Tenants et aboutissants du problème

"[The] main goal of [the] study is not that of building a top performing recognition system, but rather to verify that the use of page layout features allows obtaining satisfactory results."
Ainsi, nous avons essayé plusieurs modèles de classification afin de trouver un modèle plus précis, et nous avons finalement choisi RandomForest Classifier.

"[A secondary objective was to] perform a statistical analysis of the considered features in order to characterize the discriminating power of each of them."
Ici, nous avons déterminé que les deux variables aux plus bas scores n'étaient pas nécessaires à l'analyse, et en les retirant, nous obtenons une meilleure précisions (de 99,765% à 99,817%).

Table 3. Feature ranking according to the five considered measures. For each row, the most left numeric value indicates the best feature, while the most right value denotes the worst one.

Measure	Ranking
Chi Squared (C_S)	4 3 2 1 5 9 7 6 10 8
Relief (R_F)	5 4 1 9 3 7 6 10 8 2
Gain Ratio (I_R)	4 5 1 3 2 9 7 6 10 8
Information Gain (I_G)	4 3 2 1 5 9 7 6 10 8
Symmetrical Uncertainty (I_S)	4 3 5 1 2 9 7 6 10 8

Table 4. Overall ranking of the features.

id	feature	score
4	exploitation	44
3	lower margin	35
5	row number	34
1	intercolumnar distance	32
2	upper margin	24
9	peak number	22
7	interlinear spacing	16
8	weight	16
6	modular ratio	11
10	modular ratio/interlinear spacing	6

Tenants et aboutissants du problème

"Abstract:
The Avila data set has been extracted from 800 images of the 'Avila Bible', an XII century giant Latin copy of the Bible.
The prediction task consists in associating each pattern to a copyist."

Data Set Characteristics:	Multivariate	**Number of Instances:**	20867	**Area:**	Computer
Attribute Characteristics:	Real	**Number of Attributes:**	10	**Date Donated**	2018-06-20
Associated Tasks:	Classification	**Missing Values?**	N/A	**Number of Web Hits:**	36519

Les diapositives suivantes représentent les caratéristiques initales de notre dataset.

Tenants et aboutissants du problème

La description de ce dataset ne rentrant pas dans des détails suffisants, nous avons donc cherché son origine, et nous avons trouvé l'article de recherche associé à ce dataset.

https://www.researchgate.net/publication/221356167_A_Method_for_Scribe_Distinction_in_Medieval_Manuscripts_Using_Page_Layout_Features

Nous avons donc déduit que chaque ligne de données (tuple) correspond à M lignes, avec M = 4 dans les cas de ce dataset.

Les données proviennent d'une bible écrite avec 2 colonne par page

Tenants et aboutissants du problème

Il y a 3 sets d'attributs

1. Les attributs de la page
upper margin , lower margin , intercolumnar distance

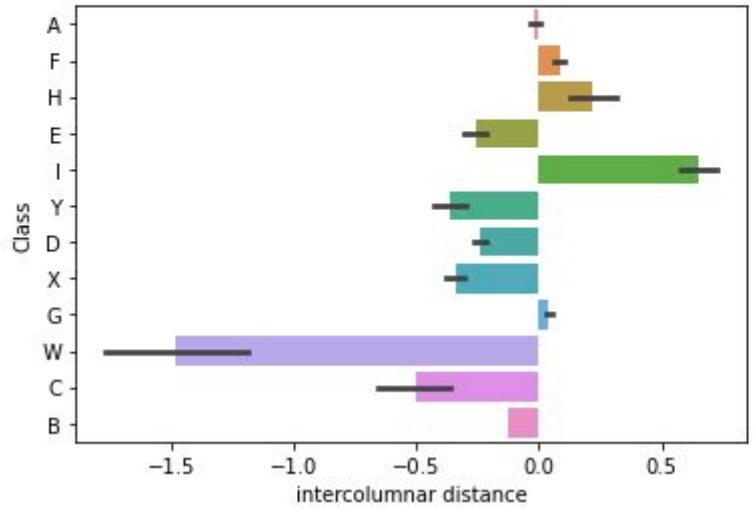
2. Les attributs de la colonne
number of rows in the column and
column exploitation coefficient

