Mémoire Projet MIDL 2

BOIREAU-DEVIER Chloé
chloe.boireau-devier@univ-tlse3.fr

FELIX Judicaël
judicael.felix@univ-tlse3.fr

KOHIL Rayane
rayane.kohil@univ-tlse3.fr

Janvier 2025

Contents

1 Introduction & Choix du sujet

Ce mémoire contient le rapport de nos recherches et découvertes dans le cadre du Projet MIDL 2. A travers ce document, nous relaterons notre choix de sujet, les différentes méthodes utilisées, nos processus et étapes, ainsi que les résultats observés, en faisant le lien avec notre problématique.

Le fil rouge de notre projet est la problématique suivante : "Quel est l'impact de la non-participation de la Russie aux jeux olympiques de Paris 2024 ?" Afin de répondre à cette question, nous avons tout d'abord fait une analyse des données qui nous ont été relayées pour ce projet, auxquelles nous avons ajouté des données que nous avons trouvées. Ensuite, à l'aide de modèles prédictifs, nous nous sommes penchés sur ce qui aurait pu arriver, ainsi que sur ce qui aurait dû arriver. Nous détaillerons tout cela dans la suite de ce mémoire.

2 Choix et traitement des données

Dans cette partie, nous détaillerons comment nous avons pré-traité les données, et comment nous les avons mises en forme pour les utiliser dans la suite de ce projet. Tout le code lié à cette partie ce trouve sous les onglets "Imports et création des DF" et "Récupération d'autres données" de notre Google Collab. Nous avons tout d'abord commencé par traiter les données qui nous avaient été partagées, c'est-à-dire les données sur les Jeux de 2020 et 2024. Nous nous sommes concentrés sur les fichiers qui contenaient le total des médailles, et nous en avons créé des objets de type DataFrame (DF.).

Les données étaient complètes, mais nous nous sommes rapidement rendus compte que les pays qui n'avaient aucune médaille n'apparaissaient pas, et nous nous sommes alors penchés sur la question d'introduire les pays participants sans médailles à nos données. Pour ce faire, il nous fallait la liste des pays participants, ou un moyen de les identifier. Pour 2024, le fichier nocs.csv indiquait tous les pays participants, et ce sur plusieurs années, on pouvait notemment y retrouver la Russie, et le Russian Olympic Comitee (ROC), qui ne participaient pas en 2024. Mais nous avons compris que la colonne "note", si elle était remplie d'un "P", signifiait que le pays participait en 2024. Nous avons pu donc obtenir les pays participants aux J.O. de 2024, et nous les avons ajoutés à notre DF.

Pour 2020, nous aovns du adopter une autre stratégie : nous n'avons pas trouvé de fichier reportant les pays participants de 2020. Il aurait été possible de le créer nous-même, mais nous avons eu une autre idée : utiliser le fichier des athlètes participants aux J.O. de 2020. En effet, les athlètes participants sont forcément rattachés à un pays participant, il nous a donc suffi de récupérer tous les pays apparaissant dans ce fichier. Nous avons donc ainsi pu trouver nos pays maquants, et les ajouter à nos données.

Après avoir obtenu nos deux DF sur 2020 et 2024, nous nous sommes doutés que ces données ne suffiraient pas ni pour faire une analyse, ni pour entraîner un modèle de prédiction. Nous avons alors cherché de nouvelles données, et nous sommes tombés sur un lien Gigasheet regroupant tous les athlètes ayant participé à des Jeux, en détaillant entre autre leur pays, leur sport, l'année et les Jeux de participation, ainsi que l'obtention ou non de médailles. En filtrant alors par Jeux, nous avons pu obtenir les médailles par pays, et tous les pays participants, en utilisant une méthode similaire à celle que nous avons utilisé pour 2020. La plus grosse difficulté, dans cette partie, a été de trouver comment identifier les médailles, car nous obtenions souvent des résultats trop élevés. Cette augmentation venait du fait que nous comptions le nombre de médailles par athlète, et donc que pour un sport en équipe, nous comptions trop de médailles (par exemple, pour l'escrime, nous comptions 3 médailles en trop par rapport au classement des médailles). Une fois ce problème corrigé, en filtrant par Jeux, sport et event, et en enlevant les doublons, nous obtenions un nombre de médailles plus cohérent.

