Apprentissage Statistique

Sujet 3 :

Prédiction d'oscillations lentes en sommeil profond

Walid Slim — Mariem Azouz

Mohamed Yassin Mejri — Oumaima Merhben

Résumé

Dans cette étude , on a exploré des données consistant en signaux EEG acquis sur le Dreem Headband, dans des conditions fictives , c'est à dire sans aucune sorte de stimulation sonore. On a procédé à des modification de notre dataset (le signal discret) afin d'obtenir une meilleure précision . Deux études : SVM et KPP sont faites sur 3 différents datasets et on a malheureusement obtenu le même résultat (même précision 0.34).

Plan

Introduction

- Contexte de défi
- Objectifs

Étude préalable des données

- Préparation de la base de données
- Étude des corrélations
- Sélection des variables

Méthodes de classification

- K - plus proches voisins
- SVM

Conclusion

Introduction :

1. Contexte de défi :

Une mesure du sommeil se fait dans un laboratoire du sommeil. Divers électrodes sont attachés à la tête pour mesurer les signaux physiologiques.

Le Dreem Headband est un appareil capable de mesurer le sommeil. Son but est d'aider les gens à suivre et à améliorer leur sommeil. Il est capable de mesurer l'activité, la position, la respiration, la fréquence cardiaque et les mouvements du cerveau tout au long de la nuit. L'appareil est également capable d'envoyer des sons par conduction osseuse. Les signaux sont analysés en ligne toute la nuit et l'appareil est capable de stimuler le son pour améliorer la qualité du sommeil profond à différentes étapes de la nuit : s'endormir, dormir profondément (oscillations lentes) et se réveiller.

Dans ce défi, les données sont des EEG signaux acquis sur le Dreem Headband dans des conditions fictives, c'est-à-dire sans aucune sorte de stimulation sonore. Ainsi, nous visons à prédire l'activité cérébrale dans des conditions normales.

2. Objectifs :

Nous essayons de prédire si une oscillation lente sera suivie d'une autre en condition fictive, c'est-à-dire sans aucune stimulation. Cela permettra de :

- Prédire l'activité normale du cerveau
- Savoir quand il est intéressant de stimuler
- Mieux quantifier l'impact d'une stimulation individuelle en comparant à ce qui se serait passé sans stimulation.

Plus précisément, la prédiction est une étiquette entre {0, 1, 2} :

0- Aucune oscillation lente ne commence dans la seconde qui suit.
1- Une oscillation lente de faible amplitude a commencé dans la seconde qui suit.
2- Une oscillation lente de forte amplitude a commencé dans la seconde qui suit.

Étude préalable des données :

1. Préparation de la base de données :

La base de données comporte initialement N=261634 observations de 1261 variables. Chaque échantillon représente 10 secondes d'enregistrement commençant 10 secondes avant la fin d'une oscillation lente. Les données fournies consistent en une matrice N x 1261.

Les variables se décomposent comme suit :

0- Nombre d'oscillations lentes précédentes
1- Amplitude moyenne des oscillations lentes précédentes
2- Durée moyenne des oscillations lentes précédentes
3- Amplitude de l'oscillation lente actuelle
4- Durée de l'oscillation lente actuelle
5- Stade de sommeil actuel
6- Temps écoulé depuis que la personne s'est endormie
7- Temps passé dans le sommeil profond jusqu'à présent
8- Temps passé en sommeil léger jusqu'à présent
9- Temps passé à dormir jusqu'à présent
10- Le temps passé dans le sommeil de veille jusqu'à présent
11- à 1260. Signal EEG pendant 10 secondes (fréquence d'échantillonnage : 125Hz -> 1250 points de données)

On a fait une étude de signal (les 1250 points) par trois méthodes :

a. On considère comme une 11^{ième} variable le minimum de signal (des 1250 points), la 12^{ième} le maximum et la 13^{ième} l'espérance, ainsi on continue nos études avec les 14 variables.

b. Une deuxième méthode consiste à faire une étude statistique au signal et on considère comme des nouvelles variables :

11. Le 5^{ième} centile
12. Le 25^{ième} centile
13. Le 75^{ième} centile
14. Le 95^{ième} centile
15. la médiane
16. L'espérance
17. Standard déviation
18. La variance
19. La norme en L2

Ainsi, on aura 20 variables.

c. Dans une troisième méthode, on a transformé les signaux en des ondelettes via la famille ‘coiflette’(une projection des 1250 variables) et on a fait une étude statistique sur le résultat du projection d’où l’obtention des 20 variables comme la méthode précédente.

