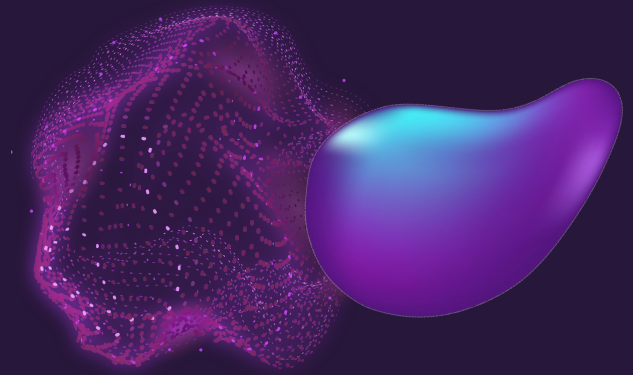


PROJET IA

Application d’étude des accidents de la route

Tristan SAEZ - Vincent LE BRENN
Adrien LEBOUCHER
2023

TABLE DES MATIERES

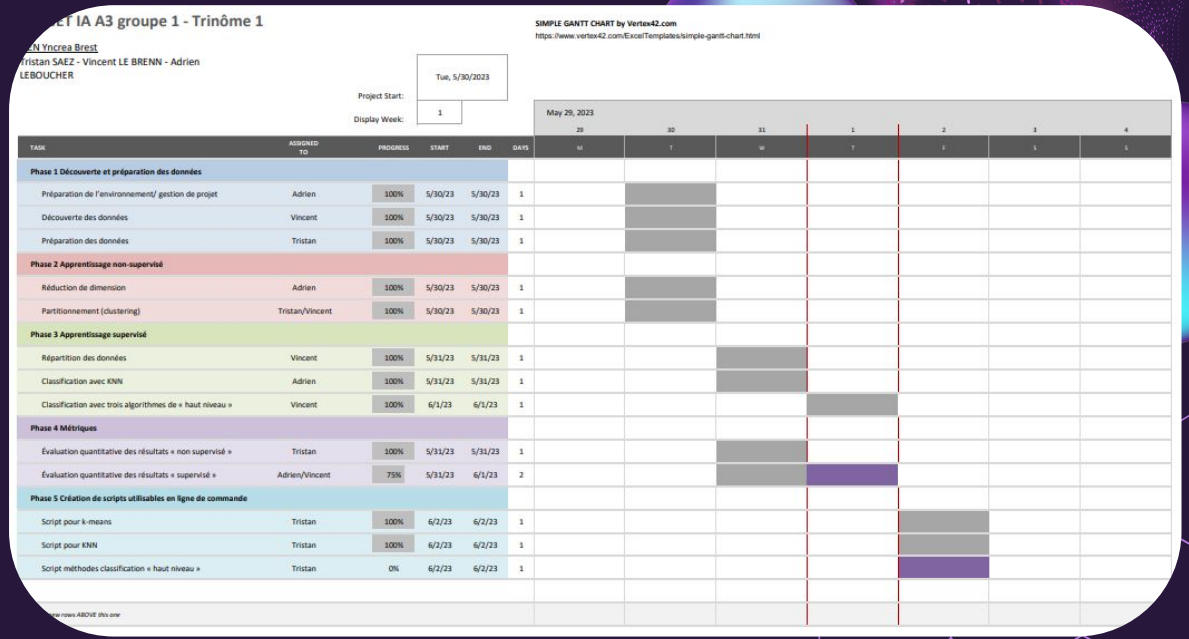
01

GESTION DE PROJET

LE PROJET

- Représentations graphiques & apprentissages à partir d'un fichier csv préalablement traité en Big Data.
- Données d'accidents corporels de la circulation routière

LES OUTILS



ET IA A3 groupe 1 - Trinôme 1

EN Yncrea Brest
Tristan SAEZ - Vincent LE BRENN - Adrien LEBOUCHER

SIMPLE GANTT CHART by Vertex42.com
https://www.vertex42.com/ExcelTemplates/simple-gantt-chart.html

