Analyse et apprentissage sur le jeu de données "Paralympic Medal"

Xuan-Vinh HO, Jules YVON

1er juin 2022

Résumé

Dans le cadre de l'UV SY09, notre mission était d'analyser le jeu de données appelé "Médaille Paralympique", et d'appliquer les méthodes d'exploration et d'apprentissage vues en cours. Ce faisant, nous avons pu les pratiquer mais surtout les mettre en persepective les unes par rapport aux autres afin de choisir celles qui correspondaient le plus à nos données.

1 Introduction

Le jeu de données est composé d'environ 20 000 lignes et 10 colonnes. Chaque ligne représente une médaille gagnée dans un sport donné par un athlète : on dispose donc de différentes informations comme le nom du sport, la nature de la médaille obtenue, le pays qui a remporté la médaille ou encore l'année où a eu lieu la compétition.

1.1 Présentation des variables

La première chose remarquable de ce jeu de donnée est la nature de ses variables : deux variables quantitatives, discrètes qui plus est, et neuf variables qualitatives, dont la plupart ne semblent pas disposer de relation d'ordre. Or la plupart des méthodes d'exploration et d'analyse de données se basent sur des données numériques, de préférences continues. C'est pourquoi nous avons essayé de faire apparaitre des quantités numériques à partir de nos variables de départ.

Chaque pays sera placé dans un espace à deux dimensions (isomorphe à la surface de la Terre), via un autre jeu de donnée répertoriant les longitudes et latitudes de tous les pays.

Pour le sexe, c'est un peu moins évident, car on évite en général d'imaginer une relation d'ordre sur un ensemble qui n'en contient naturellement pas comme ce serait le cas de l'ensemble $\{Homme, Femme\}$. Cependant, comme certaines compétitions sont mixtes, on a en réalité un ensemble de 3 éléments dont l'un est nécessairement juste au milieu des autres. C'est pourquoi on affectera -1 aux rencontres masculines, 0 aux rencontres mixtes et 1 aux rencontres féminines.

Pour la nature de la médaille, c'est à la fois plus simple car une relation d'ordre apparaît naturellement (Bronze < Argent < Or) mais l'affectation à des valeurs numériques sera forcément arbitraire : combien de médailles de bronze ou d'argent vaut une médaille d'or ? Nous avons opté pour l'affectation la plus naturelle : 1 pour les médailles de bronze, 2 pour les médailles d'argent et 3 pour les médailles d'or.

Il nous reste plusieurs variables qui ne sont ni numériques ni ordinales : certaines d'entres elles comme le nom de l'événement ou de l'athlète peut prendre plusieurs milliers de valeurs différentes, ce qui les rend difficile à visualiser. Le sport lui est beaucoup plus intéressant car il ne prend que 11 valeurs différentes. Pour ce qui est de l'entraineur (*guide*), il prend aussi un grand nombre de valeur différentes, mais il est surtout utile de noter qu'il est non défini dans une très large majorité des cas, tout comme *grp_id*.

1.2 Quelques fonctions

A l'aide de la bibliothèque Pandas et de quelques algorithmes rudimentaires, il est alors facile d'utiliser le tableau des données pour construire de nouvelles variables, ce que l'on va représenter dans ce document par des fonctions mathématiques :

On va d'abord considérer 3 fonctions qui vont de l'ensemble des pays (i.e. $\mathbb{R}^2$) dans $\mathbb{N}$: bron, arg et or qui renvoient respectivement les nombres de médailles de bronze, d'argent et d'or du pays. On utilisera aussi d'autres fonctions comme score, la somme pondérée (cf supra) des nombres de médailles de chaque catégorie, obtenue par un pays ou encore la médaille majoritaire, c'est-à-dire le type de médaille qu'il obtient le plus souvent.

Une dernière fonction intéressante est rang qui donne la position du pays dans le classement de leurs scores. Elle servira principalement à découper les pays selon des classes d'effectifs équivalents.

