Analyse du jeu de données Pokemon – Weedle's cave

Bouvier Matteo, Metais Thibault, You Yi

13 novembre 2019

Résumé

Le jeu de données Pokemon - Weedle's cave présente 800 Pokemons et leurs caractéristiques telles que présentes dans le jeu du même nom, ainsi que les issues de 50000 combats de Pokemons l'un contre l'autre. Durant ce projet, nous avons identifié les variables les plus importantes pour décrire un Pokemon et ses chances de remporter un combat. Nous avons utilisé cette étude préliminaire pour établir une méthode de prédiction de l'issue d'un combat et nous avons obtenu un taux de précision de 94% à l'aide d'un arbre de décision.

1 Introduction

Notre groupe a travaillé sur le jeu de données "Pokemon - Weedle's cave". Il se présente sous la forme de deux tables : l'une présente les caractéristiques de 800 Pokemons et l'autre présente le vainqueur de 50000 combats entre deux Pokemons. Nous avons donc choisi de dégager deux problématiques principales. Dans un premier temps, nous analyserons le tableau des Pokemons et tenteront d'établir si certaines caractéristiques présentent un avantage sensible durant un combat. Dans un second temps, nous appliqueront les méthodes vues en cours pour identifier la méthode la plus à même de prédire le vainqueur d'un combat binaire.

Pour des raisons d'esthétique, nous avons choisi d'employer des bibliothèques permettant de personnaliser facilement l'apparence des graphiques, sans jamais modifier les calculs statistiques effectués. Nous avons ainsi utilisé notamment "ggplot2", "RColorBrewer" et "gridExtra".

L'une des difficultés rencontrés très tôt a été d'affecter un couleur à chaque type de Pokemon. Les Pokemons sont en effet caractérisés par un des 18 types existants (eau, feu, combat, vol, etc.). Il a été difficile de choisir une palette de couleurs permettant de correctement distinguer chaque type, mais nous avons finalement opté pour celle présenté en Table 1.

Table 1 – Palette de couleur des types de Pokemons

Couleur	Type	Couleur	Type
#B2DF8A	Bug	#000000	Dark
#B15928	Dragon	#FFFF33	Electric
#0090FF	Fairy	#E31A1C	Fighting
#FF7F00	Fire	#CAB2D6	Flying
#EE00FF	Ghost	#057700	Grass
#FDC086	Ground	#A6CEE3	Ice
#1B9E77	Normal	#3FFF00	Poison
#F781BF	Psychic	#666666	Rock
#6A3D9A	Steel	#084DA8	Water

2 Analyse descriptive

2.1 Analyse préliminaire

2.1.1 Tableau des Pokemons

Chaque Pokemon du tableau est décrit par cinq variables qualitatives : "Name", "Type.1", "Type.2", "Generation" et "Legendary". En effet, chaque Pokemon possède au moins un type primaire parmi les 18 cités précédemment, mais peut également posséder un type secondaire (également parmi les dix-huit cités précédemment). Les Pokemons considérés ici appartiennent aux six premières générations de Pokemons, sur les huit actuellement existantes. Chaque génération ajoute au jeu un nouvel ensemble de Pokemons qui viennent enrichir l'univers. Enfin, la variable Legendary est une variable binaire, indiquant si le Pokemon est légendaire ou non. Les Pokemons légendaires ont une importance particulière et une rareté plus importante dans l'univers du jeu.

Les Pokemons sont également décrits par six variables quantitatives : "HP", "Attack", "Defence", "Sp..Atk", "Sp..Def", "Speed", correspondant respectivement aux "points de vie", à l'"attaque", à la "défense", à l'"attaque spéciale", à la "défense spéciale" et à la "vitesse" du Pokemon. Ces caractéristiques sont directement utilisées durant les combats entre Pokemons : la vitesse détermine qui inflige des dégâts en premier,

les points de vie, la défense et la défense spéciale permettent d'encaisser les dégâts subits, l'attaque et l'attaque spéciale permettent d'infliger des dégâts à l'adversaire.

2.1.1.1 Variables quantitatives Avant tout, nous avons cherché à déterminer si plusieurs des variables quantitatives décrivant les Pokemons étaient corrélées. Nous avons donc réalisé un test de corrélation selon la méthode de Kendall, puisque les distributions des variables étaient significativement différentes d'une distribution normale (test de normalité de Shapiro-Wilk, P-value $<$ 1e-6). Les coefficients de corrélation obtenus sont présentés en Table 2.

Table 2 – Résultats des tests de corrélation entre variables quantitatives.

	HP	Atk	Def	Sp.Atk	Sp.Def	Speed
HP	1					
Attack	0.41	1				
Defense	0.32	0.37	1			
Sp.Atk	0.34	0.26	0.22	1		
Sp.Def	0.36	0.22	0.45	0.42	1	
Speed	0.18	0.26	0.06	0.33	0.22	1

Les coefficients de corrélation sont tous faibles (inférieurs à 0.5 en valeur absolue), ce qui indique que les variables quantitatives sont faiblement corrélées entre elles et qu'elles pourraient toutes avoir une influence distincte sur l'issue d'un combat. Nous décidons donc de garder les six variables quantitatives pour la suite de l'étude.

2.1.1.2 Générations Nous avons ensuite cherché à savoir s'il existait des différences significatives entre les Pokemons selon les générations, sur la base de six variables quantitatives décrites précédemment. (cf Figure 14) Afin de comparer les distributions des variables quantitatives, nous avons réalisé un test de Kruskal Wallis puisqu'il permet de généraliser le test de Student à plus de deux populations et permet ainsi de tester si les moyennes de plusieurs échantillons sont les mêmes. Il a également l'avantage de ne pas supposer que les distributions soient normales et que les variances soient égales, contrairement à l'ANOVA. Chaque génération comporte entre 82 et 166 Pokemons, les données sont donc suffisantes pour obtenir de bonnes conclusions. Les résultats sont présentés en Table 3.

Aucun des tests ne passe sous le seuil de P-value $<$ 0.01 et sont donc considérés comme non significatifs.