On poursuit l’étude en utilisant chaque fois une méthode et par conséquent une nouvelle base de données.

2.Etude des corrélations et sélection des variables:

Nous avons voulu étudier les différentes corrélations entre les variables afin d’avoir une intuition concernant la variable d’intérêt. Nous procèderons de la même façon pour les trois méthodes, on va tracer la matrice de corrélation de HEATMAP pour chaque type de donnée pour visualiser la corrélation entre les variables.

a) Première méthode :

Visualisation des données : On analyse les relations de dépendance entre les variables :

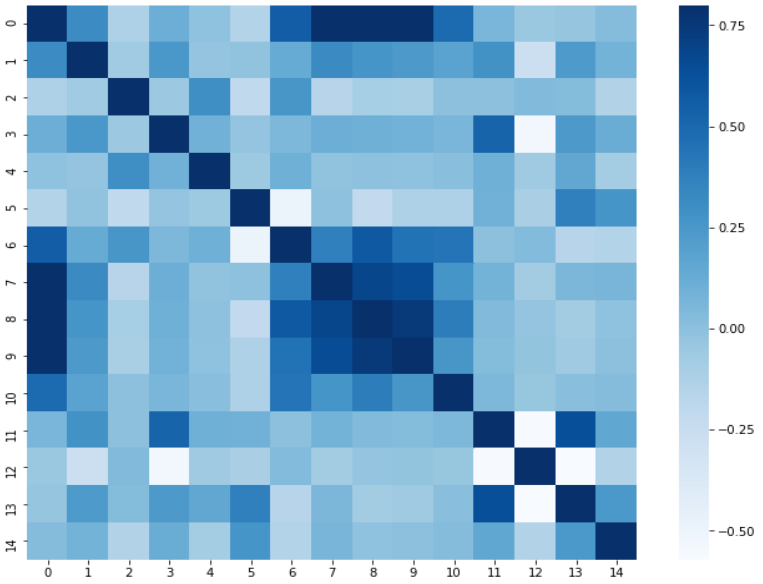

Figure 1a : Corrélation

Console 9/A

	0	1	2	...	18	19	20
0	1.000000	0.315689	-0.128312	...	-0.006362	-0.026534	0.025142
1	0.315689	1.000000	-0.072843	...	0.240043	0.234263	0.086659
2	-0.128312	-0.072843	1.000000	...	0.045484	0.034138	-0.139393
3	0.114099	0.249328	-0.047764	...	0.305166	0.239788	0.122924
4	-0.006859	-0.024140	0.298607	...	0.138234	0.153271	-0.080273
5	-0.146940	-0.011261	-0.206821	...	0.252212	0.380929	0.269614
6	0.560468	0.131540	0.261126	...	-0.087933	-0.158839	-0.143595
7	0.829811	0.320543	-0.162240	...	0.054560	0.058061	0.066746
8	0.918959	0.264989	-0.093318	...	-0.040987	-0.074256	-0.003787
9	0.820054	0.241789	-0.101938	...	-0.032219	-0.058934	0.005072
10	0.492575	0.187125	0.002726	...	0.016208	0.013757	0.023550
11	0.029305	-0.216963	-0.015665	...	-0.888476	-0.938434	-0.227932
12	0.050215	-0.181473	-0.041567	...	-0.804203	-0.925474	-0.219601
13	-0.050108	0.180621	0.025722	...	0.804530	0.932903	0.242260
14	0.003397	0.242359	0.053298	...	0.898788	0.915171	0.217087
15	-0.020054	0.013852	-0.038906	...	0.087976	0.136440	0.071921
16	0.033439	0.044022	0.022690	...	0.093809	0.086886	0.060020
17	0.000833	0.269169	0.031278	...	0.952414	0.977708	0.236309
18	-0.006362	0.240043	0.045484	...	1.000000	0.923910	0.179398
19	-0.026534	0.234263	0.034138	...	0.923910	1.000000	0.244777
20	0.025142	0.086659	-0.139393	...	0.179398	0.244777	1.000000

Figure 2a : Les valeurs numériques des corrélations

On va définir un critère expérimental de sélection des variables, on élimine les variables qui ont une corrélation inférieur à 0.1 . En respectant ce critère on élimine les variables 0,1,4,7,8,9,10 d’où la nouvelle matrice de corrélation :

