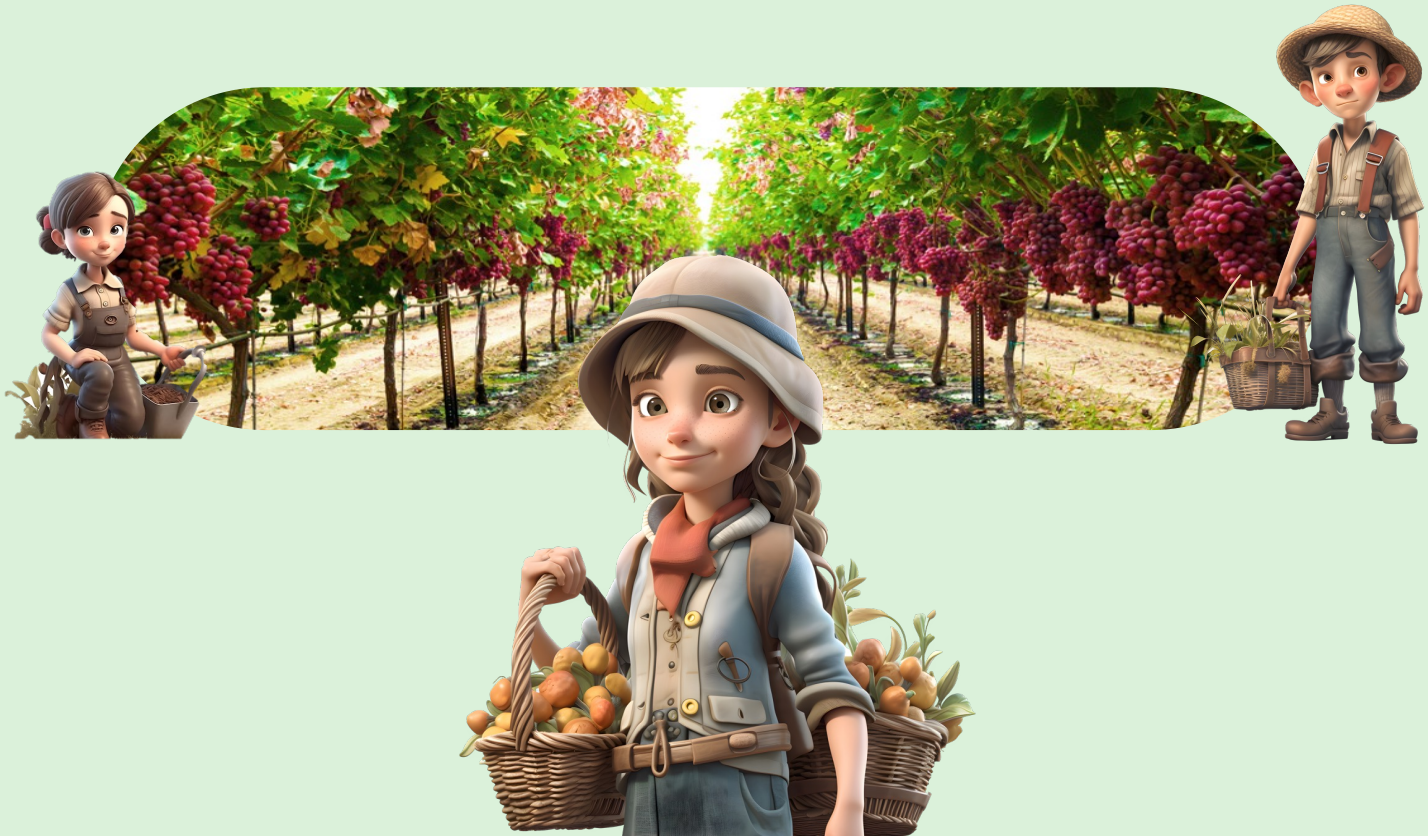
WELCOME




RAISIN CLASSIFICATION

RAISIN CLASSIFICATION

- **Huichan Lee**
- **Ly Nguyen**
- **Thu Tran**

RAISIN CLASSIFICATION

1. Introduction
2. Data overview
3. Analysis with Logistic Regression
4. Analysis with other Machine Learning models
5. Conclusions

1 Introduction

- Raisins is recognized for their nutritional richness
- By understanding the different features can aid in quality assessment, agriculture application, trade…
- The dataset was collected from a comprehensive sampling across various raisin-producing regions, including Turkey, the United States, and Greece in December , 2020.

