How to get the Bookies grumpy

Auteurs : **Guillaume Gréau, Guillaume Oillo, Georges Sleimen**

Table des matières

1 Introduction

L'objectif de notre projet consiste à prédire l'issue de tous les matches de tennis masculins professionnels (ATP) qui se sont déroulés lors des deux dernières décennies, et de comparer nos résultats avec ceux qu'auraient obtenus les bookmakers sur la même période.

Pour ce faire, nous avons d'une part enrichi notre dataset d'origine avec des données théoriquement corrélées au vainqueur de chaque rencontre, et d'autre part nous avons bien sûr utilisé différents modèles predictifs de machine learning afin d'obtenir les meilleurs résultats possibles.

Par ailleurs, tout au long du projet, nous avons veillé à toujours présenter les données de chaque observation telles qu'elles étaient disponibles au matin de chaque match à prédire, ceci dans le but d'éviter de communiquer à nos algorithmes des informations qui n'étaient pas encore disponibles le jour où la rencontre a eu lieu, et donc qui nous avantageraient par rapport aux bookmarkers que nous cherchons à « rendre grincheux » (cf. « How to get the Bookies grumpy »).

2 Présentation du dataset

Notre dataset est une version améliorée du dataset téléchargeable sur kaggle et dont le lien est fourni dans la fiche projet Datascientest (https ://www.kaggle.com/code/edouardthomas/beat-the-bookmakers-with-machine-learning-tennis).

En effet, avons utilisé la même source que l'auteur du projet kaggle, à savoir le site http ://tennis-data.co.uk/alldata.php, et avons téléchargé puis assemblé toutes les années disponibles, de 2000 à 2021 inclus. Notre dataset contient donc tous les matchs masculins joués entre le 03/01/2000 et le 21/11/2021, qui est la date à laquelle s'est joué le dernier match de l'année 2021, ainsi que les cotes pour ces mêmes matches, fournies au fil des années par plus d'une dizaine de bookmakers. Parmi les fichiers csv, il s'agit du fichier **atp_data_1.csv**

2.1 Présentation des variables du dataset

Pour chacun des matchs nous avons plusieurs informations à disposition que nous allons vous lister ci-dessous :

- Date : Date à laquelle le match s'est joué.
- Tournament : Nom du tournoi dans lequel le match a eu lieu. Par exemple Roland-Garros, Wimbledon, etc...
- Location : Ville dans lequel le tournoi est situé.
- Series : Catégorie à laquelle appartient le tournoi. Cette variable sera expliquée plus en détails dans la suite.
- Court : Type de court. Les modalités sont indoor et outdoor. Le match a eu lieu en extérieur ou en salle.
- Surface : Surface sur lequel le match a eu lieu. Les modalités sont hard, clay, grass et carpet.
- Round : Round du tournoi auquel s'est joué le match. Les modalités sont 1er tour, 2eme tour, 3eme tour, 8ème/quart de finale, demi-finale et finale.
- Format : Format du match. Les modalités sont 3 sets ou 5 sets.
- Winner : Le nom du gagnant du match.
- Loser : Le nom du perdant du match.
- WRank : Le rang (classement) du vainqueur du match.

- LRank : Le rang (classement) du perdant du match.
- WPts : Le nombre de points gagnés par le gagnant sur l'année en cours. Ce sont les points qui permettent d'établir le classement du joueur.
- LPts : Le nombre de points gagnés par le perdant sur l'année en cours.
- Wsets : Le nombre de sets remportés par le vainqueur du match.
- Lsets : Le nombre de sets remportés par le perdant du match.
- Comment : La manière dont le match s'est terminé. Les modalités sont Abandon, Forfait, Disqualifié.
- elo_winner : Points elo du vainqueur. Ils sont propre au système de classement elo et non au classement ATP (appelation du classement propre au tennis).
- elo_loser : Points elo du loser.
- proba_elo : Probabilité que le gagnant l'ai emporté en se référant au système de classement elo.
- Toutes les autres variables concernent les côtes estimées par les bookmakers. Ces variables ne sont d'aucune utilité pour prédire l'issue des matchs mais seront utilisées lorsque l'on voudra comparer nos résultats.

Expliquons plus en détails les modalités de la variable « Series ». Cette variable est constituée de 5 modalités qui sont les catégories auxquelles appartient les différents tournois : les tournois ATP 250, ATP 500, Master 1000, Grand Chelem et ATP Final Tour. Ces tournois n'ont pas le même prestige et il est primordial dans la suite de les différencier. Les tournois ATP 250 constituent les tournois les moins prestigieux du circuit. Vient ensuite les ATP 500, puis les Masters 1000, l'ATP Final Tour et pour terminer les Grand Chelems. Le circuit est principalement composé de tournois ATP 250 et 500. Il existe entre 12 et 14 Masters 1000 (cela dépend des saisons). Un seul ATP Final Tour et 4 Grand Chelems. Il est important de les différencier car il est bien plus complexe de remporter un Grand Chelem que gagner un ATP 250. Dans un tournoi ATP 250, le nombre de tours à passer pour remporter le tournoi est inférieur à celui d'un Grand Chelem et il est surtout constitué de joueurs moins performants. Leur différenciation réside également dans le nombre de points gagnés dans les tournois. Un ATP 250 offrira seulement 250 points au vainqueur et 150 au finaliste contre 2000 pour un vainqueur en Grand Chelems et 1200 au finaliste. Cet important écart est conservé pour les différents tours du tournoi : 1er, 2eme, 3eme tour, 8ème, quarts et demi-finales. Le système d'attribution des points ne sera pas présenté dans ce rapport, bien qu'utile dans la réalisation du projet, mais est simplement mentionné afin de mieux comprendre ce qui sera présenté par la suite.

Maintenant que toutes les variables ont été introduite. Nous allons présenter leur intérêt pour la suite. Parmi toutes ces variables, nous avons fait le choix de supprimer toutes les variables propres au joueur : « WRank », « LRank », « WPts », « LPts », « Wsets », «Lsets », « elo_winner », « elo_loser » et « proba_elo ». Et de ne garder que les variables propres au déroulement du match : « Tournament », « Location », « Series », « Court », « Surface », « Round », « Format », « Winner » et « Loser ».

2.2 Choix des variables à conserver

Énumérons, pour chaque variable, les raisons pour lesquelles nous ne souhaitons pas les conserver.

- La variable « Rank » (WRank/LRank) ne rends pas compte de la différence de points entre les joueurs. Prenons un exemple avec 3 joueurs, l'un classé 12ème, l'autre 13ème et le dernier 14ème. Celui qui est classé 12ème à 500 points de plus que le 13ème et le 13ème a 20 points de plus que le 14ème. Si l'on s'en tient au classement, la différence de points (ici il est question de différence de places dans le classement) entre le 12ème et le 13ème et celle entre le 13ème et 14ème sont identiques. En revanche, si on s'en tient aux points, la différence de points est plus importante entre le 12ème et le 13ème, traduisant ainsi de bien

meilleures performances pour le joueur classé 12ème. Des victoires dans des tournois plus prestigieux par exemple donc contre de meilleurs joueurs. Tandis que pour le 13ème et le 14ème les performances semblent quasiment identiques entre elles.

- Comme mentionné précédemment, la variable « Pts » (WPts/LPts) pourrait être très intéressante pour la suite mais elle ici elle est mal renseignée. Premièrement, il manque toutes les valeurs entre 2000 et 2005, ce qui constitue un quart de la statistique. De plus, le système d'attribution des points a changé à partir d'une certaine année, impliquant ainsi que les différences de points entre les 2 joueurs seront globalement plus faibles avant qu'après le changement et cela pourrait perturber l'algorithme. Il est important de garder le même système de points pour tous les matchs du dataset.

- Le système de classement Elo et ses probabilités (elo_winner/elo_loser/proba_elo) constitue un élément intéressant mais il ne nous semble pas adapter dans le cadre du tennis. Fonctionnement du classement Elo : Le classement Elo attribue au joueur, suivant ses performances passées, un nombre de points (« points Elo ») tel que deux joueurs supposés de même force aient le même nombre de points. Plus le joueur est performant et plus son nombre de points Elo est élevé. Si un joueur réalise une performance supérieure à son niveau estimé, il gagne des points Elo. Réciproquement, il en perd s'il réalise une contre-performance. (Définition Wikipédia). Le classement Elo ne semble pas adapté au tennis car il ne retranscrit pas certains paramètres. Supposons qu'un joueur soit dans une grande forme et qu'il se blesse. Après plus de 6 mois d'absence, celui-ci reprend la compétition avec peu de confiance et peu de repères mais avec un nombre de points Elo qui est exactement le même que celui qu'il avait avant sa blessure. Par conséquent, le nombre de points qu'il possède lorsqu'il revient de blessure est complétement biaisé. Nous avons pris un exemple fort pour appuyer les points faibles de ce classement mais il ne fonctionne pas que dans le cas d'une blessure. Un autre exemple, un joueur possédant un nombre de points Elo important mais qui subitement se met à enchaîner les défaites va voir son nombre de points Elo diminuer très lentement contrairement au classement ATP qui va retranscrire bien plus rapidement la méforme actuelle de ce joueur.

  Maintenant nous allons expliquer l'intérêt de garder certaines de ces variables et de l'utilisation que nous allons en faire. Nous avons décidé de créer, à partir de ces informations-là, de nouvelles variables qui nous semblaient pertinentes pour prédire l'issue d'un match.

  Mais avant de rentrer dans les détails de cette partie, nous allons résumer les différentes étapes effectuées pour nettoyer le dataset car il représente une grande partie du projet et fut déterminant pour l'implémentation des nouvelles variables.

3 Nettoyage et correction du dataset

Le nettoyage et correction a été effectué à partir du fichier **bkm_clean.ipynb** sur le dataset **atp_data_2.csv**. En sachant que ce dataset est le résultat de quelques modifications supplémentaires avec l'ajout des colonnes « Year », « Week_number » et « Year_week ». Toutes les étapes effectuées sont résumées ci-dessous :

1. Vérification de la présence de lignes dupliquées ou de valeurs non renseignée. Le dataset ne comportait ni de lignes dupliquées, ni de valeurs non renseignées.

2. La variable « Series » comprend les anciens noms des catégories ainsi que les nouveaux. Par exemple, la catégorie International correspond à ATP 250 aujourd'hui. Afin d'être cohérent pour la suite, nous avons renommé les anciens noms par les nouveaux.