Table 3 – Résultats des tests de Kruskal Wallis menés sur les caractéristiques qualitatives des Pokemons, par génération

Variable testée	P-value
HP	0.03169
Attack	0.05265
Defence	0.3078
Sp..Atk	0.1913
SP..Def	0.154
Speed	0.06138

On suppose donc que les générations sont homogènes en terme de distribution des valeurs sur les six variables quantitatives.

Pour terminer cette première analyse, nous nous sommes demandés si certaines générations présentaient plus de Pokemons légendaires que d'autres. Nous avons donc calculé les proportions de Pokemons légendaires dans les générations 1 à 6. (Table 4)

Table 4 – Proportions de Pokemons légendaires, par génération

Génération	Effectif	Proportion
1	166	3.61%
2	106	4.72%
3	160	11.25%
4	121	10.74%
5	165	9.09%
6	82	9.76%

Nous avons réalisé un test de proportions, "prop.test" sous R, afin de déterminer si ces six proportions sont égales ou si elles diffèrent significativement. Le test renvoie une P-value de 0.0788, il ne semble donc pas y avoir de différence significative entre les générations.

Pris ensemble, ces résultats semblent indiquer que les générations sont plutôt équilibrées en terme des caractéristiques décrivant les Pokemons. Il existe certes des variations au sein de chaque génération, certains Pokemons ayant des valeurs plus hautes que d'autres, mais aucune génération n'est sensiblement déséquilibrée par rapport aux autres. Nous supposons donc que l'information sur la génération n'aura pas d'impact sur la détermination du vainqueur d'un combat et nous décidons de ne pas prendre en compte dans la suite de l'étude.

2.1.1.3 Types Nous avons ensuite voulu déterminer si les types de Pokemons (Type 1 et Type 2) avaient une influence sur les variables quantitatives décrivant les

Pokemons. Nous avons représenté en Figure 1 et Figure 2 les effectifs pour chaque Type 1 et Type 2 respectivement. On peut voir clairement que les modalités des Types sont distribuées de façon très inégale dans les Pokemons. En particulier, le Type 2 présente une majorité de "None", c'est à dire de Pokemons sans Type 2.

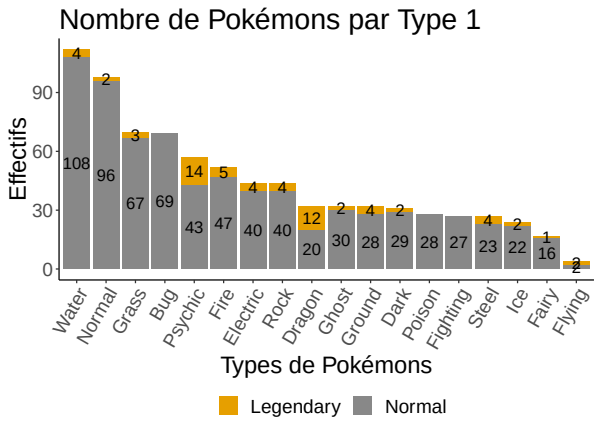

FIGURE 1 – Effectifs des modalités de Type 1

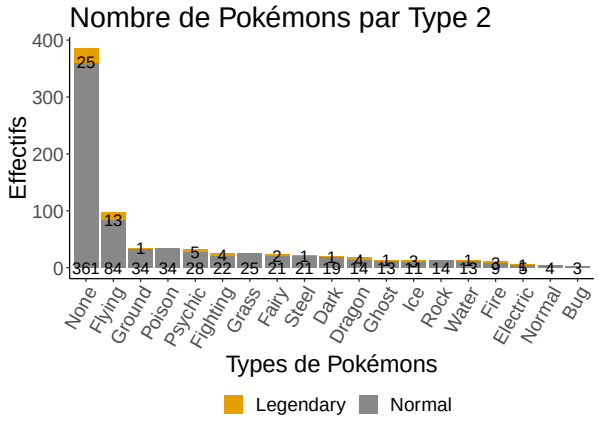

FIGURE 2 – Effectifs des modalités de Type 2

A nouveau, nous avons réalisé un test de Kruskall Wallis sur chacune des six variables quantitatives décrivant les Pokemons, groupés par Type 1, puis par Type 2. Les résultats obtenus sont présentés en Table 5.

On obtient ici des P-values très inférieures au seuil de 0.01, indiquant une forte influence du Type 1 et du Type 2 sur les caractéristiques des Pokemons, que l'on peut remarquer sur la Figure 16 en Annexe. Des résultats très similaires sont obtenus même après avoir supprimé les Types attribués à moins de 20 individus ("Fairy" : 17 individus, et "Flying" : 4 individus). Ceci

TABLE 5 – Résultats des test de Kruskal Wallis menés sur les caractéristiques quantitatives des Pokemons, selon Type 1 et Type 2

	Type 1	Type 2
Variable testée	P-value	P-value
HP	5.28e-05	2.328e-05
Attack	1.826e-12	2.885e-10
Defence	1.016e-15	1.035e-15
Sp..Atk	2.2e-16	3.468e-06
SP..Def	6.955e-06	5.006e-05
Speed	6.191e-10	8.899e-14

nous indique que l'information donnée par les six variables quantitatives et par les variables Type 1 et Type 2 sont probablement redondantes, mais il est difficile de conclure plus précisément. Nous décidons donc de maintenir les variables quantitatives et les types dans la suite de l'étude, mais nous nous attendons à ce que certaines ne soient pas utilisées par les méthodes de prédiction du vainqueur d'un combat.

On peut également remarquer que certaines modalités de Type 1 et Type 2 présentent plus de Pokemons légendaires que d'autres, variant de 0% à 50% pour Type 1 et de 0% à 25% pour Type 2.

2.1.1.4 Legendaires Nous avons donc voulu savoir si la variable "Legendary" avait un impact sur les caractéristiques quantitatives des Pokemons. Nous avons représenté en Figure 3 les valeurs prises par les variables quantitatives selon que les Pokemons soient légendaires ou non.