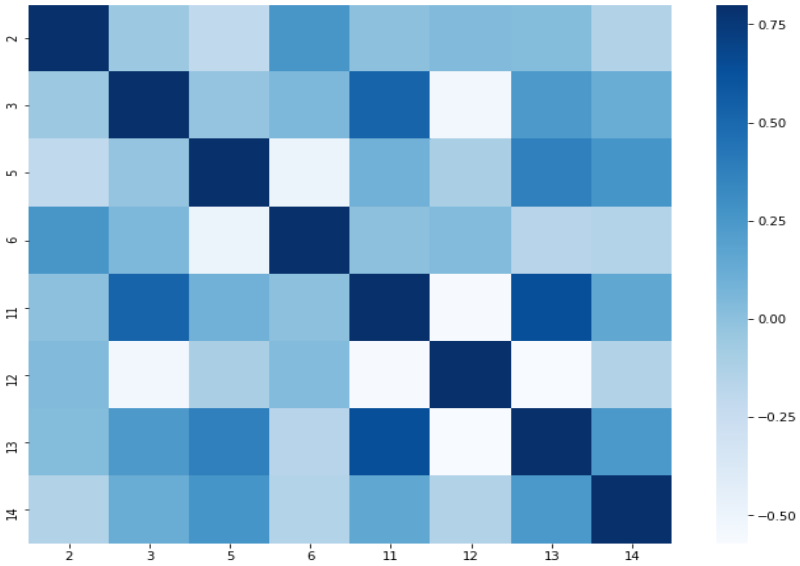

Figure 3a : nouvelle matrice de corrélation

Visualisons la représentation des variables en fonction de Y :

***Figure 4a :* Représentation des variables en fonction de Y**

b) Deuxième méthode :

