Rapport : Analyse des Performances Olympiques 2020-2024

Les Jeux Olympiques sont un événement attendu avec impatience par les passionnés de sport comme par les curieux du monde entier. Ce projet est une opportunité pour explorer les facteurs qui permettent aux pays de briller et de remporter des médailles. Le physique des athlètes, l'économie des nations, ou encore la géographie, sont autant d'éléments qui influencent les performances sportives.

En nous plongeant dans l'analyse des données des Jeux Olympiques de 2020 et 2024, nous chercherons à mieux comprendre ces facteurs de réussite. Nous chercherons à analyser les éléments qui influencent directement les performances, afin de mieux cerner ce qui peut réellement faire basculer le résultat lors de l'épreuve.

1. Méthodologie pour le traitement des données

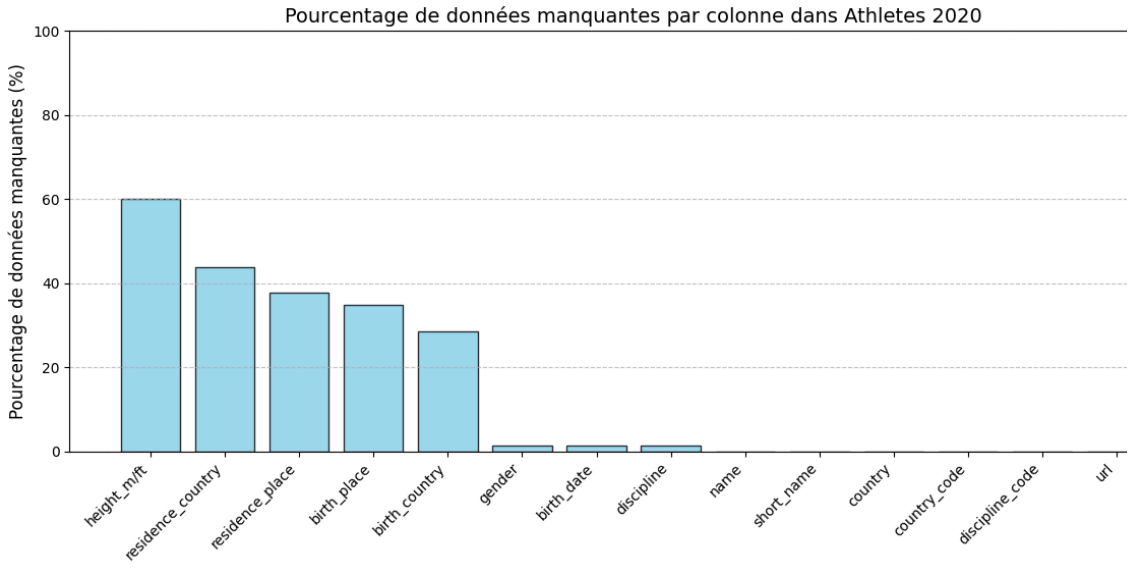
Les données utilisées dans ce projet proviennent de deux sources principales :

1. **Données fournies par le professeur :** Résultats des athlètes, médailles et informations associées des JO de Tokyo 2020 et de Paris 2024. Ces données incluent les disciplines, les pays, les genres, et d'autres caractéristiques des participants.
2. **Données externes sur le PIB :** Issues de la base de données WITS (World Integrated Trade Solution), comprenant le Produit Intérieur Brut (PIB) de différents pays pour les années 2020 et 2024. Ces données ont été récupérées pour évaluer les corrélations entre les performances olympiques et les contextes économiques.

Avant de commencer à utiliser les données nous avons dû les traiter rigoureusement pour éviter les biais et les problèmes lors de l'exécutions des modèles à venir :

- **Identification des données manquantes :**

Un diagnostic des colonnes contenant des valeurs manquantes a été effectué pour chaque dataset 2020 et 2024. Les pourcentages d'absence ont été visualisés à l'aide de graphiques afin de prioriser les colonnes à traiter.

