

# **CLASSIFICATION OF DATE FRUITS INTO GENETIC VARIETIES**



**EXST 7152: ADVANCED TOPICS IN STATISTICAL MODELING**

**SPRING 2022**

**FINAL PROJECT**

**CHRISTOPHER KUETSINYA**

## INTRODUCTION

Date fruit (*Phoenix dactylifera*), which has more than 3000 varieties worldwide, is an edible and nutritive fruit. Considered the oldest cultivated fruit in the world, fossil evidence indicates that dates fruits go back at least 50 million years ago. The date palm is of the Kingdom Plantae, Division Magnoliophyta, Class Liliopsida, Order Arecales, Family Arecaceae, and Genus Phoenix. There are 12 to 19 species of date palms with a height reach of approximately 69 to 75 feet. The fruit is a short cylindrical shape about 1-3 inches (25.4mm - 76.3mm) in length and an inch (25.4mm) in diameter. The color ranges from bright red to bright yellow, to deep purple.

## MOTIVATION

Lately, there has been increased interest in some types of varieties as snacks, inclusion in traditional desserts, oils, cosmetics, and soaps. Pharmacological studies have also shown that dates (some varieties) contain phytochemicals that may help in some disease treatment and management. With these growing interests, there is a need to properly classify the different varieties of date fruits which is important for commercial viability and pharmaceutical processing. Expert opinion is needed to distinguish date fruit varieties and species due to different nutritional values, quality differences, and tastes.

Therefore, this study aims to accurately classify different varieties of date fruits using image analysis without needing time-consuming and complex physical measurements.

## DATA DESCRIPTION

898 images of 7 different date fruit varieties; Barhee 65, Deglet Nour 98, Sukkary 204, Rotab Mozafati 72, Ruthana 166, Safawi 199, and Sagai 94 samples were used. A CVS was set for the image acquisition and captured images of the date fruits were transferred to the computer environment.



Source: Koklu, M. et al 2021

*Figure 1. The processing method of the obtained images of the palm fruit stages. (a) Initial status. (b) Cleaned status. (c) Final status.*

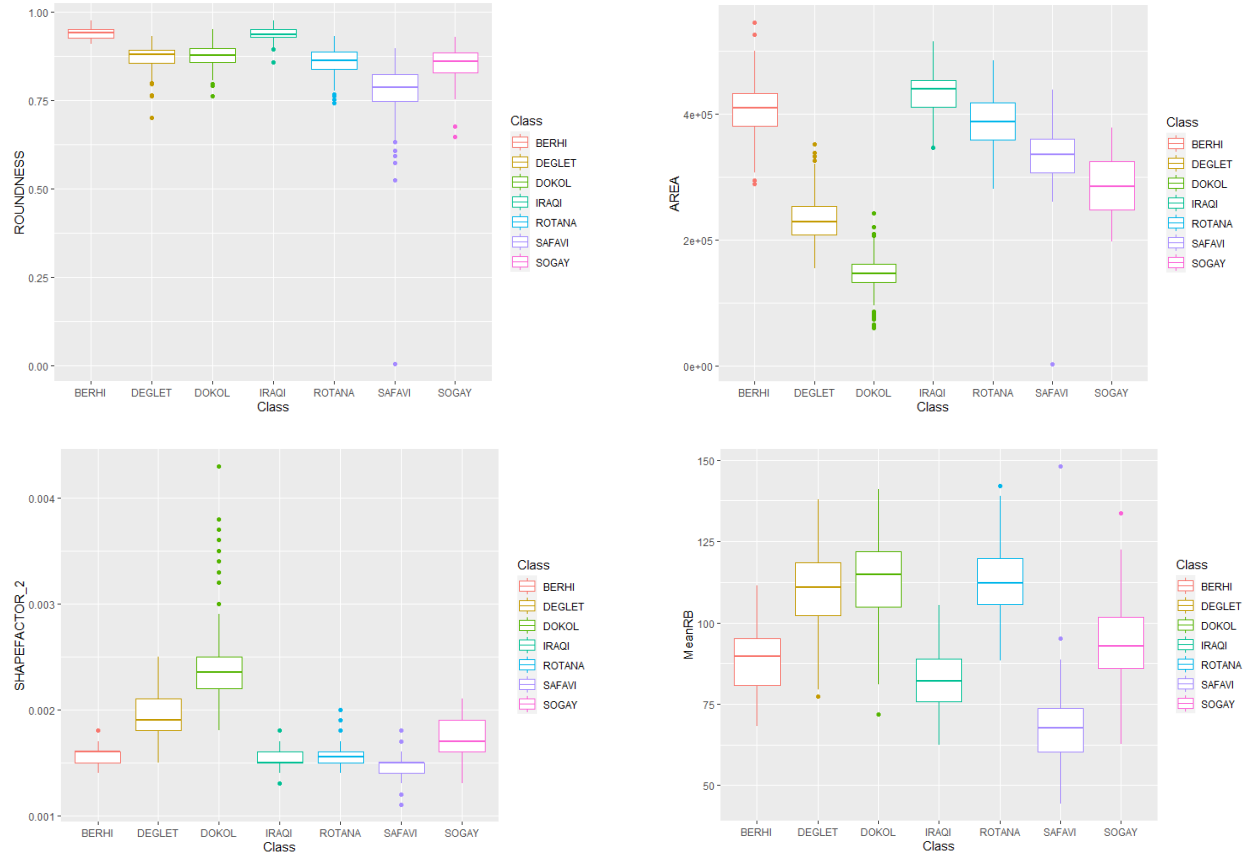
The camera used in the setup is placed on a closed box with an LED light system. The future robot S100 series smart camera used to capture images has a resolution of  $1280 \times 1024$ . A total of 34 features, which include 12 morphological, 4 shape, and 18 color features, were extracted. All the features are numerical and the response is a categorical variable with 7 levels.

