RAISIN CLASSIFICATION with Logistic Regression

By

Thu Tran
Ly Nguyen
Huichan Lee

Date

April 30, 2024

I. INTRODUCTION:

Dried grapes, or raisins globally prized for their nutritional richness, are staple crops in countries like Türkiye, the United States, and Greece. Our study examined 900 raisin grains, equally divided between Kecimen and Besni varieties. We meticulously extracted seven morphological features from these images. This project aims to develop a classification system using Logistic Regression to differentiate between the two raisin types. Additionally, we explore the effectiveness of other machine learning techniques like decision trees and random forest for comparison with the logistic model.

II. DATASET OVERVIEW:

1. Data description:

The dataset initially comprised 900 images of raisin grains, equally divided between the Besni and Kecimen varieties, cultivated in Turkey. These images serve as the foundation for our project, where researchers utilize advanced machine vision techniques to extract numerical data from the images. The data used in this project consists of 900 rows and 8 columns, described in the table below:

Table 1: Data Description

Variable name	Description
Area	The number of pixels within the boundaries of the raisin
Perimeter	The environment is measured by calculating the distance between the boundaries of the raisin and the pixels around it.
MajorAxisLength	The length of the main axis, which is the longest line that can be drawn on the raisin
MinorAxisLength	The length of the small axis, which is the shortest line that can be drawn on the raisin
Eccentricity	A measure of the eccentricity of the ellipse, which has the same moments as raisins
ConvexArea	The number of pixels of the smallest convex shell of the region formed by the raisin
Extent	The ratio of the region formed by the raisin to the total pixels in the bounding box
Class	Kecimen and Besni raisin

2. **Exploratory Data Analysis (EDA):**
   a. **Statistical summary:**

Table 2: Statistical Summary Table

STATISTICAL SUMMARY TABLE

Type	Variables	Missing	Min	Mean	Median	Max	SD
factor	Class	0	NA	NA	NA	NA	NA
numeric	Area	0	25387.00	87804.13	78902.00	235047.00	39002.11
numeric	MajorAxisLength	0	225.63	430.93	407.80	997.29	116.04
numeric	MinorAxisLength	0	143.71	254.49	247.85	492.28	49.99
numeric	Eccentricity	0	0.35	0.78	0.80	0.96	0.09
numeric	ConvexArea	0	26139.00	91186.09	81651.00	278217.00	40769.29
numeric	Extent	0	0.38	0.70	0.71	0.84	0.05
numeric	Perimeter	0	619.07	1165.91	1119.51	2697.75	273.76

This statistical summary table provides an overview of the measurement characteristics of various variables extracted from the raisin grain images dataset. For each variable, the table displays the count of missing values, as well as key descriptive statistics such as minimum, mean, median, maximum, and standard deviation. Notably, there is no missing value in this dataset.

   b. **"Class" distribution:**

The histogram illustrates the distribution of raisin types, with two distinct categories: Besni and Kecimen. It is evident from the plot that there is a noticeable difference in the frequency of occurrence between the two types, with Besni raisins appearing more frequently than Kecimen. This visualization provides a clear overview of the relative abundance of each raisin type within the dataset, highlighting the prevalence of Besni raisins compared to Kecimen.

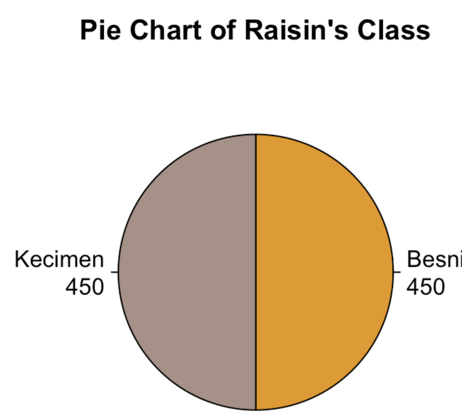

Figure 1: Raisin's class distribution

c. Variables distribution bases on “Class”:

