

Master 1 mention Mathématiques appliquées, statistiques

Classification non supervisée

année universitaire 2021-2022

Projet de classification sur le jeu de données de l’univers des Pokémons

*Réalisé par :*

***LOUARN Justine***

***RAIMBAULT Lucie***

***DHENNIN Nolwenn***

## Classification des Pokemons

Lors de ce projet nous allons étudier un jeu de données à propos des caractéristiques des Pokemons. Exploité sous forme de jeux vidéo, dessins animés ou jeux de cartes, les Pokemons sont des créatures animales qu’il faut capturer, dresser et faire combattre. Nous avons accès aux Pokemons de la 1ère à la 8ème génération. Dans ce jeu de données, une ligne est égale à un Pokemon, nous avons leur caractéristiques nominales, leurs types, des données quantitatives (attaques, points de vies, taille, etc…), des caractéristiques de dressage ainsi que les dommages reçus contre un certain type.

# ANALYSE DES DONNEES

### Nettoyage des donnés :

* **Gestion des données manquantes** : Nous décidons d'enlever l'individu qui n’a pas de valeur pour le poids (“weight\_kg”) et ainsi que celui qui n’a pas de valeur pour le cycle de l'oeuf (“egg\_cycle”). Comme ces variables nous semblent importantes à garder dans notre projet, garder ces individus pourrait fausser notre analyse.
* **Gestion des Pokemons “mega”** : Nous avons aussi souhaité enlever les Pokemons méga de notre jeu de données.
* **Gestion des formes des Pokemons/numéro du Pokemon** : Pour plus de clarté dans notre étude nous avons décidé de ne garder qu'une seule forme de Pokemon. Pour cela nous avons gardé qu’un numéro de pokédex par Pokemon.
* **Gestion des valeurs aberrantes de “against\_ice”** : Nous avons ensuite remarqué que la variable “against\_ice” avait des valeurs aberrantes à 125 (alors que la variable est normalement entre 0 et 4). Après vérification sur le site internet PokéBip, nous avons donc remplacé toutes les valeurs 125 par ¼.

# CHOIX DES VARIABLES

Nous avons décidé de ne pas prendre en compte les variables “base\_friendship” et “base\_experience” car elles comportent trop de **valeurs manquantes** et sont donc un problème pour notre analyse.

Nous ne prenons pas non plus les variables qui sont les deuxièmes caractéristiques de certaines telles que “type\_2”, “ability\_2”, “ability\_hidden” ou encore “egg\_type\_2”.

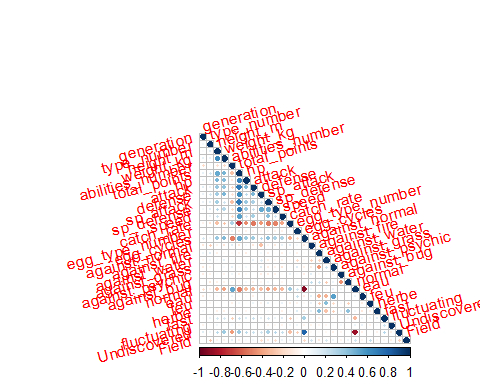
Pour les variables qualitatives qui comportent beaucoup de modalités (plus de 20 comme “species” et “ability”), nous avons choisi de ne pas les prendre en compte non plus afin de faciliter l’interprétation.

Pour les variables qualitatives ayant moins de 20 modalités, nous choisissons de faire des indicatrices sur quelques modalités nous semblant intéressantes :

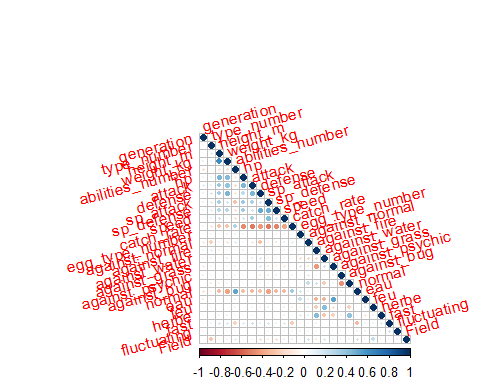
* Pour les status : nous choisissons de faire une indicatrice qui permet de savoir si le Pokemon est normal ou non.
* Pour les types : nous souhaitons garder les 3 types de base c'est-à-dire eau (“water”), feu (“fire”), et herbe (“grass”.)
* Pour la variable “egg\_type” : nous faisons 2 indicatrices sur ses modalités les plus présentes, c'est-à-dire une indicatrice sur “undiscovered” et une autre sur “field”.
* Pour les variables “against”, nous choisissons de garder seulement les modalités dont le type est le plus représenté. Ainsi, nous gardons : “water”, “normal”, “bug”, “grass”, “fire” et “psychic”.

