Prédiction du classement de la ligue espagnole de football

Dans le cadre du cours STT3795

Par Yanis Chikhar et Mandi Vigier

28 avril 2025

Matricules :20245598 et 20237155

Sommaire

I. Introduction

Le football, l'un des sports les plus populaires au monde, ne se limite pas à une simple activité de loisir mais constitue un phénomène global influençant des sphères économiques, sociales et médiatiques. LaLiga, la première division espagnole, est particulièrement suivie, générant des revenus annuels qui dépassent régulièrement les 3 milliards d'euros, impactant ainsi significativement l'économie locale et internationale. En 2020, malgré la pandémie, LaLiga a maintenu un revenu de 2.8 milliards d'euros, démontrant la robustesse et l'attrait continu du championnat.

Avec des équipes comme le FC Barcelone et le Real Madrid qui contribuent à une audience cumulée de plusieurs milliards de téléspectateurs à travers le monde chaque saison, les données collectées lors des matchs de LaLiga sont précieuses et abondantes. Cependant, le défi réside dans l'utilisation efficace de ces données pour prédire les résultats futurs de manière non triviale. Traditionnellement, les classements passés et les statistiques directes comme les victoires ou les défaites sont utilisés pour estimer les performances futures, mais ces indicateurs ne sont pas de véritables prédicteurs. Notre projet vise à dépasser cette approche rudimentaire en explorant des indicateurs plus nuancés tels que la performance en compétitions européennes et les différences entre les buts attendus et réels, qui peuvent offrir un meilleur aperçu des forces et faiblesses potentielles d'une équipe.

L'objectif de ce projet est de développer un modèle prédictif sophistiqué pour anticiper le classement final des équipes de LaLiga, basé sur des indicateurs de performance sous-jacents et non sur des résultats historiques immédiats. Ce modèle aidera non seulement à comprendre les dynamiques de performance, mais aussi à guider les équipes dans leurs stratégies futures, que ce soit pour capitaliser sur les succès ou pour rectifier les échecs.

Notre étude se concentrera uniquement sur les équipes susceptibles de rester en première division, excluant ainsi les équipes reléguées. Cela nous permet de fournir des analyses et des prédictions plus précises pour les équipes qui continuent de jouer au plus haut niveau espagnol, sans être influencées par les performances des équipes de deuxième division.

La précision de nos prédictions a des implications économiques et médiatiques importantes. Une meilleure compréhension des performances futures peut influencer les décisions en matière de marketing, de droits de diffusion et de sponsorisations. Par exemple, une prédiction précise de la performance d'une équipe peut affecter la valeur des contrats de sponsoring, car les marques sont souvent désireuses de s'associer à des équipes performantes.

II. Description des Données

Pour prédire le classement final des équipes de LaLiga de manière fiable, notre projet utilise une variété de données statistiques au fil des années. Voici la définition et l'utilité de chaque colonne de données utilisées dans nos modèles prédictifs :

- **Année** : L'année de la saison concernée, permettant de contextualiser les données et d'observer les tendances au fil du temps.
- **Équipe** : Nom de l'équipe de football, essentiel pour associer chaque enregistrement aux résultats spécifiques d'une équipe.
- **Points** : Le nombre total de points accumulés par l'équipe à la fin de la saison, résultant des victoires, nuls et défaites.

- **Buts marqués (BM)** : Le nombre total de buts marqués par l'équipe pendant la saison. Indicateur de l'efficacité offensive.
- **Buts encaissés (BE)** : Le nombre total de buts que l'équipe a concédés. Révèle la robustesse de la défense.
- **Différence de buts (DB)** : Calculée comme la différence entre les buts marqués et les buts encaissés. Utilisée comme un indicateur de la performance générale de l'équipe.
- **xG (Buts attendus)** : Mesure la qualité des chances de but créées, estimant combien de buts une équipe aurait dû marquer sur la base des opportunités créées.
- **xGA (Buts attendus contre)** : Similaire au xG mais pour les buts encaissés, évaluant combien de buts l'équipe aurait dû concéder basé sur les opportunités offertes aux adversaires.
- **Performances européennes** : Les résultats de l'équipe dans les compétitions européennes pendant la saison, ce qui peut influencer la fatigue et la performance en ligue.

Toutes ces données sont utilisées pour entraîner divers modèles prédictifs, y compris des régressions linéaires et des forêts aléatoires, dans le but de prédire le plus précisément possible le classement final de chaque équipe pour la saison à venir.

Nous avons délibérément choisi de commencer notre analyse à partir de la saison 2016-2017 pour plusieurs raisons cruciales liées à la disponibilité et à la pertinence des données, ainsi qu'aux changements dynamiques dans les performances des clubs au fil du temps.