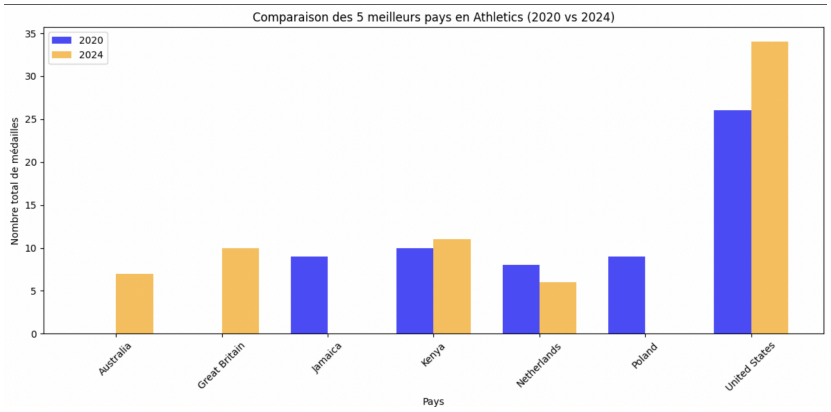


- **Stratégies adoptées pour le nettoyage :**
  - **Suppression des colonnes ou lignes non pertinentes :** Les colonnes avec un taux élevé de données manquantes ont été supprimées. Les lignes avec des valeurs critiques manquantes (ex. : gender, discipline) ont été éliminées.
  - **Imputation des valeurs manquantes :** Pour certaines colonnes numériques, comme la médiane a été utilisée pour remplacer les valeurs manquantes.
  - **Harmonisation des noms des pays :** Des corrections ont été apportées pour aligner les dénominations des pays dans différents fichiers (ex. : "United States of America" -> "United States").
  -
- **Encodage des variables catégorielles :**

Les colonnes comme gender, country, et discipline ont été transformées en valeurs numériques à l'aide d'un encodage adapté (LabelEncoder) afin d'être intégrées dans les modèles.

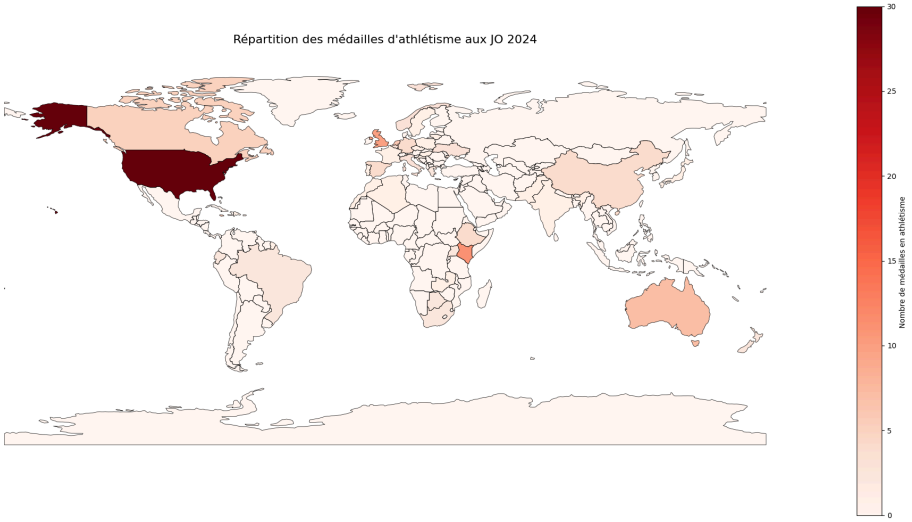
2. Analyse exploratoire

La première question à laquelle nous allons essayer de répondre est la suivante: **Quels pays sont les plus performants dans certaines disciplines spécifiques, et comment ces performances ont-elles évolué entre 2020 et 2024 ?** Pour ce faire, nous allons nous intéresser à deux disciplines en particulier: la natation et l'athlétisme. Pour ce qui est de la partie sur l'**athlétisme**, nous avons décidé d'utiliser la méthode k-mean, car le volume de données est important, rendant un CHA moins plausible. Nous avons également recherché quel pays avait eu le plus de changement en nombre de médailles entre les deux JO. De plus, cela permettra de représenter les clusters de 2020 et de 2024 et de les comparer. c'est donc ce que nous avons fait ici:

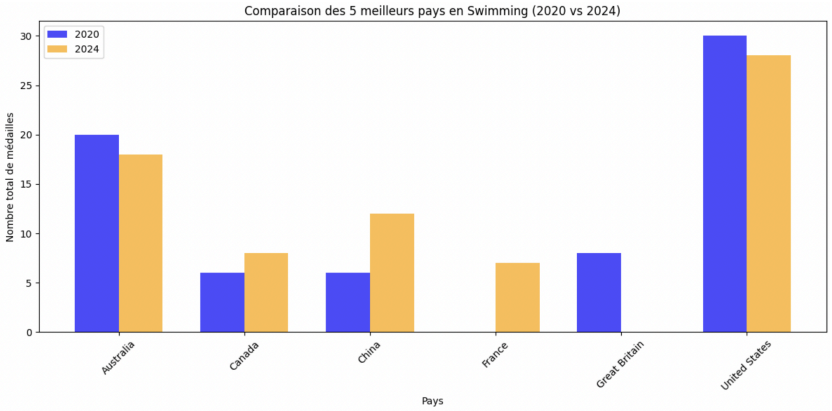


Nous avons eu certains soucis au départ car dans les jeux de données de 2020 et de 2024, certains pays ne portent pas le même nom. Nous avons par exemple United States en 2020 et United states of america en 2024. Nous avons donc dû uniformiser ces données à la main grâce à la fonction ".replace". Le problème était le même pour la Chine. J'ai donc fait mes recherches afin de connaître le nombre de médailles de la Chine dans l'année concernée afin d'identifier le nom le plus représentatif de la réalité.
Après avoir fait cela, nous trouvions toujours des résultats incohérents, proches de 2 fois plus de médailles obtenues par pays que dans les faits. On a donc fini par identifier le problème: nous ne trions pas par 'event' et donc les sports d'équipes faisaient compter plusieurs fois leurs médailles. On a donc modifié cela et avons été satisfait du résultat.
Ainsi, on voit clairement que les États-Unis dominent largement le domaine de l'athlétisme au JO et que leurs performances rapportent plus de médailles en 2024 qu'en 2020. Cela semble cohérent sachant que c'est l'une de leur discipline phare. De plus, on peut voir que le Kenya est également second chaque année,

ce qui est cohérent également sachant qu'il y a le plateau du riff au Kenya et que cette région est connue pour ses performances uniques sur les courses de moyenne et longues distances.
Pour observer de manière plus visuelle nous avons utilisé la librairie geopandas pour créer une carte qui colore proportionnellement aux nombres de médailles de chaque pays.

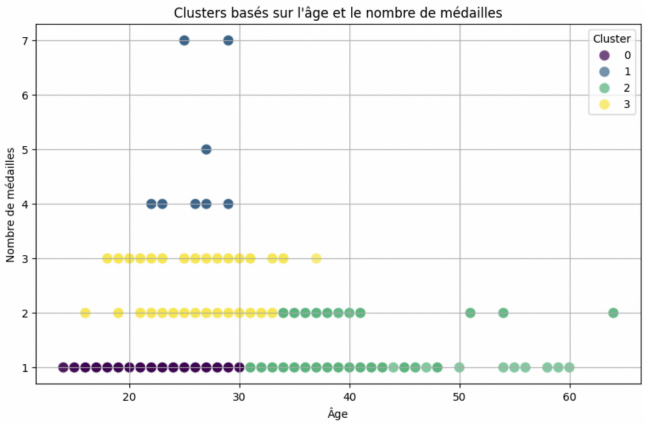


Nous avons appliqué la même méthode pour la **natation** et voici les résultats:

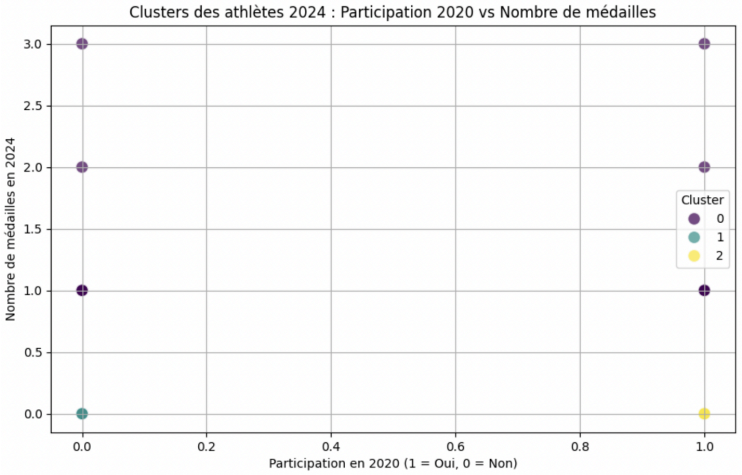


Ainsi, on voit que les États-Unis dominent largement la Natation également mais que cette fois-ci, leur performance ont légèrement été moins bonnes en 2024 par rapport à 2020. Cela est cohérent sachant qu'ils ont les meilleurs centres d'entrainement au monde notamment dans les Universités. L'Australie est également présente les 2 années et domine largement le sport avec une légère baisse du nombre de médailles en 2024, comme les États-Unis.

