Blocks Classification

THOMAS MOULY AYMERIC LEBOUCHER

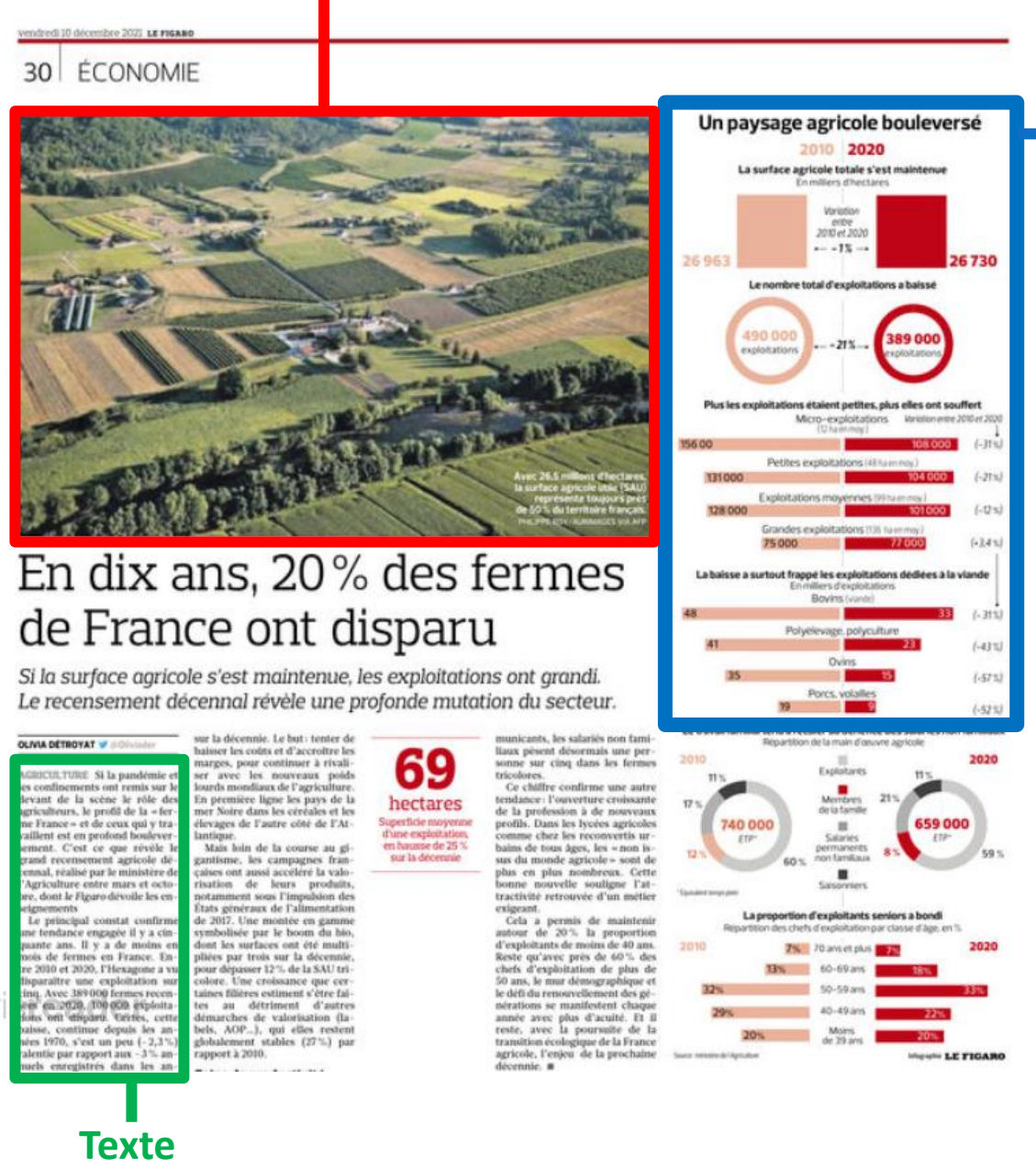
Le contexte de l’étude

- Domaine : Analyse automatique de document
- Filère : Segmentation de bloc
- Problématique : Classifier la nature d'un bloc en fonction de paramètre graphiques, géometriques.
- Chaine de protocole: Document Analysis -> Bloc segmentation -> **Bloc classification** -> Text extraction
- Intéret: Extraire l'information textuelle, conditionner un algorithme de data mining
- Documents cibles : Formulaires, chèque bancaire

