

Projet 8 : Participez à une compétition Kaggle !

Claire Gayral

Août 2021

Contents

1	Analyse exploratoire	2
1.1	Description des données	2
1.2	Variables catégorielles	2
1.3	Variables numériques	4
1.4	Les jeux d'entraînement	5
2	Classification	6

Introduction

Kaggle est une plateforme qui propose des compétitions de datascience. Parmi ces compétitions, la [compétition "Prédiction du type de couverture forestière"](#) propose un ensemble de données cartographiques, avec pour objectif de classer les différents sols des forêts observées. Les variables sont toutes définies dans [la page de présentation du projet kaggle](#). Les données sont téléchargeables depuis cette même page.

Pour répondre à la problématique, plusieurs modélisations issues d'une exploration des données vont être proposées, et plusieurs modèles de classification seront testés.

1 Analyse exploratoire

La première étape, quelque soit le modèle de classification, est de nettoyer les données, et d'en extraire des variables pertinentes.

1.1 Description des données

Chaque ligne des tables représente une cellule donnée de 30 x 30 mètres. Les données d'entraînement sont composées de 15120 observations, avec le type de couverture forestière associée, et les données de sortie de 565892 cellules.

Il s'agit donc de construire un modèle supervisé à partir :

- des 2 variables catégorielles d'entrée, qui ont respectivement 4 et 40 modes, avec un unique mode par observation.
- des 10 variables numériques
- de la classification des sols donnée sur un certain ensemble d'entraînement. Il y a 7 types de sols à déterminer, et la sortie n'est pas multi-label.

Les données d'entraînement seront divisées en train et test à raison de 20 pour comparer les différentes modélisations. Néanmoins, une fois la meilleure modélisation et le meilleur modèle de classification choisis, les pré-traitements seront relancés sur l'intégralité du jeu d'entraînement pour prédire la classification finale.

La variable de sortie est une variable catégorielle encodée de façon ordinale. Elle est distribuée uniformément sur l'ensemble d'entraînement.

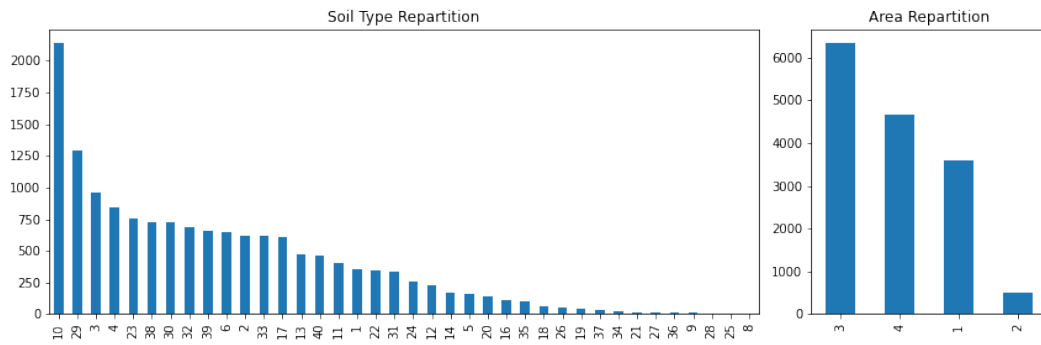
L'analyse exploratoire a été divisée en deux pour séparer le traitement des variables catégorielles et numériques.

1.2 Variables catégorielles

Les deux variables catégorielles sont encodées en "one-hot", et les classes sont bien séparées. La première variable correspond au type de sol, la deuxième à la zone géographique. Il n'est pas très clair que la première variable soit bien une variable "cartographique" et non "télé-détectée", mais comme cela n'est pas explicité, nous allons la garder.

