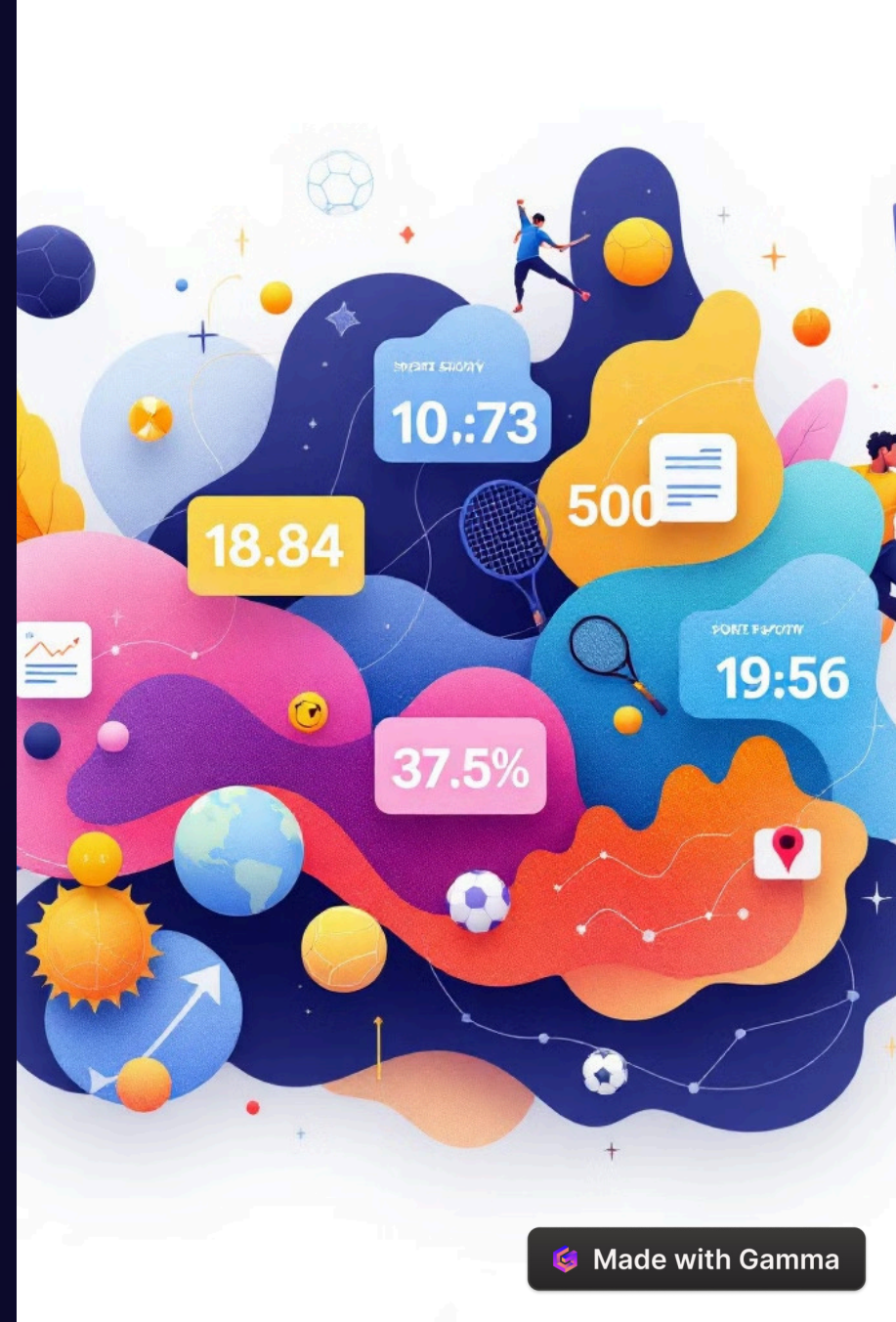


Analyse des Disciplines Sportives par Clustering et Classification

Objectif : Analyser les données des disciplines sportives via clustering et classification.

Méthodologie : Utilisation de DBSCAN, KMeans, SVM, Random Forest et KNN.