3. Les attributs de la ligne
weight, modular ratio, interlinear spacing,
modular ratio/interlinear spacing ratio and peaks number

Tenants et aboutissants du problème

```
import seaborn as sns
sns.barplot(y=data["Class"],x=data["intercolumnar distance"])
```

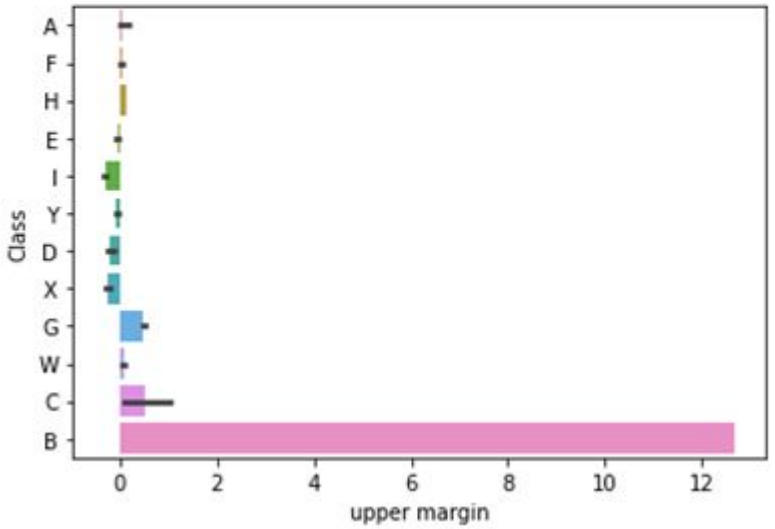
```
<AxesSubplot:xlabel='intercolumnar distance', ylabel='Class'>
```



Distance intercolumnaire

Tenants et aboutissants du problème

```
sns.barplot(y=data,x=data["upper margin"])
<AxesSubplot:xlabel='upper margin', ylabel='Class'>
```

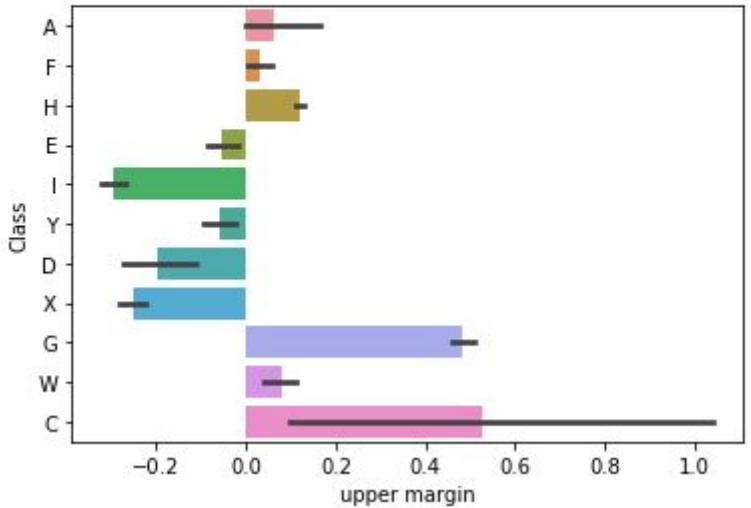


Marge supérieure

Tenants et aboutissants du problème

```
t=data[(data.Class!="B")]
sns.barplot(y=t["Class"],x=t["upper margin"])
```

```
<AxesSubplot:xlabel='upper margin', ylabel='Class'>
```

