



Université Paul Valéry Montpellier III

Département Mathématiques, Informatique Appliquées

- **Classification non supervisée :**

**Analyse des thèmes de reportages les plus représentés sur différentes chaînes,
durant la période de septembre 2018 à septembre 2020.**

Réalisé par Anthony COMBES-AGUÉRA, numéro étudiant : 22113542

- **Classification supervisée :**

**Prédiction du nombre de buts par match joué pour l'année 2017,
à partir des données de l'année 2016.**

Réalisé par Wassim HARRAGA, numéro étudiant : 22203314

Présenté à M. Arnaud Sallaberry

SOMMAIRE

1. APPRENTISSAGE NON SUPERVISÉ.....	2
1.1 - Jeu de données.....	2
1.2 - Nettoyage des données.....	4
1.3 - Clustering.....	6
2. APPRENTISSAGE SUPERVISÉ.....	8
2.1 - Jeu de données.....	8
2.2 - Nettoyage et prétraitement des données.....	10
2.3 - Modèles d'apprentissage.....	11
2.4 - Prédictions.....	14
WEBOGRAPHIE.....	15

1. APPRENTISSAGE NON SUPERVISÉ

1.1 - Jeu de données

Analyse des thèmes de reportages les plus représentés sur différentes chaînes, durant la période de septembre 2018 à septembre 2020 :

Pour obtenir notre jeu de données, nous avons exploré la plateforme de diffusion de données publiques française, data.gouv.fr (voir Figure 1). Nous avons identifié un ensemble de données qui correspondait à nos intérêts, celui-ci portant sur les thèmes de reportages de chaînes de télévision.



Figure 1 : Capture d'écran du site data.gouv.fr où se trouve notre jeu de données.

	A	B	C	D	E	F	G	H	I
1	MOIS	THEMATIQUES	Nombre de sujets de JT TF1	Nombre de sujets de JT France 2	Nombre de sujets de JT France 3	Nombre de sujets de JT Canal +	Nombre de sujets de JT Arte	Nombre de sujets de JT M6	Nombre de sujets de JT Totaux
2									
3	janvier-05	Catastrophes	214	191	88	18	40	49	600
4	janvier-05	Culture-loisirs	27	42	35	4	0	23	131
5	janvier-05	Economie	35	18	10	1	8	11	83
6	janvier-05	Education	14	12	4	3	3	8	44
7	janvier-05	Environnement	31	25	15	1	3	10	85
8	janvier-05	Faits divers	24	19	9	1	1	14	68
9	janvier-05	Histoire-hommages	36	38	24	10	23	19	150
10	janvier-05	International	52	61	54	15	108	26	316
11	janvier-05	Justice	20	27	16	11	11	7	92
12	janvier-05	Politique France	27	25	25	6	22	8	113
13	janvier-05	Santé	21	19	15	0	2	12	69
14	janvier-05	Sciences et techniques	28	11	11	4	2	7	63
15	janvier-05	Société	139	129	65	24	31	53	441
16	janvier-05	Sport	37	49	19	7	0	32	144
17	février-05	Catastrophes	42	30	13	4	14	14	117
18	février-05	Culture-loisirs	41	66	23	12	5	29	176
19	février-05	Economie	46	43	24	2	7	11	133
20	février-05	Education	27	31	22	7	7	10	104
21	février-05	Environnement	46	25	12	0	6	7	96
22	février-05	Faits divers	17	16	13	8	0	10	64
23	février-05	Histoire-hommages	14	12	12	1	9	4	52
24	février-05	International	100	110	81	26	160	32	509
25	février-05	Justice	56	52	43	15	8	26	200
26	février-05	Politique France	23	18	13	2	4	6	66
27	février-05	Santé	50	41	20	9	6	16	142
28	février-05	Sciences et techniques	17	15	10	0	4	4	50
29	février-05	Société	177	132	67	29	27	70	502
30	février-05	Sport	54	58	29	18	3	38	200

Figure 2 : Capture d'écran du tableur de notre jeu de données sans aucune modification.

Notre jeu de données d'origine contient 14 thématiques, 2647 instances pour 2 variables qualitatives et 7 variables quantitatives. Les observations correspondent ici à des thèmes de reportages sur des chaînes de télévision françaises de janvier 2005 à septembre 2020.

Les variables qualitatives incluent le mois ("Mois") et les thèmes ("Thématiques").

Les variables quantitatives comprennent le nombre total de sujets de JT pour chaque chaîne (TF1, France 2, France 3, Canal +, Arte et M6), ainsi que le total du nombre de sujets, toutes chaînes confondues ("Totaux").

Nous sommes alors confrontés à un problème, car le logiciel Orange ne détecte pas correctement le type des variables (Figure 3).