Nous en avons terminé avec la présentation des variables et des fonctions qui seront utilisées au cours de cette analyse. Celle-ci aura pour but de répondre à la question suivante : "Comment gagne-t'on des jeux paralympiques ?" Pour y répondre nous allons commencer par utiliser les méthodes non supervisées (description élémentaire, ACP, k-means) avant de passer aux méthodes supervisées (régression logistique, arbres de décision, KPPV, analyse discriminante).

2 Méthodes non supervisées

2.1 Méthodes de description élémentaire

Cette première partie va nous permettre d'y voir un peu plus clair sur les valeurs prises par les différentes variables, ainsi que sur leur répartition.

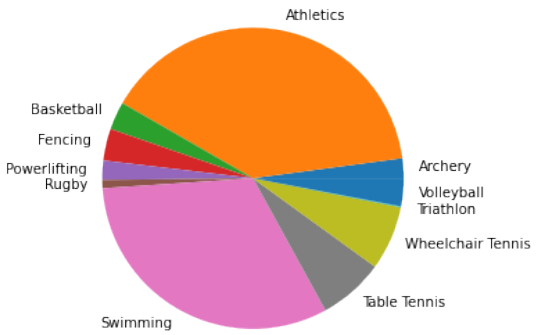


FIGURE 1 – Répartition des sports

Commençons par les sports : deux d'entre eux se distinguent clairement des autres, l'athlétisme et la natation, avec plus de 6 000 entrées chacun. Cela resservira quand on voudra faire une analyse sport par sport : les sports avec seulement quelques centaines de médailles ne constitueront pas un échantillon suffisamment gros pour que l'on puisse appliquer des méthodes d'apprentissage supervisé.

En ce qui concerne la variable année, sa répartition par année est assez uniforme, seules les années sortent 1984 et 1988 sortent du lot avec une nombre de médailles de toute nature environ 50% plus elevé

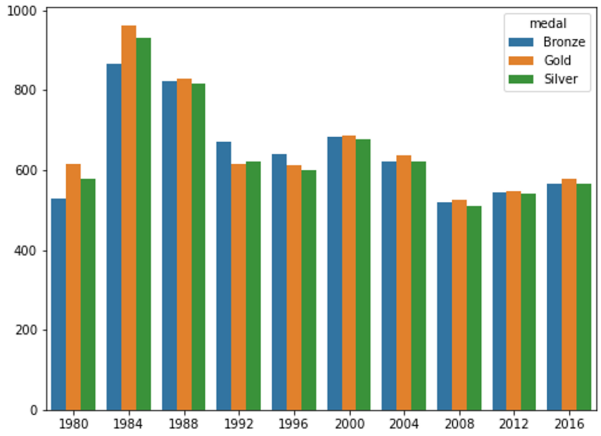


FIGURE 2 – Distribution des médailles selon les années

La répartition des types de médailles au sein d'une année tend à s'uniformiser au fil du temps : les trois quantités sont presque à égalité depuis les Jeux de 2000 alors que les médailles d'or était nettement majoritaire lors des deux premières compétitions.

Le compte du nombre de pays qui sont majoritaires dans chaque type de médaille est présenté dans le graphique suivant : on a 25% des pays qui ont majoritairement des médailles d'or, 35% qui ont majoritairement des médailles d'argent, et les 40% restant, des médailles de bronze. C'est là une information importante car elle nous indique que le fait qu'il pays ait majoritairement des médailles d'or est à la fois un indicateur de performance et un critère de démarcation vis-à-vis des autres pays.