Image

Graphique

Texte



vendredi 10 décembre 2021 LE FIGARO

30 | ÉCONOMIE

Avec 26,5 millions d'hectares, la surface agricole utile (SAU) représente toujours près de 50 % du territoire français.

En dix ans, 20 % des fermes de France ont disparu

Si la surface agricole s'est maintenue, les exploitations ont grandi. Le recensement décennal révèle une profonde mutation du secteur.

OLIVIA DETROYAT

AGRICULTURE Si la pandémie et les confinements ont remis sur le devant de la scène le rôle des agriculteurs, le profil de la « ferme France » et de ceux qui y travaillent est en profond bouleversement. C'est ce que révèle le grand recensement agricole décennal, réalisé par le ministère de l'Agriculture entre mars et octobre, dont *le Figaro* dévoile les enseignements.

Le principal constat confirme une tendance engagée il y a cinquante ans. Il y a de moins en moins de fermes en France. Entre 2010 et 2020, l'Hexagone a vu disparaître une exploitation sur cinq. Avec 389 000 fermes recensées en 2020, 100 000 exploitations ont disparu. Cette baisse, continue depuis les années 1970, s'est un peu (- 2,3 %) ralentie par rapport aux - 3 % annuels enregistrés dans les années sur la décennie. Le but : tenter de baisser les coûts et d'accroître les marges, pour continuer à rivaliser avec les nouveaux poids lourds mondiaux de l'agriculture. En première ligne les pays de la mer Noire dans les céréales et les élevages de l'autre côté de l'Atlantique.

Mais loin de la course au gigantisme, les campagnes françaises ont aussi accéléré la valorisation de leurs produits, notamment sous l'impulsion des États généraux de l'alimentation de 2017. Une montée en gamme symbolisée par le boom du bio, dont les surfaces ont été multipliées par trois sur la décennie, pour dépasser 12 % de la SAU tricolore. Une croissance que certaines filières estiment s'être faites au détriment d'autres démarches de valorisation (labels, AOP...), qui elles restent globalement stables (27 %) par rapport à 2010.

69 hectares
Superficie moyenne d'une exploitation, en hausse de 25 % sur la décennie

municants, les salariés non familiaux pèsent désormais une personne sur cinq dans les fermes tricolores.

Ce chiffre confirme une autre tendance : l'ouverture croissante de la profession à de nouveaux profils. Dans les lycées agricoles comme chez les reconvertis urbains de tous âges, les « non issus du monde agricole » sont de plus en plus nombreux. Cette bonne nouvelle souligne l'attractivité retrouvée d'un métier exigeant.

Cela a permis de maintenir autour de 20 % la proportion d'exploitants de moins de 40 ans. Reste qu'avec près de 60 % des chefs d'exploitation de plus de 50 ans, le mur démographique et le défi du renouvellement des générations se manifestent chaque année avec plus d'acuité. Et il reste, avec la poursuite de la transition écologique de la France agricole, l'enjeu de la prochaine décennie. ■

Un paysage agricole bouleversé

2010 2020

La surface agricole totale s'est maintenue
En milliers d'hectares
26 963 — Variation entre 2010 et 2020 : - 1 % — 26 730

Le nombre total d'exploitations a baissé
490 000 exploitations — - 21 % — 389 000 exploitations

Plus les exploitations étaient petites, plus elles ont souffert
Micro-exploitations — Variation entre 2010 et 2020
156 00 — 108 000 (-31 %)
Petites exploitations
131 000 — 104 000 (-21 %)
Exploitations moyennes
128 000 — 101 000 (-12 %)
Grandes exploitations
75 000 — 77 000 (+ 2,4 %)

