Projet 8 : Participez à une compétition Kaggle !

Claire Gayral

Décembre 2021 - Janvier 2022

Choix de compétition

Forest Cover Type Prediction

- prédiction sur 7 classes de couverture forestière
- cellule = 30 x 30 mètres :
  - 15120 observations annotées,
  - 565892 à prédire
- à partir de 12 variables cartographiques
  - dont 2 variables catégorielles
  - et 10 variables numériques

Introduction - Plan

1. Analyse exploratoire
   - Variables catégorielles
   - Variables numériques
2. Classification
   - Choix du modèle de classification
   - Résultats du meilleur modèle
3. Conclusion

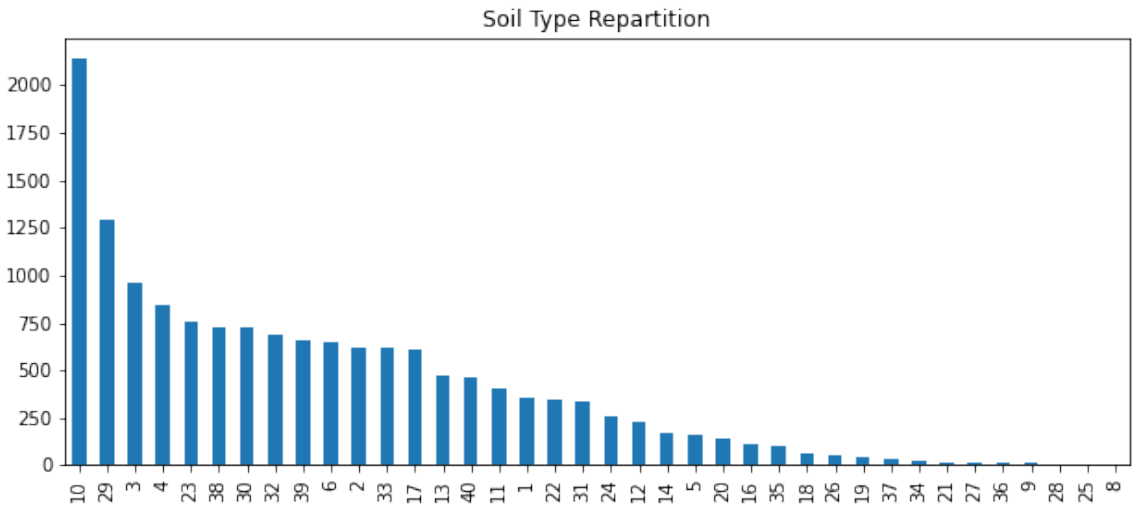
Catégorielle - univariée

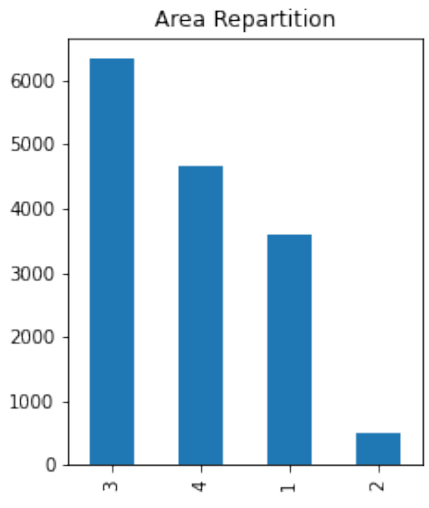
Le type de sol :

