Analyses statistiques et prédictions TOP 14

Mathieu Roig

Septembre 2024 - Août 2025

Table des matières

1 Introduction du sujet

L'objectif de ce document est d'introduire les données et les méthodes que j'ai utilisé pour étudier le championnat de TOP 14, le premier championnat de rugby à XV de France et probablement du monde.

Le travail va se décomposer en trois grandes parties, une première partie de récolte de données, qu'on cherchera à visualiser dans un second temps, puis qui donnera lieu à une partie prédiction orientée Machine Learning et modèles probabilistes.

2 Données

2.1 Collecte

L'option la plus pertinente pour récupérer les données à étudier a été de travailler sur **Wikipedia**. D'autres sites sont plus détaillés, comme notamment le site officiel du top14 mais il demande des processus plus compliqués d'authentification pour des données moins essentielles, ou encore celui de all rugby qui propose des informations sur les transferts mais difficilement exploitables.

C'est pourquoi, l'option retenue a été de travailler sur la page Wikipedia — Championnat de France de rugby 2024–2025. Parmi toutes les data disponibles sur la page, seulement 5 tableaux ont été rétenus. Un des points important à vérifier avant de récuperer des données est qu'on peut les avoir, dans l'idéal au même format, pour toutes les années. Pour être plus clair, si on a seulement accès au joueur de l'année à partir de 2022, cette statistique ne va pas pouvoir être utilisable lors de la partie prédiction car lorsque l'on va l'introduire dans nos modeles, on a besoin qu'elle soit disponible pour chaque année.

Tableaux retenus : voir figure 1

- **Présentation** : pour chaque club — dernière montée, budget (M€), classement de la saison précédente, entraîneur, stade, capacité, compétition européenne d'inscription.
- **Classement** : pour chaque club — journées jouées (26 si la saison est terminée), victoires, nuls, défaites, points marqués et encaissés, bonus, total de points en fin de saison en comptant (victoire=4, nul=2, défaite=0, +1 bonus offensif et/ou défensif).
- **Évolution** : pour chaque club — rang à la fin de chaque journée.
- **Forme** : pour chaque club — résultat par journée (victoire, nul, défaite).
- **Résultats** : scores des matchs à domicile (lignes) et à l'extérieur (colonnes).

Même si Wikipedia ne demande pas d'authentification pour récupérer les données, il y'a plusieurs contraintes qui gène l'automatisation de la collecte. Le problème principal est que toutes les pages sont différentes et n'ont pas été faites à la suite, ce qui fait qu'elles ne sont **pas du tout uniformisées**. Bien que certaines années se ressemblent, il faut donc faire au cas par cas pour récupérer les bons tableaux.

Etudier au cas par cas revient à trouver les tableaux dans chaque page, ainsi qu'à construire ceux qui n'y apparaissent pas : forme apparait à partir de 2016, évolution à partir de 2015 et présentation à partir de 2009.

Le défi principal a été de construire évolution et forme car même si on a les scores dans résultats, on ne dispose pas du calendrier des matchs donc impossible de construire les autres tableaux puisqu'on ne peut pas avoir quel score correspond à quelle journée. Pour ça j'ai dû utiliser un LLM et le site all rugby pour reconstruire, journée après journée, les tableaux manquants.

2.2 Nettoyage

Au final on a pu récupérer les 5 tableaux (présentation, classement, évolution, forme, résultats) **depuis 2005** (la première saison de top14) **jusqu'à 2025**.

Il y'a maintenant un travail de nettoyage de données à faire, car les données ont été récupérées brutes mais la conversion depuis la page Wikipédia n'est pas parfaite, on le voit sur la table présentation avec des indices au dessus des capacités, des notes, ou bien des colonnes présentes certaines années qui ne le seront pas dans d'autres.

Il faut donc passer sur toutes les données pour voir tous les cas particuliers de notes qui ont été introduites, les retirer afin de tout uniformiser.

Un autre principal problème d'uniformisation concerne le **nom des équipes**, on peut penser au cas du Montpellier RC qui est devenu le Montpellier HR en 2009, mais à des cas encore plus simples comme Stade Toulousain et Stade toulousain qui ne seraient pas reconnus de la même manière, USAP et USA Perpignan qui sont tous les deux mentionnés pour désigner le même club, ou dans le tableau résultats BAY et Aviron bayonnais qui doivent désigner aussi le même club. Il y'a donc un travail dit de mapping qui a été fait pour que tous les clubs portent un seul et même nom afin de ne pas perdre leurs données. Pour donner une idée, il y'a environ 150 noms différents pour une trentaine d'équipe.

En regardant seulement pour la saison 2024-2025 pour esquisser le format des données, on a finalement les tableaux suivants :

Club	Dernière montée	Budget (M€)	Classement précédent	Entraîneur en chef	Stade	Capacité
Stade toulousain	1907	49	1	Ugo Mola	Stade Ernest-Wallon, Toulouse	18784
Stade français Paris	1997	45	2	Laurent Labit	Stade Jean-Bouin, Paris	19607
Union Bordeaux Bègles	2011	30	3	Yannick Bru	Stade Chaban-Delmas, Bordeaux	34635
RC Toulon	2008	35	4	Pierre Mignoni	Stade Mayol, Toulon	16437
Stade rochelais	2014	37	5	Ronan O'Gara	Stade Marcel-Deflandre, La Rochelle	16689
Racing 92	2009	27.3	6	Patrice Collazo	Paris La Défense Arena, Nanterre	30680
Castres olympique	1989	25	7	Xavier Sadourny	Stade Pierre-Fabre, Castres	11778
ASM Clermont	1925	34	8	Christophe Urios	Stade Marcel-Michelin, Clermont-Ferrand	19357
Section paloise	2015	28	9	Sébastien Piquerонies	Stade du Hameau, Pau	14999
USA Perpignan	2021	22.5	10	Franck Azéma	Stade Aimé-Giral, Perpignan	14727
Lyon OU	2016	35	11	Karim Ghezal	Matmut Stadium Gerland, Lyon	35052
Aviron bayonnais	2022	30	12	Grégory Patat	Stade Jean-Dauger, Bayonne	14537
Montpellier HR	2003	28	13	Joan Caudullo	GGL Stadium, Montpellier	14392
RC Vannes	2024	19	16	Jean-Noël Spitzer	Stade de la Rabine, Vannes	11865