3. Par soucis de commodité, nous avons raccourci le nom des modalités des variables « Series » et « Round ».

4. Nous avons corrigé le nom de certains noms de tournois et de joueurs qui ont été ajouté avec un espace.

5. Tout comme la variable « Series », plusieurs tournois ont changé de nom au cours du temps, donnant lieu ici à plusieurs modalités différentes pour un seul et même tournoi. Par exemple, le tournoi d'Indian Wells qui est retranscrit sous 3 appellations différentes. Pour regrouper ces tournois, nous avons utilisé la

localisation qui constitue le seul point entre eux car en général nous avons qu'un seul tournoi par ville sauf rare exception où il peut y avoir 2 tournois mais ils sont de catégories différentes (GC, M1000, ATP500, ATP250). Nous avons donc rassemblé les tournois qui ont changés de noms en un seul nom lié à la ville dans laquelle ils se déroulent. Pour les 4 tournois du GC on leur donnera leur propre nom mais quand il s'agira d'un master 1000 on rajoutera l'attribut Master et si c'est un ATP 500 ou 250 on donnera l'attribut Championship.

6. Inversement, certains tournois ont conservé la même appellation mais ont subit un changement de court ou de surface au cours du temps. Comme par exemple, le tournoi de Madrid qui a d'abord été joué sur surface dur puis ensuite sur terre battue. Nous voulons donc faire en sorte de différencier à nouveau le nom des tournois qui ont subi ces changements.

7. Selon la catégorie à laquelle appartiennent les tournois, le nombre de tours n'est pas le même. Le problème est que le dataset est organisé de telle manière que pour tous les tournois le 1er tour sera toujours le 1R et le dernier la finale F. Les GC comptent 7 tours : 1R, 2R, 3R, 8th, Q, SF, F mais les ATP 250 n'en contiennent que 5 et l'enchaînement des tours se fait de la manière suivante : 1R, 2R, Q, SF, F. Sauf que cette notation pose un problème pour la suite. Nous voulons que la succession des tours soit cohérente avec celles des autres types de tournois. On voudrait qu'ici 1R, 2R, Q, SF, F se transforment en 3R, 8th, Q, SF, F. Nous ne voulons pas de saut dans le nom des tours.

8. Et pour terminer, nous avons réorganisé le dataset de sorte à avoir les matchs ordonnés par ordre chronologique. Regroupés ensuite par noms de tournois, c'est-à-dire que nous ne voulons pas trouver le match d'un tournoi entre 2 matchs d'un autre tournoi. Et puis ordonné par rapport à l'enchaînement des tours : 1R, 2R, 3R, 8ème, quart, demi-finale et finale.

Le dataset complètement nettoyé et corrigé est sauvegardé dans le fichier **atp_data_3.csv**.

4 Étude statistiques et choix des variables à implémenter

Les graphiques qui vont suivre ont été réalisé à partir du dataset **bkm_plot_and_create_variables.ipynb**.

4.1 Études statistiques des caractéristiques des joueurs

Maintenant, présentons les variables que nous avons décidé de créer. Tout d'abord, nous allons présenter quelques graphes permettant d'expliquer le choix de nos variables. Les graphes présentés contiennent les caractéristiques de 5 joueurs que nous avons sélectionné pour leur profil assez différent les uns des autres. Il s'agit de Federer R, Ferrer D, Cilic M, Seppi A et Simon G. Si on devait quantifier le niveau de jeu de ces joueurs, Federer représente un top player, Ferrer et Cilic sont 2 très bons joueurs mais avec des styles de jeu complétement différents, Simon est un bon joueur et Seppi peut être considéré comme un joueur de seconde zone.

Le graphique ci-dessous représente le pourcentage de victoires sur différents courts (indoor et outdoor). Peu importe le court, Federer semble tout aussi bon en indoor qu'en outdoor. Ferrer quant à lui semble meilleur en outdoor qu'en indoor contrairement à Cilic qui semble légèrement meilleur en indoor. Et surtout on peut observer que le pourcentage de victoires de Cilic surpasse celui de Ferrer en indoor et qu'en outdoor c'est l'inverse. Cela est plutôt simple à expliquer. Les matchs en intérieur offrent des conditions de jeu davantage rapide qu'en extérieur. Cilic étant un joueur plus offensif et Ferrer un joueur défensif il semble logique de trouver ce résultat.

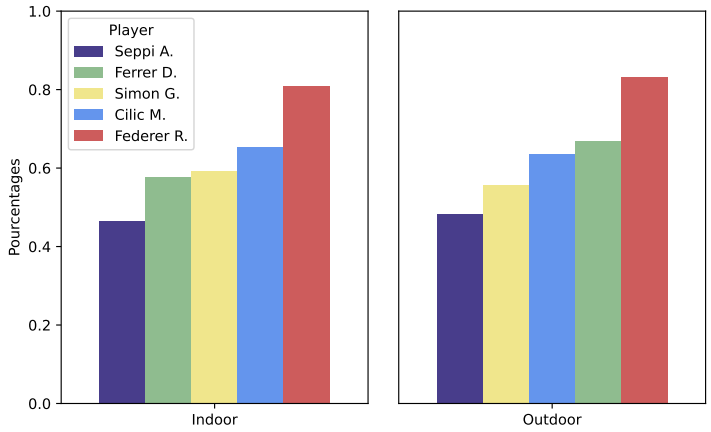

FIGURE 1 – Pourcentage de victoires sur différents courts

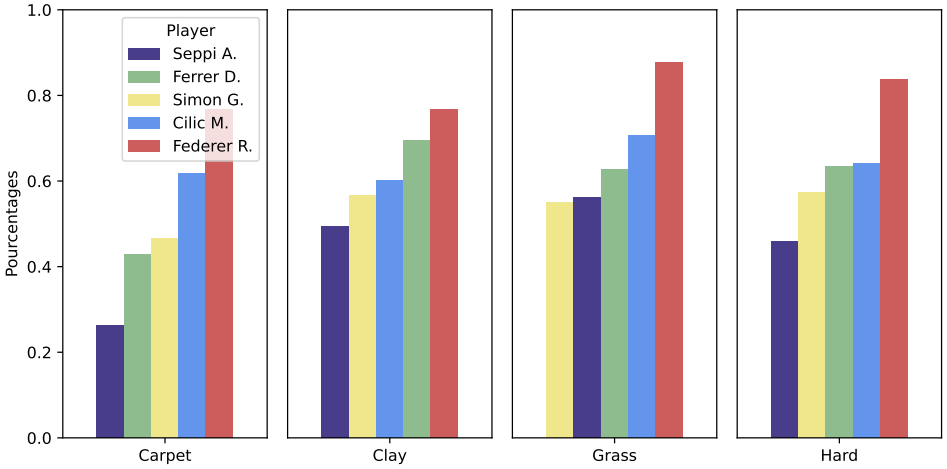

FIGURE 2 – Pourcentage de victoires sur différentes surfaces

Le graphique ci-dessus représente le pourcentage de victoires sur différentes surfaces (Carpet, Clay, Grass et Hard). Comme précédemment Federer domine largement au niveau des statistiques et on peut constater qu'il excelle sur gazon mais également sur dur. Malgré un très bon pourcentage de victoires sur terre battue, il reste tout de même moins bon que sur ces autres surfaces. Les conditions de jeu sur la terre battue sont plus lentes et favorise des styles de jeu plus défensif. C'est facilement observable pour Ferrer qui a de très bonnes stats sur terre battue, proche de celles de Federer. Encore une fois, on peut remarquer que Ferrer surpasse Cilic sur terre battue mais Cilic surpasse Ferrer sur gazon qui est une surface très rapide. Plus difficile de trancher sur le dur car il existe du dur rapide tout comme du dur lent. On remarquera également que de manière générale Simon et Seppi ont des stats moins bonnes et particulièrement Seppi. Cela confirme bien ce qui a été dit précédemment concernant le niveau de ces joueurs.

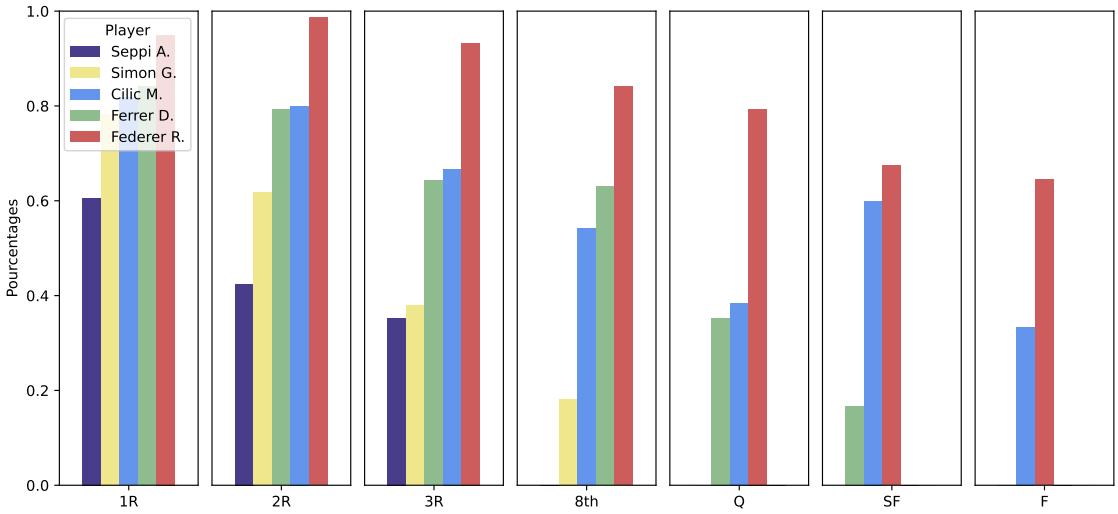

FIGURE 3 – Pourcentage de victoires à différents rounds d'un Grand Chelem

Le graphique ci-dessus représente le pourcentage de victoires aux différents tours d'un Grand Chelem. Pour rappel, les Grand Chelems constituent les tournois les plus importants du circuit. Ce qui est intéressant ici, c'est de voir à quel stade de ces tournois, les joueurs ont tendance à être éliminé. Si on regarde Seppi par exemple, il n'a jamais passé le 3ème tour. Simon n'a jamais passé les 8èmes de finale. Ferrer semble un peu plus sécuriser les 8èmes de final que Cilic mais à partir des 8èmes c'est Cilic qui semble sécuriser davantage ses matchs. À partir des quarts de finale, la plupart des adversaires sont des top players donc on pourrait penser que Cilic réussit mieux contre les top players que Ferrer.