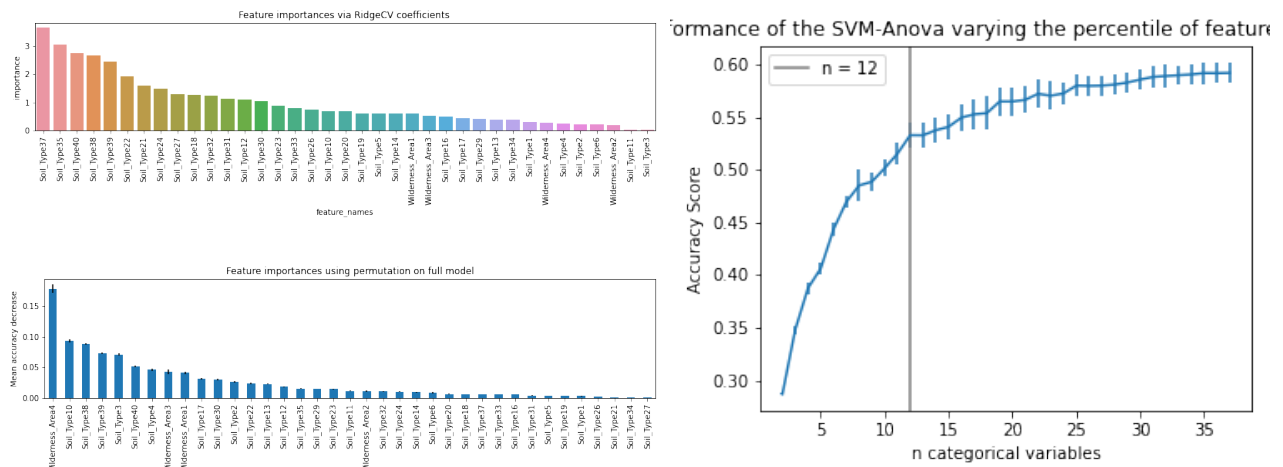
Analyse des variables

Les variables catégorielles ne sont pas réparties uniformément :

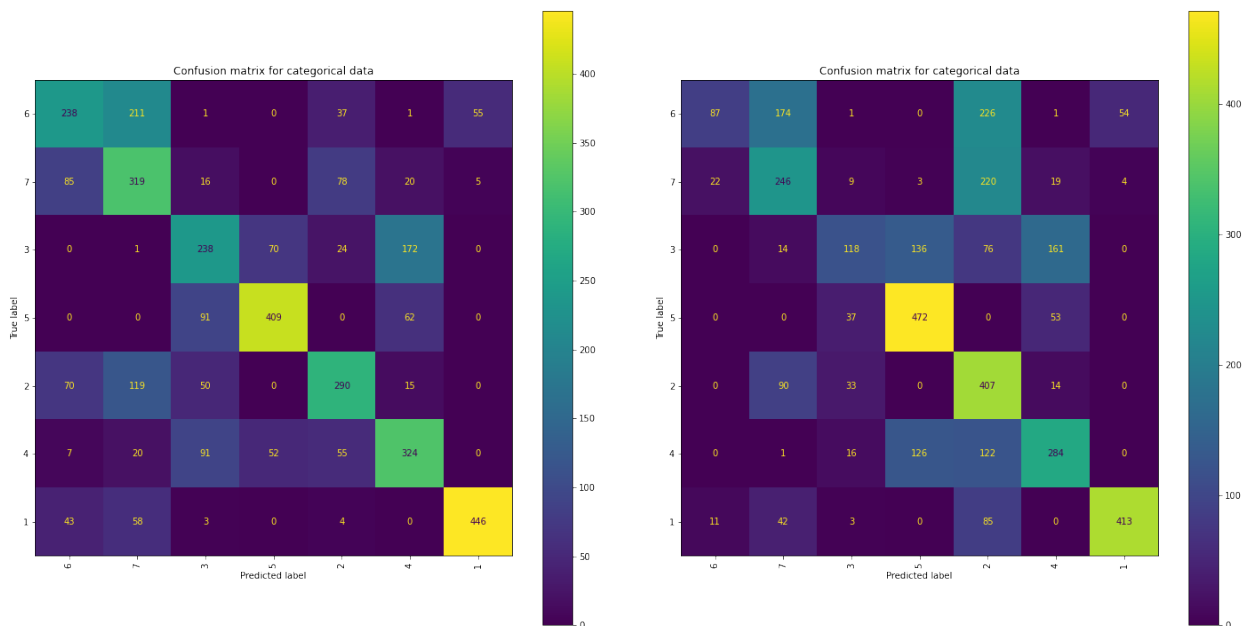


La variable catégorielle correspondant aux types de sols a beaucoup de catégories. Deux modes de type de sols ne sont pas représentés, il faudra retirer les colonnes associées.