La question n°2 que nous avons traité est la suivante: **Quels sont les profils typiques des athlètes qui gagnent des médailles (par exemple, âge moyen, taille, expérience) ?**
Nous avons décidé d'utiliser la méthode k-mean a nouveau car le volume de données est toujours aussi important mais également car les données à étudier sont continues, ce qui signifie qu'elles sont quantifiables sur une échelle.
Nous avons commencé par regarder le lien entre les nombre de médailles obtenues et l'âge. Nous avons utilisé la méthode du coude pour identifier le nombre idéal de clusters, dans notre cas, 4. On a ensuite effectué le k-mean entre le nombre de médailles et l'âge sans tenir compte de l'année. Voici les résultats:

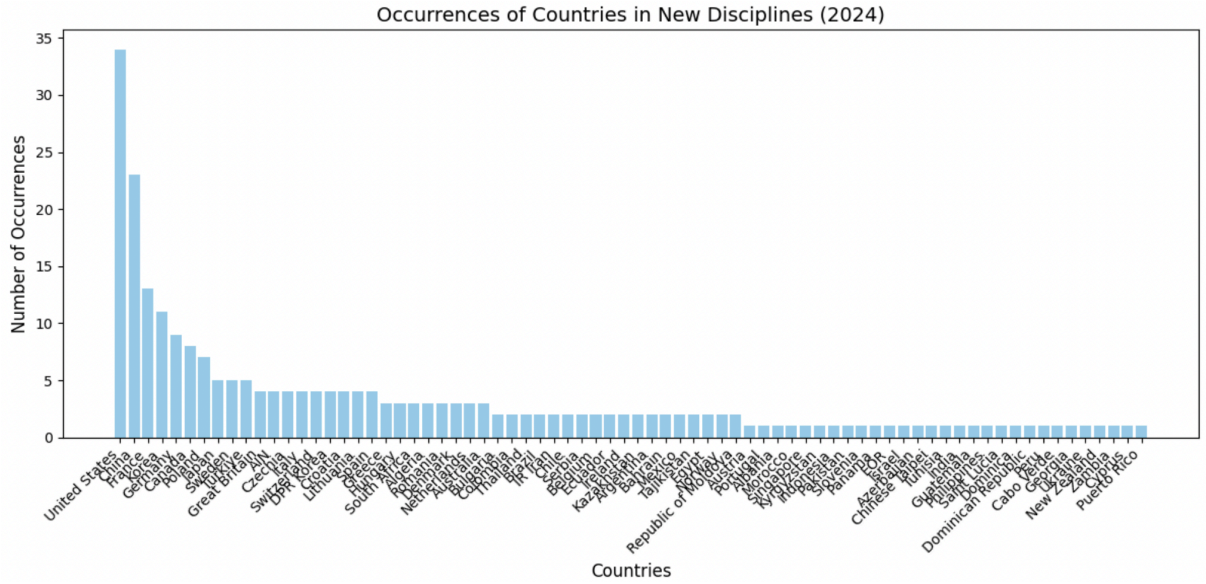


On voit que plus on vieillit moins on a de médailles, et que la meilleure tranche d'âge est entre 24 ans et 27 ans.
Nous avons ensuite voulu étudier le comportement du nombre de médailles obtenues en 2024 selon la participation ou non aux JO de 2020. On a donc à nouveau utilisé la méthode du coude pour trouver le nombre de clusters idéal (3). Voici les résultats du k-mean:



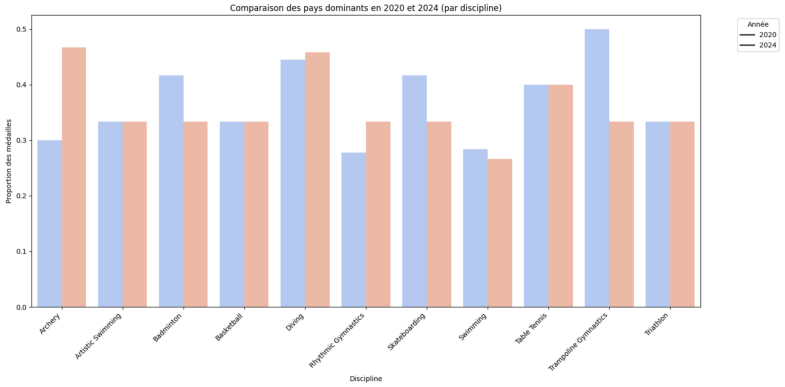
On peut donc voir que la participation en 2020 n'a pas d'impact particulier sur le nombre de médailles gagnées en 2024.

