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

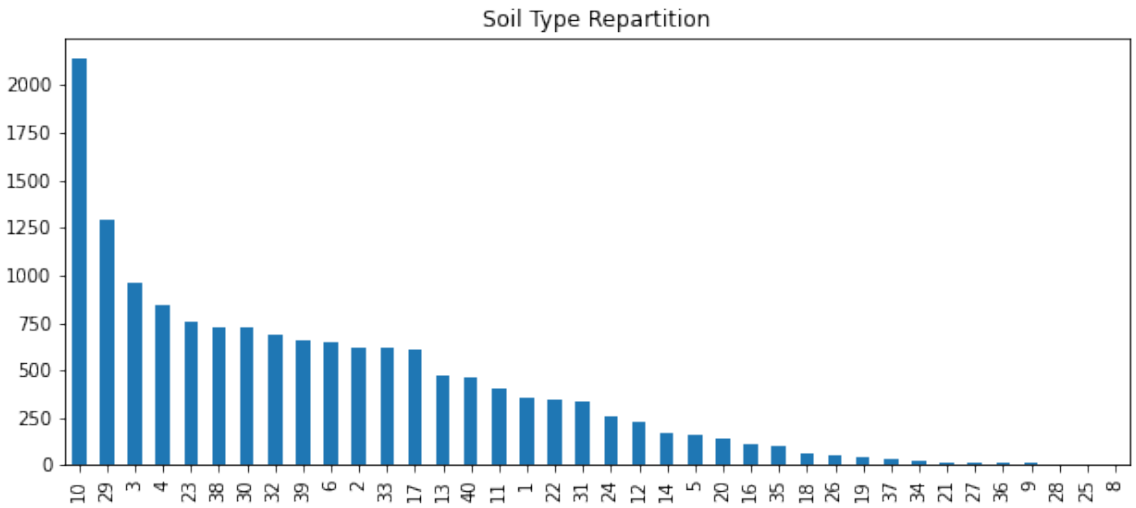
Catégorielle - univariée

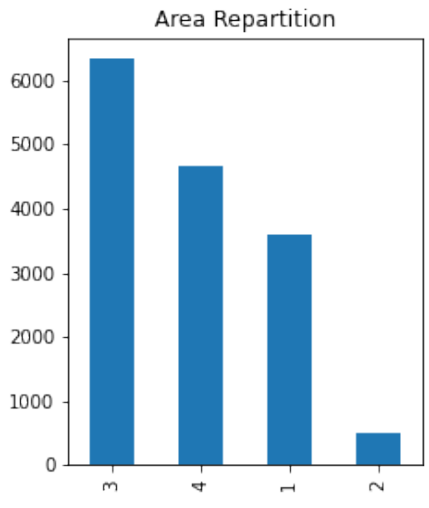
Le type de sol :