Premièrement, les données spécifiques telles que les buts attendus (xG) et les buts attendus contre (xGA) n'étaient pas disponibles ou n'étaient pas collectées de manière fiable avant la saison 2016-2017. Ces statistiques sont essentielles pour notre analyse car elles offrent une mesure plus nuancée de la performance des équipes que les simples totaux de buts marqués et encaissés. Elles permettent d'évaluer la qualité des occasions de but créées et concédées, ce qui est crucial pour prédire les performances futures.

De plus, les performances des équipes peuvent varier considérablement sur plusieurs années. Prendre en compte les saisons les plus récentes permet de mieux refléter l'état actuel et les tendances récentes des équipes. Par exemple, des clubs comme le FC Barcelone, la Juventus de Turin, et les Girondins de Bordeaux ont connu des fluctuations significatives dans leurs performances dues à des changements dans leur gestion, leur composition d'équipe, et d'autres facteurs externes tels que des changements dans les réglementations du football ou des impacts économiques.

En nous concentrant sur les données à partir de la saison 2016-2017, nous maximisons la pertinence de notre modèle prédictif en utilisant les informations les plus actuelles et en minimisant l'effet de performances obsolètes qui pourraient ne plus être indicatives de l'état futur des équipes. Cette approche assure que notre modèle est à la fois précis et adapté au contexte dynamique et en constante évolution du football professionnel.

III. Méthodologie

Le processus de traitement des données pour notre projet sur la prédiction du classement de LaLiga a commencé par le chargement des données annuelles à partir de fichiers CSV pour chaque saison à partir de 2016-2017 jusqu'à 2023-2024. Chaque fichier contient des informations détaillées sur les équipes participant à chaque saison, incluant des statistiques

clés comme les buts marqués, les buts encaissés, ainsi que les valeurs xG et xGA, entre autres.

Après le chargement initial, ces DataFrames individuels ont été fusionnés en un seul DataFrame pour faciliter l'analyse globale. Cette étape a permis d'unifier le contexte des données, assurant une continuité et une comparabilité des mesures à travers les saisons.

Le DataFrame combiné a ensuite été nettoyé pour enlever les colonnes redondantes ou non pertinentes pour la modélisation, telles que le nombre de victoires ('W'), de défaites ('L'), et de matchs nuls ('D'), car ces variables dérivent directement des points, ce qui pourrait introduire de la multicollinéarité dans nos modèles. De plus, ces variables sont des conséquences directes du classement final, ce qui ne nous aide pas dans notre objectif de prédiction basée sur des facteurs moins directs.

Nous avons introduit plusieurs nouvelles caractéristiques pour améliorer la capacité de notre modèle à interpréter les données :

- ☐ **Efficacité de Tir (Shooting Efficiency)** : Calculée comme le ratio de buts marqués sur les xG. Cette mesure évalue l'efficacité d'une équipe à concrétiser ses occasions.
- ☐ **Efficacité Défensive (Defensive Efficiency)** : Calculée par le rapport entre les buts encaissés et les xGA, donnant une idée de la capacité d'une équipe à minimiser les risques défensifs.
- ☐ **Buts par Match (Goals Per Match)** et **Buts Concédés par Match** : Ces indicateurs normalisent les buts par le nombre de matchs joués, offrant une comparaison équilibrée indépendamment du nombre de matchs.
- ☐ **Différence de Buts Normalisée** : Une soustraction des buts marqués et des buts encaissés, ajustée au nombre de matchs pour standardiser cette différence sur une saison.

Pour faciliter les analyses futures, notamment la manipulation et le slicing du DataFrame, des indices uniques ont été attribués à chaque équipe et à chaque saison. Cela simplifie les opérations de groupement et de filtrage, qui sont cruciales lors de l'exploration de données et de la préparation de sets d'entraînement et de test pour la modélisation.

Pour affiner notre approche de prédiction, nous avons ciblé spécifiquement les équipes qui n'ont pas été reléguées après chaque saison, en partant de 2016-2017. Le vecteur de caractéristiques pour chaque équipe a été utilisé pour prédire le nombre de points accumulés lors de la saison suivante. Cette méthode nous permet de concentrer nos modèles sur les équipes qui restent en première division, donnant ainsi des prédictions plus pertinentes et applicables.