La 3ème question à laquelle nous allons répondre est la suivante: **Quels nouveaux pays ou disciplines ont gagné des médailles pour la première fois en 2024 ?**
Afin d'y répondre, nous n'utilisons ni un CHA ni un k-mean. Nous allons seulement filtrer et comparer les données, cela semble plus approprié car il n'y a pas de raison de comparer le lien entre les différentes données. Nous avons donc décidé de regarder quels pays ont gagné des médailles dans de nouvelles disciplines en 2024 par rapport à 2020. Pour ce faire, on a isolé les colonnes country et discipline, puis nous avons drop les doublons lorsque j'ai fusionné les tableaux 2020 et 2024. On a ensuite compté le nombre d'occurence de chaque pays, et voici les résultats:



On voit encore une fois les États-Unis être devant les autres pays. Le nombre est très important car nous avons seulement deux années. Si on avait regardé avec de nombreux autres JO, cela aurait été plus intéressant car nous aurions pu voir si certains pays gagnent des médailles qu'ils n'ont jamais gagnées dans leur histoire.

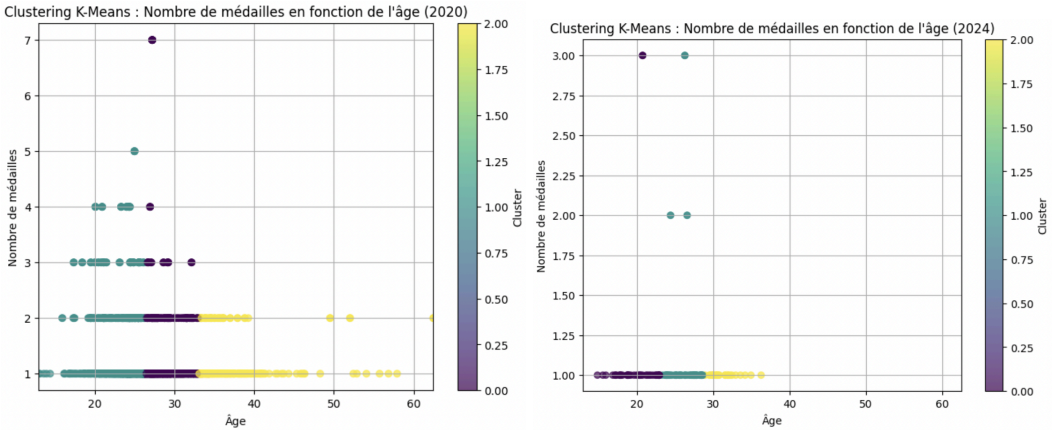
La 4ème question que nous avons traité est la suivante: **Quels sont les sports où certains pays ou continents dominent largement, et comment cette domination a-t-elle évolué entre 2020 et 2024 ?**
Afin de répondre à cette question, Nous avons fait le choix de faire des graphiques plutôt qu'un k-mean ou un CHA car encore une fois, il n'y avait pas de lien entre des variables à regarder mais plus un tri et une interprétation de données plutôt volumineuse. Nous nous sommes donc intéressés aux pays ayant gagné plus de 25% des médailles d'une même discipline par année. Pour 2020 et pour 2024 donc nous avons obtenu des graphiques. Nous avons ensuite trié les discipline apparaissant en 2020 et en 2024 avec une domination particulière d'un même pays et voici le graphique que nous avons obtenu:



On peut voir que la discipline la plus dominée et la plus constante est le plongeon. Lorsqu'on regarde le tableau disponible dans le google collab, on voit que c'est la chine qui domine largement ce domaine avec 44% des médailles..

Enfin, nous avons décidé d'éclaircir la question suivante: **Comment la moyenne d'âge des médaillés a-t-elle évolué entre 2020 et 2024 ?**
Pour répondre à ce questionnement, nous avons décidé d'adopter une approche utilisant le k-mean. En

effet, le volume de données était important et l'âge est une valeur continue donc s'utilise bien dans du k-mean. Nous avons repris le même nombre de clusters qu'à la question n°2 c'est-à-dire 3 et nous avons étudié en 2020 et en 2024 indépendamment l'âge par rapport au nombre de médaille gagné. Voici les graphiques obtenus:

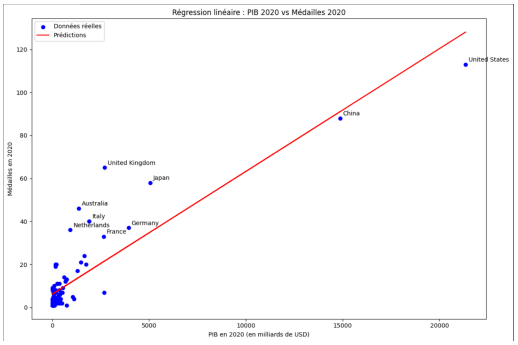


On voit que l'âge le plus cohérent en 2020 est entre 25 et 27 ans tandis que pour 2024, l'âge le plus cohérent est entre 21 et 26 ans. Ainsi, on peut voir que l'âge moyen a diminué pour 2024, ce que l'on peut remarquer avec les jeunes français ayant gagné beaucoup de médailles (Les frères Lebrun au ping pong et Léon Marchand en natation).