- 40 classes
- 2 non représentées

La zone naturelle

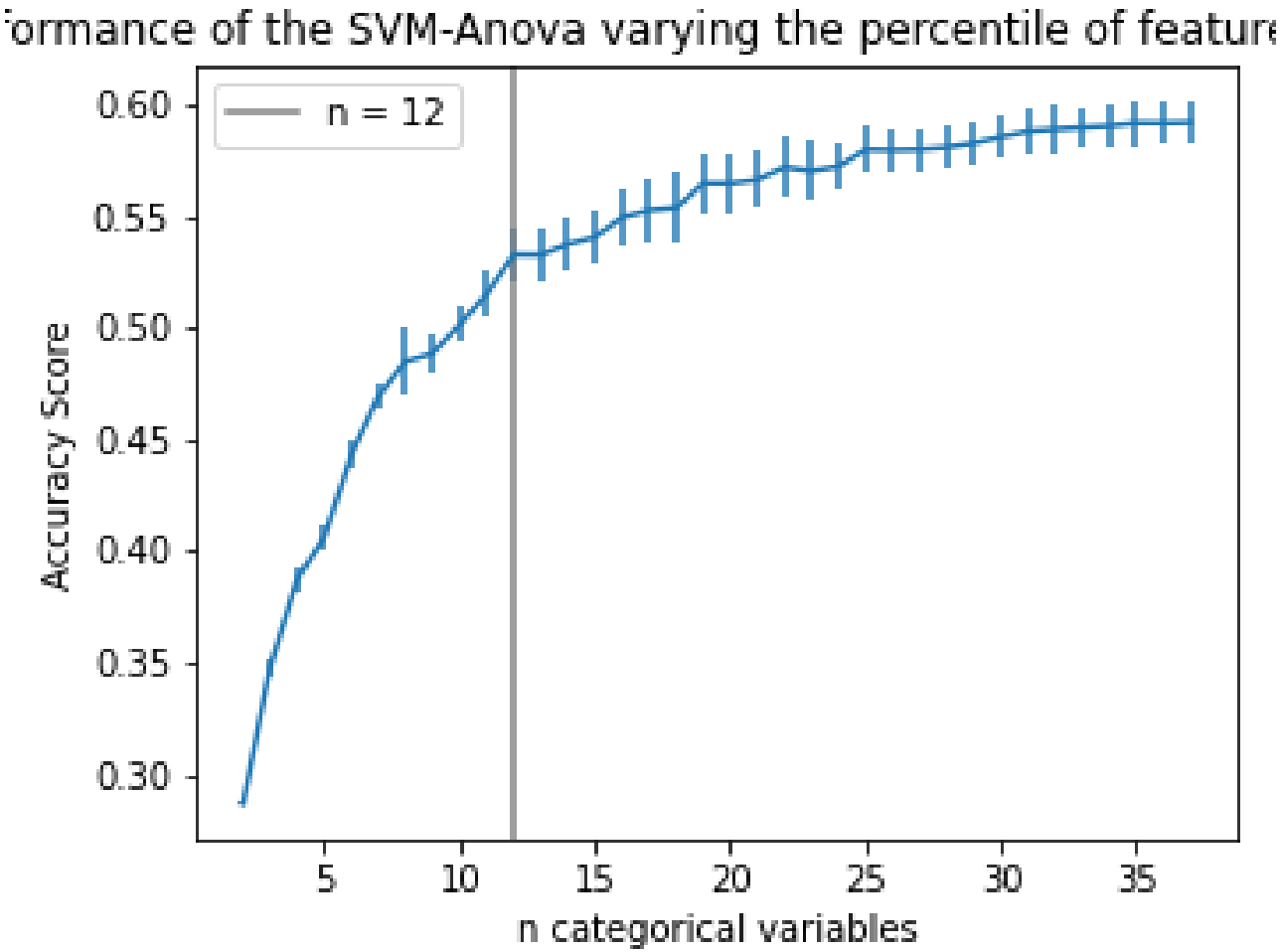
- 4 classes





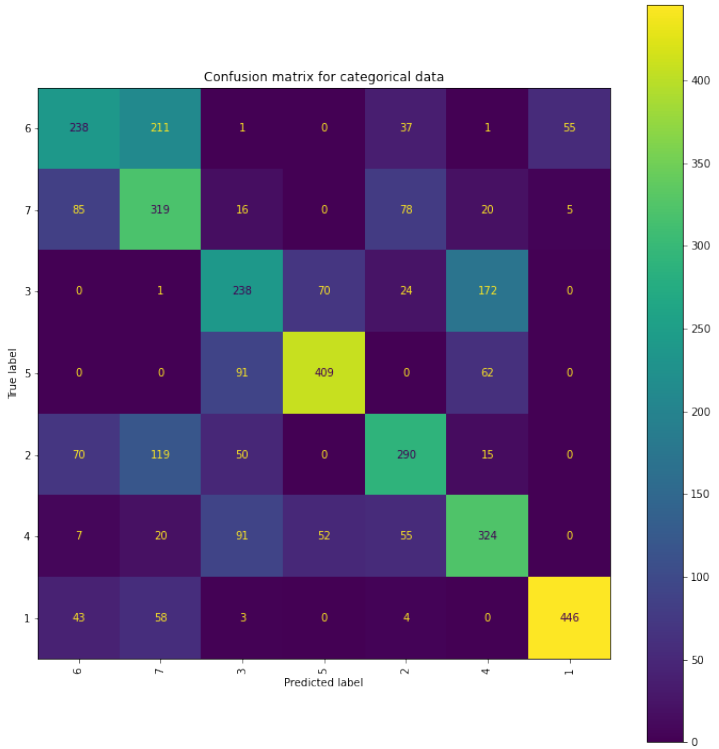
Catégorielle - sélection de variable