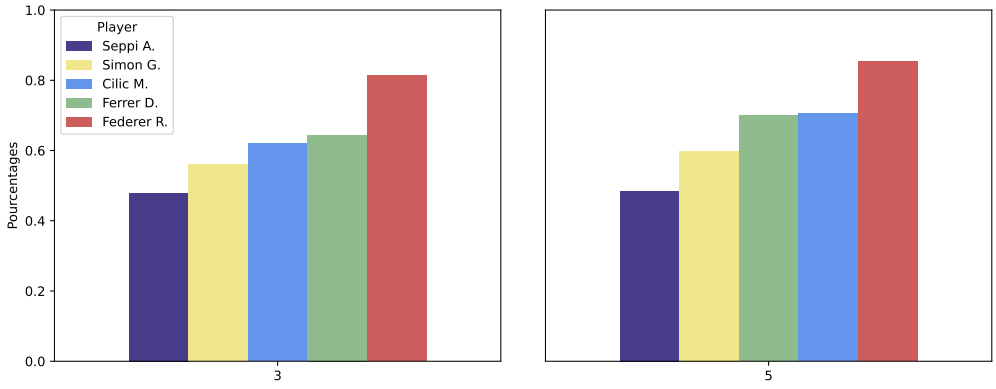

FIGURE 4 – Pourcentage de victoires dans différents formats

Le graphique ci-dessous représente le pourcentage de victoires sur différents formats de match, 3 ou 5 sets. Les résultats semblent indiquer que la plupart des joueurs sont meilleurs en 5 sets qu'en 3 sets. Ce résultat vient surtout du fait que les matchs en 5 sets sont joués en Grand Chelem, ce qui constitue seulement 4 tournois par an. Et tous les autres matchs se jouent en 3 sets donc le nombre de matchs en 3 sets est considérablement plus important. Ici ce qui est important c'est surtout de mettre en évidence la capacité qu'a un joueur à tenir sur la longueur dans un format 5 sets.

Le graphique ci-dessous représente le pourcentage de victoires dans différents Masters 1000. Si on s'attarde sur chacun des joueurs, on pourra observer que chaque joueur à ses tournois favoris. Bien qu'il ait de très bons résultats dans la plupart des tournois, on constate que Federer a tendance à performer dans les tournois de Cincinatti et Indian Wells. Cilic a tendance à performer à Cincinnati également qui est sur surface dur rapide. Ferrer performe à Madrid et Monte-Carlo qui sont des tournois sur terre battue. Simon performe à Shanghai et surtout ses statistiques sont meilleures que celle de Ferrer ou Cilic contrairement à tout ce qui a été vu précédemment. Ce tournoi semble donc vraiment lui réussir. On peut en conclure que jouer Simon à Shanghai et à Montréal n'est pas du tout comparable. Cette variable semble très intéressante car on voit énormément de variations dans les performances d'un joueur d'un tournoi à l'autre.

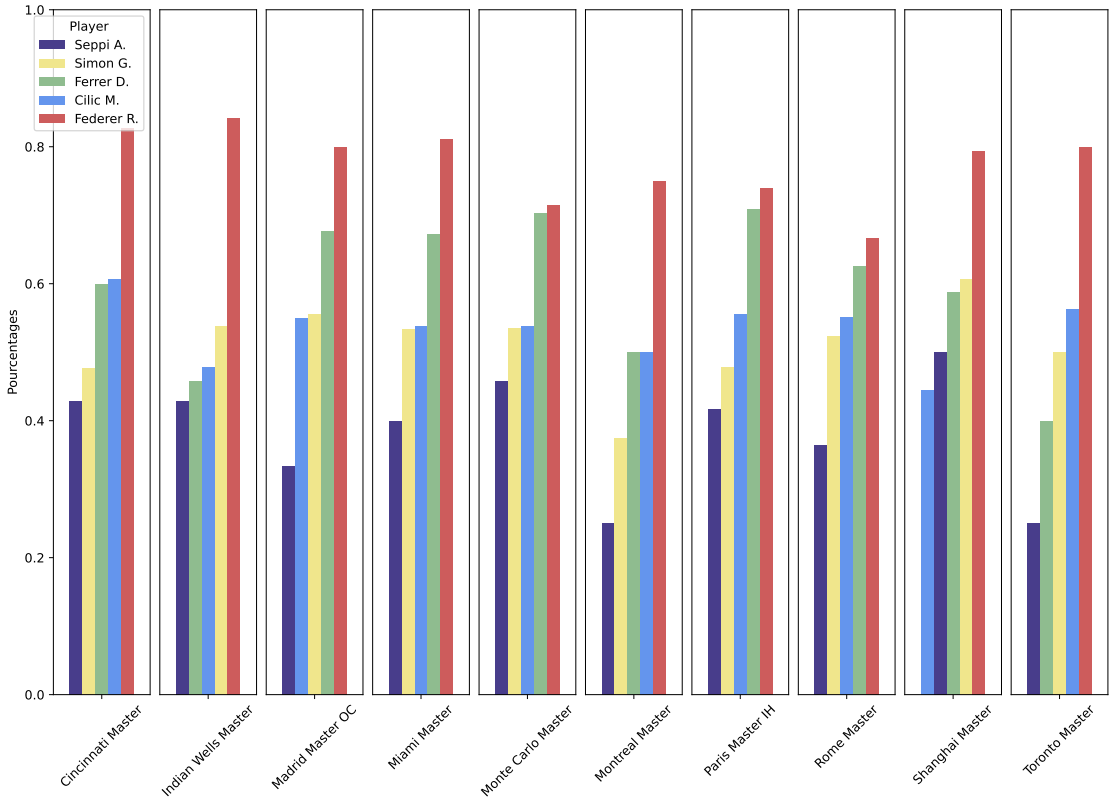

FIGURE 5 – Pourcentage de victoires dans différents Masters 1000

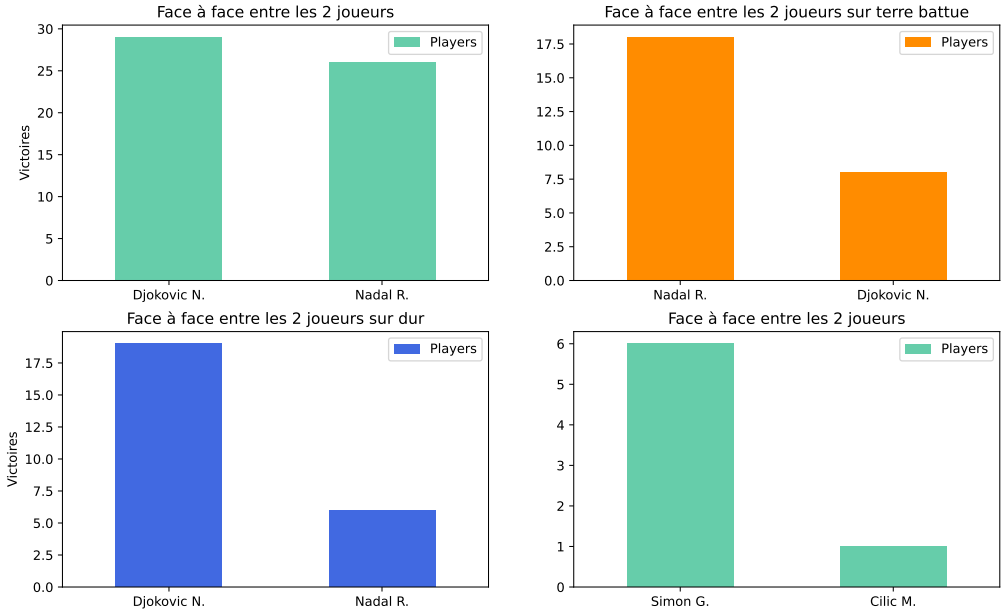

FIGURE 6 – Face à face entre différents joueurs

Le graphique ci-dessus représente les face à face entre les joueurs mais également les face à face sur différentes surfaces. Nous avons pris les 4 exemples suivants. Tout d'abord en haut à gauche on retrouve le face à face entre Djokovic et Nadal toutes surfaces confondues. On observe que Djokovic mène de quelques unités dans le face à face. En haut à droite est représenté les confrontations entre Nadal et Djokovic mais cette fois-ci sur terre battue. Ici on constate que c'est Nadal qui domine complétement sur cette surface. Maintenant en bas à gauche nous avons à nouveau leur face à face mais sur surface dur. Ici c'est Djokovic qui domine. Cet exemple est très intéressant car le face à face ne suffit pas pour montrer la supériorité d'un joueur face à un autre, il faut également prendre en compte la surface. Un dernier exemple est le face à face entre Simon et Cilic qui sont 2 joueurs dont on a déjà pu observer les statistiques précédemment. Comme on avait pu le voir, Cilic pouvait être catégorisé de très bon joueur et Simon de bon joueur. Cependant si on s'attarde sur leur face à face on peut observer une très nette domination de Simon. Simon peut alors être considéré comme la bête noire de Cilic. On pourrait alors penser que le style de jeu de Simon ne convient pas du tout à Cilic. Encore une fois, on voit tout l'intérêt de ce face à face. On ne peut pas comparer seulement les statistiques entre 2 joueurs, il est important d'avoir des informations sur leurs confrontations passées.

5 Choix des variables à implémenter

Maintenant nous allons lister toutes les variables que nous avons décidé d'implémenter.

Dans un premier temps, nous avons reproduit le système d'attribution de points du circuit ATP où le nombre de points dépend de la catégorie du tournoi et du tour auquel s'est arrêté le joueur. Par exemple, Simon a été battu en 8ème de finale d'un tournoi du GC et quitte le tournoi avec un total de 180 points. Le nombre de points n'est attribué que lorsque le joueur a perdu et donc quitte le tournoi. Grâce à ce système d'attribution des points, nous avons pu recréer le classement ATP où chaque joueur possède, avant chaque match, un certain nombre de points ATP qui dépend des résultats obtenus sur les derniers mois. Normalement le système de classement ATP

se fait seulement sur une seule année et sur toutes surfaces confondues mais nous avons décidé de l'expérimenter sous différentes formes.

Nous avons donc implémenté les variables suivantes :

- Classement glissant sur une année (toutes surfaces confondues)
- Classement glissant sur 8 mois (toutes surfaces confondues)
- Classement glissant sur 4 mois (toutes surfaces confondues)
- Classement glissant sur une année (en tenant compte des différentes surfaces)
- Classement glissant sur 8 mois (en tenant compte des différentes surfaces)
- Classement glissant sur 4 mois (en tenant compte des différentes surfaces)

Le système de classement est intéressant car il tient vraiment compte des performances obtenues dans un passé proche et surtout il permet de différencier les victoires dans une catégorie d'une autre car une victoire en demi-finale dans un tournoi ATP 250 et une victoire au même stade dans un tournoi du Grand Chelem n'a strictement rien à voir. Donc il ne prend pas en compte que les victoires, il prend surtout en compte le prestige des victoires. De manière générale, les victoires dites prestigieuses sont effectuées contre de très bons joueurs voir des top players et ce système de classement le retranscrit plutôt bien. Le fait de faire ce classement sur des périodes plus courtes permet d'évaluer encore mieux la forme actuelle car les résultats sur 4 mois semblent plus significatif que ceux sur un an mais cela ne marche pas toujours avec les changements de surface car 4 mois sur terre battue puis une reprise sur gazon n'a pas beaucoup de sens pour évaluer la forme d'un joueur. On a donc décidé d'incorporer également un classement par surface.

