Université de Lille - Année 2023 - 2024

Rapport centré sur l'obtention des médailles au Jeux Olympiques

Etudiantes de Master Systèmes d'Information et Aide à la Décision (SIAD)

Réalisé par DOUAFLIA Cyrielle
LEPERCQ Louise

Encadré par VASSSEUR Corentin

Table des matières :

1. Introduction

Les Jeux Olympiques est un événement sportif mondial qui a lieu tous les 4 ans. A chaque Jeux, les meilleurs athlètes du monde entier se réunissent pour s'affronter les uns aux autres. Pour remporter les Jeux Olympiques il faut beaucoup d'entraînement et avoir une condition physique digne d'un grand sportif. Leurs caractéristiques physique est aussi l'un des points clé pour l'obtention d'une médaille olympique. Les prochains Jeux Olympiques auront lieu à Paris en 2024.

Notre projet aura pour but de connaître les meilleures conditions physiques qu'il faut avoir pour avoir le plus de chance d'obtenir une médaille olympique pour les Jeux Olympiques 2024 de Paris. Pour répondre à cette problématique, nous avons trouvé une base de données sur le site Kaggle qui répertorie un grand nombre d'athlètes ayant participé au Jeux Olympique depuis la fin du 19è siècle.

Pour connaître les meilleures caractéristiques des athlètes pour gagner une médaille olympique pendant les jeux 2024, nous allons effectuer plusieurs modèles qui prédisent qu' un individu pourrait gagner une médaille olympique.

Dans le cadre de ce rapport, nous allons établir un plan en quatres parties. La première partie aura pour but de vérifier la qualité de notre base pour effectuer des statistiques descriptives. La seconde partie consistera à réaliser une ACM pour commencer à avoir une idée des caractéristiques que doit avoir un athlète des Jeux Olympiques. Dans une troisième partie, nous allons tester plusieurs modèles pour savoir lequel prédit le mieux l'obtention de médaille. Puis dans une dernière partie nous allons conclure le projet à l'aide des résultats de la modélisation.

2. Partie collaborative (GIT)

Afin de simplifier notre collaboration, nous avons opté pour l'utilisation de Git sur GitHub. Cette solution offre une gestion efficace des versions et un hébergement pratique du code. Le but d'utiliser Git est de travailler chacun à son rythme sur son poste et de fusionner à chaque session de travail nos travaux. Nous avons créé un repository qui va héberger notre projet.

Notre Repository : Rendu_MLOPS

Nous avons fait 2 branches :

- ***La dev*** : c'est la branche dans laquelle nous faisons les développements. C'est là où nous travaillons sur de nouvelles fonctionnalités, résolvent des bugs et effectuent des améliorations. Toutes les modifications sont apportées à cette branche avant d'être intégrées à la branche principale.

- ***La main*** : Une fois que les modifications dans la branche "dev" ont été testées et sont prêtes à être intégrées dans la version principale, nous allons fusionner (merge) les changements de la branche "dev" avec la branche "main". Cela se fait après avoir assuré la stabilité du code et passé en revue les modifications.

Notre Méthodologie de Commits :

Lors de chaque commit, nous suivons une méthodologie précise pour assurer une gestion claire et efficace du développement. Cette approche repose sur deux aspects essentiels :

- ***Titre Global des Commits :***

- Chaque commit est accompagné d'un titre global succinct qui résume la nature spécifique des tâches traitées.
- Le titre fournit une vue d'ensemble rapide des modifications apportées et permet de comprendre l'objectif principal du commit.

- ***Commentaires Détaillés sur GitHub :***

- En complément du titre, nous ajoutons des commentaires détaillés directement sur la plateforme GitHub, offrant une vision des prochaines étapes à effectuer.
- Ces commentaires visent à fournir des informations contextuelles, à partager des réflexions sur la logique de code, et à indiquer des points clés pour une meilleure compréhension du processus de développement.

3. Les bonnes pratiques (package et tests unitaires)

Le code package :

Emballer son code dans des packages, c'est comme ranger ses affaires dans des boîtes bien étiquetées. Cela rend tout plus ordonné et facile à trouver. Chaque boîte (ou package) contient des morceaux spécifiques de code, comme des outils dans une trousse à outils. Cela aide à réutiliser ces outils d'un projet à l'autre. De plus, on peut attacher des instructions (tests unitaires) sur la boîte pour s'assurer que tout fonctionne correctement. En partageant ces boîtes avec d'autres développeurs, chacun peut bénéficier des outils sans avoir à tout reconstruire.

Ici, nous avons ordonné en deux fichiers de fonctions dans le répertoire mlops_functions :

- Un fichier qui traite toutes fonctions générales de notre main.ipynb
- Un fichier associé aux fonctions des statistiques

Également, dans ce répertoire, vous retrouverez un fichier unit_test_functions.py pour les tests unitaires.

Les tests unitaires :

Les tests unitaires sont cruciaux en développement logiciel. Ils détectent rapidement les erreurs, facilitent la maintenance en évitant l'introduction de nouveaux problèmes lors de modifications, et agissent comme une documentation instantanée pour les développeurs. En assurant le bon fonctionnement de chaque composant, ils contribuent à garantir la qualité, la fiabilité et la stabilité du logiciel.

- ***test_load_data*** : Ce test vérifie si la fonction load_data charge correctement les données depuis un fichier CSV et si le DataFrame résultant est identique au DataFrame initial. Cela garantit que la fonction de chargement de données fonctionne correctement et produit les résultats attendus.

- ***test_filter_summer_season*** : Ce test s'assure que la fonction filter_summer_season filtre correctement les données pour inclure uniquement les événements liés à la saison estivale. La vérification de la forme du DataFrame résultant garantit que le filtre a produit le nombre attendu de lignes.

- ***test_create_age_classes et test_create_height_classes*** : Ces tests vérifient si les fonctions create_age_classes et create_height_classes produisent correctement des classes d'âge et de taille respectivement. La vérification des modalités uniques générées garantit que la fonctionnalité de création de classes fonctionne correctement.

Pour lancer le fichier unit_test_functions.py, il faut se mettre sur le répertoire dans lequel le fichier se trouve et lancer la commande suivante :

python unit_test_functions.py

4. Etudes des données

4.1. Présentation du fichier d'étude

Pour réaliser notre étude, nous avons une base de données dans lesquelles nous retrouvons 271 116 observations et 15 variables. Elles vont nous permettre d'avoir des informations sur l'athlète tel que leurs noms, leurs sexes, leurs poids, leurs tailles puis le pays qu'ils représentent. Ensuite, nous avons des informations concernant les jeux olympiques tels que l'année des jeux, la saison des jeux, la ville qui reçoit les jeux. Puis, nous avons les informations relatives aux performances de l'athlète tel que le sport qu'il pratique, la discipline, et pour finir s'il a reçu une médaille. Pour simplifier notre analyse, nous allons nous concentrer seulement sur les Jeux Olympique d'été soit 222 552 observations.

4.2. Qualité des données

4.2.1. Valeurs manquantes

Les valeurs manquantes sont des valeurs qui ne sont pas renseignées dans notre base de données. Elles peuvent être dues aux non renseignements de l'information ou à des oublis de saisie dans la base de données. Nous avons répertorié les valeurs manquantes grâce au langage python. Nous avons identifié que les valeurs manquantes sont indiquées par "NA" dans notre base de données. Nous avons donc compté le nombre de "NA" par variables.

Tableau : Valeurs manquantes dans notre base de données

Variables	Type de variable	Nombre de valeurs manquantes	% de valeurs manquantes
Age	NUMÉRIQUE	9 189	4,12 %
Height	NUMÉRIQUE	51 857	23,3%
Weight	NUMÉRIQUE	53 854	24,19%
Medal	CATÉGORIELLE	188 464	84,68%

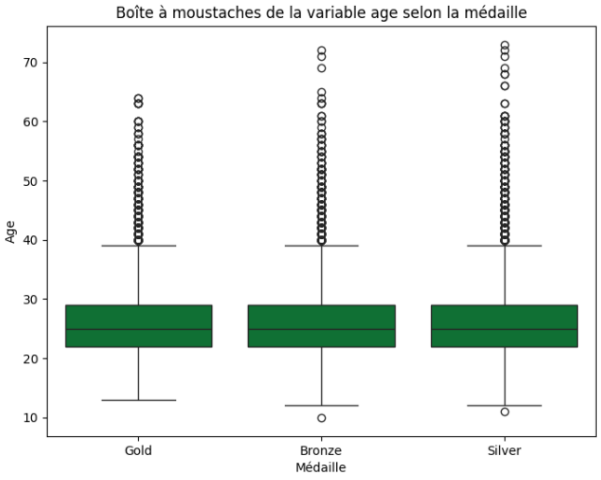
<u>Remarque</u> : Le taux de valeur manquante pour la variable "Medal" est normal car si l'individu n'a pas de médaille, cela est représenté comme une valeur manquante.

Nous avons décidé de ne pas traiter immédiatement les valeurs manquantes, puisque nous allons ajouté un filtre sur les données par la suite, et ce taux de valeurs manquantes sera moindre.

4.2.2. Valeurs aberrantes

Les valeurs aberrantes sont des valeurs qui ont une tendance différente des autres valeurs. Elles peuvent apparaître dans les bases de données suite à des erreurs de mesures, des erreurs de saisie. Pour identifier les valeurs aberrantes, nous avons réalisé des boîtes à moustaches. En effet, les boîtes à moustaches sont des représentations graphiques permettant de repérer facilement les valeurs aberrantes, si une valeur se trouve hors de la boîte à moustache, elle sera considérée comme valeurs aberrantes.

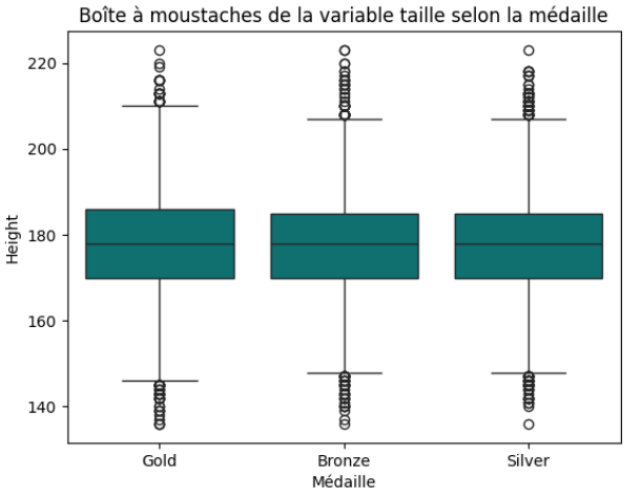
<u>Graphique</u> : valeur aberrante pour la variable âge



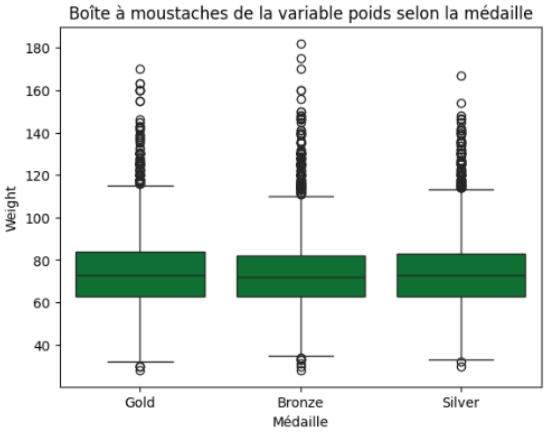
Nous pouvons voir grâce à ces boîtes à moustache qu'il existe des valeurs aberrantes pour l'âge. En effet, pour les trois types de médailles, les athlètes doivent avoir entre 11 et 40 ans pour que les données ne soient pas considérées comme aberrantes.