La baisse a surtout frappé les exploitations dédiées à la viande
En milliers d'exploitations
Bovins (viande)
48 — 33 (-31 %)
Polyélevage, polyculture
41 — 23 (-43 %)
Ovins
35 — 15 (-57 %)
Porcs, volailles
19 — 9 (-52 %)

Répartition de la main d'œuvre agricole
2010 — 740 000 ETP* — 2020 — 659 000 ETP*
Exploitants 11 % / 11 %
Membres de la famille 17 % / 21 %
Salariés permanents non familiaux 12 % / 8 %
Saisonniers 60 % / 59 %
* Équivalent temps plein

La proportion d'exploitants seniors a bondi
Répartition des chefs d'exploitation par classe d'âge, en %
2010 — 2020
70 ans et plus : 7 % / 7 %
60-69 ans : 13 % / 18 %
50-59 ans : 32 % / 33 %
40-49 ans : 29 % / 22 %
Moins de 39 ans : 20 % / 20 %

Source : ministère de l'Agriculture — Infographie LE FIGARO

Exemple

Le Dataset

Toutes les observations sont completes, il ne manque pas de valeurs

Classes: **1,2,3,4,5** respectivement **"text", "horiz. line", "graphic", "vert. line", "picture**“

5473 observations

Attributs

Attribut	Signification
Height: entier	Hauteur du block
Length: entier	Largeur du block
Area: entier.	Aire du block
Eccen: continue.	Excentricité du block (largeur / hauteur);
p_black: continue.	Pourcentage de pixels noirs sur le block
p_and: continue.	Pourcentage de pixels noirs sur le block après Run Length Smoothing Algorithm (RLSA)
mean_tr: continue.	Moyenne des transitions Blanc/Noir (blackpix / wb_trans);
blackpix: continue.	Nombre de pixels noirs sur le block original
blackand: continue.	Nombre de pixels noirs sur le block après application du RLSA
wb_trans: continue.	Nombre de transitions blanc/noir du bloc

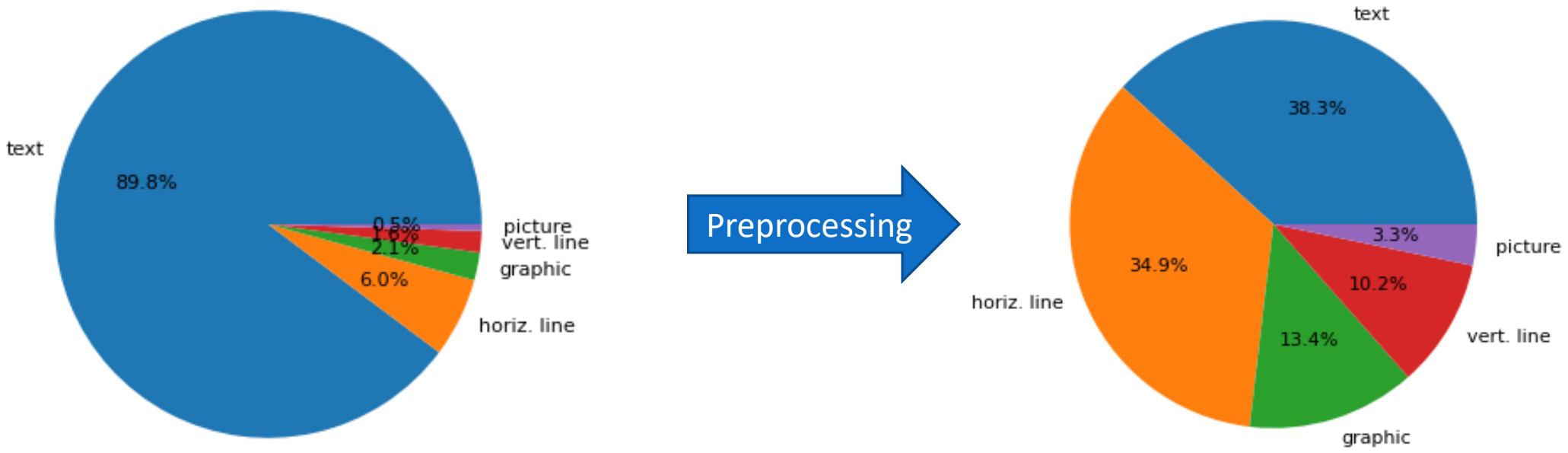
RLSA ALGORITHM

(a)