Sriwelavan Theeban, Meddas Kilian



Présentation des Données

Description

Données de performance des plongeurs dans différentes disciplines.

csv.head(10)
✓ 0.0s Python

	Start	Diver	Nationality	Gender	Discipline	Line	Official Top	AP	RP	Card	Points	Remarks	Title Event	Event Type	Day
0	1	Deborah Andollo	CUB	F	CWT	NaN	00:00	61 m	61 m	WHITE	61.0	OK	WR Attempt - ANDOLLO Deborah (CWT)	Worldrecord attempt	1994-06-12
1	1	Umberto Pelizzari	ITA	M	CWT	NaN	00:00	72 m	72 m	WHITE	72.0	OK	WR Attempt - PELIZZARI Umberto (CWT)	Worldrecord attempt	1995-09-17
2	1	Deborah Andollo	CUB	F	CWT	NaN	00:00	62 m	62 m	WHITE	62.0	OK	WR Attempt - ANDOLLO Deborah	Worldrecord attempt	1996-10-05
3	1	Michael Oliva	FRA	M	CWT	NaN	00:00	72 m	72 m	WHITE	72.0	OK	WR Attempt - OLIVA Michael (CWT)	Worldrecord attempt	1996-10-11
4	1	Alejandro Ravelo	CUB	M	CWT	NaN	00:00	73 m	73 m	WHITE	73.0	OK	WR Attempt - RAVELO Alejandro (CWT)	Worldrecord attempt	1997-08-02
5	1	Umberto Pelizzari	ITA	M	CWT	NaN	00:00	75 m	75 m	WHITE	75.0	OK	WR Attempt - PELIZZARI Umberto (CWT)	Worldrecord attempt	1997-09-13
6	1	Deborah Andollo	CUB	F	CWT	NaN	00:00	0 m	65 m	WHITE	65.0	OK	WR Attempt - ANDOLLO Deborah	Worldrecord attempt	1997-12-05
7	1	Alexandra Louzine	CZE	F	CWT	NaN	00:00	35 m	35 m	WHITE	35.0	OK	1998 WRA CWT Fresh Water by Alexandra Louzine	Worldrecord attempt	1998-09-06
8	1	Tanya Streeter	USA	F	CWT	NaN	00:00	67 m	67 m	WHITE	67.0	OK	WR Attempt - STREETER Tanya (CWT)	Worldrecord attempt	1998-09-19
9	3	Andy Le Sauce	FRA	M	CWT	NaN	00:00	0 m	56 m	WHITE	56.0	OK	Compiled rankings for year 1999	Competition	1999-01-01

	Name	Nationality	Gender	Discipline	AP	RP	Card	Points	Title Event	Event Type	Day
0	Deborah Andollo	CUB	F	CWT	61 m	61 m	WHITE	61.0	WR Attempt - ANDOLLO Deborah (CWT)	Worldrecord attempt	1994-06-12
1	Umberto Pelizzari	ITA	M	CWT	72 m	72 m	WHITE	72.0	WR Attempt - PELIZZARI Umberto (CWT)	Worldrecord attempt	1995-09-17
2	Deborah Andollo	CUB	F	CWT	62 m	62 m	WHITE	62.0	WR Attempt - ANDOLLO Deborah	Worldrecord attempt	1996-10-05
3	Michael Oliva	FRA	M	CWT	72 m	72 m	WHITE	72.0	WR Attempt - OLIVA Michael (CWT)	Worldrecord attempt	1996-10-11
4	Alejandro Ravelo	CUB	M	CWT	73 m	73 m	WHITE	73.0	WR Attempt - RAVELO Alejandro (CWT)	Worldrecord attempt	1997-08-02
5	Umberto Pelizzari	ITA	M	CWT	75 m	75 m	WHITE	75.0	WR Attempt - PELIZZARI Umberto (CWT)	Worldrecord attempt	1997-09-13
6	Deborah Andollo	CUB	F	CWT	0 m	65 m	WHITE	65.0	WR Attempt - ANDOLLO Deborah	Worldrecord attempt	1997-12-05
7	Alexandra Louzine	CZE	F	CWT	35 m	35 m	WHITE	35.0	1998 WRA CWT Fresh Water by Alexandra Louzine	Worldrecord attempt	1998-09-06
8	Tanya Streeter	USA	F	CWT	67 m	67 m	WHITE	67.0	WR Attempt - STREETER Tanya (CWT)	Worldrecord attempt	1998-09-19
9	Andy Le Sauce	FRA	M	CWT	0 m	56 m	WHITE	56.0	Compiled rankings for year 1999	Competition	1999-01-01

