



Formes et Fond


TOP14 Classification des équipes et Identification des facteurs de victoires

CLADIERE Nathan, projet 8, Formation Data Analyst OC





sommaire

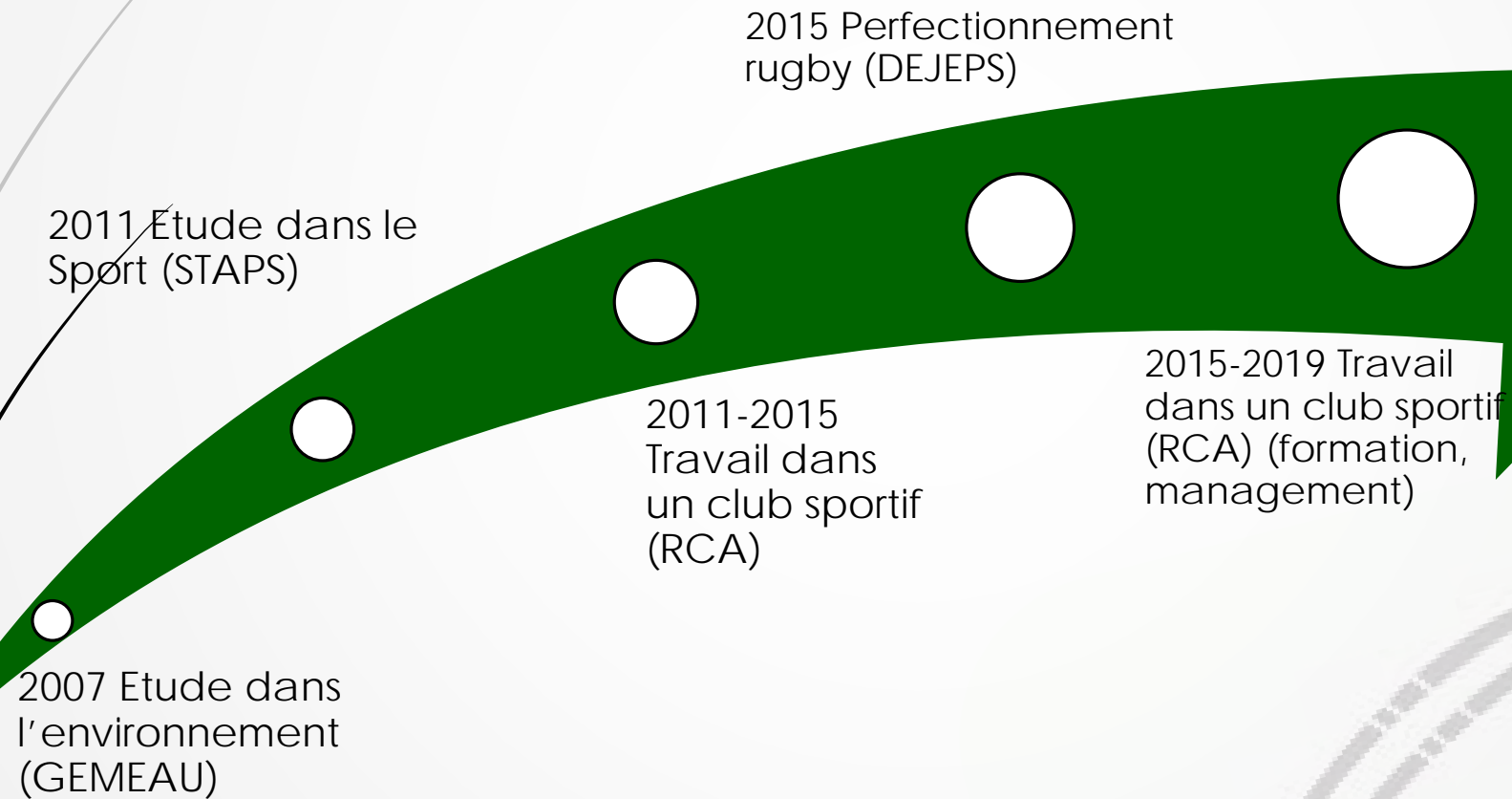
- Origine du projet
 - Le fond du projet
 - La forme du projet
 - Conclusions
- 