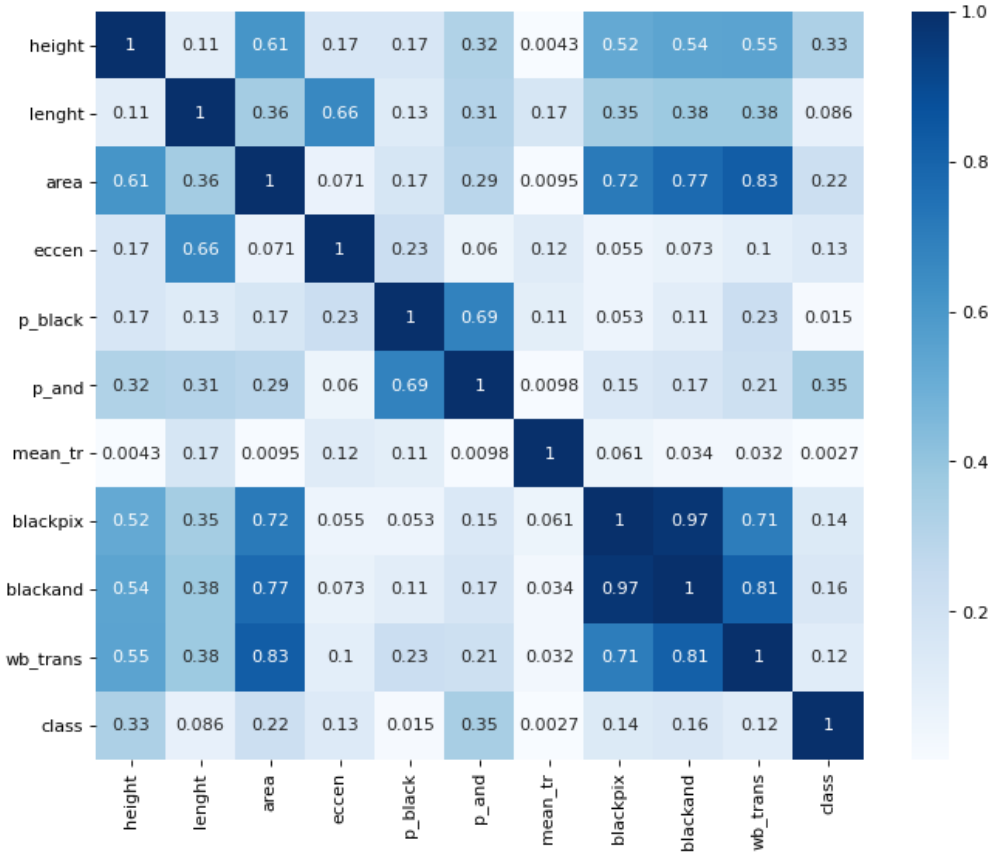
(b)