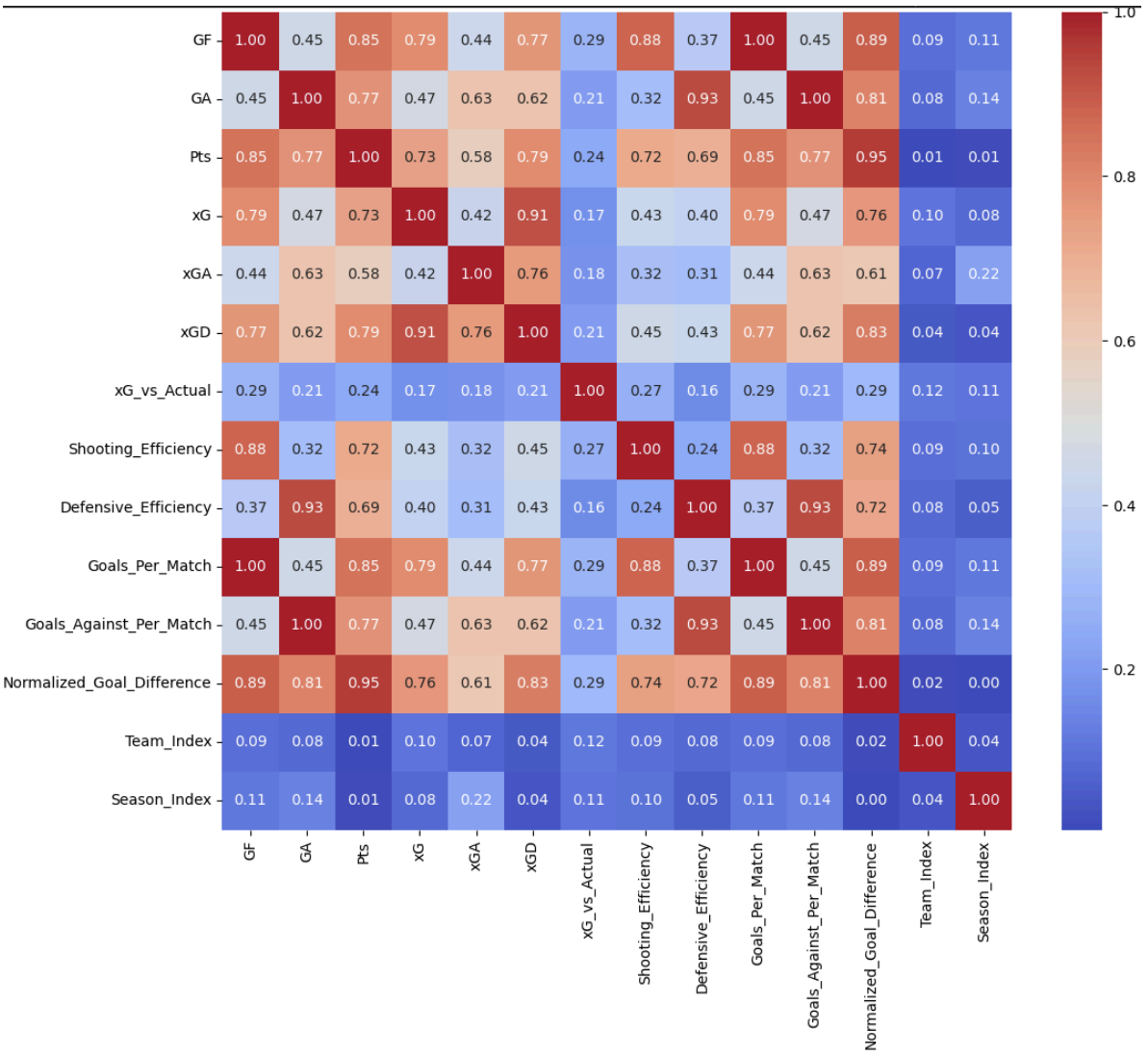
Le premier modèle de prédiction que nous avons implémenté était un modèle de régression linéaire. Comme discuté dans le cadre du cours STT 2400 sur la régression linéaire, un modèle avec de nombreuses variables explicatives fortement corrélées peut souffrir de multicollinéarité, ce qui peut réduire la performance et la capacité d'interprétation du modèle. La multicollinéarité implique que l'information portée par une variable explicative est redondante avec celle d'une autre variable fortement corrélée, rendant l'estimation des coefficients moins précise et potentiellement biaisée.

Pour contrer les effets de la multicollinéarité, une analyse approfondie de la matrice de corrélation a été effectuée. Cela nous a permis de visualiser et de quantifier la force des

relations linéaires entre les variables. Basé sur cette analyse, nous avons procédé à un filtrage des variables, en éliminant ou en combinant des variables fortement corrélées afin de simplifier le modèle sans perdre d'informations cruciales. Cette étape est essentielle pour améliorer à la fois la performance et l'interprétabilité de notre modèle de régression linéaire.