On constate une augmentation sensible des valeurs de toutes les caractéristiques des Pokemons légendaires par rapport aux non légendaires. Nous avons réalisé des tests de Wilcoxon (car la distribution n'est pas normale) pour comparer les médianes des valeurs prises par les six variables quantitatives pour les Pokemons légendaires contre les Pokemons non légendaires. En particulier, nous avons testé si les valeurs sont supérieurs chez les légendaires. Les P-values obtenues sont, dans l'ordre : 3.47e-17, 1.73e-19, 1.70e-14, 3.14e-26, 6.48e-21, 4.07e-19, indiquant que les Pokemons légendaires ont des valeurs quantitatives significativement supérieures aux autres Pokemons. La variable "Légendary" semble donc importante pour définir l'issue d'un combat.

2.1.1.5 Bilan Les résultats obtenus ci-dessus nous permettent de sélectionner les variables à utiliser dans la suite de notre étude. En effet, les six variables quantitatives sont suffisamment indépendantes pour être toutes

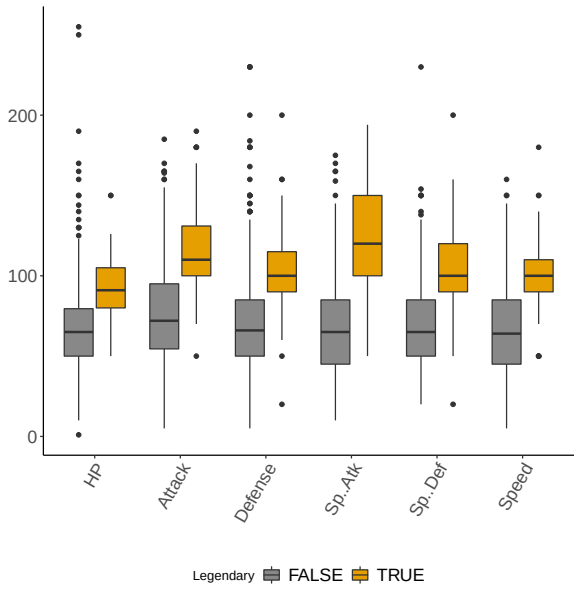

FIGURE 3 – Comparaison des valeurs prises par les variables quantitatives entre Pokemons légendaires et non légendaires

maintenues. Les informations sur Type 1, sur Type 2 et sur Legendary seront maintenue car ces variables ont une influence significative sur ces six variables quantitatives. Cependant, l'information sur la génération ne semble pas importante et peut être éliminée.

2.1.2 Tableau des Combats

Chaque combat ne comporte que trois informations : les indices des deux Pokemons s'affrontant et l'indice du vainqueur. Nous avons fusionné l'information apportée par ce tableau avec le tableau des Pokemons utilisée précédemment afin d'obtenir, pour chaque Pokemon, la proportion de combats qu'il a remporté. Nous avons également créé un nouveau tableau contenant les valeurs des six variables quantitatives pour les vainqueurs des combats et pour les perdants afin de les comparer. Il faut noter que seize Pokemons (Blastoise, Sandshrew, Wigglytuff, Poliwag, Victreebel, Magneton, Ditto, Ariados, Ursaring, Hariyama, Mega Latias, Honchkrow, Servine, Maractus, Jellicent et Pumpkaboo Small Size) ne sont pas présents dans le tableau des combats et seront donc retirés du reste de l'analyse.

2.1.2.1 Variables quantitatives Nous avons tout d'abord cherché à savoir si de fortes valeurs sur les variables quantitatives permettaient aux Pokemons de remporter plus de combats. Pour cela, nous avons utilisé la proportion de combats remportés par chaque Pokemon et nous avons mené un test de corrélation de Kendall entre cette proportion et les six variables quantitatives. Les résultats sont présentés en Table 6.

TABLE 6 – Résultats des tests de corrélation entre chaque variable quantitative et le taux de victoires.

Variables quantitatives	Tau de Kendall
HP	0.24
Attack	0.35
Defence	0.13
Sp..Atk	0.33
SP..Def	0.24
Speed	0.84

On peut voir ici que la variable "Speed" est fortement corrélée avec la proportions de combats remportés par un Pokemon, tandis que les 5 autres variables y sont surprenamment très peu corrélées, en particulier la variable "Defense". Ceci nous indique que la variable Speed sera probablement très importante pour déterminer l'issue d'un combat, tandis que les autres variables seront porbablement moins utilisées.

Nous avons également comparé les valeurs prises par les variables quantitatives entre Pokemons ayant remporté un combat ou l'ayant perdu. (Figure 4). On observe les mémes tendances que précédemment : l'écart sur la variable Speed est particulièrement marqué, tandis que les autres variables sont assez proches entre gagnants et perdants.

2.1.2.2 Legendary Nous nous sommes ensuite intéressés à l'effet de la variable "Legendary" sur l'issue du combat puisque, comme vu précédemment, les Pokemons légendaires bénéficient de meilleures valeurs sur les variables quantitatives. Nous avons donc comparé les proportions de combats gagnés de chaque Pokemon, selon qu'il soit légendaire ou non. (voir Figure 5). Pour cela, nous avons à nouveau réalisé un test de Wilcoxon pour déterminer si la médiane des proportions de combats remportées était supérieur chez les Pokemons légendaires, et nous avons obtenu une P-value de 2.51e-19. Il apparaît donc très clairement que le fait qu'un Pokemon soit légendaire lui apporte une chance supérieure de remporter un combat.

2.2 Analyse en composantes principales

Les combats se déroulant entre deux Pokemons, les variables quantitatives sont toujours à prendre par paire

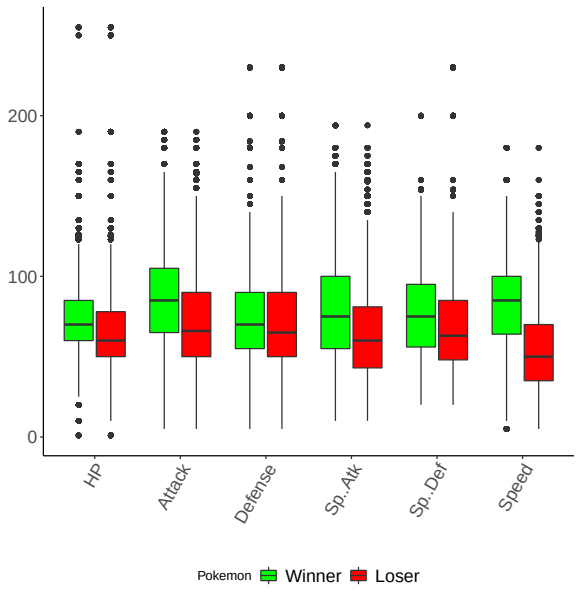

FIGURE 4 – Comparaison des valeurs prises par les variables quantitatives entre gagnants et perdants des combats.