Datavisualisation

Trop forte présence d'observations de classe texte dans le dataset



Datavisualisation

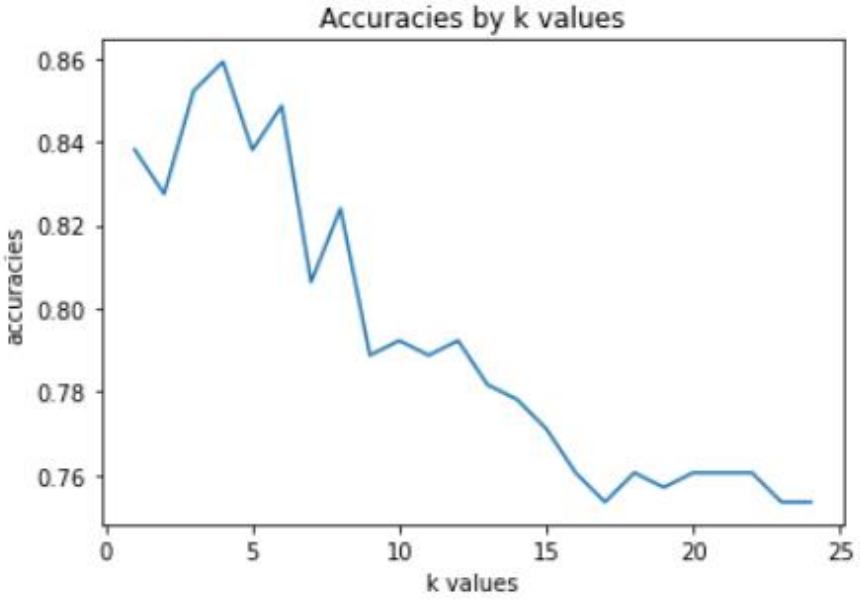


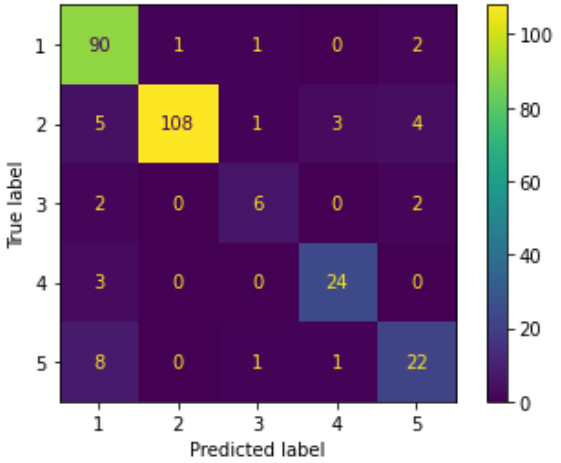
Corrélation à class:

- Height
- Area
- P_and

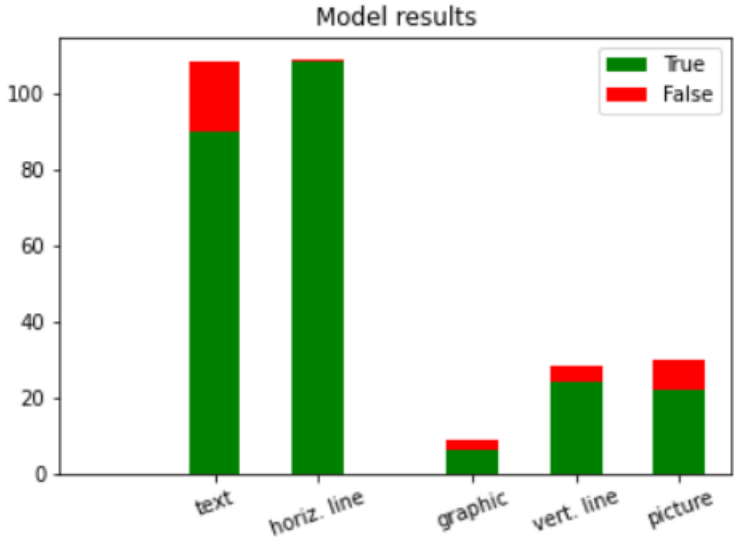
KNN

- Recherche du k optimal pour notre model
- Meilleur k= 4





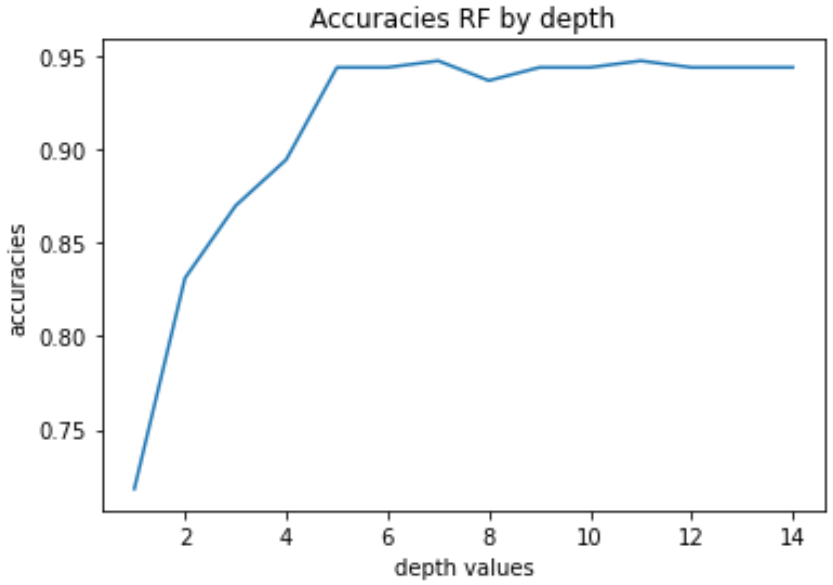
text : 83.3 %
horiz. line : 99.1 %
graphic : 66.7 %
vert. line : 85.7 %
picture : 73.3 %