Nous avons terminé le traitement des données en inscrivant le rang de chaque pays, qui était déjà présent dans certaines données, mais qu'il fallait réécrire étant donné l'ajout de nouveaux pays. Une autre difficulté de ce projet a été de réussir à recréer le type de rang qu'utilisent les J.O., avec notemment le fait que si deux pays ont le même nombre de médailles d'or, d'argent et de bronze, ils ont le même rang, et que le rang d'après "saute" (par exemple : si il y a deux pays en 2e position, le prochain pays sera 4e, et il n'y aura pas de 3e).

Pour finir, nous avons procédé à une vérification de nos résultats, en comparant nos DF obtenus avec les classements officiels. Nous nous sommes alors rendus compte que nos résultats n'étaient pas les mêmes que ceux attendus. Il y avait des médailles qui n'étaient pas au bon endroit. Nous avons donc cherché manuellement d'où venait le problème, et nous en avons tiré la conclusion suivante : les données que nous avons ne prennent pas en compte la disqualification d'athlètes et le report de médailles suite à une disqualification. Nous avons longuement réfléchi à ce problème, et étant donné la difficulté que nous avions éprouvée à trouver des données sur les Jeux, et le temps et l'effort que le changement manuel des médailles prendrait, nous avons décidé de laisser nos données telles qu'elles sont, avec les médailles qui ont été retirées plus tard. Ainsi, notre analyse et nos modélisations ne reflèteront pas les classements actuels remaniés, mais elles peuvent tout de même donner une idée du classement et de l'impact de la participation ou non d'un pays.

3 Analyse Exploratoire

Dans cette partie nous allons expliquer comment nous avons analysé les données pour comprendre l'impact de la Russie sur les JO. Tous les graphiques et codes utilisés sont dans le Google Collab cité au dessus.

Dans un premier temps, nous avons regardé quelle était la place de la Russie dans les JO précédents, et en particulier en 2020. La Russie se classe en 7e place. Pour mieux comprendre ce classement nous avons regardé plus en profondeur les médailles de la Russie. Pour ce faire nous avons utilisé le fichier medals_2020 fourni sur le moodle. Nous avons filtré toutes les lignes ayant pour country_code=ROC. Et c'est ici que nous avons eu notre premier problème. Le fichier medals_2020 contient tous les athlètes ayant obtenu une médaille, par conséquent il y a par exemple, 15 athlètes possédant une médaille d'or en Handball alors que la Russie ne possède qu'une seule médaille d'or dans cette discipline. Ce problème vient du fait qu'il y a 15 athlètes dans une équipe de hand et donc 15 athlètes récompensés dans le fichier. Nous devons donc filtrer de nouveau notre dataframe. Une fois ceci fait nous pouvons voir les sports dans lesquels la Russie a le plus de médailles. Nous en avons tracé un graphique. Mais ce n'est pas suffisant, en effet toutes les médailles n'ont pas le même impact, c'est pourquoi il faut de nouveau diviser nos résultats pour pouvoir savoir quel sport rapporte le plus de médailles d'or. Nous en avons de nouveau tracé un graphique dans le Google Collab. Avec cela nous avons remarqué que la lutte et l'escrime sont sûrement les sports les plus intéressants pour la Russie.

Dans un second temps nous nous sommes rendus compte assez vite que les médailles avaient un poids. Une médaille d'or vaut beaucoup plus qu'une médaille d'argent, si bien que dans les premières place du classement les places sont presque exclusivement décidées via les médailles d'or. L'argent et le bronze ont peu d'influence sur le classement. Nous essayons donc de comprendre pour quels pays le classement ne dépend que des médailles d'or, et si la Russie en fait partie. Pour ce faire nous avons tracé une boite à moustache afin de visualiser les pays qui sont grandement au dessus des autres en terme de nombre de médailles d'or. Grâce a ceci nous pouvons voir que les 11 premiers pays sont en dehors de la boite (et ainsi la Russie) et que par conséquent leur classement sera grandement défni par leur nombre de médailles d'or exclusivement. Une fois cette étape effectuée nous nous sommes rendus compte que ce nous faisions s'apparente beaucoup à du clustering, et nous avons décidé de laisser ça pour une partie dédiée qui sera détaillée plus tard dans ce mémoire.