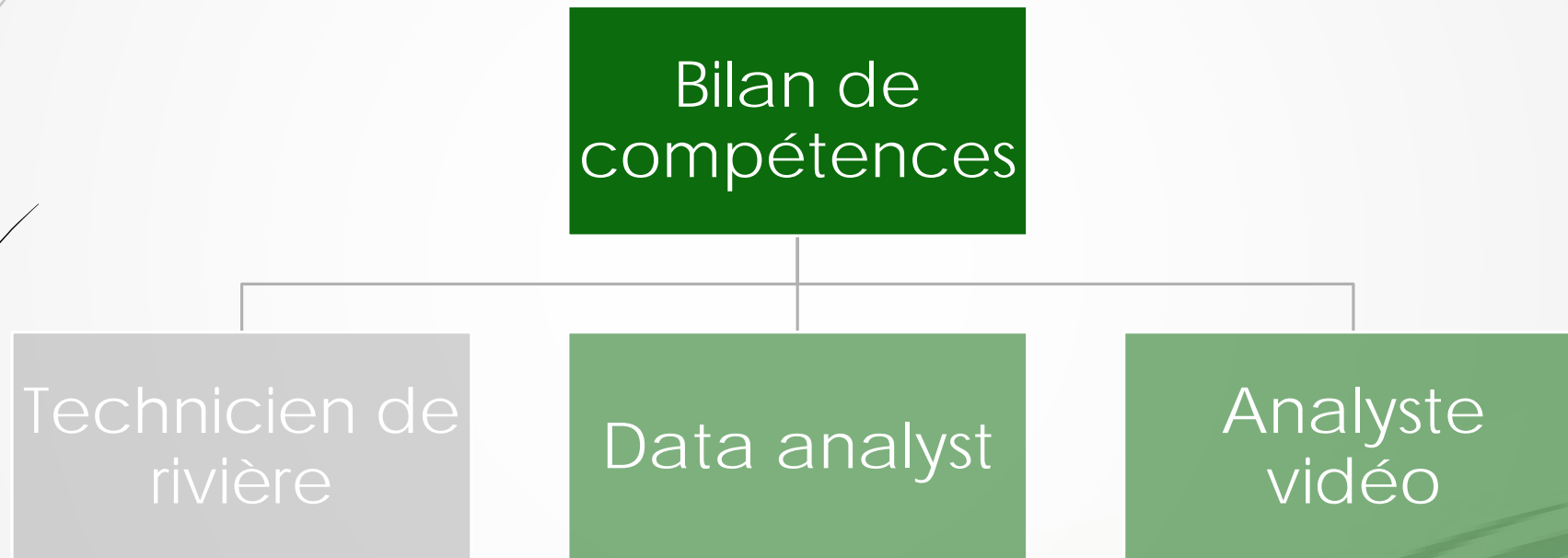
Origine du projet



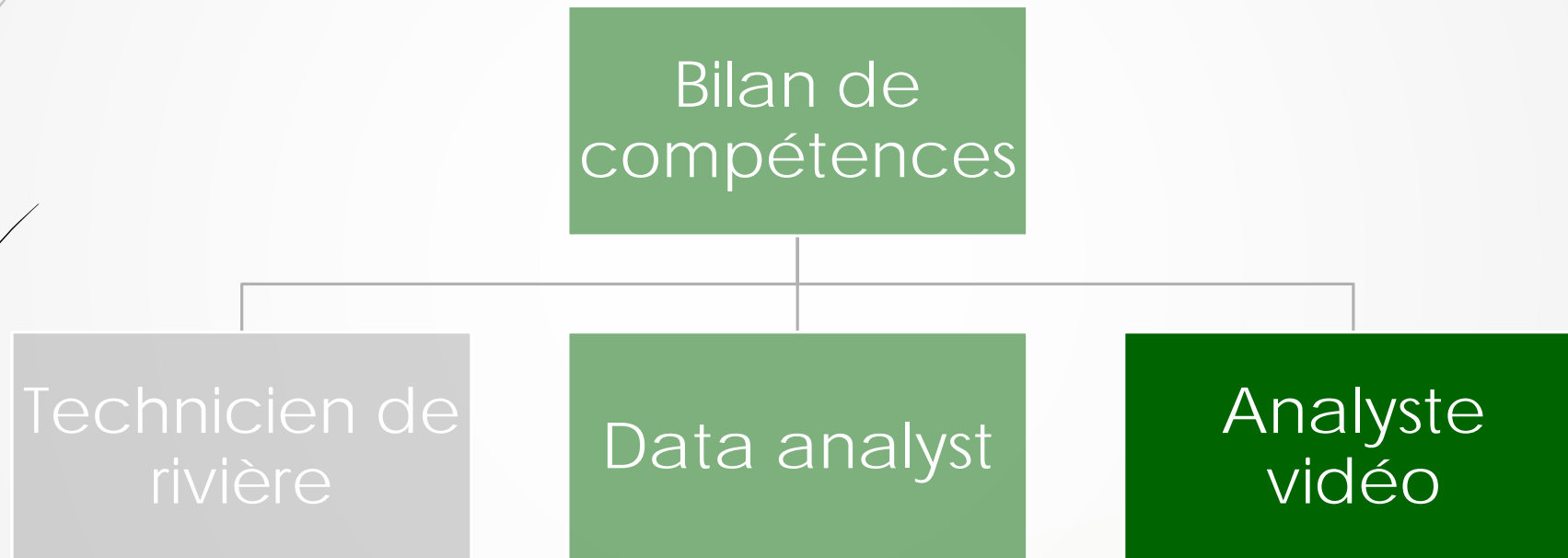
Mon parcours



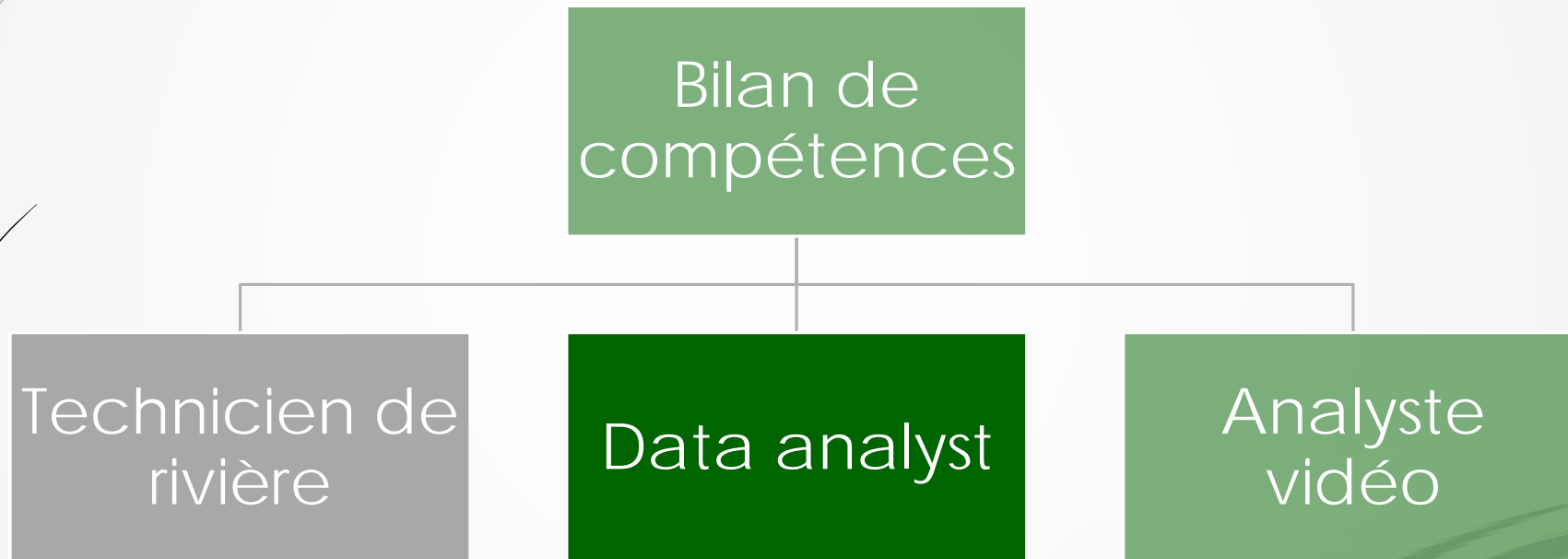
2019-2020 Riche en formations



2019-2020 Riche en formations



2019-2020 Riche en formations





Mes projets data

Sport

- Reporting statistiques match RCA
- Projet 8 et perspectives

Environnement

- Data for good

Social

- Data for good
- 



Le fond du projet



Le fond : Problématiques

Pourquoi ?

- Question personnelle: peut-on prédire le résultat d'un match de rugby?
- Données accessibles

Etudes et recherches anglo-saxonne

- Modélisation possible
précision de 80%
- Compliqué à l'international (niveau stratégique élevé, peu de rencontre)
- Manque de contexte

Problématiques

- Quels sont les facteurs de victoire en top 14 ?
- Peut-on classifier les styles de jeux en top 14 ?
- Données relatives/isolées

Le fond : Clustering des équipes