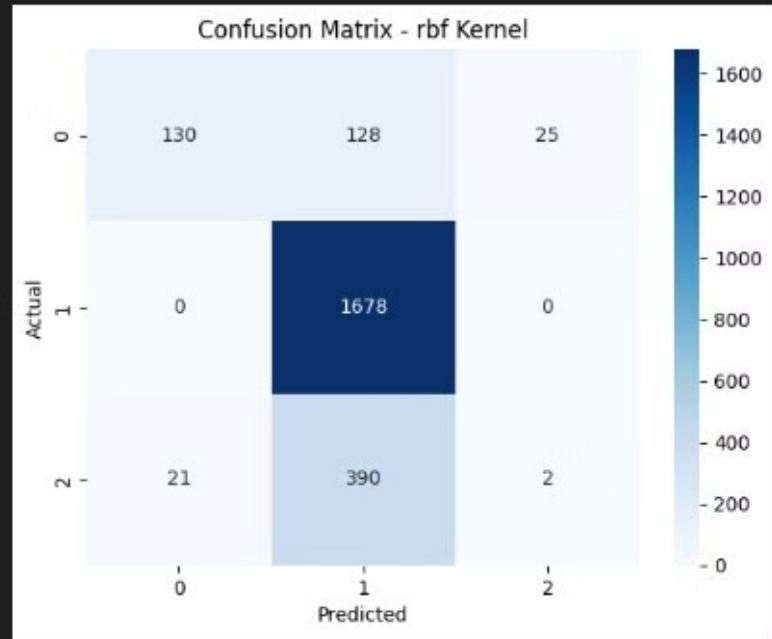
(dataset originel et modifié)

Nettoyage


Suppression des doublons, renommage et changement de type.


- `csv.drop(['Start', 'Line', 'Official Top'], axis=1, inplace=True)`
- `csv['AP']=csv['AP'].str.extract(r'(\d+)').astype(float)`
- `csv['RP']=csv['RP'].str.extract(r'(\d+)').astype(float)`
- `csv.rename(columns={'Diver': 'Name'}, inplace=True)`
- `csv.AP = csv.AP.astype(float)`
- `csv.RP = csv.RP.astype(float)`
- `csv.Day = lien inconnu_datetime(csv.Day)`

	precision	recall	f1-score	support
RED	0.86	0.46	0.60	283
WHITE	0.76	1.00	0.87	1678
YELLOW	0.07	0.00	0.01	413
accuracy			0.76	2374
macro avg	0.57	0.49	0.49	2374
weighted avg	0.66	0.76	0.69	2374