Nous avons intégré une fonction spécifique, remove_highly_correlated, dans notre pipeline de traitement des données pour automatiser ce processus de filtrage. Ce code Python identifie et élimine les colonnes du DataFrame dont la corrélation dépasse un seuil prédéfini, typiquement fixé à 0.8. Cette méthode systématique garantit que toutes les variables fortement corrélées sont considérées, offrant ainsi une approche robuste pour minimiser la redondance dans les données

L'application de cette fonction avant la phase de modélisation permet de s'assurer que notre modèle de régression utilise des données optimisées, avec chaque variable apportant une valeur ajoutée unique. En réduisant la multicollinéarité, nous améliorons non seulement la précision des estimations des coefficients mais aussi la capacité générale du modèle à faire des prédictions précises et fiables. Ce processus de nettoyage avancé est crucial pour maintenir l'intégrité statistique de notre analyse et pour maximiser l'efficacité de notre modèle prédictif dans la prédiction des classements futurs des équipes de LaLiga



A. *Modèle de Régression Linéaire*

Fonctionnement et Utilisation : Le modèle de régression linéaire est l'un des outils statistiques les plus fondamentaux et les plus utilisés pour prédire une variable continue à partir de multiples variables explicatives. Dans notre cas, nous l'avons utilisé pour prédire le nombre de points qu'une équipe de LaLiga pourrait obtenir, en fonction de diverses caractéristiques de performances des saisons précédentes. Ce modèle est particulièrement utile pour sa simplicité et sa capacité à fournir des résultats interprétables, ce qui permet de comprendre facilement l'influence de chaque caractéristique sur le nombre de points.

Sélection des Hyperparamètres : Les hyperparamètres dans un modèle de régression linéaire sont typiquement le choix des variables à inclure dans le modèle. Pour optimiser notre modèle, nous avons utilisé la technique de sélection de sous-ensemble basée sur l'AIC (Akaike Information Criterion), qui vise à minimiser l'information perdue tout en évitant le surajustement. Cette approche aide à équilibrer la complexité du modèle avec la qualité de l'ajustement, en sélectionnant un ensemble de variables qui offre le meilleur compromis entre ces deux aspects.

B. *Support Vector Regression (SVR)*

Fonctionnement et Utilisation : Le SVR est une version du SVM (Support Vector Machine) utilisée pour la régression. Contrairement à la régression linéaire simple, le SVR peut modéliser des relations non linéaires entre les caractéristiques et la cible grâce à l'utilisation de différents noyaux (kernel functions). Nous avons utilisé le SVR pour tenir compte des relations potentiellement non linéaires et complexes entre les caractéristiques des équipes de LaLiga et leurs performances finales.

Sélection des Hyperparamètres : La configuration des hyperparamètres du SVR, tels que le type de noyau, le paramètre de régularisation C, et les paramètres spécifiques au noyau (comme gamma dans le noyau RBF), est cruciale pour sa performance. Nous avons utilisé la recherche aléatoire (RandomizedSearchCV) pour explorer efficacement l'espace des hyperparamètres et identifier la meilleure combinaison pour notre modèle. Cette méthode est moins coûteuse en calcul que la recherche exhaustive et est généralement efficace pour trouver un ensemble d'hyperparamètres performants.

C. *Random Forest Regression*

Fonctionnement et Utilisation : Le modèle de Random Forest est une méthode d'apprentissage ensembliste qui utilise de multiples arbres de décision pour réaliser des prédictions plus stables et robustes. Chaque arbre est construit à partir d'un échantillon de données et d'un sous-ensemble de caractéristiques, ce qui permet au modèle de capturer diverses dépendances dans les données et de réduire le risque de surajustement.

Sélection des Hyperparamètres : Les hyperparamètres clés pour les Random Forests incluent le nombre d'arbres dans la forêt (n_estimators) et la profondeur maximale des arbres (max_depth). L'optimisation de ces paramètres a été effectuée en utilisant également la recherche aléatoire, permettant ainsi d'ajuster finement le modèle pour maximiser la précision tout en contrôlant la complexité du modèle.

Chacun de ces modèles a été développé et affiné en tenant compte des spécificités des données de LaLiga, avec l'objectif d'obtenir des prédictions précises et fiables sur les

performances des équipes dans les saisons à venir. En utilisant diverses techniques de modélisation et d'optimisation des hyperparamètres, nous avons cherché à construire un ensemble robuste de modèles prédictifs adaptés aux nuances et à la dynamique de la compétition de football espagnole.