Distinction des équipes

- Capacités et/ou envie de tenir la balle

Données isolées ou relatives

- Suivent la même logique dans le clustering
- Quelques changements dus à la stratégie

Le fond: facteurs de victoires

Modélisations

- 82 % de précisions
- Régression logistique plus précis que Randomforest

Données relatives

- Précisions augmentent nettement en relatives (10%)

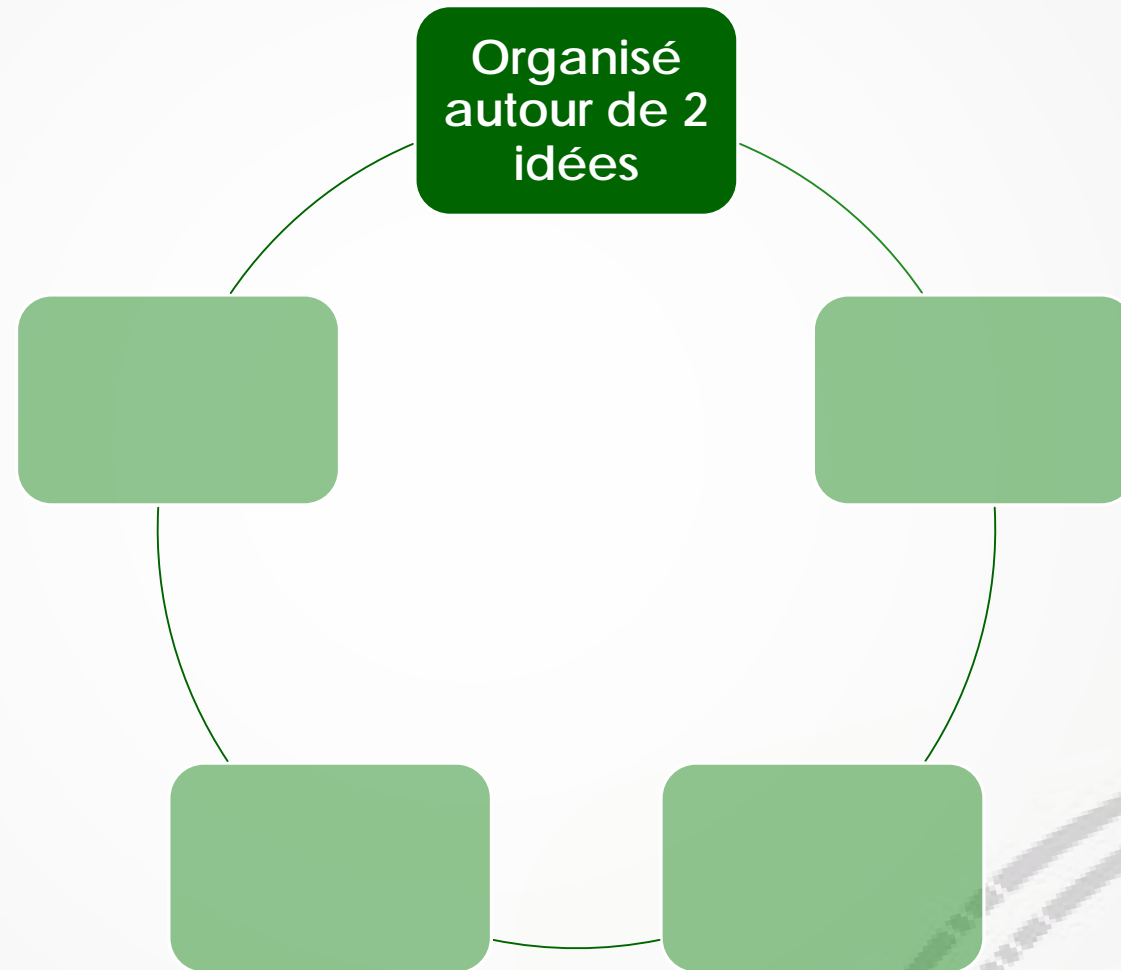
Facteurs de victoires

- Beaucoup de variables rajoutent de la précision
- 7 Variables ressortent (Turnovers, pénalités concédés, franchissements, ratio de rucks, ratio de transformations, nombre de passes)

The image features a minimalist design with several curved lines in green, black, and grey on the left side. A grey arrow points horizontally from the left edge towards the text. In the bottom right corner, there are two overlapping, curved grey shapes that resemble stylized parentheses or a swoosh.