ANOVA (chi2), Std puis SVC

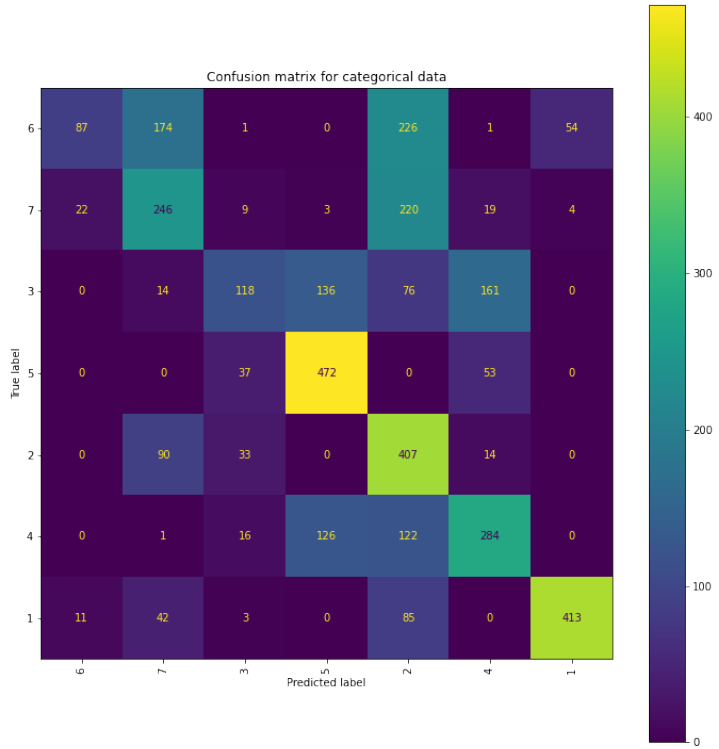


Catégorielle - sélection de variable

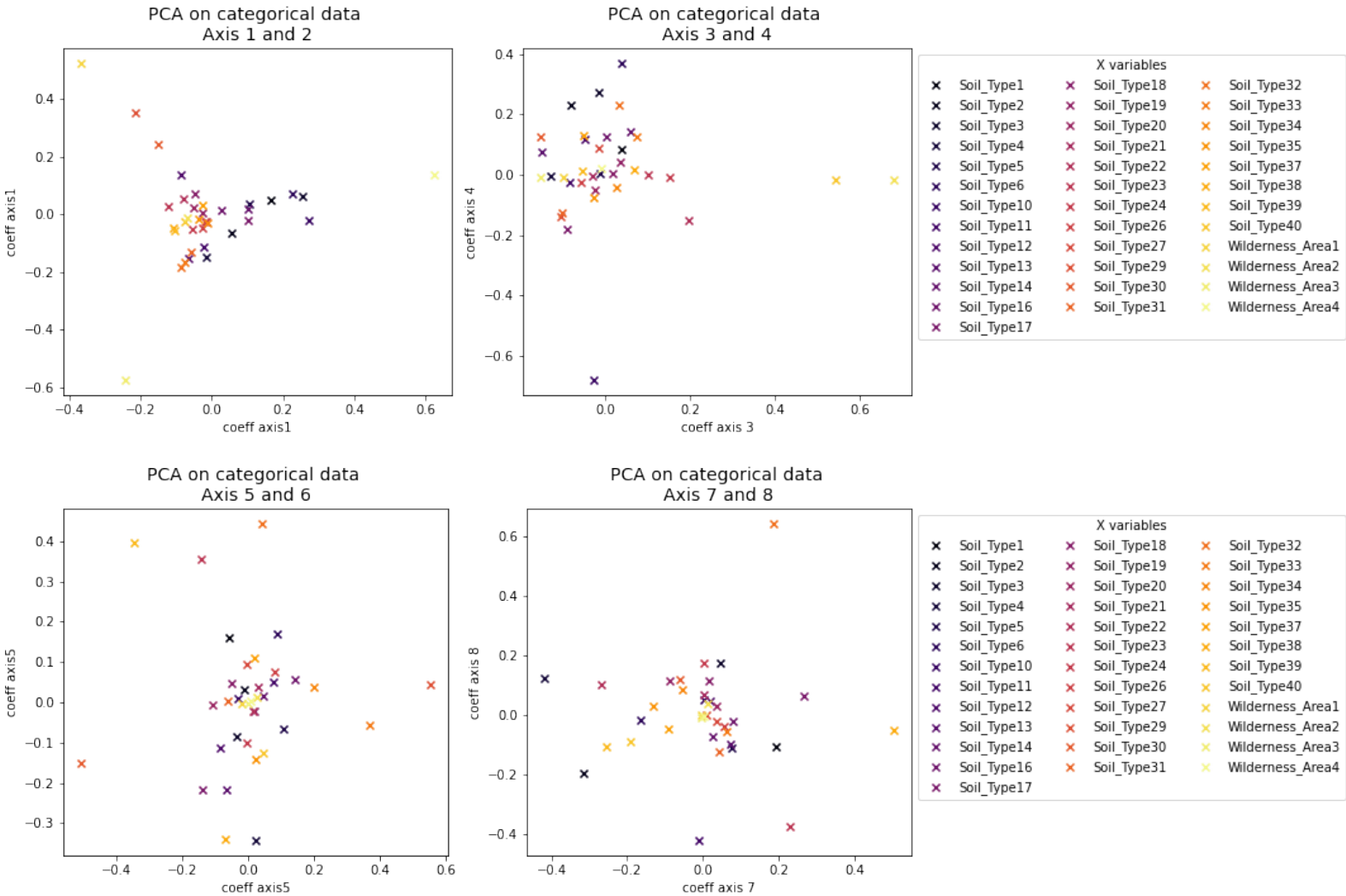
Classification avec les variables catégorielles



