to handle.
- Splitted with 70/30 ratio rule to training and testing sets respectively, thus a 30% of the data is used for final testing and 70% will use the training set for train and validation.

The dataset is clean, no missing values or duplicated entries were found. A challenge is to explain the boxplot since the “area” of the boxplot, would be defined as the number of pixels that fit into the zone taken up by an image of a bean. Additionally, the geometrical data carry no information about the bean colour; it is unfortunate, as different dry bean species tend to vary in colour.

Research Question(s)

What is your research question you aim to answer using the dataset? Be sure the research question is well defined (see project description for details).

Given a set of features extracted from the shape of the beans in images, predict the type of each bean from 7 bean types.

What are some linearly correlated features between labels?

Can we implement and compare the performance of Logistic Regression (LR) and K-Nearest Neighbor (KNN), Decision Tree and SVM classification model?

Methods

Models were prepared by using SVM, kNN, Decision Tree, and Logistic Regression in order to make classification according to features obtained from dry beans. The aforementioned models can help determine which features belongs to which class

SVM : A kernel based method for classification and regression problems, and has better generalizations.

KNN : A classification method that classifies objects based on the given k samples according to the class of the nearest neighbor. (n = 2, p = 2)

Decision tree : Uses a tree-like graphic or model to show its decisions and possible result. In this study, kfold divided the dataset into 4.

Logistic Regression : Models the probabilities for classification problems with two possible outcomes.

Findings

Sample Distribution of all types of dry beans

	Class	class_encoder	count	percent	cumulative_count	cumulative_percent
0	DERMASON	3	3546	26.052458	3546	26.052458
1	SIRA	6	2636	19.366689	6182	45.419146
2	SEKER	5	2027	14.892366	8209	60.311513
3	HOROZ	4	1928	14.165014	10137	74.476526
4	CALI	2	1630	11.975608	11767	86.452134
5	BARBUNYA	0	1322	9.712732	13089	96.164867
6	BOMBAY	1	522	3.835133	13611	100.000000

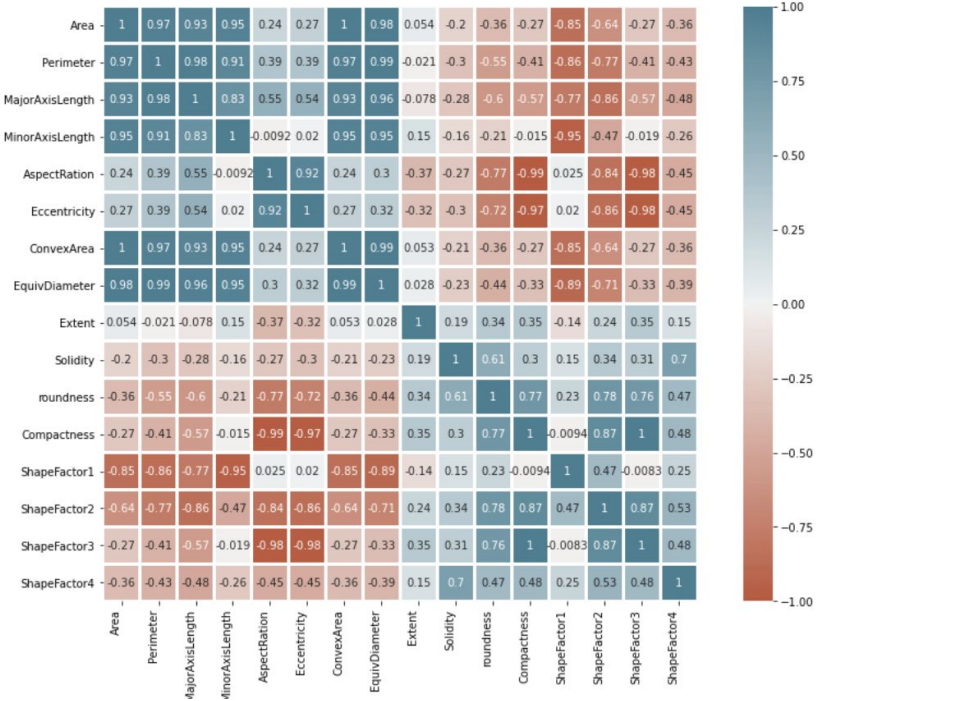
Statistical distribution of features of dry bean varieties (in pixels)