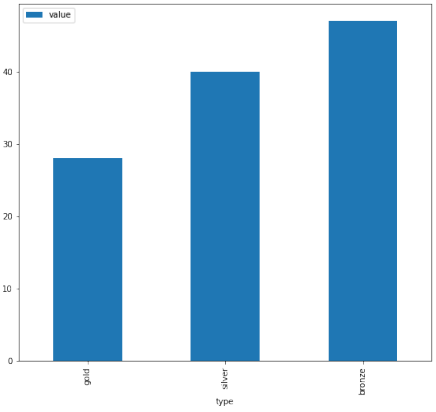


FIGURE 3 – Nombre de pays par type de médaille majoritaire

La dernière variable qu'il est important de présenter ici est celle du sexe. Nous pouvons présupposer que certains sports sont plus pratiqués par des femmes tandis que d'autres le sont plus par des hommes. Sur l'ensemble des sports, les femmes représentes 37% contre 61% pour les hommes et 1% pour les rencontres mixtes. L'annexe 1 est un graphique représentant la répartition des sexes dans chaque sport. Il nous indique quels sont les sports qui s'éloignent significativement de cette moyenne : certains sports sont pratiqués autant par les femmes que par les hommes : le triathlon (50/50), le rugby (qui est mixte à 100%), le basketball (38/36 avec 25% de sexe inconnu) et dans une moindre mesure la natation (56/44). Les sports très majoritairement masculins sont le volleyball avec 75% d'hommes, et le tennis de table, tennis-fauteuil et athlétisme, tous les trois à 66%.

Nous avons également essayé de tirer des informations des variables *grp_id* et *guide*, mais celles-ci, en plus d'être quasi-systématiquement nulles, n'ont aucune influence sur les autres variables. Elles ne seront donc pas davantage étudiées.

2.2 Analyse en composantes principales

Nous avons réalisé une analyse en composante principale sur un tableau individu-variable dont chaque individu est un pays et les variables sont or, arg, bron.

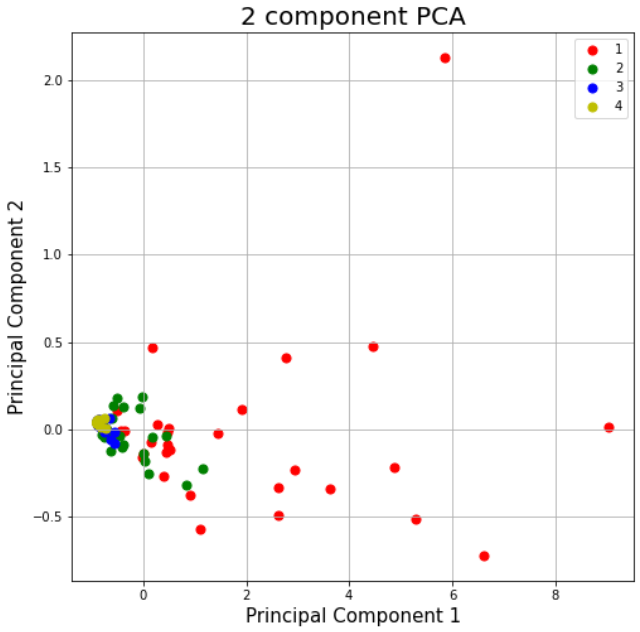


FIGURE 4 – ACP des nombres de médailles par pays

On a colorié les points du plan obtenus par cette ACP selon leur rang : ceux en jaune sont les 25% des pays qui ont le moins de médaille (en somme pondérée) puis viennent ceux en bleu et ceux en vert, jusqu'à ceux en rouge qui sont les 25% des pays ayant le plus de médaille. Cela nous permet de donner du sens au résultat obtenu, les pays les moins performants s'agglomèrent autour d'un point qui semble être (-1, 0) alors que les meilleurs s'en éloignent dans diverses directions.

Un partitionnement selon le rang ne semble pas pouvoir être deviné via la méthode des k-means, car les classes ne forment pas de blocs bien séparés mais plutôt des sortes d'arcs de cercles concentriques. Cela se confirme après avoir essayé de l'appliquer : la partition obtenue diverge fortement de celle établie via le rang des pays.

3 Méthodes supervisées

3.1 Méthode des K plus proches voisins

Cette méthode nécessite la définition d'une distance sur $\mathbb{R}^p$ où p est le nombre de variable utilisées. Ici nous allons prendre p = 3, car les variables seront la longitude, la latitude et l'année. La distance entre deux points $A = (x_a, y_a, t_a)$ et $B = (x_b, y_b, t_b)$ de $\mathbb{R}^3$ pourra alors être définie par $\sqrt{(x_a - x_b)^2 + (y_a - y_b)^2 + D^2 * (t_a - t_b)^2}$, $D \in \mathbb{R}_*^+$.