Project Start: Tue, 5/30/2023
Display Week: 1

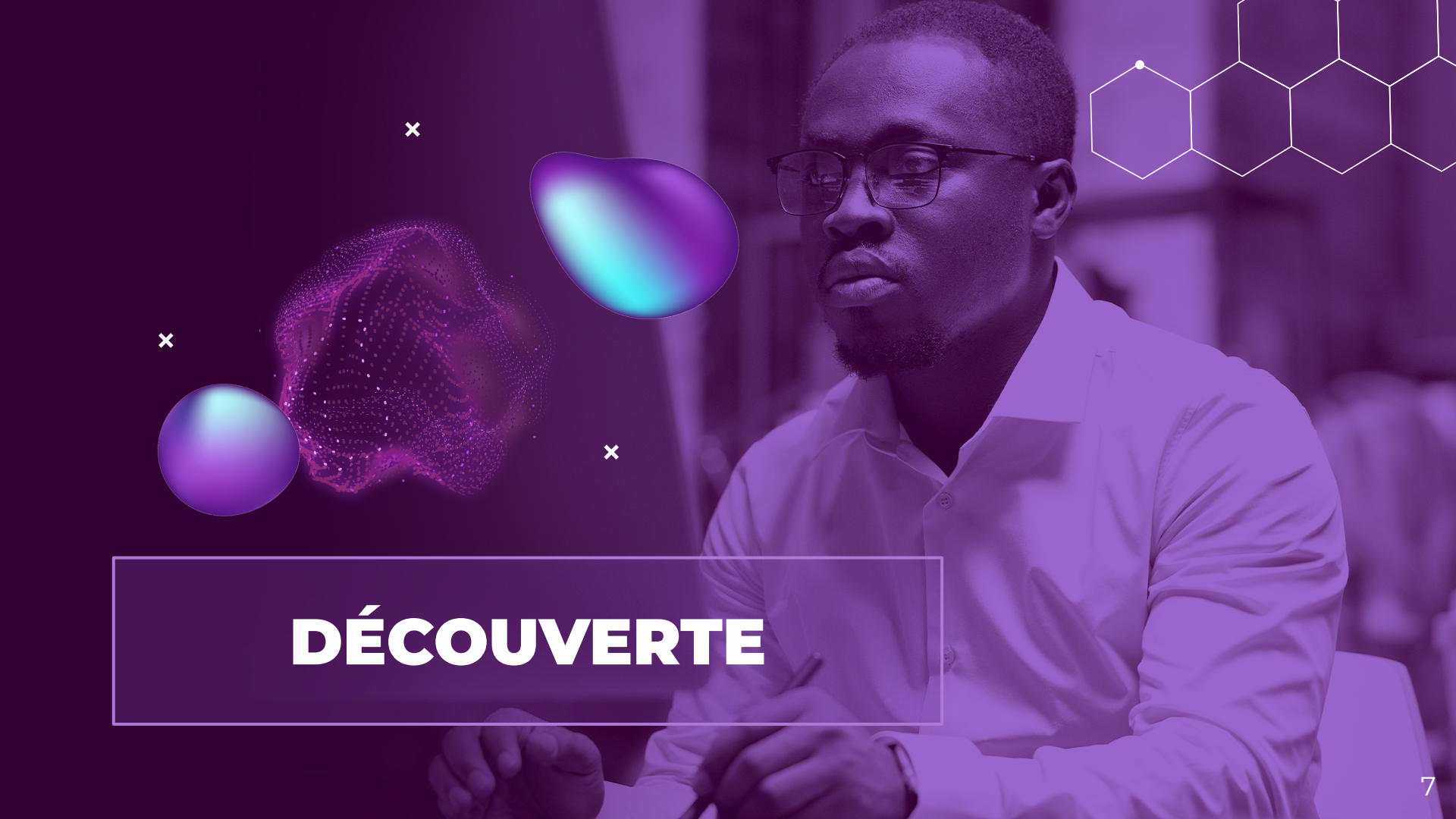
May 29, 2023

TASK	ASSIGNED TO	PROGRESS	START	END	DAYS
Phase 1 Découverte et préparation des données					
Préparation de l'environnement/ gestion de projet	Adrien	100%	5/30/23	5/30/23	1
Découverte des données	Vincent	100%	5/30/23	5/30/23	1
Préparation des données	Tristan	100%	5/30/23	5/30/23	1
Phase 2 Apprentissage non-supervisé					
Réduction de dimension	Adrien	100%	5/30/23	5/30/23	1
Partitionnement (clustering)	Tristan/Vincent	100%	5/30/23	5/30/23	1
Phase 3 Apprentissage supervisé					
Répartition des données	Vincent	100%	5/31/23	5/31/23	1
Classification avec KNN	Adrien	100%	5/31/23	5/31/23	1
Classification avec trois algorithmes de « haut niveau »	Vincent	100%	6/1/23	6/1/23	1
Phase 4 Métriques					
Évaluation quantitative des résultats « non supervisé »	Tristan	100%	5/31/23	5/31/23	1
Évaluation quantitative des résultats « supervisé »	Adrien/Vincent	75%	5/31/23	6/1/23	2
Phase 5 Création de scripts utilisables en ligne de commande					
Script pour k-means	Tristan	100%	6/2/23	6/2/23	1
Script pour KNN	Tristan	100%	6/2/23	6/2/23	1
Script méthodes classification « haut niveau »	Tristan	0%	6/2/23	6/2/23	1

02

LES DONNÉES

Découverte & Préparation

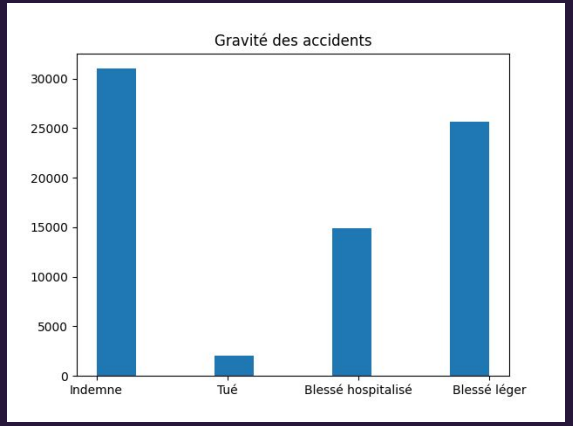


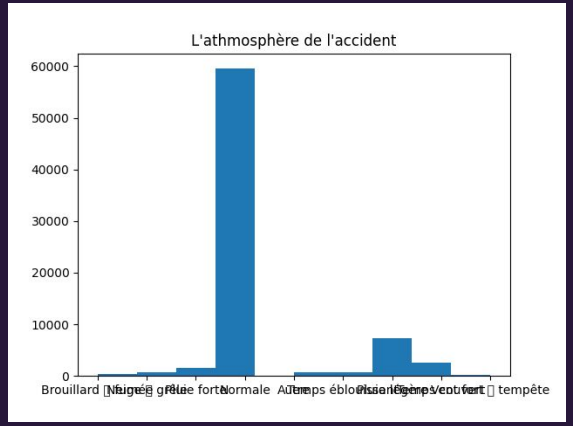
DÉCOUVERTE

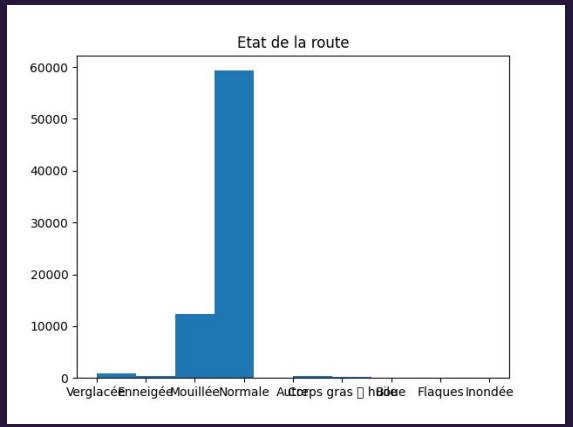
Caractéristiques de la base de donnée

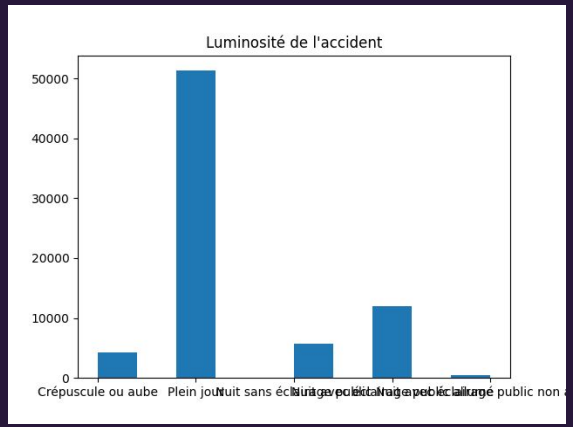
Valeur cible(descr_grav)	[1,2,3,4]
Nombre d’instance	21
Nombre d’instance par classe	73 643 -> 55 428
Taille des features	20