Main features	Sub-features		
<b>Morphological features</b>	Area	Equivalent diameter	
	Perimeter	Solidity	
	Major axis	Convex_area	
	Minor axis	Extent	
	Eccentricity	Aspect ratio	
	Roundness	Compactness	
<b>Shape features</b>	Shapefactor_1	Shapefactor_3	
	Shapefactor_2	Shapefactor_4	
<b>Color features</b>	Mean RR	Mean RB	Mean RG
	Std. dev RR	Std. dev RB	Std. dev RG
	Skew RR	Skew RB	Skew RG
	Kurtosis RR	Kurtosis RB	Kurtosis RG
	Entropy RR	Entropy RB	Entropy RG
	All Daub4 RR	All Daub4 RB	All Daub4 RG

Table 1. A list of features extracted from the 898 images, grouped into 3 main features.

## PRELIMINARY ANALYSIS

The data was checked for missing values and duplicate entries but there were none. For preliminary analysis, some features (Roundness, Area, Shapefactor\_2 & MeanRB); which were believed to have information on the general features (length, diameter & color) noted in the introduction, were plotted against the varieties to identify distinctions and similarities. The plot from the “Roundness” showed that there are similarities between the varieties Berhi and Iraqi, Deglet and Dokol, Rotana and Sogay. The variety Safavi didn’t show any similarity with others in terms of roundness. The plot showed relatively low roundness values for Safavi as compared to the varieties. For the “Area” feature plot, the Dokol variety was distinct from the others. This variety has lower area values as compared to the other varieties. The other feature plots also show some distinctions for the Dokol and Safavi varieties. In general, it was seen that some of the varieties were similar and others distinct according to different features.



*Figure 2. Side-by-side box plots for the 7 varieties using Roundness (top-left), Area (top-right), Shapefactor\_2 (bottom-left) and MeanRB (bottom-right). These plots capture some similarities and differences in the 7 varieties according to the features.*

The side-by-side boxplots captured some interesting information about some features and varieties. Given that, scatterplots of some pairwise features, that displayed distinctions in some varieties were built to identify where some varieties can be classified using the combination of two features. The scatterplot between the features “Roundness” and “MeanRB” classify the variety of SAFAVI very well to some extent. It classifies about 95% of the SAFAVI variety; when Roundness is between 0.5 to 0.85, and when MeanRB is 0 to 85.



Figure 3. Scatterplots displaying observations using Roundness & Area (left), and MeanRB & Roundness (right).

It can be said that when pairwise combinations of all the features are observed, obtaining feature ranges for classifying each variety can be ascertained. However, that means one has to look at an extremely large number of scatterplots to obtain these ranges and this is an impossible task. Hence, other methods were considered in classifying the varieties.