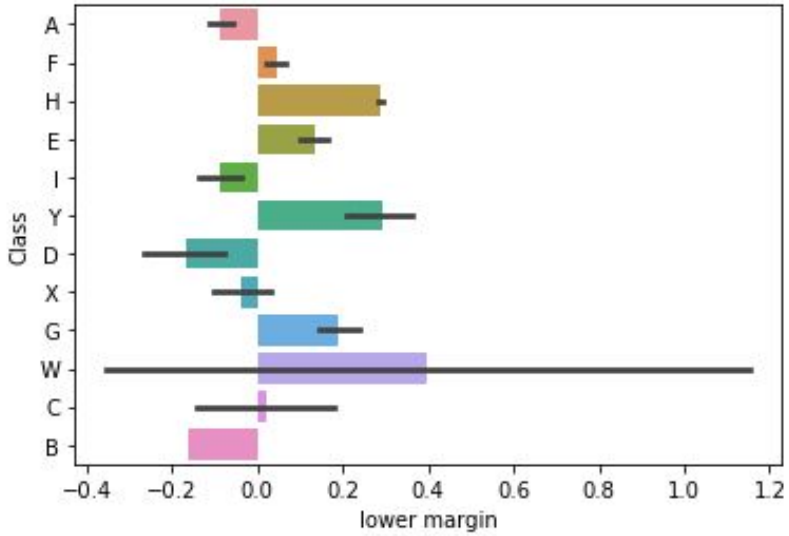


Marge supérieure, en excluant la classe B ne comportant que 5 paragraphes M

Tenants et aboutissants du problème

```
sns.barplot(y=data["Class"],x=data["lower margin"])
```

```
<AxesSubplot:xlabel='lower margin', ylabel='Class'>
```

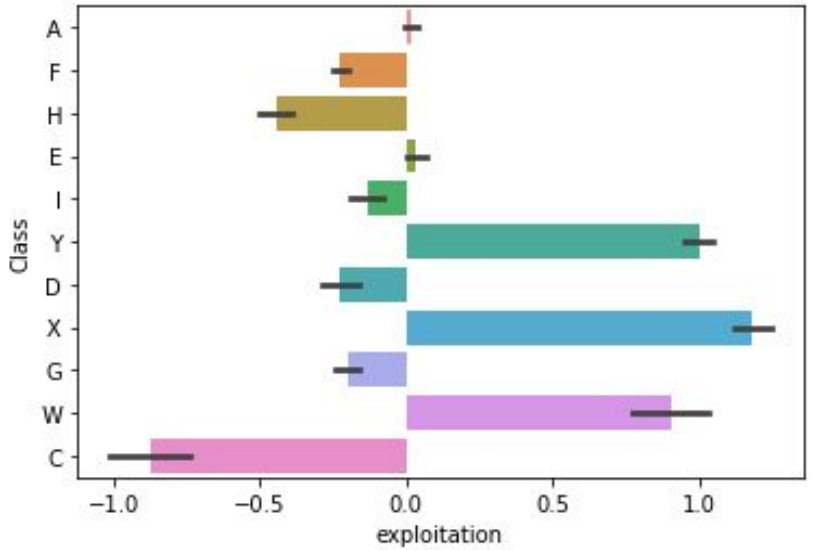


Marge inférieure

Tenants et aboutissants du problème

```
sns.barplot(y=t["Class"],x=t["exploitation"])
```

```
<AxesSubplot:xlabel='exploitation', ylabel='Class'>
```

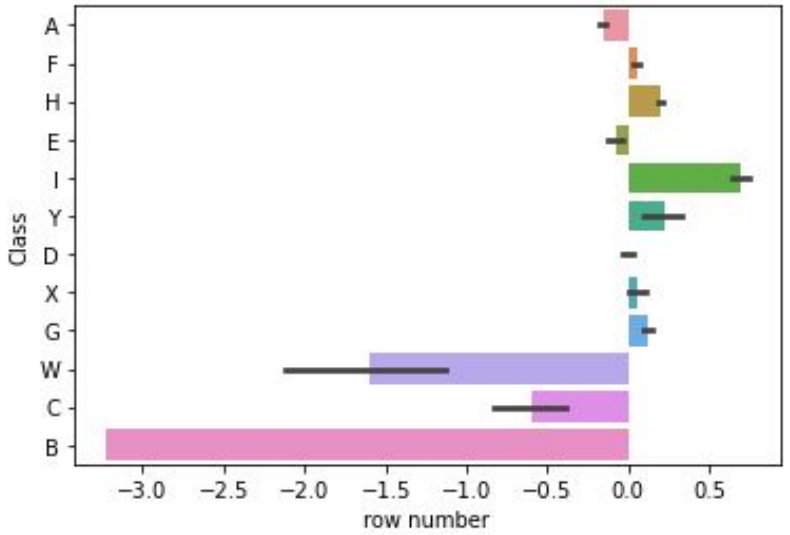


Exploitation (à quel point une colonne est remplie d'encre), en excluant B

Tenants et aboutissants du problème

```
sns.barplot(y=data["Class"],x=data["row number"])
```

```
<AxesSubplot:xlabel='row number', ylabel='Class'>
```