L'importance de chaque mode pour la prédiction de la couverture forestière peut alors permettre de choisir les variables les plus pertinentes. Plusieurs méthodes comme le ridgeCV ou les randomForestClassifiers permettent de classer ainsi les modes par importance :

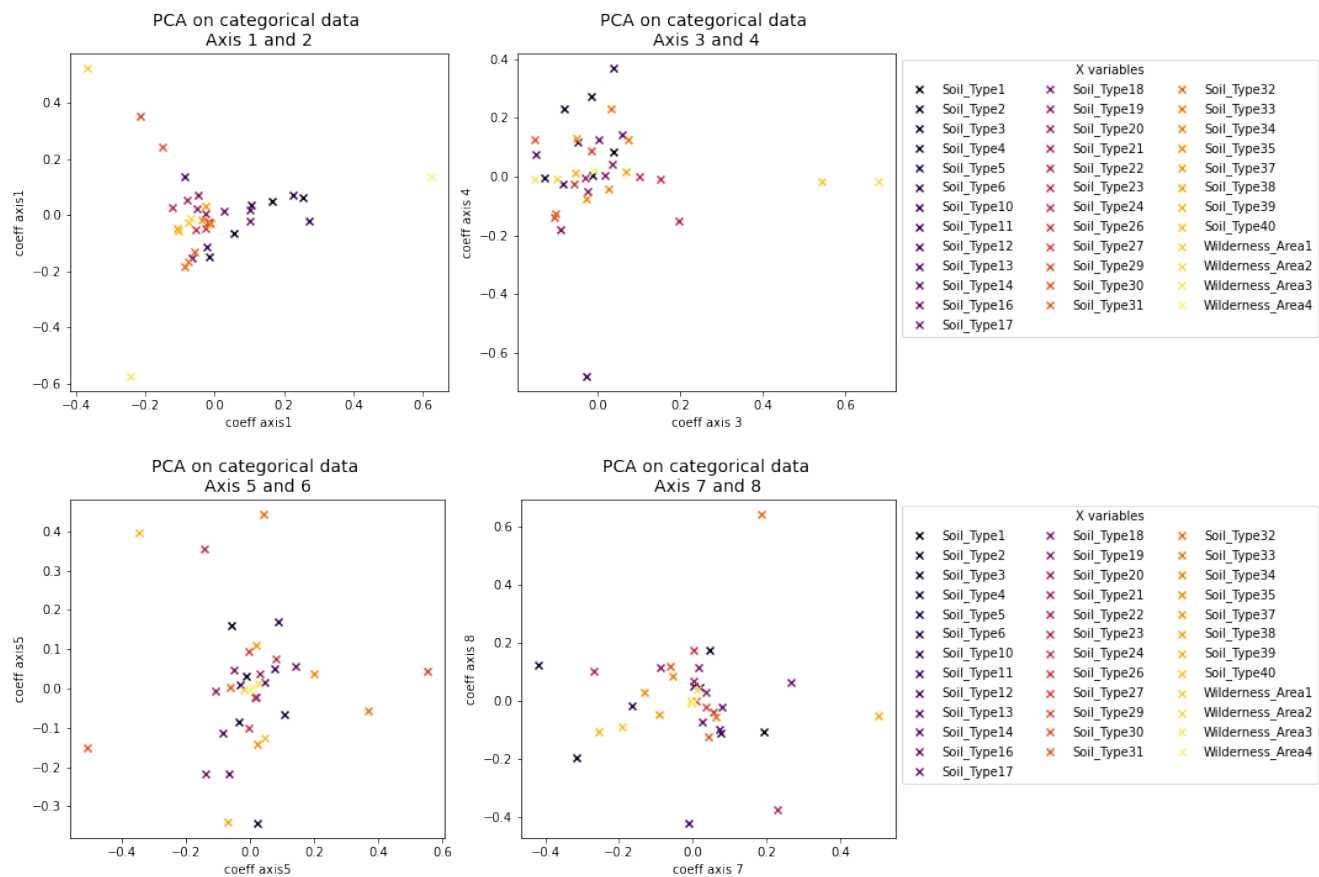


Selon les classifieurs, l'importance n'est pas la même, ce qui limite l'automatisation de ces procédés. De plus, en comparant la matrice de confusion avant et après le retrait des variables les moins importantes, on voit que l'on a juste rendu plus feignant le modèle. En effet, les modes les moins importants coïncident avec les modes les moins représentés, les retirer n'est pas forcément pertinent.