Ensuite nous avons implémenté toutes les variables qui concernent les face à face :

- Face à face (depuis toujours)
- Face à face (sur les 5 derniers matchs)
- Face à face (depuis toujours par surface)
- Face à face (sur les 5 derniers matchs par surface)

Comme on a pu le voir précédemment, la variable face à face semble être une variable pertinente. Le face à face sur les 5 derniers matchs a été ajouté car tout comme le classement sur 4 mois, il faut toujours avoir un regard sur les dernières confrontations. Supposons qu'on ait le face à face entre 2 joueurs qui ont 7 ans d'écart. Il y a de très fortes chances que le plus vieux des 2 batte le plus jeune dans leurs premières confrontations et que la tendance s'inverse par la suite. La face à face depuis leur toute première confrontation aura donc moins de sens ici. De même, il arrive parfois que certains joueurs parviennent à trouver la solution tactique face à des joueurs qui leur ont souvent posé problème par le passé.

Ensuite nous avons implémenté toutes les variables qui concernent le pourcentage de succès en prenant en compte différents facteurs :

- Pourcentage de victoires (3 dernière années)
- Pourcentage de victoires sur les différents formats 3 ou 5 sets (3 dernières années)
- Pourcentage de victoires sur les différentes surfaces (3 dernières années)

- Pourcentage de victoires sur les différents tours par catégorie de tournois (5 dernière années). Très important de différencier les tours par catégorie car comme précisé plus haut, un quart de finale en Grand Chelem et un quart de finale dans un ATP 250 n'a strictement rien à voir. À ce stade, on ne rencontre pas du tout les mêmes types de joueurs.
- Pourcentage de victoires sur les différents tournois (5 dernières années)

Encore une fois, nous avons pris sur les 3 ou 5 dernières années pour ne pas prendre en compte l'entièreté de la carrière du joueur car les résultats que peut avoir un joueur à ses débuts et en fin de carrière ne sont pas du tout comparables donc des statistiques glissantes sur les 3 ou 5 dernières années nous semblait plus pertinentes.

Ces dernières variables ont été reproduites mais sous une autre forme que le pourcentage de victoires. Au lieu d'établir un pourcentage, nous avons implémenté une variable qui soustrait la somme des défaites à la somme des victoires :

- Nombre de victoires - Nombre de défaites (3 dernière années)
- Nombre de victoires - Nombre de défaites sur les différents formats 3 ou 5 sets (3 dernières années)
- Nombre de victoires - Nombre de défaites sur les différentes surfaces (3 dernières années)
- Nombre de victoires - Nombre de défaites sur les différents tours par catégorie de tournois (5 dernière années)
- Nombre de victoires - Nombre de défaites sur les différents tournois (5 dernières années)

Parmi ces 2 méthodes, nous ne savons pas vraiment celle qui pourrait être la plus pertinente. Nous avons donc fait le choix de conserver les 2 et de voir par la suite celle qui pourrait s'avérait gagnante. Évidemment, on s'attend à ce que ces variables soient très corrélés les unes aux autres.

Et pour terminer, nous avons décidé d'implémenter des variables que l'on n'avait pas forcément étudié avant d'un point de vue statistique mais qui nous semblait être des paramètres à prendre en compte avant un match.

- Le pourcentage de victoires de matchs gagnés sur les 3 dernières semaines (représenté également sous la forme victoires-défaites)
- La série de victoires ou de défaites consécutives. Cette variable met en avant la forme et la confiance accumulée par un joueur sur les dernières semaines. Une importante série de victoires est associée à une grande forme et grande confiance. Tandis qu'une importante série de défaites est plutôt associée à une mauvaise forme et une faible confiance.
- Le nombre de matchs joués sur les 3 dernières semaines. Un nombre de matchs joués trop important peut nuire aux performances physiques d'un joueur.
- Le temps écoulé (en jours) qui sépare le dernier match et le match actuel. Un joueur qui revient de blessure et qui n'a pas joué depuis un moment voit forcément ses chances de remporter le match diminuer.

Maintenant que nous avons présentés les différentes variables et leurs implémentations dans le dataset de base, nous allons voir comment organiser le dataset de telle sorte que celui-ci soit interprétable par les modèles de prédiction.

6 Implémentation des variables

6.1 Initialisation des variables

L'implémentation de ces variables s'est déroulée en deux temps. Tout d'abord, nous avons dû créer un dataset répertoriant toutes les informations associées au match d'un joueur. Dans ce dataset, le nom du joueur représente l'index du dataset. Par exemple, pour l'index Federer R. lié à la date du 19/01/2009 il a joué un match de 1er tour (1R) à l'Open d'Australie et il a remporté le match. Cette information sur l'issue du match est contenue dans la variable « Résultats ». Elle comporte 2 modalités : V (victoire) ou D (défaite). Évidemment je n'ai précisé que certaines variables mais les variables « Location », « Series », « Round », « Court » et « Surface » sont toujours présentes. Ici nous perdons l'information sur le joueur battu ce jour-là par Federer. Nous souhaitons simplement connaître l'issue du match et les informations associées à ce match. Et c'est à partir de ce dataset que nous allons pouvoir construire la quasi-totalité des variables qui nous intéressent. Reprenons l'exemple précédent. Admettons que nous voulions ajouter une information supplémentaire à cette ligne comme par exemple le nombre de matchs remportés sur l'année écoulée. On procède de la manière suivante : on va chercher tous les matchs joués par Federer R. entre le 19/01/2008 et le 19/01/2009 et on compte toutes les lignes pour lesquelles l'issue du match a été une victoire. Nous ne rentrerons pas dans les détails de la construction d'autres variables mais de manière générale toutes les autres variables ont été créées sur le même principe excepté quelques variables implémentées directement à partir du dataset de base.

C'est donc à partir de ce dataset que nous avons pu implémenter toutes nos variables dans le dataset de base. On rappelle que dans le dataset de base, chaque ligne correspond à un match opposant 2 joueurs. Nous avons le « Winner » puis le « Loser ». Dans l'autre dataset, nous retrouvons cette information là mais sous 2 lignes distinctes avec l'un des 2 joueurs qui a un V dans la colonne Résultats et l'autre un D. Prenons par exemple dans le dataset de base, le match opposant Federer R. à Djokovic N. qui a eu lieu le 10/09/2011 en SF de l'US Open. Nous souhaitons avoir toutes les informations sur Federer R. et sur Djokovic N. de leurs matchs passés. Toutes ces informations-là sont contenues dans l'autre dataset créé au préalable. Pour Federer R., nous venons chercher à l'index Federer R. et à la date 10/09/2011 toutes les informations supplémentaires dont nous avons besoin et nous faisons de même pour Djokovic N. Nous nous retrouvons alors avec toutes les informations appartenant au passé sur Federer R. et sur Djokovic N.

Le code qui a permis de fabriquer ce construire ce dataset se trouve dans le fichier **bkm_plot_and_create_variables.ipynb**. La boucle qui permet d'initialiser toutes ces variables est comprise dans une cellule qui ne doit pas être compilé sous risque d'avoir une exécution très longue. La dernière cellule de ce notebook prend en entrée le fichier **atp_data_all_players_var.csv** qui contient tous les matchs de chacun des joueurs avec toutes les variables crées qui lui sont associées.

6.2 Préparation des variables pour les modèles de prédiction

6.2.1 Création des variables « player_A », « player_B », et « target »

Dans un premier temps, nous devons abstraire les notions de « Winner » et de « Loser » au travers d'une variable cible « target » binaire.

Pour ce faire, nous avons créé les deux variables « player_A » et « player_B » et les avons renseignées de manière aléatoire en choisissant entre le « Winner » et le « Loser ». Nous avons toutefois figé la graine aléatoire numpy afin d'être en mesure de reproduire les mêmes valeurs pour ces variables « player_A » et « player_B » à chaque éventuelle régénération du dataset.

Enfin, nous avons créé la variable cible « target » de notre dataset en utilisant la règle suivante : si « player_A » est le vainqueur du match, « target » prend la valeur 1, sinon « target » vaut 0.

6.2.2 Création des variables de type « head-to-head »

La principale difficulté pour créer ces variables réside dans l'identification d'une rencontre unique, quel que soit la position des joueurs A et B. Par exemple, un match où les joueurs A et B sont respectivement « Federer R. » et « Nadal R. » doit être comptabilisé dans le même « face-à-face » qu'un match où les joueurs A et B sont respectivement « Nadal R. » et « Federer R. ».

Ainsi, afin d'obtenir les variables souhaitées, c'est-à-dire les « h2h » « au global », « au global mais seulement sur la surface du match observé », « au global mais seulement lors des 5 dernières confrontations », et « au global mais seulement sur la surface du match observé lors des 5 dernières confrontations », nous avons dû passer par la création des variables intermédiaires définies ci-après :

- Identifiant unique pour chaque h2h qui ne change jamais pour un h2h donné, quels que soient les joueurs A et B, avec tri par ordre alphabétique de type « Federer R. - Nadal R. ».

- Variables « h2h_player_1 » et « h2h_player_2 » qui ne changent jamais pour un h2h donné, quels que soient les joueurs A et B. Par exemple, dans les confrontations entre Federer et Nadal dont l'ID est « Federer R. - Nadal R. », la variable « h2h_player_1 » fera toujours référence à Federer, c'est-à-dire au premier par ordre alphabétique.

- Variables « h2h_player_1_win » et « h2h_player_2_win » qui valent 1 ou 0 selon qui remporte le match.

- Variables « h2h_player_1_score » et « h2h_player_2_score » qui comptent le score de chaque joueur dans le h2h en sommant les variables « h2h_player_1_win » et « h2h_player_2_win » des matches précédemment joué avec l'ID du h2h du match.

- Variables « h2h_match_number » qui s'incrémente à chaque nouveau match dans le head-to-head (toutes surfaces confondues).

- Variables « h2h_surface_match_number » qui s'incrémente à chaque nouveau match dans le head-to-head sur la surface concernée par le match à prédire.

- Variables « h2h_score_last_5_player_1 » et « h2h_score_last_5_player_2 » qui comptent le score de chaque joueur dans le head-to-head lors des 5 derniers matches grâce à « h2h_match_number ».

- Variables « h2h_score_last_5_surface_player_1 » et « h2h_score_last_5_surface_player_2 » qui comptent le score de chaque joueur dans le head-to-head lors des 5 derniers matches sur la surface concernée grâce à « h2h_surface_match_number ».

Une fois créées, ces variables intermédiaires nous ont ainsi permis de générer facilement les variables de scores dans les h2h définies au point 5 pour chaque joueur, au global, par surface et sur les 5 dernières confrontations.

6.2.3 Création des variables relatives aux « points du classement ATP »

A l'aide de notre dataset décrit au point 6.1 qui répertorie toutes les informations associées au match d'un joueur, nous avons pu faire remonter les points ATP gagnés lors de chaque match par chaque joueur.

Ainsi, en sommant ces points selon les critères listés au point 5 (dernière année, derniers 8 mois ou derniers 4 mois, avec filtre spécifique sur la surface concernée ou non), nous avons pu créer pour chaque match, pour chacun des 2 joueurs, les 6 variables de comptabilisation de ses points ATP en fonction de la durée et de la surface observée.