- 40 classes
- 2 non représentées

La zone naturelle

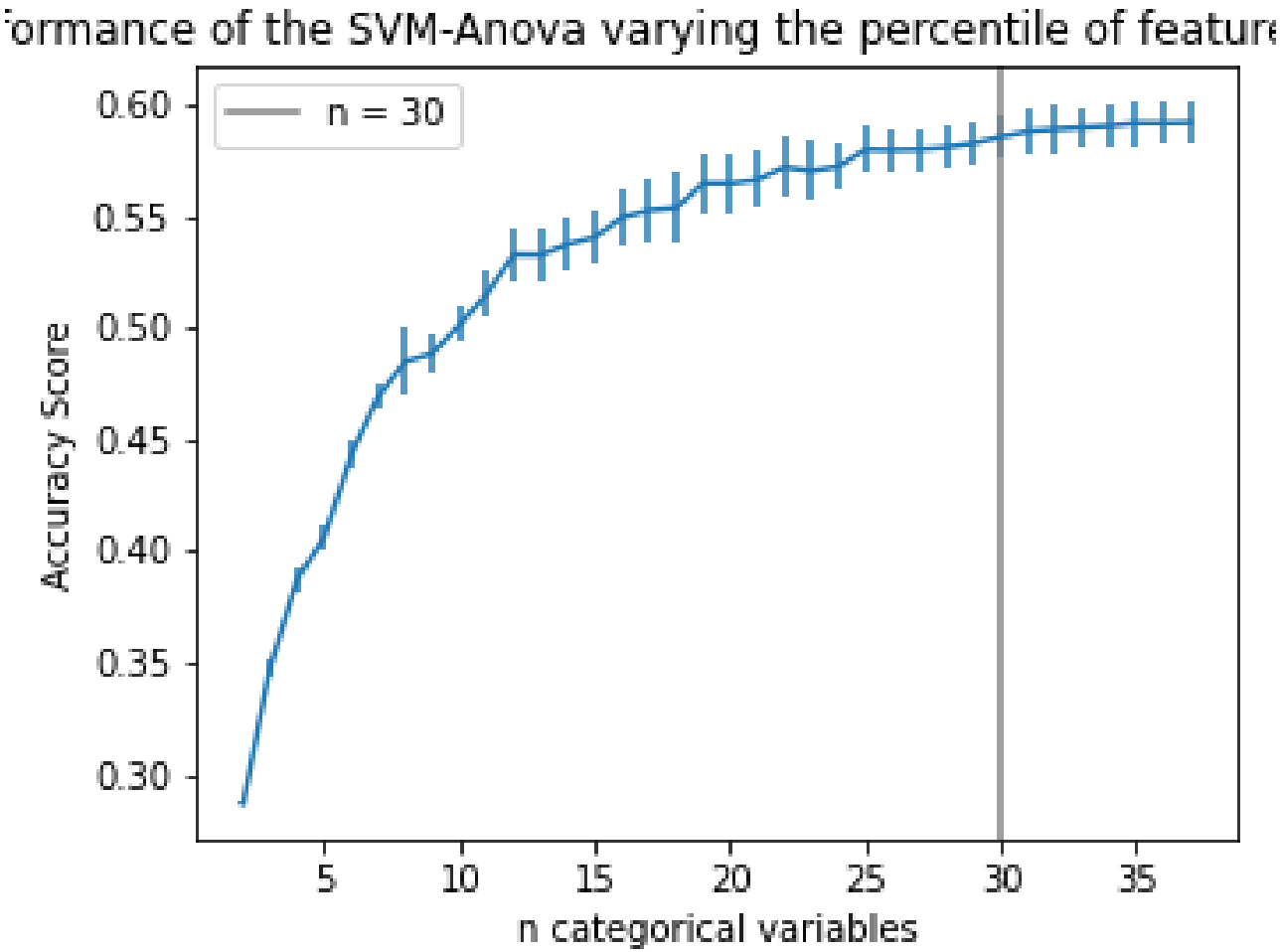
- 4 classes