3. Modélisation et prédictions

4.1 Régression

Pour commencer la partie modélisation, nous avons voulu utiliser un des modèles les plus simples : une régression linéaire. Nous étions curieux de comprendre à quel point le pib d'un pays est lié au nombre de médailles qu'il obtient.
Il a d'abord fallu trouver la liste des PIB pour chaque pays. Après quelques recherches, nous avons choisi d'utiliser la liste disponible sur le site du fond monétaire international.
Nous avons ensuite nettoyé ce nouveau jeu de données (remplacer les virgules par des points, transformer en nombres…).
La fusion des datasets n'a pas été facile mais nous avons choisi comme solution d'utiliser une bibliothèque python pour associer à chaque pays son code pays. En espérant que cette bibliothèque possède une plus grande tolérance aux noms des pays (par exemple "Korea,republic of" est reconnu comme étant la corée).
Toutefois on peut voir sur le premier graphique de la régression qu'il manque la Chine. Après avoir affiché les premières lignes, on voit que le problème vient du fait que la chine est nommée "China, People's Republic of".
Comme c'est un cas isolé, j'ai remplacé le nom dans le csv manuellement, ce qui a résolu le problème.
On obtient comme résultats (plutôt bons) :
MSE : 81.01089778105997
R2 : 0.7617000587208764
Qui correspondent à ce graphique :



On calcule ensuite les résidus et on trouve les pays qui font le plus baisser le R2 :

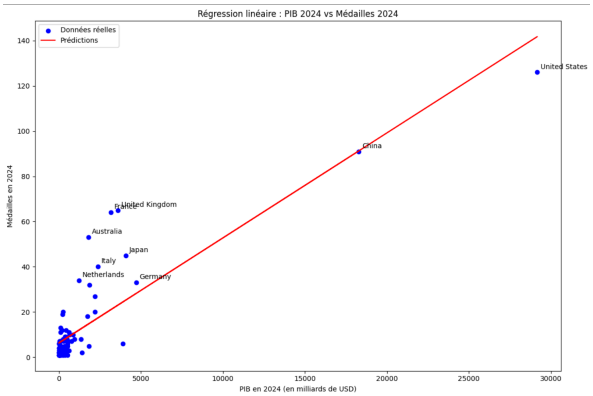
```
Pays les plus proches de la droite de régression :
   Country Name   résidus
45      Romania  0.274181
22      Georgia -0.297396
6       Belgium -0.385851
5    Azerbaijan -0.459095
51       Sweden  0.495674

Pays les plus éloignés de la droite de régression :
      Country Name    résidus
24           India -17.411829
27           Italy  21.248233
3        Australia  37.527837
21          France  40.764210
57  United Kingdom  41.443605
```

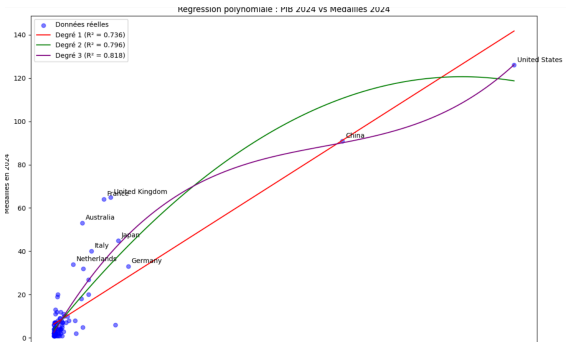
On a essayé la même régression pour 2024, et on peut considérer que les résultats de la France sont légèrement biaisés à cause de son statut de pays hôte. Si on l'enlève, on obtient les performances suivantes :