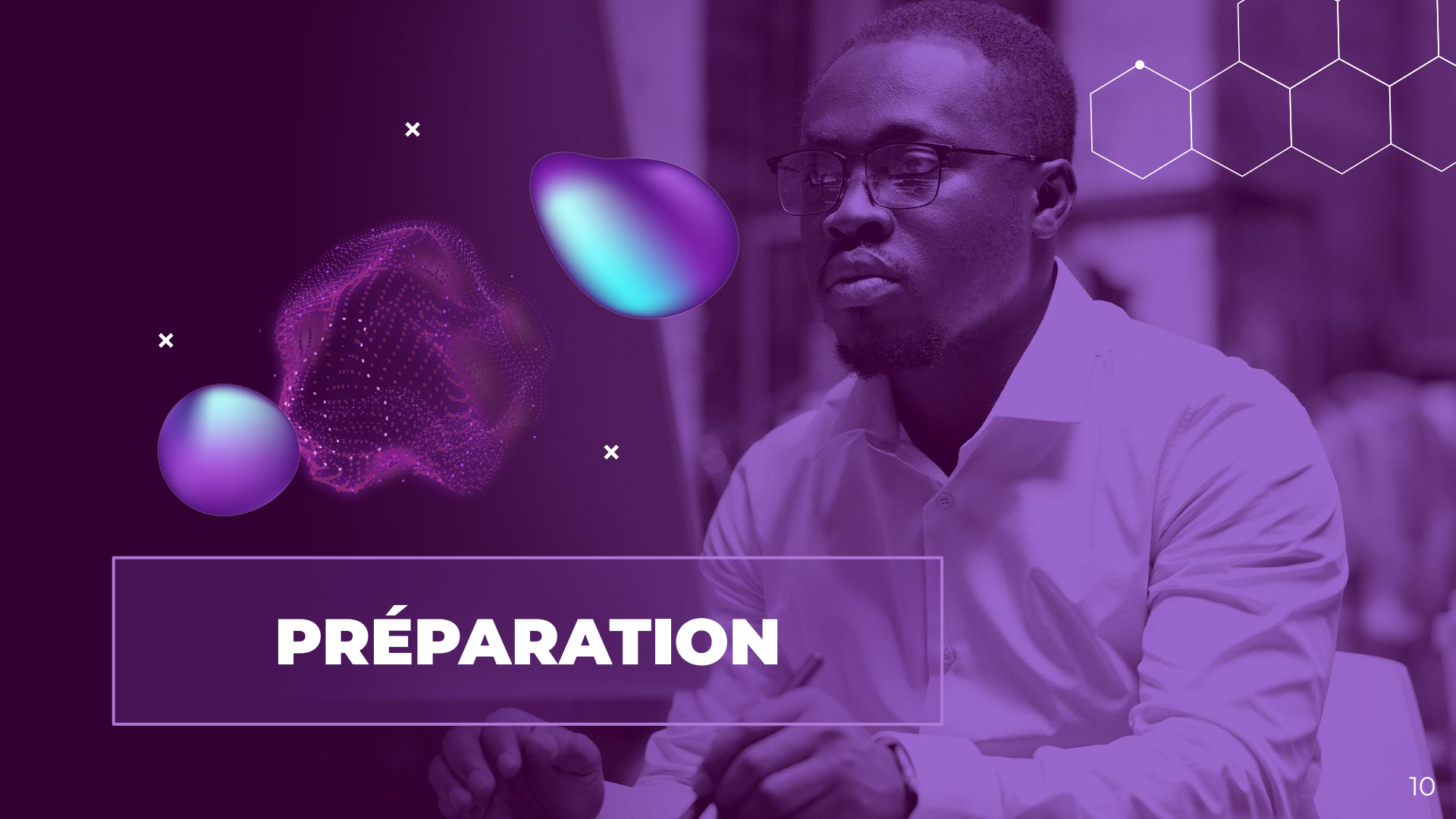
HISTOGRAMMES











PRÉPARATION

CONVERSIONS

	index	value
descr_cat_veh	0	PL seul > 7,5T
descr_cat_veh	1	VU seul 1,5T <= PTAC <= 3,5T avec ou sans remorq
descr_cat_veh	2	VL seul
descr_cat_veh	3	Autocar
descr_cat_veh	4	PL > 3,5T + remorque
descr_cat_veh	5	Cyclomoteur <50cm3
descr_cat_veh	6	Motocyclette > 125 cm3
descr_cat_veh	7	Tracteur routier + semi-remorque
descr_cat_veh	8	Tracteur agricole
descr_cat_veh	9	PL seul 3,5T <PTCA <= 7,5T
descr_cat_veh	10	Autobus
descr_cat_veh	11	Train
descr_cat_veh	12	Scooter > 125 cm3
descr_cat_veh	13	Scooter < 50 cm3
descr_cat_veh	14	Voiturette (Quadricycle à moteur carrossé) (ancier
descr_cat_veh	15	Autre véhicule
descr_cat_veh	16	Bicyclette
descr_cat_veh	17	Motocyclette > 50 cm3 et <= 125 cm3
descr_cat_veh	18	Scooter > 50 cm3 et <= 125 cm3
descr_cat_veh	19	Engin spécial
descr_cat_veh	20	Quad lourd > 50 cm3 (Quadricycle à moteur non ca
descr_cat_veh	21	Tramway
descr_cat_veh	22	Tracteur routier seul
descr_cat_veh	23	Quad léger <= 50 cm3 (Quadricycle à moteur non
descr_agglo	0	Hors agglomération
descr_agglo	1	En agglomération

Tableau de conversion non-numériques ⇔ numériques

Conversions à effectuer :

- Valeurs non-numériques en numériques
- Dates et Heures au bon format

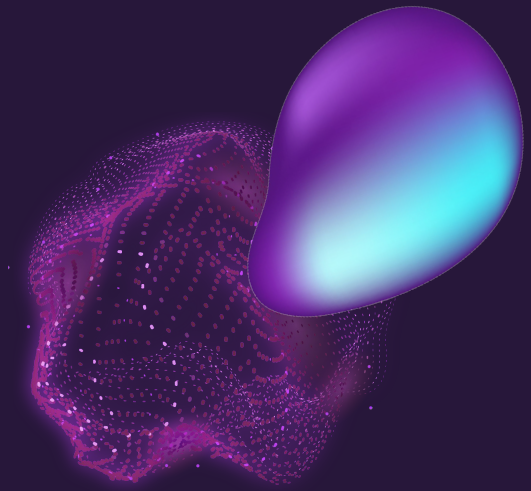
Problèmes rencontrés :