La forme du projet

La forme



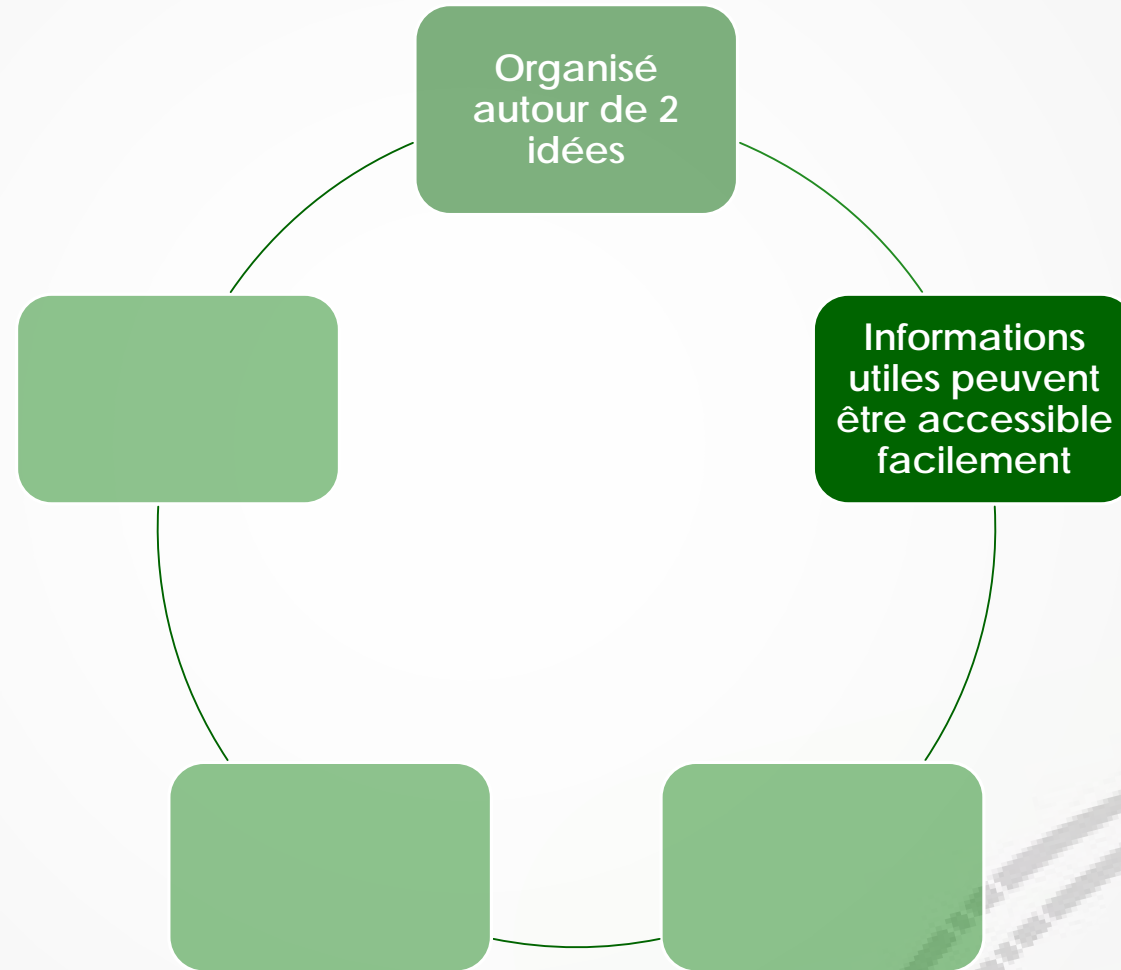


Organisation

Table des matières

1	Introduction	1
1.1	La data dans le sport	1
1.2	La data dans le rugby	2
2	Problématiques.....	2
3	Méthode	3
3.1	Démarche et outils.....	3
3.2	Les variables étudiées	4
4	Quels sont les types d'équipes	5
4.1	Regroupement des équipes	5
4.2	Plus de lisibilité avec l'analyse en composantes principales.....	6
4.2.1	Analyse des données isolées.....	6
4.2.2	Analyse des données relatives	7
4.3	Description des différents clusters (isolées)	8
4.4	Description des différents clusters (relatives)	9
4.5	Ce qu'il faut retenir des différents clusters	10
5	Facteurs de victoires	11
5.1	Quel meilleur modèle pour prédire la victoire	11
5.2	Quelles variables influence la victoire	12
6	Après ce premier projet.....	13
7	Références	14
7.1	Etudes	14
7.2	Articles	14

La forme





Information utiles

6 Conclusions et perspectives

Les analyses statistiques de cette étude ont conclu plusieurs choses, dans un premier temps:

- Il est possible de regrouper les équipes en fonction de leurs types de jeu.
- Face aux données relatives, les caractéristiques des équipes suivent la même logique.
- La clusterisation se fait principalement sur la capacité ou le désir de développer du volume de jeu.
- Cependant, il serait nécessaire d'entrer plus en détails dans la clusterisation afin d'établir plus précisément les styles de jeu des équipes.