Classification avec les 30 catégories les plus importantes

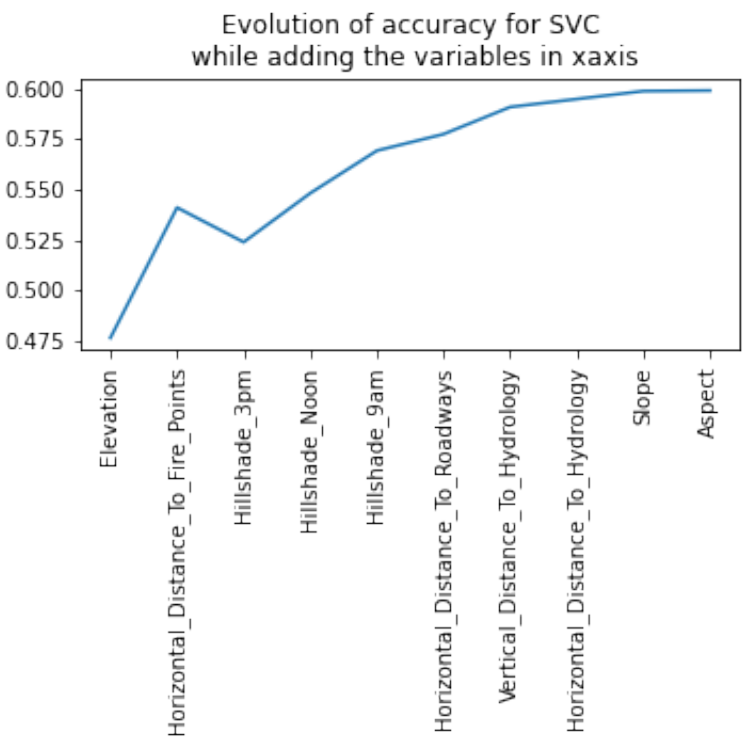


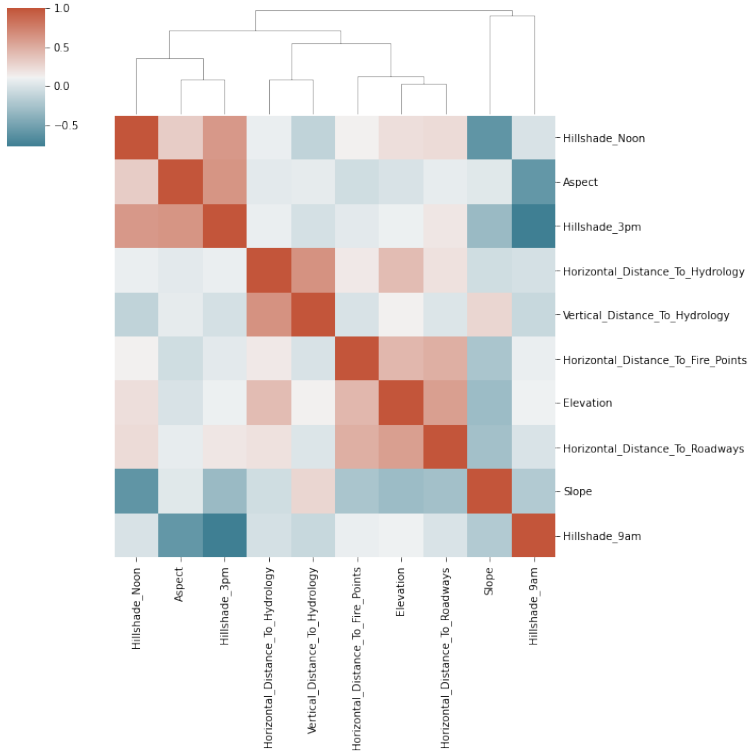
Catégorielle - Projection en dimension 20