Une autre façon de réduire la dimension de ces variables catégorielles est de les projeter dans un espace de dimension inférieure. Ainsi, la PCA maximise l'accuracy pour 20 composantes, et les 8 premiers axes sont donnés

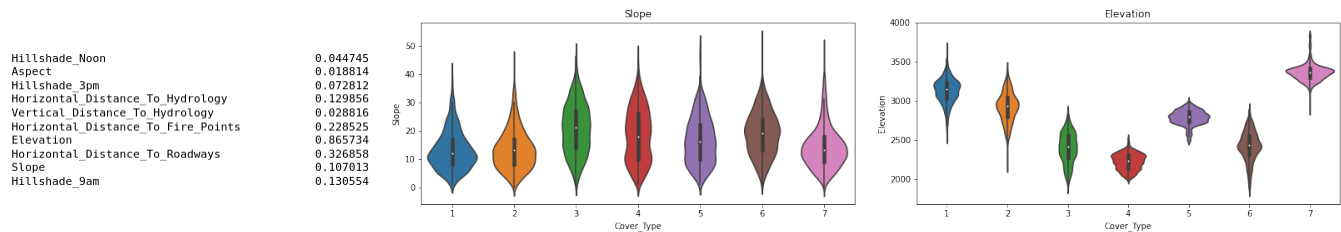
par la pondération suivante :



Les données catégorielles ainsi projetées dans des axes orthonormés de dimension 20 semblent donner des prédictions aussi bonnes qu'en dimension.

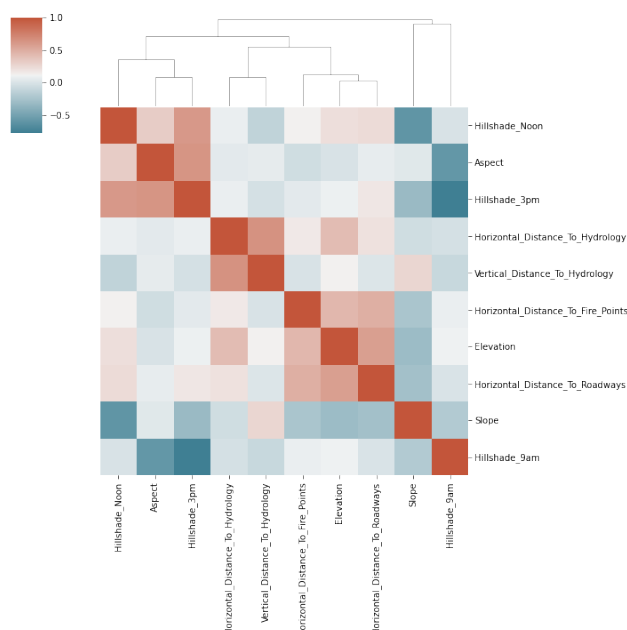
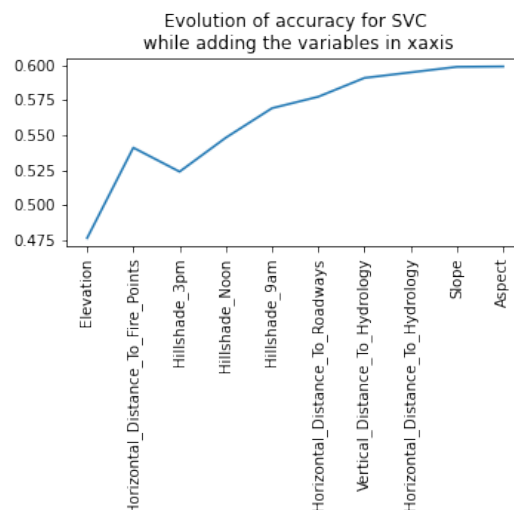
1.3 Variables numériques

Ensuite, les 10 variables numériques ne séparent pas toutes aussi bien les classes de forêt. En effet, les valeurs de η^2 sont très différentes, il s'agit un critère statistique qui mesure à quel point une variable catégorielle est séparée par une variable numérique.



