

**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)

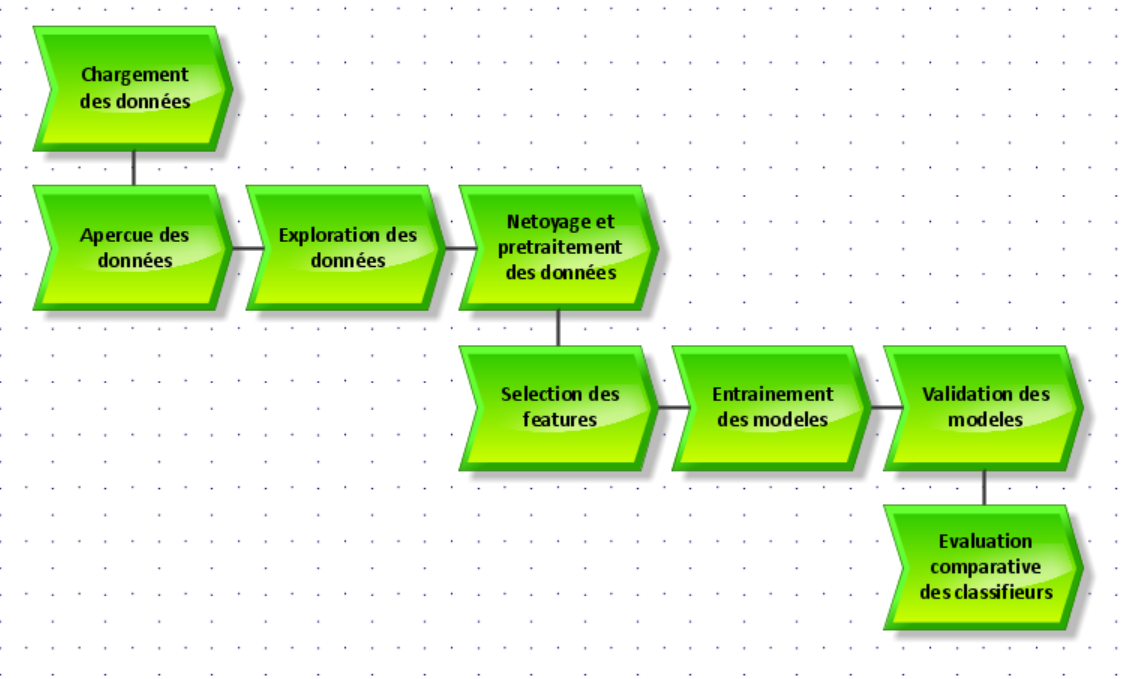
[Conclusion 7](#_Toc34719322)

# 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 succinctes dans les paragraphes suivants ce qui a été fait pour chacune des étapes de notre processus.

# CHARGEMENT DES DONNÉES

Les données se trouvant dans un fichier csv, ont été chargé grâce à Spark

# 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

L’exploration des données vise à analyser et visualiser les données de manières a en ressortir les rapports entre les différents éléments du schéma des données, d’en ressortir les proportions relatives et absolues qui pourront éventuellement nous permettre d’orienter notre prétraitement en tenant compte de la réalité contenue dans les données.

Dans cette perspective, un dessin valant mieux que mille mots, nous avons établis un ensemble de schéma. Le premier vise à déterminer les proportions des différentes valeurs possibles du label, pour savoir si les données sont équilibrées au niveau de leur proportion, ensuite nous avons établis un ensemble de schéma permettant de ressortir la distribution des différents attributs par rapport au label représenté par l’attribut « Churn », et pour le faire nous avons effectué un distinguo entre les valeurs positives (churn) de cet attribut et les valeurs négatives (no churn).