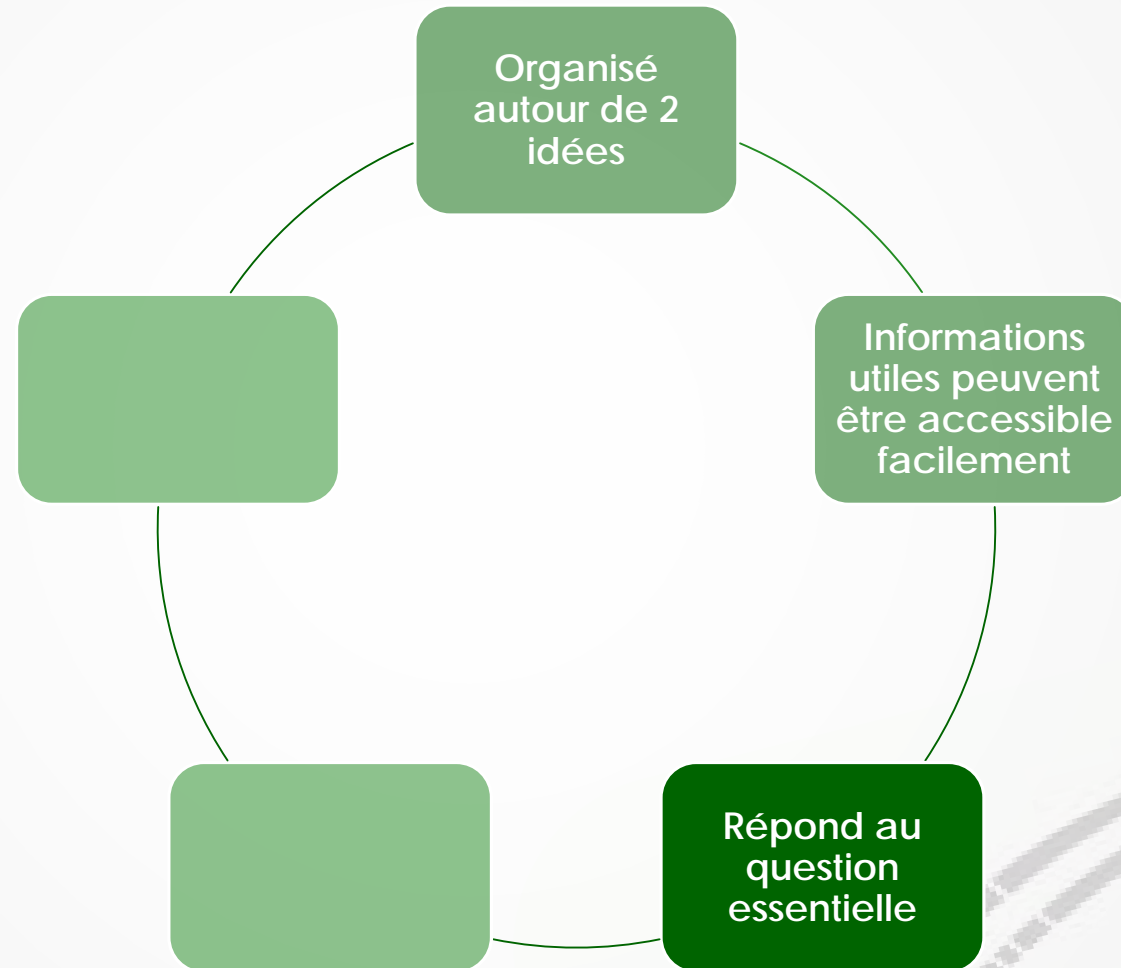
Dans un second temps :

- Bien que Bennett M(1) conclut que peu de facteur influence la victoire, on a vu à travers les différents modèles que la moindre source d'information peut rajouter de la précision.
- La régression logistique semble être un modèle mieux adapté à la prédiction de victoire.
- Les facteurs majeurs d'une victoire de Top14 sont (on retrouve aussi ses résultats dans la littérature anglo-saxonne) :
 - Les Turnovers (concedés puis acquis)
 - Les pénalités concedées par le cinq de devant
 - Les franchissements
 - Le ratio de Rucks réussis
 - Le nombre de mêlées
 - Le ratio des coups de pieds de conversions
 - Le nombre de passes
- Ces facteurs sont à considérer dans l'idée relative, il faut faire plus que l'adversaire. (Moins pour les pénalités et turnovers concedés)

Il serait intéressant de modéliser les victoires selon certains critères pour voir si les facteurs majeurs changent :

- Lieu du match (domicile/extérieur)
- Type d'équipe
- Type d'opposition

La forme





Qui ?
Quand ?

01/12/2020

TOP14

Classification des équipes et Identification des facteurs de victoires

Saison 2019-2020

Qui ?
Quand ?

2 Problématiques

Trois points vont être abordés par rapport aux équipes de top14, sur la saison 2019-20 :

	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster4
Données isolées	<ul style="list-style-type: none">• <u>SUA</u>• <u>CAB</u>• <u>CA</u>• MHR• SF• RCT	<ul style="list-style-type: none">• <u>AB</u>• SR• <u>LOU</u>• SP• <u>ST</u>	<ul style="list-style-type: none">• <u>UBB</u>	<ul style="list-style-type: none">• <u>R92</u>	<ul style="list-style-type: none">• ASM
Données relatives	<ul style="list-style-type: none">• <u>SUA</u>• <u>CAB</u>• <u>CA</u>	<ul style="list-style-type: none">• <u>ST</u>• <u>LOU</u>• <u>AB</u>	<ul style="list-style-type: none">• <u>UBB</u>• ASM• SR	<ul style="list-style-type: none">• MHR• SP• <u>R92</u>• SF• RCT	

Tableau de classification des équipes



Pourquoi ?

1.2 La data dans le rugby

Une étude recoupe toutes les études ayant été réalisées de 2007 à 2019(1) visant à déterminer les facteurs de victoires, un total de 41 articles ont été écrits à ce sujet. La plupart des articles se concentrent sur la collecte et l'analyse d'indicateurs de performances. Peu d'études se consacrent aux contextes du match (domicile, extérieur, enjeux, style d'opposition, météo).

Vingt-neuf indicateurs de performances sont utilisés à travers toutes les études et seulement quelques-uns sont communs :

- Jeu au pied
- Contre d'une touche adverse
- Essais marqués
- Points marqués au pied
- Plaquages effectués
- Turnover acquis

La dernière étude en date (2) se focalise sur la saison 2016-17 et une partie de la saison 2017-18 sur les matchs de la Premiership anglaise soit en tout 132 matchs.