	count	mean	std	min	25%	50%	75%	max
Area	13611.0	53048.284549	29324.095717	20420.000000	36328.000000	44652.000000	61332.000000	254616.000000
Perimeter	13611.0	855.283459	214.289696	524.736000	703.523500	794.941000	977.213000	1985.370000
MajorAxisLength	13611.0	320.141867	85.694186	183.601165	253.303633	296.883367	376.495012	738.860153
MinorAxisLength	13611.0	202.270714	44.970091	122.512653	175.848170	192.431733	217.031741	460.198497
AspectRation	13611.0	1.583242	0.246678	1.024868	1.432307	1.551124	1.707109	2.430306
Eccentricity	13611.0	0.750895	0.092002	0.218951	0.715928	0.764441	0.810466	0.911423
ConvexArea	13611.0	53768.200206	29774.915817	20684.000000	36714.500000	45178.000000	62294.000000	263261.000000
EquivDiameter	13611.0	253.064220	59.177120	161.243764	215.068003	238.438026	279.446467	569.374358
Extent	13611.0	0.749733	0.049086	0.555315	0.718634	0.759859	0.786851	0.866195
Solidity	13611.0	0.987143	0.004660	0.919246	0.985670	0.988283	0.990013	0.994677
roundness	13611.0	0.873282	0.059520	0.489618	0.832096	0.883157	0.916869	0.990685
Compactness	13611.0	0.799864	0.061713	0.640577	0.762469	0.801277	0.834270	0.987303
ShapeFactor1	13611.0	0.006564	0.001128	0.002778	0.005900	0.006645	0.007271	0.010451
ShapeFactor2	13611.0	0.001716	0.000596	0.000564	0.001154	0.001694	0.002170	0.003665
ShapeFactor3	13611.0	0.643590	0.098996	0.410339	0.581359	0.642044	0.696006	0.974767
ShapeFactor4	13611.0	0.995063	0.004366	0.947687	0.993703	0.996386	0.997883	0.999733

- The table shows the quantity distribution of dry bean samples in the study. “DERMASON” has the highest instance for class, while the least amount of instance is “BOMBAY”, suggesting The features are positive skewed or negative skewed.
- The minimum, maximum, mean and standard deviation data of the features obtained for all dry bean samples are displayed in the right table.

Findings

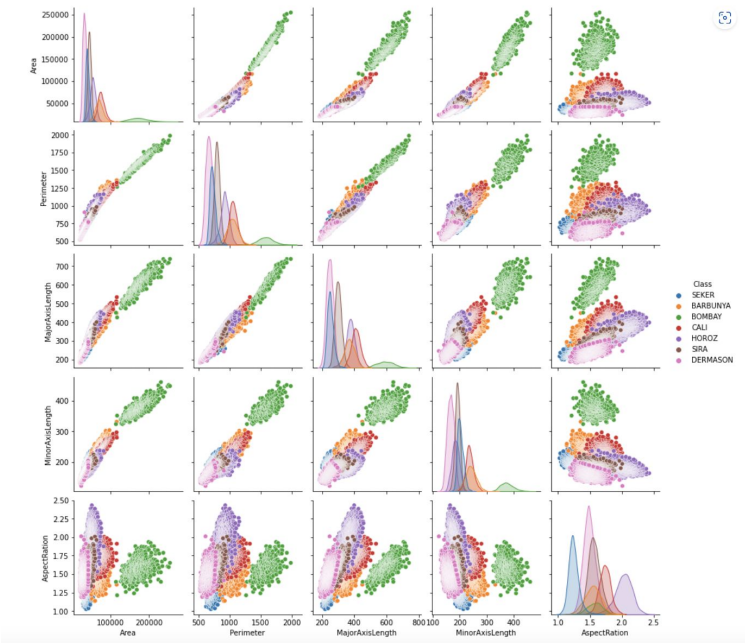
Correlation Matrix - Pearson Method using pairwise observations



- From the above correlation we see majority of the features are having higher positive or negative correlation.
- The labels for "Primeter" and and "EquivDiameter", and "ConvexArea" and "EquivDiameter" are highly positive correlated, suggesting there is a relationship between the two variables.