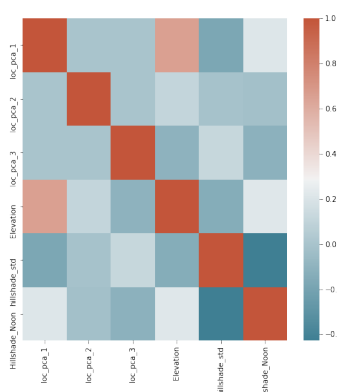
Ainsi, les classes à prédire sont plutôt bien séparées par la variable numérique "Elevation" et très mal par la variable "Aspect".

Comme pour les variables catégorielles, une sélection de variable est faite. On retrouve bien l'ordre d'importance donnée par les η^2 , où la variable "Elevation" est prépondérante, quand "Aspect" est très peu (voir pas du tout) significative. En ajoutant la variable "Hillshade 3pm", l'accuracy chute. Cela veut peut-être dire que cette variable est corrélée aux premières, et bruite plus les données qu'elle n'apporte d'informations.



Les variables numériques sont corrélées les unes aux autres. Par exemple, la variable "Vertical_Distance_To_Hydrology" est fortement liée à la variable "Horizontal_Distance_To_Hydrology", et la variable "Hillshade_3pm" est inversement corrélée à "Hillshade_9am". La plupart des modèles de classification font l'hypothèse d'indépendance entre les variables : il vaudra mieux projeter les données dans un espace où les variables ne sont pas corrélées.

Avant de retirer des variables numériques, il est important d'extraire des variables plus pertinentes à partir de ces variables numériques. Par exemple, ce n'est pas tant l'ombre aux différentes heures de la journée qui compte, mais plutôt la variation d'ensoleillement. C'est ainsi que les 3 variables concernant l'ombre sont résumées en deux variables la variation d'ombre et l'ombre moyenne.



Des variables concernant l'hydrologie, seule la distance horizontale sera conservée, car la distance totale sépare moins bien les classes de forêt, et la distance verticale est corrélée à la distance horizontale. Enfin, les 3 variables de localisation (loin des routes, des points de feux, des points d'eau) sont corrélées. Pour limiter le biais de cela, les variables sont projetées dans un espace orthonormé. Avec ces prétraitements, la matrice de corrélation indique que cette représentation des données

1.4 Les jeux d'entraînement

A partir des différents prétraitements explicités ci-dessus, différents jeux d'entraînement sont extraits :

- dataset 0 : le jeu de données initial, seules les variables constantes ont été retirées
- dataset 1 : les variables catégorielles non importantes sont retirées, et les variables numériques sont prétraitées

- dataset 2 : les variables catégorielles sont projetées dans un espace orthonormé adapté (de dimension 20), et les variables numériques sont prétraitées
- dataset 3 : les variables catégorielles sélectionnées sont projetées, et les variables numériques prétraitées sont orthonormalisées