## DATA ANALYSIS

### PRINCIPAL COMPONENT ANALYSIS

Considering the number of features and the possibility of high correlation among some features, the PCA was used to identify highly correlated features, identify relevant features, and also identify some clustering. Generally, the goal of the PCA was to reduce the dimension of the data and also find a low-dimension representation that captures as much variation in the data as possible.

#### Correlation

From the loadings plot, the circle around the loadings reflects how well the variables are described. A long arrow that touches or almost touches the circle indicates that all the information regarding that feature is captured by the PCA plot. Therefore, in identifying the correlation between two features, the corresponding loadings of the two features need to be pretty close to each other to be highly correlated. Furthermore, they need to be well explained in the PCA (loadings must be close to the circle).

With the above-mentioned idea, it was seen that Area, Eqdiasq, and Convex\_Area were highly correlated. Shapefactor\_3 and Compactness; SkewRR, KurtosisRG and SkewRG; EntropyRR and EntropyRG; and Alldaub4RB, Alldaub4RG, Alldaub4RR, MeanRG, and StdDevRR were all highly correlated.

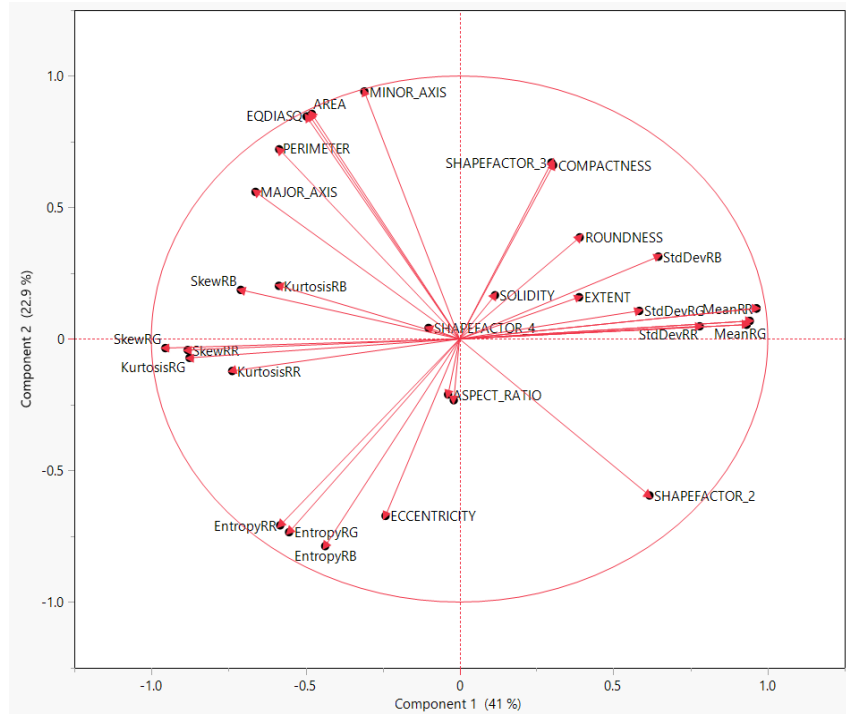


Figure 5. PCA Loading plot

## Clustering

By plotting the PC1 and PC2, clustering groups were seen with slight overlapping. PC1 and PC2 together explain 63.9% variation in the data. In general, the dimensionality of the data was reduced from 34 variables to just 15 PCs, with only the first 10 PCs explaining about 95% variation in data.