Après nous être renseignées, nous avons trouvé qu’une catégorie de Pokemon avait la particularité d’être asexué. La variable “percentage\_male” aurait pu être intéressante mais elle fait partie des variables avec des valeurs non renseignées (NA). Nous avons émis l’hypothèse que les NA correspondaient au terme « asexué », mais après vérification dans le pokebip, ce n’était pas forcément le cas à chaque fois. Par conséquent, nous avons décidé de mettre de côté cette variable.

# Corrélation des variables

Nous voulons vérifier que nos variables ne sont pas trop corrélées entre elles et par conséquent n’apportent pas la même information.

On remarque qu’il y a une forte corrélation entre la variable “undiscovered” et 2 autres variables. Cela semble logique car souvent les pokemons légendaires ont la valeur undiscovered pour leurs œufs.

Ensuite, la variable “total\_point” est aussi très corrélée à d’autres variables qui doivent être utilisées pour le calcul du total de points. Nous décidons alors de mettre de côté ces variables pour le reste de l’étude.

# Standardisation des données

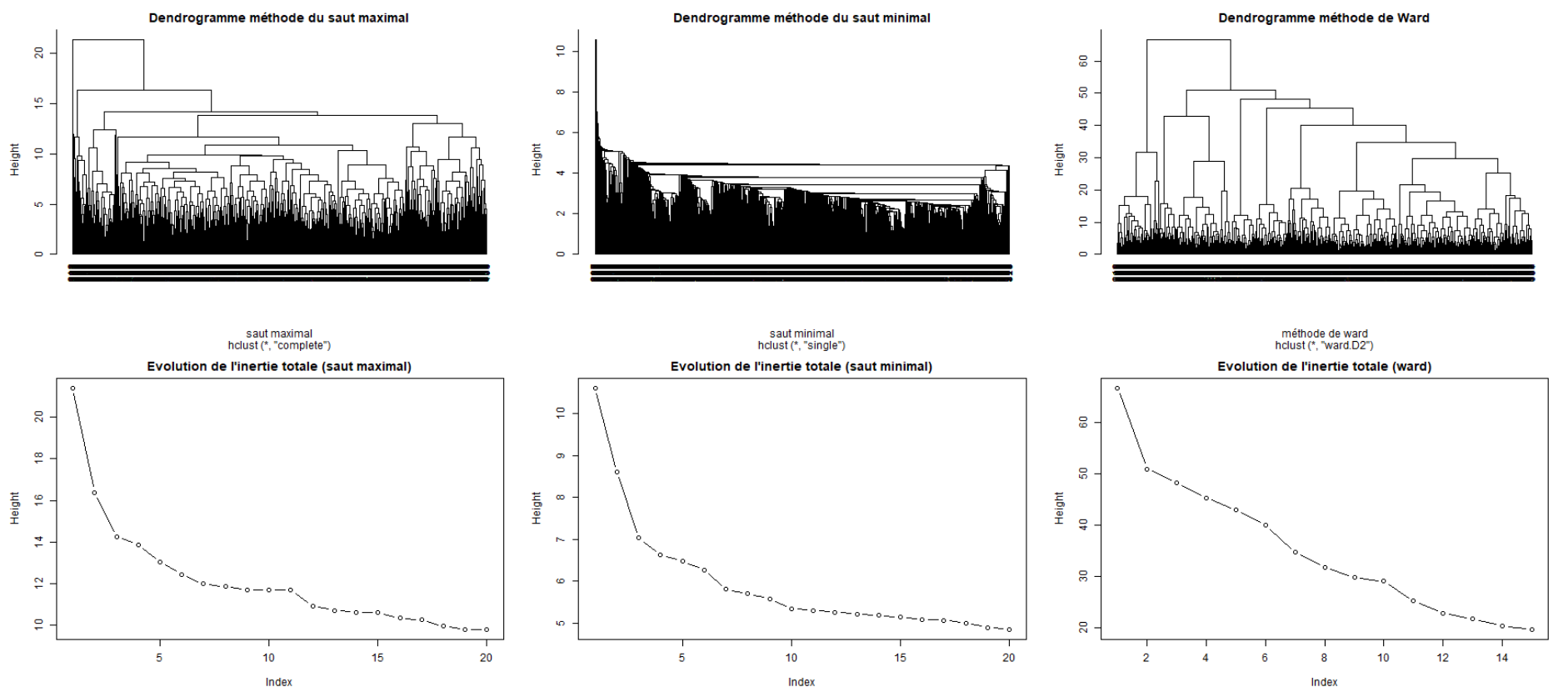
Pour mener au mieux notre analyse, il est nécessaire de vérifier si nos variables sont sur la même échelle. À l’aide de visualisations et du summary, nous constatons que ce n’est pas le cas et qu’il faut standardiser les données afin qu’elles soient à la même échelle. De plus, les variables ne sont pas de mêmes unités. Cette étape se réalise à l’aide la fonction scale() sur R.