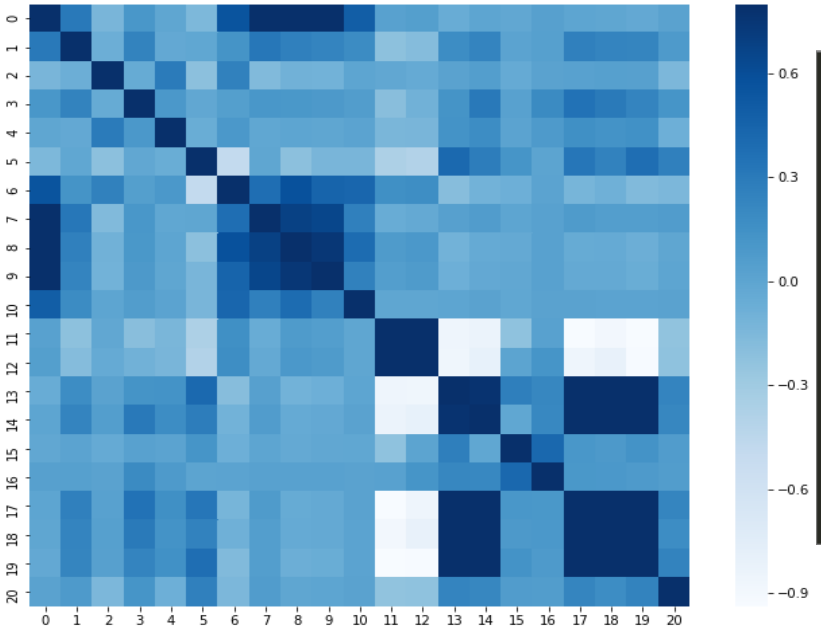

Figure 1b : Corrélation

	0	1	2	...	18	19	20
0	1.000000	0.315689	-0.128312	...	-0.006362	-0.026534	0.025142
1	0.315689	1.000000	-0.072843	...	0.240043	0.234263	0.086659
2	-0.128312	-0.072843	1.000000	...	0.045484	0.034138	-0.139393
3	0.114099	0.249328	-0.047764	...	0.305166	0.239788	0.122924
4	-0.006859	-0.024140	0.298607	...	0.138234	0.153271	-0.080273
5	-0.146940	-0.011261	-0.206821	...	0.252212	0.380929	0.269614
6	0.560468	0.131540	0.261126	...	-0.087933	-0.158839	-0.143595
7	0.829811	0.320543	-0.162240	...	0.054560	0.058061	0.066746
8	0.918959	0.264989	-0.093318	...	-0.040987	-0.074256	-0.003787
9	0.820054	0.241789	-0.101938	...	-0.032219	-0.058934	0.005072
10	0.492575	0.187125	0.002726	...	0.016208	0.013757	0.023550
11	0.029305	-0.216963	-0.015665	...	-0.888476	-0.938434	-0.227932
12	0.050215	-0.181473	-0.041567	...	-0.804203	-0.925474	-0.219601
13	-0.050108	0.180621	0.025722	...	0.804530	0.932903	0.242260
14	0.003397	0.242359	0.053298	...	0.898788	0.915171	0.217087
15	-0.020054	0.013852	-0.038906	...	0.087976	0.136440	0.071921
16	0.033439	0.044022	0.022690	...	0.093809	0.086886	0.060020
17	0.000833	0.269169	0.031278	...	0.952414	0.977708	0.236309
18	-0.006362	0.240043	0.045484	...	1.000000	0.923910	0.179398
19	-0.026534	0.234263	0.034138	...	0.923910	1.000000	0.244777
20	0.025142	0.086659	-0.139393	...	0.179398	0.244777	1.000000

Figure 2b : Les valeurs numériques des corrélations

De même, en respectant le critère choisi on élimine les variables 0,1,4,7,8,9,10,15,16 ; donc la nouvelle matrice de corrélation est :

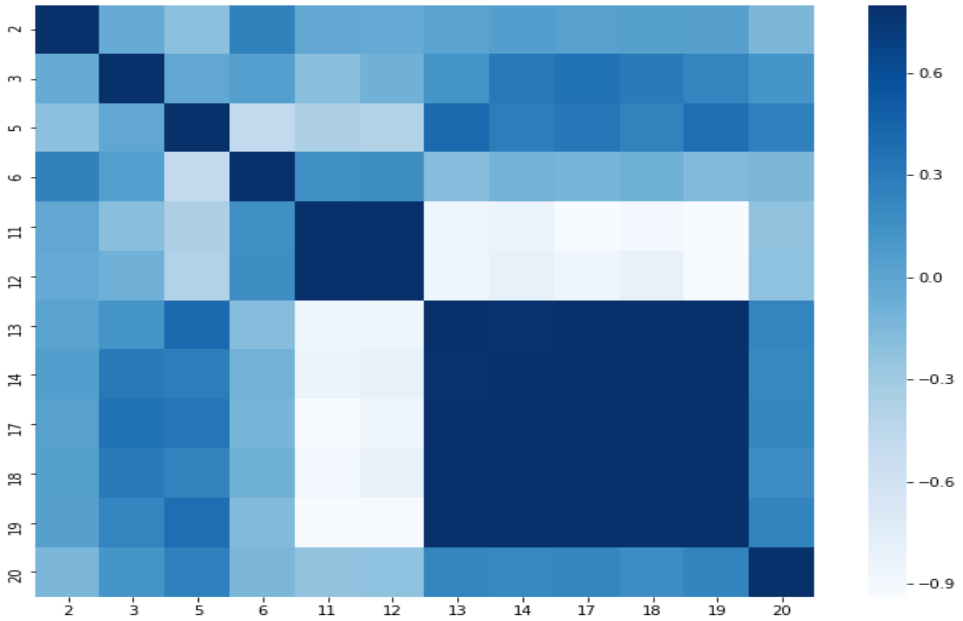

Figure 3b : nouvelle matrice de corrélation

Visualisons la représentation des variables en fonction de Y :

