

# **RAPPORT DU PROJET DE BIG DATA**

**MASTER MIAGE PARCOURS INFORMATIQUE ET INNOVATION.**

**UE: DATA MINING et BIG DATA**

**ENSEIGNANTS:** Pr Antoine TABONNE.  
Pr Olivier PERRIN

*Rédigé par Alain NGUIDJOI BELL*

## SOMMAIRE

---

Introduction .....	3
1 Description de la méthodologies .....	3
2 CHARGEMENT DES DONNÉES.....	4
3 APERCUS DES DONNEES .....	4
4 EXPLORATION DES DONNEES .....	7
5 NETTOYAGE ET PRETRAITEMENT DES DONNEES .....	16
6 SELECTION DES FEATURES.....	17
7 ENTRAÎNEMENT DES MODELES .....	17
8 VALIDATION DES MODELES .....	18
9 EVALUATION DES PERFORMANCES .....	18
Conclusion .....	18

## INTRODUCTION

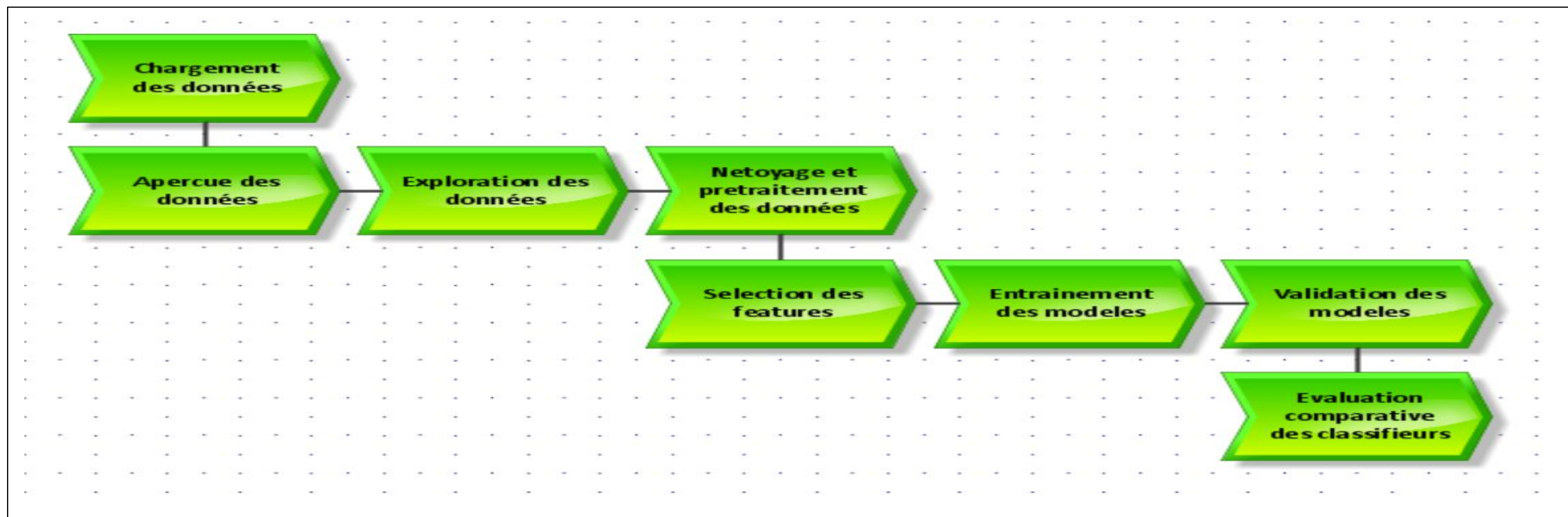
---

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 au logiciel ANACONDA, ensuite grâce aux bibliothèques seaborn et panda nous avons effectué une analyse exploratoire des données, des transformations des certaines données ont ensuite été effectuées, puis nous avons construit le pipeline de nos modèles grâce à Spark après entraînement 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 suivi 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 chargées grâce à Spark.

## APERÇUS DES DONNÉES

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

Info				
7043 instance(s)				
20 feature(s) (0.0% missing values)				
Data has no target variable.				
1 meta attribute(s)				

Columns (Double click to edit)				
	Name	Type	Role	Values
1	gender	 categorical	feature	1 female, Male
2	SeniorCitizen	 categorical	feature	0, 1
3	Partner	 categorical	feature	No, Yes
4	Dependents	 categorical	feature	No, Yes
5	tenure	 numeric	feature	
6	PhoneService	 categorical	feature	No, Yes
7	MultipleLines	 categorical	feature	No, No phone service, Yes
8	InternetService	 categorical	feature	DSL, Fiber optic, No
9	OnlineSecurity	 categorical	feature	No, No internet service, Yes
10	OnlineBackup	 categorical	feature	No, No internet service, Yes