# Classification des individus

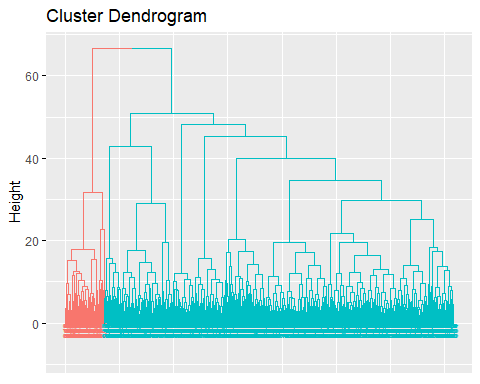
Après le pré-traitement des données, nous disposons de **895 individus** et **26 variables** quantitatives. Nous allons donc pouvoir appliquer les méthodes de classification (nécessitant des données quantitatives) directement à ce jeu de données.

Nos 26 variables étudiées sont donc : "generation", "type\_number","height\_m","weight\_kg", "abilities\_number","hp", "attack", "defense", "sp\_attack","sp\_defense","speed"n "catch\_rate", "egg\_type\_number", "against\_normal", "against\_fire", "against\_water", "against\_grass", "against\_psychic", "against\_bug", "normal", "eau", "feu", "herbe", "fast", "fluctuating" ainsi que "Field".

## Classification ascendante hiérarchique (CAH)

On effectue plusieurs classifications avec différents critères. La méthode du saut maximal, saut minimal et Ward.

D’après la distance du saut maximal on prendrait plutôt K=2 ou K=4, pour la distance du saut minimal K=3 et enfin pour la méthode de Ward K=2. En observant les dendrogrammes, on s’aperçoit rapidement que c’est le **critère de Ward** qui correspond le mieux à nos données. On voit qu’on peut partitionner nos données en 2 groupes. De plus les statistiques de Hubert et Dindex nous conforte dans le choix de **K=2.** Ci-dessous le dendrogramme de la représentation de nos groupes selon la méthode de Ward.