6.2.4 Création des variables relatives aux « ratios de victoires », aux nombres de « victoires moins défaites », aux nombres de « matches joués sur les 3 dernières semaines », ainsi qu'aux nombres de « jours écoulés depuis le dernier match joué »

Toujours en utilisant le dataset décrit au point 6.1, nous sommes ensuite venus enrichir notre dataset principal en lui intégrant les variables suivantes relatives aux :

- « ratios de victoires » (variables de type « ratioV »).

- « nombres de victoires moins défaites » (variables de type « V-D »).

- « nombres de matches joués sur les 3 dernières semaines » (variables de type « matchs_played_3w »).

- « nombres de jours écoulés depuis le dernier match joué » (variables de type « days_since_last_match_played »).

Une fois ces dernières variables intégrées, nous disposons alors de toutes les variables que nous avons choisi d'implémenter au point 5 pour chaque joueur (A et B) et pour chaque match.

6.2.5 Création du « dataset final » (« df.csv ») utilisable par les modèles de prédiction

La dernière étape avant de pouvoir lancer nos modèles de prédiction sur notre dataset consiste, d'une part, à sélectionner uniquement les variables que nous avons retenues au point 5, et d'autres part, à transformer chaque « paire de variables » en une variable unique représentant au mieux l'écart entre ces deux valeurs.

En effet, ne connaissant pas de solution pour indiquer à nos algorithmes de comparer une à une chaque variable composant une paire de variables. Par exemple, comparer les points ATP sur 1 an du joueur A avec ceux du joueur B, puis le ratio de victoires et défaites du joueur A avec celui du joueur B... Nous avons pris le parti de fusionner chaque paire de variables en une seule variables égale à la différence de la valeur de A moins celle de B lorsqu'il s'agit d'une valeur entières, et au quotient de A par B lorsqu'il s'agit d'un ratio.

C'est donc de cette manière que nous avons obtenu notre dataset final (« **df.csv** ») contenant notre variable cible et nos 25 variables explicatives représentant chacune le « delta » entre les statistiques du joueur A et celles du joueur B pour chaque paire de variables listée au point 5.

#	Column	Non-Null Count	Dtype
0	target	58664 non-null	int64
1	1_year_pts_delta	58664 non-null	float64
2	8_months_pts_delta	58664 non-null	float64
3	4_months_pts_delta	58664 non-null	float64
4	1_year_pts_delta_surface	58664 non-null	float64
5	8_months_pts_delta_surface	58664 non-null	float64
6	4_months_pts_delta_surface	58664 non-null	float64
7	h2h_player_score_delta	58664 non-null	int64
8	h2h_player_score_last_5_delta	58664 non-null	int64
9	h2h_player_score_surface_delta	58664 non-null	int64
10	h2h_player_score_last_5_surface_delta	58664 non-null	int64
11	days_since_last_match_played_delta	58664 non-null	int64
12	V-D_rang_delta	58664 non-null	int64
13	matchs_played_3w_delta	58664 non-null	int64
14	V-D_3w_delta	58664 non-null	int64
15	ratioV_3w_delta	58664 non-null	float64
16	V-D_3y_delta	58664 non-null	int64
17	ratioV_3y_delta	58664 non-null	float64
18	V-D_3y_format_delta	58664 non-null	int64
19	ratioV_3y_format_delta	58664 non-null	float64
20	V-D_3y_surface_delta	58664 non-null	int64
21	ratioV_3y_surface_delta	58664 non-null	float64
22	V-D_Series_and_Round_5y_delta	58664 non-null	int64
23	ratioV_Series_and_Round_5y_delta	58664 non-null	float64
24	V-D_tournament_5y_delta	58664 non-null	int64
25	ratioV_tournament_5y_delta	58664 non-null	float64

FIGURE 7 – L'intégralité des variables implémentées

7 Étude statistique des variables

Les graphiques qui vont suivre sont produits à partir du code qui se trouve dans le fichier **bkm_stat_study_ML .ipynb** et qui prend en entrée le dataset **df.csv**.

Maintenant que les variables sont interprétables par les modèles de prédiction, nous allons pouvoir commencer à chercher le modèle de prédiction qui nous donnera les meilleurs résultats mais avant il est nécessaire de faire une brève étude statistique de ces variables. Nous aimerions savoir quelles sont les variables qui semblent avoir le plus d'impact sur l'issue du match.

7.1 Map de corrélation entre les variables et la target

Pour ce faire, nous avons essayé de prédire l'issue de chacun des matchs en ne considérant à chaque fois qu'une seule de ces variables. Par exemple, si on prend la variable « 1_year_pts_delta ». Si la valeur est positive on attribue un 1 car le player A a remporté le match sinon 0 et on le compare aux valeurs de la target. De la même manière pour les rations, si le rapport est supérieur à 1 on attribue un 1 sinon un 0. Par moment, les valeurs peuvent avoir un delta de 0 ou un rapport de 1 et dans ce cas on attribue une valeur NaN car on considère que l'on ne peut pas se prononcer à partir de cette unique variable.

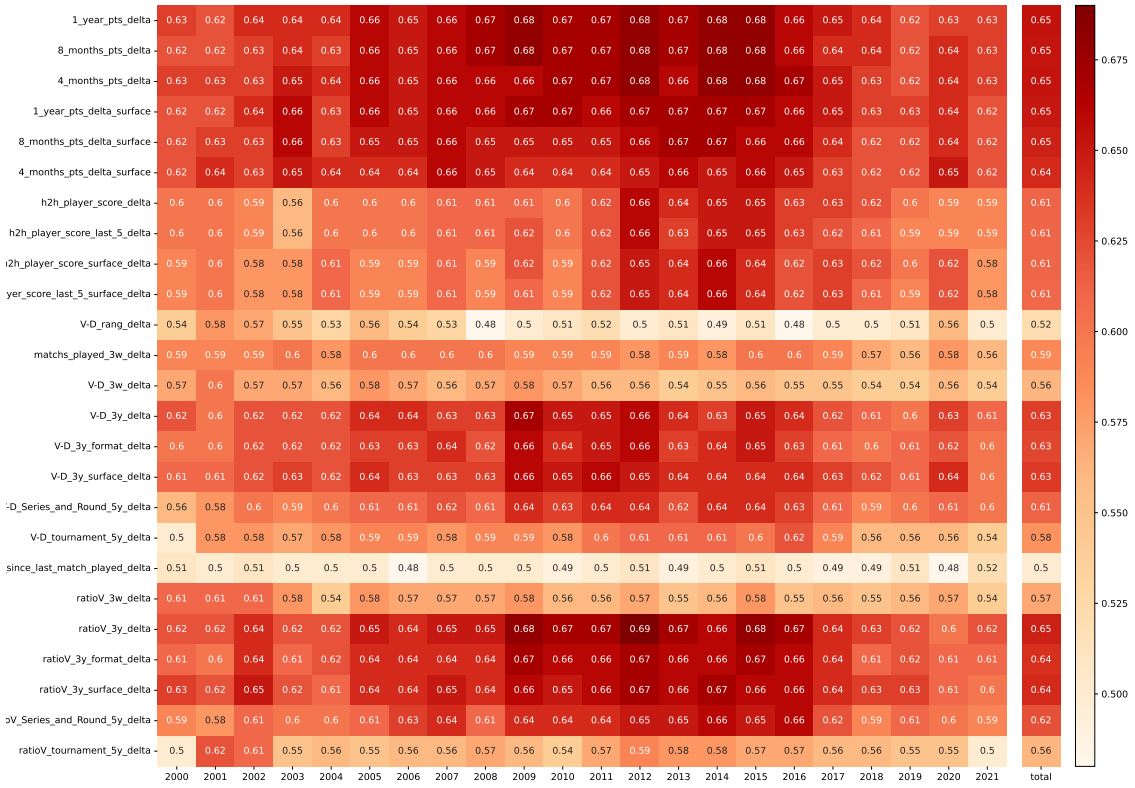

FIGURE 8 – Map de corrélation entre chacune des variables et la target en fonction des années

Ces résultats ont été retranscrit dans, ce qui peut s'apparenter à une map de corrélation mais qui n'en est pas vraiment une au sens strict puisque nous n'étudions pas la corrélation entre les variables elles-mêmes mais les corrélations entre les variables séparément et la target.

Sur l'axe des ordonnées est représenté toutes les variables que nous avons implémenté. L'axe des abscisses lui est représenté par les différentes années qui vont donc de 2000 à 2021 ainsi que la colonne totale dont les valeurs représentent la moyenne des valeurs obtenues pour chacune des années. La valeur 0.63 contenue dans le croisement entre la ligne « 1_year_pts_delta » et la colonne 2000 se traduit de la manière suivante. Si dans le dataset on venait à ne considérer que cette variable-là et que l'on souhaite faire des prédictions seulement à partir de cette variable alors pour l'année 2000 on aurait réussi à prédire 63% des matchs. Le raisonnement est le même pour toutes les autres variables.

Par exemple, pour la variable « 1_year_pts_delta » en 2000, nous sommes parvenus à partir de cette unique variable à prédire 63% des matchs. L'intérêt de regarder par année est que l'on peut observer des périodes pour lesquelles les prédictions sont meilleures.

Si on s'attarde sur la colonne totale, les variables les plus efficaces sont les variables qui prennent en compte le classement. Leur taux d'efficacité très similaire semble indiquer une forte corrélation entre elles, ce qui n'est pas très étonnant puisqu'elles ont été implémentées à partir du même classement et sur des périodes temporelles relativement proches. Leurs bons scores prouvent bien l'efficacité de ce classement ATP.

Ensuite en terme de bon score, viennent légèrement derrière les variables prenant en compte toutes les variables associées aux résultats passés, sur le format, la surface et les tours par catégories. Mais comme précédemment elles semblent corrélées entre elles au vu des scores très proches.

Juste derrière viennent se loger les face à face et encore une fois, elles semblent très corrélées entre elles, ce qui est également plutôt attendu. Nous pensons que toutes les études statistiques faites précédemment où l'on peut observer certaines différences restent des cas suffisamment rares pour vraiment faire une grosse différence. Il ne faut pas oublier que l'on a sélectionné des joueurs de niveau différent. Lorsque des joueurs de catégorie différentes s'affrontent, le résultat peut être anticipé mais lorsque des joueurs de même catégorie s'affrontent, le résultat du match devient plus indécis. Sur le top 100, nous pouvons considérer que nous avons 60 joueurs de seconde zone, 25 bons joueurs, 10 très bons joueurs et 5 top players. Le nombre de match entre joueurs de seconde zone vient fortement impacter les résultats obtenus.

La variable « match_played_3w_delta » qui est le nombre de matchs joués sur les dernières semaines semble tout de même intéressante et a très peu de chance d'être corrélée aux autres. En revanche, les 3 dernières : « V-D_rang_delta », « days_since_last_match_played_delta » et « ratioV_3w_delta » ne sont pas du tout intéressantes pour la suite, ce qui peut s'expliquer par le fait qu'un nombre de victoires et défaites de rang reste relativement faible tout comme le temps écoulé entre le match actuel et le dernier match joué. Celui-ci peut faire la différence lorsqu'un joueur se blesse mais cela reste des cas tout de même assez rares sur un aussi grand nombre de matchs.

