

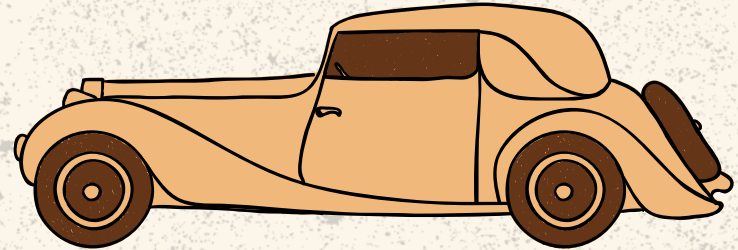
ISEN

ALL IS DIGITAL!

NANTES

Projet A3 - 1A

Antonin SOQUET
Maxence LAURENT
Martin LOBEL



Sommaire

01

Découverte et
préparation des
données

04

Scripts

02

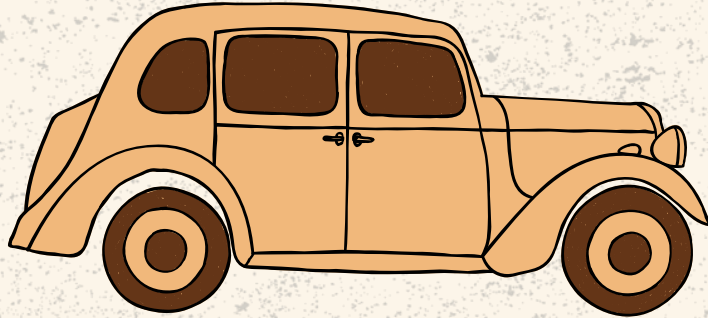
Apprentissage non-
supervisé

05

Organisation

03

Apprentissage supervisé



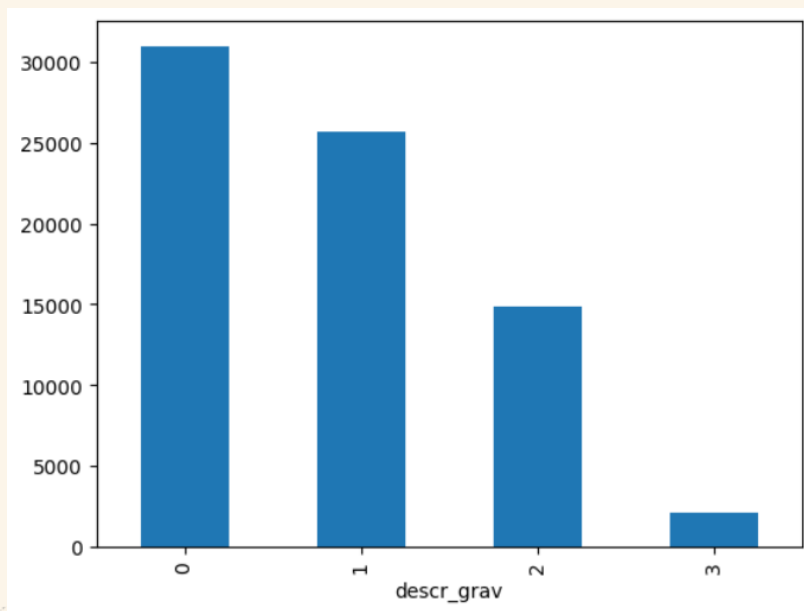
1

Découverte et préparation des données

Découverte des données

```
Valeur cible 1 : id_code_insee - 8258
Valeur cible 2 : Num_Acc - 40250
Valeur cible 3 : num_veh - 58
Valeur cible 4 : id_usa - 73643
Valeur cible 5 : date - 29950
Valeur cible 6 : ville - 8078
Valeur cible 7 : latitude.x - 1253
Valeur cible 8 : longitude.x - 1495
Valeur cible 9 : descr_cat_veh - 24
Valeur cible 10 : descr_agglo - 2
Valeur cible 11 : descr_athmo - 9
Valeur cible 12 : descr_lum - 5
Valeur cible 13 : descr_etat_surf - 9
Valeur cible 14 : description_intersection - 9
Valeur cible 15 : an_nais - 101
Valeur cible 16 : age - 101
Valeur cible 17 : place - 10
Valeur cible 18 : descr_dispo_secu - 15
Valeur cible 19 : descr_grav - 4
Valeur cible 20 : descr_motif_traj - 6
Valeur cible 21 : descr_type_col - 7
Valeur cible 22 : department_name - 89
Valeur cible 23 : department_number - 89
Valeur cible 24 : region_name - 17
```