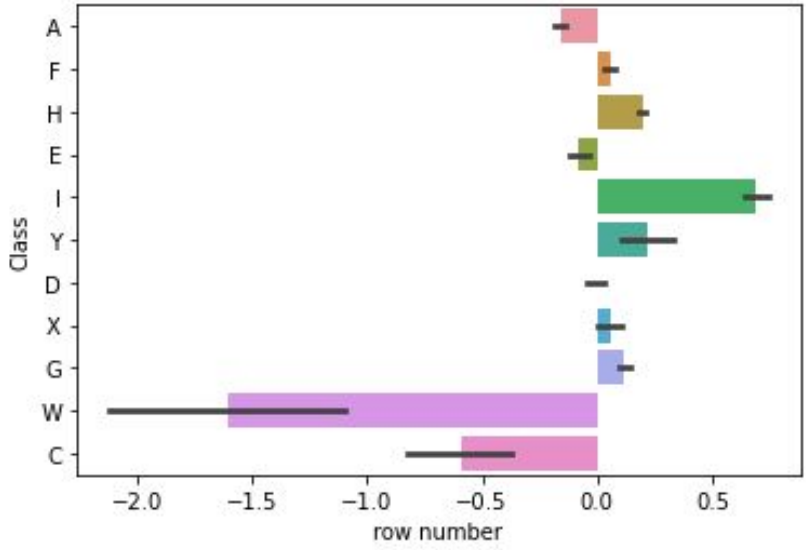


Nombre de ligne (centré, d'où les nombres négatifs)

Tenants et aboutissants du problème

```
sns.barplot(y=t["Class"],x=t["row number"])
```

```
<AxesSubplot:xlabel='row number', ylabel='Class'>
```

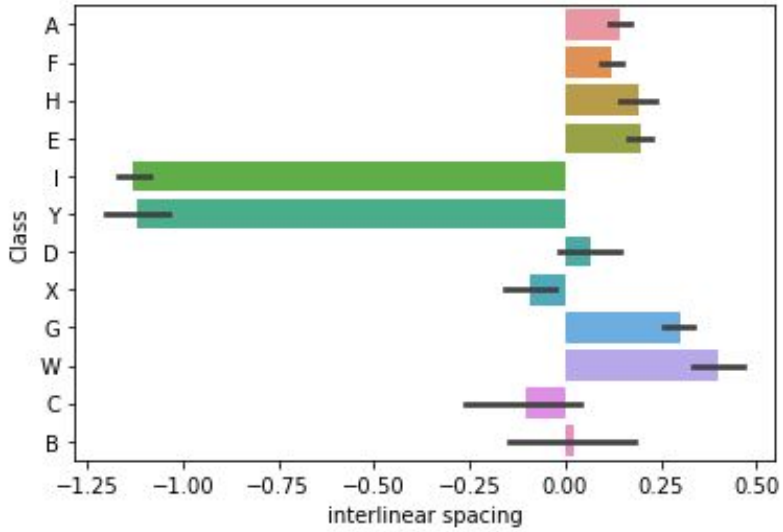


Nombre de ligne (centré), en excluant B

Tenants et aboutissants du problème

```
sns.barplot(y=data["Class"],x=data["interlinear spacing"])
```

```
<AxesSubplot:xlabel='interlinear spacing', ylabel='Class'>
```