Info				
2647 instances 1 feature (0.0% missing values) Data has no target variable. 8 meta attributes				
Columns (Double click to edit)				
	Name	Type	Role	Values
1	THEMATIQUES	 categorical	feature	Catastrophes, Culture-loisirs, Economie, Education, Environnement, Faits divers, ...
2	MOIS	 text	meta	
3	Nombre de sujets de JT (1)	 text	meta	
4	Nombre de sujets de JT (2)	 text	meta	
5	Nombre de sujets de JT (3)	 text	meta	
6	Nombre de sujets de JT (4)	 text	meta	
7	Nombre de sujets de JT (5)	 text	meta	
8	Nombre de sujets de JT (6)	 text	meta	
9	Nombre de sujets de JT (7)	 text	meta	

Figure 3 : Capture d'écran des infos de notre jeu de données, sur le logiciel Orange.

Ici, on souhaite que la variable "MOIS" soit de type catégorical et que les variables "Nombres de sujets de JT" de chaque chaîne soient de type numérique, or cela n'est pas le cas ici. De plus, le nom des variables ne sont pas assez explicites.

Nous avons ainsi dû procéder à un nettoyage de mes données.

1. APPRENTISSAGE NON SUPERVISÉ

1.2 - Nettoyage des données

Notre objectif est que le logiciel Orange nous classe les thématiques des sujets de JT les plus représentés sur la période septembre 2018 - septembre 2020. Cela nous est impossible si le logiciel ne reconnaît pas le type de nos variables.

De plus, notre jeu de données englobe une période très large, de janvier 2005 à septembre 2020.

Nous avons ainsi créé une copie du fichier .csv de mon jeu de données et procéder à un nettoyage des données :

- Suppression de la variable Totaux, celle-ci n'est pas utile dans mon cas,
- Suppression de la ligne 2 de mon tableur,
- Changement du nom des variables quantitatives pour ajouter le nom des chaînes,
- Suppression des données n'étant pas dans la période septembre 2018 - septembre 2020,
- Suppression de la variable Canal +, car la diffusion des JT a été arrêtée en 2016,
- Vérification des données, en particulier, les données manquantes et les données aberrantes.

	A	B	C	D	E	F	G
1	MOIS	THEMATIQUES	TF1 - Nombre de sujets de JT	France 2 - Nombre de sujets de JT	France 3 - Nombre de sujets de JT	Arte - Nombre de sujets de JT	M6 - Nombre de sujets de JT
2	septembre-18	Catastrophes	28	32	39	7	30
3	septembre-18	Culture-loisirs	28	22	25	20	55
4	septembre-18	Economie	78	72	45	15	36
5	septembre-18	Education	21	19	6	2	18
6	septembre-18	Environnement	29	28	23	14	15
7	septembre-18	Faits divers	12	16	32	2	27
8	septembre-18	Histoire-hommages	6	15	13	5	9
9	septembre-18	International	33	71	33	112	34
10	septembre-18	Justice	20	20	19	8	35
11	septembre-18	Politique France	32	38	41	9	27
12	septembre-18	Santé	17	15	20	7	14
13	septembre-18	Sciences et techniques	14	5	2	3	8
14	septembre-18	Société	88	95	60	38	72
15	septembre-18	Sport	13	15	12	3	30
16	octobre-18	Catastrophes	98	98	92	7	67
17	octobre-18	Culture-loisirs	39	27	17	26	59
18	octobre-18	Economie	85	67	41	8	38
19	octobre-18	Education	8	10	7	4	11
20	octobre-18	Environnement	36	40	32	17	16
21	octobre-18	Faits divers	22	28	25	6	36
22	octobre-18	Histoire-hommages	32	27	25	10	21
23	octobre-18	International	46	84	38	150	52
24	octobre-18	Justice	21	21	28	6	26
25	octobre-18	Politique France	32	38	44	7	30
26	octobre-18	Santé	19	20	21	11	30
27	octobre-18	Sciences et techniques	14	5	5	9	8
28	octobre-18	Société	98	96	69	38	80
29	octobre-18	Sport	9	11	8	0	25
30	novembre-18	Catastrophes	22	20	19	8	24

Figure 4 : Capture d'écran du tableur de notre jeu de données après nettoyage des données.

Notre jeu de données comporte maintenant 350 instances pour 2 variables qualitatives et 5 variables quantitatives. Le logiciel Orange détecte maintenant correctement le type des variables.