2 Data overview

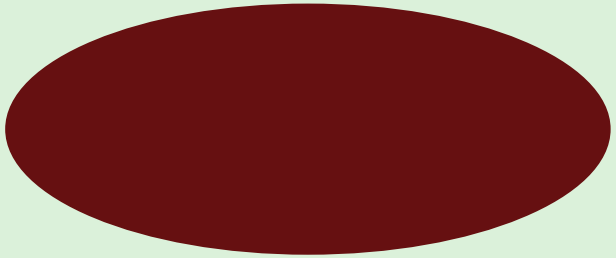
a. Data Description

- Our data has 900 rows and 8 columns.
- Here is 8 morphological features that are used in the dataset :

2 Data overview

a. Data Description

- Our data has 900 observations and 8 variables.
- Here is 8 morphological features that are used in the dataset :

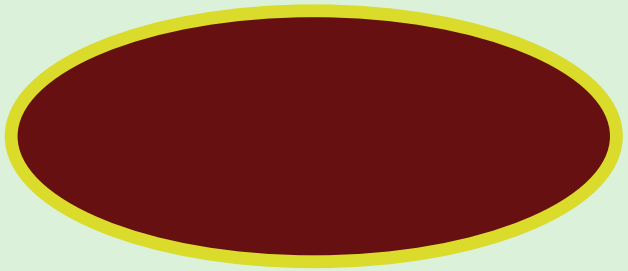


Area

2 Data overview

a. Data Description

- Our data has 900 observations and 8 variables.
- Here is 8 morphological features that are used in the dataset :

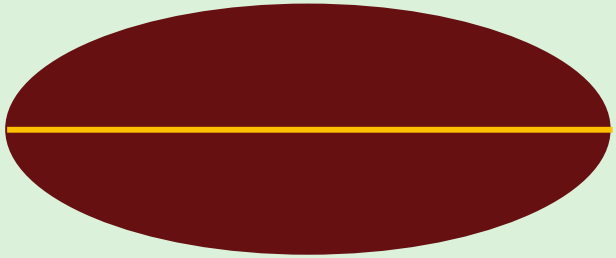


Perimeter

2 Data overview

a. Data Description

- Our data has 900 observations and 8 variables.
- Here is 8 morphological features that are used in the dataset :

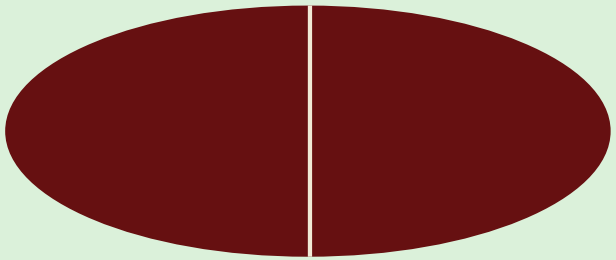


Major Axis Length

2 Data overview

a. Data Description

- Our data has 900 observations and 8 variables.
- Here is 8 morphological features that are used in the dataset :

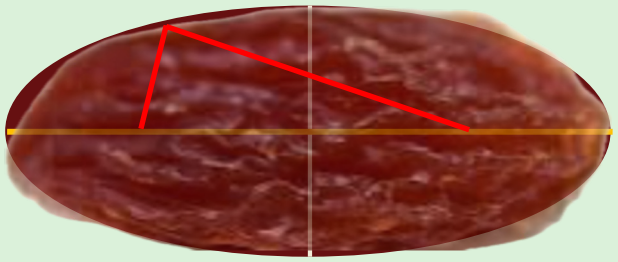


Minor Axis Length

2 Data overview

a. Data Description

- Our data has 900 observations and 8 variables.
- Here is 8 morphological features that are used in the dataset :

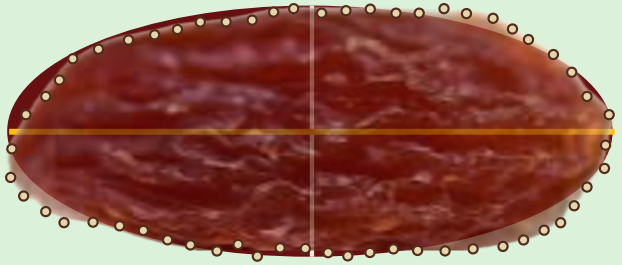


Eccentricity

2 Data overview

a. Data Description

- Our data has 900 observations and 8 variables.
- Here is 8 morphological features that are used in the dataset :