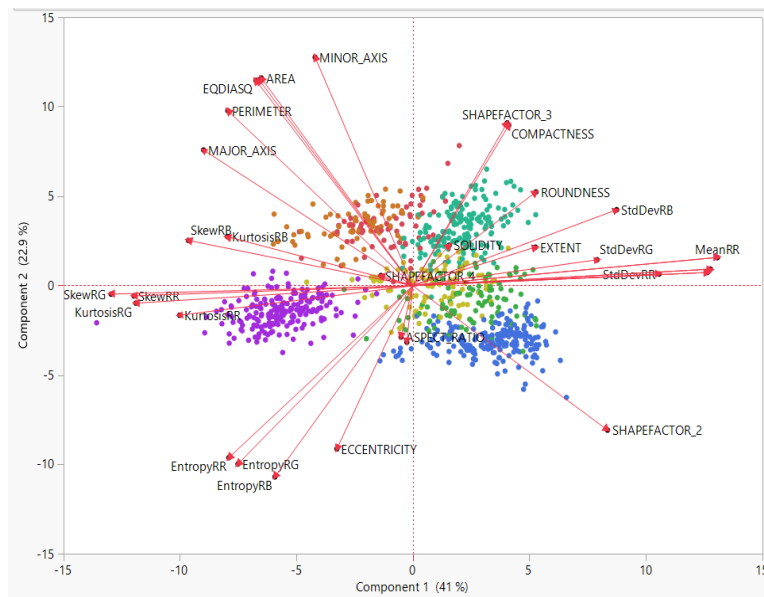


Figure 4. PCA Biplot

## LINEAR DISCRIMINANT ANALYSIS

LDA performs dimensionality reduction to reduce the number of dimensions in the dataset while retaining as much information as possible, utilizing the linear combination of original variables that provide the best possible separation between the groups.

The data was split; 70% for training and 30% for testing. LDA was conducted using all 34 features. This LDA produced a training accuracy of 96% and a test accuracy of 89%. The dimensionality was reduced to 6, with the first 2 dimensions (canonical scores) explaining about 87% variation in the data. However, collinearity in the data can cause some inaccuracies in LDA.

Therefore, some features were removed from the data to eliminate collinearity. With 15 features, another LDA was conducted. The LDA with the reduced features produced training and test accuracies of 92% and 86% respectively, which were relatively good. The first two dimensions were plotted to identify possible groupings, a confusion table was obtained on the test data.

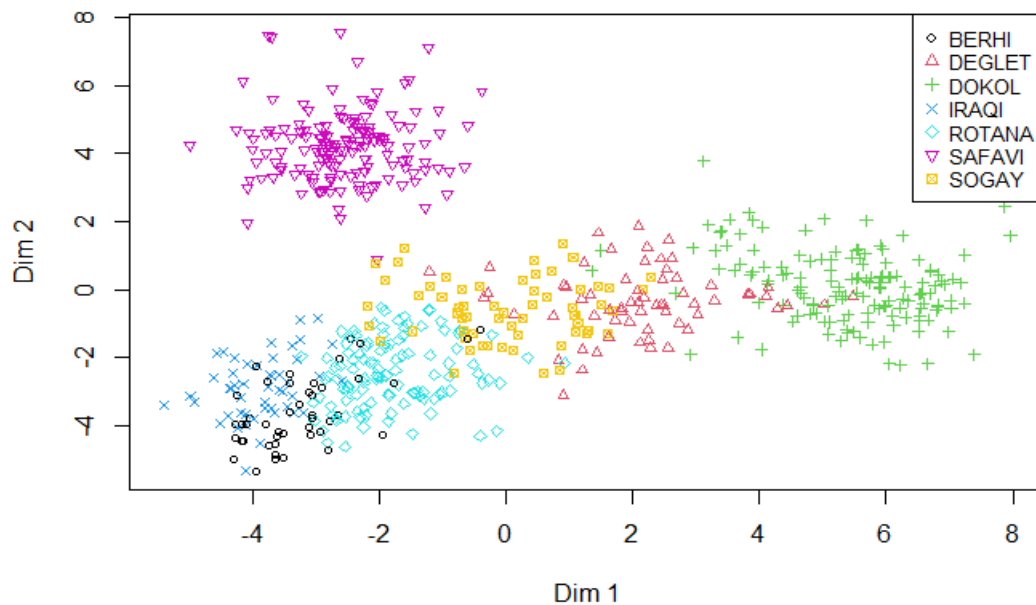


Figure 5. LDA plot of the reduced data

Projecting the points of the first two dimensions, all varieties were well grouped with some slight overlapping. The dimensionality was reduced to 6, with the first two dimensions explaining about 82% variation in the data. Other types of discriminant analyses like the QDA, FDA, MARS, and BRUTO were also explored but they all gave similar results.