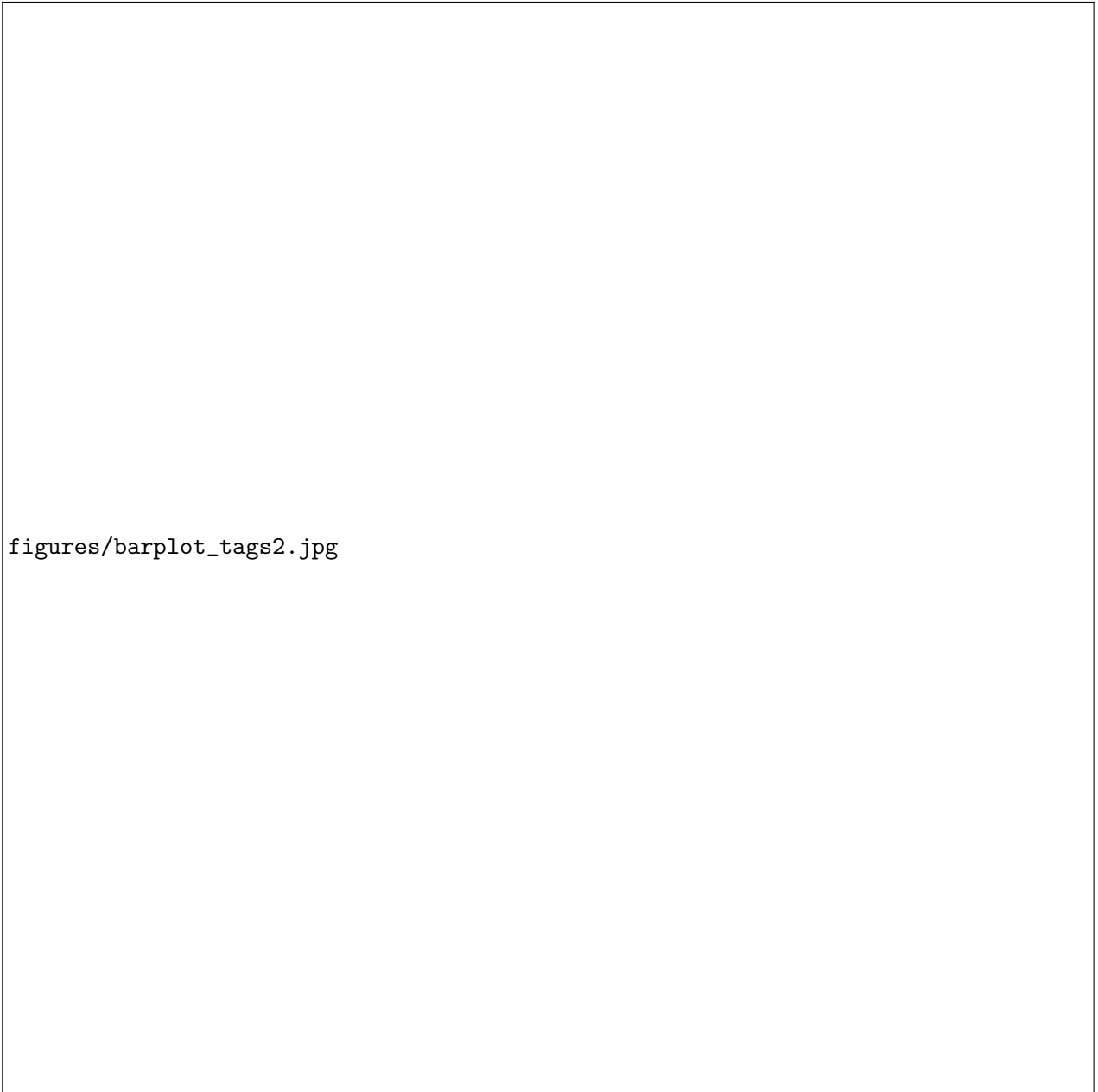
Les datasets 4, 5, 6 et 7 correspondent aux datasets précédents orthonormalisés. L'orthonormalisation permet de supprimer les corrélations entre les variables, ce qui donne de meilleurs résultats, néanmoins, la projection rend compliqué l'interprétation du rôle dans la classification de chaque variable initiale.