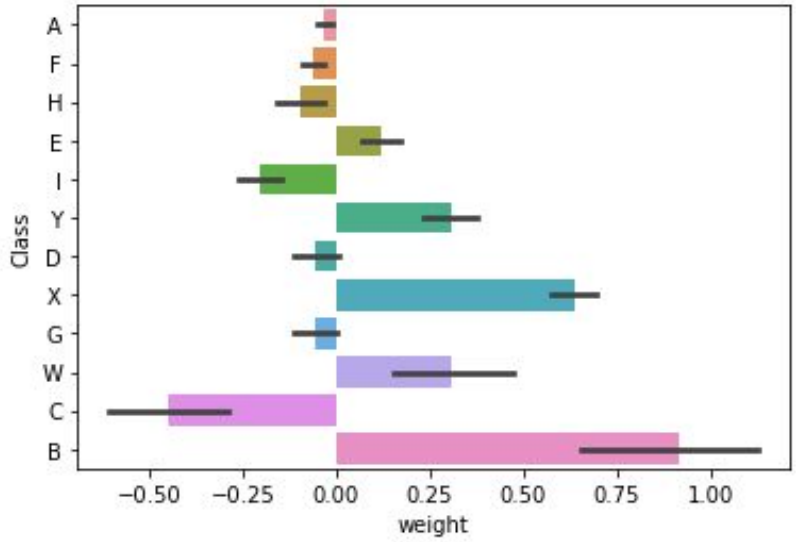


Taille d'interligne

Tenants et aboutissants du problème

```
sns.barplot(y=data["Class"],x=data["weight"])
```

```
<AxesSubplot:xlabel='weight', ylabel='Class'>
```

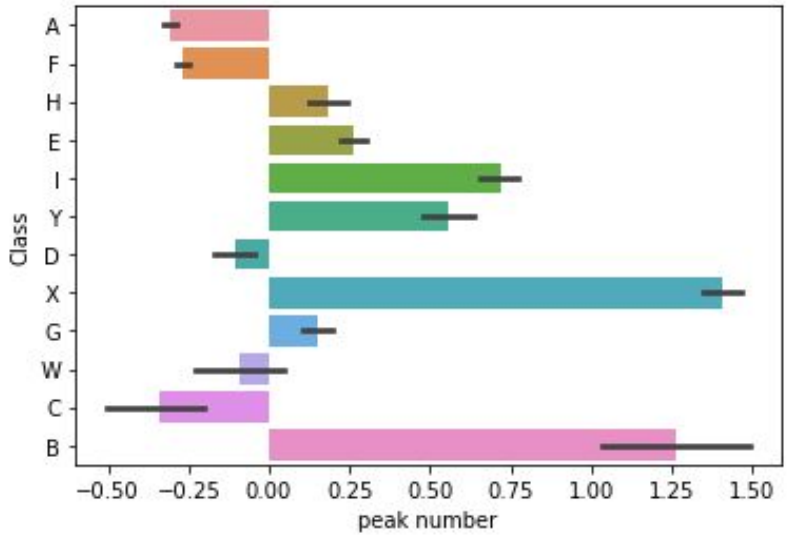


Poids (équivalent de l'exploitation, mais pour les lignes)

Tenants et aboutissants du problème

```
sns.barplot(y=data["Class"],x=data["peak number"])
```

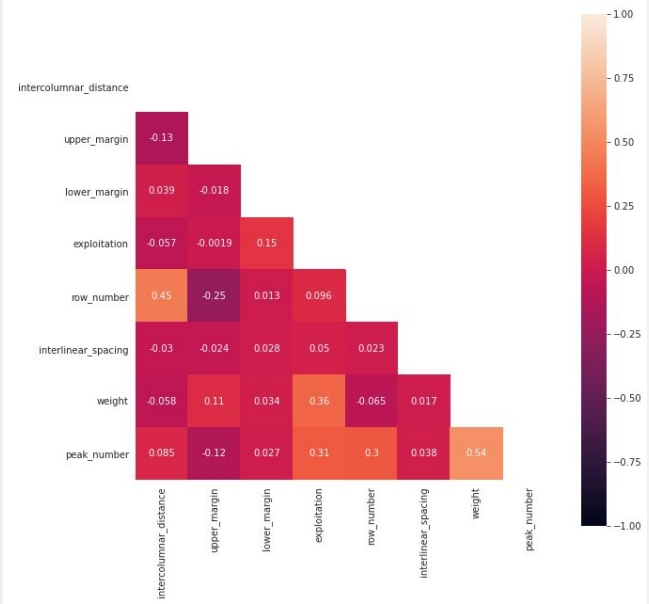
```
<AxesSubplot:xlabel='peak number', ylabel='Class'>
```



Nombre de pics (dans la projection horizontale de l'histogramme d'un tuple)