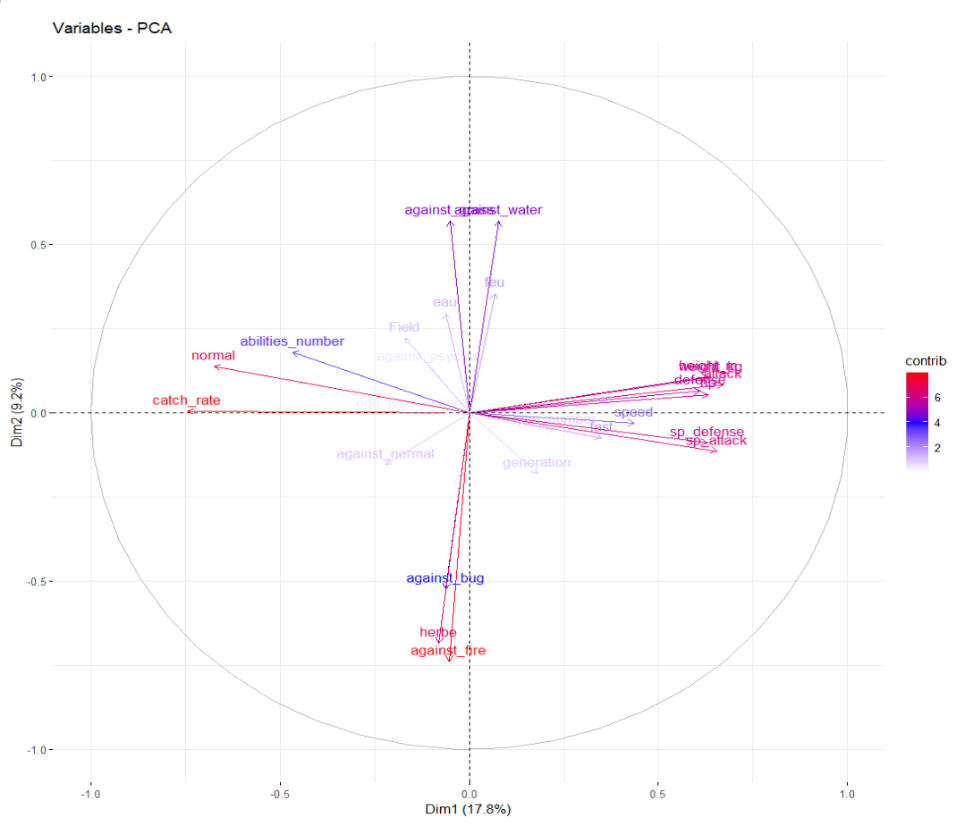
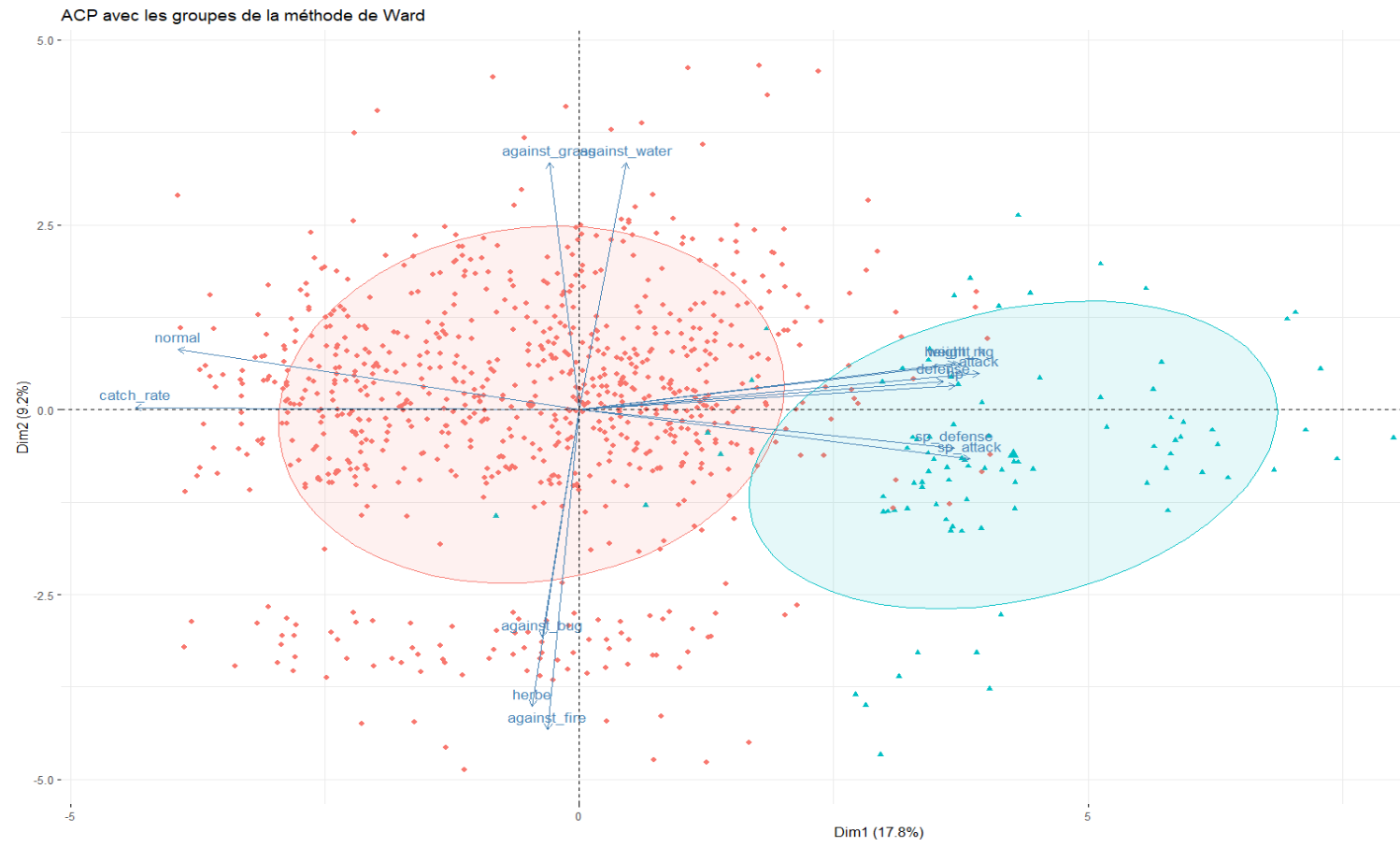
## Classification méthode des K-means