Columns (Double click to edit)				
	Name	Type	Role	Values
1	MOIS	C categorical	feature	août-19, août-20, avril-19, avril-20, décembre-18, décembre-19, février-1...
2	THEMATIQUES	C categorical	feature	Catastrophes, Culture-loisirs, Economie, Education, Environnement, Faits ...
3	TF1 - Nombre de sujets de JT	N numeric	feature	
4	France 2 - Nombre de sujets de JT	N numeric	feature	
5	France 3 - Nombre de sujets de JT	N numeric	feature	
6	Arte - Nombre de sujets de JT	N numeric	feature	
7	M6 - Nombre de sujets de JT	N numeric	feature	

Figure 5 : Capture d'écran des infos de notre jeu de données, sur le logiciel Orange après nettoyage.

Nous pouvons passer à l'étape du clustering.

1 - APPRENTISSAGE NON SUPERVISÉ

1.3 - Clustering

Pour aborder la question de recherche, nous avons conçu un dendrogramme en utilisant l'ensemble des données nettoyées. Pour cela, nous avons commencé par élaborer une matrice de distance en intégrant l'outil "Distances" à notre chaîne de traitement. Ensuite, nous avons configuré cet outil pour utiliser une mesure de distance euclidienne normalisée. Pour finir, afin de visualiser les clusters, nous avons connecté la sortie de cette étape à l'outil "Hierarchical Clustering", nous permettant ainsi de générer et d'examiner le dendrogramme.

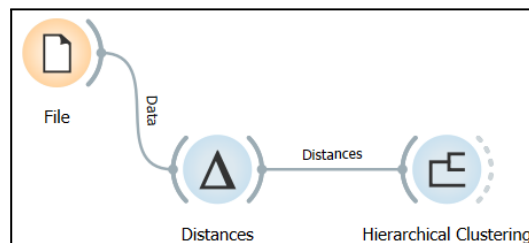


Figure 6 : Capture d'écran de la chaîne de traitement de notre jeu de données, sur le logiciel Orange.

Nous avons identifié 5 clusters :

- 1er Cluster : Ce premier cluster, englobant les mois mars, avril et mai 2020, se distingue par une attention particulière portée à la santé, alignée sur l'évolution de la pandémie de COVID-19 en France. Cette période critique a vu les médias jouer un rôle essentiel en informant et guidant le public, avec une couverture médiatique intense et continue. TF1 et France 2 ont été en tête en termes de volume de couverture, tandis qu'Arte a adopté une approche plus spécialisée sur ce sujet.
- 2ème Cluster : Ce cluster semble englober une variété de thèmes tels que l'environnement, les faits divers, et l'éducation. Cela suggère une période de programmation plus "normale" ou diversifiée par rapport au 1er Cluster.

-

7

2. APPRENTISSAGE SUPERVISÉ

2.1 - Jeu de données

Prédiction du nombre de but par matchs joués pour l'année 2017 à partir des données de l'année 2016 :

Nous avons découvert un ensemble de données sur la plateforme Kaggle (voir Figure 8), qui correspond parfaitement à nos besoins, axé sur les performances statistiques des footballeurs.