7.2 Valeurs non-prédictibles par année

Nous aimerions expliquer la faiblesse de certaines variables sur différentes périodes car il pourrait être intéressant de voir si certaines périodes ne sont pas biaisées par des évènements extérieurs. Pour cela, nous avons tracé une nouvelle map qui cette fois ne s'apparente pas du tout à une map de corrélation mais qui calcule le nombre de matchs par variable et par année sur lesquels nous ne pouvons pas nous prononcer.

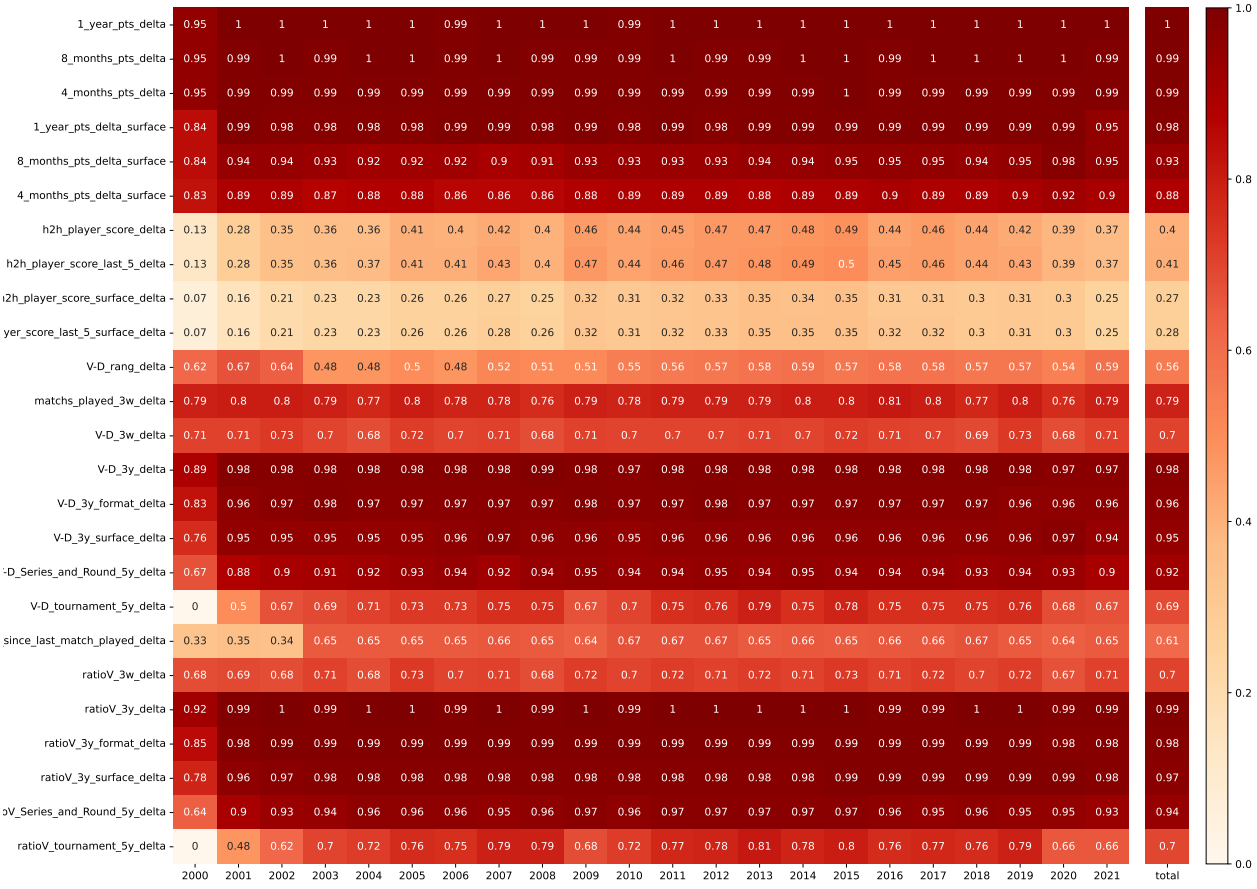

FIGURE 9 – Taux de valeurs non-prédictibles par variable en fonction des années

Si on prend par exemple les premières variables concernant le classement, on voit que dans quasiment 100% des cas nous pouvons nous prononcer, c'est-à-dire que nous avons un delta entre la variable du joueur A et le joueur B toujours soit positif soit négatif. Pour que cela n'arrive pas, il faudrait que les 2 joueurs aient le même nombre de points ce qui est hautement improbable. Si on regarde les variables liées au pourcentage ou nombre de victoires, on est également très haut mais un peu plus faible pour celles calculées avec le ratio entre les 2. On rappelle que ces variables sont calculées de manière variable_joueur_A/variable_joueur_B ce qui impose que si la variable variable_joueur_B est nulle, alors on ne peut pas calculer ce ratio.

On observe donc que les variables où l'on calcule la différence de victoires/défaites permet de prédire plus de matchs. Ces variables semblent donc plus intéressantes. Le nombre de prédictions pour les face à face est relativement faible, ce qui est lié au fait que de nombreux joueurs ne sont jamais rencontrés et possèdent donc un delta de 0. Les joueurs qui ont tendance à se jouer régulièrement entre eux sont les bons, très bons et top player qui parviennent régulièrement à atteindre au moins les 8èmes de finale et à se retrouver entre eux. Les joueurs de seconde zone sont beaucoup plus nombreux et ont tendance à être éliminé assez rapidement lors des tournois

donc moins de probabilité de jouer contre plusieurs joueurs différents. Malgré l'efficacité de cette variable (autour de 61%), celle-ci ne permet pas souvent de faire la différence. Elle reste tout de même une variable intéressante, et non négligeable puisqu'elle obtient tout de même un bon score. Les autres variables que nous avions considérées comme peu efficace semblent également peu prédictibles hormis « matchs_playerd_3w_delta », ce qui explique les très faibles scores obtenus.

À présent, observons la colonne de l'année 2000. On constate que les valeurs sont extrêmement faibles et cela peut s'expliquer très simplement par le fait que toutes les variables ont été initialisées lors de l'année 2000. Nous n'avions aucune information sur les années antérieures donc davantage de variables non prédictibles et cela est flagrant sur les face à face. On démarre à 13% en 2000 pour atteindre 49% en 2015. Il est assez visible de voir que la quantité de certaines variables prédictibles augmentent au fur et à mesure des années pour atteindre de très bons résultats à partir des années 2008-2009. Et on peut remarquer que lors des dernières années les résultats baissent sensiblement. Nous expliquons cette augmentation de valeurs prédictibles pour certaines variables et de légères baissent par le fait qu'à partir des premières années nous avions beaucoup de joueurs déjà en activité mais sur lesquels nous avions peu de statistiques puis sont arrivés pleins de nouveaux joueurs qui ont commencé leur carrière dans ces années-là. Et une grande partie de la statistique à partir de 2006 dépend de tous ces nouveaux joueurs apparus dans les premières années. Et les statistiques n'ont cessé de s'enrichir au fur et à mesure. Tout cela s'observe très bien à partir des variables de face à face qui ne cessent d'augmenter jusqu'à atteindre un pic dans les années 2014-2015 pour diminuer légèrement à nouveau. Et à son tour, cette diminution peut s'expliquer par le fait que tous les joueurs ayant débuté leurs carrières dans les premières années, la termine en moyenne dans les années 2016-2017. On peut considérer que la carrière d'un joueur de tennis peut durer de 13 à 16 ans environ. Donc des joueurs enrichis de statistiques n'apparaissent plus dans le dataset et laisse place à des joueurs avec un peu moins de statistiques. La variable sur les face à face est très intéressante pour quantifier cela car elle y est directement corrélée.

Au vu des ces observations, il peut être intéressant de sélectionner une période plus riche en statistiques afin d'avoir des résultats plus fiable.

8 Modèles de prédiction

Dans cette partie, nous allons examiner deux approches différentes pour répondre à notre problématique, celle d'obtenir de meilleurs prédictions que les bookmakers.

La première approche consiste à prédire n'importe quel match en s'appuyant seulement sur les informations que nous pouvons extraire d'un match sans avoir connaissance des matchs passés des 2 joueurs. En anonymisant les joueurs, on s'intéresse alors seulement aux statistiques du tennis. L'intérêt ici est de voir quels sont les paramètres les plus intéressant qui entrent en jeu lors d'un match et voir selon les résultats obtenus, quels sont par exemple, la surface, les tours, le type de tournoi ou encore les tournois sur lesquels il est le plus intéressant de parier.

En revanche dans la seconde approche, il ne sera plus question d'anonymiser les joueurs car nous ferons les prédictions pour chacun des matchs directement à partir des matchs passés pour les 2 joueurs tout comme le font les bookmakers et nous essaierons de nous mettre à la place d'un parieur afin de voir si notre modèle nous permet d'être gagnant sur du long terme.

8.1 Approche statistiques sur le taux de prédictibilité dans le tennis

8.2 Choix du modèle

Afin de réaliser cette première partie, nous avons utilisé le code se trouvant dans le fichier **bkm_stat_study_ML.ipynb** et prenant en entrée le dataset **df.csv** contenant toutes les variables numériques avec le delta des variables entre les 2 joueurs ainsi que la target contenant un 1 si le player A gagne et un 0 si c'est le player B qui gagne. Il est évident qu'ici nous sommes dans un cas où le modèle le plus adapté est un modèle prédictif de

classification : classe 1 ou classe 0.

Comme il a été vu dans la section précédente, la période temporelle la plus intéressante est celle entre 2005 et fin 2019. Nous avons donc décidé de découper le dataset entre début 2005 et une partie de l'année 2020 (3 mois seulement) que nous avons coupé juste avant la période de la pandémie du covid. Cette période risque d'avoir fortement perturbé la suite de la saison donc il n'y a aucun intérêt à la conserver. En ce qui concerne les données de train et de test nous avons sélectionné 80% des données pour le train et 20% pour le test. Ce découpage a été effectué de sorte à conserver l'ordre chronologique.

Le travail de preprocessing a effectué ici ne repose que sur la méthode de normalisation puisque toutes les variables sont déjà numériques. Étant donné l'écart d'ordre de grandeur entre certaines variables, nous avons jugé nécessaire de les normaliser. Il est peu probable que cette distribution suive une loi normale donc le meilleur choix semble être la méthode MinMax.

Par la suite, nous devons choisir le modèle le plus adapté à notre jeu de données et ses paramètres. Pour cela, différents modèles de classification s'offrent à nous :

- KNeighborsClassifier
- Decision Tree Classifier
- Random Forest
- SVM

Les algorithmes de Boosting et de Bagging aurait pu être utilisé tout comme ceux de Voting Classifier mais pour des raisons de performance en terme de rapidité de calcul, nous avons préféré rester sur des algorithmes moins complexe.