La modélisation de la victoire dans cette étude s'est basée sur des données mises en forme différemment :

- Les données isolées : mesurées dans les matchs (400 m parcourus, 130 plaquages, etc.).
- Les données relatives : données par rapport à l'opposition, si l'équipe A a parcouru 300 m et l'équipe B 230 m, les données seront donc, équipe A : +70 et équipe B : -70

Ces données relatives se sont montrées plus significatives dans la précision de la modélisation (prédiction de victoire).

Quoi ?

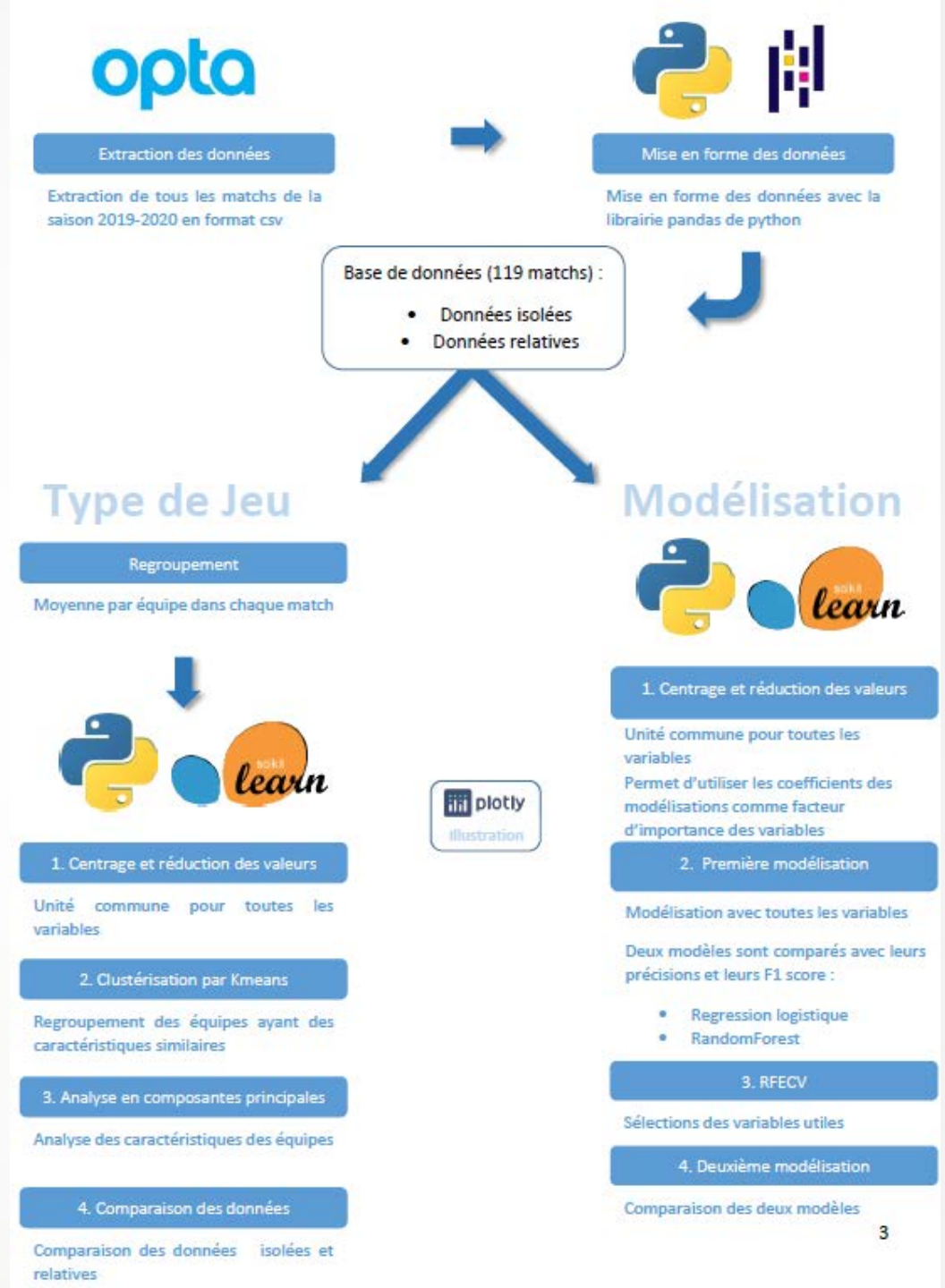
2 Problématiques

Trois points vont être abordés par rapport aux équipes de top14, sur la saison 2019-20 :