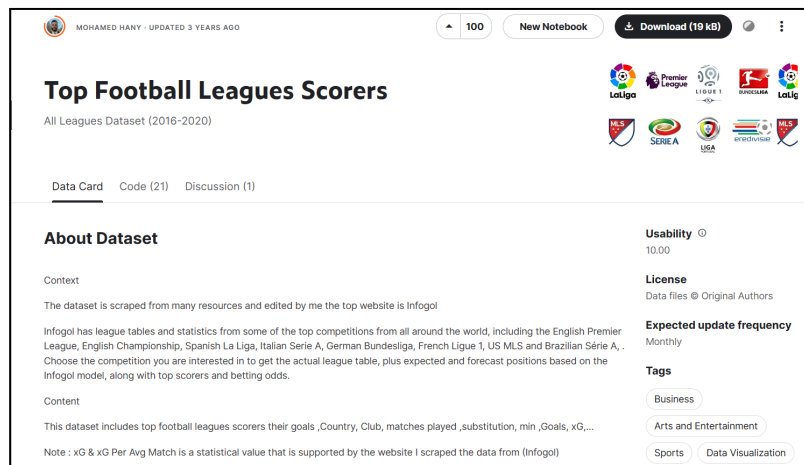


Figure 8 : Capture d'écran du site kaggle.com où se trouve notre jeu de données.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Country	League	Club	Player Names	Matches Played	Substitution	Mins	Goals	xG	xG Per Avg Match	Shots	OnTarget	Shots Per Avg Match	On Target Per Av
2	Spain	La Liga	(BET)	Juanmi Callejon	19	16	1849	11.6.62	0.34		48	202.47		1.03
3	Spain	La Liga	(BAR)	Antoine Griezmnn	36	0	3129	16.11.86	0.36		88	412.67		1.24
4	Spain	La Liga	(ATL)	Luis Suarez	34	1	2940	28.23.21	0.75		120	573.88		1.84
5	Spain	La Liga	(CAR)	Ruben Castro	32	3	2842	13.14.06	0.47		117	423.91		1.4
6	Spain	La Liga	(VAL)	Kevin Gameiro	21	10	1745	13.10.65	0.58		50	232.72		1.25
7	Spain	La Liga	(JUV)	Cristiano Ronaldo	29	0	2634	25.24.68	0.89		162	605.84		2.16
8	Spain	La Liga	(RMA)	Karim Benzema	23	6	1967	11.13.25	0.64		69	343.33		1.64
9	Spain	La Liga	(PSG)	Neymar	30	0	2694	13.13.33	0.47		105	423.7		1.48
10	Spain	La Liga	(CEL)	Iago Aspas	25	7	2354	19.13.88	0.56		78	373.15		1.49
11	Spain	La Liga	(EIB)	Sergi Enrich	31	7	2804	11.8.25	0.27		64	262.09		0.85
12	Spain	La Liga	(None)	Asduriz	27	5	2480	16.15.92	0.61		85	453.26		1.72
13	Spain	La Liga	(HUE)	Sandro Ramirez	28	2	2340	14.7.14	0.29		93	383.78		1.54
14	Spain	La Liga	(BAR)	Lionel Messi	32	2	2910	37.26.65	0.87		179	765.84		2.48
15	Spain	La Liga	(VIL)	Gerard Moreno	37	0	3361	13.8.49	0.24		82	322.32		0.9
16	Spain	La Liga	(JUV)	Morata	14	12	1392	15.9.67	0.66		55	303.75		2.05
17	Spain	La Liga	(MON)	Wissam Ben Yedder	20	11	1735	11.7.85	0.43		44	232.41		1.26
18	Spain	La Liga	(SOC)	William Jose	27	1	2102	12.8.41	0.38		69	293.12		1.31
19	Spain	La Liga	(Florin)	Andone	32	5	2984	12.11.62	0.37		99	423.15		1.34
20	Spain	La Liga	(SOC)	Cedric Bakambu	17	9	1633	10.8.08	0.47		50	262.91		1.51
21	Spain	La Liga	(RMA)	Isco	18	12	1690	10.3.91	0.22		32	151.8		0.84
22	Italy	Serie A	(LIV)	Mohamed Salah	29	2	2567	15.11.62	0.43		80	502.96		1.85
23	Italy	Serie A	(SAS)	Gregoire Defrel	23	6	2054	12	80.37		43	281.99		1.3
24	Italy	Serie A	(LAZ)	Ciro Immobile	35	1	3294	22.19.76	0.57		136	983.92		2.83
25	Italy	Serie A	(VER)	Nikola Kalinic	26	6	2648	15.15.05	0.54		90	613.23		2.19
26	Italy	Serie A	(NAP)	Dries Mertens	28	7	2671	28.21.65	0.77		148	1025.26		3.63
27	Italy	Serie A	(ATA)	Alejandro Gomez	37	0	3247	16.11.62	0.34		117	753.42		2.19
28	Italy	Serie A	(FIO)	Jose Callejon	37	0	3276	14.11.38	0.33		86	572.49		1.65
29	Italy	Serie A	(BEN)	Iago Falque	31	4	2633	12.8.04	0.29		65	422.35		1.52
30	Italy	Serie A	(CAG)	Giovanni Simeone	29	6	2764	12.12.22	0.42		76	352.61		1.2
31	Italy	Serie A	(PSG)	Mauro Icardi	34	0	3168	24.19.13	0.57		109	523.25		1.55
32	Italy	Serie A	(CRG)	Diego Falcinelli	35	1	3308	13.11.49	0.33		99	452.84		1.67
33	Italy	Serie A	(Cyril)	Cyril Theraud	31	2	2585	12.9.25	0.34		70	482.57		1.76
34	Italy	Serie A	(ROM)	Edin Dzeko	33	4	3194	29.30.6	0.91		178	995.29		2.94
35	Italy	Serie A	(NAP)	Lorenzo Insigne	35	2	3031	18.13.08	0.41		147	814.61		2.54
36	Italy	Serie A	(SAM)	Fabio Quagliarella	35	2	3030	12.13.71	0.43		112	703.51		2.19

Figure 9 : Capture d'écran du tableur de notre jeu de données sans aucune modification.

Notre jeu de données initial contient 661 sujets pour 5 variables qualitatives et 10 variables quantitatives. Il présente des statistiques détaillées sur les joueurs de football de différentes ligues, incluant des informations sur les matches joués, les buts, les tirs, et d'autres mesures de performance, pour chaque année de 2016 à 2020.

Les variables qualitatives incluent le nom du joueur ("Player Names"), le pays dans lequel joue le joueur ("Country"), la ligue où performe le joueur ("League"), le club du joueur ("Club") ainsi que l'année ("Year").