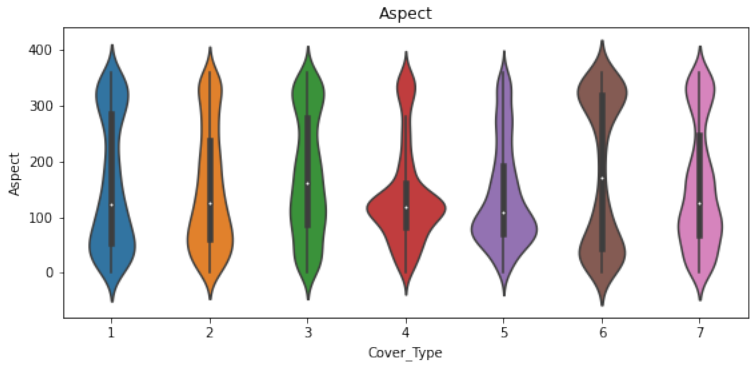
Les 10 variables numériques

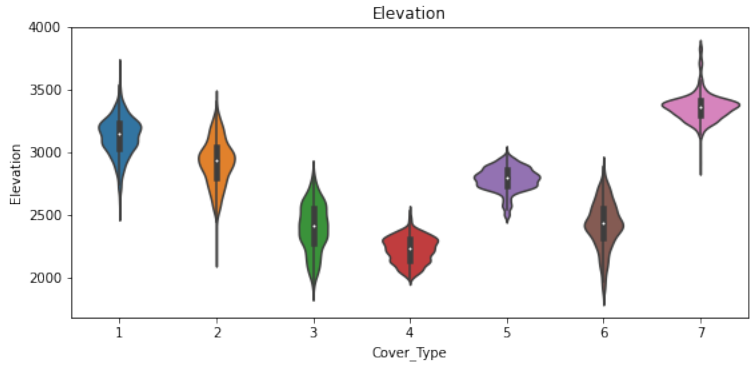
- 3 variables liées à l'ombrage (Noon, 9am, 3pm)
- 5 variables de position, dont deux liées à l'hydrologie
- 2 variables non significatives