Columns (Double click to edit)				
	Name	Type	Role	Values
11	DeviceProtection	<span>C</span> categorical	feature	No, No internet service, Yes
12	TechSupport	<span>C</span> categorical	feature	No, No internet service, Yes
13	StreamingTV	<span>C</span> categorical	feature	No, No internet service, Yes
14	StreamingMovies	<span>C</span> categorical	feature	No, No internet service, Yes
15	Contract	<span>C</span> categorical	feature	Month-to-month, One year, Two year
16	PaperlessBilling	<span>C</span> categorical	feature	No, Yes
17	PaymentMethod	<span>C</span> categorical	feature	Bank transfer (automatic), Credit card (automatic), Electronic check, Mailed check
18	MonthlyCharges	<span>N</span> numeric	feature	
19	TotalCharges	<span>N</span> numeric	feature	
20	Churn	<span>C</span> categorical	feature	No, Yes
21	customerID	<span>S</span> text	meta	

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.

	customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines	InternetService	OnlineSecurity
1	7590-VHVEG	Female	0	Yes	No	1	No	No phone service	DSL	No
2	5575-GNVDE	Male	0	No	No	34	Yes	No	DSL	Yes
3	3668-QPYBK	Male	0	No	No	2	Yes	No	DSL	Yes
4	7795-CFOCW	Male	0	No	No	45	No	No phone service	DSL	Yes
5	9237-HQITU	Female	0	No	No	2	Yes	No	Fiber optic	No
6	9305-CDSKC	Female	0	No	No	8	Yes	Yes	Fiber optic	No
7	1452-KIOVK	Male	0	No	Yes	22	Yes	Yes	Fiber optic	No
8	6713-OKOMC	Female	0	No	No	10	No	No phone service	DSL	Yes
9	7892-POOKP	Female	0	Yes	No	28	Yes	Yes	Fiber optic	No
10	6388-TABGU	Male	0	No	Yes	62	Yes	No	DSL	Yes
11	9763-GRSKD	Male	0	Yes	Yes	13	Yes	No	DSL	Yes
12	7469-LKBCI	Male	0	No	No	16	Yes	No	No	No internet service
13	8091-TTVAX	Male	0	Yes	No	58	Yes	Yes	Fiber optic	No
14	0280-XJGEX	Male	0	No	No	49	Yes	Yes	Fiber optic	No
15	5129-JLPIS	Male	0	No	No	25	Yes	No	Fiber optic	Yes
16	3655-SNQYZ	Female	0	Yes	Yes	69	Yes	Yes	Fiber optic	Yes
17	8191-XWSZG	Female	0	No	No	52	Yes	No	No	No internet service
18	9959-WOFKT	Male	0	No	Yes	71	Yes	Yes	Fiber optic	Yes
19	4190-MFLUW	Female	0	Yes	Yes	10	Yes	No	DSL	No
20	4183-MYFRB	Female	0	No	No	21	Yes	No	Fiber optic	No