Nous avons aussi réalisé une classification avec l’algorithme des **k-means**, qui est basé sur la notion de centres mobiles, en effet à chaque étape les centres des classes sont redéfinis. D’après l’évolution des inerties totales nous choisissons K=2

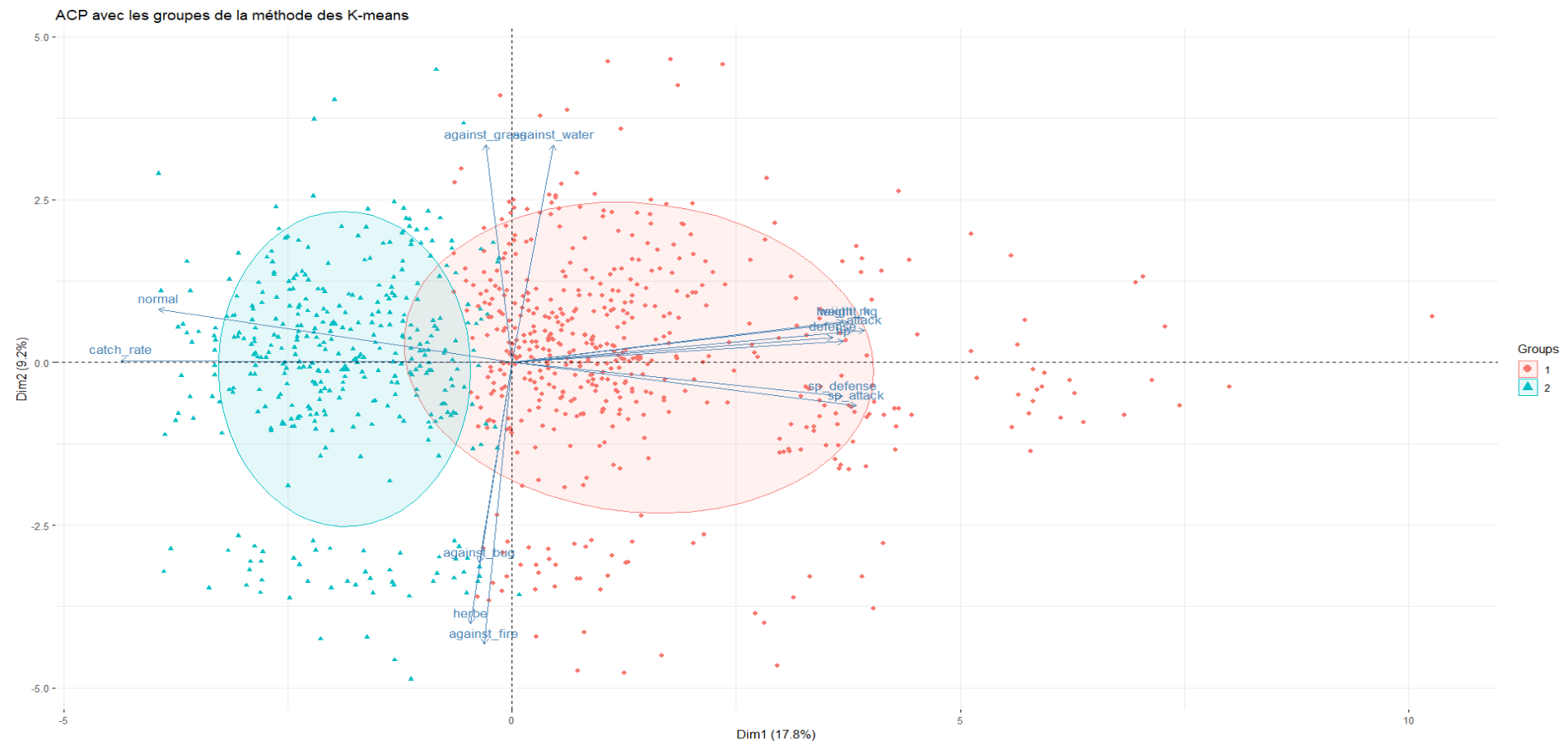
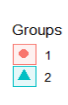
## Création des groupes