Numériques - lien avec le type de forêt



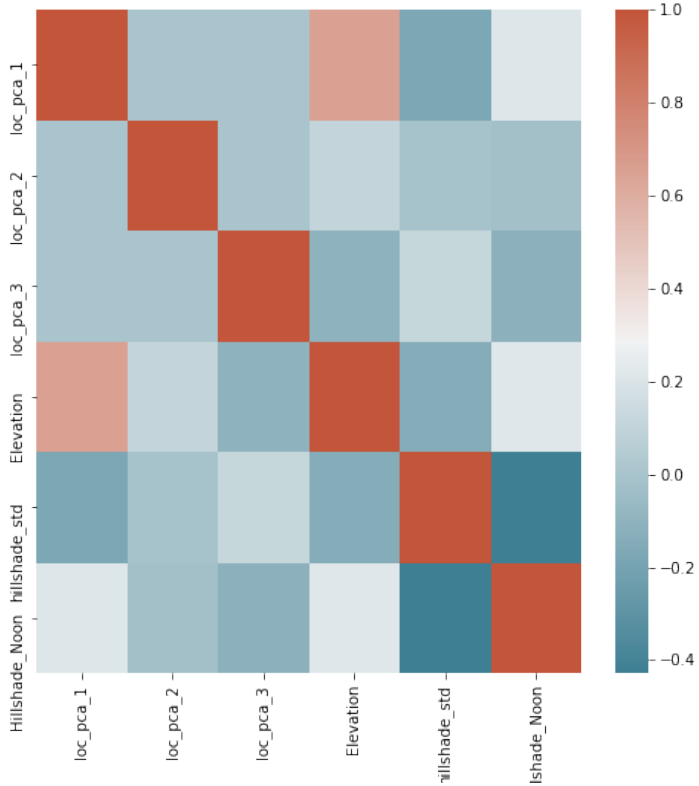



```
Hillshade_Noon                      0.044745
Aspect                              0.018814
Hillshade_3pm                       0.072812
Horizontal_Distance_To_Hydrology    0.129856
Vertical_Distance_To_Hydrology      0.028816
Horizontal_Distance_To_Fire_Points  0.228525
Elevation                           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	QDA
0	0.799735	0.701587	0.773016	0.732804	0.689683	0.775132	0.749206	0.490212	0.441005
1	0.791799	0.687831	0.802910	0.731481	0.698148	0.768783	0.742857	0.466138	0.404762
2	0.794974	0.703439	0.777513	0.740212	0.693915	0.780423	0.771693	0.633069	0.576984
3	0.813492	0.702381	0.775397	0.696561	0.682804	0.785450	0.737566	0.593122	0.409524
4	0.826190	0.719577	0.133598	0.714286	0.700265	0.710317	0.755026	0.596032	0.427778
5	0.791005	0.696561	0.807143	0.703704	0.675397	0.756349	0.739683	0.596032	0.406878
6	0.797354	0.703439	0.754233	0.705026	0.661376	0.774339	0.739947	0.607937	0.577249
7	0.791005	0.697354	0.807143	0.697884	0.677513	0.754233	0.741005	0.593122	0.405820

Les temps :

	Nearest Neighbors	Linear SVM	RBF SVM	Decision Tree	Random Forest	Neural Net	AdaBoost	Naive Bayes	QDA
0	17.019813	6.468931	59.618168	0.462922	0.687727	14.107391	3.533867	0.378033	0.706198
1	14.893226	5.541974	45.061024	0.398874	0.676596	12.019682	3.369611	0.343076	0.791143
2	14.247006	5.540740	48.553928	0.701491	0.758192	13.865766	6.933675	0.358742	0.992963
3	14.885334	5.190781	90.084564	0.931312	0.788404	14.192776	7.847549	0.363959	0.712773
4	17.430017	732.697043	74.762680	1.205033	0.987755	12.481229	12.279679	0.408645	1.078117
5	15.605935	6.514596	45.221306	0.934420	0.890597	12.406370	9.829463	0.332906	0.781618
6	15.199486	5.379707	51.099873	0.781930	0.843552	14.861792	9.541407	0.327263	0.617705
7	15.685208	6.191312	44.658276	0.857266	0.847621	13.177748	9.057373	0.364178	0.619402

Classification - Optimisation des modèles

Les scores :