TABLE 1 – Tableau présentation saison 2024

Club	J1	J2	J3	J4	J5	J6	J7	J8	J9	...	J18	J19	J20	J21	J22	J23	J24	J25	J26	Année
Aviron bayonnais	7	12	13	9	12	10	7	3	4	...	4	4	4	4	4	4	4	4	4	2024
Union Bordeaux Bègles	3	7	2	1	1	1	2	2	2	...	2	2	2	2	2	2	2	2	2	2024
Castres olympique	6	5	4	6	3	7	6	7	7	...	5	5	5	5	5	5	5	5	6	2024
ASM Clermont	1	3	3	7	6	8	5	6	6	...	7	9	6	7	6	8	7	7	5	2024
Lyon OU	5	2	7	5	7	4	8	9	10	...	6	6	7	6	8	9	11	11	11	2024
Montpellier HR	10	8	10	10	13	11	12	11	11	...	8	10	8	9	9	6	8	9	9	2024
Stade français Paris	12	11	8	13	11	13	13	12	13	...	11	11	12	12	12	13	12	12	12	2024
Section paloise	14	9	12	8	8	6	9	10	12	...	10	7	9	10	11	10	9	8	8	2024
USA Perpignan	8	13	14	12	9	12	11	13	9	...	13	13	13	13	13	12	13	13	13	2024
Racing 92	9	4	11	11	10	9	10	8	8	...	12	12	11	11	10	11	10	10	10	2024
Stade rochelais	4	10	5	4	2	3	3	4	3	...	9	8	10	8	7	7	6	6	7	2024
RC Toulon	11	6	6	3	4	5	4	5	5	...	3	3	3	3	3	3	3	3	3	2024
Stade toulousain	2	1	1	2	5	2	1	1	1	...	1	1	1	1	1	1	1	1	1	2024
RC Vannes	13	14	9	14	14	14	14	14	14	...	14	14	14	14	14	14	14	14	14	2024

TABLE 2 – Tableau évolution saison 2024

Club	J1	J2	J3	J4	J5	J6	J7	J8	J9	...	J18	J19	J20	J21	J22	J23	J24	J25	J26	Année
Aviron bayonnais	V	D	D	V	D	V	V	V	V	...	V	D	V	D	V	N	V	D	V	2024
Union Bordeaux Bègles	V	D	V	V	V	V	D	V	D	...	V	V	D	V	D	D	V	D	V	2024
Castres olympique	V	D	V	D	V	D	V	D	V	...	D	N	V	V	D	V	D	V	D	2024
ASM Clermont	V	D	V	D	V	D	V	D	V	...	D	D	V	D	V	D	V	V	V	2024
Lyon OU	V	V	D	V	D	V	D	D	D	...	V	V	D	V	D	D	D	D	D	2024
Montpellier HR	D	V	D	D	D	V	D	V	D	...	V	D	V	D	V	V	D	V	D	2024
Stade français Paris	D	V	D	D	V	D	D	V	D	...	V	V	D	D	D	D	V	D	V	2024
Section paloise	D	V	D	V	D	V	D	D	D	...	V	V	D	D	D	V	V	V	V	2024
USA Perpignan	D	D	D	V	V	D	V	D	V	...	D	D	N	V	D	V	D	D	V	2024
Racing 92	D	V	D	D	V	V	D	V	V	...	D	V	V	D	V	N	V	D	V	2024
Stade rochelais	V	D	V	V	V	D	V	D	V	...	D	N	D	V	V	V	V	V	D	2024
RC Toulon	D	V	V	V	D	D	V	D	V	...	D	V	D	V	D	D	D	V	D	2024
Stade toulousain	V	V	V	D	D	V	V	V	D	...	V	D	V	V	V	V	D	V	D	2024
RC Vannes	D	D	V	D	D	D	D	V	D	...	D	D	N	D	V	D	D	D	D	2024

TABLE 3 – Tableau forme saison 2024

Club ▼ dom. ► ext.	AB	UBB	CAS	ASM	LYO	MHR	SFP	SEC	PER	RAC	SR	RCT	ST	VAN
AB	—	36-32	33-12	31-18	28-14	28-27	21-13	27-22	21-19	32-15	37-7	18-10	12-8	38-32
UBB	30-27	—	34-29	22-18	20-22	9-6	46-26	19-6	66-12	52-34	10-21	21-17	32-24	59-28
CAS	33-3	3-13	—	34-29	30-25	30-26	35-13	24-19	27-12	31-28	28-24	28-26	28-23	32-13
ASM	26-10	32-27	54-10	—	39-31	18-22	55-20	39-7	31-13	21-23	33-19	19-18	18-35	55-33
LYO	38-49	28-26	40-38	22-30	—	32-23	35-3	27-29	17-12	34-47	53-17	27-20	17-17	36-21
MHR	42-10	46-27	21-17	10-23	22-26	—	38-32	30-3	19-13	21-17	16-0	30-38	11-20	26-24
SFP	31-27	19-46	21-10	36-6	31-30	29-20	—	39-37	24-7	40-24	22-17	10-14	21-27	34-31
SEC	51-29	22-26	33-26	20-14	29-15	40-38	30-16	—	23-6	23-33	32-18	25-21	14-22	48-24
PER	16-11	17-29	20-20	33-3	29-26	7-26	20-18	11-10	—	28-24	21-13	13-22	42-35	32-13
RAC	24-24	36-31	20-27	33-20	25-25	25-27	49-24	29-47	30-23	—	16-17	22-6	17-21	25-30
SR	29-28	32-22	12-12	20-15	43-22	47-18	35-18	49-25	38-15	21-26	—	19-15	22-19	14-23
RCT	39-19	27-10	30-28	31-24	21-10	30-17	24-6	56-25	40-19	36-24	45-26	—	16-50	54-19
ST	41-6	12-16	52-6	48-14	43-3	27-17	38-23	55-10	41-9	35-37	35-27	57-5	—	63-21
VAN	21-27	29-37	34-28	19-20	30-20	37-24	33-28	26-52	20-20	24-27	29-30	29-19	18-43	—