Par la suite nous nous sommes intéressés à la répartition des médailles d'or entre les pays. Pour ce faire nous avons tracé des graphiques camembert pour mieux le visualiser. Nous avons tracé 2012, 2016, 2020 et 2024. La première chose que l'on remarque est que les camemberts 2012 et 2016 sont très semblables avec les Etats-unis en haut du classement, avec en 2e et 3e le Royaume Unis et la Chine. Néanmoins on remarque que les Etats-Unis dominent. Alors qu'en 2020 et 2024 il y a une égalité entre la Chine et les Etats-Unis. La seule chose qui diffère réellement entre 2020 et 2024 est que les 2 premiers pays ont plus de médailles d'or (on ne regarde ici que la distribution des médailles entre les places, pas les pays qui ont pris les places). On remarque donc qu'il y a un vrai changement entre 2016-2012 et 2020-2024. On peut aussi voir que la Russie

perd quelques place entre 2016 et 2020, c'est peut-être du au fait que la Russie n'est que partiellement autorisée à participer au JO en raison de problèmes géopolitiques, ainsi certains athlètes sont absents, et par extension des chances de médailles en moins.

Pour continuer nous voulions savoir où littéralement sont allées les médailles de la Russie en 2024. C'est à dire voir pour une médaille donnée, quelle pays l'a récupérée. Nous avons donc en premier temps trié le dataframe medals_2020 pour récupérer uniquement les médailles et les sports associés à la Russie. Par la suite nous avons changé le nom des sports qui étaient différents entre le fichier de 2020 et 2024. C'est à ce moment qu'un problème est apparu, en boxe les catégories de poids ont été modifiées. Nous avons donc essayé de faire des équivalences entre les anciennes catégories et les nouvelles. Il y a eu quand même un problème, une des catégories a disparu, nous l'avons donc simplement supprimée du fichier pour éviter les errreurs (pas la solution idéale mais il est compliqué de faire autrement). Ensuite avec une simple boucle nous avons pu récupérer quels pays ont récupéré quelles médailles.

Une fois la liste des pays et médailles récupérée nous avons tracé divers graphiques pour mieux comprendre nos données. La tendance que l'on remarque le plus est que le haut du classement de 2024 est celui qui récupère le plus de médailles de la Russie. Ensuite les pays sont plus disparates. La conclusion que l'on peut avoir est que le haut du classement ne fait qu'agrandir son écart avec le bas, et que les changements intéressants peuvent s'effectuer dans d'autres endroits du classement. Nous avons néanmoins trouvé des limites à cette étude, nous ne regardons ici que les médailles directes de la Russie, or on peut s'imaginer qu'il y a un roulement. Par exemple si la Russie à la médaille d'or en 2020, alors sans la Russie en 2024 tout le classement se décale (de manière simplifiée) et ainsi nous ne regardons pas quel pays récupère la médaille d'argent, ni celle de bronze.

Pour finir notre analyse exploratoire nous avons regardé les variations dans le classement de chaque pays. C'est à dire entre 2 années a quel point le pays a bougé dans le classement. Pour ce faire nous avons mis dans un même tableau les classements de deux années que nous voulons comparer 2 à 2 (2020 et 2024 par exemple). Ensuite nous avons juste créé une colonne en plus qui contient la soustraction du classement de chaque pays de la première année au classement de la deuxième année. Ainsi avec un pays on a une colonne qui représente sa variation dans le classement.

Avec cette colonne nous avons pu facilement calculer la moyenne des variations et l'écart type. Pour la moyenne, celle qui ressort le plus est celle entre 2016-2020. Il y a en moyenne un gain de place de +3.5 contrairement aux autres qui sont aux alentours de +0.75. C'est une observation qui vient confirmer ce que l'on avait vu sur les camemberts précédemment, il y a de vraies différences dans le classement entre 2012-2016 et 2020-2024. Nous avons plusieurs suppo-

sitions de raisons à cela (non vérifiées): la pandémie du Covid-19 qui a décalé les JO de 1 an, il y a aussi le fait que la Russie ne participe pas complètement. L'écart type, quant à lui, ne change pas beaucoup.