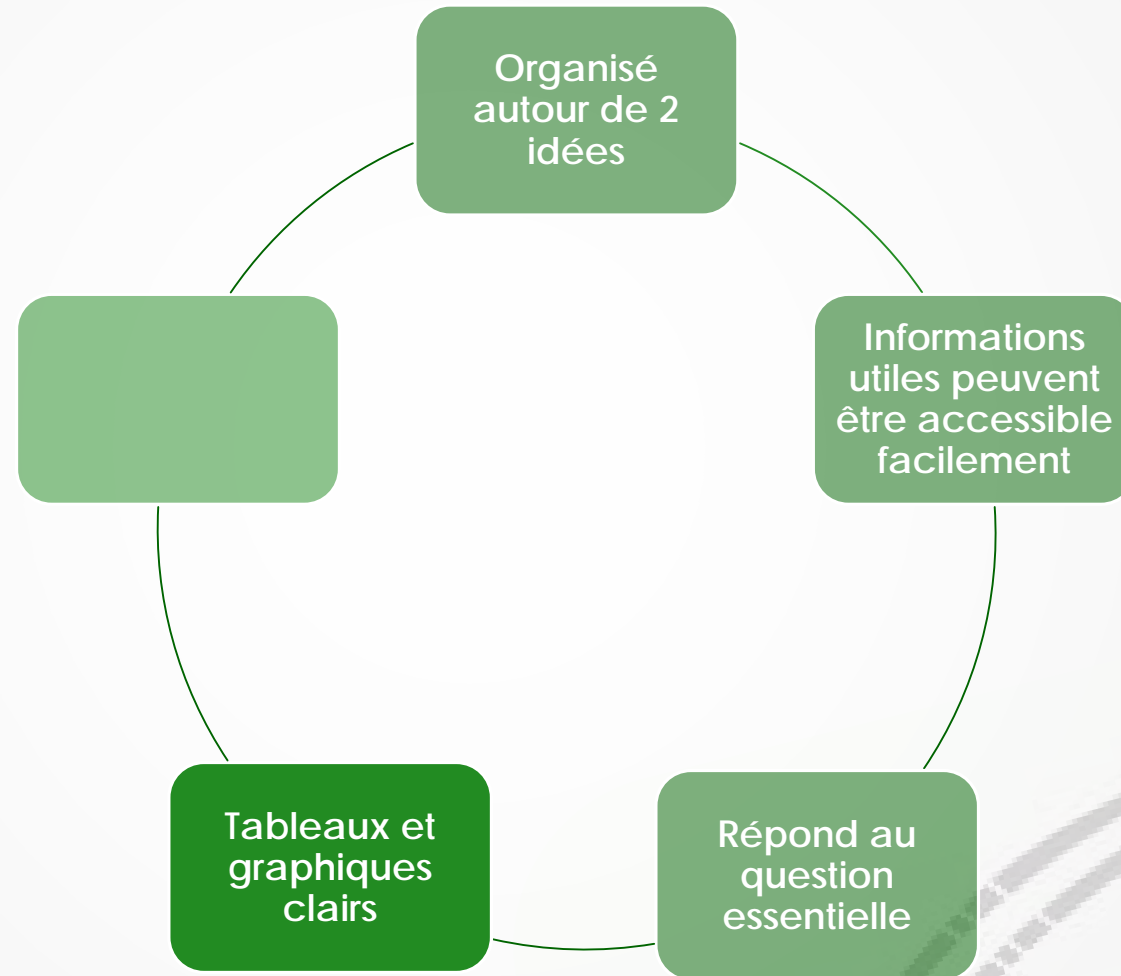
1. Par rapport aux manques de contextes sur les équipes (1), peut-on établir des types de jeux ?
 - a. Les équipes sont-elles groupables par style de jeu ?
 - b. Les équipes gardent-elles les mêmes caractéristiques avec les données relatives ?
2. En reprenant le principe de la deuxième étude (2), deux modèles de prédictions seront utilisés afin de savoir :
 - a. Quelle est la précision des modèles ? Quel modèle est le plus précis ?
 - b. Est-ce que la précision de la modélisation change avec les données relatives ?
 - c. Quels sont les facteurs de victoires pour une équipe de Top14 ?

Pour la deuxième partie, l'étude sera menée avec les points, sans les points et sans les facteurs de points (essais, transformations, et pénalités) car ces derniers prennent une importance inéluctable dans la modélisation avec les données relatives.

Comment ?

