Presentation - Python for Data Analysis

Statlog (Landsat satellite) Data Set

Morgan LEBLOND & Mehdi LEQSIOUER
DIA3

ESILV
ENGINEERING SCHOOL
DE VINCI PARIS

Ins and outs

At the beginning:

- 2 files: a training set (4435 rows) and a testing set (2000 rows)
- Features: only integers in the file (with values between ~20 and ~160)
- Target: integers between 1 and 7

At the end:

- Objective: Get the prediction of the target for the testing file based on a RandomForest model
- Accuracy: above 90%
- Visualizations

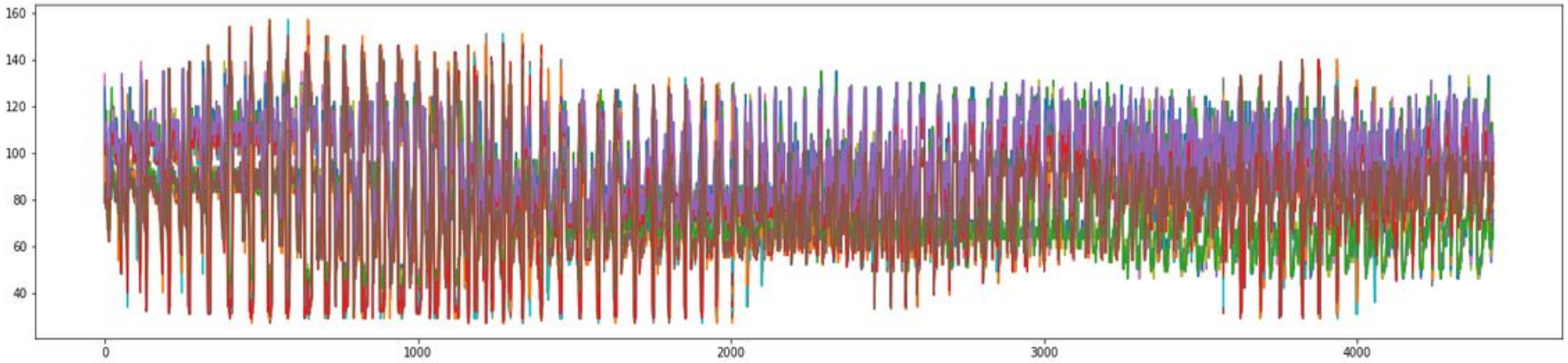
```
Accuracy: 0.9115
```

Sample of datasets

	0 int64	1 int64	2 int64	3 int64	4 int64	5 int64	6 int64	7 int64	8 int64	9 int64	10 int64	11 int64	12 int64	13 int64
0	92	115	120	94	84	102	106	79	84	102	102	83	101	126
1	84	102	106	79	84	102	102	83	80	102	102	79	92	112
2	84	102	102	83	80	102	102	79	84	94	102	79	84	103
3	80	102	102	79	84	94	102	79	80	94	98	76	84	99
4	84	94	102	79	80	94	98	76	80	102	102	79	84	99

5 rows × 37 columns

Plot of values' range of the training set



X: Number of the row (like an ID)
Y: Range of the values for all column

Classification models

- ~~Bagging~~
- ~~Boosting~~
- ~~Logistic regression~~
- **Random Forest: best compromise between execution time, accuracy and complexity**
- ~~Decision tree~~
- ~~Naive Bayes~~

List of variables

training_set = dataframe containing the training dataset

testing_set = dataframe containing the training dataset

X_test = dataframe containing all the features of the testing dataset

X_train = dataframe containing all the features of the training dataset

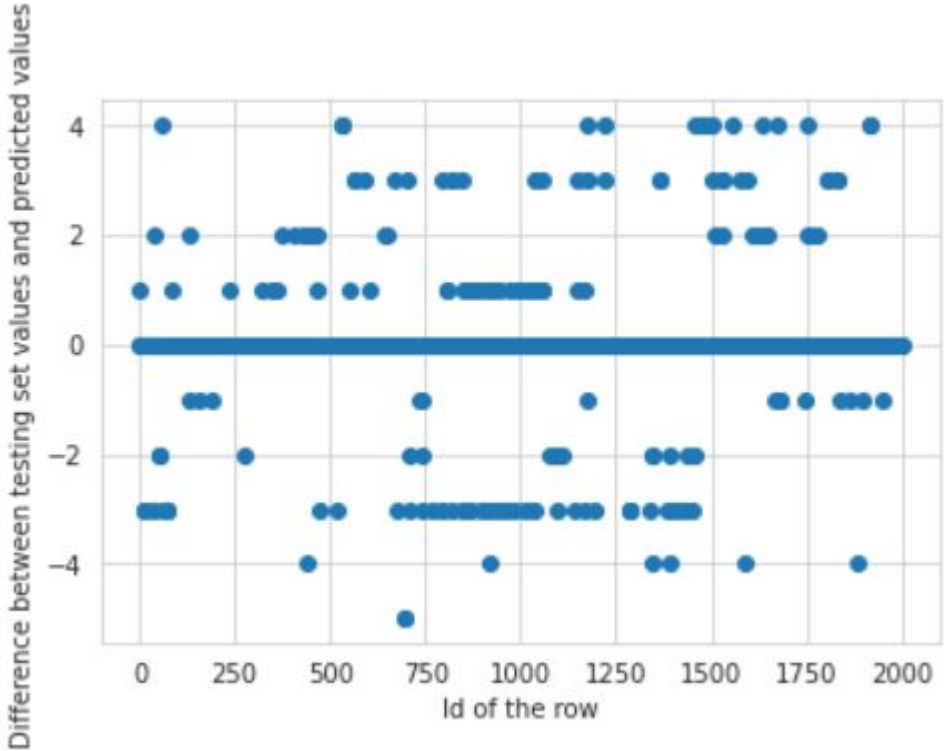
classifier = our Random Forest Classifier

y_pred = ndarray containing all the predicted targets

y_test = pandas Series containing the targets of the testing dataset

y_train = pandas Series containing the targets of the testing dataset

Results



Accuracy: 0.9115