TABLE 4 – Tableau résultats saison 2024

Rang	Club	J	V	N	D	Bo	Bd	Pm	Pe	Diff	Pts	Année
1	Stade toulousain	26	18	1	7	11	5	891	462	429	90	2024
2	Union Bordeaux Bègles	26	17	0	9	5	5	762	609	153	78	2024
3	RC Toulon	26	15	0	11	7	5	680	595	85	72	2024
4	Aviron bayonnais	26	15	1	10	2	4	632	650	-18	68	2024
5	ASM Clermont	26	13	0	13	6	5	674	627	47	63	2024
6	Castres olympique	26	13	2	11	3	4	626	658	-32	63	2024
7	Stade rochelais	26	13	1	12	5	3	617	635	-18	62	2024
8	Section paloise	26	13	0	13	4	5	682	719	-37	61	2024
9	Montpellier HR	26	12	0	14	3	5	623	609	14	56	2024
...	...	...	...	...	...	...	...	...	...	...	...	...
13	USA Perpignan	26	9	2	15	2	2	469	647	-178	44	2024
14	RC Vannes	26	7	1	18	1	5	661	891	-230	36	2024

TABLE 5 – Tableau classement saison 2024

3 Visualisation

3.1 Statistiques descriptives

Commençons par visualiser les principales statistiques des tables présentation et classement. On crée une fonction prennant en paramètre le nom de l'équipe que l'on souhaite étudier, et pour *Stade toulousain*, on obtient :

Saison	Budget (M€)	Entraîneur	Rang	J	V	D	Diff
24/25	49.0	Ugo Mola	1	26	18	7	429
23/24	62.525	Ugo Mola	1	26	16	9	173
22/23	43.7	Ugo Mola	1	26	17	8	208
21/22	37.3	Ugo Mola	4	26	15	11	206
20/21	36.6	Ugo Mola	1	26	17	8	210
19/20	37.2	Ugo Mola	7	17	8	8	37
18/19	32.0	Ugo Mola	1	26	21	3	312
17/18	30.86	Ugo Mola	3	26	16	9	124
16/17	31.5	Ugo Mola	12	26	11	15	-24
15/16	30.87	Ugo Mola	5	26	16	8	287

TABLE 6 – Statistiques clés du Stade toulousain sur les 10 dernières années

Une autre notion importante à vérifier est l'évolution du budget des équipes, en parallèle de la professionnalisation du rugby le budget moyen a plus de doublé sur les 15 dernières années :

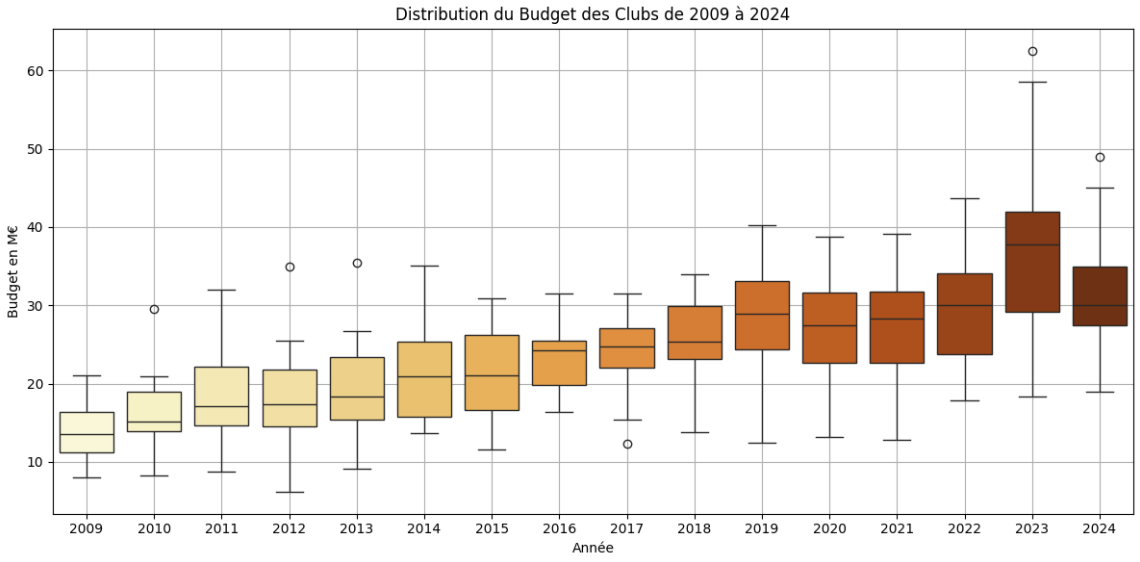

FIGURE 1 – Evolution du budget sur les 15 dernières années

Or bien que le budget moyen tend à s'élever, rappelons la forte disparité entre les clubs :

TABLE 7 – Budgets des clubs (M€) pour la saison 2024

Club	TLS	STF	ROC	TLN	LOU	CLE	UBB	BAY	PAU	MHR	RAC	CAS	PER	VAN
Budget (M€)	49.0	45.0	37.0	35.0	35.0	34.0	30.0	30.0	28.0	28.0	27.3	25.0	22.5	19.0

Enfin, visualisons l'évolution du classement à la fin de la 26ème journée des 4 clubs les plus attendus :