Les variables quantitatives comprennent le nombre total de matches joués par chaque joueur ("Matches Played"), le nombre de fois qu'un joueur a été remplacé au cours des matches ("Substitution"), le nombre total de minutes jouées par le joueur ("Mins"), le nombre total de buts marqués par le joueur ("Goals"), la mesure statistique qui évalue la probabilité qu'une occasion de but se transforme en but ("xG"), la moyenne de xG par match pour un joueur ("xG Per Arg Match"), le nombre total de tirs effectués par le joueur ("Shots"), le nombre de tirs du joueur qui ont été cadrés ("OnTarget"), la moyenne de tirs par match pour un joueur ("Shots Per Avg Match") et la moyenne de tirs cadrés par match pour un joueur ("On Target Per Avg Match").
















	Name	Type	Role	Values
1	Country	 categorical	feature	Brazil, England, France, Germany, Italy, ...
2	League	 categorical	feature	Bundesliga, Campeonato Brasileiro SÃ©ri...
3	Matches_Played	 numeric	feature	
4	Substitution	 numeric	feature	
5	Mins	 numeric	feature	
6	Goals	 numeric	feature	
7	xG	 numeric	feature	
8	xG Per Avg ...	 numeric	feature	
9	Shots	 numeric	feature	
10	OnTarget	 numeric	feature	
11	Shots Per Avg ...	 numeric	feature	
12	On Target Per ...	 numeric	feature	
13	Year	 numeric	feature	
14	Club	 text	meta	
15	Player Names	 text	meta	

Figure 10 : Capture d'écran des infos de notre jeu de données, sur le logiciel Orange.

2. APPRENTISSAGE SUPERVISÉ

2.2 - Nettoyage et prétraitement des données

Notre objectif est que le logiciel Orange nous prédit le nombre de buts par match joué pour chaque joueur pour l'année 2017 à partir de l'année 2016. Notre jeu de données contient de base 660 joueurs, or nous souhaitons garder que les données de l'année 2016. Nous allons ainsi procéder à un nettoyage des données.

Nous avons ainsi créé une copie du fichier .csv de notre jeu de données et procéder à un nettoyage des données :