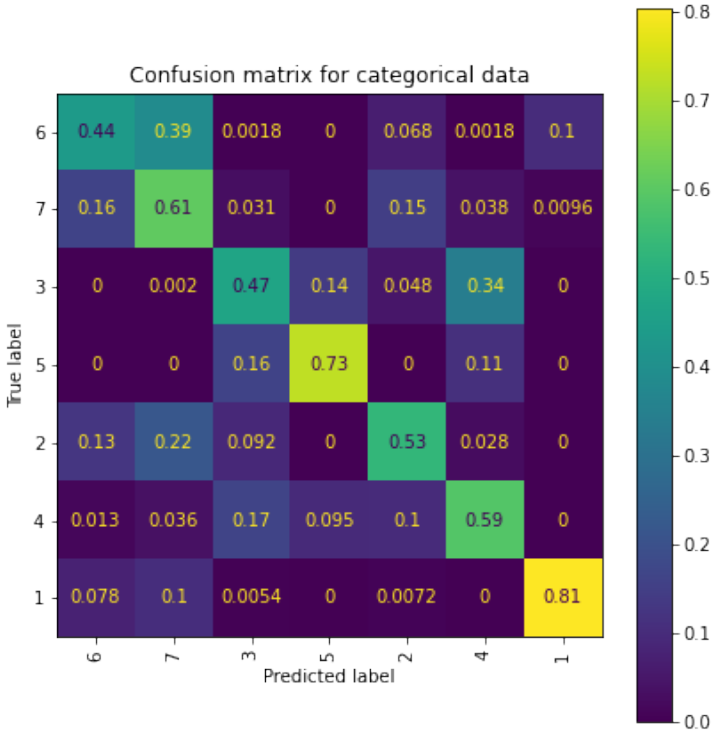
Catégorielle - sélection de variable

ANOVA (chi2), Std puis SVC

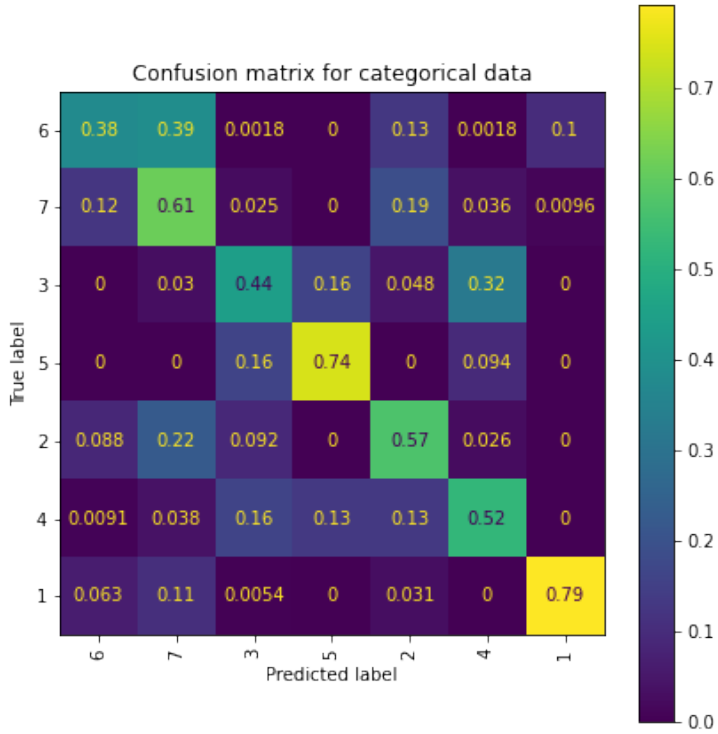