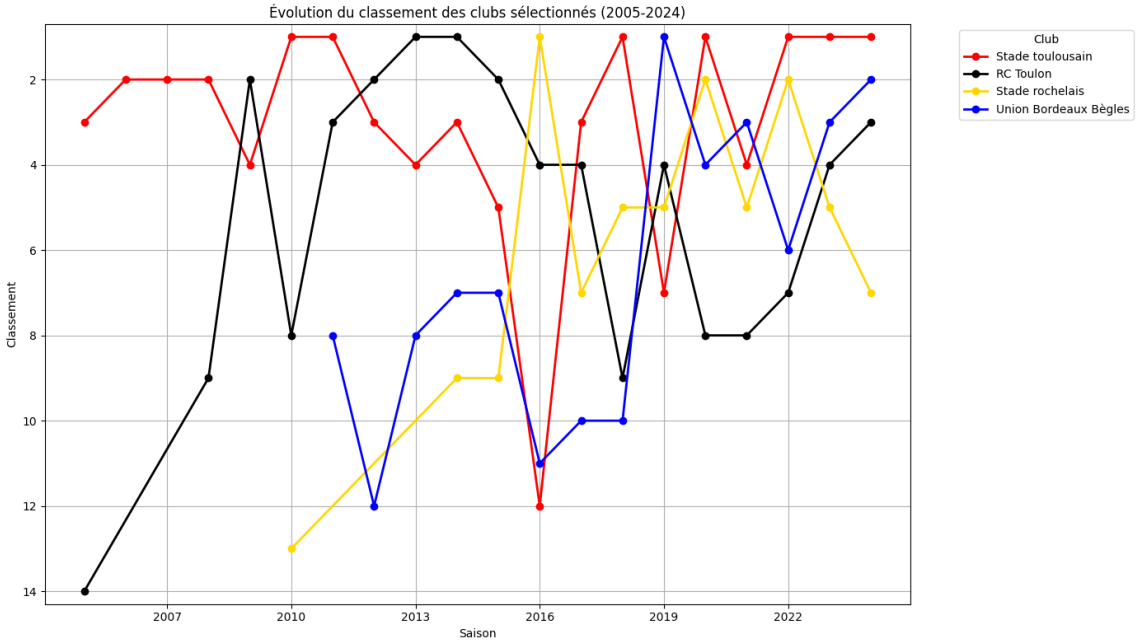

FIGURE 2 – Evolution du classement depuis le début du top14

La Figure 3 permet de voir les différentes périodes de domination des équipes, avec l'époque du grand Toulon 2013-2015, une saison en enfer en 2016 pour Toulouse puis une domination presque sans faille depuis 5 ans, ainsi qu'une croissance récente pour Bordeaux ces 3 dernières années parallèlement à un déclin de La Rochelle.

3.2 Suivis des performances

Visualisons encore quelques statistiques des tableaux récupérés précédemment, commençons par suivre le tableau évolution au cours des années, en particulier pour le *Stade toulousain* :

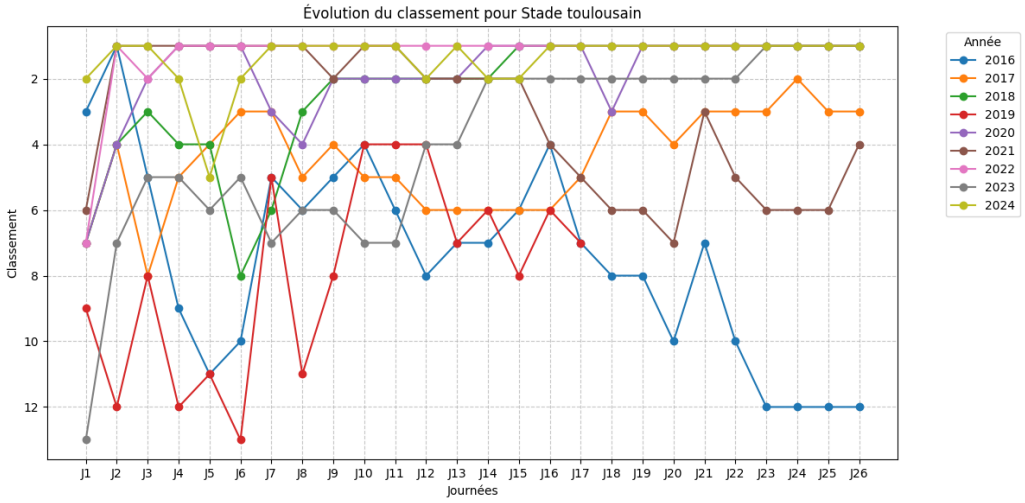

FIGURE 3 – Evolution du classement du Stade toulousain depuis 10 ans

Ainsi à part lors de la saison catastrophe de 2016, après la mi saison (donc la 13ème journée), le Stade ne descends jamais au dessous de la 8ème place. Pour donner une idée, c'est le rang minimum qu'il faut avoir à la fin de la saison pour être qualifié pour la grande coupe d'Europe.

Puis en suivant le tableau forme au cours des années, toujours pour le *Stade toulousain* :

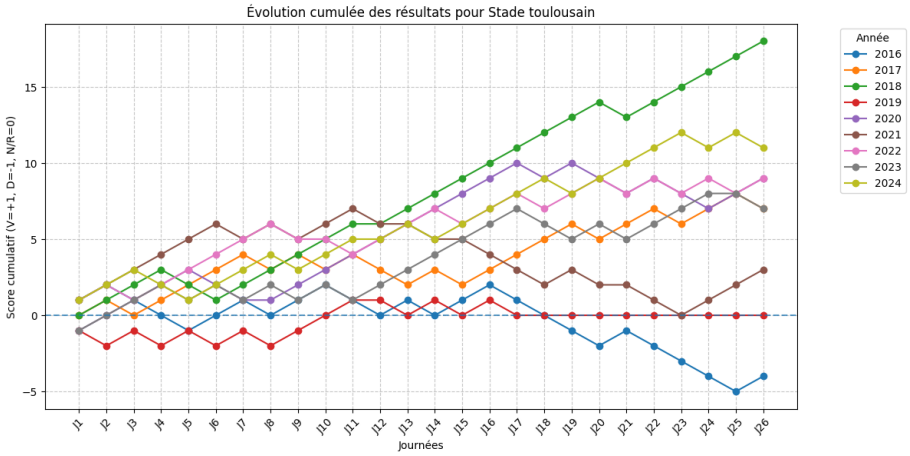

FIGURE 4 – Suivi de l'état de forme du Stade toulousain depuis 10 ans

Pour un club comme Toulouse connu pour prêter quasiment toute son équipe une à la sélection nationale, on aurait pu s'attendre à relever des paterns comme par une chute de performance entre la J14 et la J18 qui correspondent en général au moment du tournoi 6 nations, or il semblerait que ce soit en réalité l'inverse et davantage une période où l'équipe performe, ce qui montre la profondeur d'effectif du club.

4 Prédictions

4.1 Prédire le champion

Pour rappel, une saison de top14 commence par 26 semaines de championnat puis les 6 premiers au classement sont qualifiés pour les phases éliminatoires.

Le défi dans cette première partie va être de **savoir si on peut prédire à la fin des 26 journées de championnat le gagnant des phases finales**. Commençons par visualiser les précédents champions :

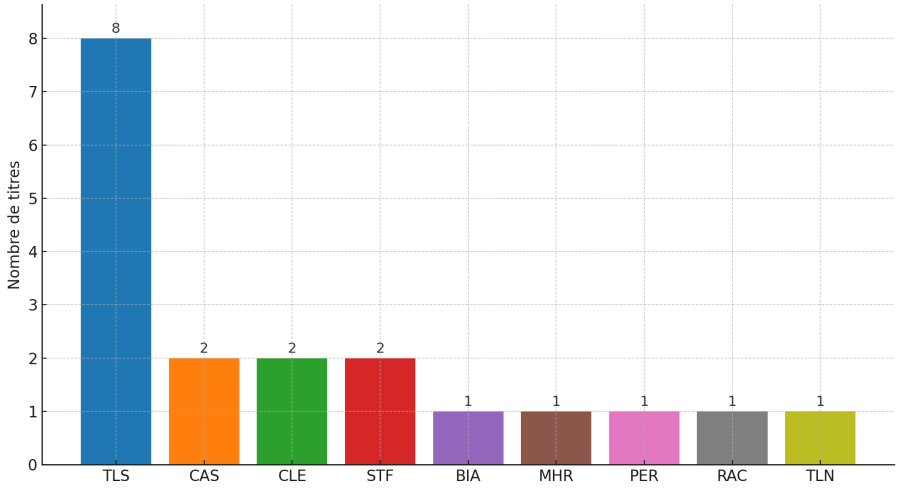

FIGURE 5 – Bouclier de Brennus depuis le début du top14

Une première idée naïve pourrait être de se dire : **les 26 journées ont été suffisantes pour dégager une hiérarchie dans les équipes, et cette hiérarchie continuera d'être la même en phases finales**, notons cette hypothèse **BASE**.

