**ddl 17 \_15 octobre 5pm**

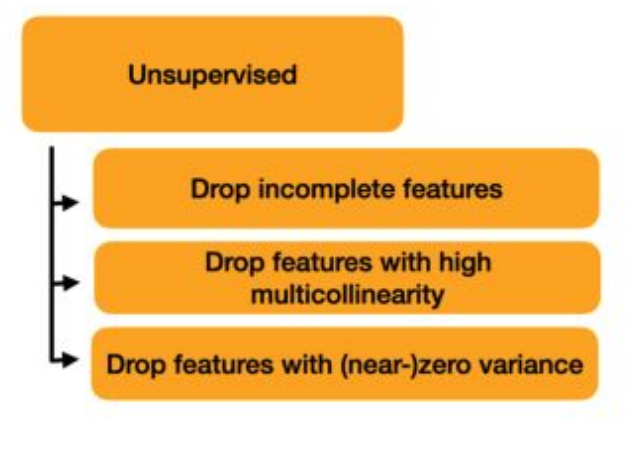
**pdf rapport 5 pages**

* **Notebook**

-visualiser le jeu de données

-comprendre data

-mettre en place les méthodes de feature selection :

  
-Voir si il ya des outilers ou pas (faire des boxplots)

Partien Méthode:

-choisir algorithmes à appliquer ( on peut choisir deux par exemple et les tester ensemble en faisant aussi des ajustements des paramètres dans chaque modele)

-choisir les métriques d’évaluation ( sont déjà donné dans le fichier de description mais on peut choisir d’autres)

-plotter les clusters ….

datat understanding :

| **#** | **Variable** | **Type** | **Signification** | **Exemple** |
| --- | --- | --- | --- | --- |
| 1 | X | numérique (1–9) | Coordonnée **est-ouest** sur la carte du parc. | 6 |
| 2 | Y | numérique (2–9) | Coordonnée **nord-sud** sur la carte du parc. | 5 |
| 3 | month | catégorielle | Mois de l’année (jan → dec) | "aug" |
| 4 | day | catégorielle | Jour de la semaine (mon → sun) | "fri" |
| 5 | FFMC | numérique | **Fine Fuel Moisture Code** — indique l’humidité des combustibles fins (feuilles, herbes sèches). Plus c’est haut → plus c’est sec → feu plus probable. | 86.2 |
| 6 | DMC | numérique | **Duff Moisture Code** — humidité des couches moyennes du sol. Haut = sec. | 26.2 |
| 7 | DC | numérique | **Drought Code** — indique la sécheresse à long terme. | 94.3 |
| 8 | ISI | numérique | **Initial Spread Index** — vitesse initiale de propagation du feu. Haut = feu qui se propage vite. | 5.1 |
| 9 | temp | numérique | Température moyenne (°C). | 18.0 |
| 10 | RH | numérique | **Relative Humidity** (%) — humidité de l’air. Bas = air sec. | 35 |
| 11 | wind | numérique | Vitesse du vent (km/h). | 2.2 |
| 12 | rain | numérique | Pluie (mm/m²). | 0.0 |
| 13 | area | numérique | **Surface brûlée** (hectares). Très asymétrique (souvent 0). |  |

**Étapes Notebook:**

Lecture du jeu de données

Affichage des informations générales sur les colonnes et les types de données

Résumé statistique des variables numériques

Vérification de la présence de valeurs manquantes

Visualisation de la distribution de toutes les variables sous forme d’histogrammes

………………………………………..

Suppression de la colonne 'rain'

Transformation logarithmique de la variable cible 'area'

Visualisation

Transformation racine carrée de 'area' pour comparaison avec la log-transformation

Visualisation après transformation racine carrée

……………………………….

Visualisation des outliers

Supprimer les outliers selon la méthode de l’IQR

This method will sadly drop 40% of the data

Encodage des variables catégorielles 'month' et 'day' en valeurs numériques

**Etude des corrrealtion entre les variables:**

**-spearman**

**-kendall**

**-Application du PCA**

**(on va tester avec 4 et 6 composantes )**

**-Algorithme**

hierarchical

gmm

optics