Une fois ces observations faites, nous avons voulu regarder de manière plus précise ces statistiques. Nous avons donc divisé le classement par groupe de 10 (les clusters n'étant pas faits à ce moment-là).Nous avons rajouté une ligne de zéros a 2012 car il y avait moins de pays, les statistiques de la neuvième ligne sont donc peu fiables. Pour chaque groupe nous avons regardé ses statistiques. Le but de cette manipulation était de voir si une partie du classement bougeait plus qu'une autre. Pour que ce soit plus facile à voir nous en avons tracé un graphique pour comparer les variations du classement entre les variations des différentes années. Ce que nous pouvons voir est que les parties du classement qui ont le plus de variations différentes se trouvent à partir de la 40-50e place. Les premiers groupes, quant à eux, varient mais pas de manière anormale.

Pour conclure cette analyse exploratoire nous pouvons commencer par dire que ces observations ont des limites, en effet nous regardons qu'un jeu de données assez petit donc il peut y avoir des biais. Néanmoins ce que nous pouvons en tirer est que la non participation de la Russie n'est pas forcément ce qui fait bouger le plus le classement, entre 2016 et 2020 il y a une grande différence, or la Russie participe toujours.

Ce que l'on peut dire par contre est que en général le haut du classement ne change pas beaucoup, mais dès que l'on se dirige vers le milieu de tableau c'est la que les vrais changements se passent. Ce qui peut paraître assez logique car c'est là que les variations de médailles ont le plus d'impact.

4 Clustering

Cette partie est dédiée à la documentation de nos procédés et difficultés quand à la création de clusters, pour regrouper les pays participants.

Comme dit dans la partie précédente, nous avons eu l'idée de regrouper les pays participants aux J.O., pour mieux étudier leur impact sur le classement, et réaliser des modèles plus précis en prenant en compte leur groupe. Nous avons réalisé des clusterings sur les années seules, mais aussi sur le groupe 2012-2016-2020 pour nous rendre compte de la tendance globale des pays.

Pour chaque clustering, nous avons procédé de la manière suivante : un CHA sur les données, puis, en analysant la distance entre les groupes sur le dendrogramme, un k-means avec le nombre de classes qui nous paraissaient pertinentes. Pour visualiser le k-means, nous avons utilisé à chaque fois deux graphiques 2D. Ces graphiques étaient suffisants pour bien visualiser les clusters faits sur des Jeux seuls, puisqu'il n'avaient que 3 dimensions, mais pour le cluster regroupant

3 années, nous avons utilisé une ACP, qui regroupe un groupe de variables en une seule (et ainsi nous pouvons afficher un graphique 2D reprennant la plupart des variables).

Pour ce qui est du choix de nos variables, nous avons expérimenté avec plusieurs paramètres pour choisir celui qui nous semblait le plus pertinent sur les données de 2020. Nous voulions former des groupes qui ordonnaient les pays en terme de "niveau", c'est à dire les pays influencés par les médailles d'or (par exemple, la France), les pays influencés par celles d'argent, ceux influencés par les médailles de bronze, et les pays sans médailles. Nous avons commencé en utilisant le nombre total de médailles d'un pays, le rang des pays, et le nombre de médailles de chaque type, mais nous obtenions des clusters avec des pays sans médailles et des pays avec 7 médailles d'or dans le même groupe, ce qui ne nous parraissait pas cohérent. Nous nous sommes alors rendus compte que nous avions un problème de pondération : en effet, une médaille d'or ne vaut pas une médaille de bronze. Nous avons donc introduit l'idée de pondérer les médailles. Nous avons commencé par chercher des pondérations usuelles pour les médailles des Jeux, mais nous avions des pays qui résussissaient à compenser malgré cela. Nous avons donc essayé arbitrairement la pondération 100:10:1 (une médaille d'or vaut 100 médailles de bronze, une médailles d'argent en vaut 10), en pensant que dans le classement, la valeur d'une médaille d'or ne peut jamais être atteinte par un nombre de médailles d'argent ou de bronze. En utilisant cette méthode, nous arrivions déjà à un résultat auquel nous nous attendions.