Catégorielle - sélection de variable

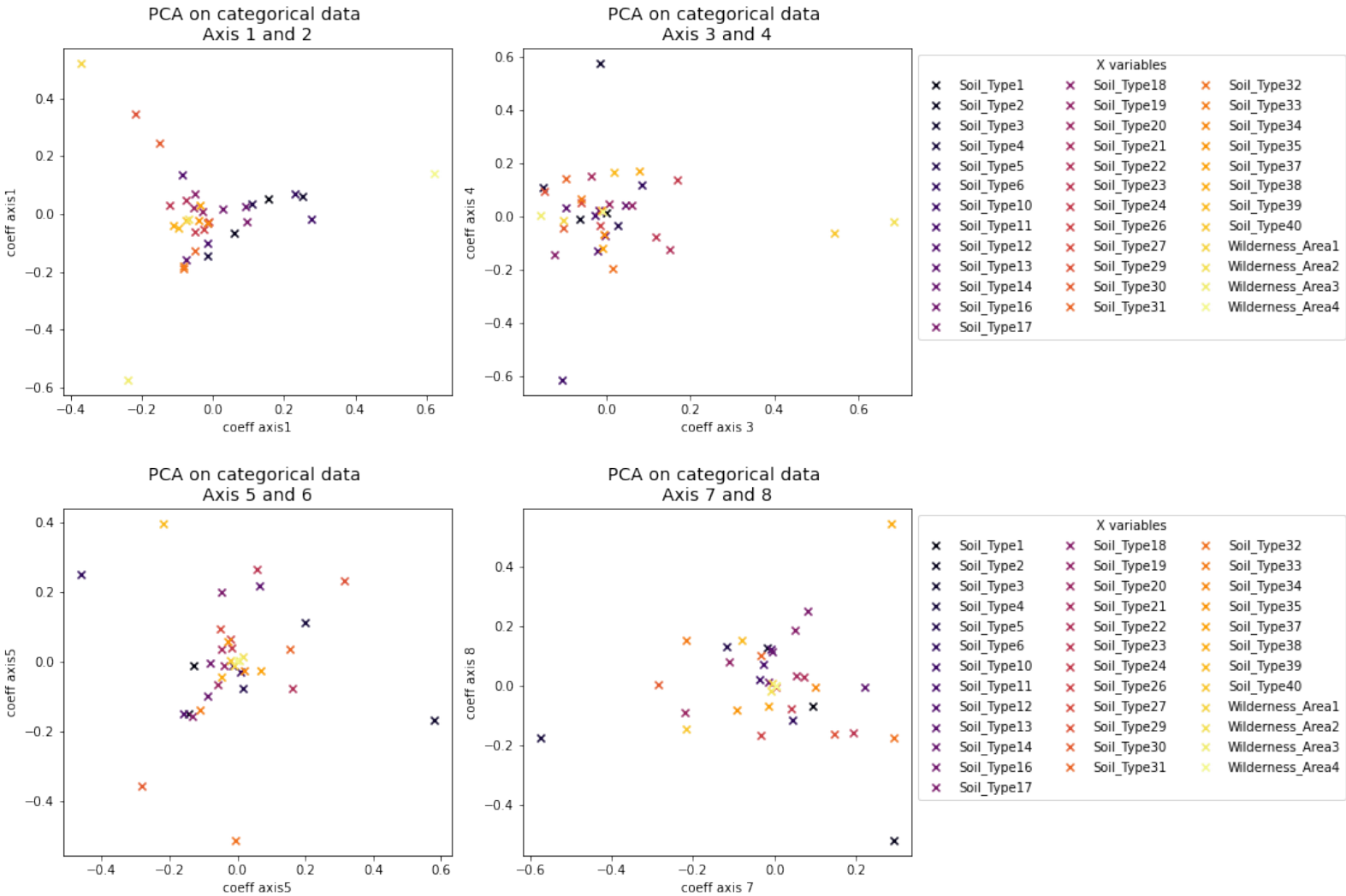
Classification avec les variables catégorielles