Graphique : valeur aberrante pour la variable taille

Nous pouvons voir que la variable taille possède des valeurs aberrantes. Les athlètes ayant une médaille ont une taille comprise entre 145 cm et 210 cm.



Graphique : Valeur aberrante pour la variable poids



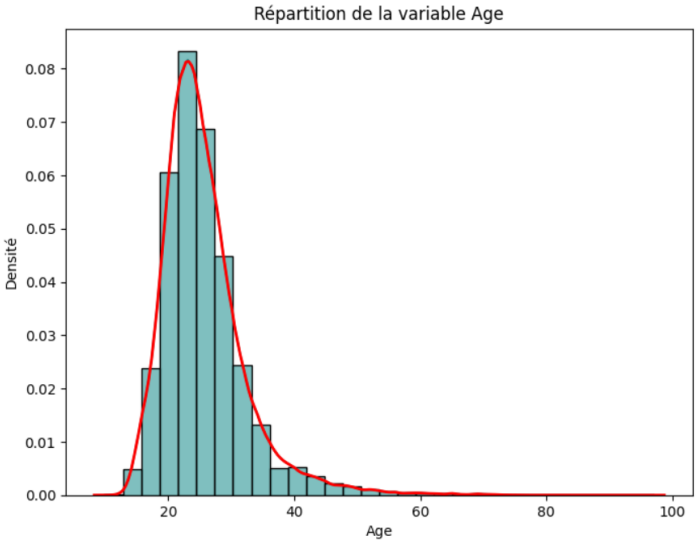
Nous pouvons constater que la variable poids possède des valeurs aberrantes. Les athlètes qui ont une médaille ont un poids entre 35 et 118 kg.

Pour l'ensemble des valeurs aberrantes identifiées, nous allons par suite réaliser des regroupements de modalités. Ceci permettra que les données ne soient plus aberrantes car les athlètes feront partie d'un groupe et n'auront plus une modalité à lui seul.

4.2.3. Analyse variable quanti pour regroupement

Nous avons décidé de faire une analyse des variables quantitatives puisque ces variables ont généralement besoin de regroupements. En effet, le fait de créer des regroupements pour les variables quantitatives va permettre de réduire le nombre de modalités pour la variable. Ceci va également améliorer l'analyse pour la suite. Pour réaliser ces regroupements, nous avons regardé la répartition de chaque variable quantitative.

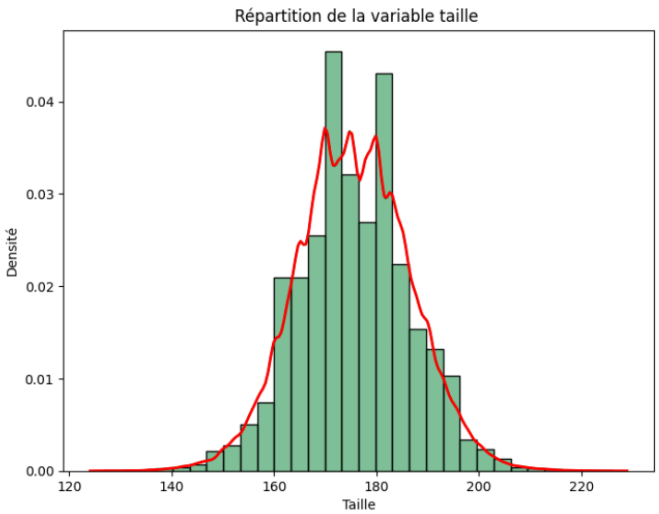
- ***Variable Age :***



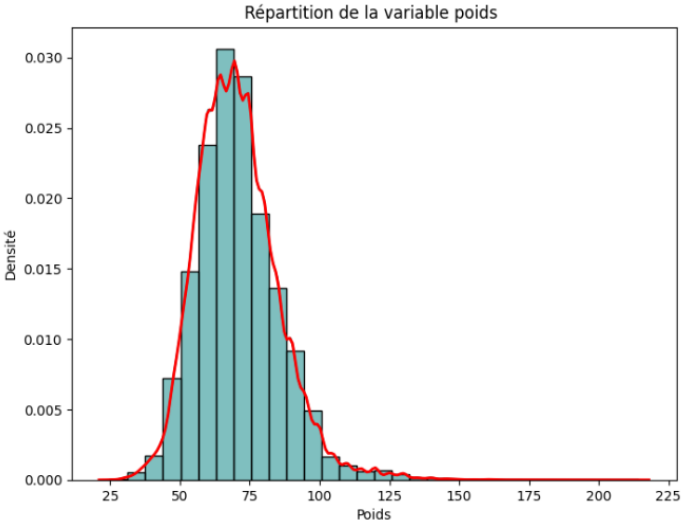
Classe âge	Nombre d'athlètes	Répartition en %
< 21 ans	39813	18,65 %
21 - 24 ans	67092	31,44 %
25 - 32 ans	81445	38,17 %
> 32 ans	25013	11,72 %

- ***Variable Taille :***

Classe taille	Nombre d'athlètes	Répartition en %
< 165 cm	25502	14,94 %
165 - 172 cm	42630	24,97 %
173 - 185 cm	72493	42,46 %
> 185 cm	30070	17,61%

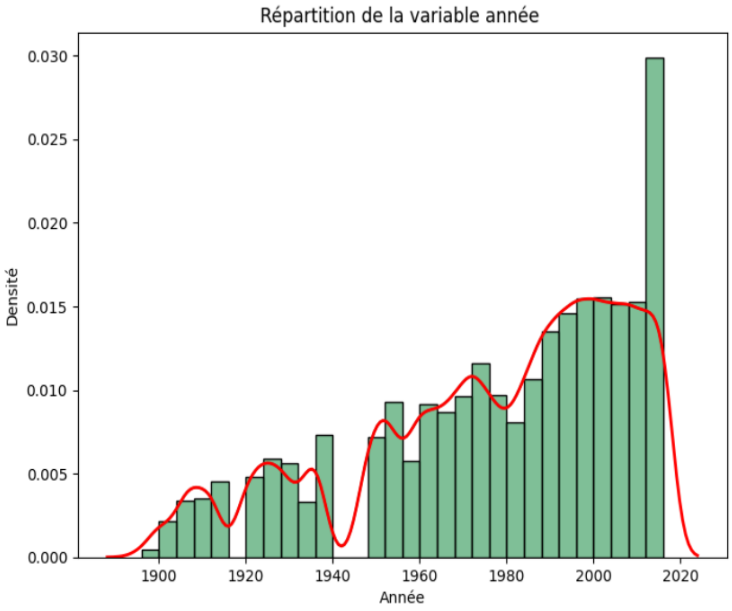


- ***Variable Poids :***



Classe poids	Nombre d'athlètes	Répartition en %
< 65 kg	60713	35,98 %
65 - 73 kg	43196	25,60 %
74 - 80 kg	28179	16,7 %
> 80 kg	36610	21,70 %

- ***Variable Année :***



Nous pouvons voir que la répartition des années n'est pas lisse dans le temps. En effet, nous pouvons remarquer deux trous, un correspondant aux années se situant entre 1914 et 1918. Ce trou peut correspondre à la période de la 1ère guerre mondiale. Puis un second trou correspondant aux années se situant entre 1939 et 1945. Ce second trou peut correspondre à la période de la 2nde guerre mondiale.

Au vu de la répartition des années, nous avons décidé de réaliser notre analyse sur les Jeux Olympiques qui ont lieu au 21ème siècle, soit les années supérieur ou égale à 2000.

4.2.4. Suppression de variables

Dans notre base de données, nous avons des variables qui sont redondantes. Effectivement, la variable “teams” et “NOC” sont deux variables qui donnent le pays représenté par l’athlète. Plus précisément, la variable “teams” est la variable avec le nom complet du pays représenté, et la variable “NOC” est l’acronyme du pays représenté, par exemple si le pays représenté est la France, la variable “teams” prendra la valeur “France” et la variable “NOC” prendra la valeur “FRA”. Nous avons donc décidé de supprimer la variable “teams” de notre base de données.

De plus, les variables “Games”, “year” et “season” sont trois variables qui donnent la même information. La variable “games” fait référence à l'année des Jeux Olympiques, suivie de la saison pendant laquelle les jeux ont lieu. Les variables “year” et “season” sont donc les mêmes informations mais en deux variables distinctes. Nous avons donc décidé de garder la variable “games”.

Comme nous avons décidé de faire notre analyse sur les Jeux d’été ayant eu lieu après les années 2000, nous allons revoir les données manquantes dont nous avons parlé dans la sous partie associée. En effet, nous n’avons pas pris de décisions puisqu’elles représentaient une certaine partie des observations. Maintenant que nous avons filtré nos données sur les JO d’été après les années 2000, nous allons rediscuter de la part qu’elles représentent dans notre nouvelle base de données. Nous avons donc une base de 67 474 observations.

Tableau : Valeurs manquantes dans notre nouvelle base de données

Variables	Type de variable	Nombre de valeurs manquantes	% de valeurs manquantes
Classe_age	CATÉGORIELLE	3	0,004 %
Classe_height	CATÉGORIELLE	654	0,96 %
Classe_weight	CATÉGORIELLE	905	1,34 %
Medal	CATÉGORIELLE	57 457	85,15 %

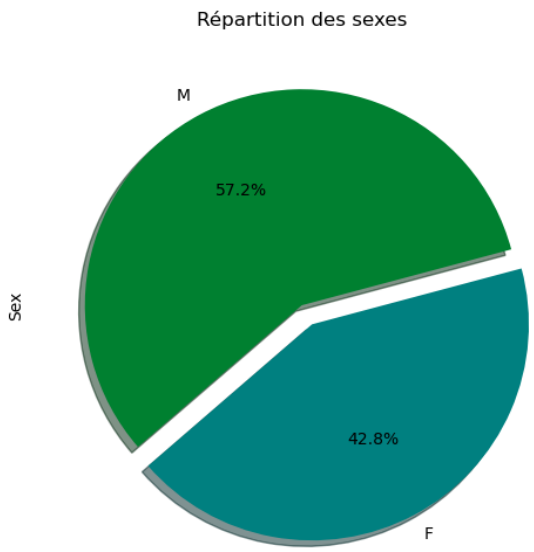
En raison d'un faible taux de valeurs manquantes, nous avons fait le choix de supprimer les valeurs manquantes pour nos 3 classes. Ainsi, notre base totale se compose de 66 450 enregistrements.