Seulement, si les groupes étaient plus cohérents, il n'étaient pas ceux auquels nous nous attendions : le groupe des pays du bas du classement regroupait jusqu'aux pays possédant 3 médailles d'or, et était considérablement plus peuplé que les autres groupes ; et nous avions beau augmenter le nombre de clusters, ce groupe ne changeait pas : c'étaient les groupes plus petits qui se fragmentaient de plus en plus. Nous nous sommes alors rendus compte de la distance entre les pays du bas/milieu du classement par rapport aux autres pays : la distance entre les pays à 0 médailles et les pays à 3 médailles d'or était plus petite que la distance entre deux pays du haut du classement.

Comme nous tenions à avoir une vue plus fragmentée du bas du classement, nous avons ensuite choisi de faire un nouveau clustering sur le bas du classement, que vous pouvez retrouver dans le Collab. Nous avons également rapidement expérimenté avec la distance de Manhattan, que nous n'avons pas gardé pour la suite, car nous étions satisfaits de nos résultats. Nous avons ensuite fait un clustering sur les données des Jeux de 2024, avec le rang puis en utilisant notre pondération 100:10:1.

Pour ce qui est du cluster regroupant 2012-2016-2020, nous avons tout d'abord testé avec toutes les données (or, argent, bronze, total et rang) sans pondérer, mais nous obtenions le même problème que précédemment, nous avons donc pondéré et ainsi obtenu le dernier clustering, qu'il est possible de retrouver

sous l'onglet "Clustering 2012-2016-2020 avec pondération".

5 Modèles prédictifs

Cette section aborde les différents modèles prédictifs que nous avons exploités ou non, avec des justifications et quelques observations.

Nous avons tout d'abord commencé par le modèle le plus simple : la régression linéaire. Nous voulons prédire le résultat de l'année X à partir des années X-4 et X-8. Nous utilisons les données sur le nombre de médailles d'or, d'argent et de bronze et nous prédisons le nombre de médailles d'or, d'argent et de bronze qui seront obtenues à l'année X, pour ensuite recréer le rang et comparer. Les hypothèses nécéssaires à la régression sont partiellement présentes, ce qui limite nos conclusions : les variables ont bien une relation linéaire, il y a plus de 20 observations par variables, la distribution des médailles est bien normale et les variables sont supposées indépendantes (nous avons demandé à Gemini comment faire un test, et il nous a donné le test du chi carré, que nous avons fait, et avec lequel nous n'avons pas refusé l'hypothèse d'indépendance). Nous ne savons pas comment mesurer l'homoscédastiscité car nous n'avons pas utilisé OLS (ce qui est une amélioration à faire). Par contre, le pairplot fait en début de partie régression suppose qu'il y a multicolinéarité.

Les résultats que nous avons obtenus n'étaient pas très surprenants : le tableau des médailles était assez fidèle au tableau réel, si ce n'est à l'exception de quelques pays comme la Grande-Bretagne qui, par ses meilleurs placements aux J.O. de 2016 et 2020, prend la 3e place, ce qui n'a pas été le cas en 2024. On remarque également que la différence entre les rangs, pour les prévisions avec et sans la Russie, ne diffère que d'un (ou de deux car il y a un autre pays qui n'apparait pas en 2024), ce à quoi nous nous attendions. Nous n'avons aucun moyen de faire comprendre à la machine que les médailles non-gagnées par la Russie ont été redistribuées, et donc le classement ne change que parce que la ligne de la Russie n'apparaît simplement pas.

Pour ce qui est de la prédiction de 2024 sans la Russie, on remarque que dans le graphique "Comparison of Predicted and Actual Ranks" du menu "Régression Linéaire", la prédiction est plutôt bonne sur le haut du classement et sur le bas du classement. C'est parce que ces points là sont les points les plus stables du classement, et qu'il y a beaucoup plus de fluctuation au milieu du tableau des médailles, ce qui le rend plus dur à prédire.

Nous avons calculé la R2 et la MSE de notre modèle qui sont respectivement approximativement: 0.877 et 2.93. On peut voir dans un premier temps que la R2 est très bonne, donc le classement est plûtot fiable. Pour la mse on peut considérer qu'elle n'est pas mauvaise. En effet nous prédisons le nombre de médailles et ici l'erreur est en moyenne +- 1.7 (racine de 2.9). Sachant que la

majorité des pays possède plusieurs médailles, on peut conclure que le modèle prédit plutôt bien.