- Sélection des données pour l'année 2016,
- Création de la variable qui donnera la tranche des buts pour l'année 2016 (“Category Goals”),
- Vérification des données, en particulier, les données manquantes et les données aberrantes,
- Suppression des variables “Year”, “Country”, “League” et “Goals”.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	
1	Country	League	Club	Player Names	Matches Played	Substitution	Mins	Goals	xG	xG Per Avg Match	Shots	OnTarget	Shots Per Avg Match	On Target Per Avg Match	Year	Goals Per Avg Match	
2	Spain	La Liga	(JUV)	Cristiano Ronaldo	29	0	2634	2524.68	0.89		162	60.5.84		2.16	2016	0.86	
3	Germany	Bundesliga	(KOL)	Anthony Modeste	34	0	3075	2519.74	0.61		101	53.3.12		1.64	2016	0.74	
4	England	Premier League	(INT)	Romelu Lukaku	36	1	3448	2518.15	0.5		110	53.03.03		1.46	2016	0.69	
5	Spain	La Liga	(ATL)	Luis Suarez	32	1	3008	2525.65	0.81		121	55.3.82		1.74	2017	0.78	
6	Spain	La Liga	(BAR)	Lionel Messi	32	1	3067	2521.63	0.67		159	68.4.93		2.11	2019	0.77	
7	Brazil	Campeonato Brasileiro SÃ©rie A	(FLA)	Gabriel Barbosa	29	0	2716	2524.59	0.86		117	62.04.09		2.17	2019	0.86	
8	Italy	Serie A	(TOR)	Andrea Belotti	34	1	3241	2619.45	0.57		130	78.3.81		2.29	2016	0.76	
9	Spain	La Liga	(JUV)	Cristiano Ronaldo	27	0	2375	26	291.16		178	76.7.12		03.04	2017	0.96	
10	Italy	Serie A	(SAM)	Fabio Quagliarella	37	0	3269	2623.06	0.67		141	57.4.1		1.66	2018	0.70	
11	Spain	La Liga	(ATL)	Luis Suarez	34	1	2940	2823.21	0.75		120	57.3.88		1.84	2016	0.82	
12	Italy	Serie A	(NAP)	Dries Mertens	28	7	2671	2821.65	0.77		148	102.5.26		3.63	2016	1.00	
13	USA	MLS	(ATA)	Josef Martinez	32	0	3050	2825.04	0.78		128	52.3.99		1.62	2019	0.88	
14	Germany	Bundesliga	(CHE)	Timo Werner	33	1	2744	2820.8	0.72		123	63.4.26		2.18	2019	0.85	
15	Italy	Serie A	(ROM)	Edin Dzeko	33	4	3194	2930.6	0.91		178	99.5.29		2.94	2016	0.88	
16	England	Premier League	(TOT)	Harry Kane	29	1	2636	2918.87	0.68		110	58.3.96		02.09	2016	1.00	
17	Italy	Serie A	(LAZ)	Ciro Immobile	33	0	2799	2919.15	0.65		110	53.3.73		1.8	2017	0.88	
18	Italy	Serie A	(PSG)	Mauro Icardi	34	0	3096	2923.14	0.71		101	53.3.1		1.63	2017	0.85	
19	Germany	Bundesliga	(BAY)	Robert Lewandowski	24	6	2247	2926.49	1.12		127	54.5.37		2.28	2017	1.21	
20	Germany	Bundesliga	(BAY)	Robert Lewandowski	31	2	2871	3030.52	0.101		143	86.4.73		2.85	2016	0.97	
21	Germany	Bundesliga	(ARS)	Pierre-Emerick Aubameyang	31	1	2894	3128.94	0.95		116	78.3.81		2.56	2016	1.00	
22	USA	MLS	(ACM)	Zlatan Ibrahimovic	31	0	2998	3122.72	0.72		159	67.05.04		2.12	2019	1.00	
23	Italy	Serie A	(JUV)	Cristiano Ronaldo	33	0	3127	3127.32	0.83		208	79.6.32		2.4	2019	0.94	
24	Spain	La Liga	(BAR)	Lionel Messi	32	4	3123	3332.54	0.99		197	95.5.99		2.89	2017	1.03	
25	France	France Ligue 1	(PSG)	Kylian Mbappe-Lottin	24	5	2488	3331.17	1.19		125	70.4.77		2.67	2018	1.38	
26	Germany	Bundesliga	(BAY)	Robert Lewandowski	31	0	2783	3431.05	0.106		138	67.4.71		2.29	2019	1.10	
27	Spain	La Liga	(BAR)	Lionel Messi	29	5	2849	3625.49	0.85		170	87.5.67		2.9	2018	1.24	
28	USA	MLS	(LAF)	Carlos Vela	33	0	3128	3625.35	0.77		167	75.05.07		2.28	2019	1.09	
29	Italy	Serie A	(LAZ)	Ciro Immobile	36	1	3371	3626.61	0.75		142	71	4		2	2019	1.00
30	Spain	La Liga	(BAR)	Lionel Messi	32	2	2910	3726.65	0.87		179	76.5.84		2.48	2016	1.16	

Figure 11 : Capture d'écran du tableau de notre jeu de données après nettoyage des données.

2. APPRENTISSAGE SUPERVISÉ

2.3 - Modèles d'apprentissage

Notre jeu de données contient actuellement 80 sujets, pour 11 variables, dont 9 variables quantitatives et 2 variables qualificatives.

Dans le logiciel Orange, la variable “Category Goals” est définie comme la cible (Target), tandis que “Player Names” est catégorisée comme métadonnée (Meta). Toutes les autres variables quantitatives sont assignées en tant que caractéristiques (Features), comme illustré dans les figures 12 et 13.

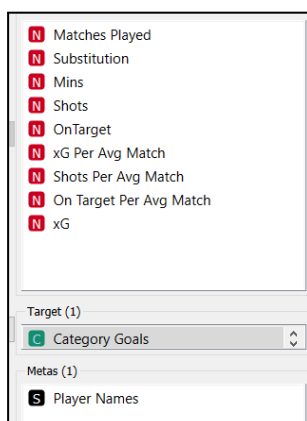


Figure 12 : Capture d'écran des infos de notre jeu de données, sur le logiciel Orange après nettoyage.

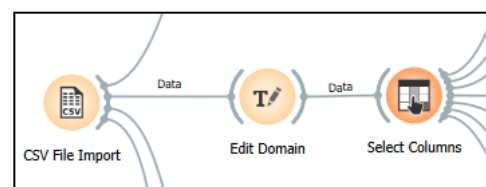


Figure 13 : Capture d'écran du début de la chaîne de traitement de notre jeu de données, sur le logiciel Orange.

Nous avons procédé à l'application de différents modèles d'apprentissage supervisé pour analyser la performance des joueurs en fonction des variables définies. Nous avons ensuite appliqué une suite d'algorithmes d'apprentissage machine pour prédire la variable cible “Category Goals”.

Chaque modèle a été évalué à l'aide du composant “Test and Score” qui fournit des métriques de performance telles que la précision, le rappel, et l'aire sous la courbe “ROC”. Cette évaluation nous a permis de comparer objectivement la performance de chaque modèle et de sélectionner le plus performant pour notre objectif de prédiction.