4.3. Statistiques descriptives

Les statistiques descriptives vont permettre de mieux connaître notre base de données. Nous procéderons en deux étapes. D'abord, nous analyserons la répartition des modalités des variables et ensuite nous les croiserons avec notre variable cible qui est la médaille

4.3.1. Statistiques univarié

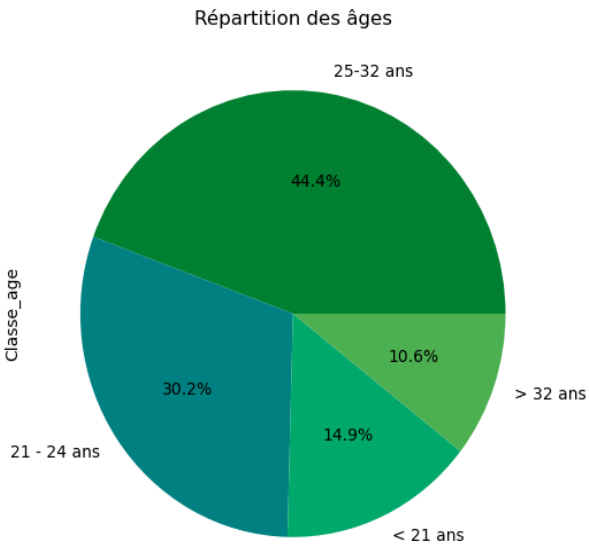
- ***Variable Sexe :***



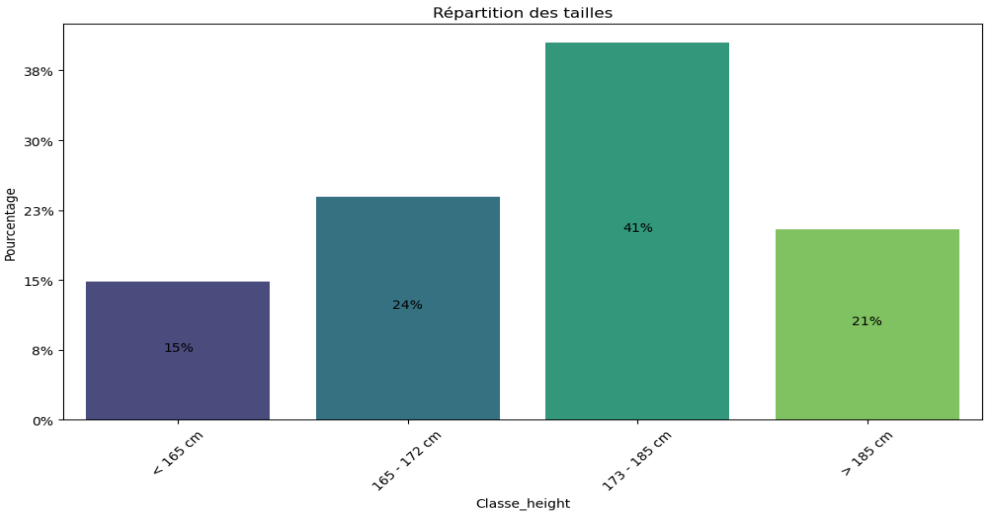
Ce graphique nous présente la répartion du sexe dans notre jeu de données. L'écart entre les hommes et les femmes n'est pas énorme, même si nous constatons plus d'homme que de femme avec un taux de 57,2%.

- ***Variable classe_age :***

Concernant la distribution de notre classe d'age, la classe la plus présente est celle des 25-32 ans avec une présence de 44,4% contre le classe la moins représentée qui est celle des plus de 32 ans avec un taux de 10,6%. D'un point de vue macro, cette tranche d'âge correspond généralement à la période où les athlètes atteignent leur pic de performance physique.



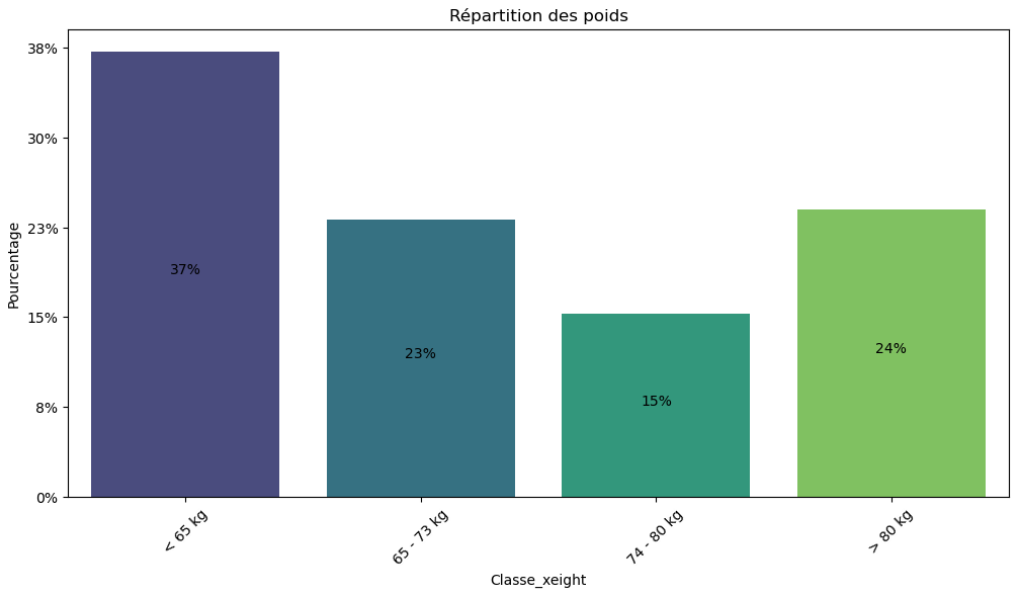
- ***Variable classe_height***



Le graphique de la répartition de catégorie de taille montre qu'un bon nombre de personnes se situe entre 173 cm et 185 cm. Cette fourchette regroupe un nombre significatif de personnes,

représentant un peu plus du double par rapport à la classe des individus mesurant moins de 165 cm.

- ***Variable Classe_weight***



Le graphique illustrant la répartition par catégorie de poids révèle des classes relativement homogènes, avec la plus grande proportion d'individus se situant dans la classe des moins de 65 kg, atteignant un taux de 37% et, la classe la moins représentée concerne les personnes pesant entre 74 et 80 kg, avec un taux de 15%.

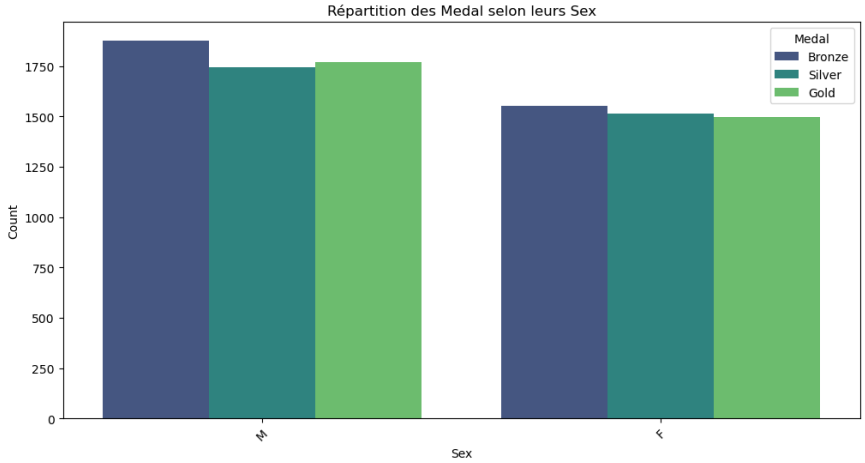
- ***Variable NOC :***

NOC	Nombre d'individus	NOC	Nombre d'individus
USA	3645	JPN	2038
AUS	2963	CAN	2018
GER	2760	ESP	1909
RUS	2738	BRA	1765
CHN	2635	UKR	1611
GBR	2344	KOR	1598
FRA	2312	NED	1383
ITA	2155	POL	1321

Ce tableau montre le nombre de sportifs olympiques pour les 16 premiers pays de notre base. Pour le podium nous retrouvons respectivement, les USA pour la première place, l'Australie pour la deuxième place et l'Allemagne en troisième position. La France se trouve en septième position avec un nombre de 2 155 participants entre 2020 et 2019.

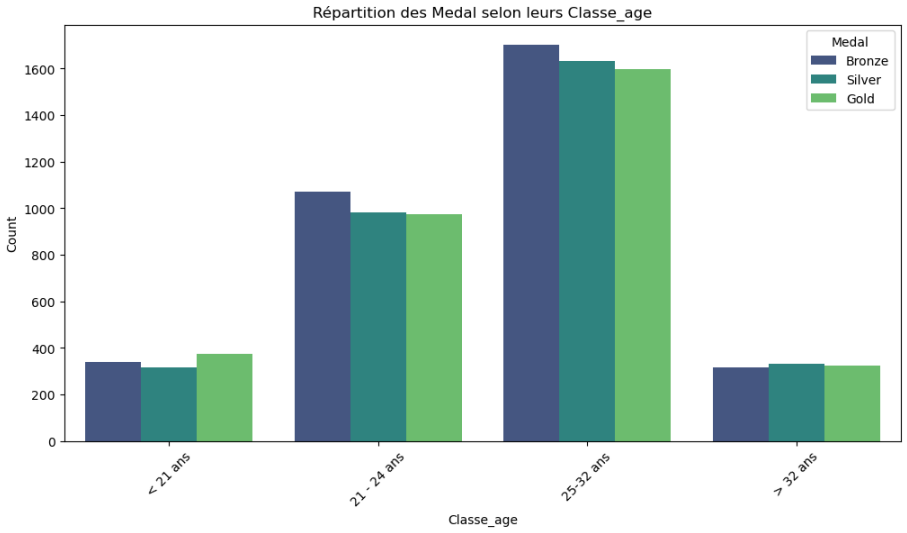
4.3.2. Statistiques bivarié

- ***Variable Medal et sex :***



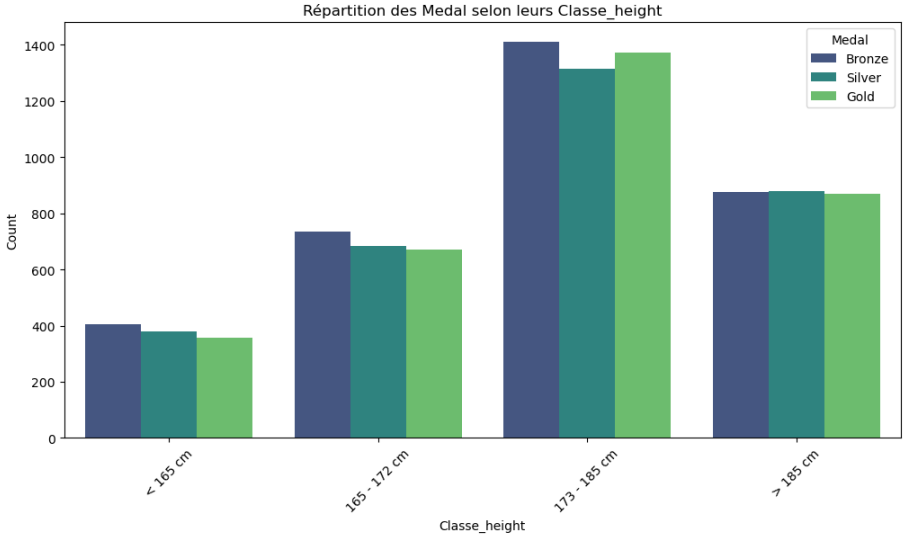
Ce graphique croise notre variable cible au sexe du sportif. Nous pouvons voir que pour les deux sexe, nous avons un nombre total de médaille de bronze légèrement supérieur que les autres médialles. En somme, nous pouvons voir que la variable sexe ne semble pas être discriminante pour la catégorie de médaille.