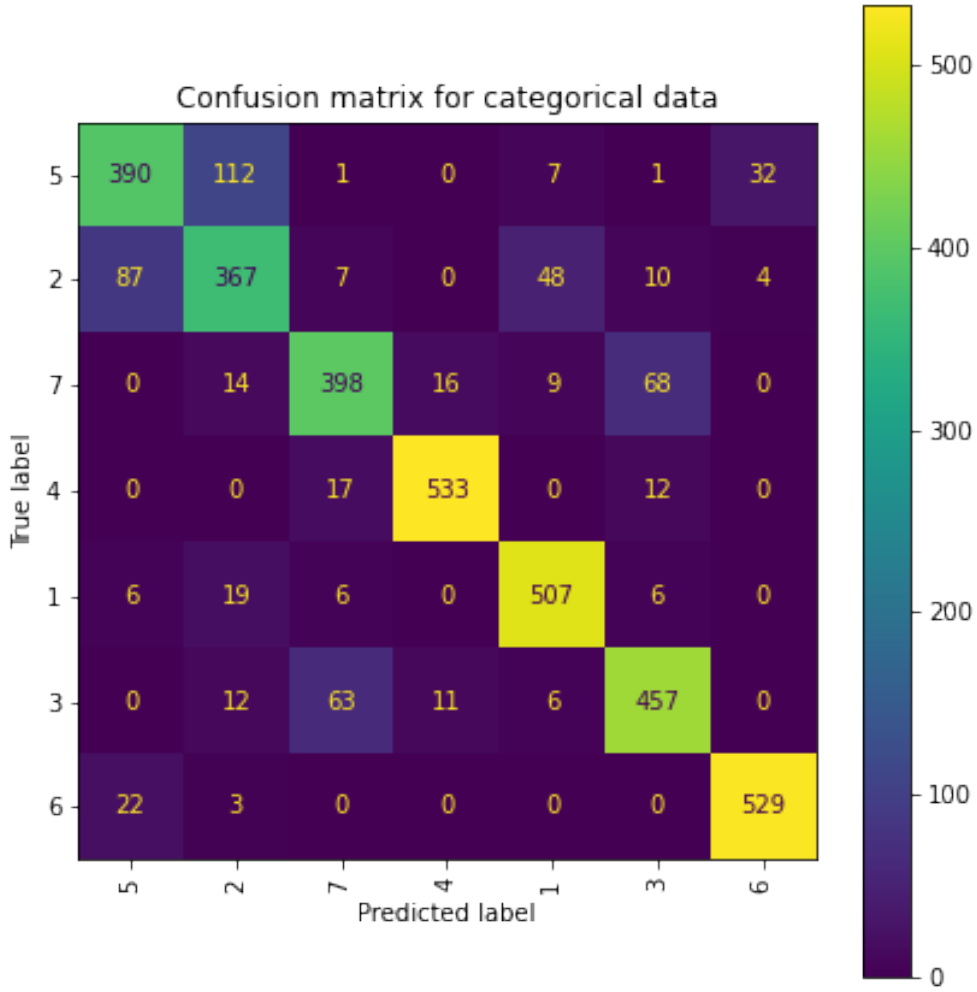
	Nearest Neighbors	Decision Tree	Neural Net	AdaBoost	SVM_rbf	SVM_linear	SVM_sigmoid
0	0.799735	0.790212	0.774074	0.760582	0.783862	0.732540	0.325661
1	0.791799	0.780423	0.773280	0.751058	0.806349	0.715873	0.308466
2	0.794974	0.797090	0.763228	0.788624	0.786508	0.726984	0.302116
3	0.814286	0.785714	0.796296	0.776190	0.816931	0.723016	0.283069
4	0.826190	0.789683	0.777778	0.782011	0.133598	0.726720	0.193651
5	0.791005	0.750794	0.787831	0.772222	0.809524	0.714815	0.311640
6	0.797354	0.761905	0.777513	0.762963	0.761905	0.726984	0.281481
7	0.804233	0.774603	0.773280	0.757672	0.787037	0.722751	0.260053

Les temps (pour le choix des paramètres) :

	Nearest Neighbors	Decision Tree	Neural Net	AdaBoost	SVM_rbf	SVM_linear	SVM_sigmoid
0	42.717831	2.703397	260.574387	48.780626	87.938108	444.921498	31.316665
1	36.485941	2.271381	197.072983	44.361669	58.521966	234.759884	25.200056
2	35.936939	7.693201	197.123844	110.268960	57.799640	202.923640	26.891430
3	35.364199	7.707678	186.278281	109.292224	55.984566	357.543533	24.856390
4	41.902307	16.223135	239.357029	214.816663	113.423352	15550.490489	30.000337
5	36.461801	11.719659	183.074093	162.381018	58.193633	74.529723	29.252898
6	36.015026	10.841118	183.170215	150.953557	61.400271	238.143942	25.600023
7	35.006364	9.547890	172.476833	136.462530	56.014080	472.914845	24.570857

Meilleur modèle - table de confusion

SVC avec un noyau RBF



Meilleurs modèles - autres métriques

	model_name	dataset	AUC	balanced_accuracy	kappa_score	matthews_corrcoef	hinge_loss	params
0	Nearest Neighbors	4.0	0.897077	0.823105	0.797108	0.798199	0.347619	3.0
1	SVM_rbf	5.0	0.909441	0.844411	0.820943	0.821033	0.306878	100.0
2	SVM_rbf	3.0	0.905777	0.838094	0.813839	0.813983	0.319048	1000.0
3	SVM_rbf	0.0	0.907178	0.840544	0.816320	0.816401	0.314815	1000.0
4	SVM_rbf	7.0	0.905558	0.837703	0.813524	0.813678	0.319577	1000.0
5	SVM_rbf	1.0	0.906514	0.839393	0.815080	0.815176	0.316931	1000.0

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 !