Python for Data Analysis : Final Project

Statlog (Landsat Satellite) Data Set

BERTONE Enora (A4-ESILV, DIA1)

CONTEXT

The database consists of the multi-spectral values of pixels in 3x3 neighbourhoods in a satellite image, and the classification associated with the central pixel in each neighbourhood.

The aim is to predict this classification, given the multi-spectral values.

In the sample database, the class of a pixel is coded as a number.

Problem type : Classification

Source : Ashwin Srinivasan

FEATURES

The dataset is composed of 37 features as follows:

- 36 features for the spectral band of each pixel (4 spectral band x 3 x 3)
- 1 features for the classification of the central pixel
- There is no empty data in the dataset

CLASSIFICATION

The central pixel can be classified in 7 classes:

- Red Soil
- Cotton crop
- Grey soil
- Damp grey soil
- Soil with vegetable stubble
- Mixture class (all types present)
- Very damp grey soil

We can see that in our dataset we don’t have any « Mixture class (all types present) »

DATA VIZUALIZATION

	Soil_Type	counts
0	Cotton crop	479
1	Damp grey soil	415
2	Grey soil	961
3	Red soil	1072
4	Soil with vegetation stubble	470
5	Very damp grey soil	1038

As we said before we can’t find any Mixture class.

We can also see that there is much more « Grey soil », « Red soil » and « Very damp grey soil »

So the result for this class will be more accurate.

DATA MODELIZATION

	Accuracy
Decision Tree Classifier	0.851217
Gaussian Naive Bayes	0.814247
Random Forest Tree	0.881817
Epsilon-Support Vector Regression	0.877348
KMeans	0.172227

We can see that the Kmeans algorythm is very bad.

The other algorithm are all in the same range but the best one is the Random Forest Tree with an accuracy of 0,88.

CONCLUSION

I’ve had some difficulty for the vizualization of the data and for use hyper parameters and improve the algorithm.

But this have been a great experience, I think that I will try to continue this project later.