- ***Variable Medal et classe_age :***



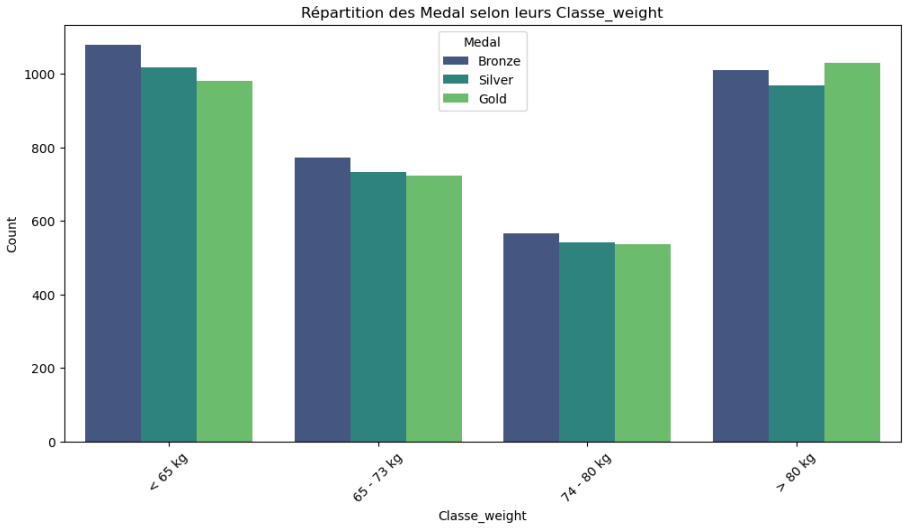
Ce graphique traite sur le lien entre les classes d'âge de sportif avec la répartition des médailles. Nous pouvons constater que les sportifs de moins de 21 ans, leur ration de médaille d'or est plus important que dans les autres catégories d'âge. Et les sportifs de plus de 32 ans étant sur le podium ont autant de chances d'avoir une médaille d'or, d'argent ou de bronze. Pour conclure, la classe d'âge semble avoir un impact sur notre variable cible.

- ***Variable Medal et classe_height :***



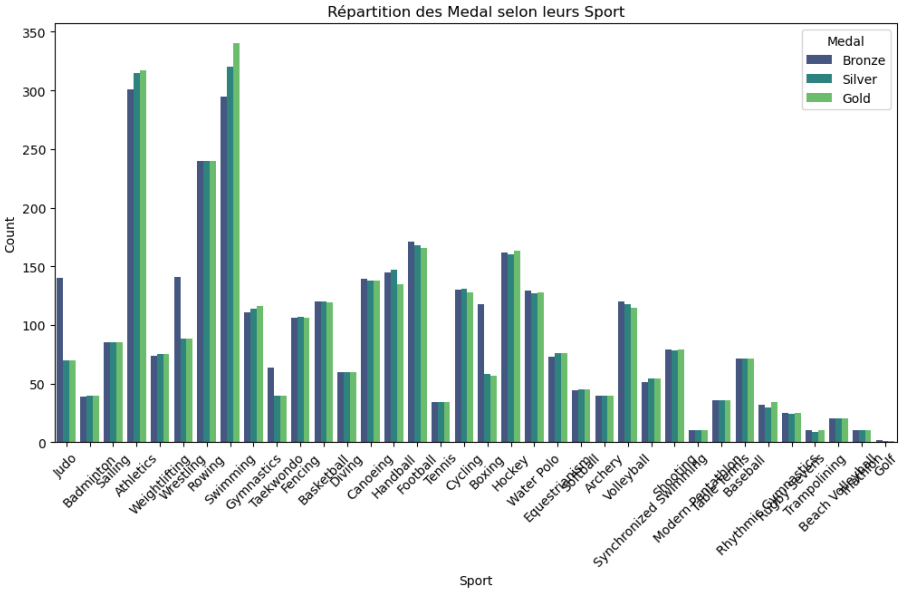
Le graphique ci-dessus présente les catégories de taille en fonction du podium. On constate une quasi parfaite égalité des chances pour les sportifs de plus de 185cm, avec un nombre total d'environ 900 pour les 3 médailles. Egalement, nous pouvons voir que les sportifs entre 173cm et 185 cm ont légèrement eu plus de médailles d'or que d'argent. En résumé, nous pouvons voir que la taille a un léger impact sur le type de médaille décerné.

- ***Variable Medal et classe_weight :***



Ce graphique met en relation la catégorie de poids du sportif au type de médaille remportée. Les sportifs ayant un meilleur ratio de médaille d'or sont ceux pesant plus de 80kg. Les trois premières catégories ont le même comportement vis-à- vis de la variable cible.

- ***Variable Medal et Sport :***



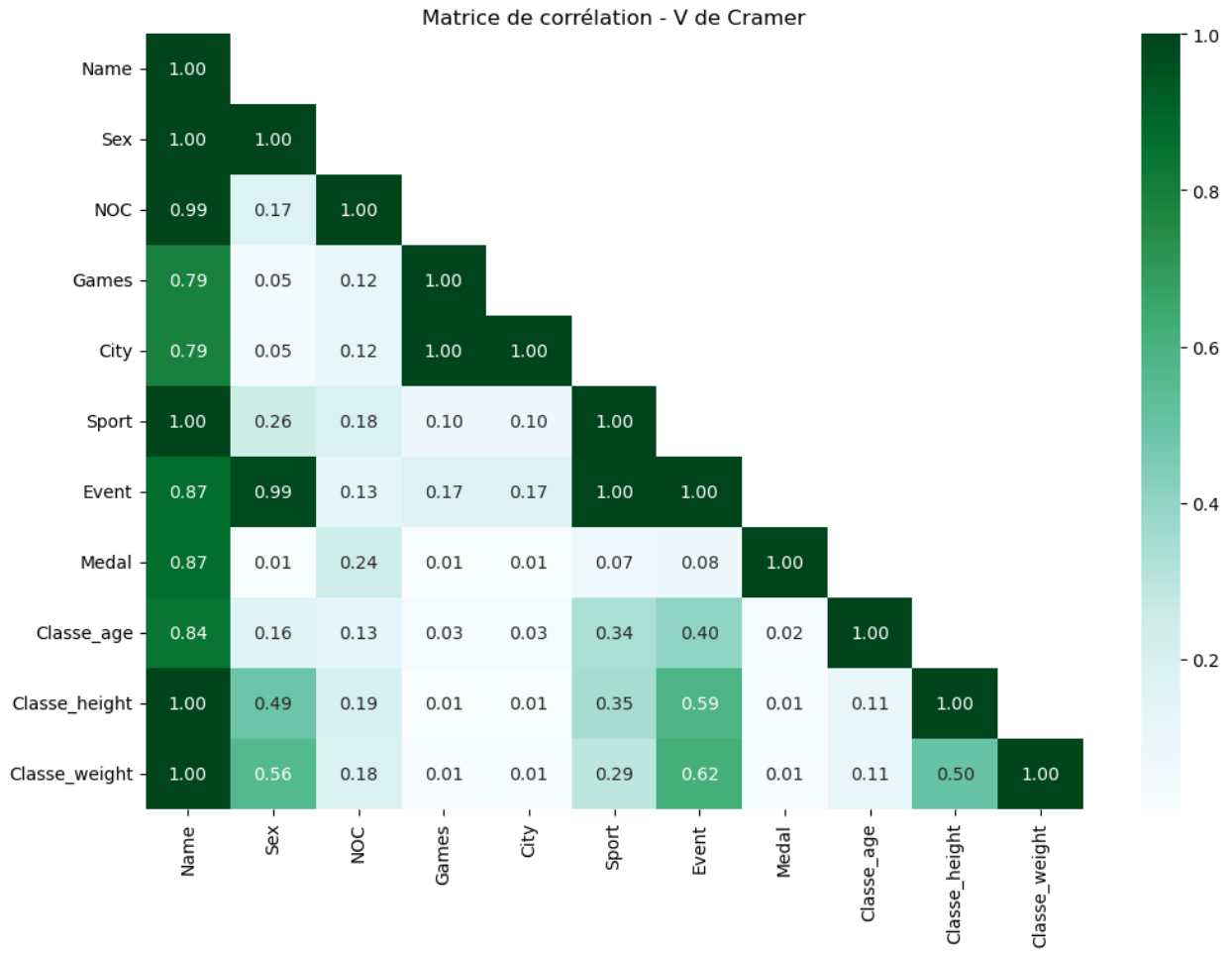
Globalement, nous pouvons voir que certains sports ont plus de catégorie et ainsi plus de médailles décernées. C'est le cas pour les sports tels que l'athlétisme ou la natation. Notamment, c'est en athlétisme ou nous pouvons constater un nombre de médaille d'or supérieur aux autres médailles.

4.3.3. Corrélation - V de Cramer

La corrélation entre les variables est un indicateur permettant de connaître les relations qu'on les différentes variables de notre base de données. En effet, nous allons pouvoir connaître les variations d'une variable associée à la variation d'une variable. Pour avoir la corrélation des

variables, nous allons calculer le V de Cramer puisque toutes nos variables sont des variables catégorielles. Plus le V de Cramer est proche de 1, plus il existe un lien fort entre les variables. A l'inverse, si le V de Cramer est proche de 0, il n'existe pas de lien entre les variables. Si les variables ont un V de Cramer supérieur à 0.5, nous allons devoir supprimer l'une des deux variables corrélés. En effet, si nous laissons l'une des deux variables, notre future analyse ne sera pas optimisée car les deux variables apporteront la même information. Nous avons réalisé un visuel pour les V de Cramer, dans celui-ci plus la couleur est foncée, plus les variables sont corrélées.

Graphique : Matrice de corrélation - V de Cramer



Au vu du résultat de la matrice de corrélation, nous pouvons dire que certaines variables sont corrélées entre elles. Nous allons donc effectuer un tableau recensant les variables corrélé entre elles, le V de Cramer, puis la décision que nous allons prendre suite à cette corrélation.

Tableau : Tableau de décision des variables corrélées

Variable 1	Variable 2	V de Cramer	Décision
Name	toutes	> 0,8	Nous allons supprimer la variable Name pour la suite de l'analyse car elle est corrélé à plus de 0.8 avec toutes les autres variable
Sex	Event	0,99	Nous allons supprimer la variable Event car elle est corrélé avec d'autre variables
Sex	Classe_height	0,49	Nous allons garder les deux variables car dans certains sports, la taille peut jouer un rôle dans l'obtention d'une médaille
Sex	Classe_weight	0,56	Nous allons garder les deux variables car dans certains sports, le poid peut jouer un rôle dans l'obtention d'une médaille
Games	City	1	Nous allons supprimer la variable City car il est plus pratique de connaître l'année de Jeux et ensuite, nous allons pouvoir rechercher où les Jeux se sont passés si besoin.
Event	Classe_height	0,59	La variable Event est déjà supprimé
Event	Classe_weight	0,62	La variable Event est déjà supprimé
Event	sport	1	La variable Event est déjà supprimé
Classe_weight	Classe_height	0,5	Les deux variables sont importantes donc, nous n'en supprimons pas

4.3.4. Choix des variables et regroupement de variable

Variables	Modalités	Regroupement
Sex	F M	

NOC	Voir annexe (cf *annexe 1*)	Asie Europe Afrique Amérique du Sud Amérique du Nord Amérique centrale Océanie
Sport	Voir annexe (cf *annexe 2*)	Combat Raquette Natation sport collectif Athlétisme Gymnastique Sur l'eau Autre sport
Medal	Gold Silver Bronze NA	
Classe_age	< 21 ans 21 - 24 ans 25 - 32 ans > 32 ans	
Classe_height	< 162 cm 165 - 172 cm 173-185 cm > 185 cm	
Classe_weight	< 65 kg 65 - 73 kg 74 - 80 kg > 80 kg	
Nombre de variable : 7	Nombre de modalité : 33	

Remarque : *Nous nous sommes rendu compte que la variable "Games" n'apportera pas d'information dans notre analyse, donc nous avons décider de ne pas l'intégrer dans la suite de notre analyse*

5. ACM

Rappelons tout d'abord que l'ACM est une application particulière de l'AFC sur un tableau représentant des individus et des variables qualitatives. De plus, elle a l'avantage de traiter des données d'enquête avec plus de 2 variables qualitatives, contrairement à l'AFC.