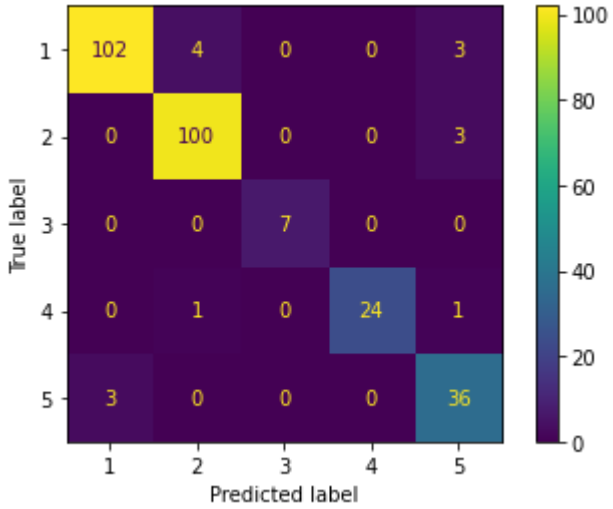


88.03 %

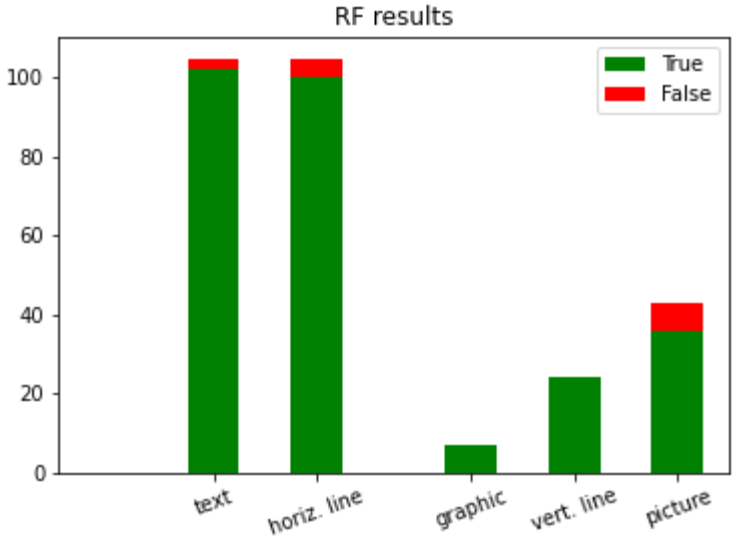
Random Forest

- Recherche de paramètre pour trouver le model optimal
- Meilleur depth= 7 ou 11



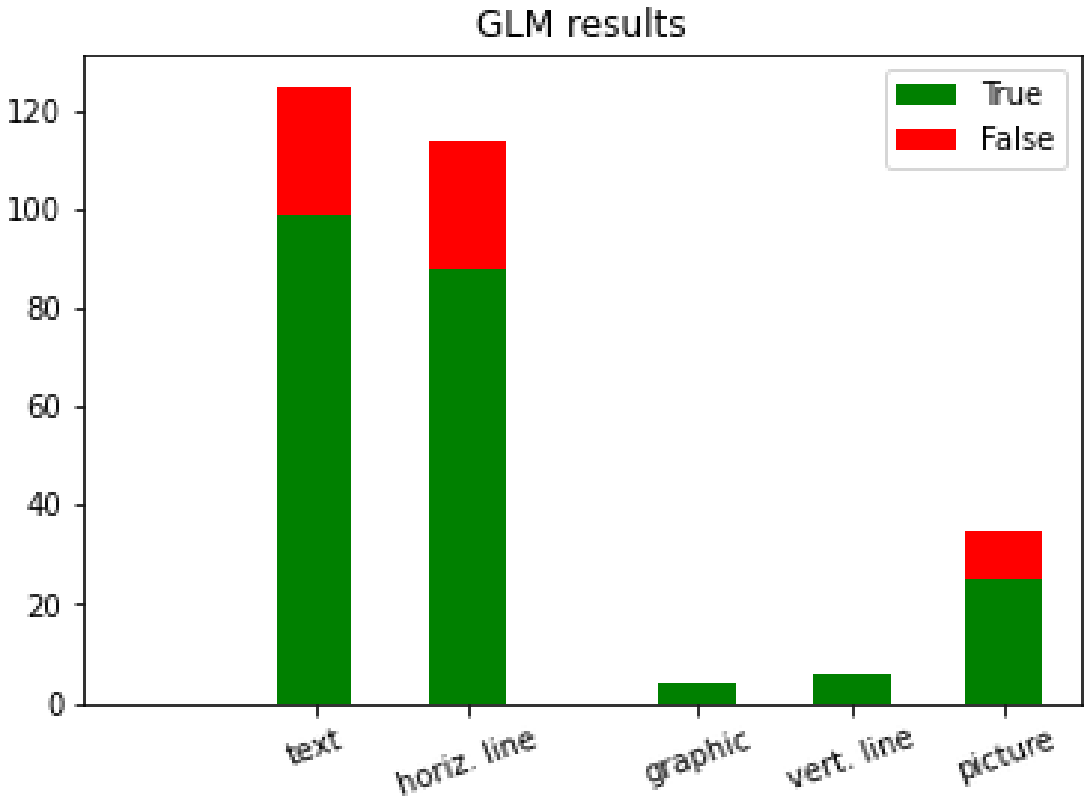


text : 97.1%
horiz. line : 95.1 %
graphic : 100 %
vert. line : 100 %
picture : 83.7 %

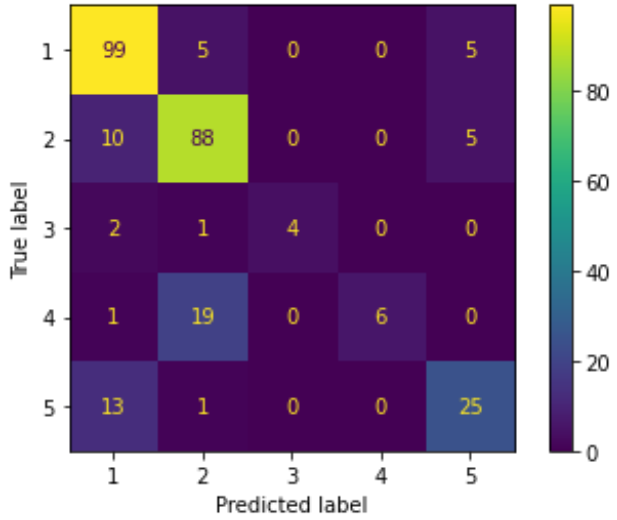


94,72 %

Logistic Regression



text : 72,2%
horiz. line : 77,2 %
graphic : 100 %
vert. line : 100 %
picture : 71,39 %



Accuracy: 78,17 %

API REST

- Input : fichier .csv ou .data

- Output : Bloc classifiés + % de texte dans le document

- aller plus loin pour un potentiel client : une analyse de document "end to end"

Appliquer un algorithme de segmentation à partir d'un fichier pdf, jpg...

Analyse automatique des 9 paramètres qui alimente le modèle

Extraire l'information textuelle