“Introduction to Machine Learning” – 2020/2021

Elham Babaei (SM3500466)

1. Problem Statement

The aim of this project is to design a classification model that is able to diagnose the species of leaves among the list of available species in *Table1*, based on two groups of features related to the shape and texture of the leaves.

Table 1. List of leaf species

Species	
Quercus suber	Primula vulgaris
Salix atrocinera	Erodium sp.
Populus nigra	Bougainvillea sp.
Alnus sp.	Arisarum vulgare
Quercus robur	Euonymus japonicus
Crataegus monogyna	Ilex perado ssp. Azorica
Ilex aquifolium	Magnolia soulangeana
Nerium oleander	Buxus sempervirens
Betula pubescens	Urtica dioica
Tilia tomentosa	Podocarpus sp.
Acer palmatum	Acca sellowiana
Celtis sp.	Hydrangea sp.
Corylus avellana	Pseudosasa japonica
Castanea sativa	Magnolia grandiflora
Populus alba	Geranium sp.

2. Performance indexes

The following performance indexes are considered in order to measure the quality of the results from a model:

Error *rate:* The proportion of false results among all predicted values for all species.

Sensitivity*:* For each species, the proportion of actual species that is correctly classified.

Precision*:* For each species, the proportion of predicted species that is correctly classified.

F_1_score: The harmonic mean of the sensitivity and precision; a number between 0 and 1.

$$F_1 = 2.\frac{precision * sensitivity}{precision + sensitivity} \qquad (1)$$

AUC*:* The Area Under the Receiver Operating Characteristic (ROC) Curve which is a graph showing the performance of a classification model at all classification thresholds; a number between 0 and 1.

For all the above indexes other than the error rate the larger the better.

3. Proposed solution

As the features from the leaf dataset are grouped by shape and texture, a simple baseline Linear Discriminant Analysis (LDA) model can be made by using the main features chosen from each group. The aim is to find a linear combination of features characterizes the classes of species.

To make a set of predictors for the final model, the other features are added to the baseline model one by one considering whether the error rate decreases.

Afterwards, three different classification techniques including LDA, Random Forest (RF) and Support Vector Machine (SVM) are compared based on the performance indexes proposed in section 2.

4. Experimental Evaluation

4.1. Data Description

The leaf dataset includes 340 rows and 16 columns in which each row represents an observation that has 16 features. The list of features is reported in *Table.2*.

*Table2. List of shape and *texture features*

Features	
Species	Maximal Indentation Depth
Specimen Number	*Lobedness
Eccentricity	*Average Intensity
Aspect Ratio	*Average Contrast
Elongation	*Smoothness
Solidity	*Third moment
Stochastic Convexity	*Uniformity
Isoperimetric Factor	*Entropy

The first step in building a model is *"Feature Selection"* that involves analyzing the features to see if there is any kind of correlation between features, any missing data, any strange values etc. The purpose is to discard irrelevant data that have negative or no effect on the model. In this regard, an auxiliary method can also be used; hence the Regularized Random Forest (RRF) (*[1]* & *[2]*) is applied on the dataset and the importance of each feature is shown in *Figure1*. The features *Solidity* and *Specimens Number* are the most and the least important ones respectively.

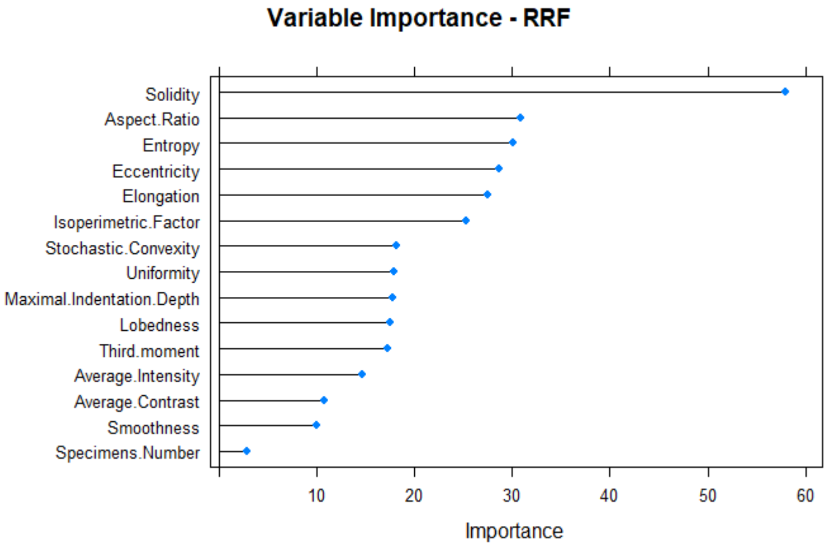

Figure1. Features importance

Figure2 turns out that most of the species have the same number of specimens which means it will not significantly impact the classifier and can be omitted. Therefore, the variable *Specimens number* is not taken into account to build the model.