- Extraire le tableau sous format .xlsx
- Latitudes et longitudes des DOM-TOM

Libraries utilisées : pandas, datetime

03

APPRENTISSAGE non-supervisé



RÉDUCTION DE DIMENSION

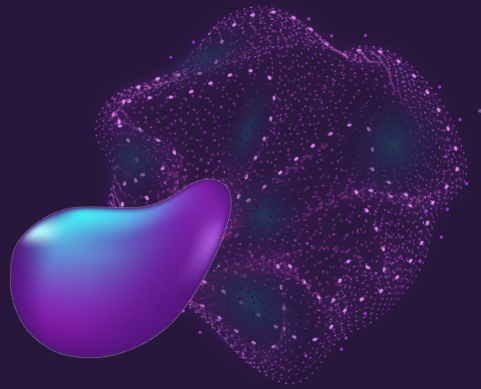


TABLEAU DE CORRÉLATION

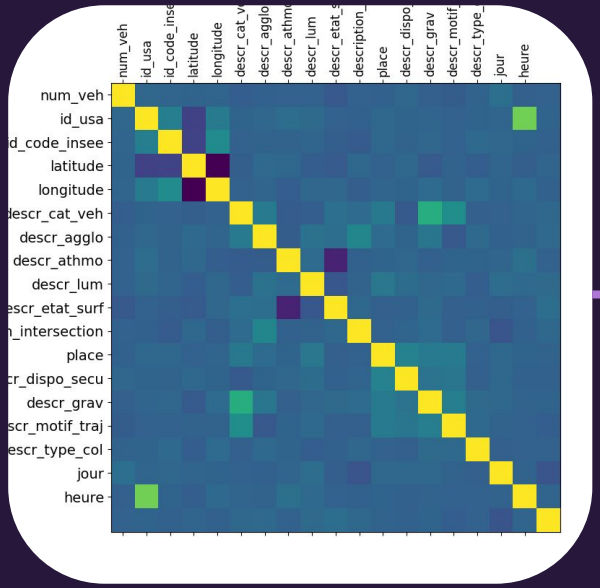
Réduction de la dimension en fonction:

- Coefficients de corrélation
- Liens entre les variables

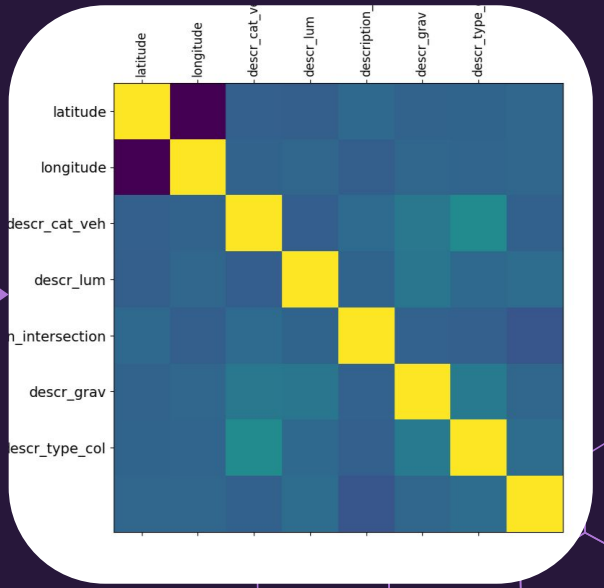
Coefficient de corrélation en fonction de la gravité des accidents

feature	Coef de corrélation	Supprimée ?
jour	0.01671	✗
id_usa	0.001948	✗
latitude	0.006072	
longitude	0.008392	
descr_athmo	0.012155	✗
id_code_insee	0.012865	✗
descr_etat_surf	0.015064	✗
descr_agglo	0.016364	✗
lescription_intersection	0.019234	
heure	0.025678	✗
descr_lum	0.030840	
num_veh	0.034247	✗
descr_motif_traj	0.054042	✗
descr_type_col	0.055053	
X (id)	0.061824	✗
place	0.110989	✗
an_nais	0.132073	
descr_dispo_secu	0.222666	✗
descr_cat_veh	0.239771	
descr_grav	1.000000	

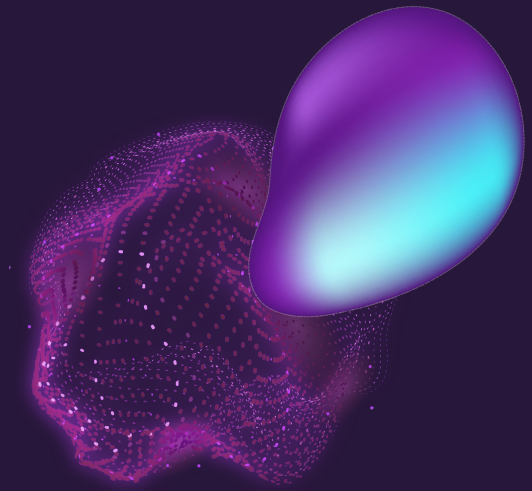
MATRICES DE CORRÉLATION



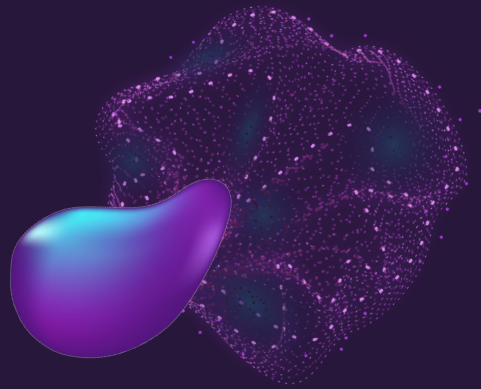
Avant réduction



Après réduction



CLUSTERING

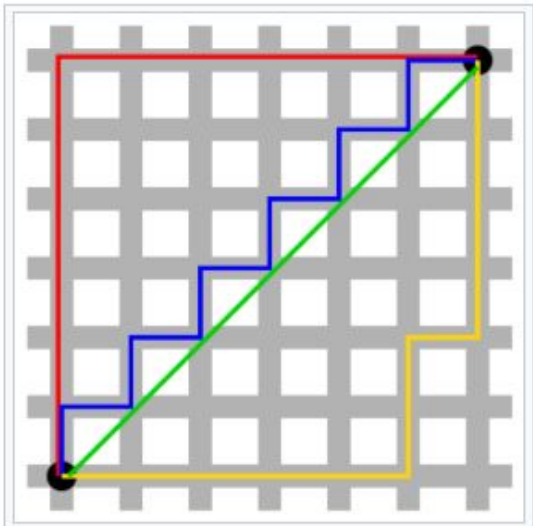


Clustering - “from scratch”