	Predicted	Actual						
		BERHI	DEGLET	DOKOL	IRAQI	ROTANA	SAFAVI	SOGAY
	BERHI	16	0	0	1	0	1	0
	DEGLET	0	19	5	0	1	0	4
	DOKOL	0	3	53	0	0	0	0
	IRAQI	1	0	0	21	0	0	0
	ROTANA	1	3	0	1	47	0	0
	SAFAVI	0	0	0	0	0	48	1
	SOGAY	2	9	0	0	4	1	28

Table 2. Confusion Table on the Test data for LDA of the reduced data

## SUPPORT VECTOR MACHINES

A support vector classifier was built using the 15 features selected to eliminate collinearity. Using the same split from the LDA, both linear and radial basis function kernels were explored on the training data. The radial basis function kernel seems to work better on this data. The SVM produced a training accuracy of 97% and a test accuracy of 90%. The classifier was projected on the first two important variables; Perimeter and EntropyRR obtained from random forest. A confusion table was also obtained.



Figure 6. SVM Classification plot



	truth						
predict	BERHI	DEGLET	DOKOL	IRAQI	ROTANA	SAFAVI	SOGAY
BERHI	16	0	0	0	0	0	0
DEGLET	0	22	1	0	1	0	7
DOKOL	0	5	57	0	0	0	1
IRAQI	2	0	0	22	0	0	0
ROTANA	2	3	0	1	50	0	1
SAFAVI	0	0	0	0	0	49	0
SOGAY	0	4	0	0	1	1	24

Table 3. Confusion Table on the Test data for SVM

For interpretability, CART was used as the global surrogate model to interpret the fitted nonlinear SVM model. Using the first two important features, the tree was constrained to have at most a two-way interaction for easy interpretation and then, a tree depth of 2 and 3 were explored and their accuracies compared.

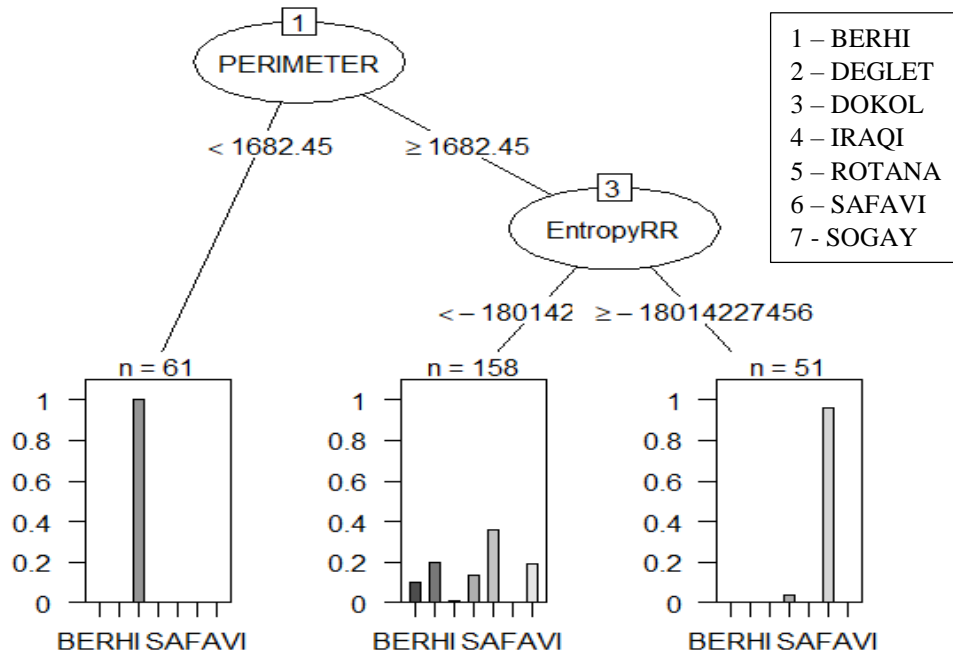


Figure 7. Trees with at most two-way interaction (Global surrogate)

The CART used to interpret the non-linear SVM produced an accuracy of 63% for trees with a maximum depth of 2 and about 74% for trees with a maximum depth of 3.