Classification avec les 30 catégories les plus importantes

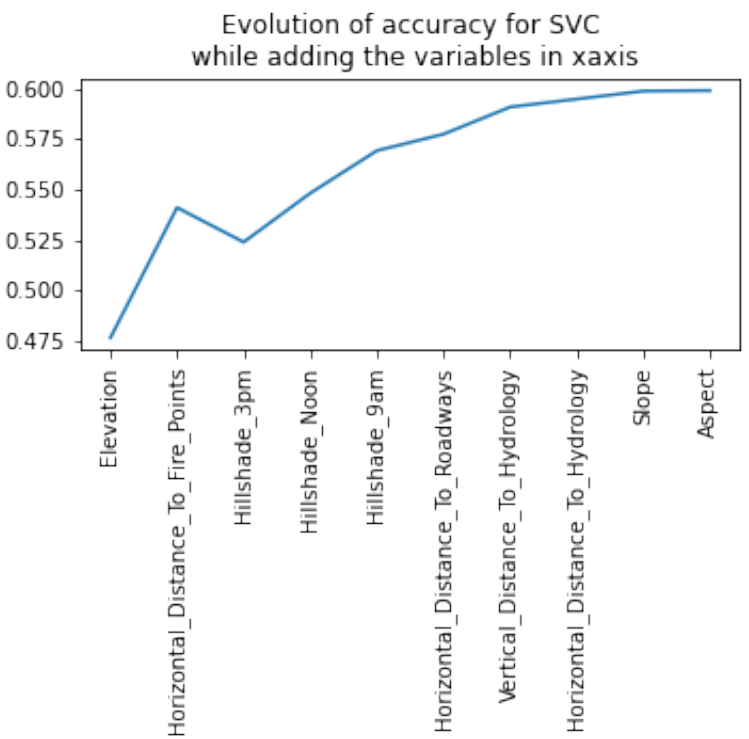


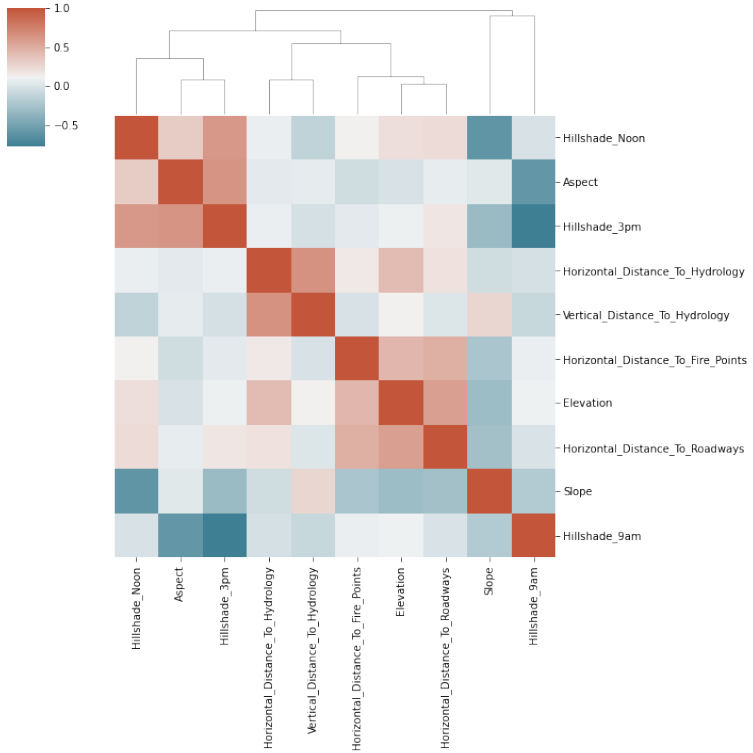
Catégorielle - Projection en dimension 20



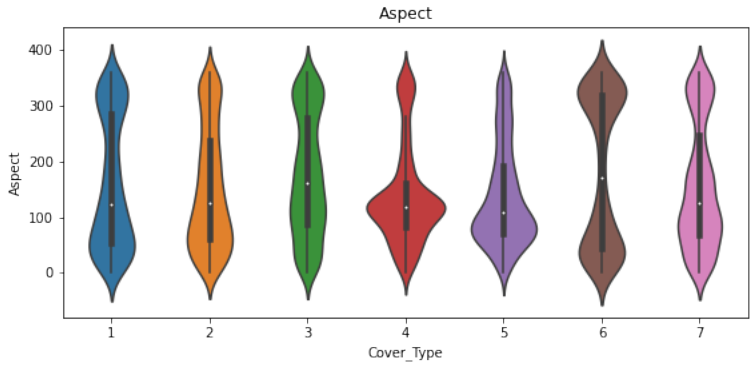
Les 10 variables numériques

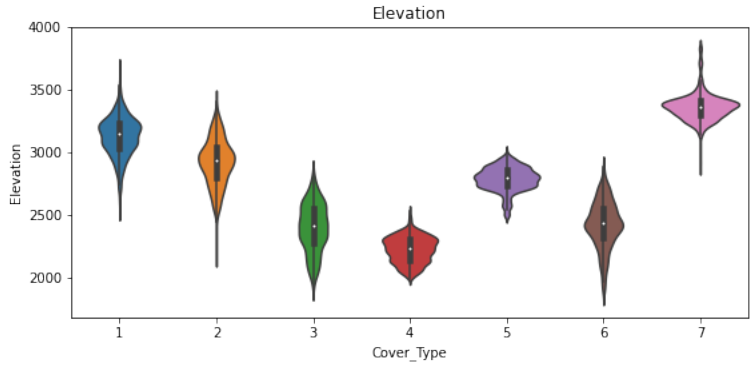
- 3 variables liées à l'ombrage (Noon, 9am, 3pm)
- 5 variables de position, dont deux liées à l'hydrologie
- 2 variables non significatives