On peut ainsi voire l'espace comme un empilement de 10 plans horizontaux, séparés par une distance D, chacun correspondant à la surface du globe à une année bien précise : 4 ans plus tard que le plan en-dessous de lui et 4 ans plus tôt que le plan au-dessus (s'ils existent).

Ce D est donc un hyperparamètre, qui correspond à comment on veut faire peser l'espacement temporel par rapport à l'espacement géographique dans le calcul des distances. On en choisira librement la valeur qui nous donne les meilleurs résultats et on pourra même analyser cette valeur optimale D_{opt} : si elle est faible, c'est que les plus proches voisins d'un pays à une année donnée seront plus facilement le même pays aux années suivantes et précédentes. Au contraire si elle est elevée, les résultats attendus d'un pays à une année donnée seront plus déterminés par les résultats des pays qui lui sont proches géographiquement à la même année.

La variable que l'on va essayer d'estimer par cette méthode est la nature de la médaille la plus fréquemment obtenue par chaque pays. Avant de commencer à l'ap-

pliquer, il faut se poser la question de la variable sport : on s'imagine bien que les pays sont bien souvent pas juste "bon en sport" mais qu'ils sont plus probablement bons dans certains sports et mauvais dans d'autres. Il pourrait alors être utile d'étudier les résultats de chaque sport indépendamment.

Les résultats obtenus sont plutôt bons : sur l'ensemble des sports, on obtient seulement 60%, mais comme nous l'avions présenti, en traitant chaque sport séparément, le taux de succès des prédictions grimpe à de très bons taux sur certains sports : 94% pour l'escrime, 95% pour le volleyball, 96% pour l'escrime, et à des taux moins satisfaisant, mais tout de même supérieurs à celui de l'ensemble des sports sur d'autres : 83% pour le tennis de table et le tennis fauteuil 75% pour la natation et 63% pour l'athlétisme. Les sports dont l'effectif est inférieur à 500 ont été écartés car un nombre minimum de données est nécessaire pour pouvoir donner du crédit à un résultat.

Ces résultats ont été obtenus pour D = 1. Nous avons testé d'autres valeurs de D pour voir si cela améliorait nos résultats, notamment ceux de la natation et de l'athlétisme, mais ceux-ci sont encore moins bons pour D = 4 et pour D = 10. Pour D = 0.2, on observe des résultats semblables à ceux de D = 1. On en conclue que les prédictions qu'on peut faire à un pays dépendent surtout des résultats des autres années de ce pays.

Bien sûr, il serait difficile d'imaginer un graphique montrant la frontère obtenue par cette méthode sur un graphique en 3D. Cependant, il est possible de trouver un moyen très visuel d'observer cette frontière : comme une des trois dimensions correspondait originellement à du temps, il serait intéressant de se faire succéder dans une image animée les graphes des différents plans horizontaux et ainsi voir se déplacer les frontières au fil des années. Malheuresement, le format de ce rapport ne le permet pas : voici toutefois quelques plans pour avoir un aperçu du résultat obtenu. On peut noter que le passage de 1996 à 2000 ne modifie pas beaucoup les affectations, contrairement au passage de 2000 à 2004.

Mentionnons pour terminer cette partie les limites liées à l'utilisation de la longitude et de la latitude dans le calcul des distances, car elle ne représente pas la distance réelle (en km) entre les pays. De plus, il y a des effets de bords du au fait qu'on ne prend pas en compte la sphéricité de la Terre.

