AIMS African Institute for Mathematical Sciences SENEGAL

MATHEMATICAL SCIENCES

DATA MINING AND BIG DATA ANALYTICS

REPORT PROJECT

Student's Names : GROUP 6

1. **WINNIE KADZO YAA**
2. **GOD'SPOWER EMMANUEL OKON**
3. **SANDRA MARION KAM TSEMO**
4. **AMISI FIKIRINI**
5. **ELIE RENE MULAMBA**
6. **KALIDOU ALIOU BALL**

Lecturer :

Prof. NDEYE NIANG KEITA

Academic year 2021-2022

Table des matières

INTRODUCTION

In this work, a public dataset has been targeted : the multi-feature dataset. This dataset consists of features of handwritten numerals ('0'–'9'). There are 200 patterns for each numeral (a total of 2,000 patterns) that have been digitized in binary images. These digits are represented the following six feature sets :

1. mfeat-fou : 76 Fourier coefficients of the character shapes ;
2. mfeat-fac : 216 profile correlations ;
3. mfeat-kar : 64 Karhunen-Love coefficients ;
4. mfeat-pix : 240 pixel averages in 2 x 3 windows ;
5. mfeat-zer : 47 Zernike moments ;
6. mfeat-mor : 6 morphological features.

We have two important tasks to do on the previous data in this project :

- First, we will do an exploratory analysis using clustering which is one of the most used methods in unsupervised learning. To do this, we will perform a PCA for each variable in order to extract the principal components : those that contribute the most to the inertia pointed by the principal axes. These variables will then be used to form a new dataset on which we will perform the k-means. **(Unsupervised Mining)**

- Secondly, after the exploratory analysis, we will perform a supervised classification to rank the handwritten figures. The objective is to predict the values taken by a nominal qualitative target variable CLASS with 10 modalities. **(Supervised Mining)**

Chapitre un

UNSUPERVISED MINING

In this section, the aim is to carry out an exploratory study including :

- Performing several PCAs considering the groups of indicators separately. The factorial designs will then be compared with the results obtained on all the variables.

- Carry out several classifications of individuals using a geometric method (e.g. hierarchical classification, by partitioning), considering the groups of indicators separately. This classification should then be compared with the classification obtained on all variables.

1.1 Exploration of all data

For the data exploration, we started by importing the data, then we visualised the data to try to see the description of each variable (mean, standard deviation, variance etc.). To check the correlation between the variables, we used a graph that explained the lineal relationship between the variables and the outliers. This gives reason to make a standardization in order to put all the data on the same scale (in the interval 0 and 1). We then applied the PCA to obtain the number of principal axes on which we projected the individuals to check the correlation between each individual and each component.

1.2 Grouped table of the six data sets

Var	Variables	Number of axis (Kaiser)	Number of axis (Elbow)	Principal component	Clusters (Kmeans)
mfeat-fou	76	20	2	2	10
mfeat-fac	216	27	3	3	10
mfeat-kar	64	20	3	3	10
mfeat-pix	240	33	3	3	10
mfeat-zer	47	11	3	3	8
mfeat-mor	6	2	2	2	10

S/N	Data	Number of Factorial Axis	Proportions
1	data_fmeat_fac	3	43.63%
2	data_fmeat_kar	3	22.60%
3	data_fmeat_mor	2	76.20%
4	data_fmeat_pix	3	34.66%
5	data_fmeat_fou	3	22.44%
6	data_fmeat_zer	3	48.24%
7	L'ensemble du dataset	4	42.14%

Working with each dataset we can see that we have 3 factorial axes for some and only two for the morphological data (mfeat-mor) and (mfeat-fou). With 3 axes it is not possible to see the scatterplot in the plane. In the same way, if we make the PCA of the global dataset we retain 4 factorial axes and with 4 factorial axes it is also impossible to see the point cloud in the plane. Therefore, the morphological data (mfeat-mor) and (mfeat-fou) are the ones that give a vizualisation of the two dimensional scatterplots.