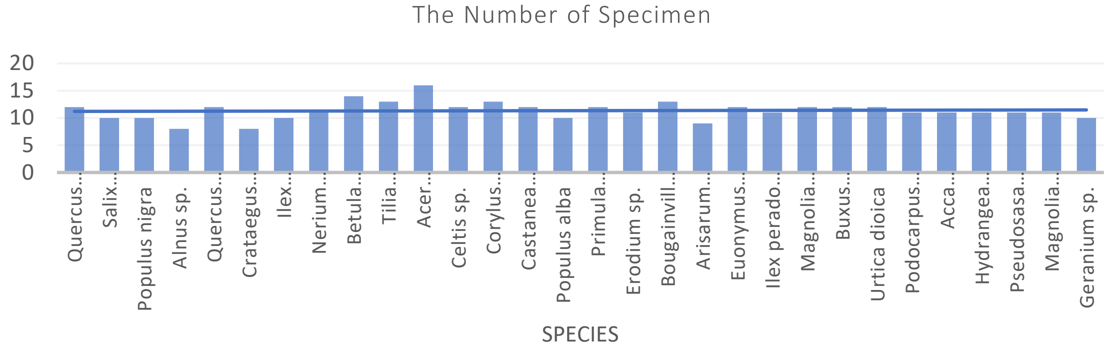

Figure2. The number of specimens for each species

4.2. Experimental Procedure

After refining the dataset as explained in section 4.1, the baseline LDA model is built by using the following features: *Solidity*, *Aspect Ratio*, *Elongation*, *Average Intensity*, *Smoothness*, and *Uniformity*. Then the other features are added to the model one by one considering whether a 10-fold Cross Validation error rate decreases. The model ended up involving all the remaining features.

Thereafter the algorithms LDA, RF, and SVM (linear kernel) are implemented and compared.

4.3. Results

The results of the mean and the standard deviation of 10-fold Cross Validation error rate are shown in *Table3*. It shows that the baseline model has been improved by adding more predictors, and LDA seems to have a better performance to diagnose the species; but it is not enough and other performance indexes are also taken into account.

Table3. 10-fold Cross Validation error rate

Model	Error rate	Standard deviation
Baseline LDA	0.34	0.08
RF	0.22	0.05
SVM	0.24	0.06
LDA	0.2	0.06

The models have been implemented on a test dataset (unseen in learning phase) chosen randomly from the leaf dataset. The values of all indexes are shown in *Tables 4* and *5*.

From *Table4* it can be understood that LDA is the prevalent algorithm since it provides a greater F_1_score in average (0.83); however, it is dominated by other methods in some cases. For example, RF works great to classify *Erodium sp* species (F1=1) while SVM and LDA are less good (F1=0.86 and F1=0.80 respectively). On the other hand, RF is unable to classify *Tilia tomentosa* species whereas SVM and LDA do it pretty well (F1=1).

As another instance, LDA is the strongest one to classify *Pseudosasa japonica* (F1=0.8), on the contrary both RF and SVM are in trouble as they represent F1< 0.6 which is not much acceptable.

Table4. Sensitivity, Precision, and F1 Score values

Species	RF			SVM			LDA		
	Sen	**Pre**	**F1**	**Sen**	**Pre**	**F1**	**Sen**	**Pre**	**F1**
Quercus suber	1	0.33	0.5	1	0.33	0.5	1	0.33	0.5
Salix atrocinera	1	0.4	0.57	1	0.4	0.57	1	0.5	0.67
Populus nigra	1	1	1	0.5	1	0.67	1	0.8	0.89
Alnus sp	1	1	1	1	1	1	1	1	1
Quercus robur	1	1	1	1	1	1	1	0.67	0.8
Crataegus monogyna	1	1	1	1	1	1	1	1	1
Ilex aquifolium	0.67	0.67	0.67	0.67	0.67	0.67	0.33	1	0.5
Nerium oleander	1	1	1	1	1	1	1	1	1
Betula pubescens	0.75	0.75	0.75	0.75	0.75	0.75	1	0.8	0.89
Tilia tomentosa	0	NA	NA	1	1	1	1	1	1
Acer palmatum	1	1	1	1	1	1	1	1	1
Celtis sp.	1	0.5	0.67	1	1	1	1	1	1
Corylus avellana	0.5	1	0.67	1	0.67	0.8	1	0.67	0.8
Castanea sativa	0.75	0.6	0.67	0.75	0.38	0.5	0.75	0.6	0.67
Populus alba	1	1	1	1	1	1	1	1	1
Primula vulgaris	1	1	1	1	1	1	1	1	1
Erodium sp.	1	1	1	1	0.75	0.86	0.67	1	0.8
Bougainvillea sp.	0.5	1	0.67	0.5	1	0.67	0.5	1	0.67
Arisarum vulgare	1	1	1	1	0.5	0.67	1	1	1
Euonymus japonicus	1	1	1	1	1	1	1	1	1
Ilex perado ssp.azorica	0.67	1	0.8	0.67	1	0.8	0.67	1	0.8
Magnolia soulangeana	0.4	0.67	0.5	0.2	0.5	0.29	0.6	0.75	0.67
Buxus sempervirens	1	1	1	1	1	1	1	1	1
Urtica dioica	1	1	1	0.67	1	0.8	0.67	1	0.8
Podocarpus sp.	0.5	0.5	0.5	0.5	0.5	0.5	0.5	1	0.67
Acca sellowiana	0.25	1	0.4	0.25	1	0.4	0.5	1	0.67
Hydrangea sp.	1	0.5	0.67	1	1	1	1	1	1
Pseudosasa japonica	0.5	0.5	0.5	0.5	0.5	0.5	1	0.67	0.8
Magnolia grandiora	1	1	1	0.5	1	0.67	0.5	0.5	0.5
Mean			**0.78**			**0.78**			**0.83**

The other performance index AUC in *Table5* offers RF as the best model. Considering the fact that the leaf dataset is not unbalanced and also ROC curve covers all the possible thresholds for classification, AUC might be more preferred than F1 to evaluate the models. Thus, it can be claimed that the good classifiers are RF, LDA, and SVM in order.

Table5. AUC values

Model	AUC
RF	0.93
SVM	0.89
LDA	0.89

References

[1] [*https://ieeexplore.ieee.org/abstract/document/6252640*](https://ieeexplore.ieee.org/abstract/document/6252640)

[2] https://www.machinelearningplus.com/machine-learning/feature-selection/#:~:text=In%20machine%20learning%2C%20Feature%20selection,important%20when%20building%20predictive%20models.