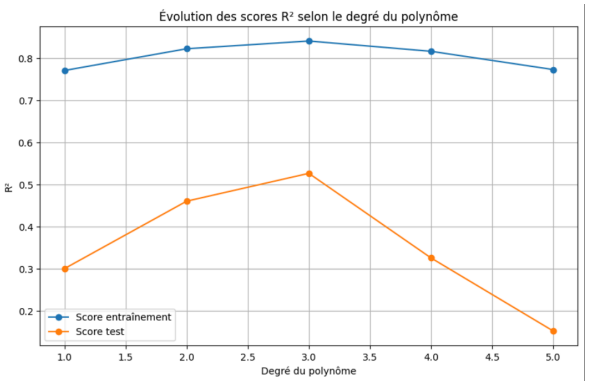
MSE : 96.50107400778737

R2 : 0.8058138800771449

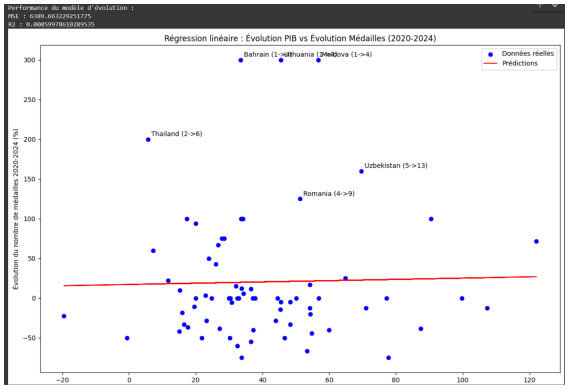


Nous avons aussi essayé une régression polynomiale, qui a donné des résultats légèrement meilleurs. En voyant la courbe, nous avons pensé que le modèle était en overfitting et voulu le vérifier, et on observe que le score en test augmente jusqu'au degré 3, ensuite le modèle devient "trop" overfit et le score en test décroît.





On peut ensuite être tenté de faire une régression avec le PIB par habitant mais cela ne marcherait pas mieux à cause de pays comme l'Andorre ou Monaco.
On choisit plutôt d'essayer avec l'évolution du pib :

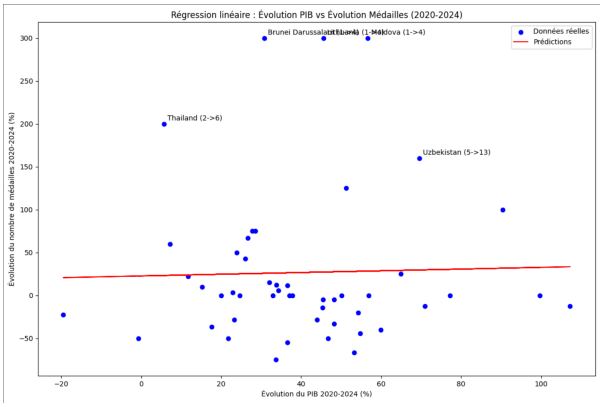


On s'aperçoit que Brunei est très haut dans les médailles. Or Brunei n'a jamais remporté de médaille olympique (mais Bahrein oui). Il y a donc un problème au niveau des codes pays. Comme on peut le voir ci-dessous, la bibliothèque python utilisée en premier avait associé aux pays leur code ISO alpha-3, qui diffère dans certains cas du code ISOC présent dans les fichiers de médailles fournis. Nous avons donc changé de bibliothèque et corrigé le problème.

Codes pays

Code CIO	Drapeau/pays	ISO 3166-1 alpha-3
BOT	Botswana	BWA
BRA	Brésil	BRA
BRN	Bahreïn	BHR
BRU	Brunei	BRN

Les résultats obtenus restent peu concluants : même si le PIB d'un pays est fortement corrélé au nombre de médailles (comme le montre les premiers graphiques), l'évolution de celui-ci ne permet de présager de l'évolution du nombre de médailles.



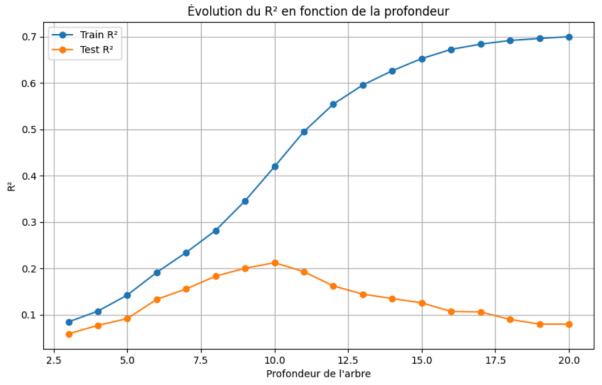
3.2 Arbres de régression

Pour la seconde partie de la modélisation, nous avons utilisé des arbres de régression pour modéliser le **nombre de médailles par pays**.
On remarque que la feature la plus importante est le nombre d'athlète par pays (on modélise le nombre de médailles qu'un pays va obtenir donc c'est normal).
Voici les résultats de ce modèle en test :
R² score : 0.696749870389781
RMSE : 14.314733847272109

On décide ensuite de tester l'overfitting du modèle, et on trouve un r2 score plutôt décent même en test. On peut aussi voir à partir de quelle profondeur l'arbre devient overfit (>=4).
Le nombre d'athlètes et de disciplines étant fortement corrélé, nous avons essayé une autre modélisation sans les deux, qui obtient un R2 très médiocre.