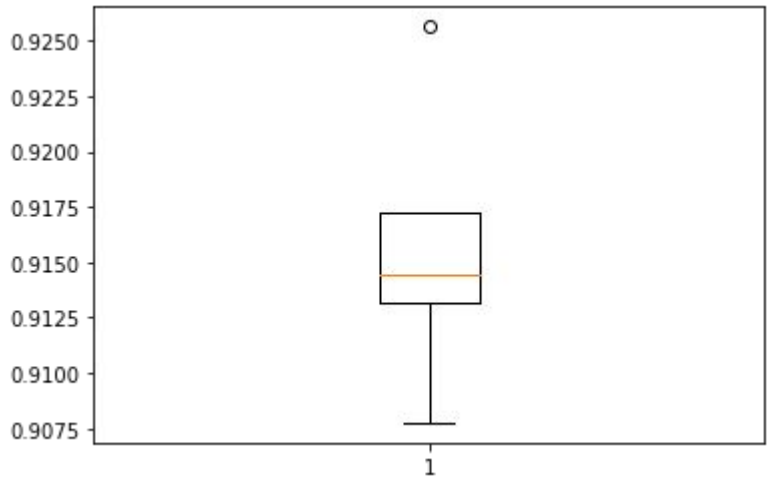
Findings

Scatterplot and their correlation between 5 variables taken from the dataset

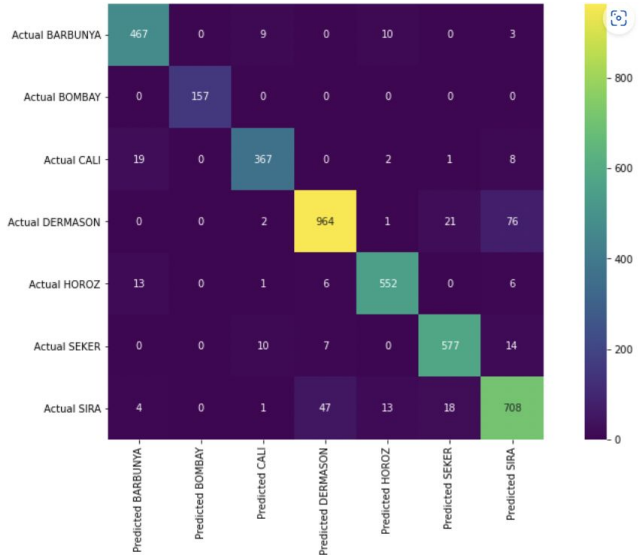


- From the pair plot, the majority of features are having very strong positive and negative linear relationship with the “Class” variable. We can also observe that that the data is either bivariate or multivariate. “AspectRaion” scattered across which does not show a strong patterns with “Class”.