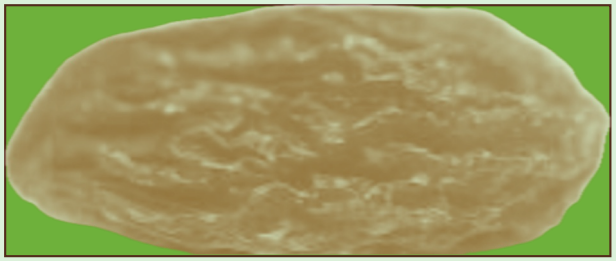


Convex Area

2 Data overview

a. Data Description

- Our data has 900 observations and 8 variables.
- Here is 8 morphological features that are used in the dataset :



Extent

2 Data overview

a. Data Description

- Our data has 900 observations and 8 variables.
- Here is 8 morphological features that are used in the dataset :

BESNI

KECIMEN

Class

2 Data overview

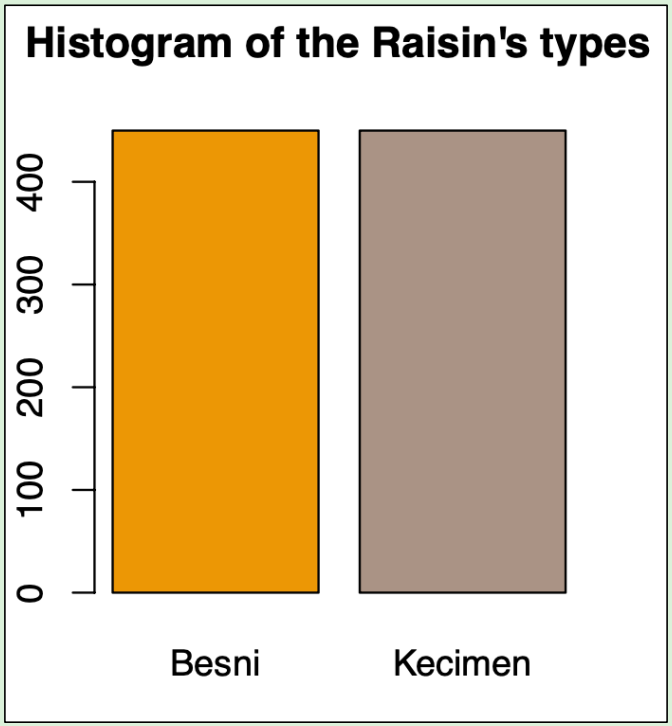
b. Exploratory Data Analysis (EDA)