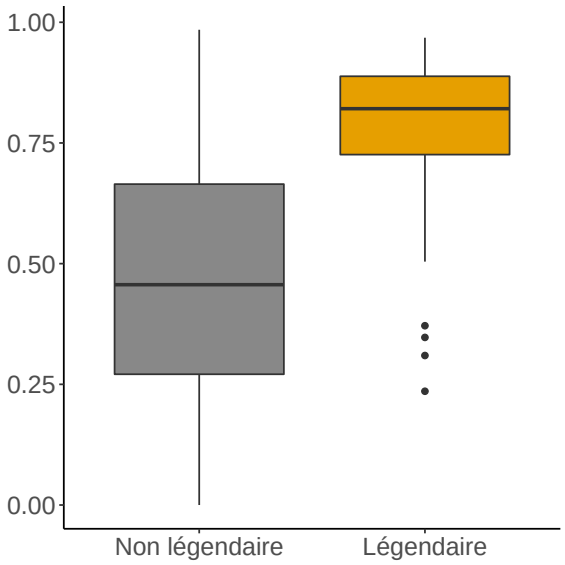

FIGURE 5 – Comparaison des proportions de combats gagnés entre Pokemons légendaires et non légendaires

(un exemplaire pour chaque Pokemon s'affrontant). En effet, comme nous l'avons vu plus haut, l'issue d'un combat dépend assez peu de la valeur prise par les variables quantitatives (sauf pour "Speed"). Nous avons donc créé de nouvelles variables en calculant l'écart entre les valeurs prises par ces variables entre gagnants et perdants. Nous espérons ainsi obtenir des variables plus à même d'indiquer l'issue d'un combat.

Ainsi, nous avons décidé de construire un nouveau tableau, en combinant les tables déjà existantes pour obtenir la différence entre les valeurs prises par les variables quantitatives (HP, Attack, Defense, Sp.Atk, Sp.Def et Speed). Nous y avons ajouté également le résultat du combat entre ces deux Pokemons sous la forme d'une variable binaire prenant la valeur 1 (le premier Pokemon a gagné) ou 2 (le second Pokemon a gagné). Cette dernière ne sera utilisée que pour indiquer si le premier ou le second Pokemon a remporté le combat. La Table 2.2 présente un exemple du tableau ainsi obtenue.

TABLE 7 – Tableau pour des études.

diff.hp	diff.att	diff.def	diff.spatt	diff.spdef
18	70	25	30	15
-5	-20	42	3	-5
-5	-27	3	-15	-15

diff.speed	résultat
62	1
-21	2
-11	2

Nous avons ensuite mené une ACP à partir de ce tableau à l'aide de la fonction "prcomp". Le résultat produit 6 composantes principales (Table 8). Tandis que le premier axe possède une valeur propre assez peu élevée (0.45), les seconde, troisième et quatrièmes composantes principales sont à peu près équivalentes avec des valeurs propres inférieures à 0.20. Cela nous indique qu'il va être difficile de mettre en exergue un petit nombre de variables afin d'obtenir une visualisation simplifiée. Cela est confirmé par la représentation graphiques en 2D selon les deux premiers axes factoriels (Figure 6) qui ne met pas en évidence une séparation totalement satisfaisante des deux groupes (vainqueurs et perdants). L'ajout d'un troisième axe factoriel n'apporte pas de gain réel sur la séparation des groupes.

On peut supposer que ces résultats mitigés sont dus au fait que l'on ne peut utiliser que les variables quantitatives pour réaliser l'ACP. Il aurait donc été intéressant de réaliser une Analyse Factorielle des Données Mixtes (AFDM), qui permet, grossièrement, de géné-

raliser l'ACP à l'analyse de variables quantitatives et qualitatives simultanément. Cette méthode dépassant le cadre du cours de SY09, nous avons préféré utiliser les autres méthodes vues en cours.

Table 8 – Variance expliquée par les axes factoriels.

Résultats	PC1	PC2	PC3	PC4	PC5	PC6
Proportion of Variance	0.45	0.18	0.13	0.12	0.07	0.05
Cumulative Proportion	0.45	0.64	0.77	0.88	0.95	1.00

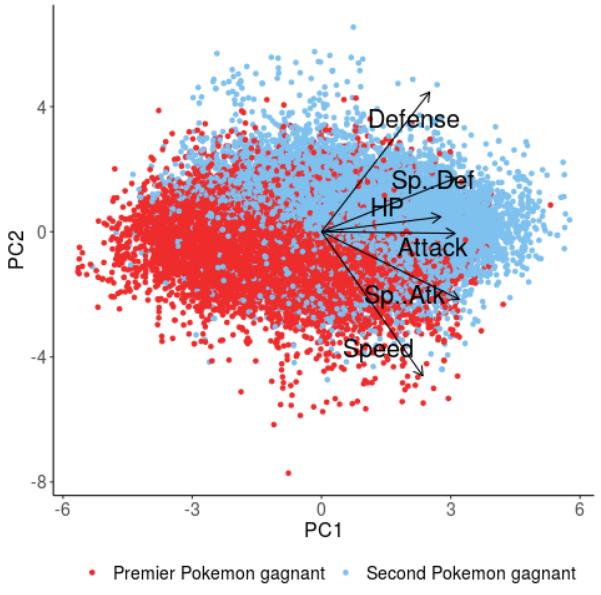

Figure 6 – Représentation graphique en composantes principales

3 Méthodes non-supervisées

3.1 Motivations et format des données

Nous avons tenté de réaliser une clusterisation non supervisée sur ces données afin de quantifier les performances que pouvaient produire ces méthodes. Le nombre de données étant important, il était difficile d'estimer si leur chevauchement visibles sur le premier plan factoriel de l'ACP était important et donc si une frontière de décision simple pouvait être trouvée ou non. Afin d'appliquer ces méthodes, nous avons décidé d'utiliser à nouveau le tableau construit lors de l'ACP. Il se compose donc des différences entre les caractéristiques numériques des deux Pokemons qui s'affrontent.