Nombre d'instances : 73643



Préparation des données

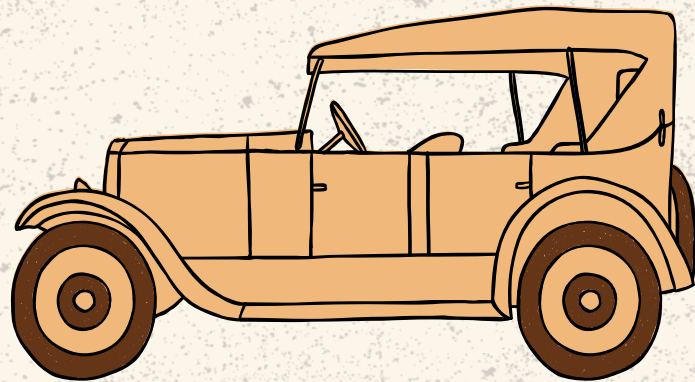
Traitement de la date

```
date
20090803130000
20090227223000
20090113160000
20090113160000
20090209201500
```

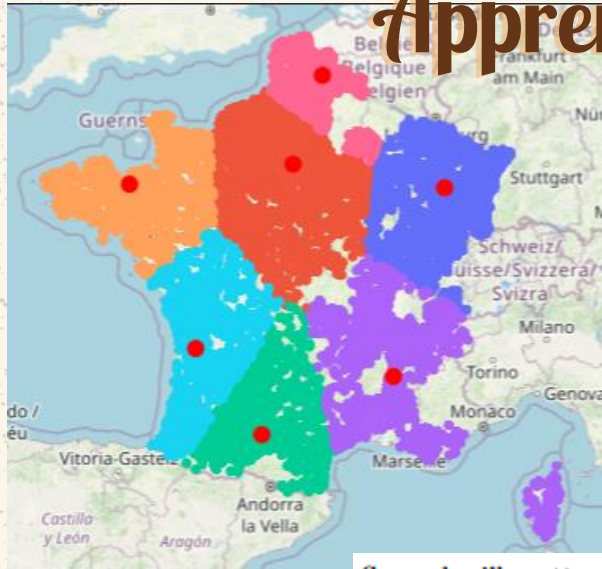
- Transformation des codes_insee en entier, en remplaçant les A et B par 0
- On transforme la colonne 'descr_type_col', 'descr_dispo_secu' en entier
- On supprime la colonne 'num_veh' et 'ville' car elles ne nous sont pas utiles
- On supprime la colonne 'department_name' car elle ne nous est pas utile

2

*Apprentissage
non-supervisé*



Apprentissage non-supervisé

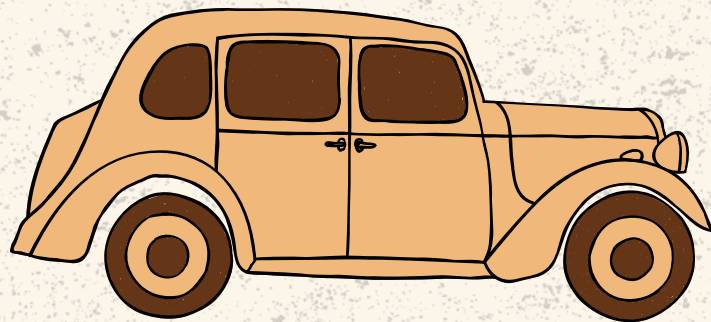


Score de silhouette sur les données manuelles avec Kmeans recodé : -0.04255662179491376
Score de silhouette sur les données manuelles avec Kmeans scikit : 0.56718671213116044
Score de silhouette sur les données PCA avec Kmeans scikit : 0.8674884745935049

Score de Calinski-Harabasz sur les données manuelles avec Kmeans recodé : 6.264823420237192
Score de Calinski-Harabasz sur les données manuelles avec Kmeans scikit : 47433.09353543223
Score de Calinski-Harabasz sur les données PCA avec Kmeans scikit : 9363097.980487494

Score de Davies-Bouldin sur les données manuelles avec Kmeans recodé : 383.76158566961936
Score de Davies-Bouldin sur les données manuelles avec Kmeans scikit : 0.7204794124231818