Différentes méthodes dans le calcul des distances:

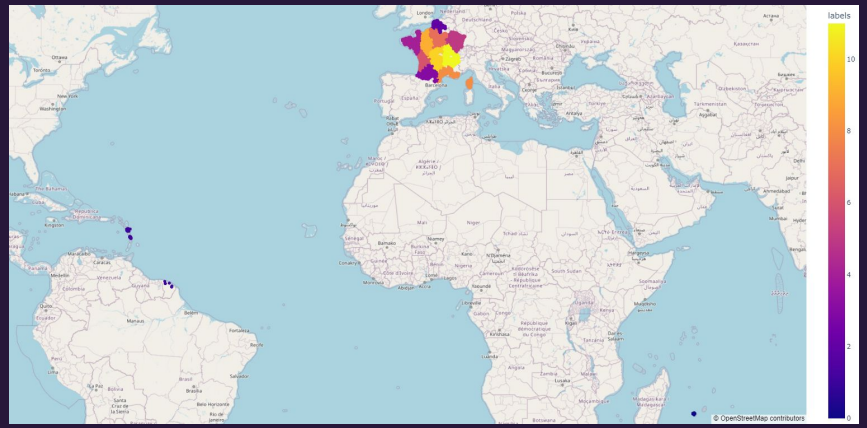
- L1(Méthode Manhattan)	$d(A,B) = \lvert X_B - X_A \rvert + \lvert Y_B - Y_A \rvert$
- L2(Méthode Euclidienne)	$d(A,B) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$
- Haversine	$= 2r \arcsin\left(\sqrt{\sin^2\left(\frac{\varphi_2 - \varphi_1}{2}\right) + \cos(\varphi_1)\cos(\varphi_2)\sin^2\left(\frac{\lambda_2 - \lambda_1}{2}\right)}\right)$

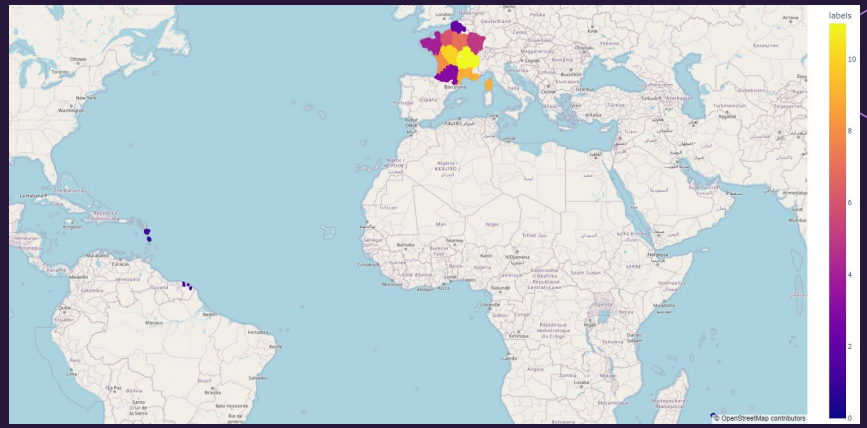
Contexte d’utilisation des calculs de distance



Distance de Manhattan (chemins rouge, jaune et bleu) contre distance euclidienne en vert.

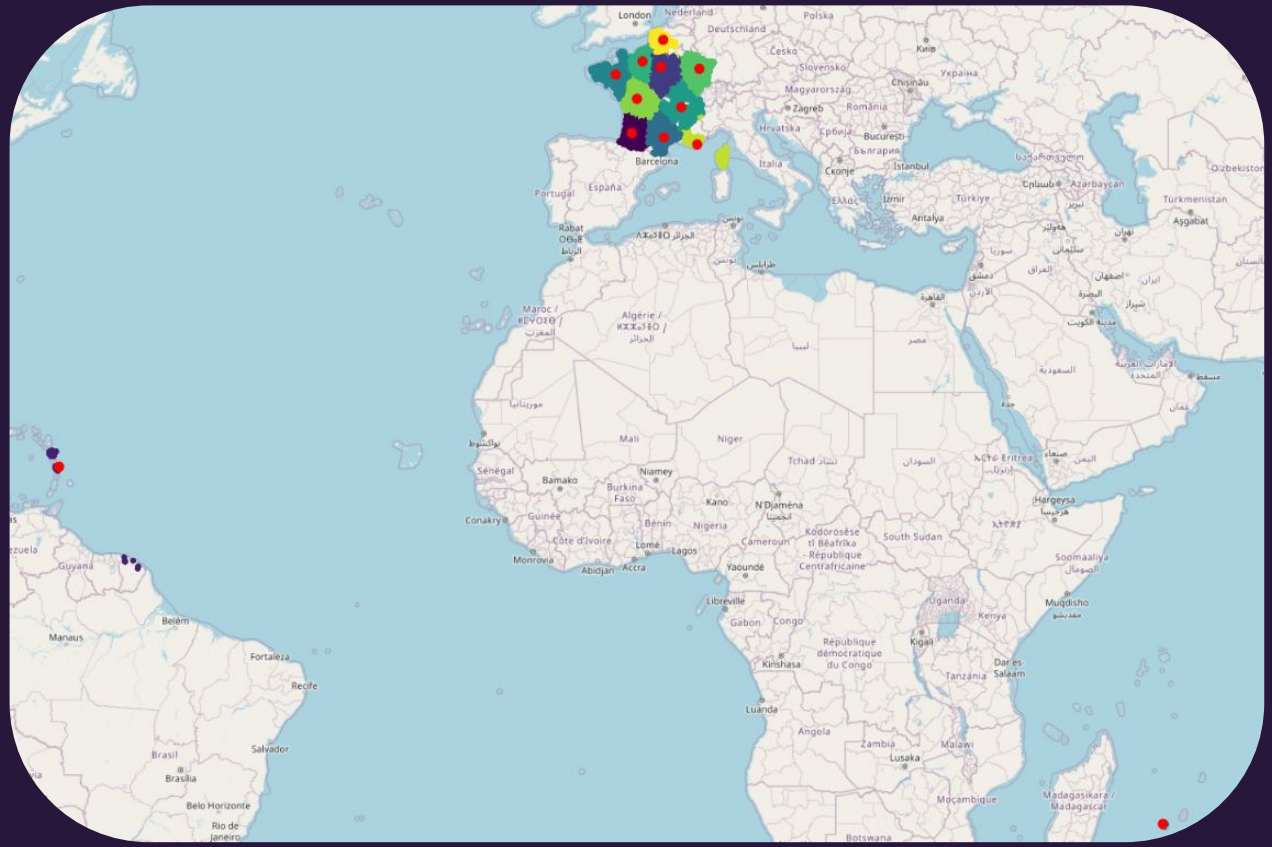
Résultats de cluster via scratch





-Etude avec 12 Clusters
-les clusters sont très similaires
-On utilise une projection mercator