3.2 K-means

Nous n'avons que 6 variables (différences de attaque, défense, attaque spéciale, défense spéciale, vitesse et point de vie) à analyser pour faire le clustering. Il n'est donc pas indispensable d'utiliser les composantes obtenues par ACP. Comme il n'y a que deux résultats possibles pour un combat (gagnant ou perdant), nous avons fixé le nombre de clusters (k) à 2.

Étant donné que nous avons 6 dimensions, il n'est pas évident de représenter le résultat par schéma. Pour mieux visualiser le résultat, nous utilisons un dataset relativement petit, et les deux premiers axes qui correspondent aux différences de point de vie et d'attaque. Dans ces graphes, la forme des points indique le résultat du combat donné, et leur couleur le résultat de clustering par les K-means. Nous remarquons que, sur la Figure 7, les deux clusters se superposent largement. Si nous somme sur deux dimensions, nous aurons une frontière simple et assez linéaire entre ces deux clusters, quitte à classifier de manière inexacte tous les points qui seraient au delà de cette frontière. Vue que nous avons 6 dimensions, ce n'est plus possible de représenter une frontière nettement. En faisant plusieurs fois l'analyse, nous obtenons une précision d'environ 69%, ce qui n'est pas satisfaisant.

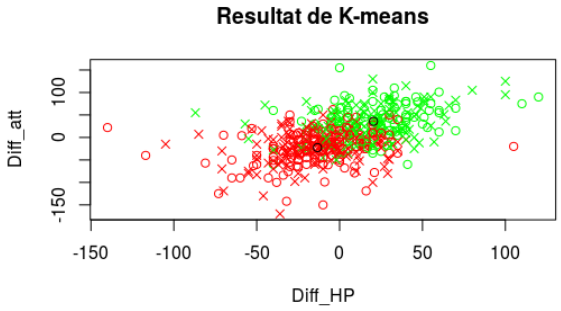

FIGURE 7 – Représentation graphique du classement par la méthode des K-means

ce qui est bien meilleur.

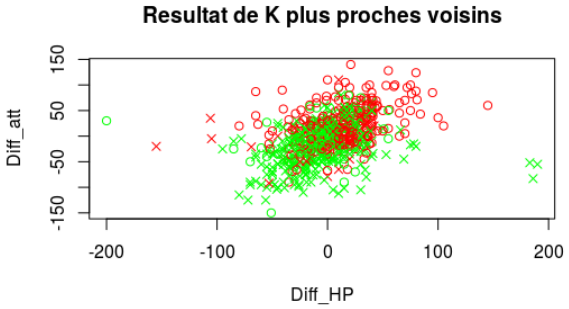

FIGURE 8 – Représentation du classement par la méthode des K plus proches voisins

4 Méthodes supervisées

4.1 Motivations

Puisque nous avions en notre possession des données avec les résultats des combats, nous pouvions utiliser les méthodes supervisées et ainsi chercher des frontières de décision plus complexes et peut être obtenir des résultats plus satisfaisants. Pour ces méthodes, nous avons utilisé le même tableau que précédemment.

4.2 Méthode des K plus proches voisins

La méthode des K plus proches voisins est également basée sur les distances entre les points, mais utilise aussi l'information de classe. Cette méthode a pour avantage de pouvoir identifier plusieurs "foyers" pour une même classe. Néanmoins, dans notre cas, il semblerait que seulement un foyer principal pour chaque classe soit présent. Puisque la méthode prend en compte la classe des points environnants (c'est une méthode plus locale que les k-means), nous devrions avec cette méthode obtenir une frontière de décision moins linéaire qu'avec cette dernière (si le nombre de plus proches voisins est suffisant) tout en évitant le sur-apprentissage (en gardant un nombre de plus proches voisins raisonnable).

Afin de trouver le nombre optimal de voisins à considérer, nous avons testé différentes valeurs. Nous avons remarqué que le résultat variait peut à partir de $k > 5$.

Nous avons utilisé la même méthode pour représenter le résultat obtenu que pour les K-means (Figure 8). Si nous sommes sur deux dimensions, la frontière obtenue est plus complexe (contrairement aux k-means qui était plutôt simple et linéaire). Cette flexibilité de la frontière de décision amène le résultat à environ 89% de réussite,

4.3 Analyse discriminante

Étant donné les résultats précédents, ils nous a semblé évident que se baser sur une frontière de décision simple n'était pas suffisant pour obtenir des résultats satisfaisants. Une autre manière d'obtenir une frontière de décision complexe est d'effectuer une analyse discriminante sur ces données : en faisant des hypothèses sur les paramètres de la loi normale multidimensionnelle, on peut obtenir les fonctions de probabilités conditionnelles à chaque classe et ainsi en déduire une estimation de la fonction de probabilité conditionnelle à un point donné.

Nous avons pour cela utilisé trois types d'analyse avec des hypothèses et propriétés différentes : une analyse discriminante quadratique, une analyse linéaire et une analyse naïve bayésienne. L'analyse discriminante quadratique se sert des estimateurs de vraisemblance afin de déterminer les paramètres de la loi associée, alors que l'analyse discriminante linéaire considère que la matrice de variance est commune à toutes les classes. Enfin, le classificateur bayésien naïf effectue une analyse quadratique en conservant seulement la diagonale des matrices de covariance.

Quant au classificateur bayésien naïf, il fait l'hypothèse de l'indépendance des variables entre-elles conditionnellement aux classes.