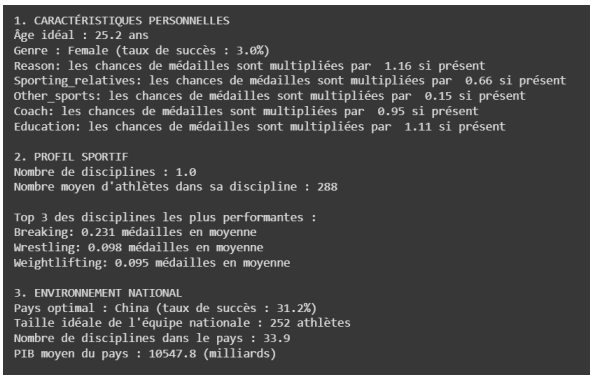
Ensuite, nous avons modélisé le **nombre de médailles d'un athlète** en entraînant le modèle sur une année et en le testant sur une autre. Or, il y a très peu de données communes donc les résultats sont mauvais.

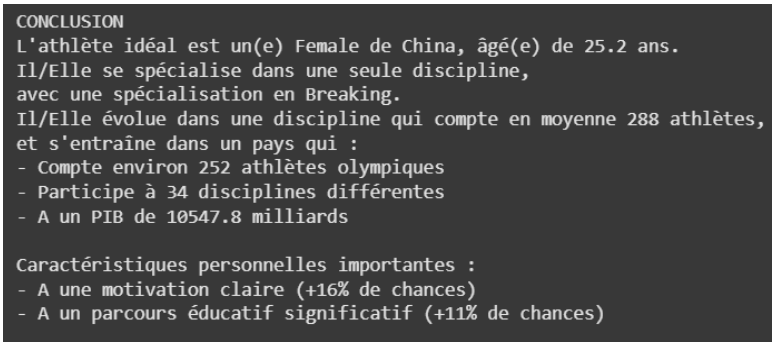
Nous avons donc terminé par une approche qui se base sur les deux jeux de données pour utiliser le plus de features possibles.
Certaines données, comme par exemple le nombre de sports pratiqués par l'athlète, ont été encodées comme suit : 1 si il pratique différents sports, et 0 sinon. Il a fallu recommencer le traitement des données car on n'avait initialement pas prévu d'utiliser ce genre de données. On garde le modèle qui obtient le meilleur score en test pour éviter qu'il soit trop overfit. On peut d'ailleurs clairement voir sur le graphique que l'overfitting devient trop important au-delà de 10 de profondeur.



A la suite de cette modélisation nous avons pu définir les caractéristiques de l'athlète idéal et répondre à notre problématique.

Voici les résultats obtenus :





On peut remarquer que le sport qui rapporte le plus de médailles est le Breakdance, probablement parce qu'il s'agit d'un sport nouveau dans lequel on retrouve peu d'athlètes. On retrouve aussi un pib très important comme on l'avait vu dans les régressions. Avoir suivi un cursus scolaire et avoir une raison de concourir sont également des bonus. Il faut cependant prendre du recul sur ces valeurs. Par exemple, on remarque que notre athlète idéal ne dispose d'aucun coach, mais cela ne veut pas dire qu'un athlète performe en moyenne moins bien avec un coach, simplement qu'il y a plus d'athlètes médaillés qui n'ont pas mentionné le fait d'être coachés plutôt que le contraire.

5. Conclusion

Cette analyse des performances olympiques de 2020 et 2024 a permis de mieux comprendre les facteurs influençant la réussite des nations et des athlètes. L'étude a révélé des corrélations notables entre les performances sportives et des variables telles que le PIB des pays, la taille des délégations et les caractéristiques individuelles des athlètes. Les analyses de clustering ont mis en lumière des regroupements significatifs parmi les athlètes et les disciplines, tandis que les modèles de régression et les arbres de décision ont confirmé l'impact du contexte économique et structurel des pays sur leur succès olympique.

Enfin, la modélisation de l'athlète parfait a permis de définir un profil optimal combinant des caractéristiques personnelles, sportives et environnementales favorisant la réussite olympique.