Numériques - lien avec le type de forêt



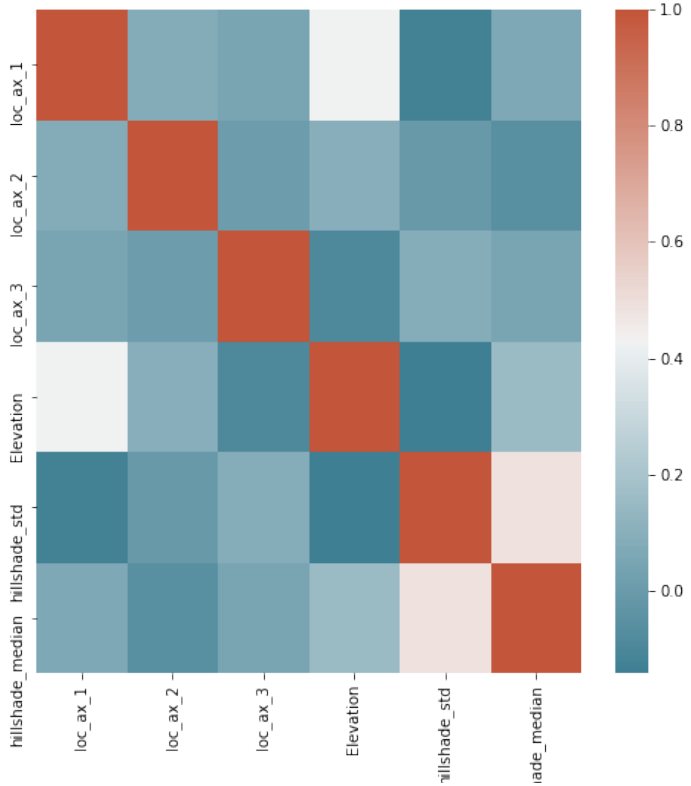



```
Hillshade_Noon                      0.044745
Aspect                              0.018814
Hillshade_3pm                       0.072812
Horizontal_Distance_To_Hydrology    0.129856
Vertical_Distance_To_Hydrology      0.028816
Horizontal_Distance_To_Fire_Points  0.228525
Elevation                           0.865734
Horizontal_Distance_To_Roadways     0.326858
Slope                               0.107013
Hillshade_9am                       0.130554
Name: eta, dtype: float64
```

Numériques - Création de variable

- Hillshade :
  - Médiane
  - Écart-type
- Hydrologie :
  - Distance totale
  - Variation

  $\hookrightarrow$ horizontale
- Distances horizontale orthonormalisée



Classification - Une première idée

Les scores :

	Nearest Neighbors	Linear SVM	RBF SVM	Decision Tree	Random Forest	Neural Net	AdaBoost	Naive Bayes
0	0.799735	0.724074	0.741005	0.733069	0.674603	0.774868	0.749471	0.480688
1	0.807143	0.710053	0.751323	0.730952	0.690741	0.753968	0.737831	0.464021
2	0.552116	0.608201	0.648677	0.571958	0.555026	0.646296	0.503704	0.454233
3	0.557672	0.633598	0.665079	0.663228	0.612169	0.673810	0.539418	0.593386
4	0.826190	0.722751	0.715608	0.716402	0.698942	0.713757	0.763757	0.593386
5	0.818254	0.710317	0.713757	0.725132	0.718519	0.692328	0.768783	0.630423
6	0.605820	0.618783	0.628571	0.628042	0.657407	0.630159	0.625926	0.638360
7	0.655026	0.643122	0.631746	0.596032	0.639418	0.618783	0.599471	0.631481

Les temps :