Tenants et aboutissants du problème

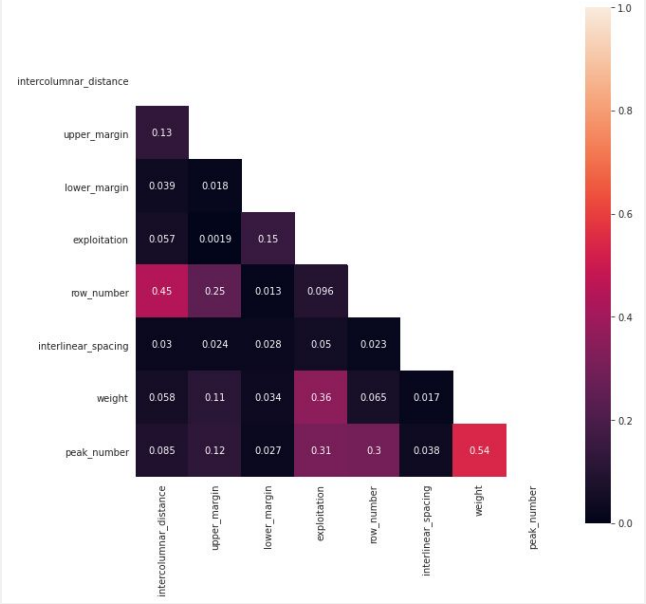
```
corr = data.corr()
mask = np.zeros_like(corr)
mask[np.triu_indices_from(mask)] = True
with sns.axes_style("white"):
    f, ax = plt.subplots(figsize=(10, 10))
    ax = sns.heatmap(corr, mask=mask, square=True, annot=True, vmin=-1, vmax=1)
```



Heatmap

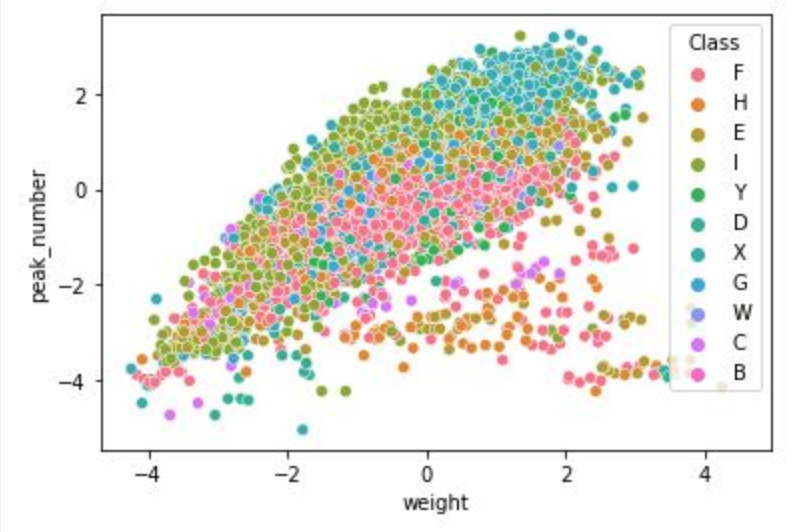
Tenants et aboutissants du problème

```
corr = data.corr().abs()
mask = np.zeros_like(corr)
mask[np.triu_indices_from(mask)] = True
with sns.axes_style("white"):
    f, ax = plt.subplots(figsize=(10, 10))
    ax = sns.heatmap(corr, mask=mask, square=True, annot=True, vmin=0, vmax=1)
```



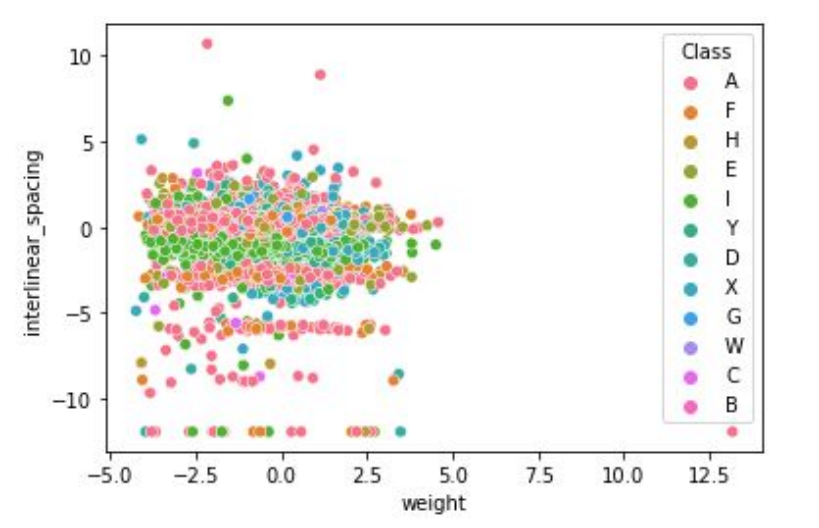
Heatmap (des valeurs absolues)