## Proportion des churn et des no churn

|  |
| --- |
|  |
|  |

## Variation de l’attribut TotalCharges par churn et no churn

|  |
| --- |
|  |
|  |

## Variation de l’attribut MonthlyCharges par churn et no churn

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

## Variation de l’attribut tenure par churn et no churn

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

## Variation de l’attribut dependents par churn et no churn

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

## Variation de l’attribut deviceProtection par churn et no churn

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

## Variation de l’attribut gender par churn et no churn

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

## Variation de l’attribut internetService par churn et no churn

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

## Variation de l’attribut multipleLines par churn et no churn

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

## Variation de l’attribut OnlineBackup par churn et no churn

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

## Variation de l’attribut OnlineSecurity par churn et no churn

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

## Variation de l’attribut partner par churn et no churn

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

## Variation de l’attribut PhoneService par churn et no churn

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

## Variation du seniorCitizen par churn et no churn

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

## Variation du streamingMovies par churn et no churn

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

## Variation du Streaming TV par churn et no churn

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

## Variation du support technique par churn et no churn

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

## Matrice de corrélation des attributs de type numériques

|  |
| --- |
|  |
|  |

# NETTOYAGE ET PRETRAITEMENT DES DONNEES

Dans cette étape nous avons effectué les opérations suivantes :

* Evaluation et nettoyage des champs

Notamment de l’attribut « TotalCharges » ayant 11 lignes non numériques.

* Mis ajours du type des champs numériques
* Remplacement pour les attributs « **OnlineSecurity** », « **OnlineBackup** », « **DeviceProtection** », « **TechSupport** », « **StreamingTV** », « **StreamingMovies** » les occurrences de « No internet service » par « No ».
* Remplacement pour l’attribut « **MultipleLines** », les occurrences de « No phone service » par « No », car le fait qu’il n’y ait pas de service téléphonique indique bien qu’il n’y a pas de ligne multiple.
* Suppression la colonne « CustomerID » qui n’a aucune valeur sémantique et ne représente les l’identifiant des clients.
* Suppression les doublons éventuels sur les données ainsi nettoyées.
* Création des transformateurs de donnée
* Construction le pipeline de transformateur
* Initialisation du pipeline et exécution de la transformation des données par le pipeline.

# SELECTION DES FEATURES

Cette partie il a été question de sélectionner les attributs encore appelés features, les plus significatifs dans le jeu de donnée, et pour cela nous avons effectué une sélection a priori basée sur un test de chi2, en retenant le nombre de features les plus significatifs (compte tenu de leurs p-value) qui maximisent la performance. Il nous a donc fallu une fonction objective qui calcule cette performance en fonction du nombre de features sélectionné, et nous avons retenu deux fonctions donnant deux indicateurs ; le premier indicateur est **AUCROC** et le deuxième est **AUCPR**. Nous avons utilisé le sélecteur « **ChiSqSelector »** de Spark en lui fournissant le nombre de features, pour chaque sélection faite nous avons évaluer la performance de la prédiction faite par un classificateur basé sur la régression logistique.

En tenant compte du déséquilibre des données nous avons pour notre cas choisi de tenir compte plutôt de l’indicateur AUCPR (Area Under Curve Précision Recall) car il permet de mieux tenir compte de la classe minoritaire dans l’évaluation des performances.

# ENTRAINEMENT DES MODELES

Apres la détermination du nombre de features, le sélecteur de features « **ChiSqSelector »** sera ajouter au pipeline de transformation les données, qui seront retraitées avec en sortie les données avec le nombre de features sélectionnés dans l’étape précédente. Ensuite nous avons créé les deux classificateurs l’un pour **la régression logistique** et l’autre pour le **Random Forest** que nous avons entrainé sur 70% des données puis valider avec 30% restant.

# VALIDATION DES MODELES

Apres l’entrainement du modèle, nous avons valider le modèle avec 30% des donnée et obtenue une prédiction pour chacun des classificateurs.

# EVALUATION DES PERFORMANCES

Pour évaluer les deux modèles nous avons utilisés un qui nous a permis d’avoir deux métrique, l’AUC ROC et l’AUCPR, Ainsi nous avons obtenus les résultats suivants :

|  |  |  |
| --- | --- | --- |
| Classificateur | areaUnderROC | areaUnderPr |
| Random Forest |  |  |
| Régression logistique |  |  |

# 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"**.