On a ici une règle très simple, mais qui en regardant les 20 dernières années marche en réalité très bien :
— Le premier de la saison régulière est **champion** : 11 fois sur 19, soit **57,8 %**.
— Il est **champion ou finaliste** : 14 fois sur 19, soit **73,7 %**.

La question est donc de savoir si il est possible de battre BASE en apprenant des patterns indétectables pour l'homme via des méthodes de Machine Learning qui permettrait de prédire avec précision qui va être le champion de l'année, connaissant la saison régulière.

Features Le premier défi va être de sélectionner des statistiques pertinentes parmi les données qu'on a pu récupérer pour prédire qui sera le champion :
— *Classement* nous donne le nombre de points, le différentiel de score, le rang
— *Forme* nous permet de voir comment l'équipe a commencé la saison et comment elle a fini, sa série de victoire la plus longue
— *Evolution* nous donne le % de victoire contre des équipes du top6

— *Résultats* permet de savoir le % de victoire à domicile et à l'extérieur
— On peut rajouter aussi des variables indicatrices de si le club a été champion ou finaliste les années précédentes

On résume alors toutes nos données dans un tableau qui a l'allure suivante :

Année	Club	Pts	Diff	Rang	Forme (5)	Streak max	...	Champ $t-1$	% top6	% dom.	% ext.
2005	ASM Clermont	63	8	8	12	2	...	0	0.33	0.85	0.23
2005	Aviron bayonnais	43	-155	12	4	2	...	0	0.25	0.54	0.08
2005	Biarritz Olympique	90	344	1	12	6	...	0	0.50	0.92	0.54
2005	CA Brive	51	-122	9	8	2	...	0	0.25	0.69	0.08
2005	CS Bourgoin-Jallieu	67	75	6	4	5	...	0	0.30	0.77	0.31
⋮	⋮	⋮	⋮	⋮	⋮	⋮	...	⋮	⋮	⋮	⋮
2024	Stade français Paris	45	-158	12	8	2	...	0	0.25	0.77	0.00
2024	Stade rochelais	62	-18	7	16	5	...	0	0.50	0.77	0.23
2024	Stade toulousain	90	429	1	12	4	...	1	0.60	0.85	0.54
2024	USA Perpignan	44	-178	13	8	2	...	0	0.25	0.69	0.00
2024	Union Bordeaux Bègles	78	153	2	8	5	...	0	0.70	0.85	0.46

TABLE 8 – Tableau features pour les prédictions

4.1.1 Régression logistique

Avec ces statistiques, on commence par l'approche la plus simple possible : on adopte une *régression logistique* avec validation *Leave-One-Season-Out* (**LOSO**) : pour chaque saison t, le modèle est appris sur toutes les saisons $\neq t$ puis, pour chaque club de la saison t, on estime la probabilité d'être champion. Le champion prédit est le club dont la probabilité estimée est la plus élevée. On obtient les valeurs suivantes :

Année	1er favori	Proba	2e favori	Proba	Vainqueur officiel	Proba	Succès
2024	Union Bordeaux Bègles	0.79	Stade toulousain	0.75	Stade toulousain	0.75	0
2023	Stade toulousain	0.85	Stade français Paris	0.84	Stade toulousain	0.85	1
2022	Stade toulousain	0.87	Stade rochelais	0.83	Stade toulousain	0.87	1
2021	Castres olympique	0.95	Montpellier HR	0.71	Montpellier HR	0.71	0
2020	Stade toulousain	0.90	Racing 92	0.74	Stade toulousain	0.90	1
2018	Stade toulousain	0.97	Lyon OU	0.74	Stade toulousain	0.97	1
2017	Montpellier HR	0.92	Racing 92	0.87	Castres olympique	0.12	0
2016	Stade rochelais	0.95	ASM Clermont	0.68	ASM Clermont	0.68	0
2015	ASM Clermont	0.92	Montpellier HR	0.79	Racing 92	0.66	0
...	...	...	...	...	...	...	...

TABLE 9 – Prédiction du champion de la saison par régression logistique

Jusqu'à 2018 (2019 exclu car Covid) on a 4 succès sur 6, en revanche lorsqu'on prend jusqu'à 2005 on a seulement **42% de précision du champion** et **74% du top2**, ce qui reste en dessous de BASE donc **aucun intérêt pour la regression logistique**.

4.1.2 Random Forest

La limite de la régression logistique précédente est son caractère essentiellement linéaire. Pour capturer des relations non linéaires et des interactions entre variables, on va passer sur un modèle de *Random Forest* pour classification. Le principe est d'entraîner un grand nombre d'arbres de décision sur des échantillons *bootstrapés* du jeu d'entraînement, en ne considérant qu'un sous-ensemble aléatoire de variables à chaque séparation. En classification, la prédiction est obtenue par *vote majoritaire* des arbres, et la probabilité estimée pour un club est la moyenne des proportions de la classe positive (=1 i.e champion) dans les feuilles terminales.

On garde parcontre le même schéma **LOSO** : pour chaque saison t, la forêt est apprise sur toutes les saisons $\neq t$, puis on estime pour chaque club de la saison t, sa probabilité d'être champion. Le club avec la probabilité la plus élevée est le *champion prédit*.

On obtient avec cette méthode en prenant jusqu'à 2005 environ **57,8%** de précision pour le champion, et seulement **63,2%** pour le top2.
A la différence de la régression logistique précédente, la méthode de Random Forest demande de faire plusieurs choix d'hyperparamètres, les 2 résultats précédents sont obtenus avec les hyperparamètres conseillés mais en cherchant à les optimiser par *grid search*, on plafonne à **57,8%** pour le champion et **73,7%** pour le top2, ce qui est au final **exactement les résultats de BASE**.