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

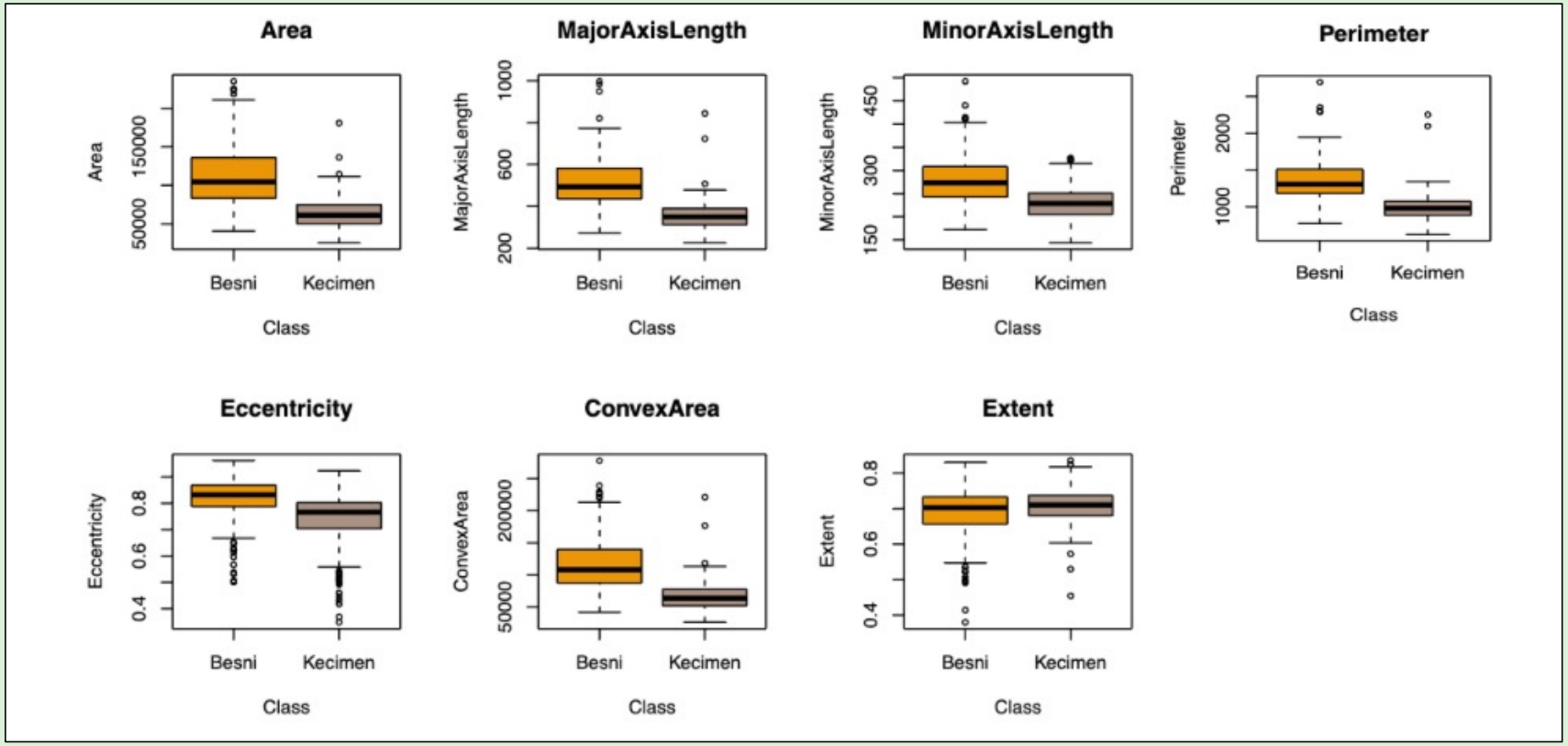
2 Data overview

b. Exploratory Data Analysis (EDA)



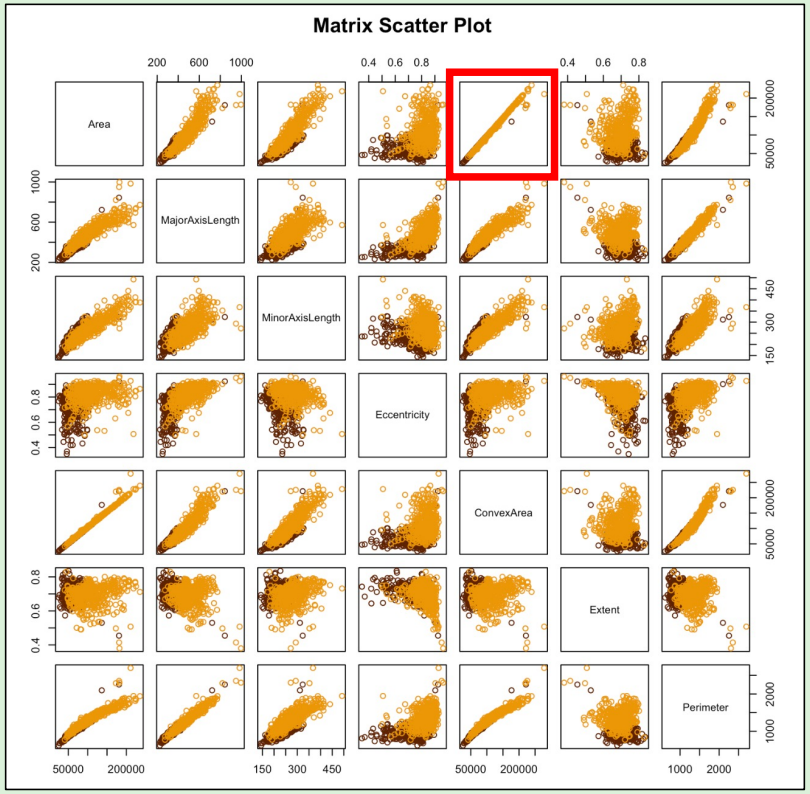
2 Data overview

b. Exploratory Data Analysis (EDA)