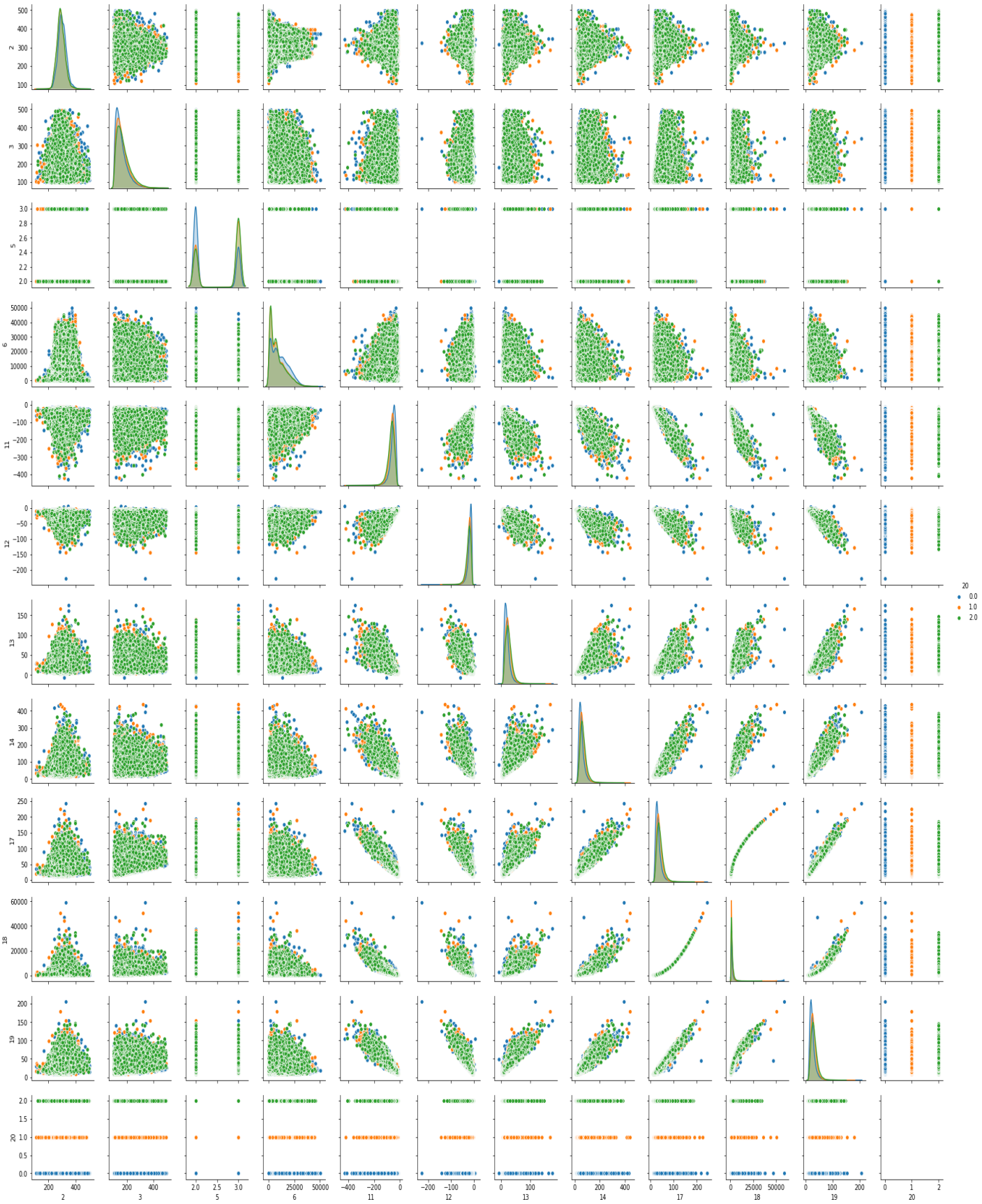

Figure 4b : **Représentation des variables en fonction de Y**

c) Troisième méthode :