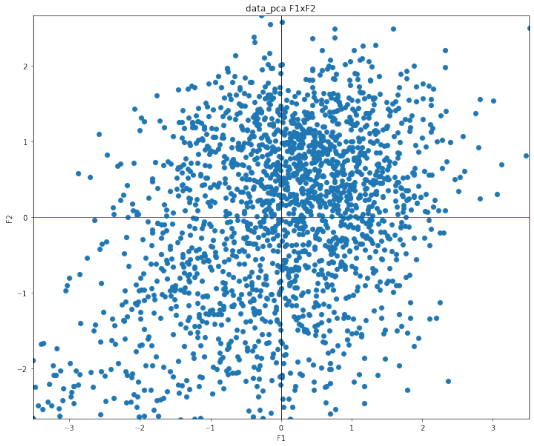


FIGURE 1.1 – Graphical representation of individuals along F1 and F2 axis

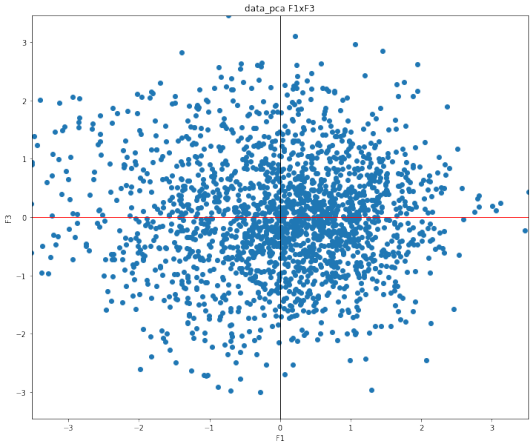


FIGURE 1.2 – Graphical representation of individuals along F1 and F3 axis

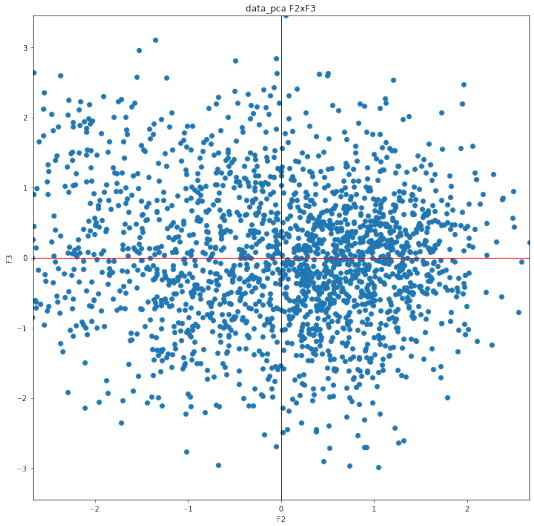


FIGURE 1.3 – Graphical representation of individuals along F1 and F4 axis

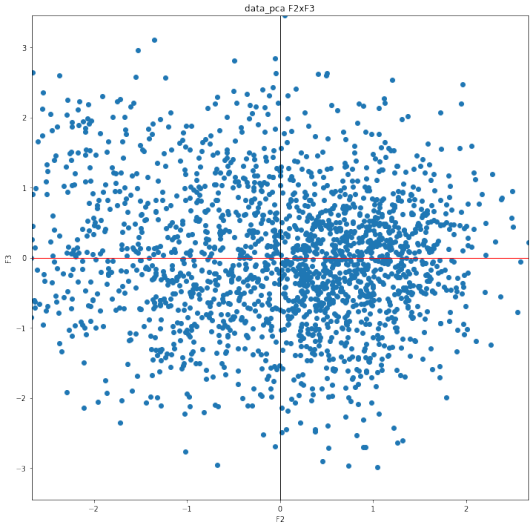


FIGURE 1.4 – Graphical representation of individuals along F2 and F3 axis

Chapitre Deux

SUPERVISED MINING

The aim is to predict the class variable which values in intervals from 0 to 9 corresponding to the handwritten digits shown in the image. This interval is the target that we have added in the general dataset which contains initially 6 variables.