Afin d'obtenir une meilleure estimation des performances de ces méthodes, nous avons réalisé une boucle afin de générer aléatoirement différents ensembles d'apprentissage et de test. Nous obtenons ainsi environ 85% de précision pour les deux premières méthodes contre environ 77% pour la dernière. Cela peut s'expliquer de

par le fait que l'hypothèse de normalité sur la distribution des variables est fausse dans notre cas. Quand à la variance des variables conditionnellement aux classes, elle semble être sensiblement la même étant donné que les résultats ne diffèrent quasiment pas entre les deux premières méthodes.

Enfin, nous avons réalisé les graphes des courbes de niveau des fonctions de probabilité d'appartenance aux classes des trois méthodes selon diverses variables. Nous avons joint à ce rapport les graphes correspondants aux deux premières variables afin de les visualiser plus aisément et ainsi voir dans quelles proportions les hypothèses posées influent sur l'allure de ces courbes. (Figures 9, 10 et 11).

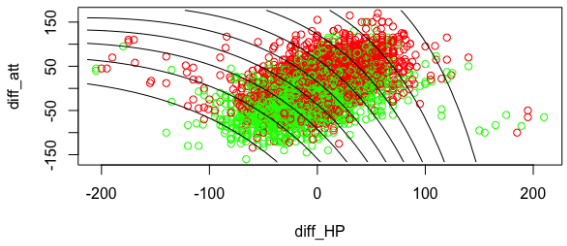

FIGURE 9 – Représentation des courbes de niveau selon les deux premières variables avec adq

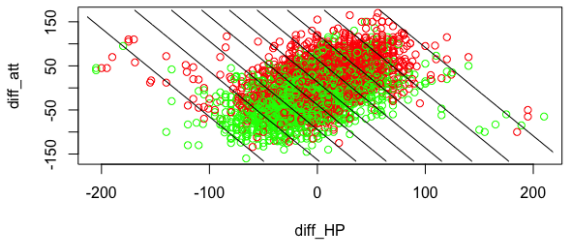

FIGURE 10 – Représentation des courbes de niveau selon les deux premières variables avec adl

4.4 Régression linéaire

Une autre méthode pouvant être utilisée pour approcher les probabilités d'appartenance aux classes est celle de la régression linéaire. Au lieu de faire des hypothèses sur les distributions conditionnelles, cette méthode estime directement les probabilités d'appartenance aux classes.

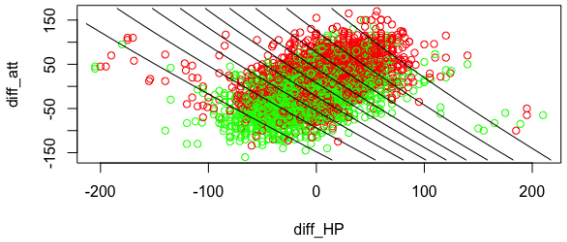

FIGURE 11 – Représentation des courbes de niveau selon les deux premières variables avec nba

L'utilisation de cette méthode pourrait palier aux défaut de l'analyse discriminante utilisée précédemment faisant l'hypothèse que les distributions conditionnelles suivent une loi normale, auxquelles s'ajoutent les hypothèses posées pour les analyses linéaire et bayésienne naïve.

Il est d'autant plus logique de l'utiliser au vu du nombre de classes qui est utilisé ici, c'est à dire 2 : la simplicité du modèle qui en découle en fait un choix populaire.

Après apprentissage sur un jeu de données défini aléatoirement, on obtient une prédiction précise à 89%, ce qui montre une légère progression par rapport à l'analyse discriminante effectuée précédemment. Les hypothèses posées précédemment avaient du influer de manière négative sur les résultats, d'où l'intérêt de ne pas les prendre à la légère, notamment l'hypothèse de normalité qui était commune aux trois méthodes de l'analyse discriminante.

On utilise ensuite les paramètres obtenus pour représenter la fonction de probabilité d'appartenance aux classes. Un graphique représentant les courbes de niveau de cette fonction selon les deux premières variables est représenté Figure 12.

4.5 Arbres de décision

Certaines caractéristiques intéressantes n'avaient pas pu été utilisées jusque là car elles étaient qualitatives et donc incompatibles avec les méthodes utilisées. L'utilisation d'arbres de décision pourrait nous permettre de mettre à profit ces valeurs et ainsi d'améliorer les résultats de nos prédictions.

De plus, on voit clairement dans l'analyse prélimi-

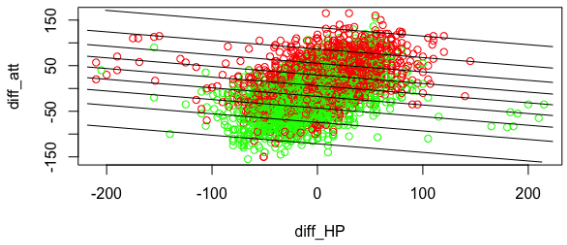

FIGURE 12 – Représentation des courbes de niveau selon les deux premières variables

naire que, tandis que la génération n'influe pas significativement sur le résultat d'un combat, la variable "Legendary" s'est révélée être importante dans la détermination du vainqueur. Ainsi, un grand avantage de cette méthode est qu'elle nous permettra d'exploiter cette variable qualitative inutilisée jusqu'à maintenant.

Afin d'effectuer cette l'analyse par arbre de décision, nous avons ajouté à notre tableau les variables binaires Légendaire1 et Légendaire2 indiquant si les Pokemons s'affrontant sont légendaires ou non.

La bibliothèque qui nous permet d'obtenir un arbre de décision prend plusieurs paramètres en entrée. D'après la méthode repeatedcv, on répète une validation croisée avec un nombre de répétitions donné. Ce nombre de répétitions correspond au nombre de fois où l'arbre sera construit avant de les combiner pour le résultat final. A chaque répétition, nous découpons l'ensemble d'apprentissage en 5 sous parties sur lesquelles sera réalisée une validation croisée. On donne dans les paramètres le critère à adopter lors du développement de l'arbre : ici le critère de Gini est utilisé.