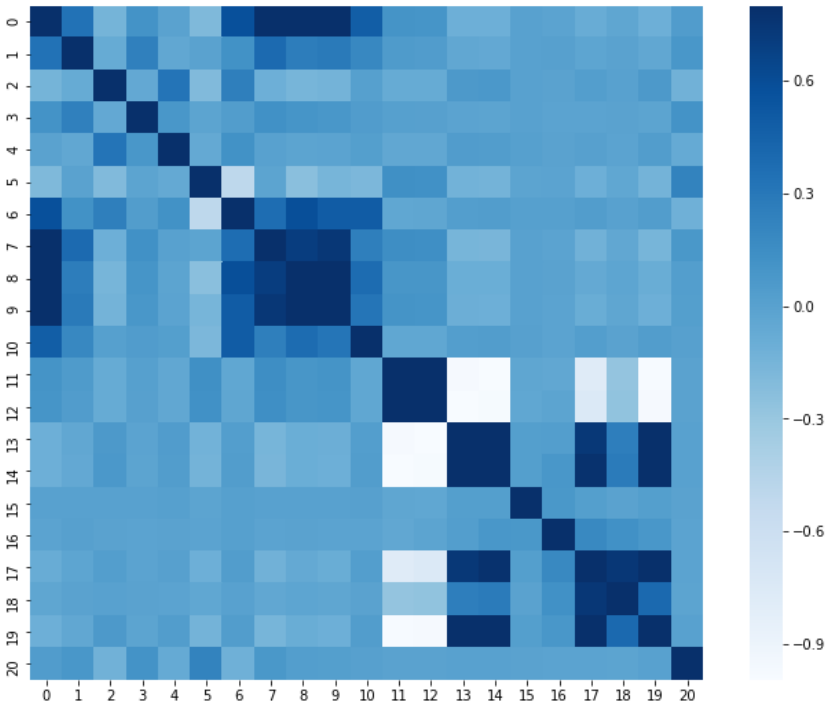

Figure 2c : Corrélation

	0	1	2	...	18	19	20
0	1.000000	0.338001	-0.140457	...	-0.030536	-0.105707	0.047817
1	0.338001	1.000000	-0.075369	...	0.000131	-0.043382	0.091750
2	-0.140457	-0.075369	1.000000	...	0.003541	0.064434	-0.121741
3	0.114010	0.249871	-0.048759	...	-0.005223	-0.013832	0.115223
4	-0.005271	-0.041603	0.328274	...	-0.011547	0.038206	-0.068066
5	-0.186283	-0.002563	-0.188462	...	-0.035127	-0.132612	0.230138
6	0.580915	0.121642	0.258856	...	0.005033	0.038435	-0.116080
7	0.822301	0.397095	-0.109822	...	-0.043243	-0.150527	0.079755
8	0.939330	0.272594	-0.151002	...	-0.022683	-0.089683	0.029063
9	0.883902	0.293938	-0.138653	...	-0.031030	-0.104565	0.027070
10	0.477914	0.194868	0.018492	...	-0.005715	0.038263	0.013509
11	0.109714	0.051245	-0.071218	...	-0.277058	-0.981824	-0.003155
12	0.105257	0.046665	-0.070730	...	-0.262026	-0.972493	-0.003631
13	-0.105011	-0.045041	0.068908	...	0.264545	0.971936	0.004235
14	-0.110012	-0.051285	0.071865	...	0.284187	0.983215	0.003946
15	0.002381	0.003524	0.006855	...	0.000554	0.027385	0.000511
16	-0.006332	0.010117	-0.003978	...	0.127087	0.087695	-0.010771
17	-0.079773	-0.020912	0.034044	...	0.746364	0.862563	-0.006952
18	-0.030536	0.000131	0.003541	...	1.000000	0.406524	-0.013279
19	-0.105707	-0.043382	0.064434	...	0.406524	1.000000	0.001137
20	0.047817	0.091750	-0.121741	...	-0.013279	0.001137	1.000000

Figure 2c : Les valeurs numériques des corrélations

De même, en respectant le critère choisi on élimine les variables 0,1,4,7,8,9,10,11,12,13,14,15,16,17,18,19; donc la nouvelle matrice de corrélation est :

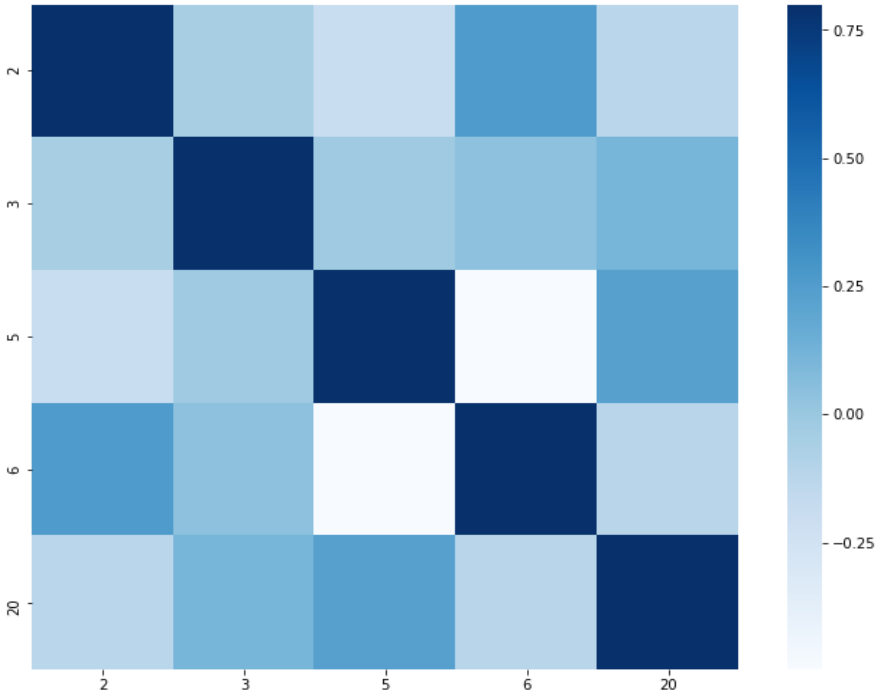

Figure 3c : nouvelle matrice de corrélation

Visualisons la représentation des variables en fonction de Y :