2 Data overview

b. Exploratory Data Analysis (EDA)



3 Analysis with Logistic Regression

a. Variable selection:

Full model:

$$\mathbf{Log}\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 \text{ . Area } + \beta_2 \text{. Major Axis Length} + \beta_3 \text{ . Minor Axis Length} +$$

$$\beta_4 \text{ . Eccentricity} + \beta_5 \text{ . Convex Area} + \beta_6 \text{ . Extent} +$$

$$\beta_7 \text{ . Perimeter}$$

3 Analysis with Logistic Regression

a. Variable selection:

Stepwise backward model (model1):

$$\mathbf{Log}\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 \text{ . Area } + \beta_2 \text{. Major Axis Length} + \beta_3 \text{ . Minor Axis Length} +$$

$$\beta_5 \text{ . Convex Area} + \beta_7 \text{ . Perimeter}$$

3 Analysis with Logistic Regression

a. Variable selection:

Stepwise backward model (model1):

$$\mathbf{Log}\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 \text{ . Area } + \beta_2 \text{. Major Axis Length} + \beta_3 \text{ . Minor Axis Length} + \beta_5 \text{ . Convex Area} + \beta_7 \text{ . Perimeter}$$

Remove high correlated variable (model2):

$$\mathbf{Log}\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 \text{ . Area } + \beta_2 \text{. Major Axis Length} + \beta_3 \text{ . Minor Axis Length} + \beta_7 \text{ . Perimeter}$$

3 Analysis with Logistic Regression

b. Transformation:

Transformation model (transform1):

$$\mathbf{Log}\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 \cdot \frac{Area^{-1/5}-1}{-1/5} + \beta_2 \cdot \frac{MajorAxisLength^{-1/3}-1}{-1/3} + \beta_3 \cdot \frac{MinorAxisLength^{-1/3}-1}{-1/3} + \beta_7 \cdot \frac{Perimeter^{-1/2}-1}{-1/2}$$

3 Analysis with Logistic Regression

b. Transformation:

Transformation for model2 (transform1):

```
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
```

3 Analysis with Logistic Regression

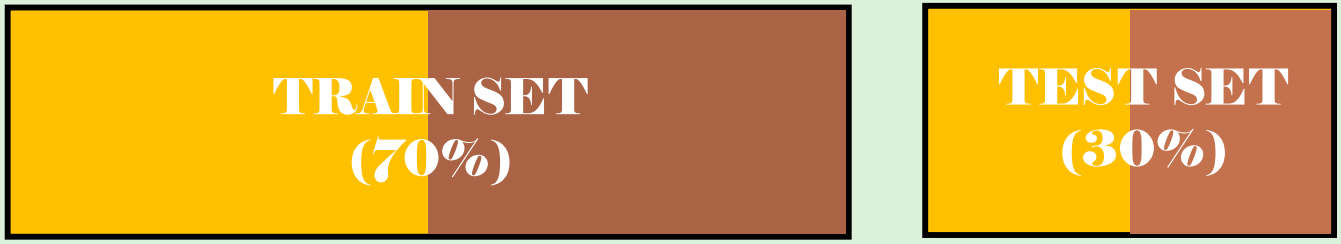
c. Evaluation:

Comparision table:

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

4 Analysis with other Machine Learning models

a. Split data:



b. Machine learning models:

- Logistic regression (Final model – model2)
- Decision Tree
- Random Forest

4 Analysis with other Machine Learning models

c. Evaluation:

MODELS COMPARATIVE TABLE				
Models	Accuracy	Sensitivity	Specificity	AUC
Logistic Regression	0.870	0.933	0.807	0.934
Decision Tree	0.867	0.933	0.800	0.867
Random Forest	0.859	0.919	0.800	0.932

5 Conclusions

- Through this project, we have done logistic regression to classify 2 types of raisin (Besni and Kecimen).
- The final model is selected by using the backward variable selection and eliminating the high correlated variable.
- The model uses only 4 predictors (Area, Major Axis Length, Minor Axis Length, and Perimeter) to predict “Class” of raisin
- Transformation model doesn’t work well for this dataset.
- As comparing with other machine learning method, logistic regression seems to be the most efficient model to predict the types of raisin. (With Accuracy score 87%, and AUC 93.4%)
- Model can be improved by comparing to other machine learning method and applied tuning .

THANK YOU