Cela mérite alors de vérifier si au final on ne prédit pas exactement le premier de chaque saison comme champion avec notre modèle, et en comparant avec notre table classement on constate que **c'est le cas 18 fois sur 19**.
Ainsi en optimisant les hyperparamètres de la Random Forest, le modèle converge vers une stratégie très proche de notre hypothèse BASE ("*le champion est le 1er de la saison régulière*"). Cela s'explique par le fait que cette règle simple est déjà très performante compte tenu des données disponibles. Pour aller au-delà, il faudrait soit davantage de données (plus de saisons), soit des variables plus discriminantes liées aux phases finales, afin que le modèle puisse apprendre des régularités plus fines que ce simple proxy.
Au final la Random Forest ne nous apprend pas davantage de patterns que BASE.

4.1.3 XG Boost

Une dernière approche à envisager est celle du *Gradient Boosting*, et en particulier son implémentation efficace *XGBoost*. Contrairement à la Random Forest qui construit de nombreux arbres en parallèle et agrège leurs votes (méthode de bagging), le boosting repose sur une logique étape après étape : chaque nouvel arbre est appris de manière à corriger les erreurs des arbres précédents.

Ce principe rend XGBoost plus flexible et plus puissant pour capturer des relations complexes et subtiles entre les variables. Comme pour les méthodes précédentes, on utilise la validation **LOSO** : pour chaque saison t, le modèle est entraîné sur toutes les saisons $\neq t$ et les probabilités de titre sont estimées pour chaque club.

Après optimisation des hyperparamètres par *grid search* de nouveau, on obtient une précision de **68,4%** pour le champion (13 sur 19 contre 11 sur 19 précédemment) et **100%** pour le top2. On a alors bien réussi à dépasser BASE, que ce soit pour le champion qui était notre objectif, mais on arrive également à prédire les 2 finalistes avec 100% de réussite :

Année	1er favori	Proba	2e favori	Proba	Vainqueur officiel	Proba	Succès
2024	Stade toulousain	0.565	Union Bordeaux Bègles	0.332	Stade toulousain	0.565	1
2023	Stade toulousain	0.672	Stade rochelais	0.087	Stade toulousain	0.672	1
2022	Stade rochelais	0.856	Stade toulousain	0.847	Stade toulousain	0.847	0
2021	Montpellier HR	0.643	Castres olympique	0.506	Montpellier HR	0.643	1
2020	Stade toulousain	0.894	Stade rochelais	0.666	Stade toulousain	0.894	1
2018	ASM Clermont	0.402	Stade toulousain	0.249	Stade toulousain	0.249	0
2017	Castres olympique	0.080	RC Toulon	0.051	Castres olympique	0.080	1
2016	ASM Clermont	0.112	RC Toulon	0.105	ASM Clermont	0.112	1
2015	Racing 92	0.375	ASM Clermont	0.372	Racing 92	0.375	1
2014	Stade français Paris	0.055	RC Toulon	0.019	Stade français Paris	0.055	1
2013	RC Toulon	0.126	Castres olympique	0.097	RC Toulon	0.126	1
2012	Stade toulousain	0.796	Castres olympique	0.148	Castres olympique	0.148	0
2011	Stade toulousain	0.783	Montpellier HR	0.092	Stade toulousain	0.783	1
2010	Stade toulousain	0.893	USA Perpignan	0.058	Stade toulousain	0.893	1
2009	USA Perpignan	0.706	ASM Clermont	0.273	ASM Clermont	0.273	0
2008	USA Perpignan	0.943	Stade toulousain	0.045	USA Perpignan	0.943	1
2007	ASM Clermont	0.877	Stade toulousain	0.147	Stade toulousain	0.147	0
2006	Stade français Paris	0.738	ASM Clermont	0.139	Stade français Paris	0.738	1
2005	Stade français Paris	0.543	Biarritz Olympique	0.169	Biarritz Olympique	0.169	0

TABLE 10 – Prédictions du champion de la saison par XGBoost

Deux questions se posent avant de nous satisfaire de nos résultats. La première est de savoir si on est pas dans un **cas d'overfitting**. Avec aussi peu de données l'overfitting est un risque réel mais ce problème est mis de côté par le choix de la validation LOSO puisque pour estimer le champion de notre saison t, on utilise aucune donnée de la saison en question.
La deuxième question est aussi liée au faible nombre de saisons, et revient à se demander si on a pas eu un **coup de chance** d'augmenter notre précision car en réalité on a réussi à prévoir seulement 2 champions de plus.

Pour vérifier si cette amélioration est significative, on peut réaliser un test statistique (test binomial) qui compare la probabilité d'obtenir 13 succès sur 19 saisons avec XGBoost contre 11 sur 19 avec la Random Forest. La p-value obtenue est d'environ **0.41**, ce qui est bien supérieur au seuil classique de 0.05.

Pour comprendre ce chiffre, on a utilisé un *test binomial* : il s'agit de comparer le nombre de succès obtenus avec XGBoost (13 saisons correctement prédites sur 19) à ceux de la Random Forest (11 sur 19). Sous l'hypothèse nulle que les deux modèles ont en réalité la même probabilité de succès, chaque saison où leurs prédictions diffèrent peut être vue comme un tirage à pile ou face. Dans ce cas, obtenir que XGBoost fasse mieux dans 2 saisons de plus correspond à une fluctuation que l'on peut calculer précisément avec la loi binomiale. Le calcul conduit à une probabilité de 0.41 d'observer un tel écart (ou un écart plus grand) simplement par hasard.

Ce résultat est important car il signifie que l'amélioration brute (68% contre 58%) **n'est pas statistiquement significative** : on ne peut pas rejeter l'idée que les deux modèles aient en réalité la même performance et que la différence observée ne soit qu'un effet du hasard dû au faible nombre de saisons (19 seulement). En d'autres termes, même si XGBoost obtient deux saisons de plus correctement prédites, cela ne suffit pas à conclure qu'il est réellement meilleur que notre hypothèse BASE : avec davantage de données on pourrait

confirmer ou au contraire invalider cette tendance.