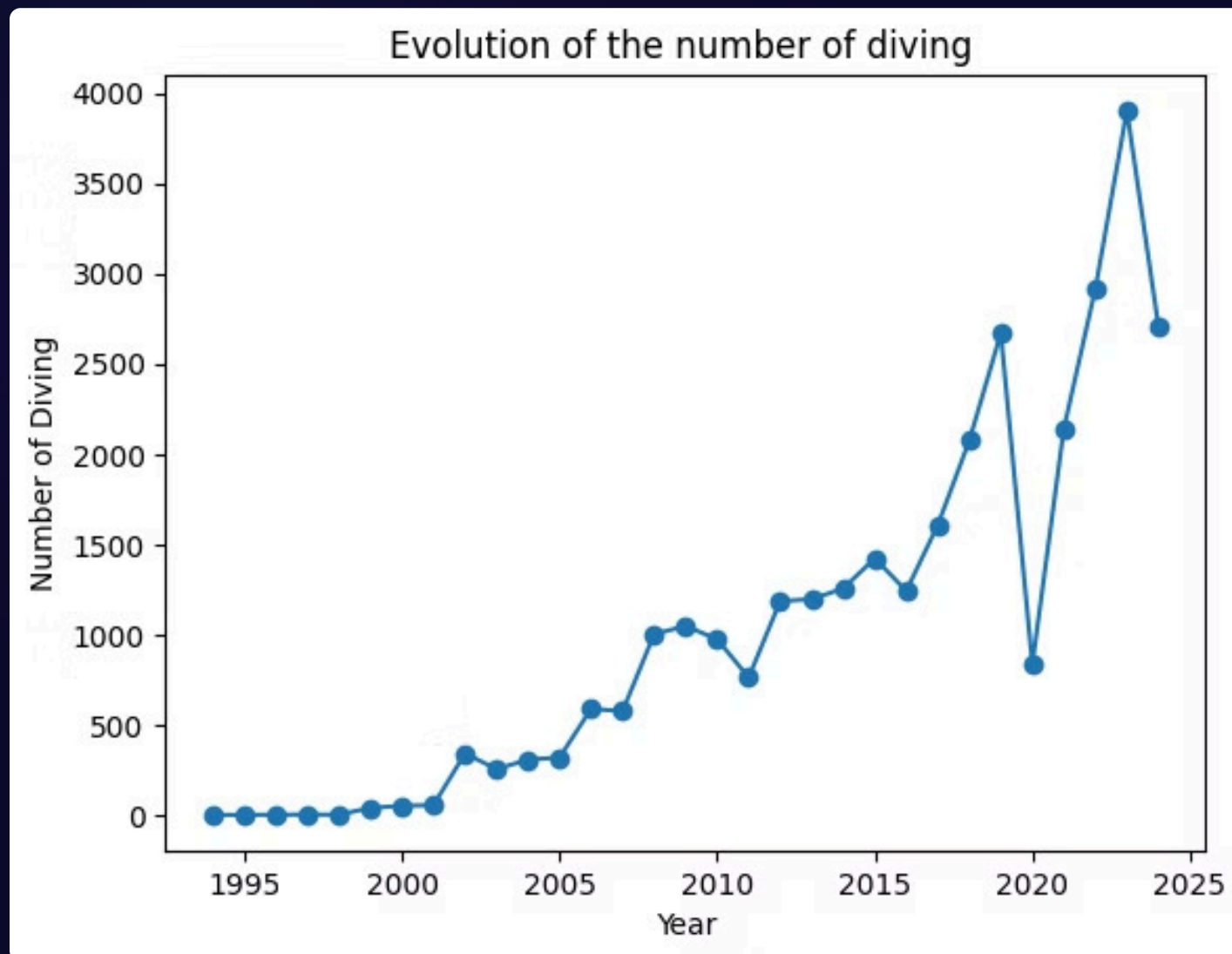
Matrice de Corrélation

 Relation Significative
Forte corrélation entre
experience_dive et Points.

 Faible Corrélation
Peu de lien entre Gender_codes
et Points.

Évolution du Nombre de Plongées au Fil des Années

Ce graphique illustre la progression du nombre de plongées enregistrées de 1995 à 2025. Il met en évidence une croissance exponentielle, avec une forte augmentation après 2010 et des variations notables autour de 2020 possiblement liées à des événements comme le COVID-19.