Clustering - scikit-learn



Études avec 12 clusters

- Les points rouges représentent les centroïdes
- Étude effectué également avec 3, 5 et 8 clusters

Problèmes rencontrés :

- Affichage des centroïdes

Libraries utilisées : sklearn, plotly, numpy

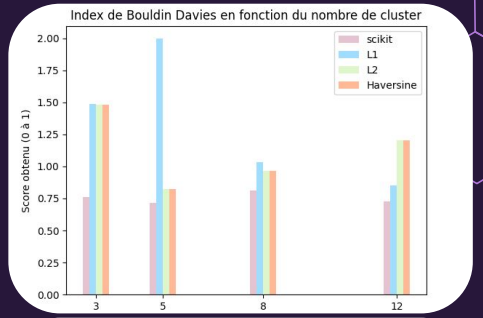
Clustering - Evaluation quantitative

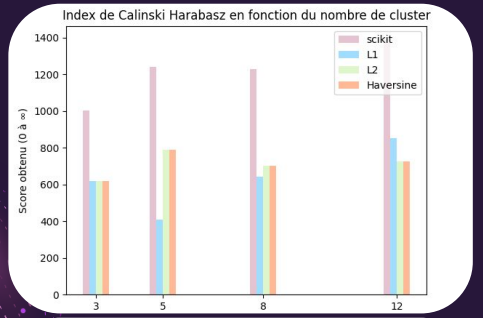
Nombre de cluster				
3				
Méthode	scikit	L1	L2	Haversine
Metric	Score	Score	Score	Score
Coefficient de silhouette	0.4396002394042591	0.3099044630729121	0.3066844697028553	0.3066844697028553
Index de Calinski Harabasz	1002.1034962689257	619.8313597309689	617.8453619262061	617.8453619262061
Index de Bouldin Davies	0.759965228551648	1.4849925933563453	1.4822945609067348	1.4822945609067348

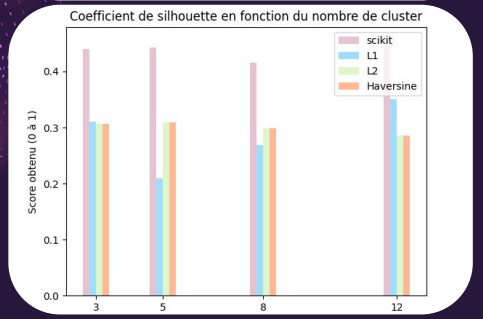
Nombre de cluster				
5				
Méthode	scikit	L1	L2	Haversine
Metric	Score	Score	Score	Score
Coefficient de silhouette	0.44375880274301455	0.20855898682664645	0.3088980353693082	0.3088980353693082
Index de Calinski Harabasz	1242.0986175669645	407.9566647218841	790.2322194733962	790.2322194733962
Index de Bouldin Davies	0.7179728279735788	1.9968265749287402	0.8234573037440249	0.8234573037440249

Nombre de cluster				
8				
Méthode	scikit	L1	L2	Haversine
Metric	Score	Score	Score	Score
Coefficient de silhouette	0.41307080643298366	0.26924551509454425	0.29935651624751863	0.29935651624751863
Index de Calinski Harabasz	1230.1642992694271	642.0362689580388	700.0842140392815	700.0842140392815
Index de Bouldin Davies	0.8160015678900034	1.033938304118349	0.9646450498949747	0.9646450498949747

Nombre de cluster				
12				
Méthode	scikit	L1	L2	Haversine
Metric	Score	Score	Score	Score
Coefficient de silhouette	0.4622257479910289	0.35067055901472927	0.28521006951152567	0.28521006951152567
Index de Calinski Harabasz	1366.3215155872706	851.2028692835758	724.7676272093031	724.7676272093031
Index de Bouldin Davies	0.7528093020232839	0.8516612387507877	1.2022495277774983	1.2022495277774983

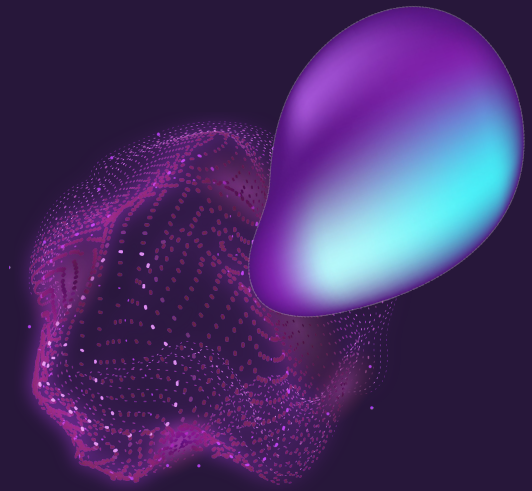




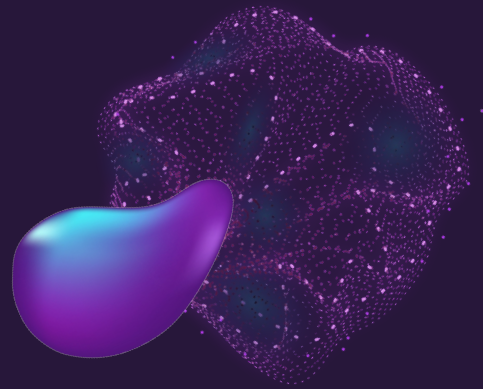


04

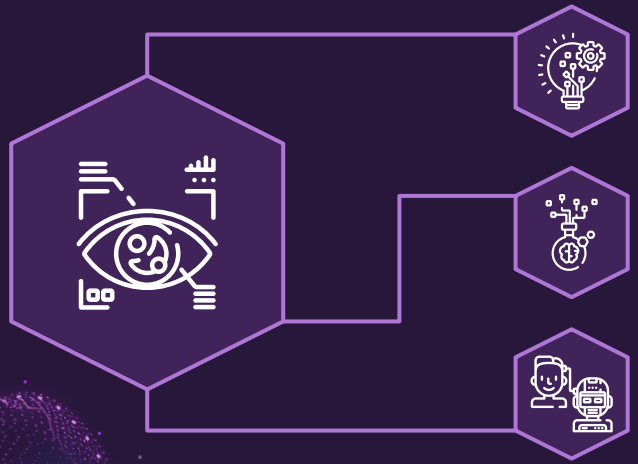
APPRENTISSAGE supervisé



CLASSIFICATION kNN



ETAPES



“ From Scratch”