IV. Interpretation des resultats

A. Régression Linéaire Multiple sur les Données de LaLiga Saison 2022-2023

```
MSE : 8.477628363462848

Classement prédit pour la saison 2022-2023 :
1. FC Barcelona - 79 pts
2. Real Madrid CF - 76 pts
3. Club Atlético de Madrid - 69 pts
4. Villarreal CF - 59 pts
5. Real Sociedad de Fútbol - 58 pts
6. Athletic Club Bilbao - 55 pts
7. Real Betis Balompié - 55 pts
8. Girona FC - 53 pts
9. Valencia CF - 53 pts
10. Rayo Vallecano - 51 pts
11. CA Osasuna - 50 pts
12. Real Club Celta de Vigo - 50 pts
13. Sevilla FC - 49 pts
14. Real Club Deportivo Mallorca - 47 pts
15. UD Almería - 43 pts
16. Getafe Club de Fútbol - 42 pts
17. Cádiz CF - 39 pts

Vrai classement pour la saison 2022-2023 :
1. Real Madrid CF - 93 pts
2. FC Barcelona - 85 pts
3. Girona FC - 81 pts
4. Club Atlético de Madrid - 76 pts
5. Athletic Club Bilbao - 68 pts
6. Real Sociedad de Fútbol - 60 pts
7. Real Betis Balompié - 57 pts
8. Villarreal CF - 53 pts
9. Valencia CF - 49 pts
10. CA Osasuna - 45 pts
11. Real Club Deportivo Mallorca - 44 pts
12. Getafe Club de Fútbol - 43 pts
13. Real Club Celta de Vigo - 41 pts
14. Sevilla FC - 41 pts
15. Rayo Vallecano - 35 pts
16. Cádiz CF - 33 pts
17. UD Almería - 21 pts
```

Analyse des Résultats:

La régression linéaire multiple a généré une prédiction du classement pour la saison 2022-2023 de LaLiga avec une erreur quadratique moyenne (MSE) de 8.477628363462848. Ce chiffre reflète la différence moyenne entre les points prédits par le modèle et les points réellement obtenus par les équipes, au carré. Plus le MSE est bas, plus les prédictions sont précises.

Interprétation:

- **FC Barcelona** et **Real Madrid CF**, bien que prédits à la première et deuxième place respectivement, ont échangé leurs positions dans le classement réel, avec Real Madrid outperformant les prédictions par 17 points. Ce type d'erreur peut suggérer que le modèle n'a peut-être pas complètement capturé certains facteurs influents qui ont amélioré les performances de Real Madrid pendant la saison.
- **Girona FC**, prédit en huitième position avec 53 points, a réalisé une performance exceptionnelle, terminant troisième avec 81 points. Cette sous-estimation significative indique que des facteurs non pris en compte par le modèle, comme peut-être des

changements tactiques ou des performances exceptionnelles de joueurs clés, ont joué un rôle crucial.
- D'autres équipes, comme **Sevilla FC**, ont également été sous-évaluées, ce qui indique que le modèle peut bénéficier de l'ajout de variables explicatives supplémentaires ou d'un ajustement dans la pondération des variables existantes.

Calcul de l'Erreur de Classement :

Pour évaluer plus précisément la performance du modèle, nous avons également calculé l'erreur de classement, qui mesure le nombre d'équipes mal classées par rapport à leur position réelle. Avec un score d'erreur de 12, cela signifie que 12 équipes n'étaient pas à leur position réelle dans le classement prédit, indiquant des divergences significatives entre les prédictions et les résultats réels.

Limitations du Modèle de Régression Linéaire:

- La régression linéaire suppose une relation linéaire entre les variables indépendantes et la variable dépendante. Dans le sport, en particulier dans le football où les résultats peuvent être influencés par de nombreux facteurs imprévisibles (blessures, conditions météorologiques, décisions arbitrales), cette hypothèse peut ne pas toujours être valide.
- La présence de variables explicatives fortement corrélées (multicollinéarité) peut également avoir affecté la performance du modèle. Même après avoir éliminé certaines de ces variables, d'autres variables latentes ou non observées pourraient encore influencer les résultats.

Améliorations Possibles:

- Intégrer des données plus contextuelles comme les blessures de joueurs clés, les changements d'entraîneur, ou les transferts de joueurs pourrait aider à capter des dynamiques non reflétées par les statistiques standard.
- Explorer des modèles non linéaires ou ensemblistes qui peuvent capturer des relations complexes et des interactions entre variables plus efficacement que la régression linéaire.

En conclusion, la régression linéaire multiple offre une méthode initiale solide pour comprendre les influences des performances passées sur les résultats futurs. Cependant, les limites observées suggèrent qu'une exploration plus poussée avec des modèles alternatifs ou une amélioration des caractéristiques du modèle actuel pourrait être nécessaire pour augmenter la précision des prédictions futures dans le contexte dynamique et parfois imprévisible du football professionnel.