2 Classification

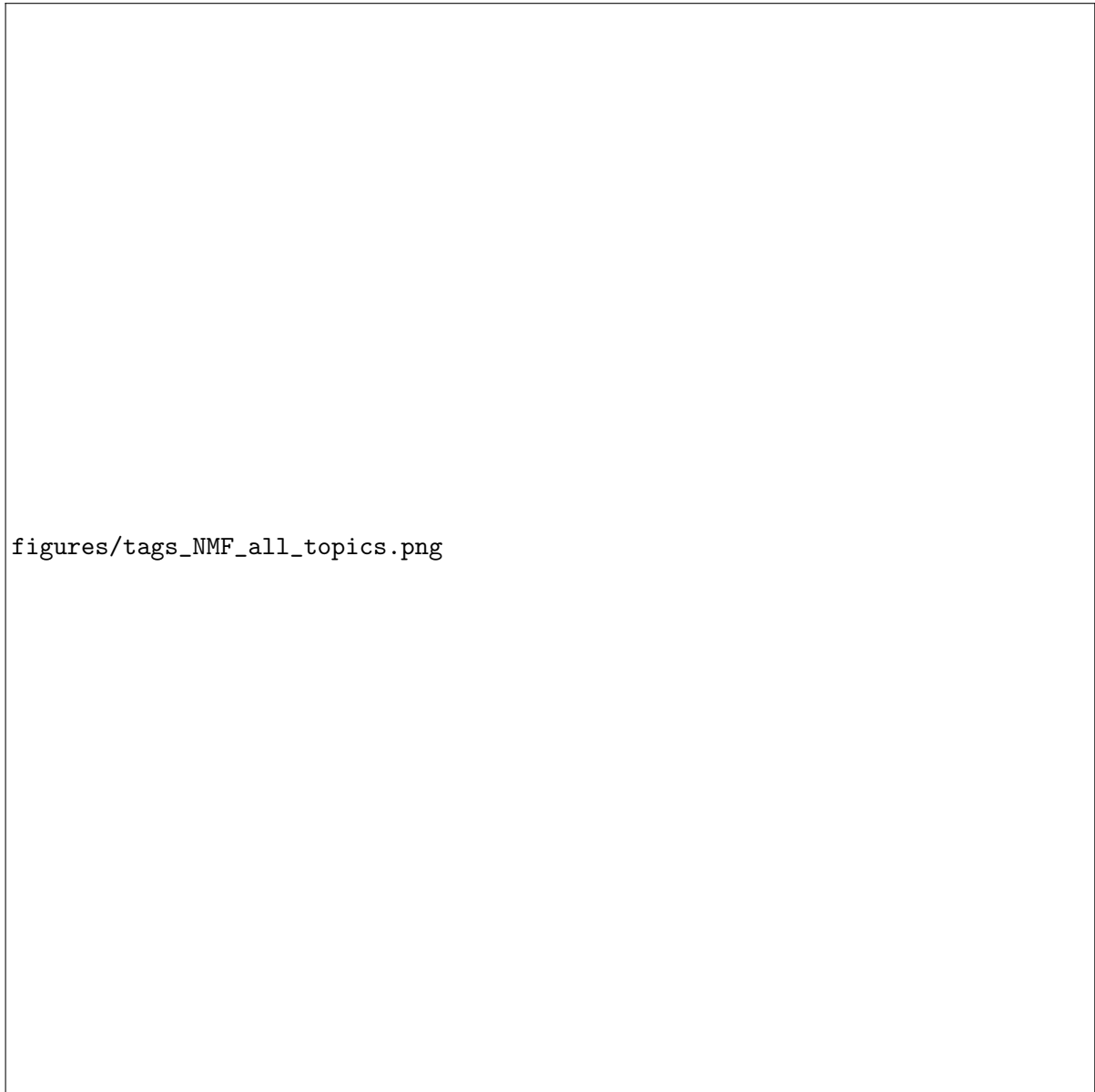
Références

Annexe

Distribution des tags :



Topics pour la NMF sur les tags :



figures/tags_NMF_all_topics.png

Projection des tags dans 2 dimensions :

figures/tags_NMF_PCA_coeffs12.jpg

Choix des hyper-paramètres par Cross Validation (CV) pour la LDA sur le corpus :

figures/corpus_LDA_CV_optim_model_MSE.png

figures/corpus_LDA_CV_optim_model_time.png

Projection LDA du corpus :

figures/coprus_LDA_small_topics.png

Les courbes issues de la matrice de confusion pour la classification binaire :