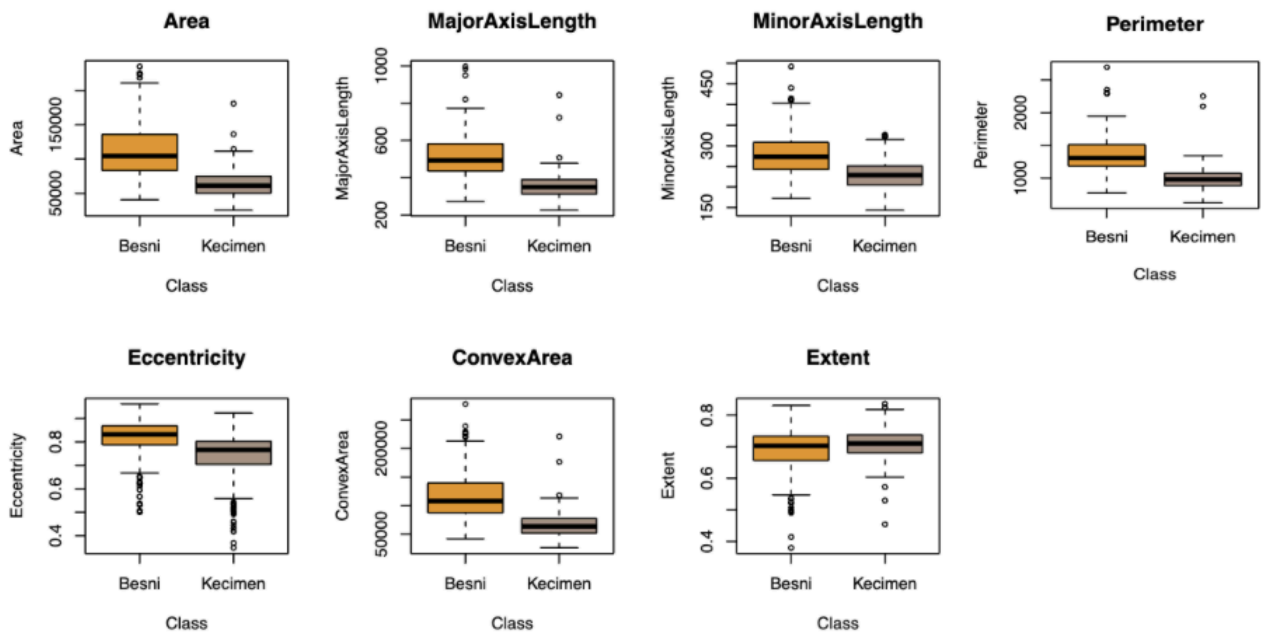

Figure 2: Side-by-side boxplot of morphological features

In Figure 2, we observe the side-by-side box plots showcasing the morphological features of both Besni and Kecimen raisin grains. These plots provide a comprehensive comparison, revealing notable distinctions between the two varieties. Across various measurements such as area, major axis length, minor axis length, eccentricity, convex area, extent, and perimeter, Besni consistently demonstrates a higher median and wider range when juxtaposed with Kecimen. This suggests intriguing differences in the morphological characteristics between the two types of raisins, potentially indicating diverse genetic backgrounds or environmental influences. Furthermore, Figure 3 offers additional insights by presenting sample images of the raisin varieties, allowing for a more holistic understanding of their visual distinctions. Through the combination of statistical analyses and visual representations, we gain a deeper understanding of the nuanced differences between Besni and Kecimen raisins, enhancing our ability to discern and classify these varieties accurately.

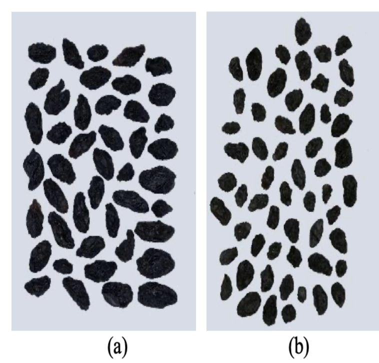

Figure 3: Sample image of raisin varieties used in the study ((a) Besni, (b) Kecimen)

d. Variables matrix:

The matrix scatter plot in figure illustrates the pairwise relationships between morphological features extracted from raisin grains. It reveals various degrees of correlation between features, with some displaying strong positive correlations, as evidenced by the upward-sloping trend lines. Multicollinearity appears evident among certain features (specially between ‘Area’ and ‘Convex Area’), where high correlation between predictors may pose challenges in regression analysis, potentially leading to inflated standard errors and inaccurate coefficient estimates. Identifying and addressing multicollinearity is crucial for ensuring the reliability of predictive models derived from these features