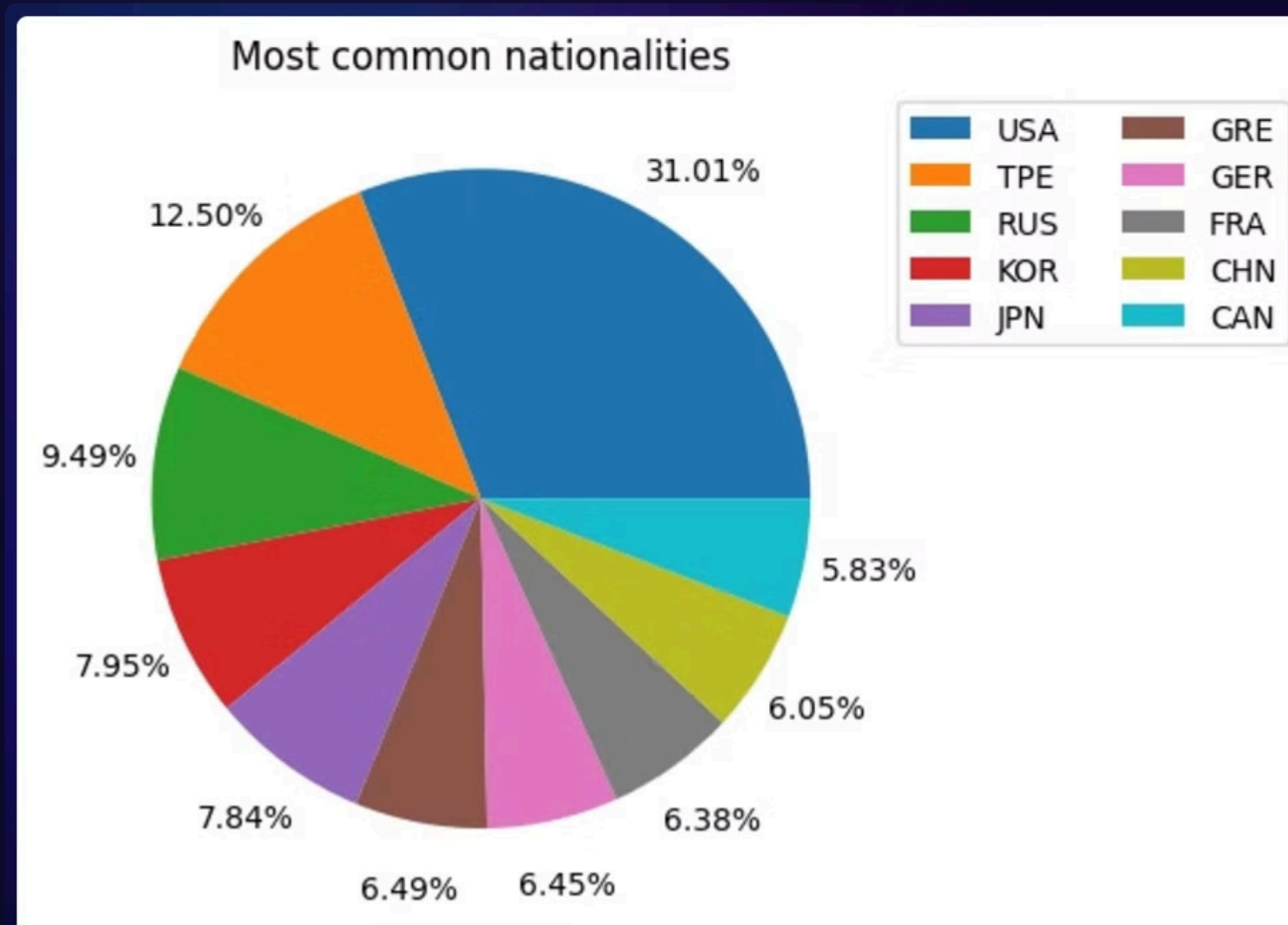


Cette tendance justifie l'analyse approfondie des performances en plongée sur une période significative pour identifier les facteurs de succès et les défis rencontrés.