Nous avons aussi tenté de faire un essai de prédictions rudimentaires en fonction des clusters. Le but était de récupérer à partir des clusters faits précédemment la classe de chacun des pays. Ensuite le modèle attribuait juste une place aléatoire a l'intérieur de son cluster. Et la combinaison de toutes les places et de tous les clusters nous donnait un classement. Par exemple si l'on prend le cluster du haut du classement (Chine, USA), le modèle attribura la première place à l'un des deux aléatoirement et la deuxième a l'autre. Ce modèle n'est pas très cohèrent, et très rudimentaire. Les résultats ne sont pas bon, mais nous avons quand même décidé de le laisser pour montrer l'idée.

Nous avons également essayé de faire une régression polynomiale, sans succès.

Nous avons ensuite, comme il nous a été suggéré de le faire, tenté de faire une régression logistique. Cette dernière a pour but de prédire une variable binaire, donc notre choix de prédiction s'est porté sur l'appartenance ou non à un cluster. Le but était donc de prédire, à partir des clusters trouvés pour les années précédentes, à quel cluster allait appartenir un pays en 2024. Après quelques essais, il est apparu qu'il fallait transformer les données de sorte qu'on obtienne des variables numériques indépendantes. Pour cela, c'est la vectorisation en vecteurs de valeurs binaires qui a été utilisée. Au début, nous pouvions constater que le score d'accuracy était plutôt bon avec le dataset de test, quelque soit le cluster. C'était encore mieux pour les clusters prédits pour 2024, mais nous nous sommes aperçus que le score était trompeur, car en effet, après avoir regardé en détail les prédictions, le modèle prédisait uniquement le cluster 0. Le score d'accuracy était bon uniquement car les données étaient nombreuses et très regroupées autour du cluster 0, ce qui fait que le modèle ne prédisait aucun pays dans le cluster 3 par exemple, alors qu'il y avait 2 pays dans le cluster 3, mais comparé au grand nombre de pays, cette erreur était considérée comme moindre. L'échec de cette méthode pourrait être dû au fait que les conditions sur les données ne sont pas remplies pour faire une régression linéaire, ou bien qu'il fallait introduire des pondérations en fonction du cluster (par ex. x10 pour le cluster 3, x5 pour le 2, etc), mais nous n'aurions pas pu reporter ces pondérations sur 2024 puisque cela suggère de connaître les clusters, soit ce que nous cherchons à prédire.

Pour ce qui est des arbres, nous avons tout d'abord écarté l'arbre de classification, car nous pensions que le seul moyen de l'utiliser serait de créer une classe par rang, ce qui est non seulement illisible mais impossible dans notre situation. En effet, il est impossible de savoir à l'avance combien de rang il y aura à l'issue des Jeux. Et dans le cas où nous utiliserions moins de classes, cela reviendrait à un clustering, que nous avons déjà fait. Pour ce qui est de l'arbre de régréssion, nous l'avions gardé en tête mais nous l'avons également écarté : soit l'arbre prédirait correctement un rang, mais il serait alors nécéssairement en overfit,

soit il donnerait une idée du rang, ce que nous avons déjà avec le clustering. Les arbres de décision ne nous apportant aucune information supplémentaire, nous avons donc décidé de les laisser de côté.

Nous avons finalement essayé de faire des réseaux de neurones. Cette partie nous a posé problème, car nous avons fait de multiples essais, mais nous ne parvenions pas à obtenir de modèle qui soit correct. Nous avons pour la plupart des cas obtenus des score très faibles, même sur les datasets de train, ce qui signifie que nous étions en underfit. Les idées qui nous semblaient intéressantes mais qui n'ont pas abouti étaient notamment d'inclure pour un DF pour les données des deux JO précédents sur les médailles par type et le rang. Ayant obtenu un score très faible, nous avons conclu que beaucoup de données seraient nécessaires ce qui signifiait qu'il fallait également tout corriger manuellement, car beaucoup d'erreurs et de biais sont présents sur les données que nous avons pu trouver en ligne, principalement à cause des médailles supprimées (pour cause de disqualification). Néanmoins après de nombreux essais nous avons réussi a créer un réseau de neuronne avec une R2 de 0.75, ce qui est plutôt encourageant. Néanmoins le résulat nous étonne lorsque l'on regarde a l'oeuil il y a peut être une erreur dans le calcul. C'est en effet parce que nous n'avons pas regardé l'accuracy, qui elle est de 0.02, donc très faible. Notre réseau est donc moins performant que la régression linéaire mais arrive a prédire le haut du classement assez bien. Il semblerait que plus l'on augmente Dense(128, activation='relu'), plus le modèle est précis. Mais il y a une limite évidente ou le modèle ne s'amèliore plus beaucoup. On peut donc dire que la régression linéaire est plus interréssante dans notre cas. La mse est très haute dans le réseau de neuronnes mais c'est normal car il y beaucoup de pays mal prédits.