Pour chacun des modèles, le déroulement a été le suivant. Tout d'abord nous avons effectué un GridSearchCV sur chacun des modèles en testant plusieurs paramètres. Et à partir des résultats obtenus nous avons utilisé la méthode KBest afin de sélectionner la combinaison de variables qui offrait le meilleur score.

Les résultats obtenus pour chacun des modèles sont les suivants :

```
Meilleurs paramètres : {'metric': 'manhattan', 'n_neighbors': 19}
Le meilleur score obtenue est : 0.6418366100867017 pour 5 features
```

FIGURE 10 – Best paramètres et variables du KNeighborsClassifier

```
Meilleurs paramètres : {'criterion': 'gini', 'max_depth': 5}
Le meilleur score obtenue est : 0.6376846989864452 pour 25 features
```

FIGURE 11 – Best paramètres et variables du Decision Tree Classifier

```
Meilleurs paramètres : {'max_features': 'log2', 'min_samples_split': 38}
Le meilleur score obtenue est : 0.6483087068018073 pour 22 features
```

FIGURE 12 – Best paramètres et variables du Random Forest

Parmi les 4 modèles de prédiction, le Random Forest est le plus performant. Les scores des différents modèles sont extrêmement proches d'un modèle à l'autre et en particulier le Random Forest et le SVM. Mais on optera

```
Meilleurs paramètres : {'C': 1, 'gamma': 0.5, 'kernel': 'rbf'}
Le meilleur score obtenue est : 0.6480644767370863 pour 13 features
```

FIGURE 13 – Best paramètres et variables du SVM

pour le modèle Random Forest pour prédire nos résultats même si nous pourrions tout aussi bien choisir le SVM. Ce résultat peut alors être comparé à celui obtenu par les bookmakers.

8.3 Comparaison de notre modèle avec celui des bookmakers

Sur le même échantillon de test, le principal bookmaker, à savoir Bet365, obtient 67% de bonnes prédictions contre 65% pour notre meilleur modèle. L'échantillon de test s'étale entre début 2017 et début 2020 et on peut observer pour les variables les plus corrélées qui sont les variables propres au classement ATP, que nous avons sur cette période une moyenne de 0.635. Également sur cette période, nous avons que l'on obtenais un score de 0.648. On voit donc que l'ajout de toutes les autres variables performent très légèrement le score de notre algorithme.

Avec un algorithme plus puissant, il aurait peut-être était possible de se rapprocher un peu plus des résultats obtenus par les bookmakers.

8.4 Quelles sont les paramètres à prendre en compte pour maximiser ses chances de gagner ?

Maintenant, parmi les différentes surfaces, les différentes catégories de tournois ou encore les tournois, il serait intéressant de voir ceux sur lesquels le taux de prédiction est le plus important mais également le plus faible.

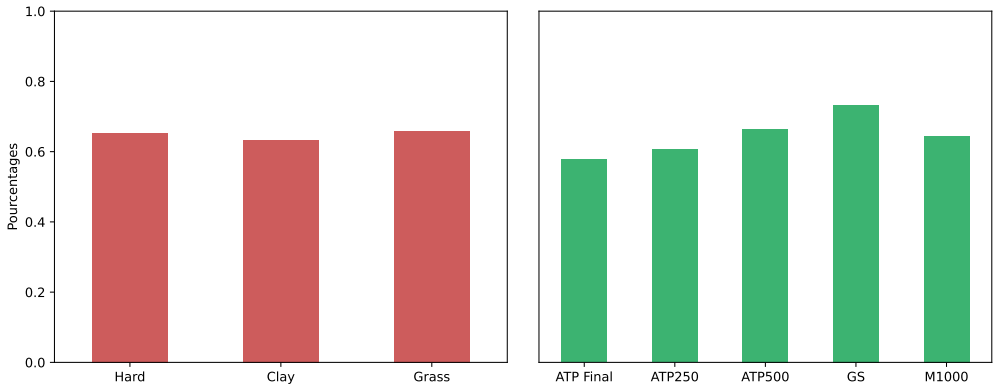

FIGURE 14 – Pourcentage de bonnes prédictions effectué par le modèle par surface (à gauche) et par catégorie de tournois (à droite)

Ci-dessus, on peut observer à gauche un taux de réussite assez similaire entre les différentes surfaces. À droite en revanche, nous pouvons observer que le taux de réussite est très important, entre 75 et 80% pour les Grands Chelems. Celui des ATP 500 et Master 1000 est également important contrairement à celui d'un ATP 250, ce qui s'explique très facilement. La grande majorité des joueurs qui jouent dans cette catégorie sont des joueurs de seconde zone avec des statistiques très proches les unes des autres. Des écarts de points ATP très faibles et donc très proches ainsi que des face à face inexistant. Le résultat obtenu pour l'ATP Final est également très attendu

car c'est un tournoi réunissant les 8 meilleurs joueurs au monde donc ce sont des joueurs avec des statistiques très similaires aussi. En revanche, ce tournoi n'a lieu qu'une fois par an alors que le circuit est principalement composé de tournois ATP 250 donc on comprend bien que ce sont principalement ces matchs qui tirent le score de notre modèle vers le bas. On peut conclure que les tournois pour lesquels la mise est la plus sûre sont les tournois du Grands Chelems.

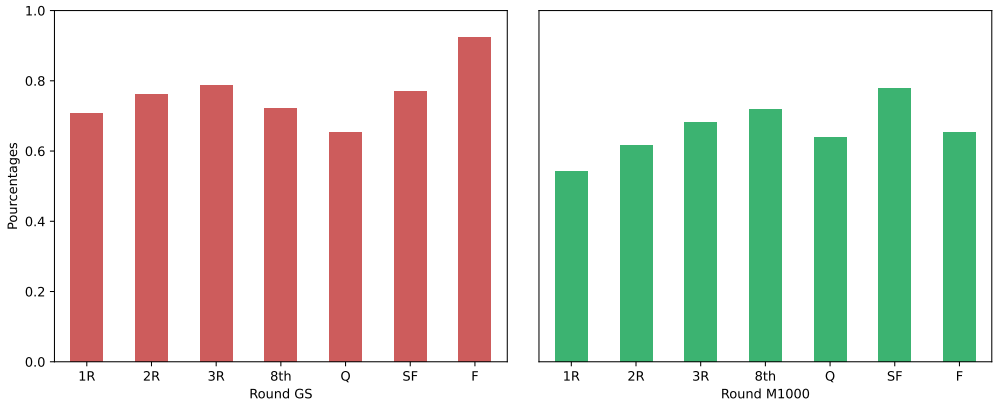

FIGURE 15 – Pourcentage de bonnes prédictions effectué par le modèle par round en Grand Chelem (à gauche) et en Master 1000 (à droite)

Ici, à gauche est représenté le taux de réussite par tour en Grand Chelem et à droite celui pour les Masters 1000. On observe assez rapidement que les distributions ne sont pas les mêmes. À gauche, il semblerait que la mise la plus sûre serait surtout pour la finale (90-95%) mais également pour le 3ème tour et la demi-finale (75-80%). Cependant, il paraît beaucoup plus risqué de miser sur un match qui a lieu en quart. Ce qui est intéressant ici, c'est que les taux de réussite pour les Masters 1000 ne sont pas les mêmes que pour les Grands Chelems. Cette fois, les tours les plus sûres sont les demi-finales et les 8èmes de finales. On constate donc que d'une catégorie de tournoi à une autre, nous avons des résultats très variables et que les mises ne doivent pas être distribuées de la même manière selon les tournois si on souhaite maximiser les gains.

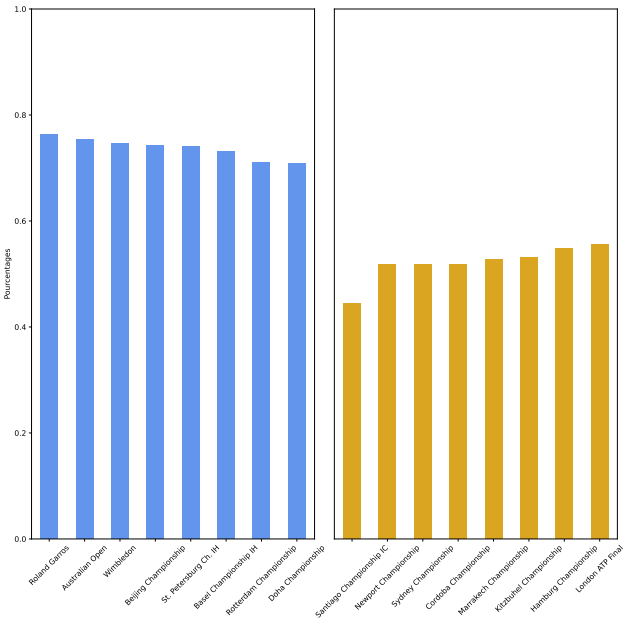

FIGURE 16 – Les meilleurs (à gauche) et les plus mauvais (à droite) pourcentage de bonnes prédictions effectué par le modèle par tournois

Et pour finir, nous avons également étudié le cas pour les différents tournois afin de voir si certains tournois se démarquaient, par leur taux de réussite, plus que d'autres. À gauche est représenté les taux obtenus pour les 8 tournois avec les taux les plus élevés et à gauche ceux avec les taux les plus faibles. On remarquera que 3 Grands Chelems apparaissent dans ce top 8, ce qui est en accord avec ce qui a été observé plus haut (le taux de victoires en GC était le plus élevé). Pour les 8 tournois, on se trouve avec entre 70 et 80% de taux de bonnes prédictions, ce qui est un très bon résultat. En revanche, à droite sont concernés les tournois sur lesquels il ne vaut mieux pas miser. On se retrouve entre 50 et 60%, ce qui revient presque à considérer la mise comme un simple jeu de hasard. Évidemment parmi ces tournois, on retrouve principalement des tournois ATP 250, ce qui fait une fois de plus écho à ce qui a été dit précédemment.

8.5 Taux de prédictibilité sur l'issue d'un match

8.5.1 Peut-on faire confiance au bookmaker ?

Dans cette seconde partie, nous avons utilisé le code se trouvant dans le fichier **bkm_ML2.ipynb** et prenant en entrée le dataset **atp_data.csv**. Ce fichier est le fichier de départ nettoyé et corrigé mais dans lequel se trouve les variables d'origine ainsi que les côtes données par différents bookmakers.

En se concentrant sur la colonne « B365W » on trouve que le bookmaker a prédit :

- correctement le joueur gagnant (22 306 fois)
- incorrectement le joueur gagnant (9212 fois)
- aucune décision prise, c'est-à-dire que les côtes des 2 joueurs étaient les mêmes (673 fois)

Si on ne tient pas compte des cas où les 2 joueurs ont la même côte et en considérant les côtes de chacun des match. En misant 1 euros sur chacun des matchs on obtient par le calcul suivant :

sommes des côtes gagnantes - (nombre des côtes gagnantes + nombre de côtes perdues)

est de -1717.73 euros sur un total de 31518 matchs.