Cette méthode permet également de :

- Situer des individus ayant de nombreuses modalités proches
- Voir les spécificités, les modalités remarquables
- Rapprocher des individus ayant une modalité rare en commun

Pour projeter l'ACM de notre base, nous avons dû traiter les valeurs manquantes de notre variable cible. Ainsi, nous avons remplacé les NA par "pas-de-médaille" afin de sortir une catégorie supplémentaire et facile à la lecture du graphique.

5.1. Valeurs propres

Les valeurs propres aussi appelé λ sont des indicateurs de notre ACM. En effet, ils vont nous permettre de connaître la contribution de chaque axe factoriel sur l'ensemble de nos données. Le nombre d'axe factoriel peut se calculer de la manière suivante :

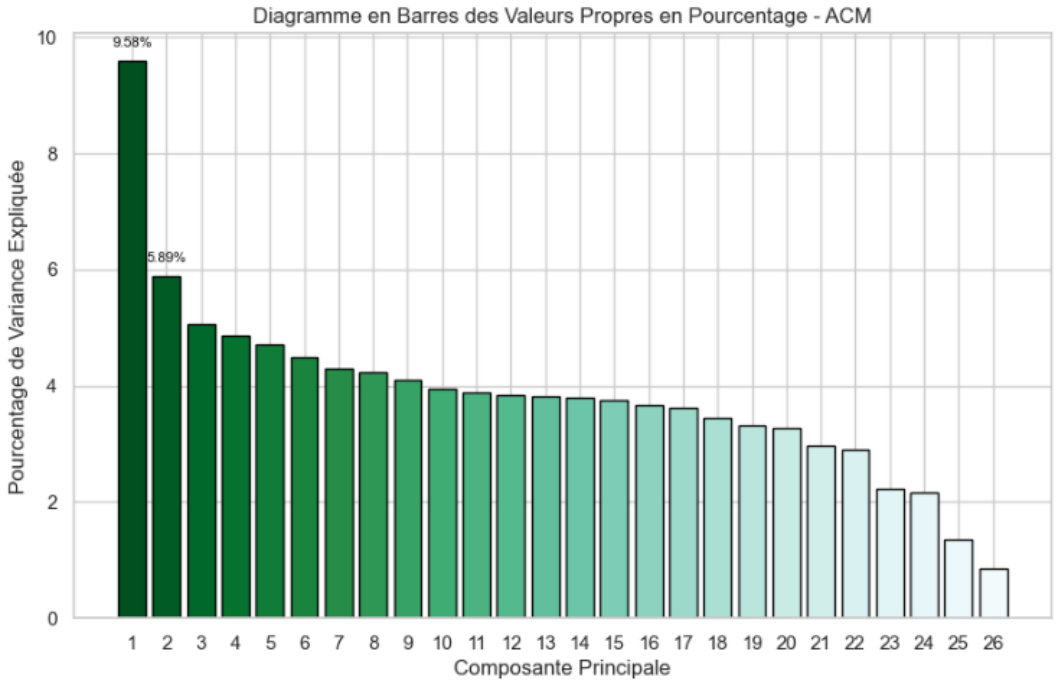
Nombre d'axe factoriel = nombre de modalité - nombre de variable

= 33 - 7

= 26

L' ACM que nous avons effectué comprendra donc 26 axes, soit un nombre équivalent de valeurs propres. La représentation ci-dessous montre donc la part d'inertie totale qu'explique chacun de ces axes.

Graphique : Diagramme en barre représentant les valeurs propres

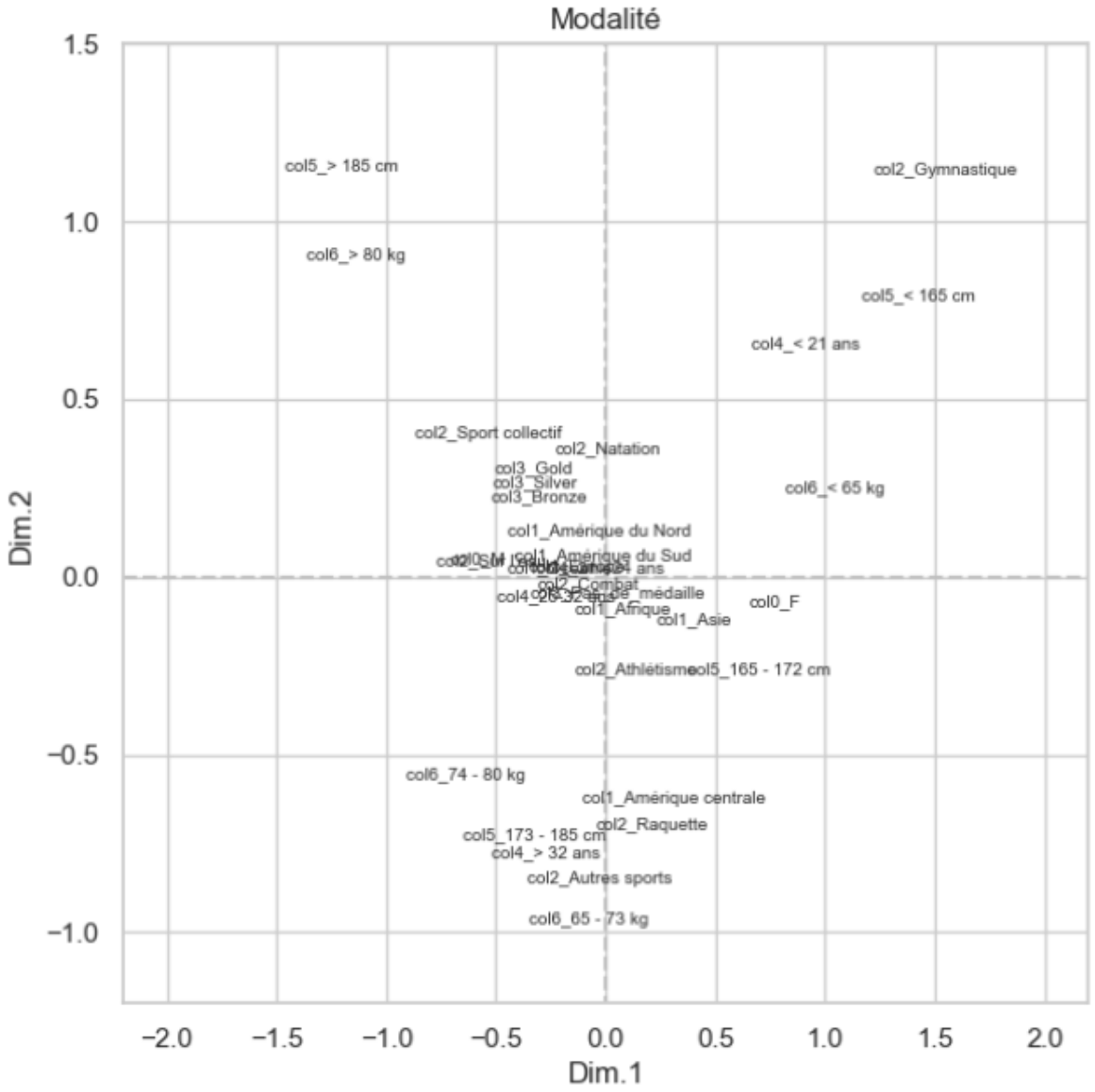


Nous pouvons voir grâce à cette représentation graphique que notre premier plan factoriel représente 15,47% de l'inertie totale. Au regard de l'allure des valeurs propres de la dimension supérieur à 3 n'apporteront pas de meilleures informations, donc nous allons étudier seulement le premier axe factoriel.

5.2. Représentation graphique de l'ACM

La représentation graphique de l'ACM va permettre d'avoir un visuel sur les potentiel caractéristiques physique que doit avoir un athlète selon sa discipline pour avoir une chance d'obtenir une médaille olympique.

Graphique : ACM



Nous pouvons voir grâce à l'ACM que les athlètes auront les mêmes caractéristiques pour l'obtention d'une médaille au Jeux Olympique que ce soit une médaille d'or, d'argent ou de bronze. En effet, nous pouvons voir les trois médailles les unes à côté des autres. Nous pouvons voir que les variables concernant la taille et le poid vont contribuer à la construction des axes.

En addition, nous pouvons voir que selon les disciplines les athlètes n'auront pas les mêmes conditions physiques. Les gymnastes seront généralement des athlètes de petite taille (moins de 1m65) et étant jeunes (moins de 21 ans).

Au contraire, les athlètes pratiquant un sport collectif ou de la natation auront généralement une grande taille (plus de 1m85) et ayant un poids supérieur à 80kg.

Les athlètes pratiquant un sport de raquette ou un autre sport seront généralement âgés de plus de 32 ans, étant de taille moyenne (entre 1m73 et 1m85).

Nous pouvons aussi voir que les athlètes représentant un pays d'Amérique centrale, d'Asie, ou d'Afrique ne remportent généralement pas de médaille au Jeux Olympique.

Nous allons maintenant étudier les contributions de chaque modalité dans la construction de notre ACM. Pour cela, nous allons faire deux tableaux dans lesquels nous allons répertorier les modalités qui contribuent le plus à la construction de l'axe correspondant, soit les contribution supérieur à la contribution moyenne (1/33) *100= 3%.

Tableau : Contribution de l'axe 1

Variable	modalité	Contribution axe 1 (%)
Classe_weight	col6_< 65 kg	16,35
Classe_weight	col6_> 80 kg	12,50
Classe_height	col5_< 165 cm	12,14
Classe_height	col5_> 185 cm	11,95
Sex	col0_F	10,32
Sport	col2_Gymnastique	7,92
Sex	col0_M	7,72
classe_age	col4_< 21 ans	4,99
Classe_height	col5_165 - 172 cm	4,68

Nous pouvons voir que les variables qui contribuent à l'axe 1 sont les variables "Classe_wieght", "Classe_height", "Sex", "Sport", "classe_age". Particulièrement, deux modalités de la variable concernant le poids contribuent le plus à la construction de l'axe 1. Puis, deux modalités de la variable concernant la taille. Cette contribution indique que ces deux variables sont importantes pour l'obtention d'une médaille au Jeux Olympique.

Tableau : Contribution de l'axe 2

Variable	modalité	Contribution axe 1 (%)
Classe_height	col5_> 185 cm	17,96
Classe_weight	col6_65 - 73 kg	14,08
Classe_height	col5_173 - 185 cm	13,87
Classe_weight	col6_> 80 kg	12,79
Sport	col2_Autres sports	7,16
Sport	col2_Gymnastique	7,00
Classe_height	col5_< 165 cm	6,10
classe_age	col4_< 21 ans	4,17
classe_age	col4_> 32 ans	4,13

Nous pouvons voir que deux modalités de la variable concernant la taille de l'athlète contribue le plus à la construction de l'axe 2 de notre ACM. Puis, deux modalités de la variable concernant le poids de l'athlète contribue le plus à la construction de l'ACM.