Pour résumer les réponses aux 2 questions, **la validation LOSO nous protège de l'overfitting, mais la taille de l'échantillon limite fortement la portée des comparaisons statistiques**.
Néanmoins l'utilisation du modèle XGBoost reste très intéréssante car **on apporte un nouveau signal**, le modèle comprend bien qu'en général être premier fait de l'équipe la favorite : lorsqu'il prédit que le premier est champion il a juste 8 fois sur 9, mais là où il se distingue c'est qu'il arrive également à prédire le champion quand ce n'est pas le premier 5 fois sur 9 (dont la finale de 2022 où La Rochelle est prédit champion, qui n'était pas si loin d'arriver ...) et encore plus impressionnant il arrive à prédire les 2 finalistes avec 100% de réussite. Ces résultats encourageants restent à vérifier avec davantages de saisons ou de données, notamment sur les joueurs.

4.2 Prédire la saison

4.2.1 Introduction et vérification

Une autre piste à étudier est de **savoir si il est possible de prédire avant que la saison commence qui va être le champion, et de voir comment cette prédiction evolue au cours du temps**.

Dans cette idée, on va introduire un autre système de prédiction, pas basé sur du machine learning mais sur un modèle probabiliste, on va parler de **modèle d'Elo**.

Le principe est le suivant : **on attribue à chaque club un nombre qui représente leur état actuel de forme, son Elo, et la différence d'Elo entre deux clubs lors d'une rencontre leur attribue plus ou moins de chance de gagner**.

Pour calculer cet Elo, on utilise plusieurs statistiques :
- D'abord le rang (après phases éliminatoires) et le differentiel de points des deux saisons précédentes,
- Pour chacune des statistiques on va accorder 70% de poids à la saison qui vient de se terminer et 30% à la précédente (puisque les progressions s'effectuent en général dans une continuité autour d'un projet de groupe, qui dans certains cas croît, dans d'autres s'éteint),
- On attribue aussi une importance à l'état de forme des équipes avant la fin de la saison, on regarde les 5 derniers matchs et en fonction des performances de l'équipe on ajuste son Elo.
- Il faut également faire un choix de comment décider l'Elo du promu puisque ses données ne sont pas disponibles pour au moins la saison précédente, une solution pour son différentiel de points est la formule avec une pénalité à la fin :

$$\mathrm{Diff}_{\mathrm{promu},t} = \mu_{\mathrm{bot}} - \texttt{promo\_penalty}\,\sigma_{\mathrm{bot}} - \Delta_{\mathrm{niveau\_ProD2}}\,.$$

- On ajoute une note de mercato d'intersaison sur 10 pour ajouter des données exogènes aux matchs :

$$\mathrm{Elo}_i = \mathrm{Elo}_i + \texttt{gamma} \cdot \mathrm{NoteMercato}_i$$

- On considère qu'il y'a surtout 3 grosses équipes : Toulouse, Bordeaux, Toulon puis une course entre toutes les autres jusqu'au promu qui est considéré détaché des autres (le top14 tend à être une ligue fermée). C'est pourquoi on va compresser autour de ces blocs là avec la formule :

$$\mathrm{Elo}_i = \mu_{\mathrm{bloc}} + \tau_{\mathrm{bloc}}\,(\mathrm{Elo}_i - \mu_{\mathrm{bloc}})$$

avec des facteurs τ_{bloc} spécifiques à chaque blocs [1-3], [4-13], [14].

Vérifions sur les données de la saison 2024/2025 si les Elo initiaux sont pertinents :

Club	$\boldsymbol{Elo_i}$	$\mathbf{classement\_prédit_i}$	$\mathbf{classement\_final_i}$
Stade toulousain	1559.14	1	1
RC Toulon	1548.85	2	3
Union Bordeaux Bègles	1546.96	3	2
Stade rochelais	1540.05	4	7
Racing 92	1527.82	5	10
Stade français Paris	1523.63	6	12
ASM Clermont	1512.86	7	5
Castres olympique	1501.82	8	6
Lyon OU	1497.16	9	11
Section paloise	1497.09	10	8
Aviron bayonnais	1496.36	11	4
USA Perpignan	1485.20	12	13
RC Vannes	1483.73	13	14
Montpellier HR	1469.53	14	9

TABLE 11 – Elo de pré-saison (Elo_i), rang prédit ($\mathrm{classement\_prédit}_i$) et rang final 2024 ($\mathrm{classement\_final}_i$).

En comparant le rang prédit (qui est juste la position relative entre les Elo) et le classement final, **il y'a bien le top3 qui se dégage mais la prédiction entre 4 et 13 est effectivement complexe**, la percée de l'aviron bayonnais a été sous évaluée, sûrement en même temps que son mercato (l'arrivée de Tuilagui a été sous-estimée par beaucoup alors que sans lui quelle aurait été la saison de Maqala ...). Il y'a donc un enjeu à ajuster la note de mercato la plus proche de la réalité possible car elle drive énormément l'Elo prédit.

4.2.2 Saison 2025/2026

Pour la saison 2025/2026, on va utiliser les notes de mercato déduites des analyses de Aymeric Milan (journaliste rugby) :

USP	LAR	UBB	LOU	TOU	BAY	RCT	PAU	MHR	CAS	RAC	SFP	ASM	USM
9.0	8.0	8.0	7.5	6.5	6.0	5.5	5.5	5.5	5.0	5.0	4.5	4.5	4.0

TABLE 12 – Notes de mercato intersaison été 2025

Avec ces notes en main, comme on vient de faire pour la saison test 2024/2025, on peut calculer maintenant le classement pour la saison prochaine, la saison 2025/2026 :

Club	Elo_i	$\mathrm{classement\_prédit}_i$
Stade toulousain	1577.56	1
Union Bordeaux Bègles	1561.78	2
Stade rochelais	1544.21	3
RC Toulon	1522.94	4
Aviron bayonnais	1510.92	5
ASM Clermont	1508.28	6
Section paloise	1506.93	7
USA Perpignan	1498.87	8
Racing 92	1497.74	9
Castres olympique	1495.94	10
Montpellier HR	1491.88	11
Lyon OU	1491.06	12
Stade français Paris	1482.52	13
US Montauban	1454.37	14

TABLE 13 – Elo de pré-saison ajusté (Elo_i) et rang prédit ($\mathrm{classement\_prédit}_i$) pour la saison 2025/2026

Le dernier est logiquement Montauban, il aurait été difficile de les prévoir à une autre place. On constate que Bayonne confirme dans les 6, puis une forte montée de Perpignan dû à son mercato de cet été. La fin de série se joue entre Montpellier, Lyon et le Stade Français qui hormis Lyon semblent avoir des projets sportifs en fin de vie.