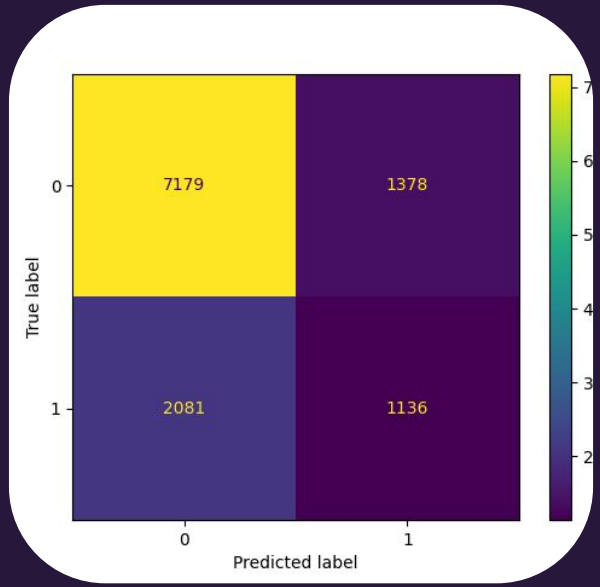
Avec scikit-learn

Métriques

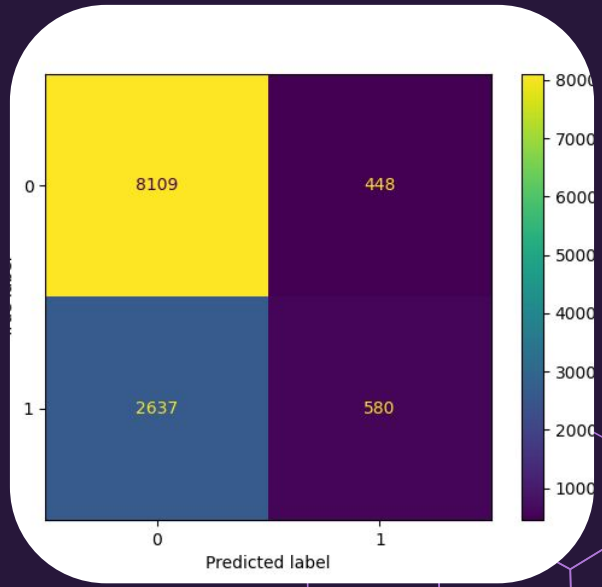
EVALUATION QUANTITATIVE

	K_neighbors = 3	K_neighbors = 48
Taux d'apprentissage	0.7062170885000849	0.7379819942245626
Précision	0.6135697545151042	0.6594043498947395
Rappel	0.5960431408620842	0.563968700228848