Analyse des nationalités les plus représentées

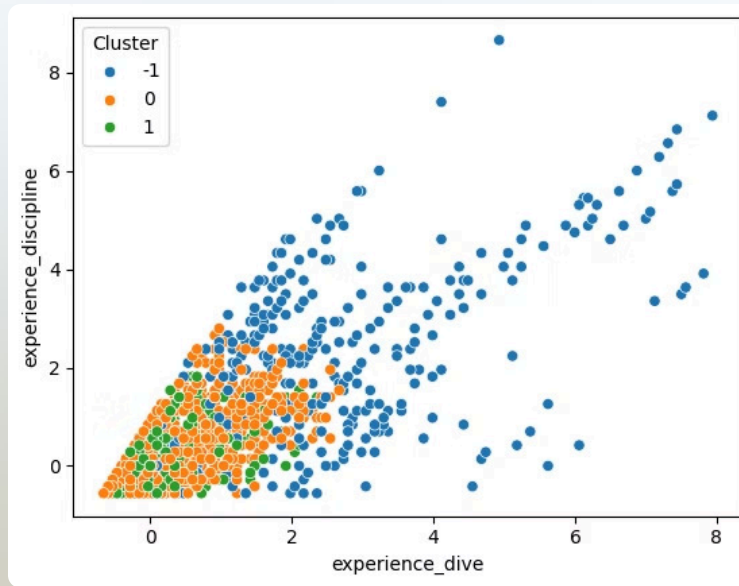
Ce diagramme circulaire illustre la répartition des nationalités les plus fréquentes parmi les participants. Les États-Unis (USA) dominent avec 31.01%, suivis de Taïwan (TPE) avec 12.50%. D'autres nationalités significatives incluent la Russie (RUS) avec 9.49%, la Corée (KOR) avec 7.95%, et le Japon (JPN) avec 7.84%. Les autres nationalités se partagent des proportions inférieures, mais restent représentées, montrant une certaine diversité.

Cette distribution pourrait refléter l'intérêt culturel et les infrastructures disponibles pour ce sport dans ces pays.



Données sélectionnées pour le clustering et la classification

	AP	Nationality	Gender	Discipline	experience_dive	experience_discipline	Points	target
0	42.0	GER	M	FIM	1.0	1.0	42.0	WHITE
1	32.0	GER	M	CWT	2.0	1.0	32.0	WHITE
2	43.0	GER	M	CWT	3.0	2.0	43.0	WHITE
3	34.0	CAN	M	CWT	1.0	1.0	34.0	WHITE
4	32.0	CAN	M	CWT	2.0	2.0	32.0	WHITE
...
31712	60.0	NaN	F	FIM	NaN	NaN	60.0	WHITE
31713	49.0	NaN	F	CWT	NaN	NaN	0.0	RED
31714	61.0	NaN	F	CNF	NaN	NaN	61.0	WHITE
31715	63.0	NaN	F	CNF	NaN	NaN	63.0	WHITE
31716	81.0	NaN	F	CWT	NaN	NaN	81.0	WHITE