3



Apprentissage
supervisé

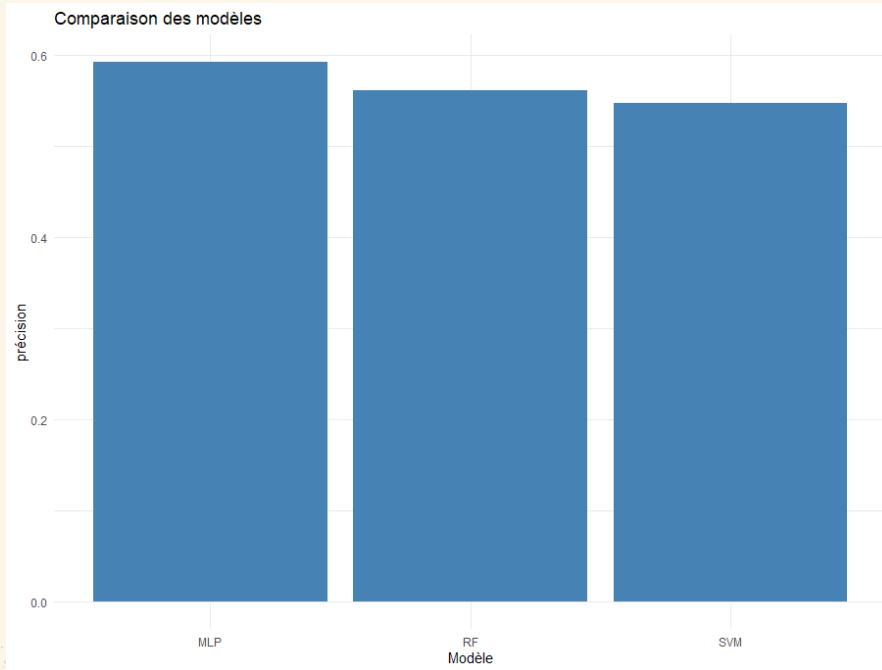
Partie Antonin

Répartition de bases de donnée :

- Holdout
- Leave One Out
- From scratch

+ creation d'un sample
et nettoyage des
données

Précision



Moyenne SVM: 0.5479394392015752

Moyenne RF: 0.5610428406544912

Moyenne MLP: 0.5936316111073393

KNN

From Scratch

Accuracy ~ 48%

Plus simple d'implémentation

Temps d'exécution : ~ 1 seconde

Avec Scikit-learn

Accuracy ~ 48%

Plus complexe d'implémentation

Temps d'exécution : ~ 2 minutes

KNN

From Scratch

Accuracy: 48.4 %

```
[0,  
0,  
0,  
0,  
0,  
0,  
2,  
0,  
0,  
0,  
2,  
2,  
0,  
0,  
0,  
0,  
0,  
2,  
2,  
1,  
0,  
2,  
0,  
1,  
0,  
2,  
...  
2,  
1,  
0,  
2,  
...]
```

Avec Scikit-learn

```
Gravité prédite : [0 0 1 ... 1 2 0]  
Accuracy pour la prédiction de gravité : 43.25 %  
Gravité prédite : [0 0 0 ... 0 2 0]  
Accuracy pour la prédiction de gravité : 47.0 %  
Gravité prédite : [0 0 0 ... 1 2 0]  
Accuracy pour la prédiction de gravité : 47.15 %  
Gravité prédite : [0 0 0 ... 1 0 0]  
Accuracy pour la prédiction de gravité : 47.85 %  
Gravité prédite : [0 0 0 ... 0 0 0]  
Accuracy pour la prédiction de gravité : 48.5 %  
Gravité prédite : [0 0 0 ... 0 0 0]  
Accuracy pour la prédiction de gravité : 49.1 %  
Gravité prédite : [0 0 0 ... 0 0 0]  
Accuracy pour la prédiction de gravité : 48.55 %  
Gravité prédite : [0 0 0 ... 0 0 0]  
Accuracy pour la prédiction de gravité : 48.699999999999996 %  
Gravité prédite : [0 0 0 ... 0 0 0]  
Accuracy pour la prédiction de gravité : 48.699999999999996 %  
Gravité prédite : [0 0 0 ... 0 0 0]  
Accuracy pour la prédiction de gravité : 49.0 %  
Gravité prédite : [0 0 0 ... 0 0 0]  
Accuracy pour la prédiction de gravité : 48.4 %  
Gravité prédite : [0 0 0 ... 0 0 0]  
Accuracy pour la prédiction de gravité : 47.8 %  
Gravité prédite : [0 0 0 ... 0 0 0]  
Accuracy pour la prédiction de gravité : 48.3 %
```

Grid Search

- Paramètre de grille
- Entraînement du modèle
- Recupération du resultat
- Souvegarde du Meilleur modèle
- Affichage des resultats
- Matrice de confusion
- Tableau des valeurs

```
Meilleures parametres SVM: {'C': 10, 'gamma': 'scale', 'kernel': 'rbf'}
Meilleur estimateur: SVC(C=10)
SVM Accuracy: 0.5449915110356537
Precision: 0.5070794374471712
Recall: 0.5449915110356537
F1 Score: 0.4761175146157662
```