Remarque *: nous avons mis en annexe le tableau complet des contributions (cf annexe 3)*

L'ACM est une étape permettant de se faire une idée sur les caractéristiques que doit avoir un athlète pour gagner une médaille au Jeux Olympiques. Nous allons maintenant faire une modélisation pour en savoir plus sur les caractéristiques que doivent avoir les athlètes.

6. Modélisation

La modélisation est une étape importante dans pour anticiper la prédiction d'obtenir une médaille aux Jeux Olympiques. Cette étape de modélisation va consister à réaliser plusieurs modèles pour utiliser le meilleur.

Pour choisir le meilleur modèle, nous allons comparer les R^2 de chaque modèle car il est un indicateur de performance du modèle. Plus le R^2 est élevé, plus le modèle sera performant.

Ensuite, nous allons utiliser la matrice de confusion pour calculer certains indicateurs de performance.

- ***Accuracy***

Définition : L'accuracy représente la proportion d'observations correctement prédites par le modèle par rapport à l'ensemble des observations. C'est le ratio des vrais positifs et vrais négatifs sur l'ensemble des prédictions.

$$\frac{\text{Vrais positifs + Vrais négatifs}}{\text{Total des prédictions}}$$

Interprétation : Une accuracy élevée indique que le modèle a une bonne capacité globale à faire des prédictions correctes.

- ***Précision globale***

Définition : La précision mesure la proportion de prédictions correctes parmi toutes les prédictions positives faites par le modèle. C'est une mesure de l'exactitude des prédictions positives.

$$\frac{\text{Vrais positifs}}{\text{Vrais positifs + Faux positifs}}$$

Interprétation : Une précision élevée indique que le modèle a une faible propension à faire des fausses alarmes lorsqu'il prédit une observation positive.

- ***Sensibilité***

Définition : La sensibilité mesure la proportion de vrais positifs parmi toutes les observations réellement positives. C'est une mesure de la capacité du modèle à identifier les cas positifs réels.

$$\frac{\text{Vrais positifs}}{\text{Vrais positifs + Faux négatifs}}$$

Interprétation : : Une sensibilité élevée indique que le modèle a une forte capacité à détecter les cas positifs réels.

- ***Spécificité***

Définition : La spécificité mesure la proportion de vrais négatifs parmi toutes les observations réellement négatives. C'est une mesure de la capacité du modèle à éviter de faire des fausses alarmes pour les cas négatifs.

$$\frac{\text{Vrais négatifs}}{\text{Vrais négatifs + Faux positifs}}$$

Interprétation : Une spécificité élevée indique que le modèle a une forte capacité à éviter les fausses alarmes lorsqu'il prédit une observation négative.

- ***L'erreur α :***

Définition : L'erreur α mesure la proportion de faux négatifs parmi les faux négatifs et les vrais positifs. C'est une mesure qui permet de connaître à quel point les modèles se trompent pour les faux négatifs.

$$\frac{\text{Faux négatifs}}{\text{Vrais négatifs + Faux négatifs}}$$

Interprétation : Un taux d'erreur α élevé indique que le modèle prédit que l'observation est négative alors qu'en réalité il est positif.

- **L'erreur β :**

Définition : L'erreur β mesure la proportion de faux positifs parmi les faux positifs et les vrais négatifs. C'est une mesure qui permet de connaître à quel point les modèles se trompent pour les faux positifs.

$$\frac{\text{Faux positifs}}{\text{Vrais négatifs + Faux positifs}}$$

Interprétation : Un taux d'erreur β élevé indique que le modèle prédit que l'observation est positive alors qu'en réalité il est négatif.

- **L'erreur moyen :**

Définition : L'erreur moyen mesure la proportion de faux positifs et de faux négatifs parmi toutes les prédictions. C'est une mesure qui permet de connaître à quel point les modèles se trompent.

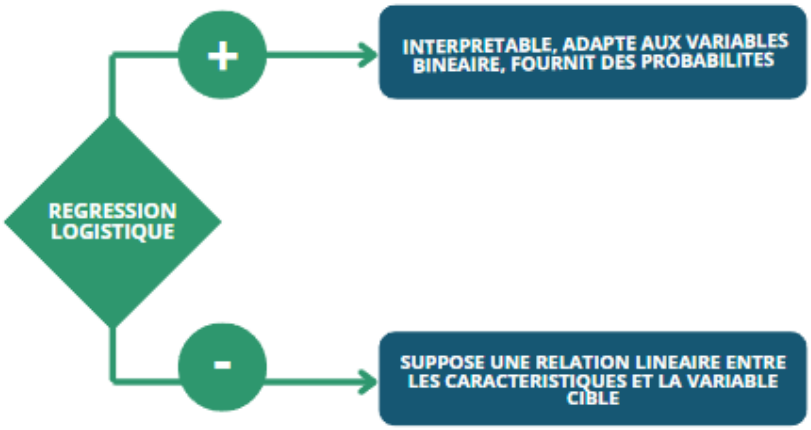
$$\frac{\text{Faux positifs + Faux négatifs}}{\text{Total des prédictions}}$$

Interprétation : Un taux d'erreur moyen élevé indique que le modèle prédit mal.

6.1. Modèle 1 : Régression logistique

- ***Qu'est un modèle de régression logistique ? :***

Un modèle de régression logistique permet de prédire des variables catégorielles. Pour la prédiction, il va calculer la probabilité qu'un individu appartienne à une certaine catégorie. Dans notre cas, il va calculer la probabilité d'obtenir une médaille aux Jeux Olympiques.



Pour réaliser le modèle de régression logistique, nous allons utiliser la fonction LogisticRegression() sur python. Nous avons un R^2 sur le jeu d'entraînement à 0,85 et un R^2 sur le jeu de test à 0,85. C'est un résultat satisfaisant car notre modèle prédit aussi bien sur le jeu d'entraînement que sur le jeu de test.

- ***Matrice de confusion :***

Régression Logistique		Valeurs prédites		Total
		Médaille	Pas médaille	
Valeurs réelles	Médaille	11 296	3	11299
	Pas médaille	1 985	6	1991
Total		13281	9	13290

La matrice se lit comme suit :

Sensibilité :

Le modèle prédit que 11 296 athlètes ont une médaille, et en réalité c'est le cas.

Spécificité :

Le modèle prédit que 6 athlètes n'ont pas de médaille, et en réalité c'est le cas.

Risque α :

Le modèle prédit que 3 athlètes n'ont pas de médaille, alors qu'en réalité ils en ont une.

Risque β :

Le modèle prédit que 1985 athlètes ont une médaille, alors qu'en réalité ils n'en ont pas.

Taux d'erreur α	Taux d'erreur β	Taux d'erreur moyen
$\tau_\alpha = \frac{3}{11299} \times 100 = 0,026\%$	$\tau_\beta = \frac{1\,985}{1991} \times 100 = 99,69\%$	$\tau = \frac{3+1985}{13290} \times 100 = 14,9\%$

Ces taux d'erreur s'interprètent comme suit :

Taux d'erreur α :

Le modèle prédit que 0,026% des athlètes n'ont pas de médaille alors qu'en réalité ils en ont une.

Taux d'erreur β :

Le modèle prédit que 99,69 % des athlètes ont une médaille alors qu'en réalité ils n'en ont pas.

Taux d'erreur moyen :

Le modèle prédit que 14,9% des individus ne sont pas prédit correctement.

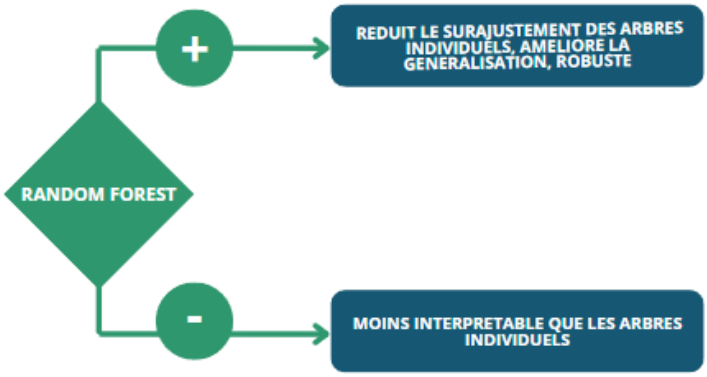
Autres indicateurs :

Sensibilité	$\frac{11296}{11299} \times 100 = 99,97\%$	**Accuracy**	$\frac{11296+6}{13290} \times 100 = 85,04\%$
Spécificité	$\frac{6}{1991} \times 100 = 0,3\%$	**Précision globale**	$\frac{11296}{13281} \times 100 = 85,05\%$

6.2. Modèle 2 : Random Forest

- ***Qu'est ce qu'un modèle Random Forest :***

Un modèle random forest est un modèle d'apprentissage automatique et a pour but de croiser le résultat de plusieurs arbres de décision. Chaque arbre possède un sous-ensemble de données d'entraînement aléatoire et les prédictions seront le résultat des croisements pour les prédictions finales.



Pour réaliser le modèle de random forest, nous allons utiliser la fonction RandomForest() sur python. Nous avons un R^2 sur le jeu d'entraînement à 0,85 et un R^2 sur le jeu de test à 0,849.

- ***Matrice de confusion :***

Random Forest		Valeurs prédites		Total
		Médaille	Pas médaille	
Valeurs réelles	Médaille	11174	125	11299
	Pas médaille	1873	118	1991
Total		13047	243	13290

La matrice se lit comme suit :

Sensibilité :

Le modèle prédit que 11 174 athlètes ont une médaille, et en réalité c'est le cas.

Spécificité :

Le modèle prédit que 118 athlètes n'ont pas de médaille, et en réalité c'est le cas.

Risque α :

Le modèle prédit que 125 athlètes n'ont pas de médaille, alors qu'en réalité ils en ont une.

Risque β :

Le modèle prédit que 1873 athlètes ont une médaille, alors qu'en réalité ils n'en ont pas.

Taux d'erreur α	Taux d'erreur β	Taux d'erreur moyen
$\tau_\alpha = \frac{125}{11299} \times 100 = 1,1\%$	$\tau_\beta = \frac{1\,873}{1991} \times 100 = 94,07\%$	$\tau = \frac{125 + 1873}{13290} \times 100 = 15\%$

Ces taux d'erreur s'interprètent comme suit :

Taux d'erreur α :

Le modèle prédit que 1,1 % des athlètes n'ont pas de médaille alors qu'en réalité ils en ont une.

Taux d'erreur β :

Le modèle prédit que 94,07 % des athlètes ont une médaille alors qu'en réalité ils n'en ont pas.

Taux d'erreur moyen :

Le modèle prédit que 15 % des individus ne sont pas prédit correctement.