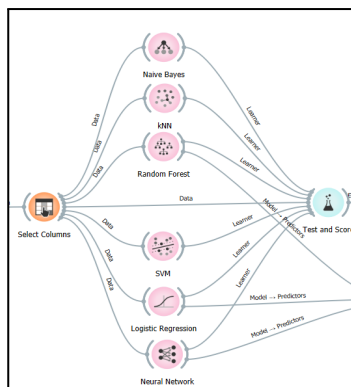


Figure 14 : Capture d'écran de la partie modélisation de la chaîne de traitement de notre jeu de données, sur le logiciel Orange.

Voici un résumé des résultats d'évaluation des modèles :

“Random Forest” et “Logistic Regression” ont affiché d'excellentes performances avec une précision de classification autour de 98.8% et des scores “AUC” proches de la perfection. “kNN” a montré des résultats solides, mais un peu inférieurs aux modèles précédents, avec une précision de classification de 91.2%. “Naive Bayes” a eu la performance la plus faible en précision de classification avec 77.5%, bien que son “AUC” soit élevé. “SVM” et “Neural Network” ont eu des scores parfaits dans toutes les métriques, ce qui est inhabituel dans la pratique et peut indiquer un surajustement.

En général, tous les modèles sauf “Naive Bayes” ont montré de très bonnes performances.

Evaluation results for target (None, show average over classes) ▾						
Model	AUC	CA	F1	Prec	Recall	MCC
Random Forest	1.000	0.988	0.987	0.988	0.988	0.936
kNN	0.964	0.912	0.887	0.920	0.912	0.450
Naive Bayes	0.983	0.775	0.815	0.925	0.775	0.499
SVM	1.000	1.000	1.000	1.000	1.000	1.000
Logistic Regression	0.991	0.988	0.987	0.988	0.988	0.936
Neural Network	1.000	1.000	1.000	1.000	1.000	1.000

Figure 15 : Capture d'écran des résultats affichés dans Test and Score.

Nous avons ensuite relié le “Test and Score” à une “Confusion Matrix” et nous avons obtenu la confirmation de ces résultats. “SVM” et “Neural Network” ne se trompent pas.

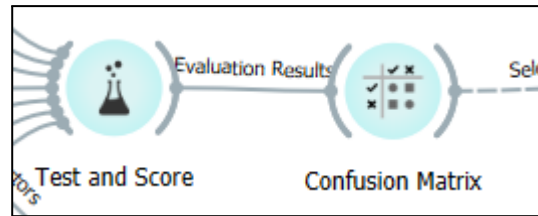


Figure 16: Capture d'écran de la partie évaluations des résultats de notre jeu de données.
sur le logiciel Orange.

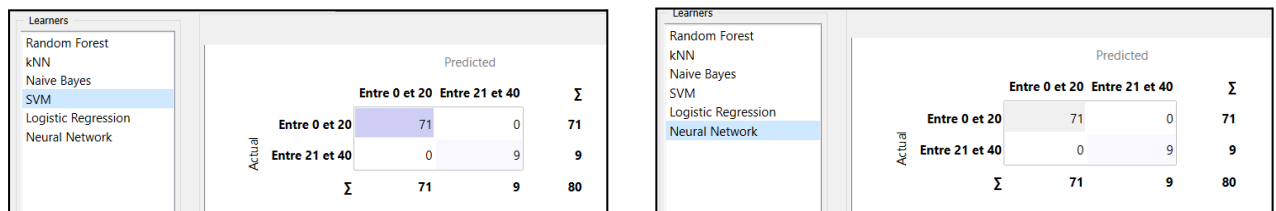


Figure 17 : Captures d'écran des résultats affichés dans Confusion Matrix pour les modèles SVM et Neural Network.

2. APPRENTISSAGE SUPERVISÉ

2.4 - Prédictions

Nous avons ajouté un fichier .csv avec la variable “Category Goals” à prédire par le logiciel Orange.

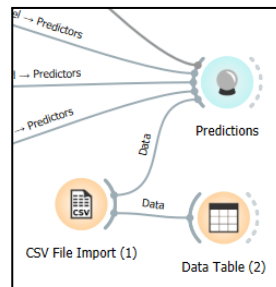


Figure 18 : Capture d'écran de la partie sur les prédictions de notre chaîne de traitement.