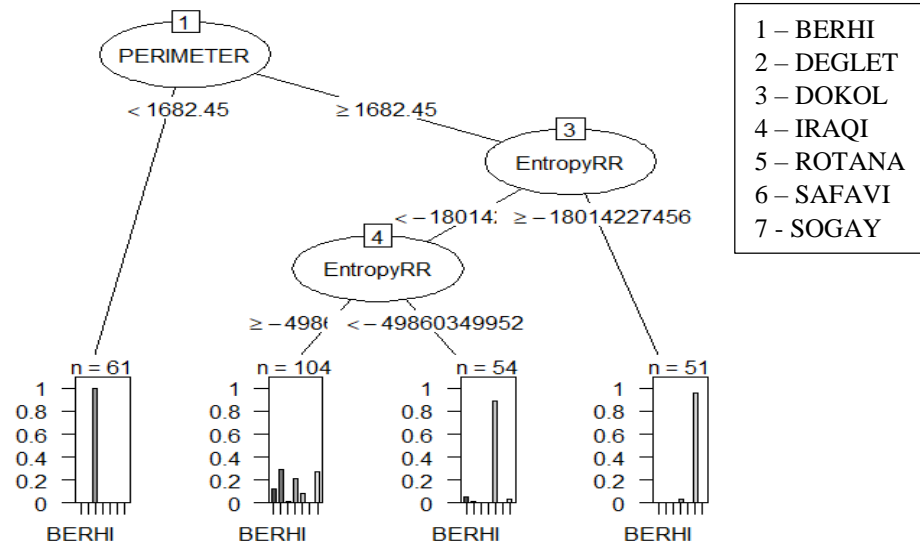


Figure 8. Trees with at most three-way interaction (Global surrogate)

From the global surrogate models, using two features (the 2 most important features) can help classify 2 to 3 varieties substantially, when the maximum depth of the tree was increased from 2 to 3. Introducing more interactions and increasing the maximum number of depths can increase the classification significantly.

## CONCLUSION

The preliminary analysis captured some interesting information on the range of values for some varieties taking into account some pairwise features. The PCA was also able to form some clusters and also reduce the dimensionality of the data from 34 features to just 15 PCs. Using those 15 PCs as new predictors can perform very well for prediction purposes.

From the onset, it was known that classifying some varieties will be more difficult than others. Considering the physical appearance of some varieties like ROTANA, SAFAVI, and DOKOL, their colors, shape, and size help to easily classify them, whilst for other varieties, there are close similarities. Therefore, it makes sense that Perimeter and EntropyRR are the first two most important features. In this data, entropy refers to color image segmentation using Red, Green, and Blue color combinations.

The DEGLET variety is termed as the mother of all dates because of its close similarities to most varieties and can usually be correctly differentiated by its taste. Therefore, looking back at the confusion tables; LDA plot, and the trees, it can be seen that most misclassifications are obtained from the DEGLET variety. The three-way interaction tree also classifies ROTANA, SAFAVI, and DOKOL very well. However, the other varieties are seen to be in the same node.

In general, the LDA and the SVM does well in classifying the 7 varieties of date fruits with SVM performing best.

## REFERENCE








Murat Koklu, Ramazan Kursun, Yavuz Selim Taspinar, Ilkay Cinar, "Classification of Date Fruits into Genetic Varieties Using Image Analysis", Mathematical Problems in Engineering, vol. 2021, Article ID 4793293, 13 pages, 2021. <https://doi.org/10.1155/2021/4793293>

## APPENDIX

**The 15 features selected to eliminate collinearity (Redundant variables were also removed)**

Features				
AREA	PERIMETER	MAJOR_AXIS	MINOR_AXIS	ECCENTRICITY
ROUNDNESS	SHAPEFACTOR_2	SHAPEFACTOR_3	MeanRR	MeanRB
StdDevRB	SkewRB	KurtosisRB	EntropyRR	EntropyRB

### Samples of date fruits in this study and their features

Date fruit type	Images	Color and size	Origins
Barhee		It is amber in color at harvest and then turns a golden-brown color. It is small to medium in size with a hard shell	Basra, Iraq
Deglet Nour		It is a medium- to large-sized date fruits variety that matures from amber to dark brown after harvest	Not specific
Sukkary		It is golden in color and is a medium-sized date fruits variety	Al Qassim region, Saudi Arabia
Rotab Mozafati		It has a full, dark brown appearance. It is a medium-sized and fleshy date variety	Kerman, Iran
Ruthana		It has brown and gold colors. It is a medium-sized date fruit variety	Madinah, Saudi Arabia
Safawi		It has a dark black cherry color and the tips are brown. It is a medium-sized date fruit variety	Madina, Saudi Arabia
Sagai		The tips are dry, golden in color, and the undersides are brown and soft. It is a medium-sized date variety	Arabian Peninsula, especially Saudi Arabia