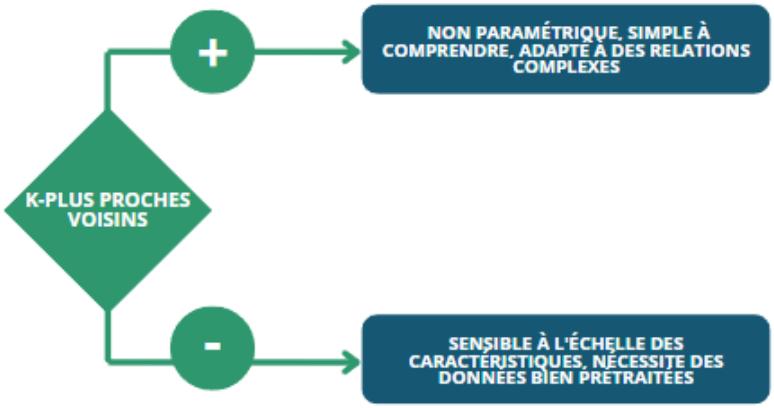
Autres indicateurs :

Sensibilité	$\frac{11174}{11299} \times 100 = 98,89\%$	**Accuracy**	$\frac{11174+118}{13290} \times 100 = 84,96\%$
Spécificité	$\frac{118}{1991} \times 100 = 5,9\ \%$	**Précision globale**	$\frac{11174}{13047} \times 100 = 85,64\ \%$

6.3. Modèle 3 : K-plus proches voisins

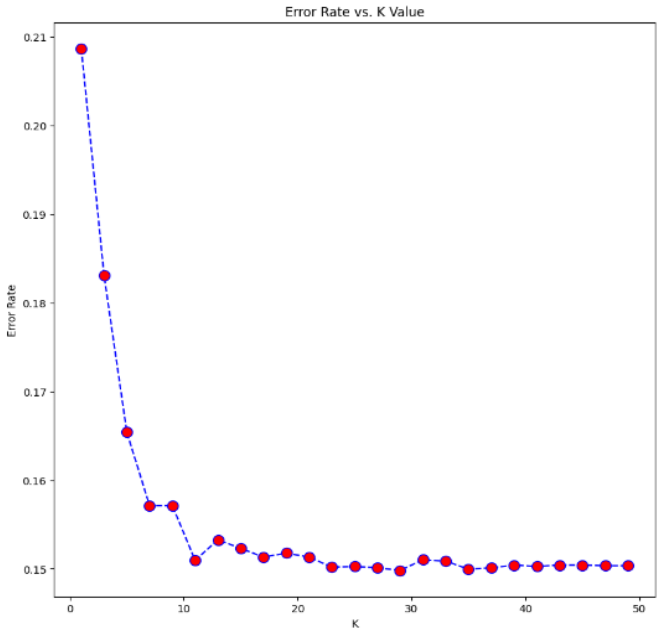
- ***Qu'est un modèle des K plus proches voisins? :***

Le modèle des K plus proche voisins aussi appelé KNN, est un algorithme d'apprentissage supervisé. Il va utiliser la proximité des autres individus pour pouvoir effectuer la classification. Avant de pouvoir réaliser le modèle, il faudra déterminer le nombre de voisins que l'algorithme doit prendre en compte.



Pour obtenir le nombre de voisin à prendre en compte, nous avons réalisé un graphique permettant de voir le nombre de voisin conseiller pour cette approche.

Graphique : Le choix du KNN le plus optimal



Ce graphique nous montre le k ou le taux d'erreur est le plus minimisé. Ici, nous pouvons constater que le taux le plus bas se situe lorsque k = 29.

Pour réaliser le modèle de KNN, nous allons utiliser la fonction KNeighborsClassifier() sur python. Nous avons un R^2 sur le jeu d'entraînement à 0,85 et un R^2 sur le jeu de test à 0,85.

- ***Matrice de confusion :***

K-plus proche voisin		Valeurs prédites		Total
		Médaille	Pas médaille	
Valeurs réelles	Médaille	11200	99	11299
	Pas médaille	1892	99	1991
Total		13092	198	13290

La matrice se lit comme suit :

Sensibilité :

Le modèle prédit que 11 200 athlètes ont une médaille, et en réalité c'est le cas.

Spécificité :

Le modèle prédit que 99 athlètes n'ont pas de médaille, et en réalité c'est le cas.

Risque α :

Le modèle prédit que 99 athlètes n'ont pas de médaille, alors qu'en réalité ils en ont une.

Risque β :

Le modèle prédit que 1892 athlètes ont une médaille, alors qu'en réalité ils n'en ont pas.

Taux d'erreur α	Taux d'erreur β	Taux d'erreur moyen
$\tau_\alpha = \frac{99}{11299} \times 100 = 0,87\%$	$\tau_\beta = \frac{1\,892}{1991} \times 100 = 95,02\,\%$	$\tau = \frac{99 + 1892}{13290} \times 100 = 14,9\%$

Ces taux d'erreur s'interprètent comme suit :

Taux d'erreur α :

Le modèle prédit que 0,87% des athlètes n'ont pas de médaille alors qu'en réalité ils en ont une.

Taux d'erreur β :

Le modèle prédit que 95,02% des athlètes ont une médaille alors qu'en réalité ils n'en ont pas.

Taux d'erreur moyen :

Le modèle prédit que 14,9 % des individus ne sont pas prédit correctement.

Autres indicateurs :

Sensibilité	$\frac{11200}{11299} \times 100 = 99,12\%$	**Accuracy**	$\frac{11200+99}{13290} \times 100 = 85\ \%$
Spécificité	$\frac{99}{1991} \times 100 = 4,97\ \%$	**Précision globale**	$\frac{11200}{13092} \times 100 = 85,55\ \%$

6.4. Comparaison des modèles et choix du meilleur modèle

Pour comparer les modèles et déterminer les meilleurs modèles, nous allons prendre en compte premièrement le R^2 du jeu de test. Deuxièmement, nous allons prendre en compte les taux d'erreur moyen.

Tableau : Récapitulatif des indicateurs de performances des trois modèles

	Régression logistique	Random Forest	KNN
R^2 données d'entrainements	0,85	0,85	0,85
R^2 données de test	0,85	0,849	0,85
Accuracy	85,04 %	84,96 %	85 %
Taux d'erreur moyen	14,9 %	15 %	14,9 %
Taux d'erreur α	0,026 %	1,1 %	0,87 %
Taux d'erreur β	99,69 %	94,07 %	95,02 %
Sensibilité	99,97 %	98,89 %	99,12 %
Spécificité	0,3 %	5,9 %	4,97 %
Classement des modèles	1	3	2

Nous pouvons voir que le modèle Random Forest est légèrement moins performant sur le jeu de test. Donc nous plaçons ce modèle en troisième position.

Concernant les deux autres modèles, nous avons regardé le taux d'erreur moyen et l'accuracy du modèle. Ils possèdent tous les deux un taux moyen de 14,9%, donc le choix du modèle modèle va se jouer sur le résultat de l'accuracy. L'accuracy du modèle de la régression logistique est très légèrement plus élevé. Donc le meilleur modèle est le modèle de régression logistique.

6.5. Interprétation du meilleur modèle

Le modèle de régression logistique étant le meilleur modèle, nous allons interpréter les résultats obtenus. Pour cela, le modèle de régression logistique nous donne des coefficients directeurs pour chaque modalité intégrée dans le modèle. Ces coefficients vont nous permettre de savoir si les modalités ont une influence sur le fait que l'athlète aura une médaille au Jeux Olympiques.

Tableau : Coefficient obtenu avec le modèle de régression logistique et de leur significativité

Paramètre	Estimations	Pr > khi-2
intercept	-3,408	
Sex_M	-0,398	<0.001
NOC_Amérique centrale	-1,403	<0.001
NOC_Amérique du Nord	1,596	<0.001
NOC_Amérique du Sud	0,662	<0.001
NOC_Asie	0,904	<0.001
NOC_Europe	0,998	<0.001
NOC_Océanie	1,082	<0.001

Sport_Autres sports	0,512	<0.001
Sport_Combat	1,029	<0.001
Sport_Gymnastique	0,327	<0.001
Sport_Natation	0,527	<0.001
Sport_Raquette	0,174	0.017
Sport_Sport collectif	1,371	<0.001
Sport_Sur l'eau	0,998	<0.001
Classe_age_25-32 ans	0,051	0.267
Classe_age_< 21 ans	-0,300	<0.001
Classe_age_> 32 ans	-0,085	0.151
Classe_height_173 - 185 cm	0,150	0.035
Classe_height_< 165 cm	-0,086	0.149
Classe_height_> 185 cm	0,250	0.001
Classe_weight_74 - 80 kg	0,106	0.100
Classe_weight_< 65 kg	0,023	0.392
Classe_weight_> 80 kg	0,188	0.011

Nous pouvons voir que la quasi-totalité des modalités intégrées dans notre modèle sont significatives. Une modalité est est significative lorsque le Pr > khi-2 indique < 0.001, cependant les modalité qui n'ont pas cet indicateur presque significative. En effet, nous pouvons dire qu'une modalité est significative lorsque l'indicateur du khi-2 est inférieur à 0,2. Donc seule la modalité "< 65 kg" de la variable "Classe_weight" n'est pas significative. Cependant, nous ne pouvons pas l'enlever de notre analyse car les deux autres modalités sont significatives.

Ensuite, dans ce tableau, nous pouvons remarquer un paramètre s'appelant "intercept", celui-ci correspond à notre individu de référence. Notre individu de référence est celui qui possède les caractéristique suivante :

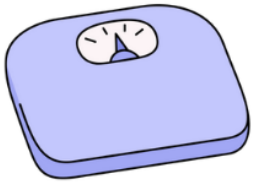
sex :
femme

NOC :
Afrique

Sport :
Athlétisme

Classe_age :
21 - 24 ans

Classe_height :
165 - 172 cm



Classe_weight :
65 - 73 kg

Pour interpréter le modèle de régression logistique, il faut adapter la valeur de d'estimations en passant à l'exponentielle.

Tableau : Interprétation des coefficients du modèle de régression logistique

Paramètre	Estimations	EXP(Estimation)	Probabilité
intercept	-3,408	0,033	0,032
Sex_M	-0,398	0,672	
NOC_Amérique centrale	-1,403	0,246	
NOC_Amérique du Nord	1,596	4,933	
NOC_Amérique du Sud	0,662	1,939	
NOC_Asie	0,904	2,470	
NOC_Europe	0,998	2,714	
NOC_Océanie	1,082	2,952	

Sport_Autres sports	0,512	1,668
Sport_Combat	1,029	2,799
Sport_Gymnastique	0,327	1,387
Sport_Natation	0,527	1,694
Sport_Raquette	0,174	1,191
Sport_Sport collectif	1,371	3,940
Sport_Sur l'eau	0,998	2,713
Classe_age_25-32 ans	0,051	1,053
Classe_age_< 21 ans	-0,300	0,741
Classe_age_> 32 ans	-0,085	0,918
Classe_height_173 - 185 cm	0,150	1,162
Classe_height_< 165 cm	-0,086	0,918
Classe_height_> 185 cm	0,250	1,284
Classe_weight_74 - 80 kg	0,106	1,112
Classe_weight_< 65 kg	0,023	1,023
Classe_weight_> 80 kg	0,188	1,207

- La probabilité d'obtenir une médaille aux Jeux Olympiques pour d'individu de référence est de 0,032.

- Le fait d'être un homme divise par 1,4 (1/0,672) les chances d'obtenir une médaille par rapport à une femme.

- Le fait que l'athlète représente un pays de l'Amérique du Nord multiplie par 4,9 les chances d'obtenir une médaille par rapport à un athlète représentant un pays d'Afrique.