Matrice de confusion :

```
[[204  31   1   0]
 [104 114   1   0]
 [ 55  59   3   0]
 [ 10   5   2   0]]
```

	mean_fit_time	std_fit_time	mean_score_time	std_score_time	param_C \
0	0.123796	0.007306	0.066400	0.003440	0.1
1	0.147599	0.011160	0.068401	0.001952	0.1
2	0.131794	0.004925	0.065200	0.003310	1
3	0.153598	0.003506	0.063608	0.001197	1
4	0.150010	0.002818	0.058397	0.000800	10
5	0.208605	0.009769	0.062793	0.001725	10

	param_gamma	param_kernel	params \
0	scale	rbf	{'C': 0.1, 'gamma': 'scale', 'kernel': 'rbf'}
1	auto	rbf	{'C': 0.1, 'gamma': 'auto', 'kernel': 'rbf'}
2	scale	rbf	{'C': 1, 'gamma': 'scale', 'kernel': 'rbf'}
3	auto	rbf	{'C': 1, 'gamma': 'auto', 'kernel': 'rbf'}
4	scale	rbf	{'C': 10, 'gamma': 'scale', 'kernel': 'rbf'}
...			
2	0.528662	0.522718	0.007524 4
3	0.562633	0.553291	0.006660 2
4	0.560510	0.557113	0.008343 1
5	0.528662	0.523142	0.008867 3

Fusion des modèles

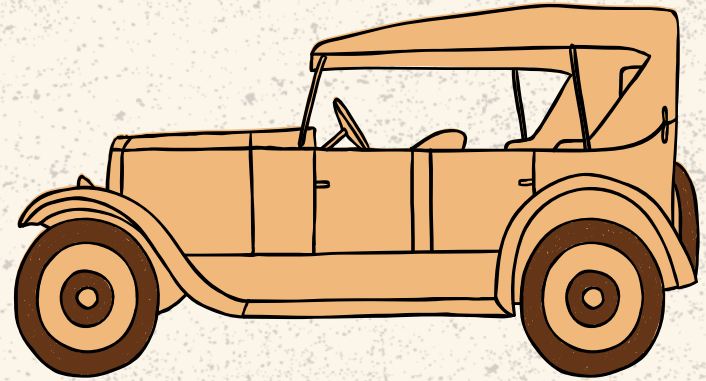
SVM 0.5449915110356537

RandomForestClassifier 0.6027164685908319


MLPClassifier 0.534804753820034

VotingClassifier:0.5959252971137521

4



Scripts

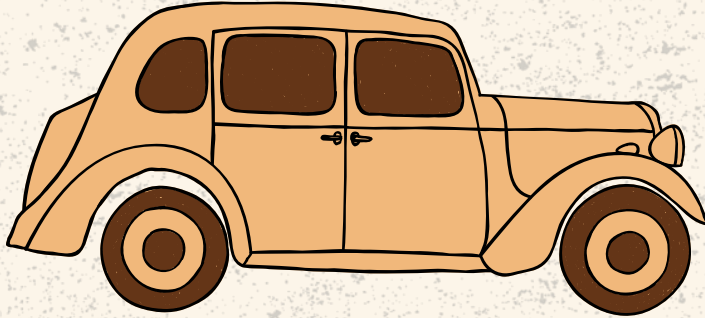
```
>  centroids.csv
latitude,longitude
47.91495500190913,-2.0904344591446837
-60.73254678714826,14.816047108433763
6.8738700000000096,43.65660000000002
48.533997432550045,6.596730368436347
45.9068979986129,4.688783619340094
48.82010688008867,2.308558506372794
-7.105427357601002e-15,5.773159728050814e-15
44.20638240751755,0.1343483566325676
48.011828416180315,0.24488457277117304
43.40616239590007,4.8993150044842695
55.389890839694665,-21.040717557251835
50.41362953730266,2.9151366409612596
1.7139606992188945,46.70095234375011
```

python3 partieX.py args

```
{"gravite": 0}
```

```
{"cluster": {"latitude du centroid": 47.91495500190913, "longitude d
u centroid": -2.090434459144684}, "accident": {"latitude de l'accide
nt": 48.0, "longitude de l'accident": -3.0}}
```

5



Organisation



Diagramme de Gantt, sur GitHub

Antonin	Maxence	Martin
Répartition des tâches	Répartition des tâches	Répartition des tâches
Apprentissage supervisé	Découverte et préparation des données	Découverte et préparation des données
Rédaction du rapport	Apprentissage non-supervisé	Apprentissage supervisé
Scripts py	Setup du Git	Rédaction du rapport
Préparation diapositive	Scripts py	Setup du Git (Gantt)
	Préparation diapositive	Scripts py