Nous avons réalisé 2 Analyses en Composantes Principales (ACP) avec les algorithmes des kmeans et de ward. Les individus y sont représentés sur les deux premiers axes principaux.

Nous avons conservé les **14 variables** qui contribuent le plus dans l’ACP.



Le premier plan factoriel explique 26,97 % de la variance. Les données ne sont pas très bien représentées, ce qui peut expliquer que les groupes paraissent un peu mélangés.

En comparant les 2 ACP, nous trouvons que la séparation des 2 groupes est plus optimale avec les **kmeans**. En effet dans la représentation de Ward des individus du groupe 1 se retrouvent au milieu de l’ellipse du groupe 2. Avec l’algorithme des kmeans, cette tendance est moins présente.

**Axe 1**

L’axe 1 semble opposer les Pokemons normaux et facile à capturer, aux pokemons forts. En effet, ils ont une bonne défense ainsi qu’une bonne attaque, et physiquement ils sont grands et lourds.

**Axe 2**

L’axe 2 oppose les résistances aux combats. Ce résultat est assez logique. En effet, un pokemon qui résiste bien à un type d’attaque(exemple : eau) est sûrement moins résistant à un autre type(exemple : feu).

**Les groupes**

Concernant les groupes, ils semblent se séparer en fonction de l’axe 1.

D’un côté on trouve un groupe de pokemons “normaux” avec un taux de capture important et de l’autre, des pokemons plus résistants avec de meilleures compétences en attaque et en défense.

## Conclusion

En conclusion, notre classification permet d’opposer les pokemons «normaux», qui se font facilement capturer aux pokemons avec de fortes capacités de combats.

On aurait pu choisir des variables différentes lors de nos choix des variables. Par exemple : on ne s’est pas beaucoup focalisé sur les variables « against ». Nous avions des variables corrélées entres elles, par conséquent, nous avons réalisé une sélection. Ainsi, le choix de différentes variables aurait pu avoir un impact sur notre classification.

Certaines variables auraient pu être intéressantes si elles n’avaient pas comportées des variables manquantes (le pourcentage mâles par exemple).

À partir simplement du dendrogramme avec la méthode de Ward, nous aurions pu penser prendre 3 groupes pour notre classification. Cependant, il est vrai qu'après observation de l'inertie totale, il était plus raisonnable de garder 2 groupes.

Pour aller plus loin, nous aurions pu réaliser une classification mixte car lors de notre CAH avec la méthode de Ward, on observait des individus mal classés. La classification mixte aurait permis de réduire ce problème. En effet, la classification mixte combine les approches hiérarchiques et centres mobiles pour tirer parti de leurs avantages respectifs.