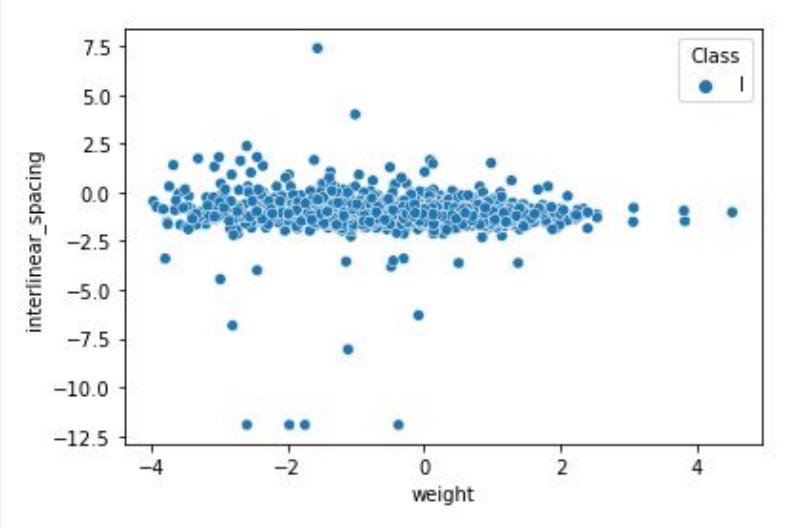
Tenants et aboutissants du problème



Grâce à nos heatmap, on peut isoler les corrélations,
que l'on peut visualiser avec des scatterplots.
Dans cet exemple, on peut observer ce qui semble être une corrélation linéaire,
entre le nombre de pics et le poids.

Tenants et aboutissants du problème





Sinon, on peut comparer des sous-graphes à leurs graphes d'origine.
Dans cet exemple, et en considérant notre requête "à quelle classe appartient ce (nouveau) tuple",
on peut conjecturer que quel que soit le poids d'un tuple,
un espace d'interligne compris entre 0 et -2 indique la possibilité d'appartenance à la classe I.

Variables créées

Colonnes trompeuses

Table 4. Overall ranking of the features.

id	feature	score
4	exploitation	44
3	lower margin	35
5	row number	34
1	intercolumnar distance	32
2	upper margin	24
9	peak number	22
7	interlinear spacing	16
8	weight	16
6	modular ratio	11
10	modular ratio/interlinear spacing	6

Suite à l'étude initial et confirmer par nos testes nous avons décidé de ne pas prendre en compte les caractéristique de "modular ratio" et "modular ratio/interlinear spacing"
Ce changement permet des résultat clairement supérieur.

Variables créées

Valeur suspicieuse

```
datatr.sort_values(by="lower margin",ascending=False)
```

	intercolumnar distance	upper margin	lower margin	exploitation	row number	modular ratio	interlinear spacing	weight	peak number	modular ratio/ interlinear spacing	Class
6619	0.000000	386.000000	50.000000	0.168104	0.000000	53.000000	83.000000	0.275032	44.000000	0.638020	A
10199	-3.498799	-0.063555	7.458681	0.129002	-4.922215	0.148790	0.031425	-1.382921	-1.619716	0.240139	W
3705	-3.498799	-0.063555	7.458681	0.129002	-4.922215	1.145386	0.861934	-0.567979	-0.434820	0.692291	W
1798	-3.498799	-0.063555	7.458681	0.129002	-4.922215	0.024215	0.635431	-0.009106	-0.029461	-0.200211	W
818	-3.498799	-0.063555	7.458681	0.129002	-4.922215	-0.058835	1.088436	-0.828456	-0.341275	-0.480408	W

Cette valeur nous semble suspecte dû à des valeurs très nettes (finissant par .0) dans plusieurs champs se retrouvant aussi être les maximums dans ces dernières.
De plus la présence de cette valeur fait varier la moyenne et l'écart-type ne donnant plus un dataset centré réduit comme il était indiqué.