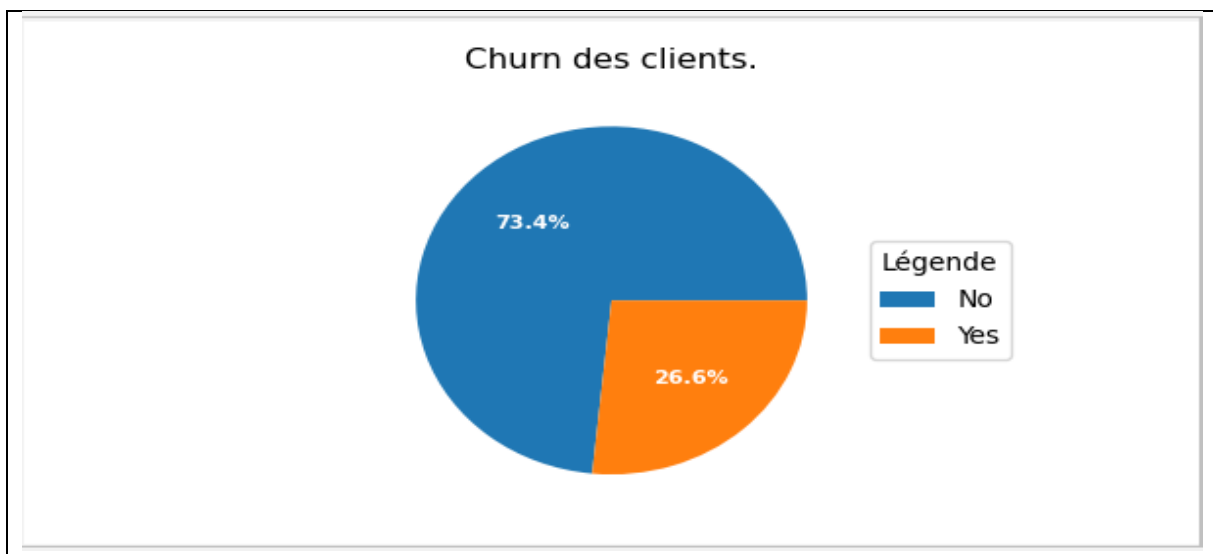
## EXPLORATION DES DONNEES

---

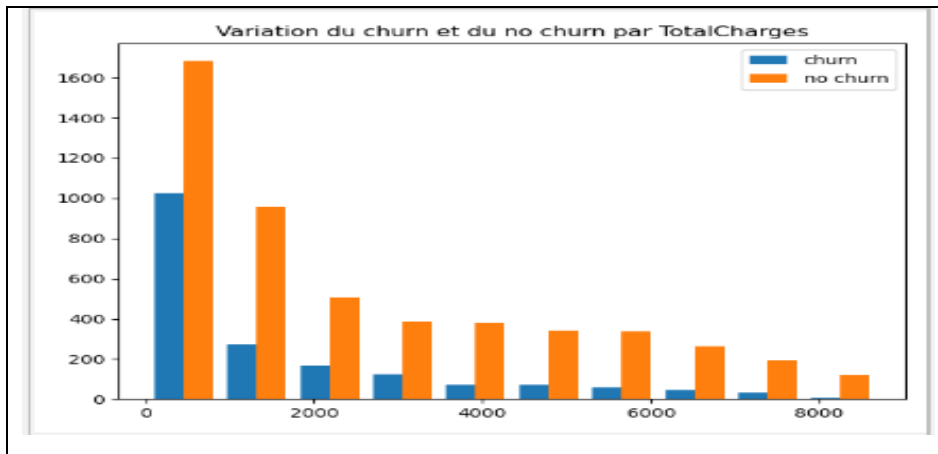
L'exploration des données vise à analyser et visualiser les données de manière à en ressortir les rapports entre les différents éléments du schéma des données, d'en ressortir les proportions relatives et absolues qui pourront éventuellement nous permettre d'orienter notre prétraitement en tenant compte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

Dans cette perspective, un dessin valant mieux que mille mots, nous avons établi 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 établi 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).

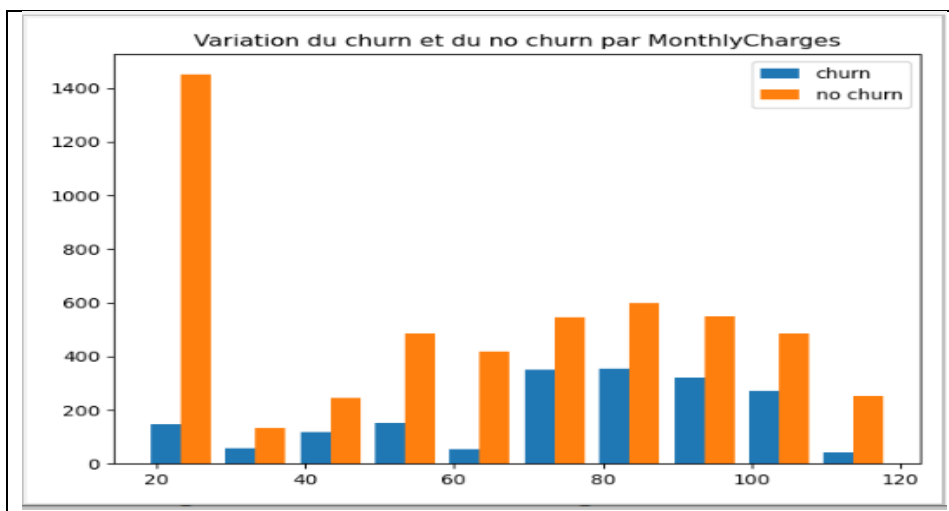
### 1.1 PROPORTION DES CHURN ET DES NO CHURN



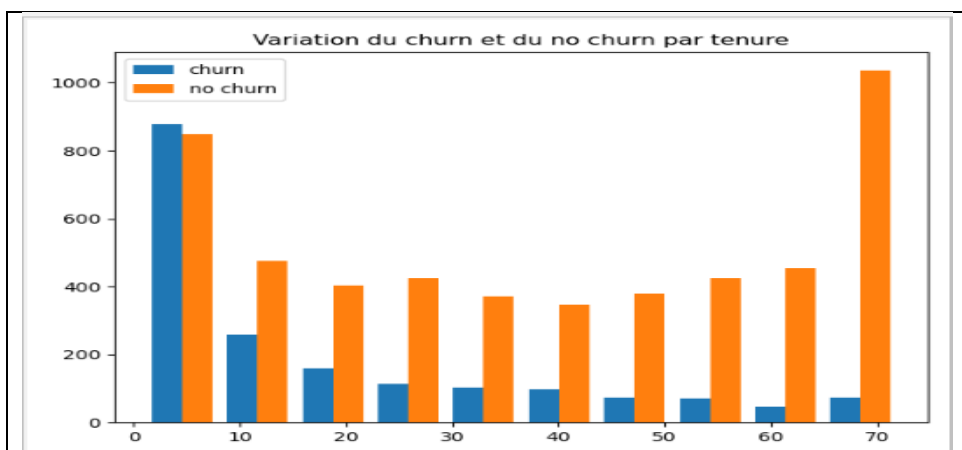
## 1.2 VARIATION DE L'ATTRIBUT TOTALCHARGES PAR CHURN ET NO CHURN



## 1.3 VARIATION DE L'ATTRIBUT MONTHLYCHARGES PAR CHURN ET NO CHURN



## 1.4 VARIATION DE L'ATTRIBUT TENURE PAR CHURN ET NO CHURN

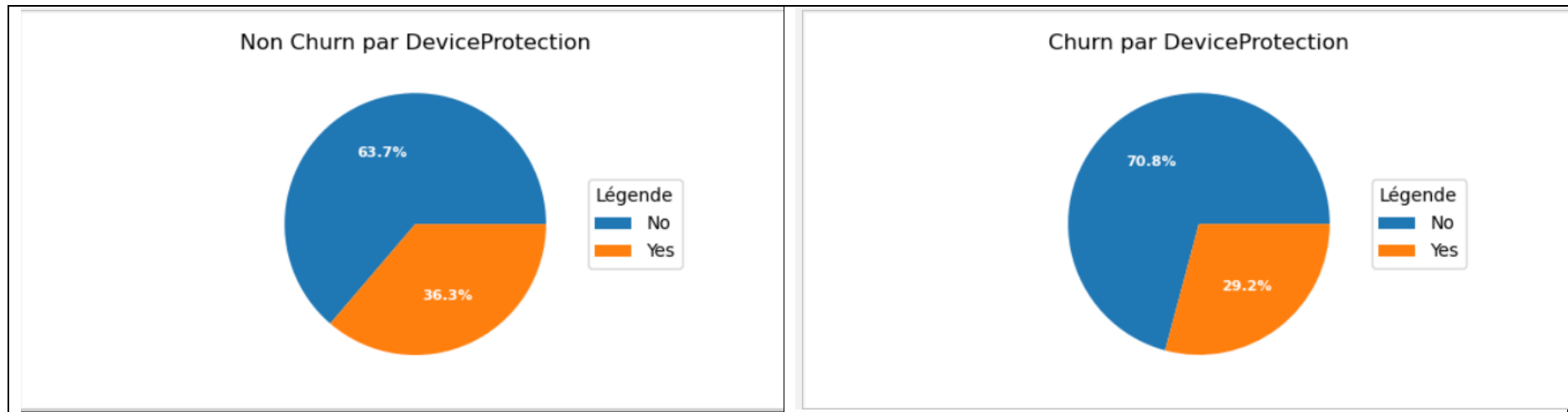




## 1.5 VARIATION DE L'ATTRIBUT DEPENDENTS PAR CHURN ET NO CHURN



## 1.6 VARIATION DE L'ATTRIBUT DEVICEPROTECTION PAR CHURN ET NO CHURN



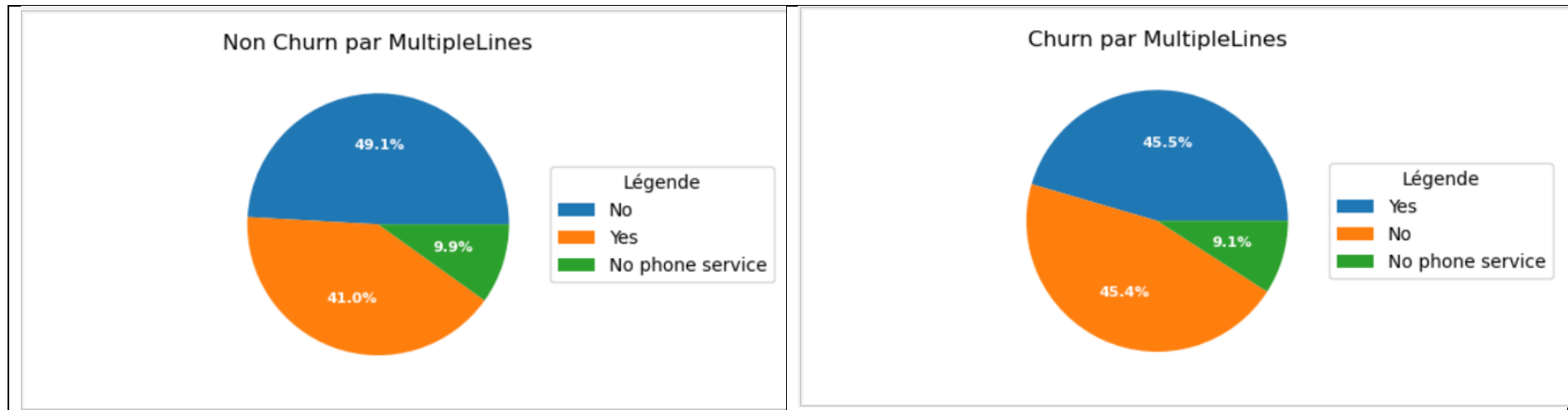
## 1.7 VARIATION DE L'ATTRIBUT GENDER PAR CHURN ET NO CHURN



## 1.8 VARIATION DE L'ATTRIBUT INTERNETSERVICE PAR CHURN ET NO CHURN



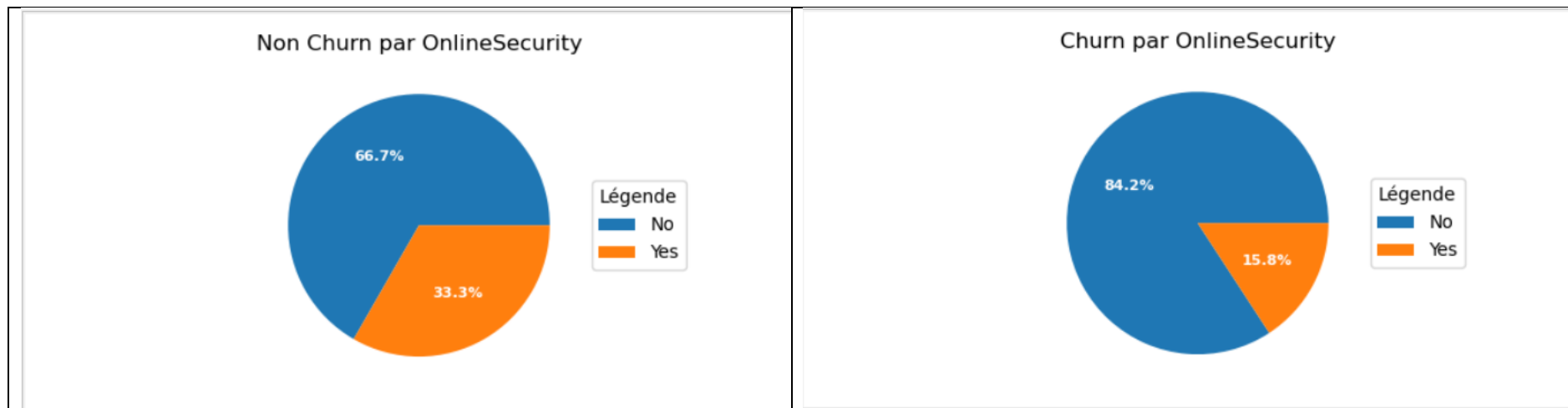
## 1.9 VARIATION DE L'ATTRIBUT MULTIPLELINES PAR CHURN ET NO CHURN



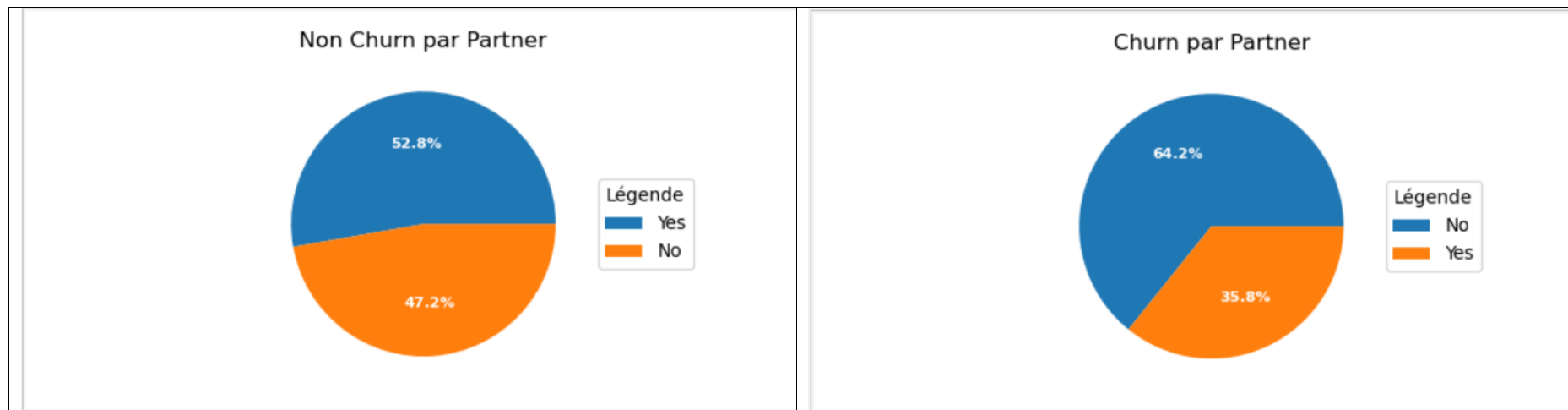
### 1.10 VARIATION DE L'ATTRIBUT ONLINEBACKUP PAR CHURN ET NO CHURN



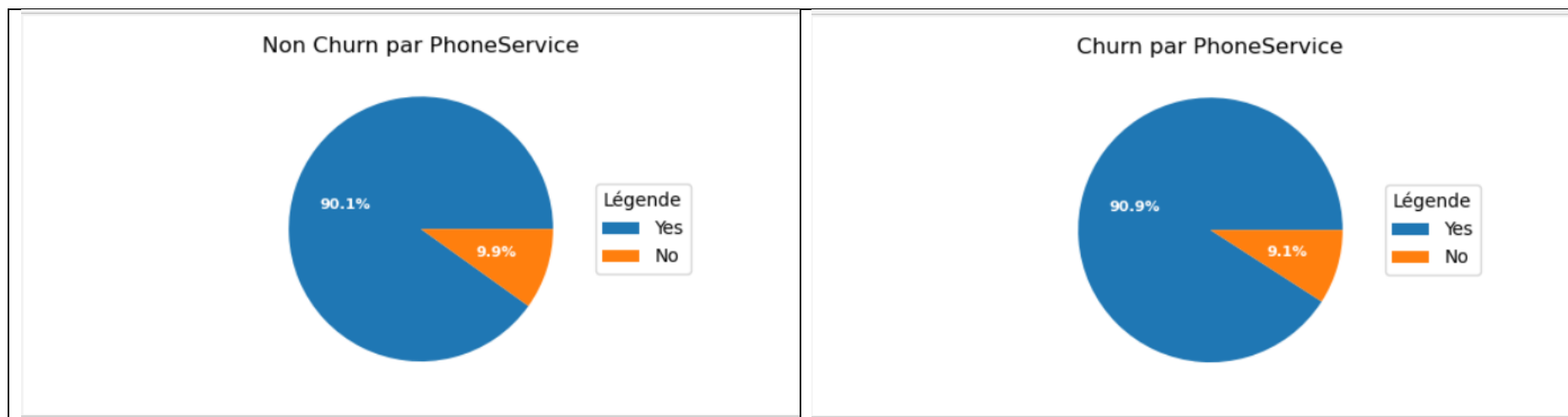
### 1.11 VARIATION DE L'ATTRIBUT ONLINESECURITY PAR CHURN ET NO CHURN



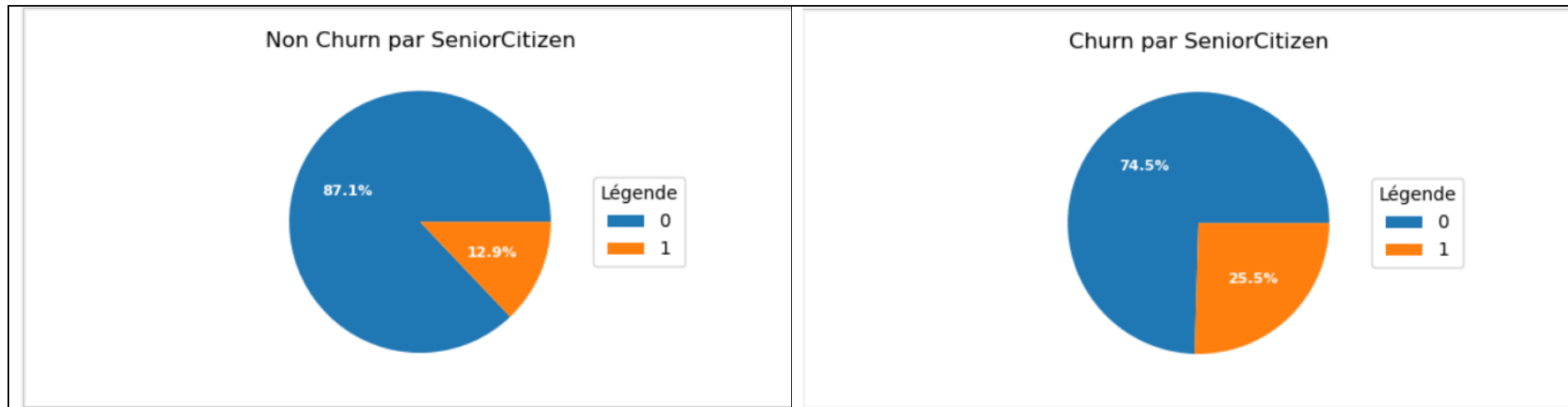
### 1.12 VARIATION DE L'ATTRIBUT PARTNER PAR CHURN ET NO CHURN



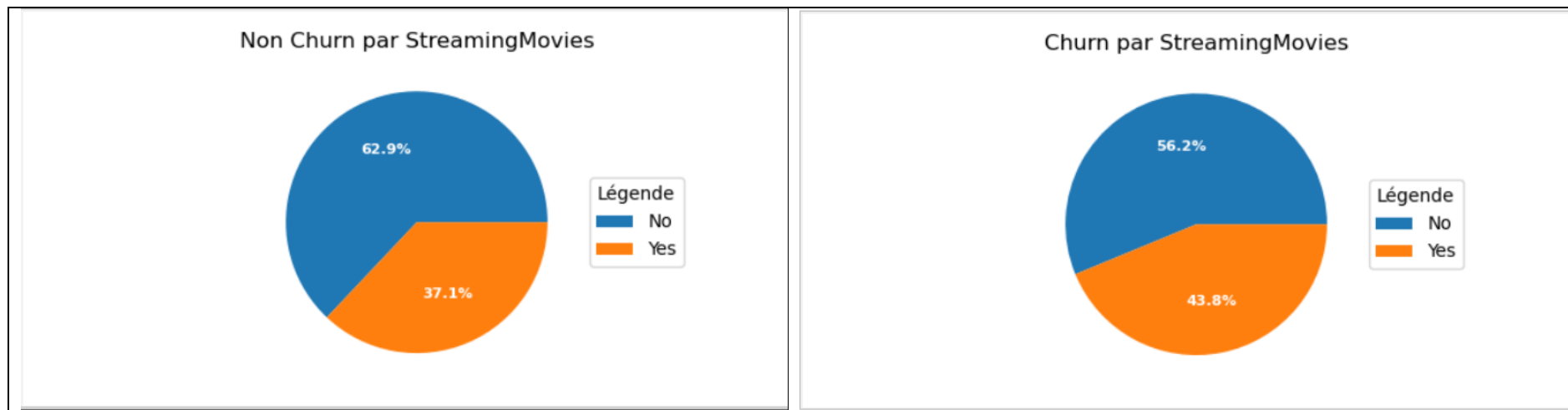
### 1.13 VARIATION DE L'ATTRIBUT PHONESERVICE PAR CHURN ET NO CHURN



### 1.14 VARIATION DU SENIORCITIZEN PAR CHURN ET NO CHURN



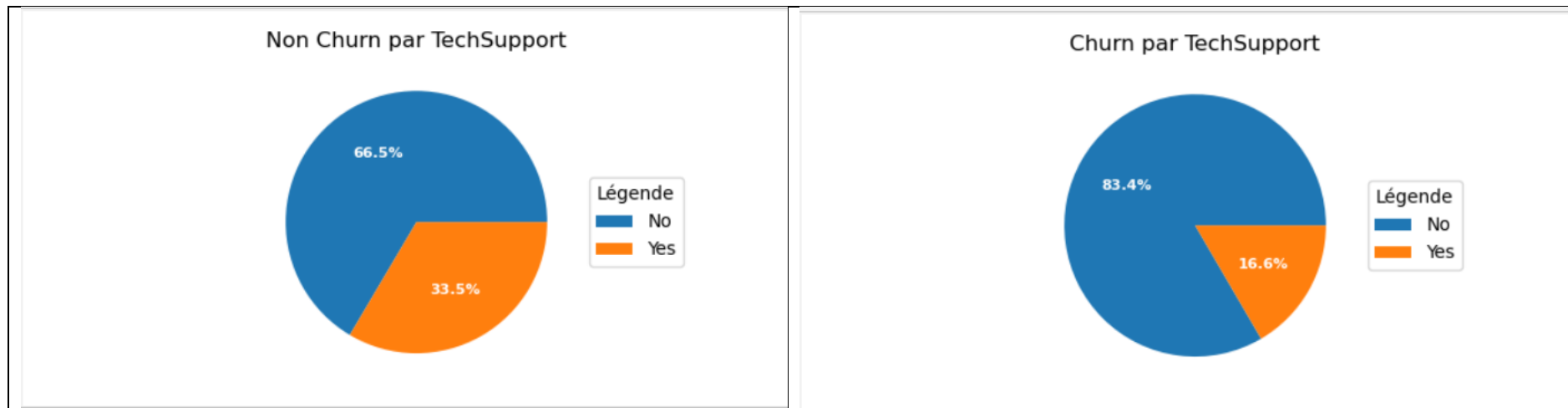
### 1.15 VARIATION DU STREAMINGMOVIES PAR CHURN ET NO CHURN



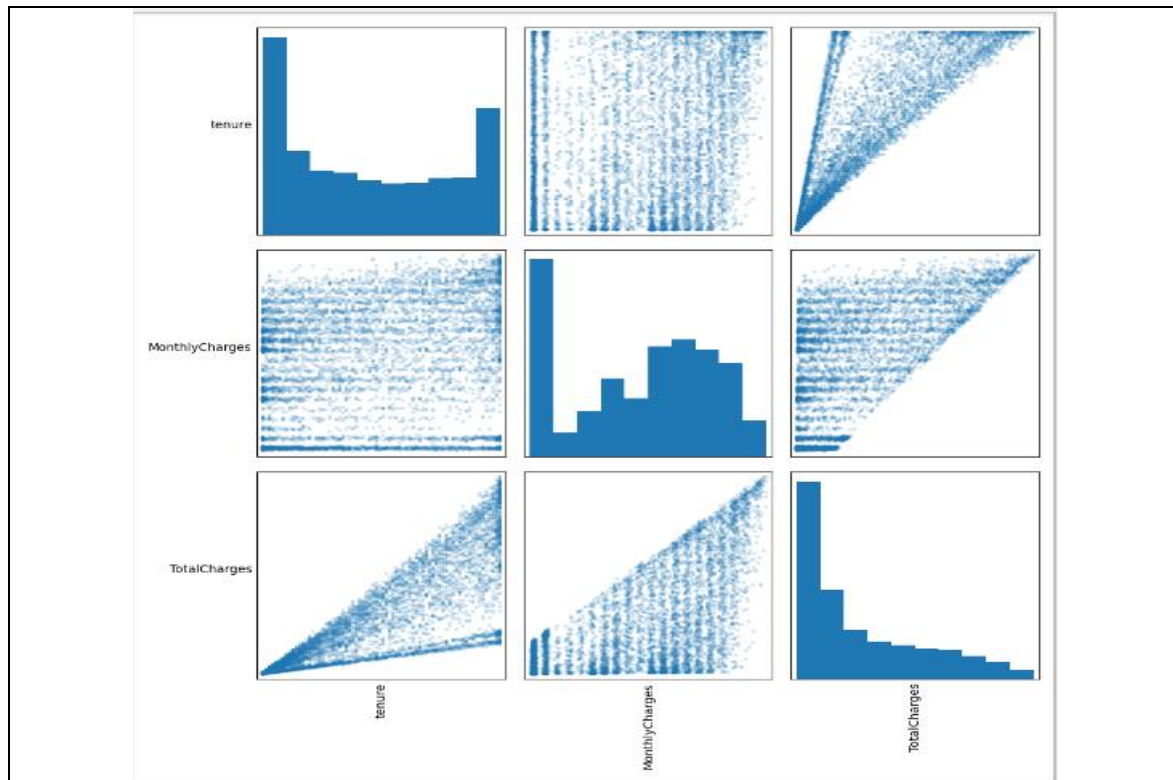
### 1.16 VARIATION DU STREAMING TV PAR CHURN ET NO CHURN



### 1.17 VARIATION DU SUPPORT TECHNIQUE PAR CHURN ET NO CHURN



## 1.18 MATRICE DE CORRELATION DES ATTRIBUTS DE TYPE NUMERIQUES



## 1.19 ANALYSE DES VARIATIONS DES ATTRIBUTS DU JEU DE DONNEE

La première figure présente un déséquilibre important dans la distribution des clients, avec 73,4% de donnée correspondants à des clients qui n'effectuent pas de Churn contre 24,6% qui le réalise. Ce déséquilibre des données peut affecter grandement le modèle issu des différents entraînements de classificateurs, notamment dans la prédiction, il est possible que des clients effectuant le Churn réellement soit classifiés comme ne le faisant pas, il sera donc important lors de l'évaluation des classifieurs de tenir compte de ce déséquilibre dans les données notamment dans le choix de l'estimateur.

Les différentes variations des attributs présents dans les données ne permettent pas de remarquer clairement une Independence entre les attributs et l'attribut cible (label), de manière à supprimer les attributs ne dépendants pas du label, à l'exception de l'attribut phoneService pour lequel plus de 90% des données sont pour des clients ayant un PhoneService. Ainsi pour sélectionner les features nous avons effectué une analyse a posteriori avec une sélection des features les plus significatifs basés sur un test de chi2 et une évaluation des performances en déterminants le nombre de features qui maximisent l'indicateur de performance **AUCPR** (areaUnderPrecisionRecall) choisi grâce à sa capacité à mieux représenter les performances compte tenu du jeu de donné déséquilibré.

## 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 à jours 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.

Le fichier source « **dataPreprocessing.py** » comporte le code python de la démarche de preprocessing et de nettoyage .

## 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  $\chi^2$ , 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 évalué 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 Precision Recall) car il permet de mieux tenir compte de la classe minoritaire dans l'évaluation des performances.

Le fichier source « **FeatureSelection.py** » comporte le code python de la démarche de sélection.

## ENTRAINEMENT DES MODELES

---

Après 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 un classificateur pour la **régression logistique** et un autre pour le **Random Forest** que nous avons entraînés sur 70% des données puis validé avec 30% des données restant.

## VALIDATION DES MODELES

---

Après l'entraînement du modèle, nous avons validé le modèle avec 30% des données et obtenu une prédiction pour chacun des classificateurs.

## EVALUATION DES PERFORMANCES

---

Pour évaluer les deux modèles nous avons utilisé deux indicateurs de performance l'un basé sur la Précision et le Recall (PR) et l'autre basé sur le ROC, les courbes de ces deux définissent avec l'axe des abscisses une aire appelée AUC ROC et AUCPR. Spark grâce à la classe « BinaryClassificationEvaluator » nous a permis d'avoir ces deux métriques sur la prédiction faite par chacun des classificateurs utilisés.

Les résultats obtenus sont mentionnés dans le tableau suivant :

Classificateur	AreUnderRoc	AreaUnderPR	Accuracy	Erreur	Matrice de confusion
Random Forest	0.82564381	0.63640964	0.82564381	0.17435619	[1436. 111.] [328. 244.]
Régression logistique	0.82330165	0.65298081	0.82330165	0.17669835	[1382. 107.] [320. 253.]

Les métriques de performance telles que le **ROC** et ses corollaires **AUC** et **areUnderRoc** basé sur le taux de TPR (True Positive Rate) semble tendre vers 1 avec un faible taux d'erreur ce qui peut sembler démontrer que nos deux classifieurs ont une bonne performance, or on observe cependant que la métrique basée sur la **Précision** et le **Recall** à savoir l'**areaUnderPR** tend beaucoup plus vers 0,5 que vers 1 ce qui semble démontrer une performance pas bonne pour nos deux classifieurs, cette différence s'explique lorsqu'on observe les matrices de confusion respective l'on observe que nos deux classifieurs ont un nombre de faux négatifs très importants, il y a donc là aussi un déséquilibre, que la métrique basée sur la Précision et le Recall semble faire transparaître.

Ainsi on peut dire que les métriques fondées sur le ROC et son AUC ne sont pas porteuses de la performance lorsque les données sont déséquilibrées et que celle basée sur le PR et le Recall sont plus adaptées car elles sont plus sensibles au déséquilibre des données.

Ainsi en se basant sur la précédente observation on peut dire que la régression logistique est quelque peu plus performante que le Random Forest, même si les deux classifieurs ont des performances très proches.

## CONCLUSION

---

Nous avons pu parcourir de façon contrainte l'ensemble des activités nécessaires au traitement des données en big data, notamment de l'importance de la phase de prétraitement, nous avons aussi perçu l'impact du déséquilibre des données sur les classifieurs et la nécessité de mettre en place des mesures pour pallier au déséquilibre des données nous semble intéressante.