MATRICES DE CONFUSION

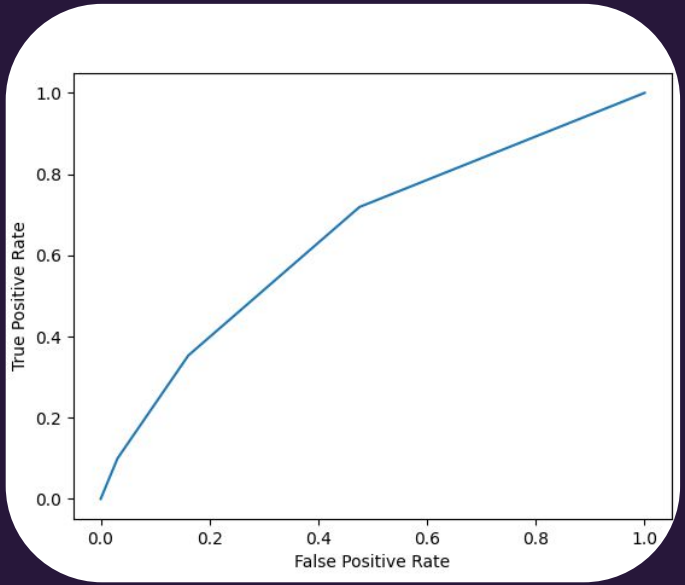


K = 3

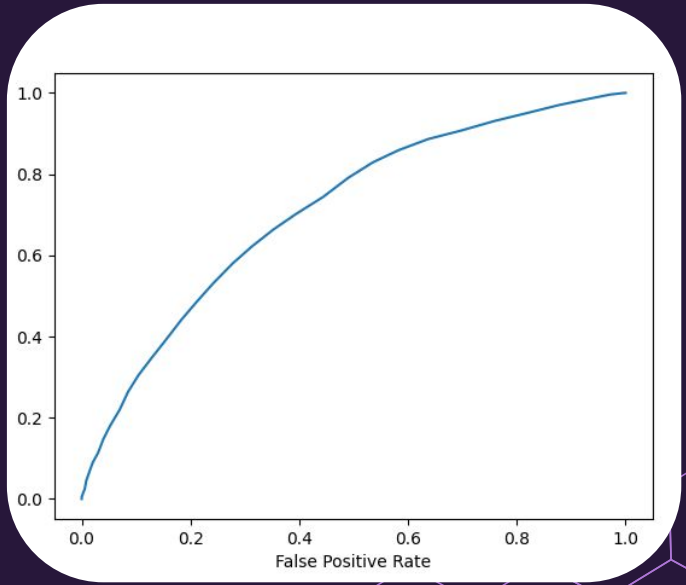


K = 48

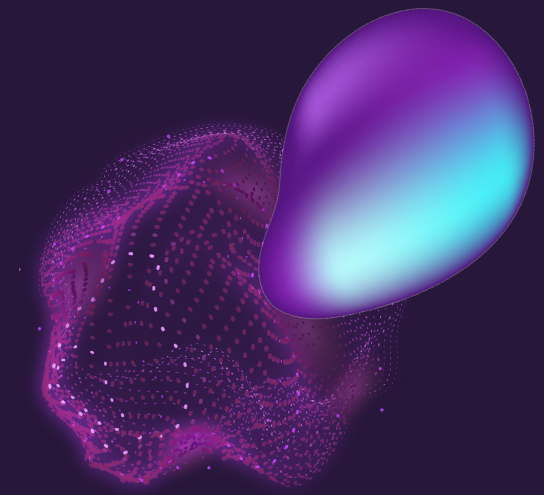
COURBES ROC



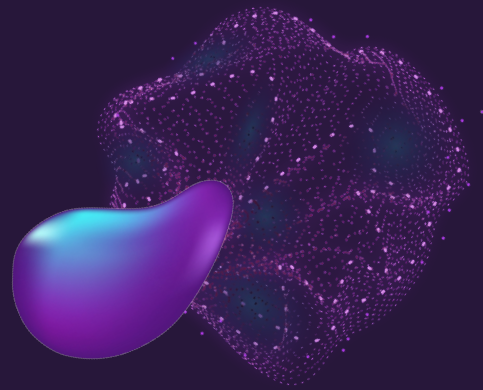
K = 3



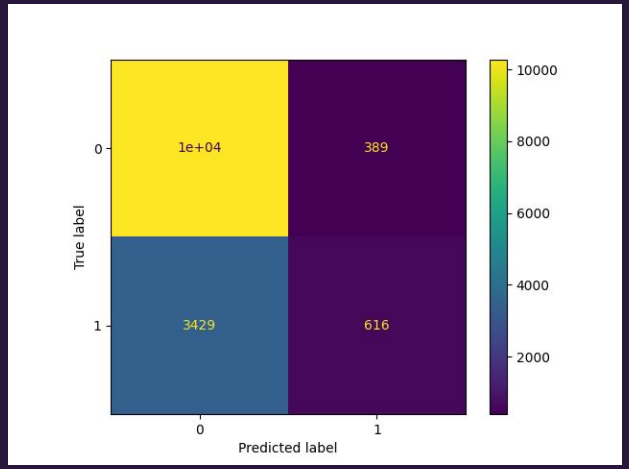
K = 48

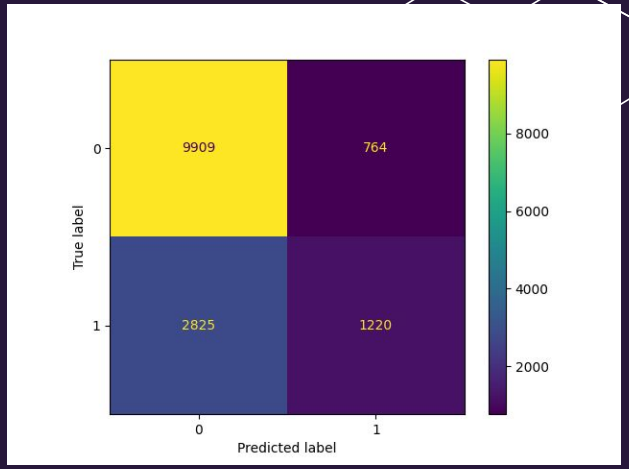


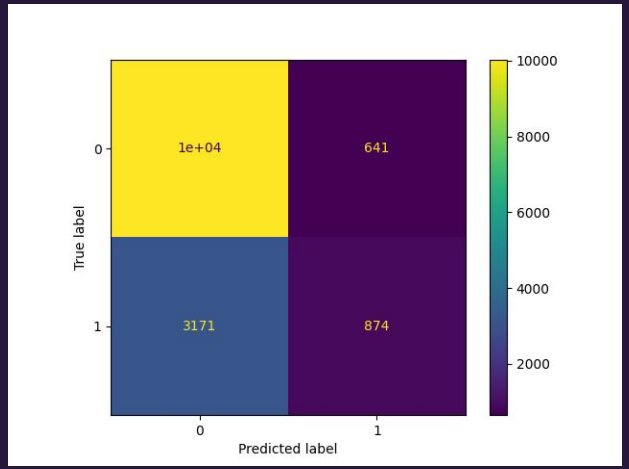
ALGORITHMES
“ HAUT NIVEAU ”

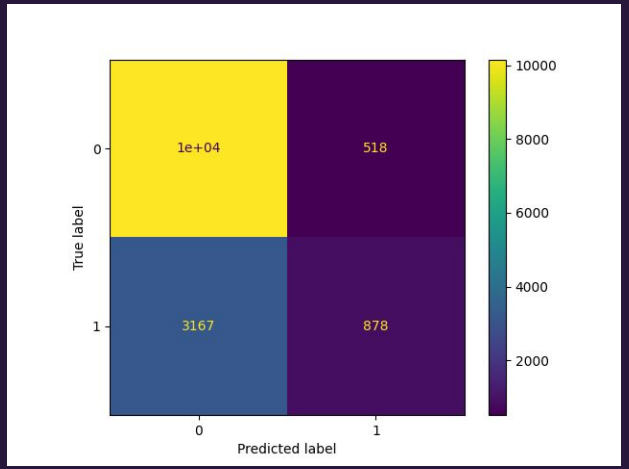


Matrices de confusion









Hold out avec GridSearch	Accuracy	Précision	F1-Score
Support Vector Machine	0.740	0.712	0.678
Random Forest	0.756	0.733	0.725
Multi layer perceptron	0.740	0.709	0.695
Vote majorité	0.7496	0.725	0.702

Tableau comparatifs des différentes classifications Hold out

Tableau comparatifs des différentes classifications leave one out

	Accuracy	Precision	Rappel	F1-Score
SVM	0.541	1.0	0.541	0.541
Random Forest	0626	1.0	0.626	0.626
MLP	0.5385	1.0	0.5385	0.5385

05

LES SCRIPTS

Scripts - utilisation

Lancer un script

.\scripts.sh -m kmean [latitude] [longitude] [centroïdes]

.\scripts.sh -m knn [info_accident] [nom_du_csv]

.\scripts.sh -m classification [info_accident] [méthode]

- Script shell avec utilisation du flag -m pour le choix du mode
- Intégration à du contenu web grâce à un export .json

Problèmes rencontrés :

- Étude du langage de programmation bash
- Transfert des arguments en python
- Renvoie d'un fichier json au bon format

Libraries utilisées : pandas, numpy, sys

CONCLUSION

Avez-vous des questions ?