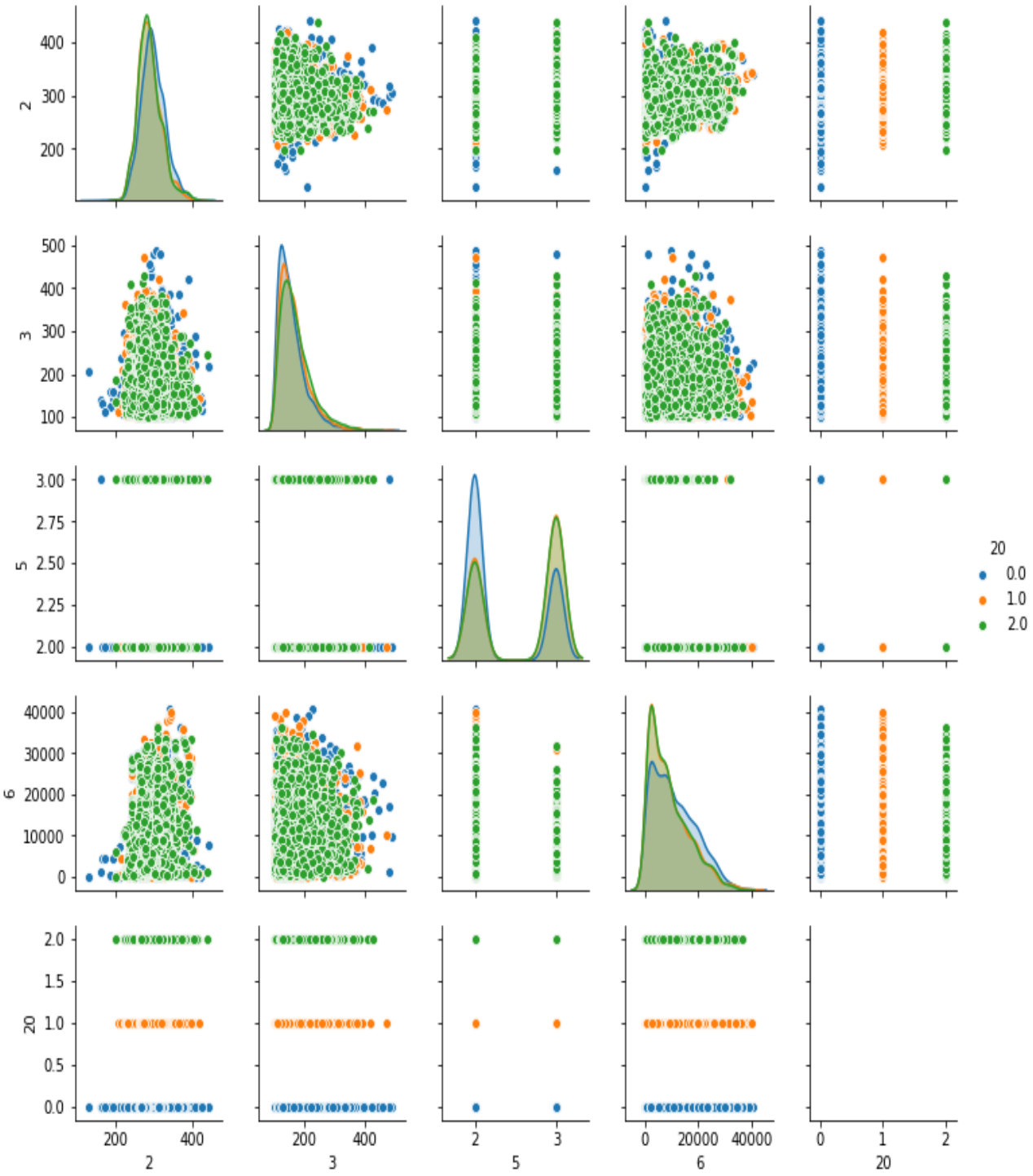

Figure 4c : **Représentation des variables en fonction de Y**

Méthode de classification :

1. K pp :

La première méthode testée est celle des k-Plus Proches Voisins vue en cours. Le paramètre k va devoir être déterminé en évaluant sur l'échantillon de test le modèle obtenu avec l'échantillon d'apprentissage pour différentes valeurs de k qui doivent être comprises entre 1 et le nombre d'individus dans l'échantillon le plus petit, ici et en général l'échantillon de test. Nous entraînons le modèle pour chaque valeur de k sur l'échantillon d'apprentissage, et mesurons son erreur sur l'échantillon de test. La figure 5 montre l'erreur obtenue pour certaines valeurs de k. Nous observons que l'erreur est minimal pour k=400 .