Variables créées

```
datats.describe().loc[['mean', 'std', 'min', 'max']]
```

	intercolumnar distance	upper margin	lower margin	exploitation	row number	modular ratio	interlinear spacing	weight	peak number	modular ratio/ interlinear spacing
mean	-0.000852	0.003396	0.005181	0.002616	-0.006365	-0.008886	0.002350	-0.010259	-0.008691	-0.000678
std	1.008551	0.955257	0.992430	0.991443	1.007876	1.000360	0.966827	0.996431	1.001240	0.992928
min	-3.498799	-2.426761	-3.210528	-5.440122	-4.922215	-7.450257	-11.935457	-4.090167	-4.737863	-6.719324
max	11.819916	19.470188	7.458681	3.987152	1.066121	12.315569	4.901228	4.580832	3.213413	11.911338

```
datatr.describe().loc[['mean', 'std', 'min', 'max']]
```

	intercolumnar distance	upper margin	lower margin	exploitation	row number	modular ratio	interlinear spacing	weight	peak number	modular ratio/ interlinear spacing
mean	0.000852	0.033611	-0.000525	-0.002387	0.006370	0.013973	0.005605	0.010323	0.012914	0.000818
std	0.991431	3.920868	1.120202	1.008527	0.992053	1.126245	1.313754	1.003507	1.087665	1.007094
min	-3.498799	-2.426761	-3.210528	-5.440122	-4.922215	-7.450257	-11.935457	-4.247781	-5.486218	-6.719324
max	11.819916	386.000000	50.000000	3.987152	1.066121	53.000000	83.000000	13.173081	44.000000	4.671232

Moyenne , écart-type , min et max des dataset tr et ts

Variables créées

Le dataset de l'étude avait était près traité afin d'obtenir un dataset centrée-réduite puis séparer en "avila-tr" et "avila-ts".

```
data = pd.concat([datatr,datats])
data = data.drop(labels='modular_ratio/interlinear_spacing', axis=1)
data = data.drop(labels='modular_ratio', axis=1)
data
```

	intercolumnar_distance	upper_margin	lower_margin	exploitation	row_number	interlinear_spacing	weight	peak_number	Class
0	0.266074	-0.165620	0.320980	0.483299	0.172340	0.371178	0.929823	0.251173	A
1	0.130292	0.870736	-3.210528	0.062493	0.261718	1.465940	0.636203	0.282354	A
2	-0.116585	0.069915	0.068476	-0.783147	0.261718	-0.081827	-0.888236	-0.123005	A
3	0.031541	0.297600	-3.210528	-0.583590	-0.721442	0.710932	1.051693	0.594169	A
4	0.229043	0.807926	-0.052442	0.082634	0.261718	0.635431	0.051062	0.032902	F
...	...	...	...	...	...	...	...	...	...
10432	-0.128929	-0.040001	0.057807	0.557894	0.261718	-0.044076	1.158458	2.277968	X
10433	0.266074	0.556689	-0.020434	0.176624	0.261718	0.597681	0.178349	0.625350	G
10434	-0.054866	0.580242	0.032912	-0.016668	0.261718	0.371178	-0.985508	-0.403638	A
10435	0.080916	0.588093	0.015130	0.002250	0.261718	-0.270579	0.163807	-0.091823	F
10436	0.377169	0.014957	0.381439	0.292753	0.261718	-0.006326	-0.494919	-0.247731	H

20866 rows × 9 columns

Afin d'avoir accès à plus de donnée nous avons choisi de regrouper ces deux datasets.

Contexte de l'étude

"There are some entirely new approaches emerged in the last few years [in the domain of digital palaeography], which have been made possible by the combination of powerful computers and high-quality digital images."

Contexte de l'étude

"However promising, all these approaches haven't yet produced widely accepted results, both because of the immaturity in the use of these new technologies, and of the lack of real inter-disciplinary research: palaeographers often missing a proper understanding of rather complex image analysis procedures, and scientists being unaware of the specificity of medieval writing and tending to extrapolate software and methods already developed for modern writings."

Contexte de l'étude

"Such kind of information would allow palaeographers to find further confirmation of their hypothesis and to concentrate their attention on those sections of the manuscript which have not been reliably classified."