Findings - Logistic Regression

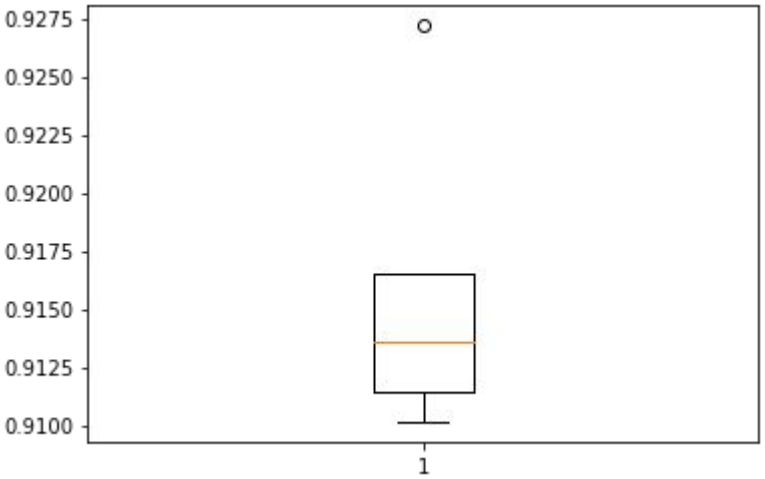


Logistic Regression Confusion Matrix

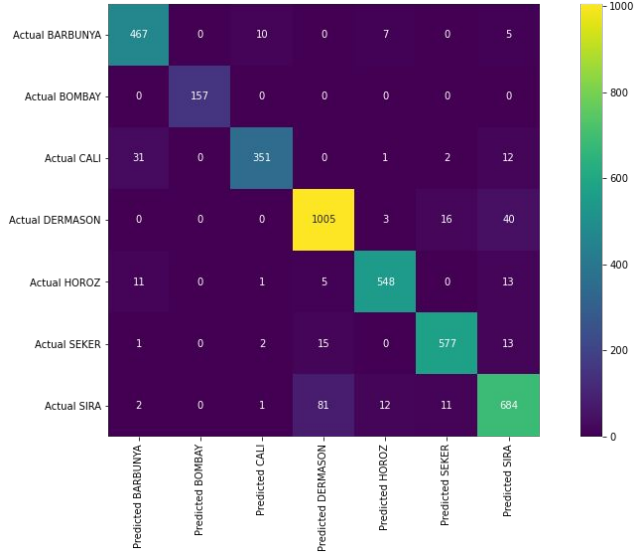


- Performance on training data varies between 0.9075 to 0.925.
- Accuracy rate of Barbunya, Bombay, Cali, Dermason, Horoz, Seker, and Sira were 95.5%, 100%, 92.4%, 90.6%, 95.1%, 95.5%, and 89.5%.
-

Findings - KNN

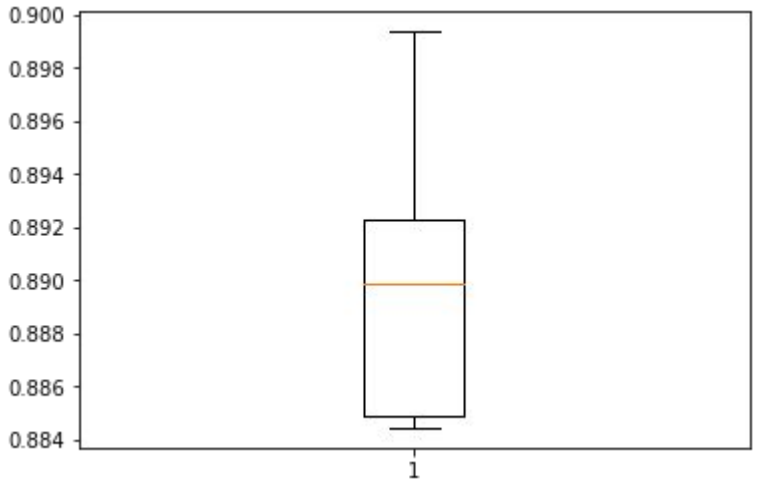


KNN Confusion Matrix

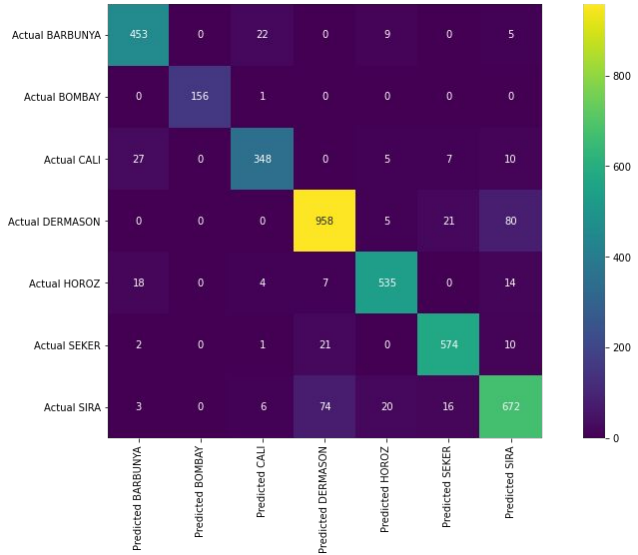


Performance on training data varies between 0.91 to 0.928.