figures/ROC_prec_recall_naivebayes_BernoulliNB.png

figures/ROC_prec_recall_naivebayes_ComplementNB.png

figures/ROC_prec_recall_naivebayes_MultinomialNB.png

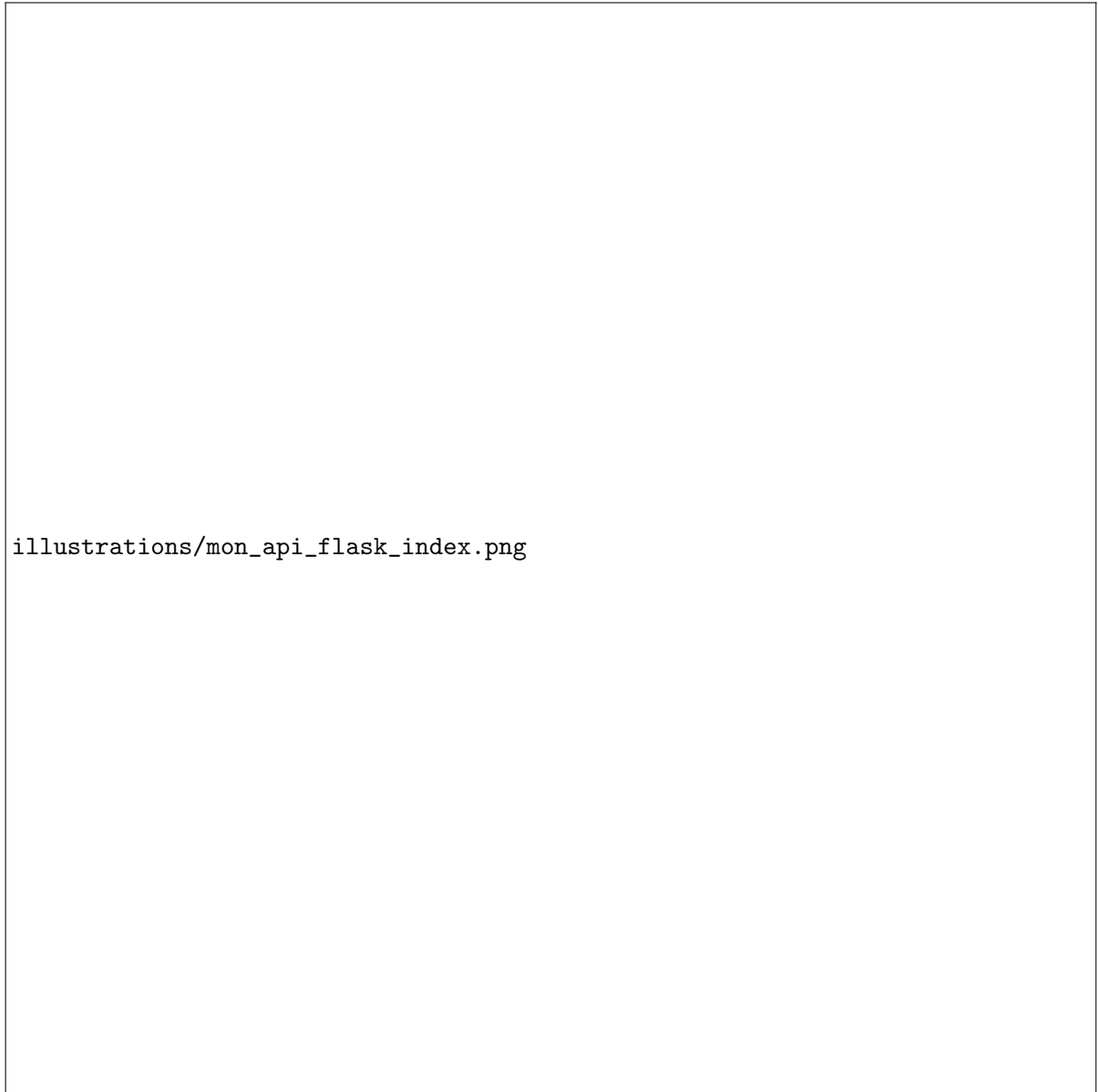
figures/ROC_prec_recall_gradient_boosting_exponential.png

figures/ROC_prec_recall_gradient_boosting_deviance.png

Matrice de contingence, gradient boosting multi-class avec un label pris au hasard parmi les tags de chaque poste :

figures/confusion_matrix_multiclass_GradientBoostingClassifier.png

L'API produite avec flask :
La page d'accueil



illustrations/mon_api_flask_index.png

Formulaire

illustrations/mon_api_create.png

Résultat

illustrations/mon_api_results.png