On obtient ainsi une précision de 94% avec cette méthode. Le tableau joint (Table 9) représente le nombre de succès et d'échecs quant à la prédiction de l'issue du combat. On remarque ainsi que l'utilisation des variables supplémentaires semble avoir un impact significatif sur les résultats des prédictions.

Pour évaluer l'impact du nombre de répétitions et de la validation croisée, nous avons également réalisé des tests sans validation croisée puis avec validation croisée et avec une seule répétition. Les résultats montrent que les performances sont sensiblement les mêmes : ainsi il semblerait que le choix du jeu de données d'apprentissage n'aie que peu d'importance dans notre cas.

On peut représenter l'arbre obtenu sous la forme d'un graphique pour le visualiser (Figure 13). Une constatation importante est que la variable "Legendary" n'a pas été utilisée : ainsi l'influence de cette variable est, contre ce que nous pensions, négligeable dans le cadre de cette méthode par rapport à d'autres variables. Il est finalement probable que l'information apportée par la variable "Legendary" soit redondante avec les informations apportées par les variables quantitatives. La méthode de l'arbre de décision semble donc être la plus appropriée à notre cas, en se contentant des variables quantitatives.

TABLE 9 – Tableau pour des études.

Prédiction/Référence	1	2
1	9048	778
2	390	9784

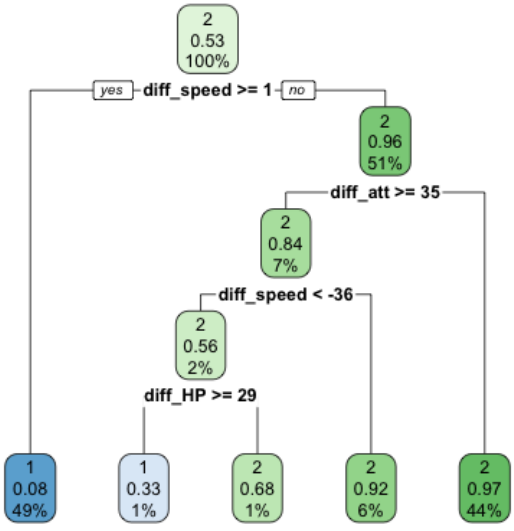

FIGURE 13 – Représentation graphique de l'arbre de décision obtenu

Les classes utilisées sont les mêmes que précédemment : les nombre 1 et 2 indiquent si le combat est gagné par le premier ou le second Pokemon. Sur ce graphique, le premier nombre correspond à la classe avec la probabilité la plus élevée à cette étape. Le second nombre est la probabilité pour que le second Pokemon gagne le combat. Le troisième représente le ratio de l'effectif au noeud choisi par rapport à l'effectif total.

5 Conclusion

Dans le cadre de cette étude, nous nous étions fixés comme objectif de parvenir à prédire, avec une bonne précision, l'issue de combats entre Pokemons. Dans un premier temps, nous avons tenté d'identifier les variables les plus importantes à considérer. Dans un second temps, nous avons appliqué les méthodes apprises en cours de SY09 et nous avons déterminé que la plus adaptée à notre cas est celle des arbres de décision. Cette méthode nous a apporté un taux de réussite de 94%, ce qui est très satisfaisant.
Nous avons été surpris par le fait que les variables qualitatives ne sont finalement pas utiles, contrairement à ce que l'on aurait pu d'abord penser. Nous expliquons cela par le fait que les variables quantitatives et qualitatives sont probablement fortement redondantes, ce que nous n'avions pas pu déterminer avec certitude dans la première partie. De plus, l'information sur la rapidité du Pokemon semble cruciale. Les Pokemons les plus rapides semblent avoir de bien meilleures chances de remporter leur combat !

6 Annexes

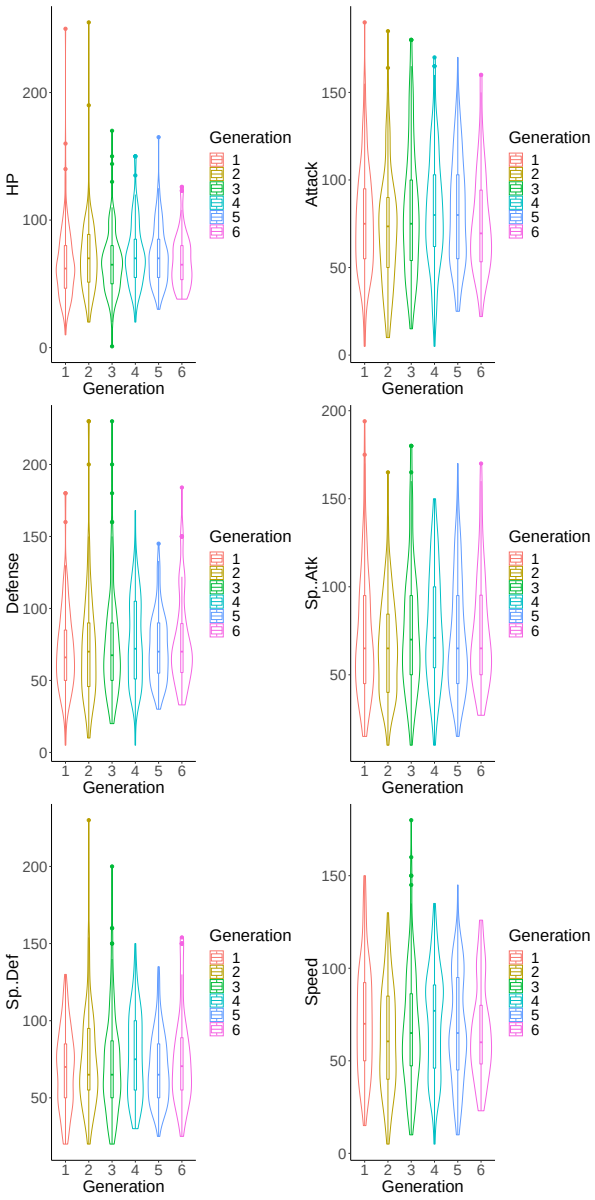

FIGURE 14 – Variables quantitatives décrivant les Pokemons, groupés par génération