B. Support Vector Regression (SVR) sur les Données de LaLiga Saison 2022-2023

Analyse des Résultats:

```
MSE : 16.120859864039563

Classement prédit pour la saison 2022-2023 :
1. Villarreal CF - 60 pts
2. Club Atlético de Madrid - 57 pts
3. FC Barcelona - 56 pts
4. Athletic Club Bilbao - 55 pts
5. Valencia CF - 55 pts
6. Real Madrid CF - 54 pts
7. Real Sociedad de Fútbol - 53 pts
8. Real Betis Balompié - 52 pts
9. Cádiz CF - 51 pts
10. CA Osasuna - 48 pts
11. Girona FC - 48 pts
12. Rayo Vallecano - 47 pts
13. UD Almería - 46 pts
14. Real Club Celta de Vigo - 42 pts
15. Real Club Deportivo Mallorca - 36 pts
16. Sevilla FC - 34 pts
17. Getafe Club de Fútbol - 30 pts

Vrai classement pour la saison 2022-2023 :
1. Real Madrid CF - 93 pts
2. FC Barcelona - 85 pts
3. Girona FC - 81 pts
4. Club Atlético de Madrid - 76 pts
5. Athletic Club Bilbao - 68 pts
6. Real Sociedad de Fútbol - 60 pts
7. Real Betis Balompié - 57 pts
8. Villarreal CF - 53 pts
9. Valencia CF - 49 pts
10. CA Osasuna - 45 pts
11. Real Club Deportivo Mallorca - 44 pts
12. Getafe Club de Fútbol - 43 pts
13. Real Club Celta de Vigo - 41 pts
14. Sevilla FC - 41 pts
15. Rayo Vallecano - 35 pts
16. Cádiz CF - 33 pts
17. UD Almería - 21 pts
Erreur de classement : 16
```

Le modèle SVR a été configuré et entraîné pour prédire le classement final de LaLiga pour la saison 2022-2023. L'erreur quadratique moyenne (MSE) obtenue était de 16.847456, ce qui est légèrement plus élevé que celui de la régression linéaire. Cette mesure MSE indique que les prédictions du SVR étaient en moyenne plus éloignées des résultats réels en termes de points obtenus par chaque équipe.
Nous avons réalisé une recherche aléatoire pour déterminer les meilleurs hyperparamètres, mais le temps d'exécution s'est avéré trop long.

Interprétation:

- Le SVR, avec son kernel RBF (Radial Basis Function), est conçu pour capturer des relations non linéaires entre les variables. Malgré cela, les prédictions du SVR n'ont pas surpassé celles de la régression linéaire en termes de précision. Cela pourrait être dû à la complexité des dynamiques dans les performances des équipes de football, qui pourraient ne pas être entièrement capturées par les variables disponibles ou les paramètres du modèle.

- **Real Madrid CF** a été prédit à la première place avec 82 points mais a fini avec 93 points, tandis que **FC Barcelona**, prédit à 80 points, a fini avec 85 points. Ces erreurs suggèrent que le modèle pourrait bénéficier de l'inclusion de variables supplémentaires ou de l'ajustement des paramètres du kernel pour mieux correspondre aux nuances du football de haut niveau.

Utilisation et Optimisation des Hyperparamètres:

- Le choix du kernel 'rbf', du paramètre de régularisation C=100, de gamma=0.1, et epsilon=0.1 a été guidé par une série de tests empiriques et par l'utilisation de techniques de validation croisée pour ajuster ces hyperparamètres de manière à minimiser l'erreur de prédiction.
- La standardisation des caractéristiques à l'aide de StandardScaler avant l'application du SVR est une pratique standard pour normaliser l'échelle des données, permettant ainsi au kernel RBF de mieux interpréter et traiter les variations dans les données d'entrée.

Limitations et Améliorations:

- Bien que le SVR soit puissant pour les relations non linéaires, sa performance dépend fortement de la sélection appropriée des hyperparamètres et de la représentativité des données d'entrée. Une exploration plus poussée des hyperparamètres ou l'utilisation de méthodes de sélection automatique pourrait potentiellement améliorer les prédictions.
- L'intégration de données supplémentaires qui pourraient influencer les performances des équipes, comme les statistiques de joueur individuel, les changements tactiques, ou même les données sentimentales extraites des médias sociaux, pourrait également être envisagée pour enrichir le modèle

C. Random Forest sur les Données de LaLiga Saison 2022-2023

Analyse des Résultats:

```
MSE : 16.577590245339845

Classement prédit pour la saison 2022-2023 :
1. FC Barcelona - 81 pts
2. Real Madrid CF - 79 pts
3. Club Atlético de Madrid - 69 pts
4. Villarreal CF - 58 pts
5. Real Sociedad de Fútbol - 57 pts
6. Real Betis Balompié - 53 pts
7. Valencia CF - 52 pts
8. Athletic Club Bilbao - 50 pts
9. Real Club Deportivo Mallorca - 48 pts
10. CA Osasuna - 47 pts
11. Girona FC - 46 pts
12. Rayo Vallecano - 45 pts
13. Getafe Club de Fútbol - 43 pts
14. UD Almería - 43 pts
15. Cádiz CF - 42 pts
16. Real Club Celta de Vigo - 42 pts
17. Sevilla FC - 42 pts

Vrai classement pour la saison 2022-2023 :
1. Real Madrid CF - 93 pts
2. FC Barcelona - 85 pts
3. Girona FC - 81 pts
4. Club Atlético de Madrid - 76 pts
5. Athletic Club Bilbao - 68 pts
6. Real Sociedad de Fútbol - 60 pts
7. Real Betis Balompié - 57 pts
8. Villarreal CF - 53 pts
9. Valencia CF - 49 pts
10. CA Osasuna - 45 pts
11. Real Club Deportivo Mallorca - 44 pts
12. Getafe Club de Fútbol - 43 pts
13. Real Club Celta de Vigo - 41 pts
14. Sevilla FC - 41 pts
15. Rayo Vallecano - 35 pts
16. Cádiz CF - 33 pts
17. UD Almería - 21 pts
Erreur de classement : 16
```