Le second atout de la méthode d'Elo permet de **pouvoir attribuer un pourcentage de victoire à chacun des clubs qui se rencontrent en fonction de leur différence d'Elo**. C'est pourquoi on va utiliser notre méthode pour prévoir les prochaines rencontres.

Calcul des probabilités de match. L'idée est simple : à partir des deux Elo et d'un avantage domicile, on transforme ces informations en trois probabilités (domicile, nul, extérieur). On va utiliser une version douce du modèle de Davidson, qui gère explicitement le nul.

— **Entrées.** Pour un match, on part de l'Elo de l'équipe à domicile E_{dom} et de l'équipe visiteuse E_{ext}. On utilise trois hyperparamètres :
  — H : avantage domicile (en points Elo) ;
  — s : "pente" de l'échelle (valeur typique autour de 400) ;
  — ν : paramètre qui ouvre la porte au nul (si $\nu = 0$, il n'y a plus de nuls).

— **On applique l'avantage domicile.**

$$E^{\star}_{\text{dom}} = E_{\text{dom}} + H, \qquad E^{\star}_{\text{ext}} = E_{\text{ext}}.$$

— **On passe sur une échelle multiplicative.** Plus un Elo est grand, plus la "force" associée est élevée :

$$A = 10^{\,E^{\star}_{\text{dom}}/s}, \qquad B = 10^{\,E^{\star}_{\text{ext}}/s}.$$

— **On calcule les probabilités.** Le terme ν augmente la masse autour de l'égalité :

$$D \;=\; A + B + 2\,\nu\,\sqrt{AB}, \qquad P(\text{dom}) = \frac{A}{D},\; P(\text{nul}) = \frac{2\,\nu\,\sqrt{AB}}{D},\; P(\text{ext}) = \frac{B}{D}.$$

Une question qui se pose est de **trouver ces hyperparamètres H, s et ν**. Surtout au rugby qui est un sport de territoire, l'avantage de jouer à la maison est un facteur décisif, c'est pourquoi on peut pas prendre le même H que dans un match de foot ou de handball où l'avantage n'est pas le même.
L'idée alors pour trouver ces hyperparamètres va être de parcourir toutes les données qu'on a déjà, de 2005 à 2025, où on connait à chaque fois le classement final, pour trouver ces paramètres qui correspondent le mieux aux paramètres réels. Pour résumer, on va alors prendre les H, s et ν qui correspondent le plus à ceux qui auraient dû être utilisés pour prédire les saisons précédentes.

On donne à titre indicatif : H = 70, s = 400 et ν = 0.04.
D'ailleurs, l'allocation de poids $70\,\%$ à la saison $t-1$ et $30\,\%$ à la saison $t-2$ pour prédire la saison t vient du même algorithme d'optimisation.

Avec toutes ces informations, on peut alors prédire les pourcentages de victoire de chacun des clubs de la première journée :

Domicile	P_{dom}	**Extérieur**	P_{ext}
Stade français Paris	67.4	US Montauban	29.8
USA Perpignan	59.5	Aviron bayonnais	37.6
Castres olympique	59.7	Section paloise	37.4
Lyon OU	60.6	Racing 92	36.5
Montpellier HR	55.5	RC Toulon	41.5
Union Bordeaux Bègles	65.4	Stade rochelais	31.8
ASM Clermont	47.3	Stade toulousain	49.7

TABLE 14 – Probabilités de victoire (en %) pour la J1 de la saison 2025/2026

On remarquera logiquement que **la somme des 2 probabilités ne fait pas 1** puisque l'évènement match nul est de probabilité positive.

Enfin le dernier atout de notre modèle est **sa faculté à pouvoir être mis à jour**, puisqu'on a défini un Elo initial, ce dernier va pouvoir **être actualisé en fonction des résultats des différentes journées**.

Mise à jour des Elo après chaque journée. Nous partons d'un *Elo de départ* (pré-saison) et nous l'actualisons à l'issue de chaque journée à partir des scores réels. L'idée est de rapprocher le rating de ce qui s'est effectivement passé, tout en tenant compte de l'ampleur de la victoire et des attentes avant match.

- **Entrées.** Pour la journée J, on dispose :
  - des Elo en début de journée ;
  - du calendrier avec les scores au format `"xx-xx"`.
- **Étape 1 — Attente avant match.** Pour chaque affiche, on recalcule les probabilités $(P_{\text{dom}}, P_{\text{nul}}, P_{\text{ext}})$ avec les Elo *courants* et les hyperparamètres (H, s, ν) (même modèle que ci-dessus).
- **Étape 2 — Score attendu vs observé.**
  - *Score attendu domicile* : $E_h = P_{\text{dom}} + 0{,}5\,P_{\text{nul}}$.
  - *Score observé domicile* : $S_h = 1$ (victoire), $0{,}5$ (nul), 0 (défaite).

  L'équipe visiteuse a $E_a = P_{\text{ext}} + 0{,}5\,P_{\text{nul}}$ et $S_a = 1 - S_h$.

— **Étape 3 — Facteur "marge".** Une large victoire doit compter davantage qu'un succès d'un point. On utilise un multiplicateur g croissant avec l'écart au score (et modéré quand l'écart Elo était déjà grand).

— **Étape 4 — Mise à jour symétrique.** On applique une correction proportionnelle à l'écart $(S-E)$:

$$\Delta = K \times g \times (S_h - E_h), \qquad \begin{cases} \text{Elo}_{\text{dom}} \leftarrow \text{Elo}_{\text{dom}} + \Delta, \\ \text{Elo}_{\text{ext}} \leftarrow \text{Elo}_{\text{ext}} - \Delta. \end{cases}$$

Le paramètre K règle la vitesse d'apprentissage (plus K est grand, plus l'Elo bouge vite).

— **Étape 5 — Application "en bloc".** Pour éviter tout biais d'ordre, on *gèle* les Elo au début de la journée, on calcule Δ pour tous les matchs, puis on applique toutes les corrections d'un coup. On obtient un tableau du type :

$$\big[\ \text{Club},\ \text{Elo\_J}(J-1),\ \text{Elo\_J}J,\ \Delta\ \big],$$

trié par Elo après J.

— **Enchaînement.** Les probabilités de la journée $J+1$ sont recalculées à partir des Elo mis à jour après J. On répète ce cycle tout au long de la saison.

Les différents pronostics seront à retrouver avant chaque journée sur : page X de l'ane catalan du top14.