- Le fait que l'athlète ait moins de 21 ans divise par 1,3 les chances d'obtenir une médaille aux Jeux par rapport à un athlète ayant entre 21 et 24 ans.

- Le fait qu'un athlète pratiquant un sport collectif multiplie par 3,9 les chances d'obtenir une médaille aux Jeux Olympiques par rapport à un athlète pratiquant de l'athlétisme.

6.6. Scoring

Le scoring est une méthode permettant de connaître les meilleures modalités que doivent avoir les athlètes pour avoir le plus de chance d'obtenir une médaille au Jeux Olympiques. Ou au contraire, les modalités qui n'est pas favorable à l'obtention d'une médaille au Jeux Olympiques.

Tableau : Scoring

Paramètre	Estimations	Pr > khi-2	minimum	maximum	ecart	score	Meilleur
Sex_F	0		-0,398	0	0,398	7,05%	7,05%
Sex_M	-0,398	<0.001				0,00%	
NOC_Afrique	0		-1,403	1,596	2,999	24,86%	53,14%
NOC_Amérique centrale	-1,403	<0.001				0,00%	
NOC_Amérique du Nord	1,596	<0.001				53,14%	
NOC_Amérique du Sud	0,662	<0.001				36,60%	
NOC_Asie	0,904	<0.001				40,89%	
NOC_Europe	0,998	<0.001				42,55%	
NOC_Océanie	1,082	<0.001				44,04%	
Sport_Athlétisme	0		0	1,371	1,371	0,00%	24,29%
Sport_Autres sports	0,512	<0.001				9,06%	
Sport_Combat	1,029	<0.001				18,23%	
Sport_Gymnastique	0,327	<0.001				5,80%	
Sport_Natation	0,527	<0.001				9,34%	
Sport_Raquette	0,174	0.017				3,09%	
Sport_Sport collectif	1,371	<0.001				24,29%	
Sport_Sur l'eau	0,998	<0.001				17,68%	
Classe_age_21 - 24 ans	0		-0,300	0,051	0,352	5,32%	6,23%
Classe_age_25-32 ans	0,051	0.267				6,23%	
Classe_age_< 21 ans	-0,300	<0.001				0,00%	
Classe_age_> 32 ans	-0,085	0.151				3,81%	
Classe_height_165 - 172 cm	0		-0,086	0,250	0,336	1,52%	5,96%
Classe_height_173 - 185 cm	0,150	0.035				4,18%	
Classe_height_< 165 cm	-0,086	0.149				0,00%	
Classe_height_> 185 cm	0,250	0.001				5,96%	
Classe_weight_65 - 73 kg	0		0	0,188	0,188	0,00%	3,33%
Classe_weight_74 - 80 kg	0,106	0.100				1,88%	
Classe_weight_< 65 kg	0,023	0.392				0,40%	
Classe_weight_> 80 kg	0,188	0.011				3,33%	
					5,644		100,00%

Pour réaliser le scoring, nous avons repéré le minimum et le maximum des estimations de chaque variable. Puis nous avons calculé l'écart entre le minimum et le maximum en valeur absolue. Pour pouvoir calculer le score de chaque modalité, nous devons calculer la somme des écarts.

Pour finir le score de chaque modalité se calculent en faisant la différence entre l'estimation et le minimum, le tout divisé par l'écart total calculé au préalable.

Nous pouvons voir que notre score est correct car la somme de toutes les meilleures modalités fait 100%.

Remarque :

- *Nous avons mis en vert les scores correspondant au modalité favorisant l'obtention d'une médaille aux Jeux Olympiques.*
- *Nous avons mis en rouge les scores correspondant aux modalités défavorisant l'obtention d'une médaille aux Jeux Olympiques.*

7. Résultat de la modélisation

Au vu des résultats du score, nous pouvons établir un profil d'un athlète qui serait amené à obtenir une médaille aux Jeux Olympiques, et un profil d'un athlète qui ne serait pas amené à obtenir une médaille aux Jeux Olympiques.

- ***<u>Athlète favorable à obtenir une médaille :</u>***

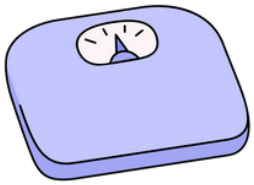
sex :
femme

NOC :
Amérique du Nord

Sport :
Sport collectif

Classe_age :
25 - 32 ans

Classe_height :
> 185 cm



Classe_weight :
> 80 kg

- ***<u>Athlète défavorable à obtenir une médaille :</u>***

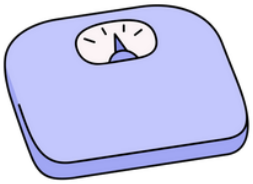
sex :
Homme

NOC :
Amérique Centrale

Sport :
Athlétisme

Classe_age :
< 21 ans

Classe_height :
> 165 cm



Classe_weight :
65 - 73 kg

8. Conclusion

En conclusion, la gestion de la partie Git au début du projet a présenté des défis avec un prise en main compliquée, mais l'adoption de bonnes pratiques et l'utilisation de branches distinctes ont facilité la collaboration. La mise en place d'un repository sur GitHub, avec des branches "dev" et "main", a permis un développement efficace et une intégration régulière des modifications.

En ce qui concerne la modélisation, l'étude des performances des athlètes aux Jeux Olympiques 2024 a été complexe en raison du niveau élevé des sportifs. Malgré cela, trois modèles ont été évalués : la régression logistique, le Random Forest et les k-plus proches voisins (KNN). Les tests ont révélé que la régression logistique était le modèle le plus performant en termes d'accuracy sur le jeu de test.

L'approche collaborative a également été appliquée aux bonnes pratiques de développement, avec l'utilisation de packages et de tests unitaires. Les fonctions ont été regroupées dans des packages, facilitant la réutilisation et la maintenance du code. Les tests unitaires ont été mis en place pour assurer la fiabilité du code.

Enfin, l'interprétation du modèle de régression logistique a permis de dégager des insights sur les caractéristiques associées à la probabilité d'obtenir une médaille. Par exemple, être un homme, représenter un pays d'Amérique du Nord, participer à un sport collectif, et avoir une taille entre 173 et 185 cm sont des facteurs positivement corrélés à la probabilité de médaille.

Le scoring a été utilisé pour identifier les modalités les plus favorables et défavorables à l'obtention d'une médaille. Ces résultats permettent d'établir un profil d'athlète favorable à la réussite aux Jeux Olympiques.

Annexe

Annexe 1 : modalité de la variable NOC

Afrique	EGY ,CMR ,ALG ,RSA ,MAR ,ERI ,SUD ,ETH ,CIV ,KEN ,NGR ,ANG ,DJI ,SOM ,GHA ,UGA ,TUN ,MRI ,NIG ,LBR ,LBA ,NAM ,RWA ,SEY ,COM ,BEN ,LIB ,CAF ,MAD ,CHA ,GAB ,BOT ,MOZ ,CPV ,ZIM ,SEN ,MTN ,COD ,GUI ,CGO ,TOG ,SLE ,GEQ ,TAN ,MLI ,SWZ ,BDI ,STP ,GBS ,ZAM ,MAW ,GAM ,SSD ,LES ,BUR
Amérique centrale	NCA ,HON ,BIZ ,GUA ,VIN ,CRC ,ESA ,PAN ,CAM
Amérique du Nord	CUB ,MEX ,USA ,CAN ,SKN ,PUR ,DMA ,DOM ,TTO ,BER ,JAM ,IVB ,CAY ,BAR ,BAH ,GRN ,LCA ,HAI ,ANT ,ISV
Amérique du Sud	ARG ,BRA ,CHI ,GUY ,COL ,VEN ,SUR ,PAR ,URU ,PER ,ECU ,BOL ,ARU
Asie	CHN ,AZE ,BRN ,IRQ ,QAT ,PAK ,IRI ,KAZ ,BRU ,KUW ,MAS ,INA ,UZB ,UAE ,KGZ ,TJK ,JPN ,TUR ,SRI ,ARM ,SYR ,GEO ,BAN ,JOR ,PLE ,KSA ,IND ,TKM ,NEP ,MGL ,MDV ,SGP ,YEM ,OMA ,PHI ,TLS ,ISR ,THA ,KOR ,PRK ,LAO ,HKG ,MYA ,AFG ,TPE ,BHU ,VIE
Europe	FIN ,ROU ,NOR ,NED ,FRA ,EST ,ESP ,ITA ,RUS ,BLR ,GRE ,IRL ,BEL ,GER ,SUI ,SWE ,POR ,GBR ,SLO ,POL ,LTU ,UKR ,CRO ,DEN ,MLT ,AUT ,SCG ,HUN ,CYP ,MDA ,BUL ,LAT ,SRB ,ISL ,BIH ,SVK ,SMR ,CZE ,LUX ,AND ,MKD ,MON ,MNE ,ALB ,LIE ,KOS
Océanie	AUS ,NZL ,PLW ,ASA ,SAM ,IOA ,FIJ ,VAN ,GUM ,ROT ,FSM ,PNG ,COK ,NRU ,MHL ,AHO ,KIR ,TUV ,TGA ,SOL

Annexe 2 : modalité de la variable Sport

Combat	Judo,Wrestling,Taekwondo,Fencing,Boxing
Natation	Swimming,Synchronized Swimming,
Raquette	Badminton,Tennis,Table Tennis
Sport collectif	Basketball,Handball,Football,Hockey,Water Polo,Softball,Volleyball,Baseball,Rugby Sevens,Beach Volleyball
Athlétisme	Athletics,Modern Pentathlon,Triathlon
Gymnastique	Gymnastics,Rhythmic Gymnastics,Trampolining,
Sur l'eau	Sailing,Rowing,Diving,Canoeing
Autres sports	Weightlifting,Cycling,Equestrianism,Archery,Shooting,Golf

Annexe 3 : Tableau des contributions

		col_contrib_dim1	col_contrib_dim2
Sex	col0_F	10,32	0,13
	col0_M	7,72	0,09
NOC	col1_Afrique	0,02	0,04
	col1_Amérique centrale	0,02	0,16
	col1_Amérique du Nord	0,00	0,13
	col1_Amérique du Sud	0,00	0,01
	col1_Asie	1,21	0,18
	col1_Europe	0,35	0,02
	col1_Océanie	0,12	0,00
Sport	col2_Athlétisme	0,15	0,80
	col2_Autres sports	0,00	7,16
	col2_Combat	0,03	0,00
	col2_Gymnastique	7,92	7,00
	col2_Natation	0,00	1,10
	col2_Raquette	0,09	1,68
	col2_Sport collectif	1,95	1,82
	col2_Sur l'eau	1,29	0,02
Medal	col3_Bronze	0,19	0,17
	col3_Gold	0,21	0,30
	col3_Pas_de_médaille	0,11	0,12
	col3_Silver	0,20	0,23
classe_age	col4_21 - 24 ans	0,01	0,01
	col4_25-32 ans	0,91	0,08
	col4_< 21 ans	4,99	4,17
	col4_> 32 ans	0,31	4,13
classe_height	col5_165 - 172 cm	4,68	1,05
	col5_173 - 185 cm	1,72	13,87
	col5_< 165 cm	12,14	6,10
	col5_> 185 cm	11,95	17,96
classe_weight	col6_65 - 73 kg	0,06	14,08
	col6_74 - 80 kg	2,49	3,12
	col6_< 65 kg	16,35	1,48
	col6_> 80 kg	12,50	12,79

contribution moyenne
3,03