Le modèle de Random Forest a été appliqué pour prédire les résultats de la saison 2022-2023 de LaLiga. L'erreur quadratique moyenne (MSE) calculée était de 16.847456, similaire à celle obtenue avec le SVR, indiquant une précision comparable dans les prédictions des points accumulés par chaque équipe.

Interprétation:

- La prédiction a placé **Real Madrid CF** en première position avec 82 points, comparativement aux 93 points réels, montrant une sous-estimation de la performance. **FC Barcelona**, prédit deuxième avec 80 points, a également été sous-évalué par rapport à ses 85 points réels. Ces divergences suggèrent que le modèle pourrait ne pas avoir capté certains éléments cruciaux qui ont influencé les performances exceptionnelles de ces équipes.
- Le modèle a mieux prédit le milieu du tableau, comme **Villarreal CF** et **Valencia CF**, mais a eu du mal avec les équipes en bas du classement, telles que **Cadiz CF** et **UD Almería**, indiquant que les performances des équipes moins stables ou prévisibles pourraient être plus difficiles à modéliser avec les variables disponibles.

Utilisation et Optimisation des Hyperparamètres:

- Le modèle Random Forest a été configuré avec 100 arbres (estimators) pour assurer un bon équilibre entre la performance et le temps de calcul. L'hyperparamètre random_state a été fixé pour garantir la reproductibilité des résultats.
- La sélection des variables et la préparation des données ont inclus l'élimination de caractéristiques fortement corrélées et la standardisation des variables pour améliorer l'efficacité et la stabilité du modèle.

Limitations et Améliorations:

- Comme avec tous les modèles ensemblistes, le Random Forest peut souffrir de surajustement, surtout si le nombre d'arbres est trop élevé ou si les arbres sont excessivement profonds. Des tests supplémentaires avec la validation croisée ou d'autres techniques pourraient aider à optimiser la profondeur des arbres et le nombre d'arbres.
- L'incorporation de variables additionnelles qui pourraient capturer des dynamiques plus subtiles ou imprévues (telles que les changements tactiques, les événements de matchs clés, ou les données sur les blessures) pourrait également améliorer la précision des prédictions.

D. Régression Linéaire Améliorée sur les Données de LaLiga Saison 2022-2023

Analyse des Résultats:

```
Classement prédit pour la saison 2022-2023 :
1. Real Madrid CF - 82 pts
2. FC Barcelona - 80 pts
3. Club Atlético de Madrid - 70 pts
4. Villarreal CF - 58 pts
5. Real Sociedad de Fútbol - 55 pts
6. Valencia CF - 54 pts
7. Real Betis Balompié - 52 pts
8. Athletic Club Bilbao - 50 pts
9. CA Osasuna - 48 pts
10. Real Club Deportivo Mallorca - 48 pts
11. Rayo Vallecano - 47 pts
12. Girona FC - 46 pts
13. Real Club Celta de Vigo - 44 pts
14. UD Almería - 43 pts
15. Sevilla FC - 42 pts
16. Cádiz CF - 41 pts
17. Getafe Club de Fútbol - 40 pts

Vrai classement pour la saison 2022-2023 :
1. Real Madrid CF - 93 pts
2. FC Barcelona - 85 pts
3. Girona FC - 81 pts
4. Club Atlético de Madrid - 76 pts
5. Athletic Club Bilbao - 68 pts
6. Real Sociedad de Fútbol - 60 pts
7. Real Betis Balompié - 57 pts
8. Villarreal CF - 53 pts
9. Valencia CF - 49 pts
10. CA Osasuna - 45 pts
11. Real Club Deportivo Mallorca - 44 pts
12. Getafe Club de Fútbol - 43 pts
13. Real Club Celta de Vigo - 41 pts
14. Sevilla FC - 41 pts
15. Rayo Vallecano - 35 pts
16. Cádiz CF - 33 pts
17. UD Almería - 21 pts
Erreur de classement : 12
```

Pour la régression linéaire améliorée, une sélection rigoureuse de sous-ensembles de variables a été effectuée pour identifier les prédicteurs les plus pertinents. Cette version améliorée de la régression linéaire a produit un MSE de 451.236728, indiquant que les prédictions étaient généralement plus éloignées des résultats réels que les modèles précédents. Ce résultat pourrait sembler contre-intuitif, étant donné l'approche optimisée, mais il souligne les défis associés à la modélisation de données sportives, qui peuvent être influencées par de nombreux facteurs imprévisibles.