	Neural Network	Random Forest	Logistic Regression	SVM	Player Names	Category Goals	Matches_Played	Substitution	Mins	xG	xG Per Avg Match	Shots	OnTarget	ots Per Avg Matc	Target Per Avg h
1	0.99 : 0.01 → Entre 0 et 20	1.00 : 0.00 → Entre 0 et 20	1.00 : 0.00 → Entre 0 et 20	0.99 : 0.01 → Entre 0 et 20	Willian Josa	?	32	2	2869	13.89	0.46	89	29	2.95	0.96
2	1.00 : 0.00 → Entre 0 et 20	1.00 : 0.00 → Entre 0 et 20	1.00 : 0.00 → Entre 0 et 20	0.99 : 0.01 → Entre 0 et 20	Simone Zaza	?	23	10	2163	12.07	0.53	73	24	3.21	1.05
3	0.35 : 0.65 → Entre 21 et 40	0.61 : 0.39 → Entre 0 et 20	0.67 : 0.33 → Entre 0 et 20	0.00 : 1.00 → Entre 0 et 20	Robert Lewand...	?	24	6	2247	26.49	1.12	127	54	5.37	2.28
4	0.99 : 0.01 → Entre 0 et 20	1.00 : 0.00 → Entre 0 et 20	1.00 : 0.00 → Entre 0 et 20	0.98 : 0.02 → Entre 0 et 20	Portu	?	35	2	3081	10.7	0.33	49	25	1.51	0.77
5	0.04 : 0.96 → Entre 21 et 40	0.90 : 0.10 → Entre 0 et 20	0.88 : 0.12 → Entre 0 et 20	0.00 : 1.00 → Entre 0 et 20	Paulo Dybala	?	26	7	2407	11.4	0.45	114	45	4.5	1.78
6	0.99 : 0.01 → Entre 0 et 20	1.00 : 0.00 → Entre 0 et 20	0.99 : 0.01 → Entre 0 et 20	0.98 : 0.02 → Entre 0 et 20	Mikel Oyarzabal	?	31	4	2864	9.35	0.31	65	32	2.16	1.06
7	0.99 : 0.01 → Entre 0 et 20	1.00 : 0.00 → Entre 0 et 20	0.99 : 0.01 → Entre 0 et 20	1.00 : 0.00 → Entre 0 et 20	Maxi Gomez	?	35	1	3168	18.01	0.54	83	37	2.49	1.11
8	0.08 : 0.92 → Entre 21 et 40	0.71 : 0.29 → Entre 0 et 20	0.69 : 0.31 → Entre 0 et 20	0.00 : 1.00 → Entre 0 et 20	Luis Suarez	?	32	1	3008	25.65	0.81	121	55	3.82	1.74
9	0.15 : 0.85 → Entre 21 et 40	0.81 : 0.19 → Entre 0 et 20	0.00 : 1.00 → Entre 21 et 40	0.00 : 1.00 → Entre 0 et 20	Lionel Messi	?	32	4	3123	32.54	0.99	197	95	5.99	2.89
10	0.23 : 0.77 → Entre 21 et 40	0.71 : 0.29 → Entre 0 et 20	0.04 : 0.96 → Entre 21 et 40	0.00 : 1.00 → Entre 0 et 20	Cristiano Ronal...	?	27	0	2375	29	1.16	178	76	7.12	3.04

Figure 19 : Capture d'écran des prédictions.

Après observation des prédictions on peut observer que le modèle “Neural Network” n’a fait aucune erreur. Cependant, le modèle “SVM” se trompe sur plus de la moitié des joueurs car il met à tout le monde, entre 0 et 20 buts marqués. Ces erreurs montrent qu’un modèle de classification n’est jamais sûr à 100% et qu’il est très difficile de prédire le nombre de buts marqués par un joueur.

Les résultats obtenus illustrent la complexité inhérente à la prédiction des performances sportives et démontrent l'importance d'une évaluation rigoureuse des modèles d'apprentissage machine. Ce projet a non seulement renforcé notre compréhension des nuances de la modélisation prédictive mais a également mis en lumière la nécessité d'une interprétation prudente des résultats pour guider les décisions basées sur les données.

WEBOGRAPHIE

Kaggle - Top Football Leagues Scorers Dataset

Hany, M. Top Football Leagues Scorers [Data set]. Kaggle

<https://www.kaggle.com/datasets/mohamedhanyyy/top-football-leagues-scorers>

Site Web de Kaggle

Kaggle. Homepage. <https://www.kaggle.com/>

data.gouv.fr - Classement Thématique des Sujets de Journaux Télévisés

Classement thématique des sujets de journaux télévisés (Janvier 2005 - Septembre 2020) [Data set]. data.gouv.fr.

<https://www.data.gouv.fr/fr/datasets/classement-thematique-des-sujets-de-journaux-televises-janvier-2005-septembre-2020/>

Site Web de data.gouv.fr

data.gouv.fr. Homepage. <https://www.data.gouv.fr/fr/>

Site Web d'Orange Data Mining

Orange Data Mining. Homepage. <https://orangedatamining.com/>