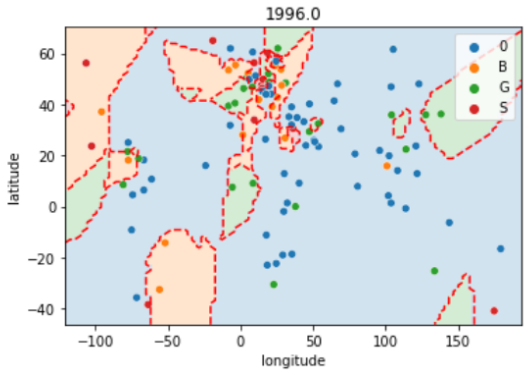


FIGURE 5 – K plus proches voisins (Athletics, 1996)

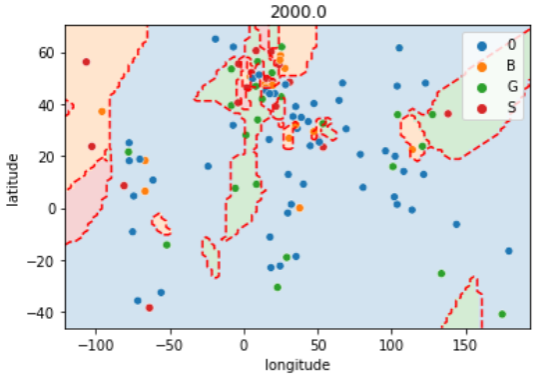


FIGURE 6 – K plus proches voisins (Athletics, 2000)

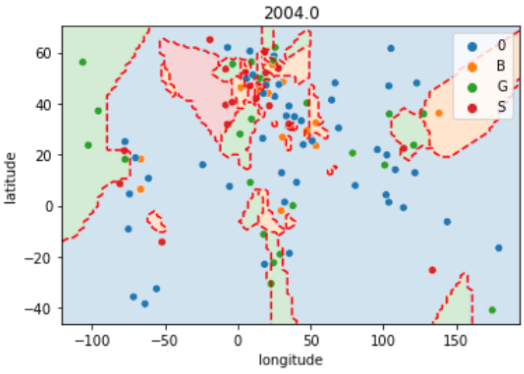


FIGURE 7 – K plus proches voisins (Athletics, 2004)

3.2 Arbres de décision

La dernière méthode d'apprentissage supervisé que nous avons réussi à appliquer avec succès et celle de l'arbre de décision. Elle permet de prédire la classe d'un pays selon ses valeurs or, bron, arg, et ce de la manière la plus transparente qui soit, un simple algorigramme, qu'on trouvera en annexe 2.

Avec cette méthode, un pays est affecté à la classe 1 (parmi les 25% les pays avec le meilleur score), s'il a 25 médailles d'or ou plus et s'il a soit 38 médailles de bronze ou moins ou alors 68 médailles de bronze ou plus. Et qu'il est de la classe 4 (25% des pays avec le moins bon score) s'il a strictement moins de 3 médailles d'or et strictement moins de 10 médailles de bronze.

Ces résultats sont suprenants, on s'attendrait à voir apparaître la variable argent pour ces deux classes comme par exemple : "un pays est affecté à la meilleure classe s'il a au moins tant d'or et tant d'argent". Cependant, c'est bien celle ci qui maximise la précision quand on sépare nos données en un ensemble d'apprentissage et un ensemble de validation (70% / 30%). Pour s'en assurer, on a fait varier le paramètre 'random_state' de la méthode de 1 à 100 pour trouver celui qui nous donnait la meilleure précision. Elle est atteinte en 83 et vaut 74%.

3.3 Autres méthodes

Pour terminer, nous allons aborder briévement les autres méthodes que nous avons testé sur nos données mais qui ont montré une efficacité très limitée.

L'analyse discriminante, selon le sport étudié, nous donnait soit une région triviale (c'est-à-dire qu'elle affectait quasiment le monde entier à la valeur '0'), soit des régions non triviales mais dont la précision, de l'ordre de 45%, était assez décevante.

La régression logistique a aussi été essayée sans grands succès. On a voulu prédire si un pays était majoritairement en or selon son score. Il parait assez évident au vu de la figure 10 que ce modèle n'est pas très fiable. La courbe obtenue atteint 0.2 vers les pays aux scores très faibles et 0.65 vers les pays aux scores les plus élevés.