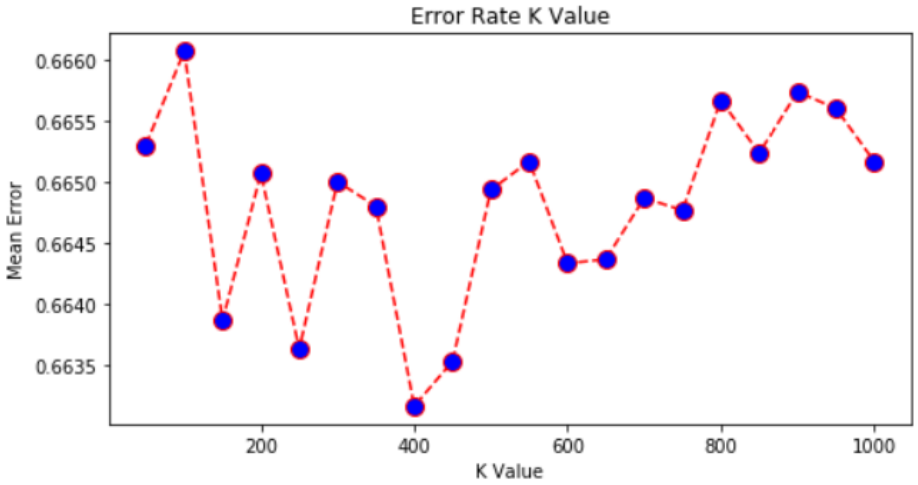

Figure 5 : Erreur des k-PPV pour différentes valeurs de k

Cependant, en analysant les prédictions, on obtient les résultats suivants avec les trois méthodes :

	precision	recall	f1-score	support
0	0.33	0.66	0.44	345
1	0.27	0.08	0.12	339
2	0.31	0.20	0.25	316
avg / total	0.30	0.32	0.27	1000

Résultat avec la 1ière méthode

	precision	recall	f1-score	support
0	0.34	1.00	0.50	10103
1	0.00	0.00	0.00	9993
2	0.00	0.00	0.00	9904
avg / total	0.11	0.34	0.17	30000

Résultat avec la 2ième méthode

	precision	recall	f1-score	support
0	0.34	0.71	0.46	10103
1	0.33	0.20	0.25	9993
2	0.33	0.08	0.13	9904
avg / total	0.33	0.33	0.28	30000

Résultat avec la 3ième méthode

2.SVM :

Nous avons tenté d'obtenir des résultats plus probants avec la classe des méthodes considérées comme une des meilleures méthodes de classification et qui se comporte bien en grande dimension.

De même , en appliquant l'SVM sur les 3 méthodes , les différents base de données n'ont pas porté leurs fruits vus qu'ils n'ont pas permis de séparer l'échantillon d'une manière précise. La précision de l'SVM reste 0.34.

	precision	recall	f1-score	support
0	0.33	0.66	0.44	345
1	0.27	0.08	0.12	339
2	0.31	0.20	0.25	316
avg / total	0.30	0.32	0.27	1000

Résultat avec la 1ière méthode

	precision	recall	f1-score	support
0	0.34	1.00	0.50	10103
1	0.00	0.00	0.00	9993
2	0.00	0.00	0.00	9904
avg / total	0.11	0.34	0.17	30000

Résultat avec la 2ième méthode

	precision	recall	f1-score	support
0	0.34	1.00	0.50	10103
1	0.00	0.00	0.00	9993
2	0.00	0.00	0.00	9904
avg / total	0.11	0.34	0.17	30000

Résultat avec la 3ième méthode

Conclusion :

Pour conclure , la faible accuracy(0.34) pour les deux méthodes (svm et kpp) peut être expliquer par la difficulté de dégager des variables pertinentes. En outre, vu la grandes dimension de dataset, les algorithmes étaient couteux d'où le besoin de travailler sur des échantillons réduits.