Nous sommes donc face à plusieurs possibilités :

- Le bookmaker a fait une erreur en prédisant le gagnant
- Le bookmaker a prédit le gagnant correctement mais avec une marge bénéficiaire en sa faveur
- Le bookmaker a prédit le gagnant correctement mais il a remplacé la côte du gagnant par la côte du perdant afin de tromper le parieur

Dans la suite nous allons donc créer un algorithme qui prédit le joueur gagnant, puis on comparera le résultat de notre algorithme avec celui du bookmaker B365.

8.5.2 Prédire l'issue d'un match

Les variables qui seront utilisées dans notre modèle de prédiction sont : « Tournament », « Series », « Court », « Surface », « Winner », « Loser », « WRank », « LRank », « elo_winner », « elo_loser », « proba_elo ».

La méthode de prédiction du joueur gagnant passe par plusieurs étapes :

- Étape 1 : On prend une observation et on prend un des 2 joueurs qui a joué n matchs auparavant donc on crée un nouveau dataset temporaire qui contient n lignes et qui correspond aux précédents matchs joués par le joueur. On lance l'algorithme sur les j lignes " ensemble d'entraînement formant 80% du dataset trouvé" et on fait une prédiction sur la ligne numéros j+1 en sauvegardant le résultat trouvé dans une liste qui s'appelle prédiction.
- Étape 2 : On colle la ligne numéros j+1 avec ces informations au dataset trouvé plus haut et on relance de nouveau l'algorithme sur les 80% des observations pour faire une prédiction sur la ligne j+2 puis on sauvegarde le résultat trouvé dans la liste prédiction.
- Étape 3 : On répète les étapes sur les lignes restantes du dataset pour prédire à la fin l'issue du match pour le joueur sur la ligne target.
- Étape 4 : On compare les résultats dans la liste de prédiction avec les résultats réels pour obtenir un seuil de confiance sur la prédiction effectuée. Par exemple, si l'algorithme a bien prédit le joueur gagnant au niveau de la ligne 'target', c.à.d si il a prédit que le joueur winner du dataset a gagné le match, dans ce cas on prendra en considération un seuil de confiance et si il est supérieur à 70% par exemple, cela signifie qu'on peut avoir une grande confiance dans cette prédiction.

Prenons un exemple, on lance l'algorithme sur la ligne d'indice 32400 en utilisant le classifieur KNN, on obtient que l'algorithme a prédit correctement le gagnant de ce match qui est Haas T. avec un seuil de confiance de 81%.

On peut observer ci-dessous quelques prédictions effectuées sur différentes observations. À la ligne d'index 32436, notre algorithme avait prédit que c'était le perdant qui allait gagner alors que le bookmaker avait prédit que ce serait le gagnant donc ici c'est un exemple typique où la prédiction de l'algorithme est meilleure que la nôtre. En revanche, aux lignes d'index 32436 et 32438, ce sont nos prédictions qui sont meilleures.

Index	prédiction de l'algorithme	prédiction du bookmaker "B365"	B365W	B365L
32445	1	1	1.72	2.00
32436	0	1	1.22	4.00
32438	1	0	3.40	1.30
43417	1	0	4.33	1.2
43426	0	0	2.00	1.8

FIGURE 17 – Prédictions effectuées sur différentes observations

8.5.3 Prédiction sur une partie du dataset et analyse du résultat

Dans cette partie, on lance l'algorithme sur 8111 observations afin de raccourcir un maximum le temps d'exécution tout en essayant de conserver un maximum de statistiques.

Le résultat que l'on obtient est le suivant :

	bonne prediction	mauvaise prédiction
prediction de l'algorithme	0.79 %	0.21 %
prediction du bookmaker	0.75 %	0.25 %

FIGURE 18 – Comparatif avec les résultats des bookmakers

On observe que nos résultats sont meilleurs que ceux des bookmakers. Et ci-dessous, on peut également regarder comment évolue notre pourcentage de bonnes prédictions en fonction de la taille du jeu de données et le comparer à celui des bookmakers.

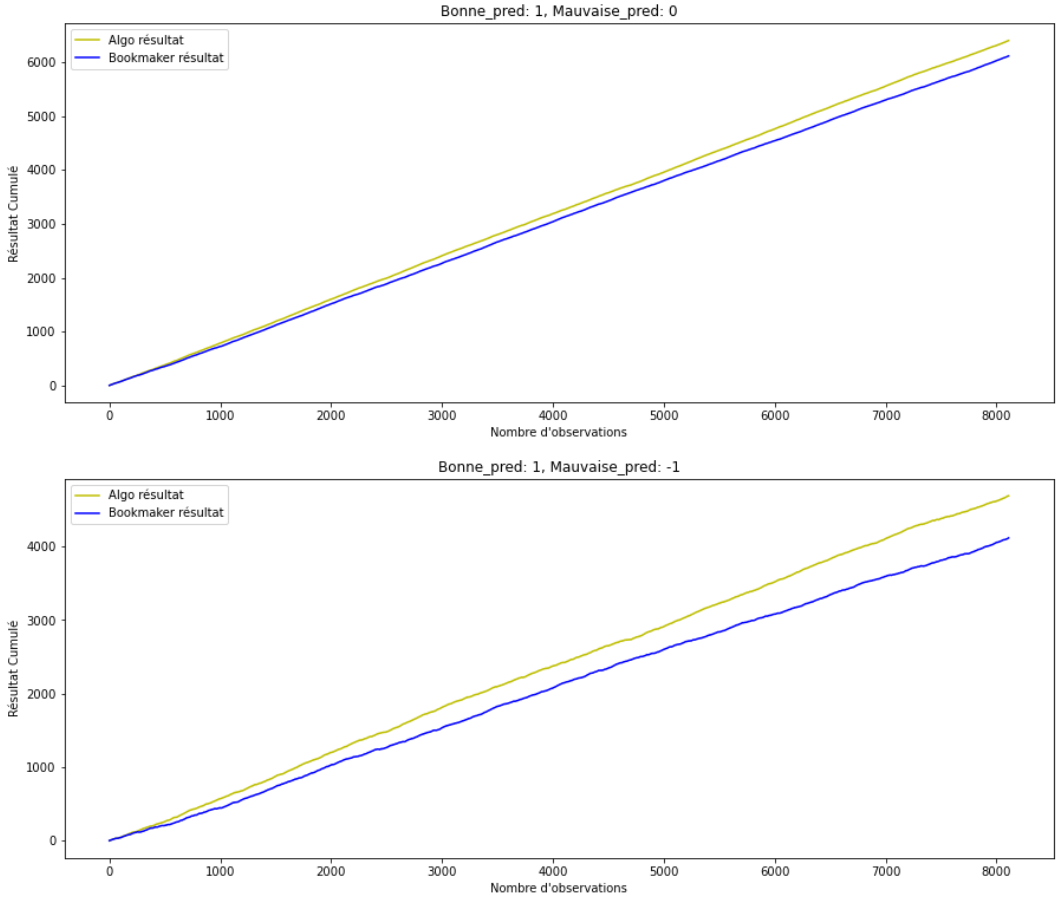

FIGURE 19 – Comparatif avec les résultats des bookmakers en fonction de la taille du jeu de données

Le graphe du haut correspond à +1 si la prédiction est bonne, 0 sinon. Et en bas, cela correspond à +1 si la prédiction est bonne, -1 sinon. En prenant l'exemple d'une côte 2 qui doublerait notre mise. En pariant 1 euros, on en récupérerait 2 en gagnant (+1) sinon on perdrait notre mise (-1). On voit alors que pour 8000 observations, on est proche des 5000 euros de gains.

En réalité les côtes ne sont pas toujours les mêmes donc si on prend en compte les différentes côtes proposées par les parieurs, on obtiendrait le résultat suivant. Si sur ces 8111 matchs on parie 1 euros sur chaque match prédit gagnant par notre algorithme alors le montant total de gain s'élève à 2993 euros contre seulement 42 euros si on se fie au modèle des bookmakers.

9 CONCLUSION

Les résultats de ces 2 approches que nous avons suivi doivent être comparé car vous aurez remarqué que les résultats diffèrent très significativement.

Dans la première approche nous nous intéresserons qu'aux statistiques entre 2 joueurs de tennis et non pas des matchs passés réalisés donc nous avons des informations générales sur le joueur et pas des informations spécifiques. Prenons par exemple, le cas d'un joueur comme Federer qui remporte Wimbledon en 2012. En 2013, il perd au second tour. Le second modèle va pouvoir prendre en compte cette information lorsqu'il jouera à Wimbledon en 2014, en conséquence il aura des informations que le premier modèle n'aura pas.

De plus, le premier modèle se base sur tous les matchs et ce quelque soit le nombre de matchs joués par le joueur auparavant. Si 2 joueurs s'affrontent et que les 2 n'ont joués que 10 matchs impliquant des statistiques très pauvres donnera donc lieu à des matchs très indécis et donc une probabilité de faire une mauvaise prédiction très grande. Et ces types de matchs sont très présents dans le dataset car comme nous l'avons dit, le dataset contient 60% de joueurs de seconde zone qui apparaissent sur le circuit principal que très peu de temps et se retrouvent parfois rapidement rétrogradés sur le circuit inférieur appelé Challenger Tour (équivalent de la ligue 2 au foot) et donc leurs matchs n'apparaîtront plus dans le dataset. Et là où le premier modèle diffère énormément du second, c'est que le second ne prend en compte que les matchs où le joueur a joué déjà plus de 50 matchs donnant lieu à des matchs dont l'issue est nettement moins indécise.

Cette différence nous permet d'introduire parfaitement la partie ouverture de cette conclusion où nous aurions aimé combiner ces 2 modèles où la prédiction pour un joueur de gagner un match repose directement sur ses matchs passés tout en utilisant la richesse de toutes les variables crées du premier modèle mais également toutes les informations que nous avons à disposition grâce aux différentes variables étudiées comme le tour en Grand Chelem où il est le plus intéressant de parier, etc.... Un modèle comme celui-ci nous permettrait de maximiser nos chances de gain.

En ce qui concerne les variables implémentées nous pourrions fournir encore plus de statistiques en faisant varier les années sur lesquelles celles-ci sont calculées comme le nombre de victoires calculé sur 3 ans, nous pourrions le calculer sur 2, 4, 5 etc et voir quelles variables seraient les plus corrélées à notre target.

Et pour terminer, nous aurions aimé essayer d'autres modèles de prédiction plus efficace comme le Boosting, Bagging ou encore Voting Classifier ainsi que des réseaux de neurones.

On voit que beaucoup de choses peuvent être encore faites sur un sujet comme celui-ci et qu'un mois et demi sur un projet comme celui-ci est bien trop insuffisant pour parvenir à accrocher les résultats obtenus par les bookmakers.