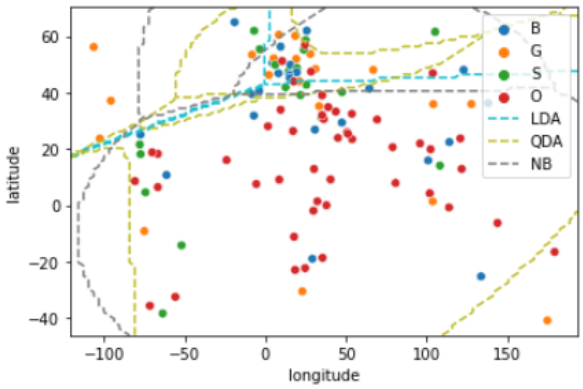


FIGURE 8 – ALD : Basketball (précision 45%)

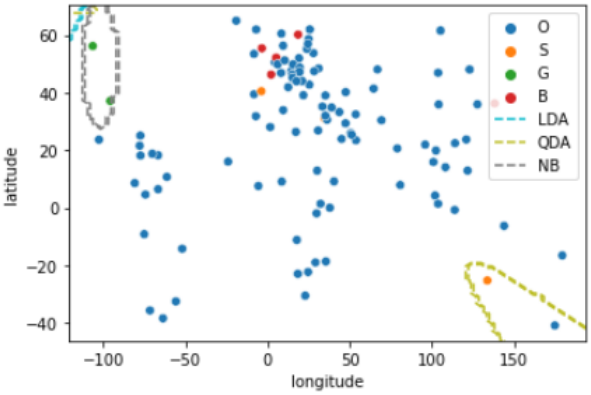


FIGURE 9 – ALD : Natation (précision 90%)

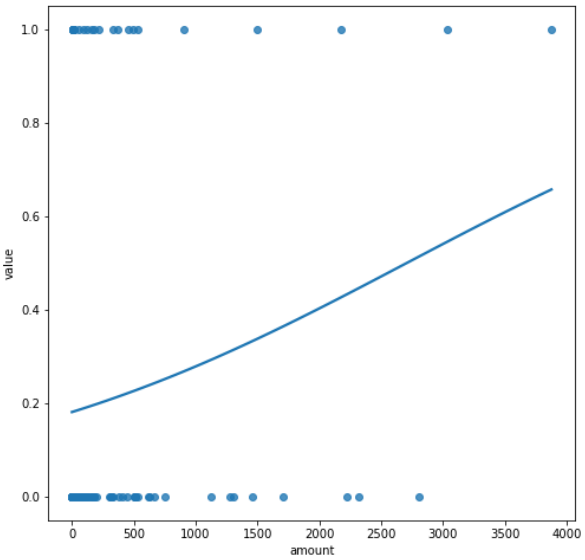
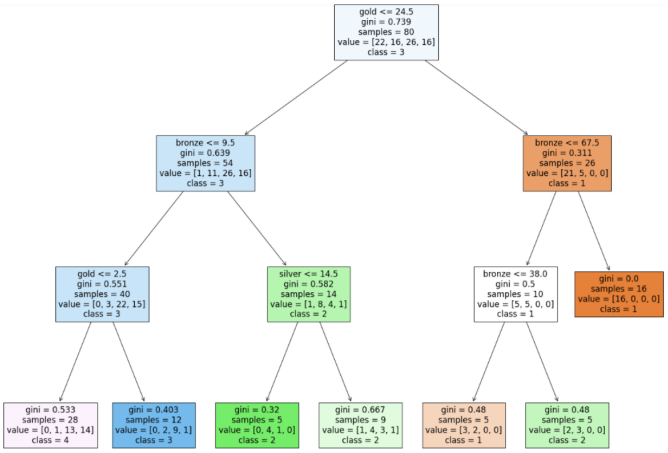


FIGURE 10 – Régression logistique (Majorité Or)