Clustering avec DBSCAN



Groupes Principaux

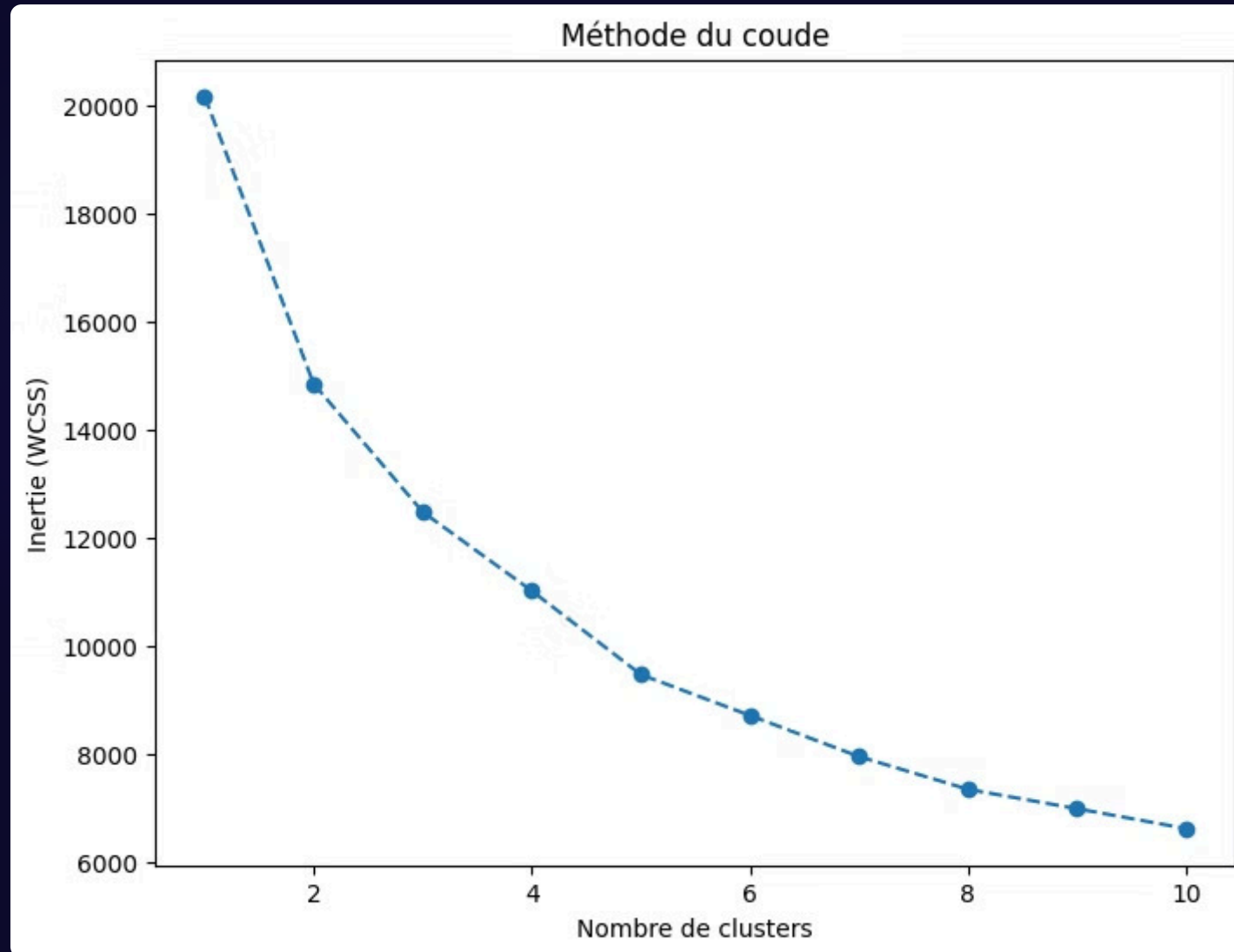
Identification de 3 groupes principaux (Cluster -1, 0, 1). Le cluster -1 représentant le bruit (pas suffisamment de voisin d'après l'algo)



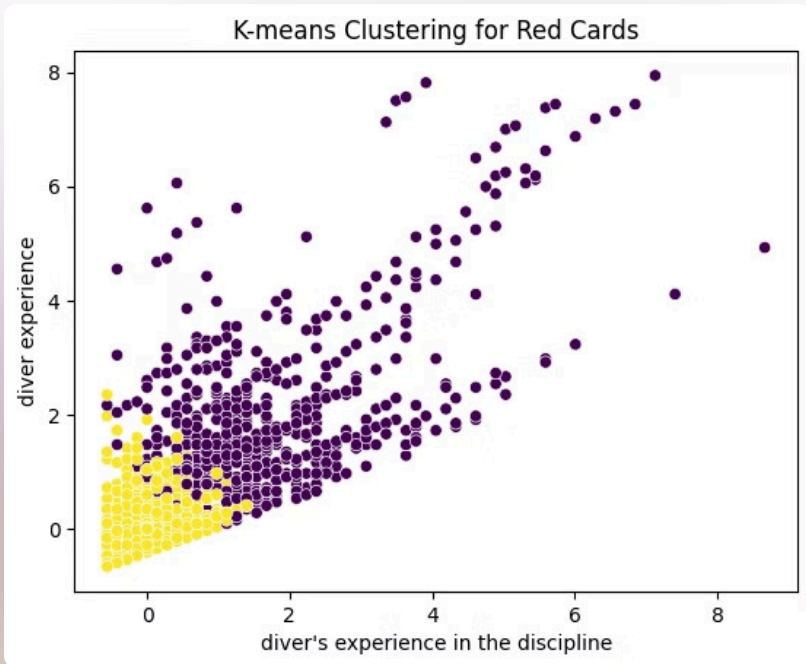
Limite

Sensibilité aux paramètres `eps` et `min_samples`.

