

**MASTER MIAGE PARCOURS INFORMATIQUE ET INNOVATION.**

**UE: DATA MINING et BIG DATA**

**ENSEIGNANTS: Pr Antoine TABONNE.**

**Pr Olivier PERRIN**

***Rédigé par Alain NGUIDJOI BELL***

Rapport DU PROJET  
 DE BIG DATA

Sommaire

[Introduction 4](#_Toc34719312)

[1 Description de la méthodologies 4](#_Toc34719313)

[2 CHARGEMENT DES DONNÉES 4](#_Toc34719314)

[3 APERCUS DES DONNEES 4](#_Toc34719315)

[4 EXPLORATION DES DONNEES 6](#_Toc34719316)

[5 NETTOYAGE ET PRETRAITEMENT DES DONNEES 6](#_Toc34719317)

[6 SELECTION DES FEATURES 6](#_Toc34719318)

[7 ENTRAINEMENT DES MODELES 6](#_Toc34719319)

[8 VALIDATION DES MODELES 7](#_Toc34719320)

[9 EVALUATION DES PERFORMANCES 7](#_Toc34719321)

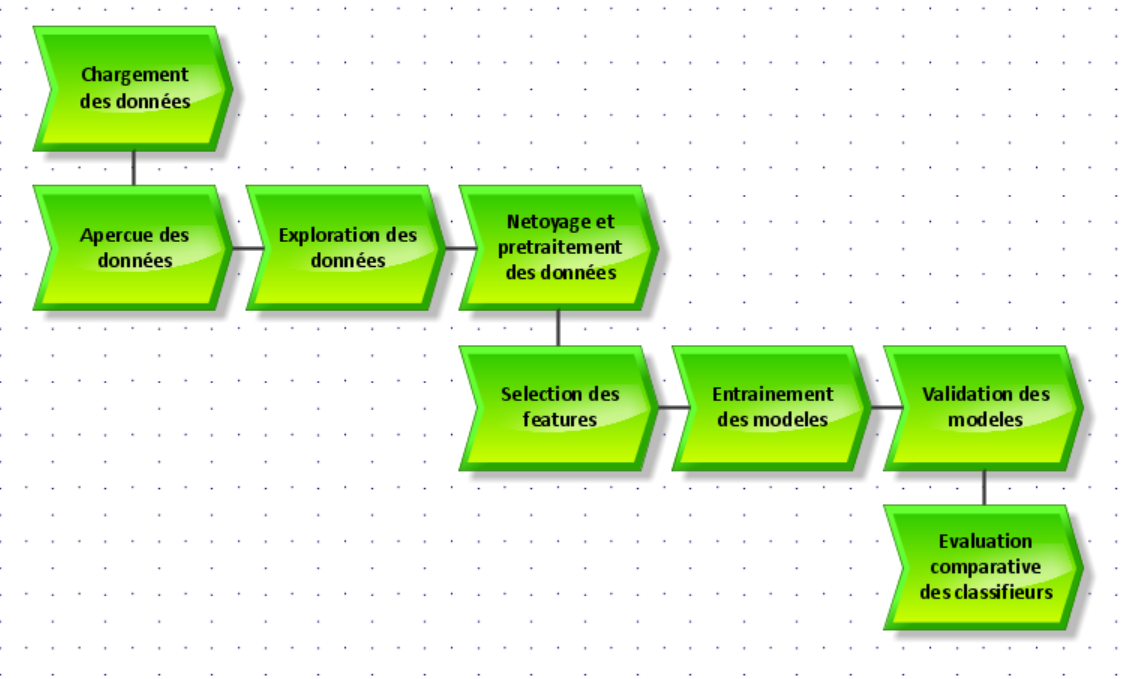
[Conclusion 7](#_Toc34719322)

# Introduction

Ce document résume les étapes de mise en œuvre du projet de Big Data. Ce projet consiste à modéliser le taux de désabonnement ou churn grâce à deux algorithmes d’apprentissages que sont la régression logistique et le Random Forest, puis de comparer les résultats obtenus. Pour réaliser ce travail, nous avons tout d’abord effectué un bref aperçu des données grâce aux logiciel ANACONDA, ensuite grâce aux bibliothèques seaborn et panda nous avons effectués une analyse exploratoire des données, des transformations des certaines donnée ont ensuite été effectué, puis nous avons construit le pipeline de nos model grâce à Spark après entrainement de ces dernières nous avons terminé par une analyse comparative des performances de nos deux modèles.

# MÉTHODOLOGIES

Dans le cadre de l’analyse des données, nous avons suivis un processus en plusieurs étapes représenté dans le schéma suivant :



Nous allons décrire de façons succincte dans les paragraphes suivants ce qui a été fait pour chacune des étapes de notre processus.

# CHARGEMENT DES DONNÉES

# APERCUS DES DONNEES

En chargeant le fichier de donnée avec l’application Orange 3 du Logiciel ANACONDA, on obtient le tableau récapitulatif suivant :

|  |
| --- |
|  |
|  |
|  |

Ce tableau nous présente les types de donnée et les valeurs possibles de chaque features, en prenant en compte que notre label est représenté par la colonne **Churn** et que le champs **customerID** ne représente que les utilisateurs, On peut conclure que notre ensemble de donnée comporte :

19 Features dont trois numériques et 16 catégorielles, et un Label de type catégoriel.

Si le récapitulatif nous indique bien qu’il n’y’a pas de valeur manquante ce qui est bien confortable pour les valeurs catégoriques ou les valeurs possibles sont déterminées, on ne connait pas les valeurs possibles des features de type numérique qui peuvent avoir des valeurs nulles ou vide.

|  |
| --- |
|  |

# EXPLORATION DES DONNEES

# NETTOYAGE ET PRETRAITEMENT DES DONNEES

# SELECTION DES FEATURES

# ENTRAINEMENT DES MODELES

# VALIDATION DES MODELES

# EVALUATION DES PERFORMANCES

# Conclusion

Au terme de ces deux travaux pratiques, force est de constater d’une part, la richesse des fonctionnalités de **POWERPLAY** bien qu’il ne soit pas possible de modifier les courbes à afficher à l’aide de fonction mathématiques personnelles, et que les définitions de sous catégories soient restrictives sur le nombre de contraintes, sa richesse dans la création de rapport a permis la réalisation du travail demandé.

D’autre part l’exécution manuelle de l’algorithme de K-Means utilisé dans le cadre de Clustering en Big Data, nous a permis de comprendre réellement son utilité ainsi que son fonctionnement.

Il aurait été souhaitable qu’il en fut ainsi pour l’ensemble des algorithmes utilisés en Big Data, et plus particulièrement les réseaux de neurones; de manière à maitriser les méandres de l’apprentissage en profondeur, dont les applications sont très nombreuses de nos jours, néanmoins force est de constater que le déroulement de cette algorithme ainsi que son apprentissage nécessite la maitrise d’un ensemble d’outils mathématiques sans lesquels ils n’est pas possible de véritablement les comprendre, ce qui confirme la pensée de **Severinius BOECE (480-524, Rome)** , selon laquelle : **"Tout a été créé par les nombres qui étaient le modèle exemplaire dans l'esprit du créateur"**.