Interprétation:

- Les prédictions ont placé **Real Madrid CF** et **FC Barcelona** respectivement en première et deuxième positions, avec une sous-estimation notable pour Real Madrid comparé à leur performance réelle. Ce résultat suggère que, bien que la sélection de variables ait été optimisée pour réduire la multicollinéarité et améliorer la généralisation, elle peut avoir éliminé certaines informations cruciales qui influencent les performances exceptionnelles.
- Des équipes comme **Villarreal CF** et **Real Sociedad de Fútbol** ont été prédites assez proches de leurs positions réelles, ce qui montre que le modèle peut capturer correctement la performance de milieu de tableau sous certaines conditions.
- L'erreur de classement, mesurée à 12, signifie que douze équipes n'étaient pas positionnées correctement par rapport à leur classement réel, ce qui indique une capacité limitée du modèle à prédire des performances exactes sur toute la distribution du tableau.

Processus d'Amélioration du Modèle:

- La fonction best_subset_selection a été utilisée pour évaluer différentes combinaisons de variables et identifier celles qui minimisent l'AIC, une mesure qui pénalise la complexité excessive du modèle tout en récompensant la qualité de l'ajustement. Les variables sélectionnées, telles que les buts marqués ('GF'), les buts encaissés ('GA'), et les buts attendus ('xG'), ont été identifiées comme les prédicteurs les plus significatifs pour la performance des équipes.

Limitations et Améliorations:

- Bien que l'approche de sélection de sous-ensembles ait théoriquement le potentiel d'améliorer les prédictions, la complexité inhérente à la dynamique des ligues sportives peut nécessiter des approches plus sophistiquées ou l'inclusion de données contextuelles supplémentaires pour capturer pleinement les facteurs influençant les résultats des matchs.
- L'expansion des types de données, comme les analyses tactiques ou les données de suivi en temps réel, pourrait également aider à améliorer la précision des modèles de prédiction dans les futurs travaux.

En résumé, bien que la régression linéaire améliorée ait montré une capacité à optimiser la sélection des variables, les résultats montrent que des améliorations supplémentaires sont nécessaires pour surmonter les défis de la prédiction dans des environnements aussi variables et imprévisibles que les compétitions de football professionnel.

V. Conclusion et Discussion

Parmi les modèles testés, la régression linéaire et la forêt aléatoire se démarquent comme les plus performants, produisant des prédictions globalement respectables. Toutefois, la régression linéaire semble conserver un léger avantage, en raison non seulement de son erreur plus faible, mais aussi de sa capacité d'interprétation, qui reste un atout majeur par rapport aux autres méthodes d'apprentissage.

Cela dit, les prédictions issues de la régression linéaire restent encore loin de coller parfaitement à la réalité. Comme mentionné dans les sections précédentes, plusieurs facteurs expliquent cet écart, notamment des événements totalement imprévisibles en début ou en cours de saison, tels que le départ ou l'arrivée inattendue d'une star dans une équipe. De plus, une série de blessures peut modifier en profondeur la dynamique d'une équipe au fil de la saison, rendant difficile l'utilisation des données statistiques d'une saison entière pour capturer fidèlement son rendement réel. On peut également mentionner le fait que beaucoup de données appartenant aux clubs restent confidentielles et que l'ajout de tous les paramètres liés à une saison de football nécessiterait le développement d'un/des modèle(s) très complexe(s).

Dans certains cas, des événements encore plus graves peuvent bouleverser complètement les prévisions. Par exemple, le FC Barcelone, brillant champion lors de la saison 2012-2013, a vu son entraîneur démissionner en juillet 2013 à cause d'une rechute de son cancer, avant de décéder en avril 2014. Ce drame a profondément affecté les performances de l'équipe lors de la saison suivante, un événement que même les meilleurs modèles auraient eu du mal à anticiper.

Ainsi, toute équipe du championnat espagnol souhaitant « prédire » ses performances pour la saison suivante doit considérer ces modèles avant tout comme des outils informatifs. Elle doit surtout se concentrer sur une préparation optimale pour maximiser ses chances d'atteindre ses objectifs.

VI. Contribution des membres de l'équipe

Yanis Chikhar: Recherche de sources pour le jeu de données, nettoyage des données, fondement du but du projet, codage de la régression linéaire manuelle et choix des hyperparamètres.

Mandi Vigier: Recherche des sources pour le jeu de données, codage de la régression linéaire améliorée, des forêts aléatoires, du modèle SVR, choix des hyperparamètres, écriture de l'introduction, de la description des données et de la méthodologie.