Méthode du Coude



Exemple de graphique illustrant la méthode du coude pour le KMeans (point d'inflexion choisi comme nombre optimal de clusters).



KMeans Clustering

Groupes Définis

Clusters bien définis selon
experience_dive et
experience_discipline.

Robustesse

Plus stable que DBSCAN face
aux variations.

Défaut

Moins facile à comprendre.

Classification - Résultats SVM

Kernel: linear, Accuracy: 0.9518212621770437
Kernel: poly, Accuracy: 0.9307496823379924
Kernel: rbf, Accuracy: 0.9561626429479034

95.18%

Précision Kernel Linear

Performance supérieure du kernel linéaire.

3

Kernels Comparés

Analyse des kernels linear, poly, et rbf.

Classification - Comparaison des Modèles

1

Random Forest

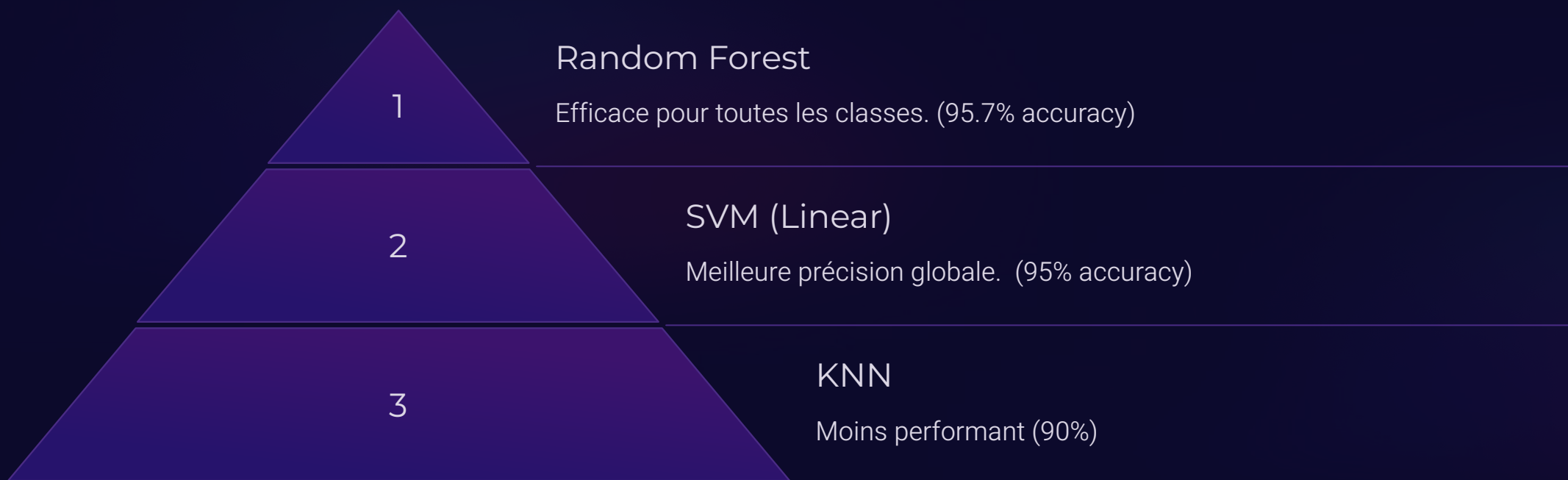
Précision de 95.15%, bon équilibre entre classes.

2

KNN

Précision de 81.92%, difficulté avec les classes déséquilibrées.

Analyse Comparative des Modèles



Conclusion et Perspectives

1

Clustering

DBSCAN pour exploration, KMeans pour segmentation robuste.

2

Classification

SVM avec kernel linear est le plus performant.

3

Recommandations

Améliorer la représentation des classes minoritaires.