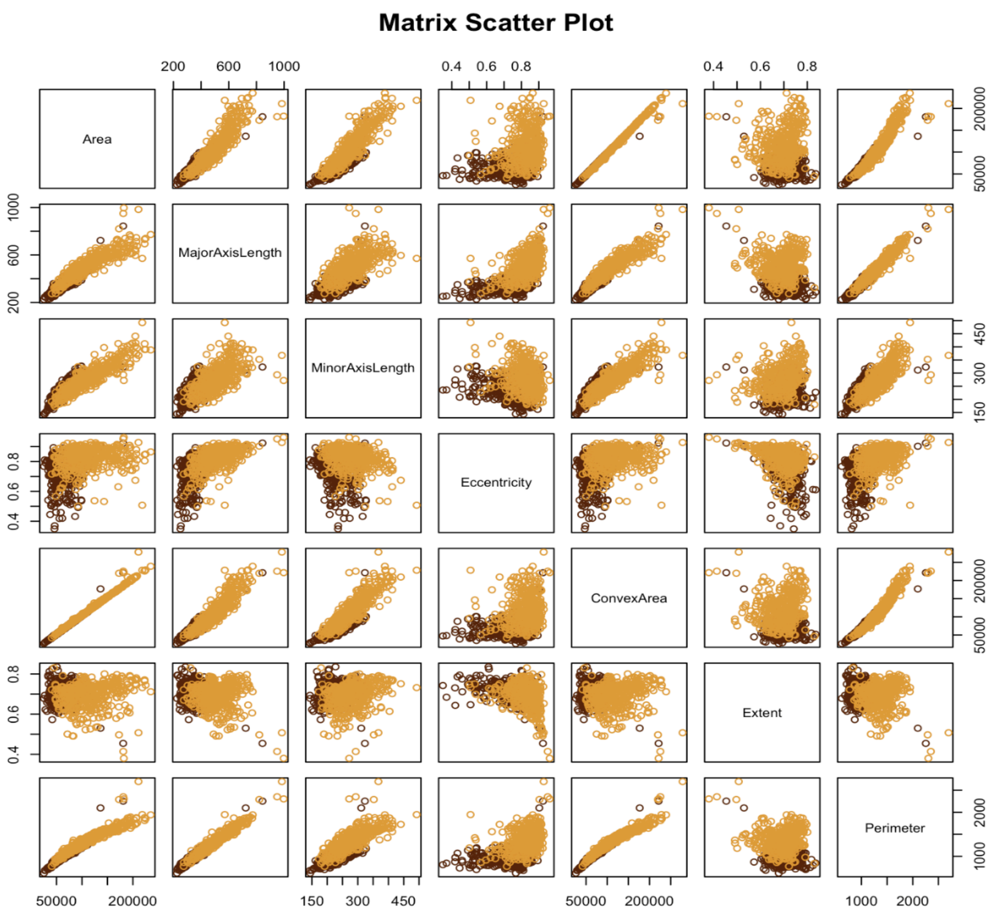

Figure 4: Matrix scatter plot of morphological features

III. ANALYSIS WITH LOGISTIC REGRESSION:

1. Variable selection:

Through the application of stepwise backward variable selection with AIC, we constructed the logistic model (model1), which included five significant predictors variables: Area, MajorAxisLength, MinorAxisLength, ConvexArea, and Perimeter (Figure 5: Model 1 summary). However, our analysis uncovered a significant correlation between Area and Convex Area (Table 3: VIF score). To address this issue and refine the model for improved interpretability and accuracy, we made the decision to exclude Convex Area from the model, resulting in the updated model (see figure 6: Model 2 summary). This adjustment aimed to minimize the impact of multicollinearity, ensuring the model's reliability and enhancing its interpretability and predictive accuracy.

2. Transformation:

In our analysis, we explored transforming the predicted variables to enhance the accuracy of outcome prediction. Despite our efforts, the transformed predictors in the model were found to be statistically insignificant (See Figure 7: Transform model summary). This unexpected result prompted us to reconsider our approach and explore alternative methods.

3. Evaluation and final model:

Table 4: Multiple logistic regression models

MULTIPLE LOGISTICS REGRESSION MODELS

Models	Num.Predictors	Accuracy	AUC	AIC
fullmodel	7	0.8577778	0.9279111	624.9814
model1	5	0.8555556	0.9278864	621.6371
model2	4	0.8611111	0.9333531	628.6019
transform1	4	0.8700000	0.9345333	616.6606
nullmodel	0	0.5000000	0.5000000	1249.6649

Although 'transform1' displayed promising performance across various metrics, the summary table revealed that the coefficients lacked statistical significance, indicating potential issues with the model. Conversely, the full model demonstrated strong classification capabilities for the raisin dataset; however, its evaluation metric fell short compared to 'model2'. Consequently, 'model2', featuring four predictors (Area, MajorAxisLength, MinorAxisLength, Perimeter), emerged as the preferred final model due to its balanced performance and predictive efficacy.

IV. ANALYSIS WITH OTHER CLASSIFICATION MODELS:

1. Cross-validation:

In the subsequent phase of our analysis, we employ cross-validation to partition the data into a 70% training set and a 30% test set. This crucial step ensures that our models are trained on a sufficiently large portion of the data while still retaining an independent subset for evaluation. By comparing the performance of our final logistic regression model with other machine learning methods like decision trees and random forests, using metrics such as accuracy, specificity, sensitivity, and AUC (Area Under the Curve), we gain valuable insights into their predictive capabilities. Cross-validation plays a pivotal role in this process, as it helps mitigate the risk of overfitting and provides a more robust assessment of model performance by validating its generalizability on unseen data. This ensures that our predictive models are reliable and effective in real-world scenarios, enhancing their utility and applicability.