Findings - Decision Tree

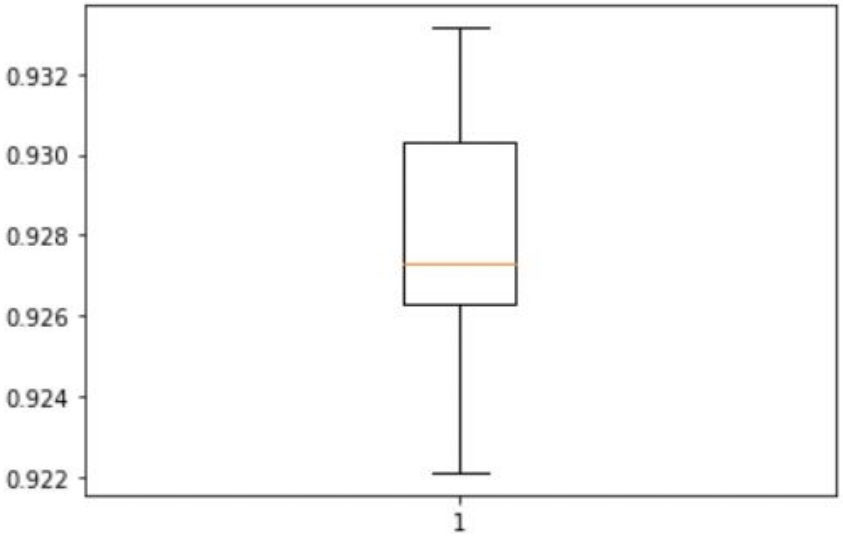


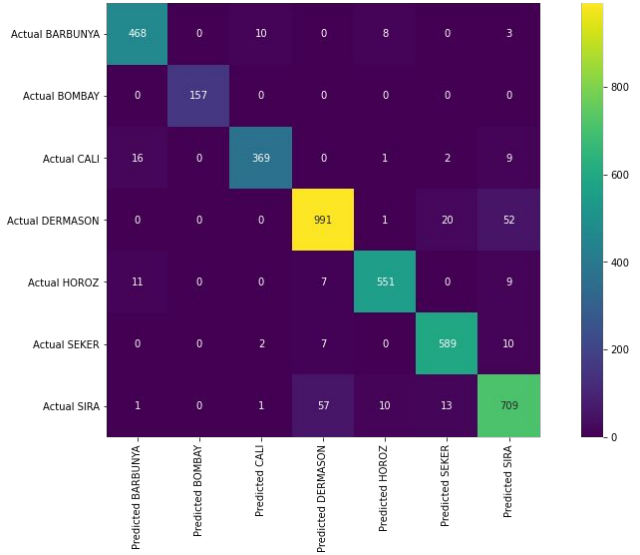
DT Confusion Matrix



- Performance on training data varies between 0.884 to 0.9.

Findings - SVM





- Performance on training data varies between 0.922 to 0.933.
- SVM model has the best value with 93.2% of accuracy, and obtained the best values among all calculated performance metrics.
- Accuracy rate of Barbunya, Bombay, Cali, Dermason, Horoz, Seker, and Sira were 95.7%, 100%, 92.9%, 93.1%, 95.3%, 96.9%, 89.6%.
- **Bombay** variety can be fully classified with 100% accuracy, while **Sira** has the lowest classification performance among all variety.

Findings

Performance values obtained for Logistic Regression, KNN, Decision Tree, and SVM

	Algorithm	Precision	Recall	F1 Score	Accuracy
0	Logistic Regression	0.922484	0.921645	0.921711	0.921645
1	KNN	0.924222	0.924094	0.924101	0.924094
2	SVM_linear	0.927841	0.927767	0.927768	0.927767
3	Decision Tree	0.897428	0.896915	0.897089	0.913565

- The accuracy, error rate, precision, specificity, recall, F1 score classification performance metrics are calculated by using confusion matrix of each model.
- We have a model of recall and accuracy is the same. It might be coincidence, and this means your model is somehow "balanced", that is, its ability to correctly classify positive samples is same as its ability to correctly classify negative samples.

Limitations

In this study, the variables related to the shape and size characteristics of the bean were taken from two dimensional images. The third dimension of beans are not included in the study. By providing the data of structure axis of the bean, such as coefficient of variance in shape and size, may increase the success rate of classification. If the market demands a more accurate classification system for dry beans, perhaps I would like to examine the color saturation levels between dry bean and the result is not collinear with the class feature (Not highly correlated with “Class”).

Conclusions

In this study, I find Sira variety has the lowest classification performance compared to Dermason using all four classification models. The fact that the flatness and roundness features of the Dermason and Sira are similar that makes the result more robust. SVM outperformed any other models and achieved 92.8% on accuracy rate.

Acknowledgements

Where did you get your data? Did you use other informal analysis to inform your work? Did you get feedback on your work by friends or colleagues? Etc. If you had no one give you feedback and you collected the data yourself, say so.

I do not receive any feedback from anyone else , and this dataset is collected from UCI repository.

References

If applicable, report any references you used in your work. For example, you may have used a research paper from X to help guide your analysis. You should cite that work here. If you did all the work on your own, please state this.

Visualizing Data with Pairs Plots in Python | by Will Koehrsen | Towards Data Science

Multiclass classification of dry beans using computer vision and machine learning techniques | Request PDF (researchgate.net)

pandas - Python Data Analysis Library (pydata.org)