	Nearest Neighbors	Linear SVM	RBF SVM	Decision Tree	Random Forest	Neural Net	AdaBoost	Naive Bayes
0	14.784027	0.191472	8.091215	0.138105	0.230682	4.571458	0.745726	0.145967
1	8.262091	0.149591	4.523373	0.129786	0.224175	2.888409	0.650367	0.132854
2	7.925421	0.142454	4.089891	0.128564	0.214091	2.539925	0.608353	0.130020
3	8.094473	0.199251	4.549675	0.133181	0.208550	2.879220	0.656897	0.133263
4	15.136565	0.314168	9.311963	0.215078	0.272194	3.598449	2.085076	0.146518
5	8.427706	0.182357	5.858842	0.159040	0.247225	2.492870	1.348735	0.131550
6	8.065719	0.169573	4.893156	0.157488	0.248050	2.503127	0.920149	0.131746
7	8.255832	0.158323	5.160444	0.157020	0.221932	3.020187	0.988942	0.131386

Classification - Optimisation des modèles

Les scores :

	Nearest Neighbors	SVM_rbf	SVM_linear	Decision Tree	Neural Net	AdaBoost
0	0.799735	0.841799	0.724339	0.789418	0.779365	0.762169
1	0.807143	0.842857	0.711905	0.782011	0.772222	0.749735
2	0.552116	0.432011	0.607143	0.540741	0.578042	0.503968
3	0.557672	0.438624	0.636772	0.589683	0.637037	0.551852
4	0.826190	0.846032	0.722751	0.785714	0.765344	0.790476
5	0.818254	0.843651	0.709259	0.809259	0.743651	0.796296
6	0.633069	0.662434	0.618519	0.573545	0.629101	0.627513
7	0.655820	0.689418	0.643386	0.587037	0.683598	0.579365

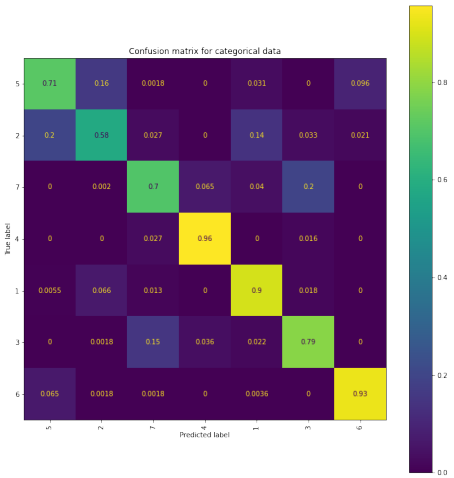
Les temps (pour le choix des paramètres) :

	Nearest Neighbors	SVM_rbf	SVM_linear	Decision Tree	Neural Net	AdaBoost
0	0.799735	0.841799	0.724339	0.789418	0.779365	0.762169
1	0.807143	0.842857	0.711905	0.782011	0.772222	0.749735
2	0.552116	0.432011	0.607143	0.540741	0.578042	0.503968
3	0.557672	0.438624	0.636772	0.589683	0.637037	0.551852
4	0.826190	0.846032	0.722751	0.785714	0.765344	0.790476
5	0.818254	0.843651	0.709259	0.809259	0.743651	0.796296
6	0.633069	0.662434	0.618519	0.573545	0.629101	0.627513
7	0.655820	0.689418	0.643386	0.587037	0.683598	0.579365