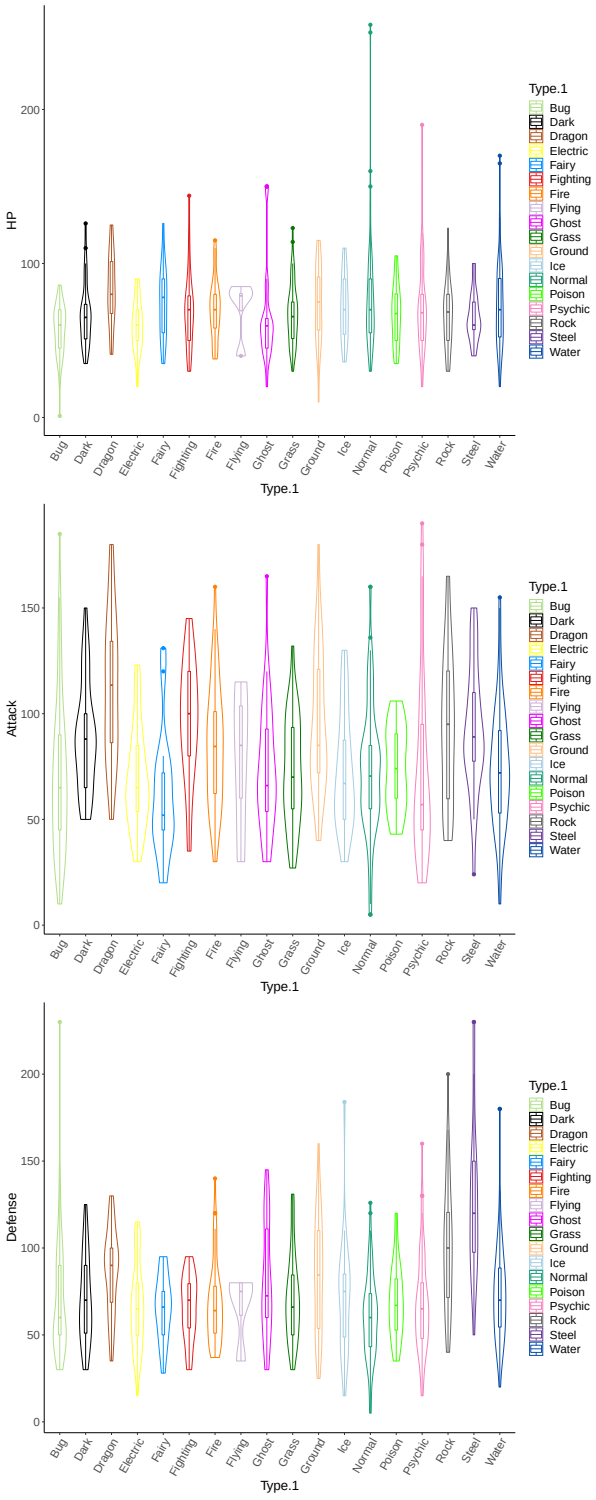

FIGURE 15 – Variables quantitatives décrivant les Pokemons, groupés par type 1 (partie 1)

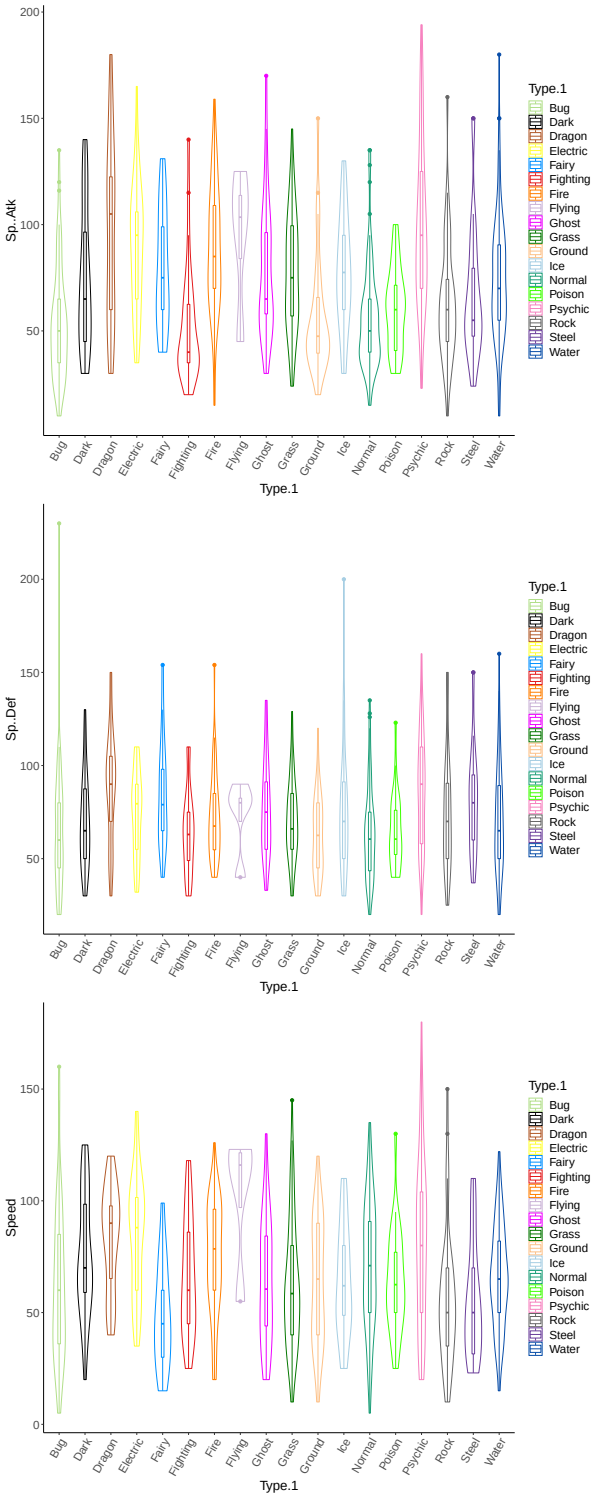

FIGURE 16 – Variables quantitatives décrivant les Pokemons, groupés par type 1 (partie 2)