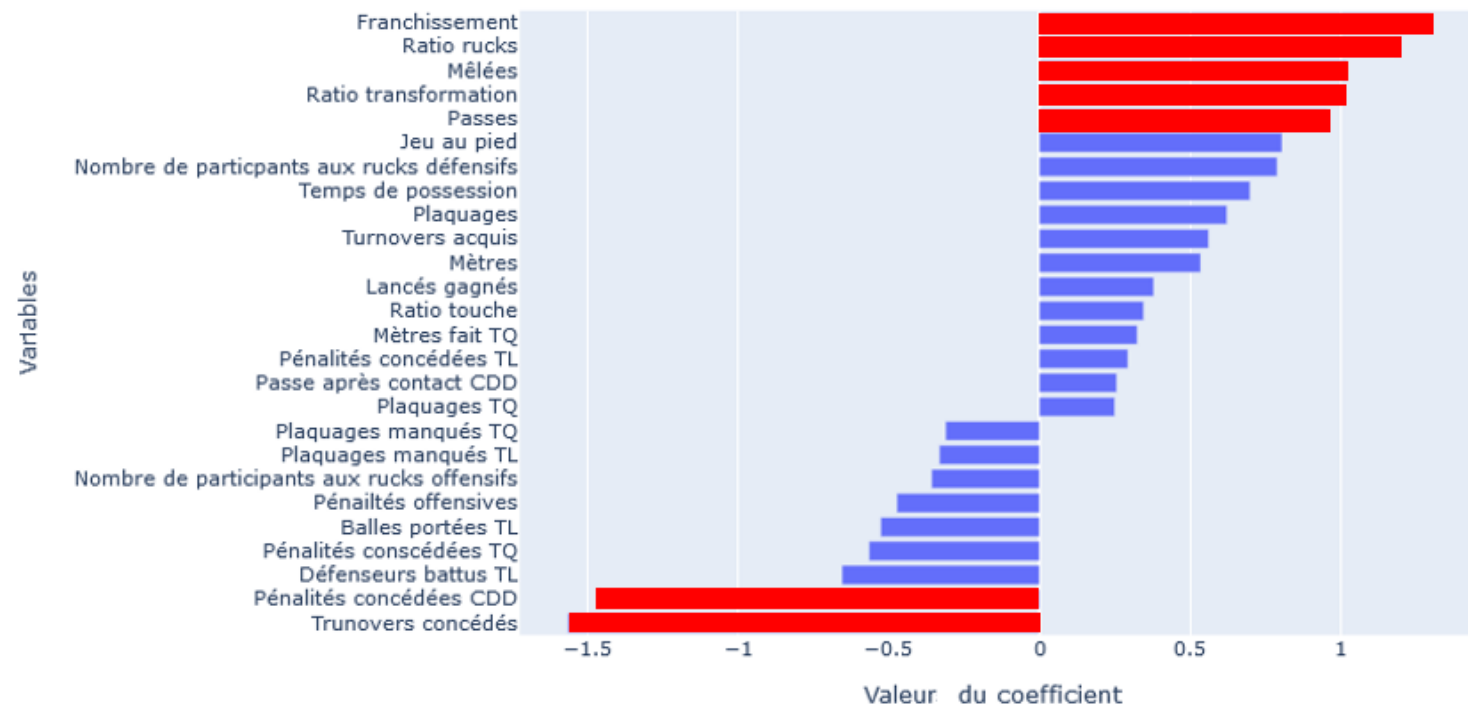


La forme



Clarté

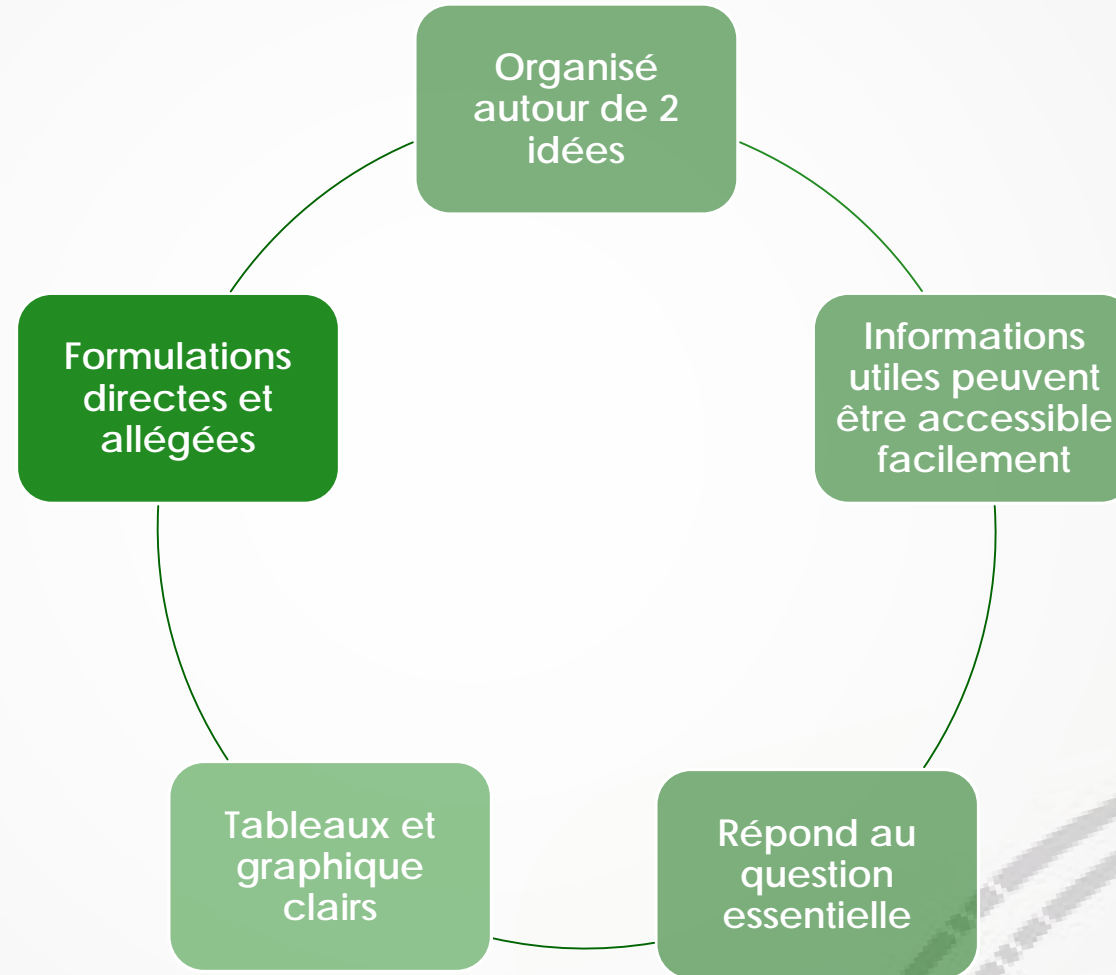
Valeurs des coefficients des facteurs majeurs d'une victoire en top14



	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Données isolées	<ul style="list-style-type: none"> • <u>SUA</u> • <u>CAB</u> • <u>CA</u> • MHR • SF • RCT 	<ul style="list-style-type: none"> • <u>AB</u> • SR • <u>LOU</u> • SP • <u>ST</u> 	<ul style="list-style-type: none"> • <u>UBB</u> 	<ul style="list-style-type: none"> • <u>R92</u> 	<ul style="list-style-type: none"> • <u>ASM</u>
Données relatives	<ul style="list-style-type: none"> • <u>SUA</u> • <u>CAB</u> • <u>CA</u> 	<ul style="list-style-type: none"> • <u>ST</u> • <u>LOU</u> • <u>AB</u> 	<ul style="list-style-type: none"> • <u>UBB</u> • <u>ASM</u> • SR 	<ul style="list-style-type: none"> • MHR • SP • <u>R92</u> • SF • RCT 	

Tableau de classification des équipes

La forme



Formulation

► Utilisation de liste :

● Analyse de la performance :

- Quantification et analyse des points forts et des points faibles
- Optimisation des temps d'entraînements

● Prévention des blessures :

- Les données biométriques et physiologiques permettent de déceler l'état de forme

► Style direct (Présent)

L'analyse en composantes principales **permet** de regrouper plusieurs variables entre elles. L'études du cercle de corrélations et l'analyse de la valeur des variables sur les composantes **servent** à définir les composantes.

► Pas de longs paragraphes



Conclusion



Ce que m'a apporté ce projet

- Mise en conditions réelles avec des données :
 - ACP
 - Clustering
 - Modélisation
- Seul face aux données :
 - Multitudes d'options
 - Pas de « bons » choix
- Perspectives:
 - Approfondir le clustering
 - Modéliser avec des critères
 - Envisager une régression linéaire multiple