Here are the steps that we followed :

1. Combine the 6 variables to get one whole dataset
2. We standardised the data so that we can put the data in the same scale.
3. We added back the target value to the dataset.
4. We split the dataset into two parts(70per cent training data and 30per cent Testing data)
5. We applied the classification methods. In our case we used two methods for comparison of the results. These methods are Neural Networks and Random Forests. The methods were selected because of the following reasons :

a. **Random Forests**. Is a set of decision trees where each tree is independent and has a part of the dataset as a sample and can predict the output. We have chosen it because it works very well with tabular data(data in table format) and initially our data was in table form. RF works independently.

b. **Neural Networks**. It works with neurals grouped in layers and processes data in each layer. Each layer forwards to the next layers and the last layer makes decisions. NN works dependently.

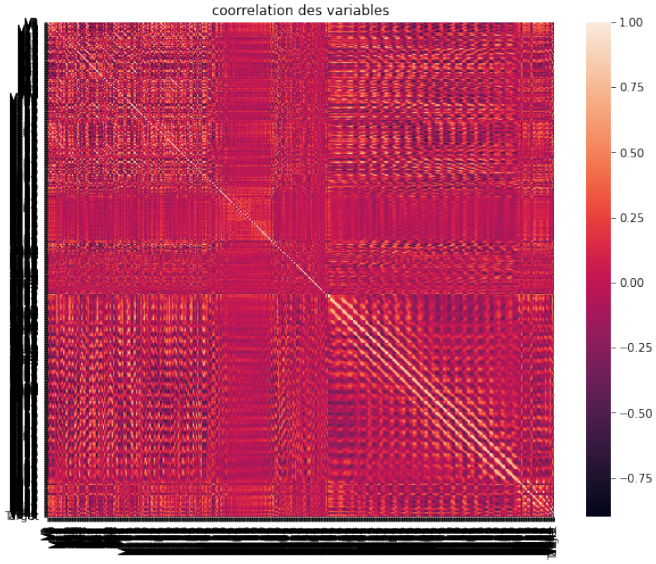


FIGURE 2.1 – Correlation of variables

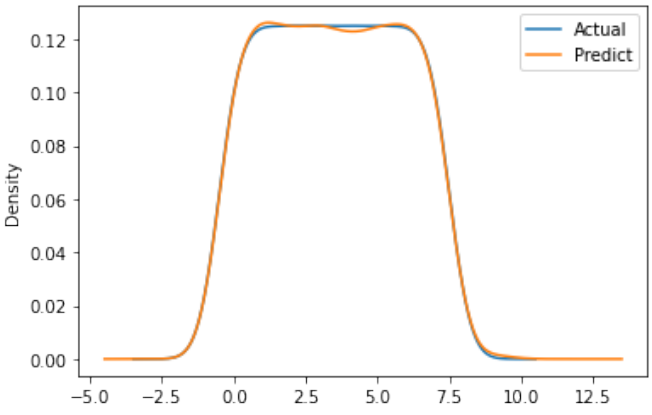
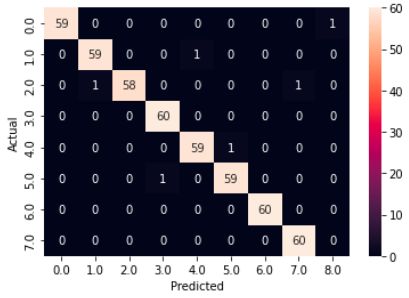


FIGURE 2.2 – Evolution of the density of predicts and Trues values



	precision	recall	f1-score	support
0.0	1.00	0.98	0.99	60
1.0	0.98	0.98	0.98	60
2.0	1.00	0.97	0.98	60
3.0	0.98	1.00	0.99	60
4.0	0.98	0.98	0.98	60
5.0	0.98	0.98	0.98	60
6.0	1.00	1.00	1.00	60
7.0	0.98	1.00	0.99	60
8.0	0.00	0.00	0.00	0

FIGURE 2.3 – Confusion matrix

CONCLUSION

In part 1 :

After running Kmeans on all our groups and on our initial dataset, we notice that among these groups, it is mfeatmor that classifies each number well. And the initial dataset classifies more than all these groups.

In part 2 :

In this project we presented case studies using the following classification models : Neural Network and Random Forest. All these models seem to give good scores of around 0.99. In the case of the discriminant analysis, the high score does not allow any conclusion. This is why the exploration of other methods confirm the results of the discriminant analysis model. Also when we look at the density between the predicted values and the true values we notice an almost identical pattern.