Meilleurs modèles - Métriques sur test

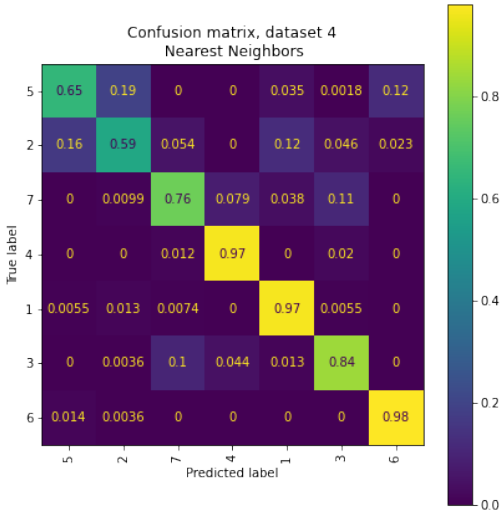
	model_name	dataset	AUC	balanced_accuracy	kappa_score	matthews_corrcoef	hinge_loss	params
0	Nearest Neighbors	4	0.897077	0.823105	0.797108	0.798199	0.347619	3.0
1	SVM_rbf	4	0.908989	0.843602	0.820310	0.820458	0.307937	10000.0
2	SVM_rbf	5	0.907559	0.841151	0.817516	0.817785	0.312698	10000.0
3	SVM_rbf	0	0.906730	0.839780	0.815395	0.815489	0.316402	1000.0
4	SVM_rbf	1	0.907340	0.840820	0.816631	0.816706	0.314286	1000.0

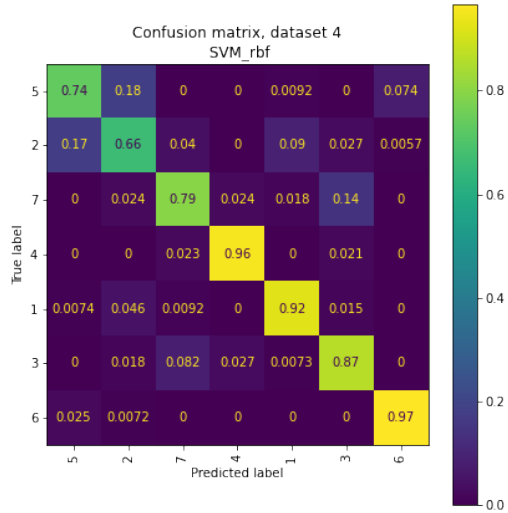
Light GBM :

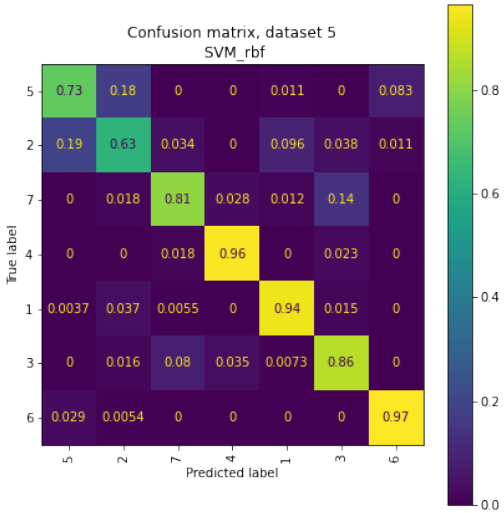


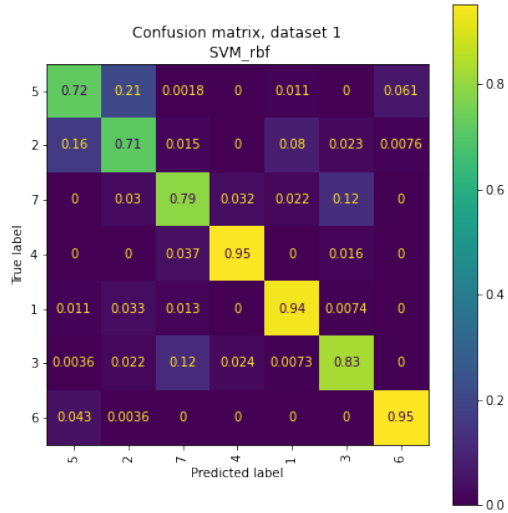
AUC	0.880141
balanced_accuracy	0.794028
kappa_score	0.763458
matthews_corrcoef	0.763962
hinge_loss	0.405291

4 Meilleur modèle - tables de confusion









Conclusion

Résumé

- Petite dimension, des variables peu significatives
- Deux façons de modéliser le problème : avec ou sans pré-traitements
- Classification par SVR avec un noyau gaussien

Améliorations et suite :

- Comparer avec et sans pré-traitements
- Améliorer le modèle en combinant les analyses, pour mieux caractériser les deux classes mal séparées

Merci pour votre écoute !