4 Conclusion

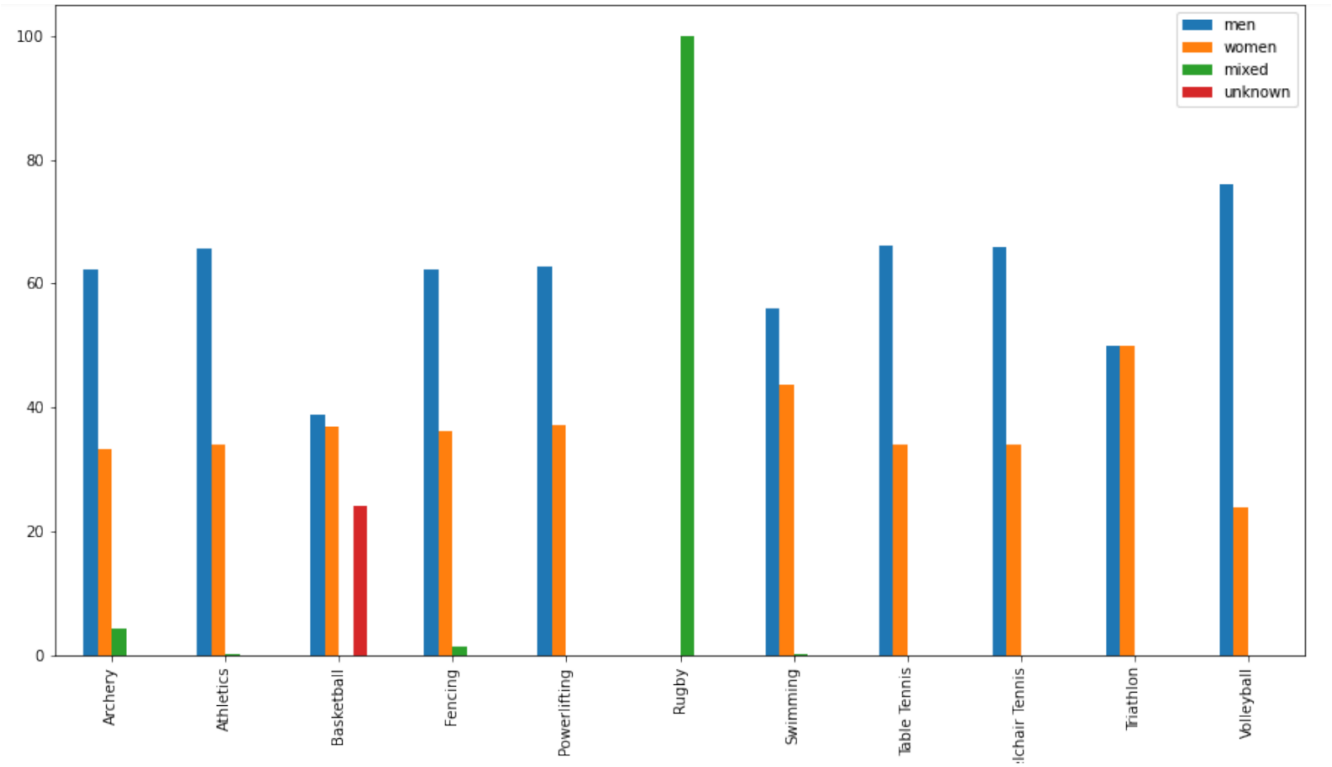
L'étude du jeu de données "Paralympic Medal" a été particulièrement difficile à cause de la nature de ses variables : la plupart étaient qualitatives et il a fallu faire preuve d'imagination pour se ramener à des variables quantitatives afin d'utiliser les méthodes vues en cours. Malgré ces difficultés, nous avons réussi à mener à bien plusieurs analyses, avec des méthodes variés et sur des variables différentes, avec un certain succès, puisque la plupart des méthodes nous donne un bon taux de réussite (entre 75% et 95%).

Pour répondre brièvement à la problématique formulée en début de rapport, nous avons vu que ce qui détermine le succès d'un pays aux jeux paralympiques à un sport donné va surtout dépendre des performances de ce pays dans ce sport aux années précédentes, et dans une bien moindre mesure, des performances des pays qui lui sont le plus proche.



Annexe 2 : Arbre de décision pour déterminer la classe d'un pays en fonction de or, arg, et bron

5 Annexes



Annexe 1 : Pourcentage de médailles par sport et par sexe