6 Conclusion

Pour conclure ce mémoire nous allons revenir sur notre problématique initiale: quel est l'impact de la Russie sur le classment des médailles des JO 2024. Nous avons traité cette question en partie dans l'analyse exploratoire. Nous partions avec comme idée que l'abscence de la Russie change le classement du haut du tableau. Néanmoins après analyse nous avons pu observer quelque chose de différent. Dans un premier temps entre 2020 et 2024 l'abscence de la Russie ne semble pas faire bouger de manière anormale le haut du classement, si ce n'est augmenter l'écart des deux premiers avec ceux d'en dessous. Ensuite nous avons vu qu'il y a de grande différence dans la répartition des médailles entre les années 2012, 2016 et 2020,2024. Ainsi le réel changement dans le classement intervient en 2020. Le classement semble donc changer pour d'autres raisons que l'abscence de la Russie. En nous re-centrant sur notre problématique, nous avons aussi vu que finalement, les écarts de classement anormaux surviennent plus au milieu/bas du classement qu'en haut. Ainis nos aprioris ont été changé, le classement n'a pas eu les mouvements attendus. Il reste néanmoins difficile

d'estimer si la raison de ces mouvements est dû a la Russie ou a une autre cause. C'est pourquoi nous avons regardé plus en profondeur comment ont été réparties les médailles de la Russie.

Pour répondre à nos problématiques, nous avons sélectionné des données de sources externes sur les JO précédant ceux de 2020, et après les avoir traités pour qu'elles soient utilisables dans notre projet, nous avons commencé une amorce de modélisation avec le clustering. Nous avons durant nos recherches en analyse exploratoire réalisé que faire des clusters était nécessaire pour avoir une idée du niveau de la Russie mais aussi pour avoir une idée de la dynamique du tableau. De savoir par exemple quels types de pays auront un classement majoritairement défini par leur médaille nombre de médaille d'argent. Ce qui nous a permis d'aboutir a un grand groupement en cluster sur plusiers années de tous les pays participants cohérent. Ainsi nous avons chaque pays classé dans une classe adéquate, ce qui nous fournit une compréhension supplémentaire du classement.

Pour ce qui est des modèles de prédictions, nous n'avons pas mené toutes nos idées à terme à cause de difficultés à recueillir et traiter assez de données pour réaliser une bonne prédiction du classement. Nous avons néanmoins pu réaliser une régression linéaire qui semble fonctionner de manière correcte, et amorcé un réseau de neurones, en plus d'autres essais. Et nous avons exploré de nombreuses pistes, ce qui nous a permis notamment de rejeter l'idée d'utiliser les arbres de décision. Les autres modèles était soit trop difficiles à mettre en place soit peu performants.

En finalité, ce projet nous a permis de comprendre en détail l'impact de la non-participation de la Russie aux JO 2024 sur le classement, et de manipuler nos cours d'intelligence artificielle afin de mettre en place des modèles de prédictions.

7 Sources

Notre collab : https://colab.research.google.com/drive/1JqjvWgiQoxYj9OoiN-asNxkSgHWGJZj1?usp=sharing

Les données supplémentaires : https://www.topendsports.com/events/summer/medal-tally/rankings-weighted.htm

Le site sur la pondération des médailles : https://www.topendsports.com/events/summer/medal-tally/rankings-weighted.htm

Un site expliquant comment faire le réseau de neurones https://www.cours-gratuit.com/tutoriel-python/tutoriel-python-matriser-les-rseaux-de-neurones-avec-keras#_Toc58080506