2. Comparative analysis:

Table 5: Comparative table

MODELS COMPARATIVE TABLE

Models	Accuracy	Sensitivity	Specificity	AUC
Logistic Regression	0.8703704	0.9333333	0.8074074	0.9339369
Decision Tree	0.8666667	0.9333333	0.8000000	0.8666667
Random Forest	0.8592593	0.9185185	0.8000000	0.9321262

From the comparative table, it's evident that the final model, implemented with logistic regression, outperforms both the decision tree and random forest models across various metrics. The final model achieved the highest accuracy, sensitivity, and specificity values, indicating its superior ability to correctly classify instances. Moreover, it exhibited the highest AUC (Area Under the Curve), implying excellent overall performance in distinguishing between different classes. This suggests that logistic regression is the most effective model among the tested approaches for predicting raisin classes in our analysis.

V. CONCLUSION:

Our primary objective was to employ logistic regression for classifying raisin types. Initially, we utilized the complete set of morphological features in our model, subsequently refining it by removing predictors deemed insignificant or prone to multicollinearity. Upon finalizing the model, we employed cross-validation to ensure robustness and compared our logistic model with alternative machine learning techniques. The resulting analysis revealed logistic regression's effectiveness in predicting raisin classes, as demonstrated in the latest comparative table. This comprehensive approach allowed us to identify logistic regression as a valuable tool for accurately classifying raisin types, offering insights into its practical utility and efficacy within the context of our study.

Figures

Figure 4: Model 1 summary

```
glm(formula = Class ~ Area + MajorAxisLength + MinorAxisLength +
    ConvexArea + Perimeter, family = binomial, data = raisin)

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)      6.7317525  4.4215966   1.522 0.127891
Area            -0.0004777  0.0001159  -4.120 3.79e-05 ***
MajorAxisLength  0.0467310  0.0156343   2.989 0.002799 **
MinorAxisLength  0.0788838  0.0212508   3.712 0.000206 ***
ConvexArea       0.0003990  0.0001127   3.540 0.000401 ***
Perimeter       -0.0360055  0.0063523  -5.668 1.44e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1247.66  on 899  degrees of freedom
Residual deviance:  609.64  on 894  degrees of freedom
AIC: 621.64

Number of Fisher Scoring iterations: 7
```

Figure 5: Model 2 summary

```
glm(formula = Class ~ Area + MajorAxisLength + MinorAxisLength +
    Perimeter, family = binomial, data = raisin)

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)      3.3473160  4.2360894   0.790   0.4294
Area            -0.0001113  0.0000562  -1.981   0.0476 *
MajorAxisLength  0.0321112  0.0148931   2.156   0.0311 *
MinorAxisLength  0.0682675  0.0209138   3.264   0.0011 **
Perimeter       -0.0219106  0.0044520  -4.921 8.59e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1247.7  on 899  degrees of freedom
Residual deviance:  618.6  on 895  degrees of freedom
AIC: 628.6

Number of Fisher Scoring iterations: 7
```

Figure 6: Summary of transformation model

```
Call:
glm(formula = Class ~ I((Area^(-1/5) - 1)/(-1/5)) + I((MajorAxisLength^(-1/3) -
    1)/(-1/3)) + I((MinorAxisLength^(-1/3) - 1)/(-1/3)) + I((Perimeter^(-1/2) -
    1)/(-1/2)), family = binomial, data = raisin)

Coefficients:
                                       Estimate Std. Error z value     Pr(>|z|)
(Intercept)                             1662.99     291.67   5.702 0.0000000119 ***
I((Area^(-1/5) - 1)/(-1/5))              -65.64      90.47  -0.726        0.468
I((MajorAxisLength^(-1/3) - 1)/(-1/3))   106.12      84.36   1.258        0.208
I((MinorAxisLength^(-1/3) - 1)/(-1/3))   107.08      66.29   1.615        0.106
I((Perimeter^(-1/2) - 1)/(-1/2))        -986.90     177.39  -5.564 0.0000000264 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1247.66  on 899  degrees of freedom
Residual deviance:  606.66  on 895  degrees of freedom
AIC: 616.66

Number of Fisher Scoring iterations: 6
```

Table

Table 3: VIF Score table

Variables	**Area**	**MajorAxisLength**	**MinorAxisLength**	**ConvexArea**	**Perimeter**
VIF score	429.6	68.3	51.7	431.3	65.3

Citation:

CINAR I., KOKLU M. and TASDEMIR S., (2020), Classification of Raisin Grains Using Machine Vision and Artificial Intelligence Methods. Gazi Journal of Engineering Sciences, vol. 6, no. 3, pp. 200-209, December, 2020.

Code: