

Projet Machine Learning

Article 1: A Customer Churn Prediction using Pearson Correlation Function and K Nearest Neighbor Algorithm for Telecommunication Industry

Article 2: Sequential Feature Selection in Customer Churn Prediction Based on Naive Bayes

Article 3: Customer Churn Prediction, Segmentation and Fraud Detection in Telecommunication Industry

Travail réalisé par : Ayadi Yassine - Bhar Hane - Dahmen Omar
Loukil Eya - Soffelgil Nour

Table des matières

I-Introduction.....	5
II- Organigramme du modèle de prédition du client	5
III- La méthodologie utilisée	5
IV- Etapes de travail :	6
1- Compréhension du problème métier.....	6
2- Compréhension des données	6
a- Articles 1 et 2:.....	6
b- Article 3.....	12
3- Préparation des données.....	19
a- Article 1-2.....	19
b- Article 3.....	24
4-Modélisation:.....	24
a-Article 1.....	24
b-Article 2	25
c-Article 3 :.....	28
5-Evaluation :.....	28
a- Article 1.....	28
b- Article 2	32
c-Article3 :.....	34
VIII - Outils de travail.....	38
IX- Conclusion	40

Table des figures :

Figure 1: Organigramme du modèle de prédictions du client	5
Figure 2: Cycle de méthodologie CRISP	6
Figure 3 : Liste des attributs	7
Figure 4 : Affichage des 5 premières lignes	7
Figure 5: Customer Churn en pourcentage	8
Figure 6: Gender en fonction de Churn	8
Figure 7: Observation des données uniques	9
Figure 8 : Observation de la tenure des abonnés	9
Figure 9: Observation de la tenure en fonction de Churn	10
Figure 10 : Boîte à moustaches	10
Figure 11: Visualisation globale des données.....	11
Figure 12 : Visualisation des tenures groups triées	11
Figure 13 : Représentation graphique des variables numériques	12
Figure 14 : « Churn » en fonction du « gender »	13
Figure 15 : Modalités des variables	13
Figure 16 : Carte de Californie en fonction du Churn	14
Figure 17: Villes de Californie en fonction du Churn	15
Figure 18 : Nombre de « Churn_reason »	15
Figure 19 : Contrat en fontion du Churn	16
Figure 20 : Tech support en fonction churn	16
Figure 21: Tenure en fonction de churn	17
Figure 22: Churn score en fonction du churn	17
Figure 23 : Paiement en fonction du churn	18
Figure 24 : Valeurs aberrantes	18
Figure 25 : cinq premières lignes des attributs centrés réduits	19
Figure 26 : Matrice de corrélation des variables numériques	19
Figure 27 : Les différentes modalités des variables catégoriel.les	20
Figure 28 : Les variables catégorielles encodées	20
Figure 29 : Matrice de corrélation des attributs	21
Figure 30 : Densité de "Tenure"	22
Figure 31: Droite d'Henry de "Tenure"	22
Figure 32 : Densité de "Monthly Charges"	23
Figure 33 : Droite d'Henry de "Monthly Charges"	23
Figure 34 : Corrélation avec la variable cible	24
Figure 35 : Choix de la valeur k et la métrique	25
Figure 36 : Courbe Roc Naives Bayes	26
Figure 37 : f1 score	26
Figure 38 : Matrice de confusion	27
Figure 39 : Nombre de clusters	28
Figure 40 : Matrice de confusion KNN	29
Figure 41 : Matrice de confusion Random Forrest	29

Figure 42 : Matrice de confusion Logistic Regression	30
Figure 43 : Matrice de confusion SVM	30
Figure 44 : Tables des scores	31
Figure 45 : Courbe Roc Article1	31
Figure 46 : Tableau modèle de performance	32
Figure 47 : Accuracy des modèles	33
Figure 48 : AUC des modèles	33
Figure 49 : Ordonnancement de « roc_auc »	34
Figure 50 : Figure 50: Ordonnancement de « Accuracy »	34
Figure 51 : f1 score article 3	34
Figure 52 : Matrice de confusion arbre de décision	35
Figure 53 : Affichage de 5 premières lignes	35
Figure 54 : Représentation 3D des clusters	36
Figure 55 : La taille de chaque cluster 1	36
Figure 56: La taille de chaque cluster 2	37
Figure 57 : Tenure months en fonction de monthly charges	37
Figure 58 : Cluster 0	37
Figure 59 : Cluster 1	38
Figure 60 : Cluster 2	38

I-Introduction

Le désabonnement des clients est l'une des mesures les plus importantes à évaluer pour une entreprise en croissance. Bien que ce ne soit pas la mesure la plus favorable, c'est un chiffre qui peut donner à votre entreprise la dure vérité sur sa rétention de la clientèle. Et vu que ce sujet est à la pointe de l'actualité on a été demandé de comprendre et expliquer trois différents articles qui traitent ce phénomène.

II- Organigramme du modèle de prédition du client

On dit très souvent que le client est le roi, exceptionnellement dans le domaine de l'industrie de télécommunication. Le satisfaire c'est une priorité, et ne pas le perdre c'est un objectif. Ceci parce que reconvertis un client est de 5 à 6 fois moins cher d'avoir un nouveau client. La science de données et spécialement la Machine Learning est un bon moyen pour prédire un client de la compagnie de télécommunication. Notre objectif c'est de ne pas effectuer une perte de dépenses extraordinaires de fidélisation et de ré-acquisition des clients. On peut aussi épargner les coûts de publicité, de planification et de budgétisation.

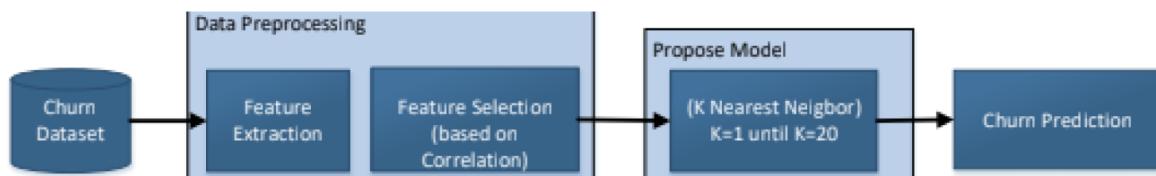


Figure 1 : Organigramme du modèle de prédictions du client

III- La méthodologie utilisée

En Machine Learning, suivre une méthodologie de travail est obligatoire. Dans notre cas, on va appliquer celle de **CRISP**. Cette dernière se compose de 6 étapes :

- 1- La compréhension du problème métier
- 2- La compréhension des données
- 3- La préparation des données (Classement des données, nettoyage des données, recodage des données)
- 4- La modélisation (Le choix, le paramétrage, le test, l'enchainement des algorithmes)
- 5- L'évaluation
- 6- Le déploiement

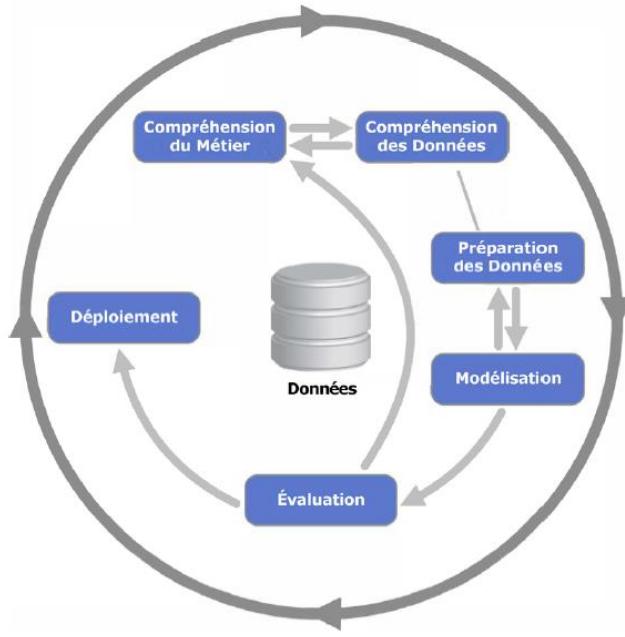


Figure 2 : Cycle de méthodologie CRISP

IV- Etapes de travail :

1- Compréhension du problème métier

Pour comprendre le métier en question, il nous était indispensable de lire les trois articles, décrypter tous leurs détails et effectuer des recherches sur internet. Dans notre cas, c'est l'industrie de télécommunication qui est en question.

- On a souligné les mots clés des articles étudiés :

Churn prediction = prédiction du désabonnement

-On a compris que la perte des abonnés ou de la clientèle est directement lié à la satisfaction du client. Il est évident que le coût de l'acquisition d'un client est beaucoup plus élevé que le coût de la fidélisation d'un client, ce qui fait de la rétention des clients un prototype compréhensif. Il n'existe pas de modèle standard permettant de résoudre avec précision les problèmes de clientèle dans ce domaine.

-On a constaté que la machine Learning se révèle être un moyen efficace d'identifier et prédire le désabonnement des clients. On peut anticiper leur rupture avant sa production.

2- Compréhension des données

a- Articles 1 et 2 :

La première étape consistait essentiellement d'avoir une vue globale de tous les attributs présents dans le Dataset (taille 1.1MB) ainsi que leurs types et descriptions :

Attributes	Type	Description
customerID	object	Unique number to represent customer
gender	object	Customer gender
SeniorCitizen	int64	Customer status based on age
Partner	object	Customer status based on partner
Dependents	object	Customer status based on dependency
tenure	int64	Customer period on using services from telco company
PhoneService	object	Status of having phone service
MultipleLines	object	Status of having multiple lines
InternetService	object	Status of having internet service
OnlineSecurity	object	Status of having online security service
OnlineBackup	object	Status of having online backup service
DeviceProtection	object	Status of having device protection service
TechSupport	object	Status of having technical support service
StreamingTV	object	Status of having streaming tv service
StreamingMovies	object	Status of having streaming movies service
Contract	object	Status of having contract
PaperlessBilling	object	Status of billing
PaymentMethod	object	Method of payment by customer
MonthlyCharges	float64	Customer monthly charges
TotalCharges	object	Customer total charges

Figure 3 : Liste des attributs

- On a 21 attributs dont 18 sont de type « Object », 2 sont de type « int64 » et 1 est de type « float64 ».

-Suite à l'affichage des premières lignes du Dataset on a pu identifier notre variable cible (Churn) ayant des valeurs possibles :

Yes pour les clients qui sont encore abonnés dans la compagnie

No pour les clients qui ne sont plus abonnés dans la compagnie

	customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines	InternetService	OnlineSecurity	... DeviceProtection	1
0	7590-VHVEG	Female	0	Yes	No	1	No	No phone service	DSL	No	...	No
1	5575-GNVDE	Male	0	No	No	34	Yes	No	DSL	Yes	...	Yes
2	3668-QPYBK	Male	0	No	No	2	Yes	No	DSL	Yes	...	No
3	7795-CFOCW	Male	0	No	No	45	No	No phone service	DSL	Yes	...	Yes
4	9237-HQITU	Female	0	No	No	2	Yes	No	Fiber optic	No	...	No

TechSupport	StreamingTV	StreamingMovies	Contract	PaperlessBilling	PaymentMethod	MonthlyCharges	TotalCharges	Churn
No	No	No	Month-to-month	Yes	Electronic check	29.85	29.85	No
No	No	No	One year	No	Mailed check	56.95	1889.5	No
No	No	No	Month-to-month	Yes	Mailed check	53.85	108.15	Yes
Yes	No	No	One year	No	Bank transfer (automatic)	42.30	1840.75	No
No	No	No	Month-to-month	Yes	Electronic check	70.70	151.65	Yes

Figure 4 : Affichage des 5 premières lignes

Un résultat en pourcentage de la prédiction du désabonnement des clients dans la société :

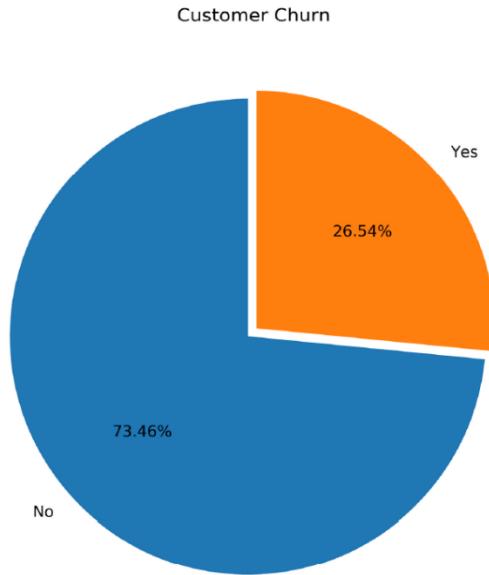


Figure 5 : Customer Churn en pourcentage

-73.46 % des clients sont restés dans l'entreprise .

-26.54 % des clients ont quitté l'entreprise .

Un histogramme en pourcentage de la prédiction du désabonnement des clients par sexe :

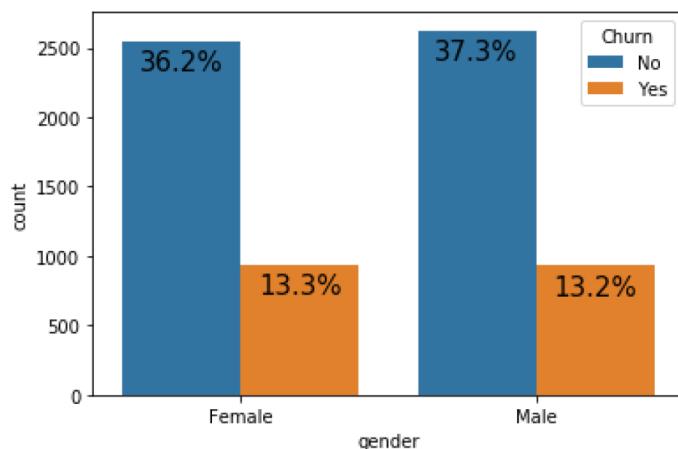


Figure 6 : Gender en fonction de Churn

-36.2% des femmes churn

-37.3% des hommes churn

-13.3% des femmes non churn

-13.2% des hommes non churn

On a vérifié les éléments uniques. On a également affiché les attributs catégoriaux de notre Dataset

```
customerID      7043
gender          2
SeniorCitizen   2
Partner         2
Dependents     2
tenure          73
PhoneService    2
MultipleLines   3
InternetService 3
OnlineSecurity  3
OnlineBackup    3
DeviceProtection 3
TechSupport     3
StreamingTV     3
StreamingMovies 3
Contract        3
PaperlessBilling 2
PaymentMethod   4
MonthlyCharges  1585
TotalCharges    6531
Churn           2
dtype: int64
[ 'SeniorCitizen',
  'Partner',
  'Dependents',
  'PhoneService',
  'MultipleLines',
  'InternetService',
  'OnlineSecurity',
  'OnlineBackup',
  'DeviceProtection',
  'TechSupport',
  'StreamingTV',
  'StreamingMovies',
  'Contract',
  'PaperlessBilling',
  'PaymentMethod']
```

Figure 7: Observation des données uniques

On a divisé la variable Tenure en des groupes d'années pour mieux comprendre nos données :

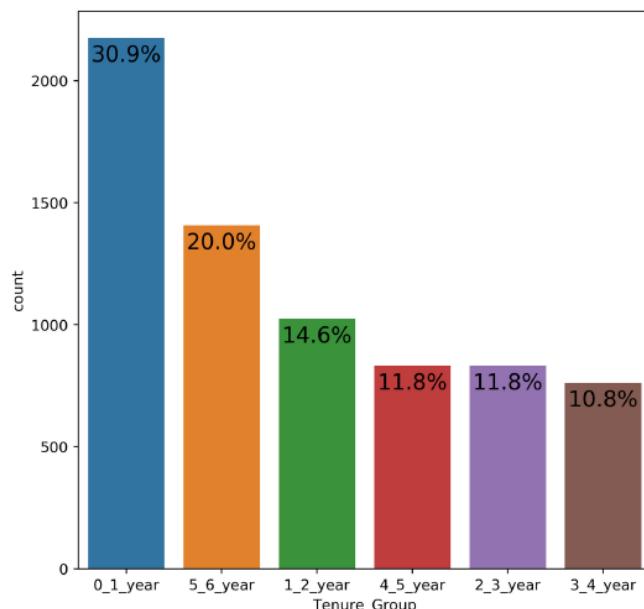


Figure 8 : Observation de la tenure des abonnés

On constate que la majorité des clients appartient au groupe de 12 mois (30.9%) et (20%) des clients ont une valeur de Tenure groupe entre 5 et 6 ans.

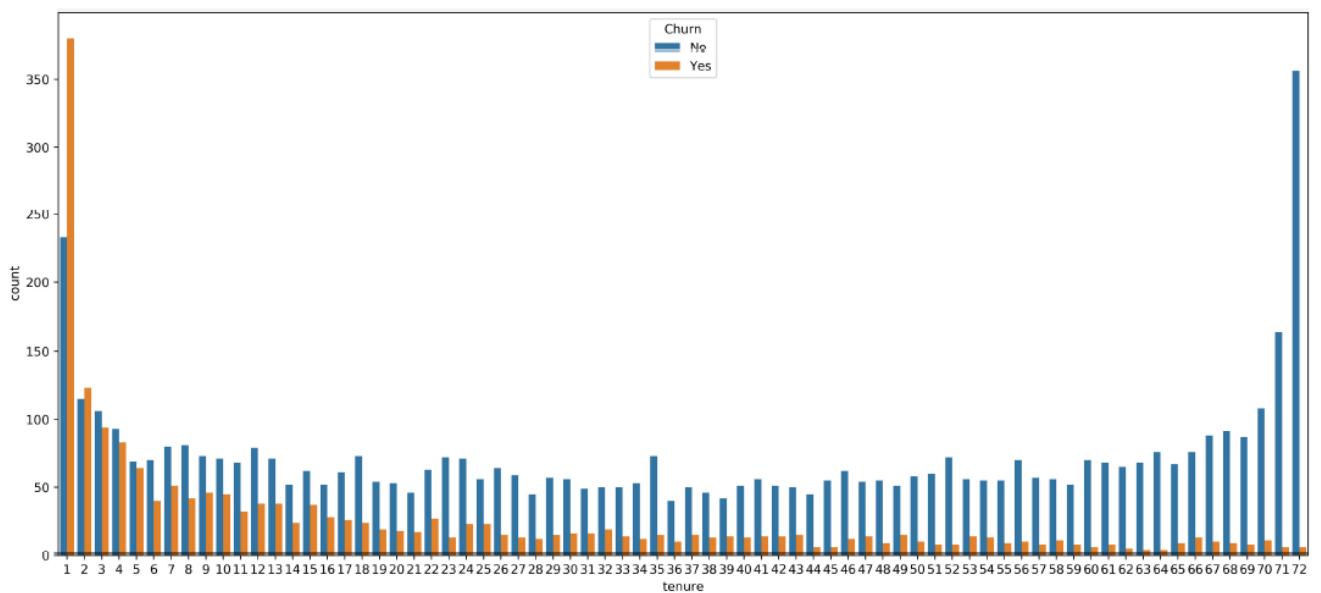


Figure 9 : Observation de la tenure en fonction de Churn

On constate d'après la figure 9 que la possibilité que le client quitte est élevée pour les valeurs de tenure moins importantes alors que la possibilité que le client reste est élevée pour les valeurs de tenures plus importantes.

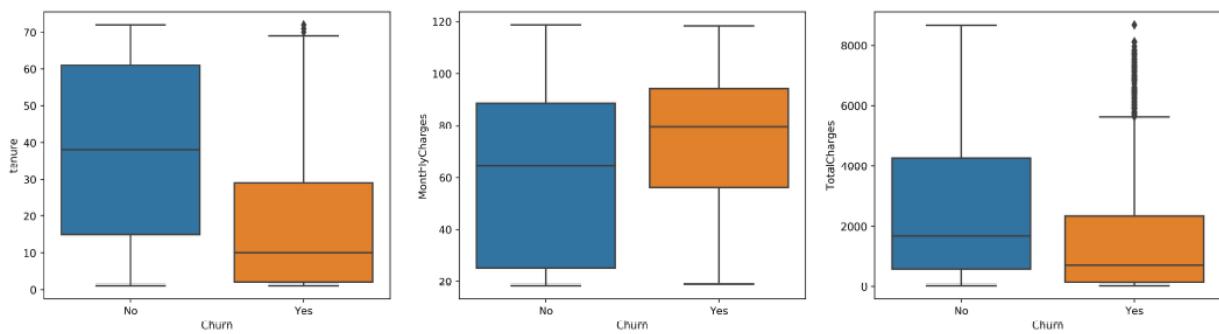


Figure 10 : Boîte à moustaches

Les boîtes à moustaches sont utiles car elles fournissent un résumé visuel des données permettant aux chercheurs d'identifier rapidement les valeurs moyennes, la dispersion de l'ensemble de données et les signes d'asymétrie. On constate l'apparition des points aberrants (outliers) pour les variables « tenure » et « totalcharges » .

Toujours dans le cadre de la préparation des données, on a ordonné les tenures groupes selon la période. Ceci nous est efficace pour comprendre nos données et constater d'où vient l'anomalie.

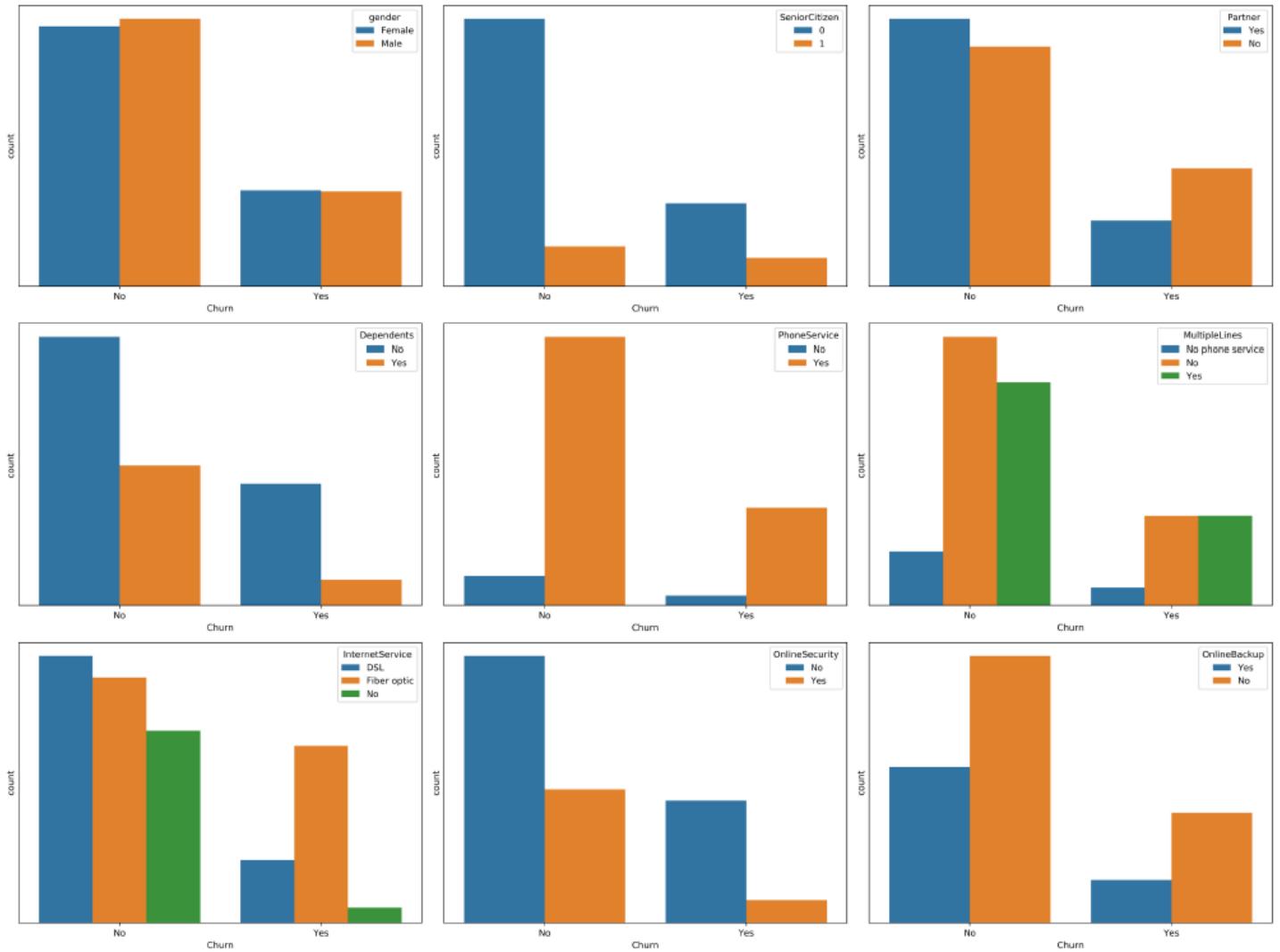


Figure 11: Visualisation globale des données

Cette figure globale nous aide à voir comment des différents facteurs affectent la décision des clients de rester ou de quitter l'entreprise.

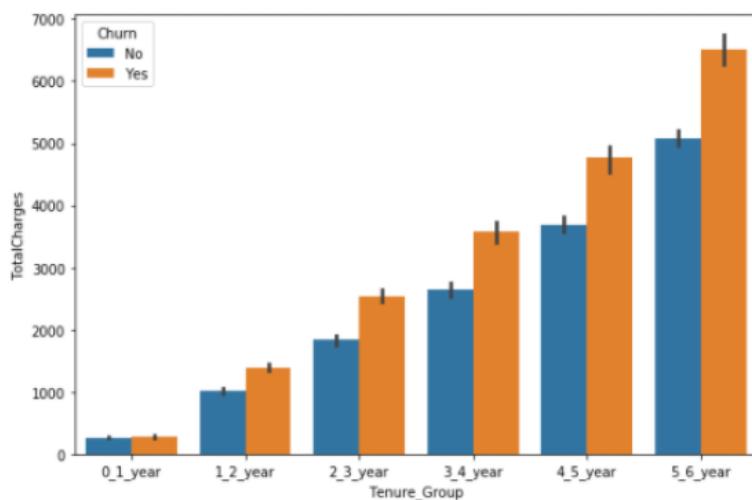


Figure 12 : Visualisation des tenures groups triées

On distingue la période et les dépenses augmentent proportionnellement.

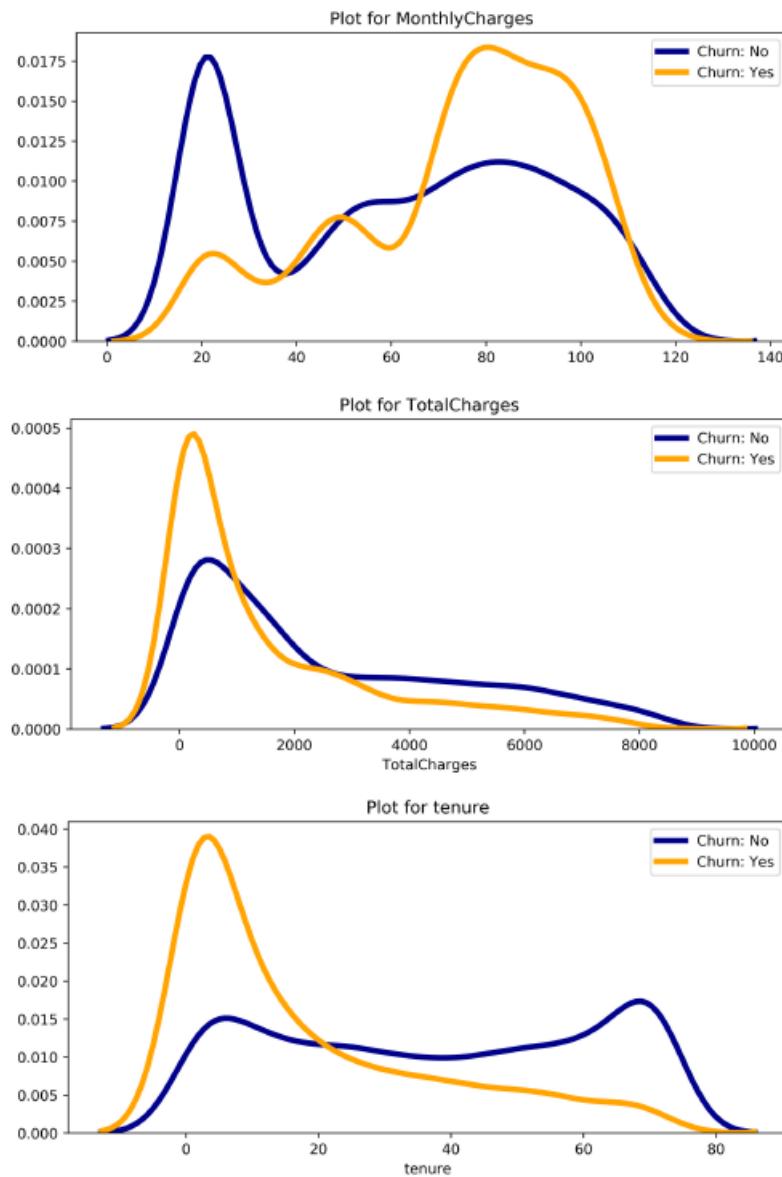


Figure 13 : Représentation graphique des variables numériques

À partir des graphiques ci-dessus, nous pouvons conclure que:

- Les utilisateurs récents sont plus susceptibles de quitter .
- Les utilisateurs avec des charges mensuelles plus élevées sont également plus susceptibles de quitter.

b- Article 3

Afin de réaliser l'étude sur cet article on dispose d'une Dataset (de taille --)ayant plus de variables que celle des deux premiers articles.

Après avoir consulter les différents variables et leurs différentes valeurs nous avons constaté que les variables suivantes sont dispensables pour la prédition :

- Count (toujours égale à 1).
- State (toujours égale à United States).

-On a vérifié les données manquantes dans la colonne « Churn Reason », on les a remplacées avec la valeur la plus fréquente.

-Pour la colonne « totales charges » on commençait par une consultation des cases vides , puis on a remplacé les 11 espaces trouvées de valeurs manquantes par nan, ensuite on a supprimé les valeurs manquantes, enfin on a changé le type de la variable TotalCharges de « object » à « float64 »

-Ci-dessous un histogramme décrivant le churn value en fonction du sexe :

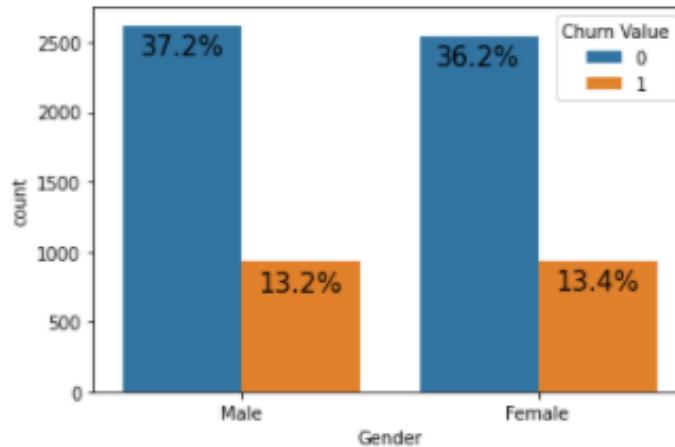


Figure 14 : « Churn » en fonction du « gender »

-On a sélectionné ensuite les variables catégorielles tout en considérant « No internet Service » comme « No » pour éviter toute redondance.

```
City : ['Los Angeles' 'Beverly Hills' 'Huntington Park' ... 'Tulelake'
    'Olympic Valley' 'Redcrest']
Gender : ['Male' 'Female']
Senior Citizen : ['No' 'Yes']
Partner : ['No' 'Yes']
Dependents : ['No' 'Yes']
Phone Service : ['Yes' 'No']
Multiple Lines : ['No' 'Yes']
Internet Service : ['DSL' 'Fiber optic' 'No']
Online Security : ['Yes' 'No']
Online Backup : ['Yes' 'No']
Device Protection : ['No' 'Yes']
Tech Support : ['No' 'Yes']
Streaming TV : ['No' 'Yes']
Streaming Movies : ['No' 'Yes']
Contract : ['Month-to-month' 'Two year' 'One year']
Paperless Billing : ['Yes' 'No']
Payment Method : ['Mailed check' 'Electronic check' 'Bank transfer (automatic)'
    'Credit card (automatic)']
Churn Label : ['Yes' 'No']
Churn Reason : ['Competitor made better offer' 'Moved' 'Competitor had better devices'
    'Competitor offered higher download speeds'
    'Competitor offered more data' 'Price too high' 'Product dissatisfaction'
    'Service dissatisfaction' 'Lack of self-service on Website'
    'Network reliability' 'Limited range of services'
    'Lack of affordable download/upload speed' 'Long distance charges'
    'Extra data charges' "Don't know" 'Poor expertise of online support'
    'Poor expertise of phone support' 'Attitude of service provider'
    'Attitude of support person' 'Deceased']
```

Figure 15 : Modalités des variables

-On a importé la carte de Californie tout en choisissant les latitudes et longitudes des villes de Californie et voici le résultat :

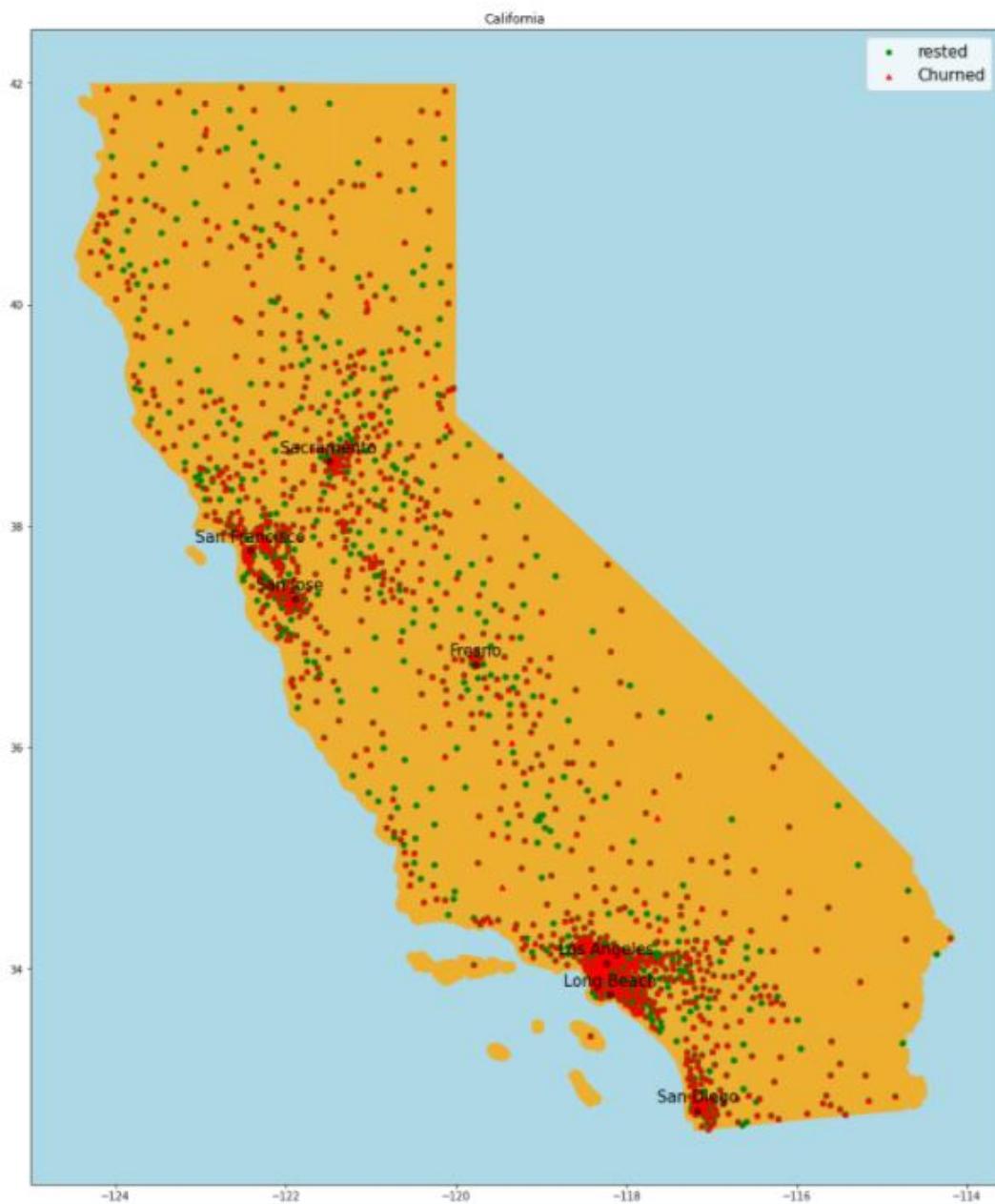


Figure 16 : Carte de Californie en fonction du Churn

D'après la figure ci-dessus on comprend que la plupart des clients qui ont quitté l'entreprise sont concentrés dans les villes : San Francisco , Los Angles , Long Beach. C'est pour cette raison qu'on a cherché à ordonner les villes par nombres de clients « churned », illustré par la figure suivante.

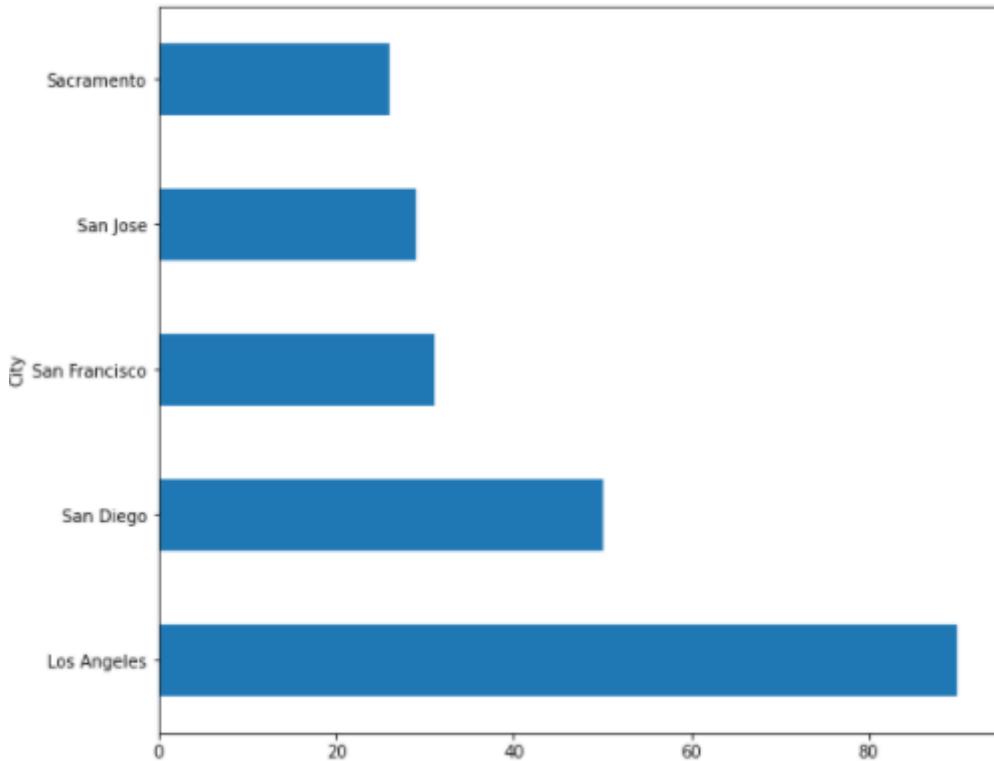


Figure 17: Villes de Californie en fonction du Churn

-Pour mieux comprendre les données disponibles on a décidé le nombre des clients en fonction de la raison qui les a poussés à quitter l'entreprise

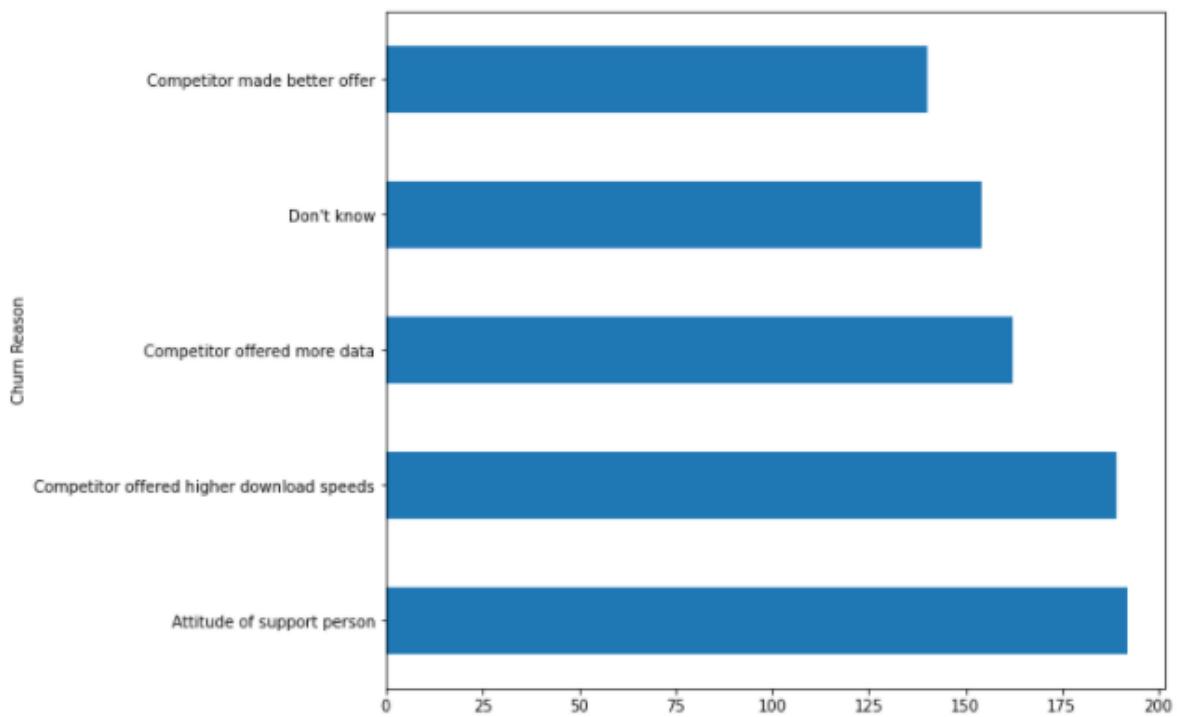


Figure 18 : Nombre de « Churn_reason »

-On conclue que l'attitude du personnel et les vitesses de téléchargement élevées sont les raisons principales pour lesquelles les clients ont quitté l'entreprise.

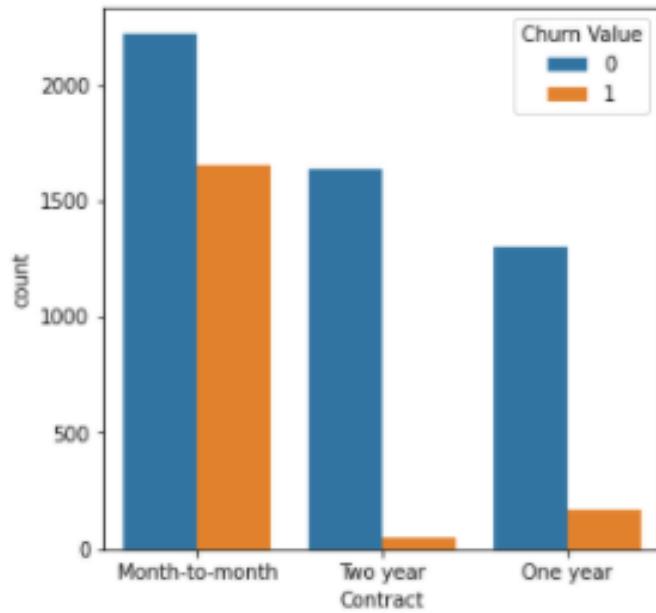


Figure 19 : Contrat en fonction du Churn

-Dans le but de savoir si le contrat signé par le client affecte son état dans l'entreprise ou pas , il s'est avéré il est plus probable que le client quittera l'entreprise, s'il a signé un contrat par mois .

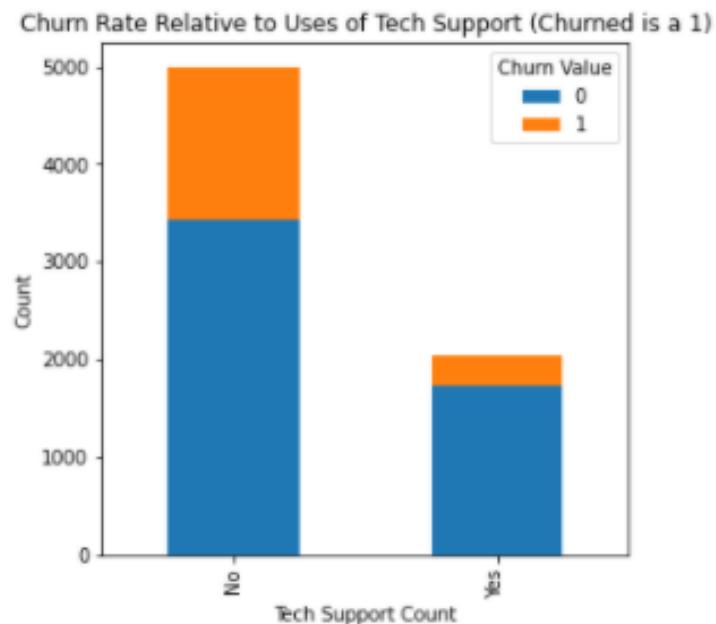


Figure 20 : Tech support en fonction churn

-Nous pouvons voir que les non-churners utilisent le "Tech Support" plus souvent que les clients « churn ».

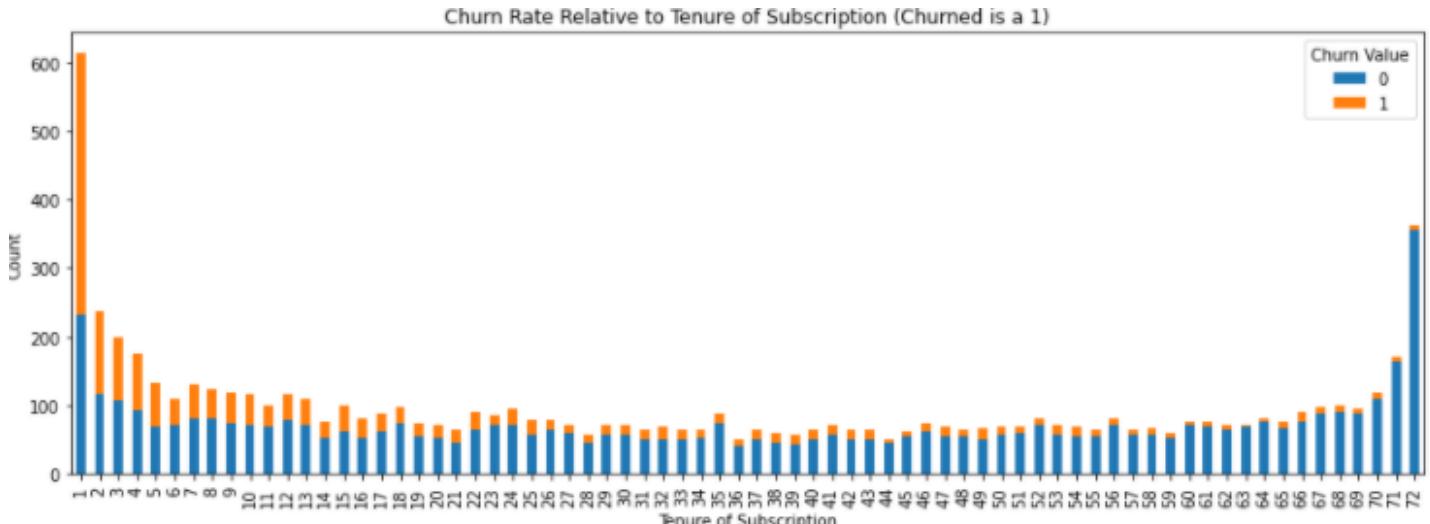


Figure 21 : Tenure en fonction de churn

-Nous pouvons clairement voir que plus un client reste longtemps abonné, moins il est susceptible de changer d'abonnement.

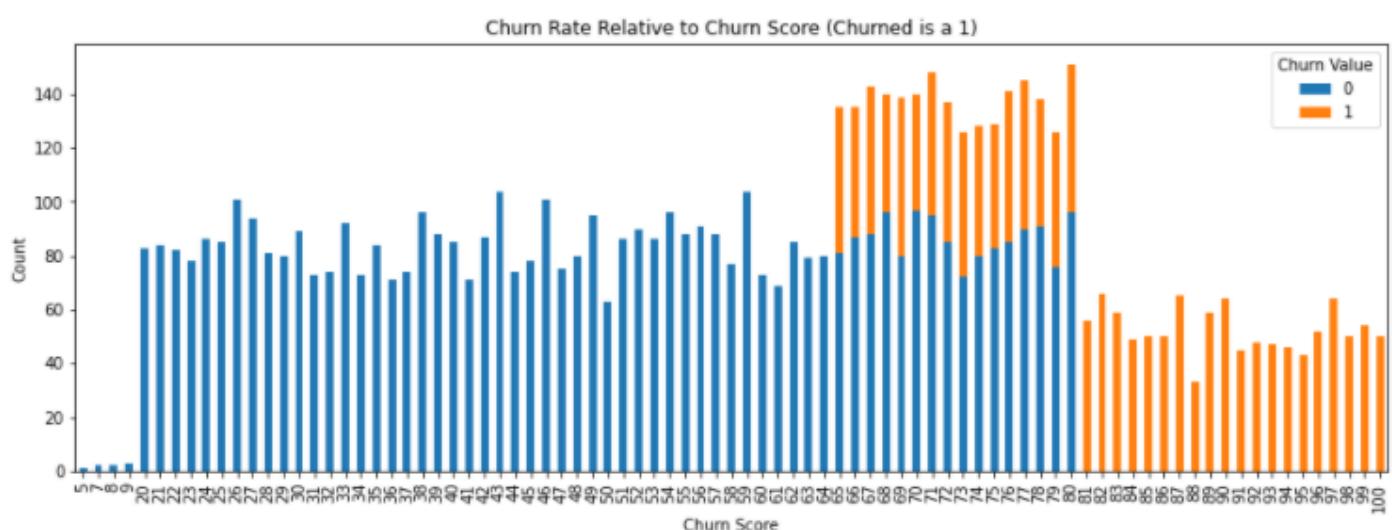


Figure 22 : Churn score en fonction du churn

-Nous pouvons constater que plus un client possède un "Churn Score" important plus il risque de quitter l'entreprise.

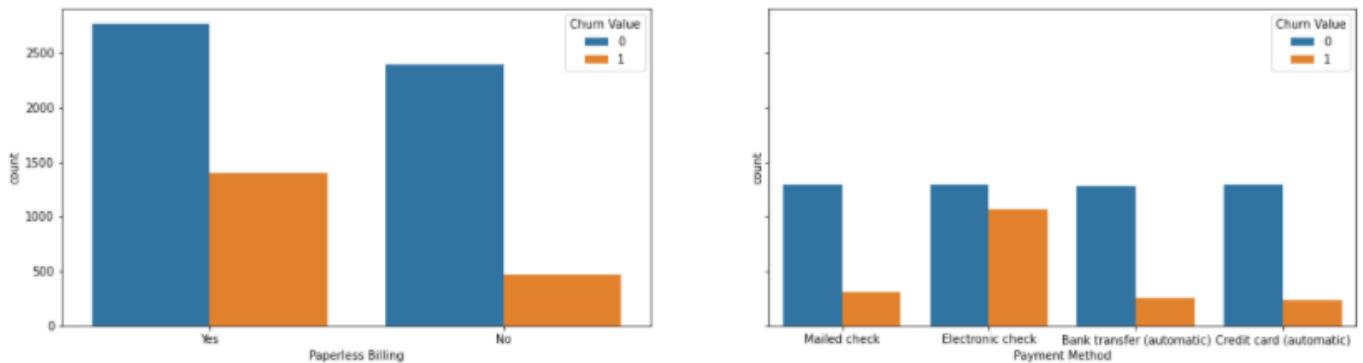


Figure 23: Paiement en fonction du churn

-Ces deux histogrammes servaient à démontrer s'il existait une relation entre la méthode de paiement et le taux de churn , nous pouvons voir que les clients qui utilisent la facturation électronique sont beaucoup plus enclins à abandonner .

-Affichage des valeurs aberrantes des colonnes « Tenure Months », « Monthly Charges » et « Total Charges » :

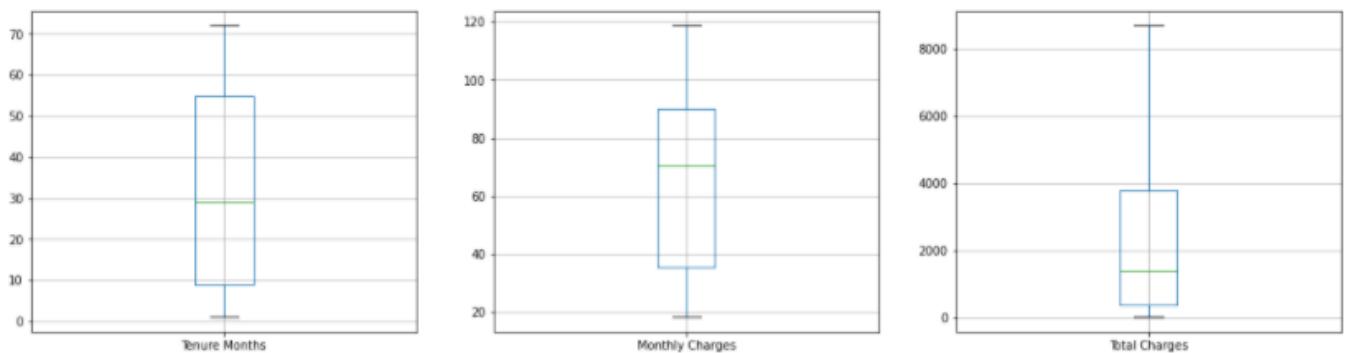


Figure 24: Valeurs aberrantes

-Le boxplot ne présente aucune valeur aberrante dans la distribution des différentes variables affichées

3- Préparation des données

a- Article 1-2

On doit préparer nos données afin d'avoir de bons résultats de la prédiction de nos clients. On a commencé par voir la qualité des données. On doit faire de sorte que nos données soient valides, pertinentes, complètes, consistantes et uniformes. Pour ce faire, on a commencé par vérifier au niveau de notre table s'il existe des observations dupliquées, des variables clés dupliquées, manquantes, nulles, positives ou négative Les spécificités du processus de préparation des données varient selon le secteur d'activité, l'entreprise et les besoins, mais le principe demeure le même. On fait une collecte de données, une découverte et une évaluation, un nettoyage, une transformation, un enrichissement et finalement un stockage. L'étape de centrage réduction est indispensable car les variables ne sont pas sur un même plan. Pour ce faire, on applique la formule du centrage réduction soustrait la moyenne en on divise sur l'écart-type. Le but est d'avoir une relation entre les attributs et les rendre visibles.

	tenure	MonthlyCharges	TotalCharges
0	-1.280248	-1.161694	-0.994194
1	0.064303	-0.260878	-0.173740
2	-1.239504	-0.363923	-0.959649
3	0.512486	-0.747850	-0.195248
4	-1.239504	0.196178	-0.940457

Figure 25: Affichage des cinq premières lignes des attributs centrés réduits

On a dû afficher la matrice de corrélation entre les variables numériques pour pouvoir différencier les variables fortement corrélées.

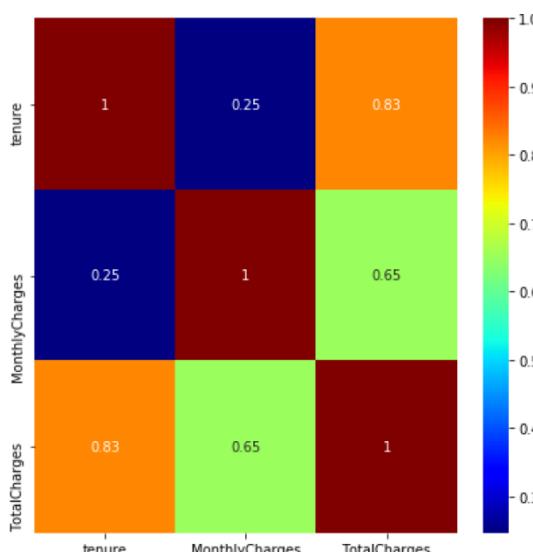


Figure 26 : Matrice de corrélation des variables numériques

D'après la matrice de corrélation on constate une forte corrélation entre les variables « Tenure » et « Totalcharges » donc pour éviter la redondance, on a supprimé la variable « Tenure ».

```
gender : ['Female' 'Male']
SeniorCitizen : [0 1]
Partner : ['Yes' 'No']
Dependents : ['No' 'Yes']
PhoneService : ['No' 'Yes']
MultipleLines : ['No phone service' 'No' 'Yes']
InternetService : ['DSL' 'Fiber optic' 'No']
OnlineSecurity : ['No' 'Yes' 'No internet service']
OnlineBackup : ['Yes' 'No' 'No internet service']
DeviceProtection : ['No' 'Yes' 'No internet service']
TechSupport : ['No' 'Yes' 'No internet service']
StreamingTV : ['No' 'Yes' 'No internet service']
StreamingMovies : ['No' 'Yes' 'No internet service']
Contract : ['Month-to-month' 'One year' 'Two year']
PaperlessBilling : ['Yes' 'No']
PaymentMethod : ['Electronic check' 'Mailed check' 'Bank transfer (automatic)'
'Credit card (automatic)']
```

Figure 27 : Les différentes modalités des variables catégorielles

On remarque que les variables dont les valeurs "No" et "No internet service" expriment le même concept donc il serait mieux de considérer "No internet service" comme "NO".

L'étape suivante est d'encoder toutes les variables catégorielles afin d'obtenir un tableau de valeurs numériques.

gender	SeniorCitizen	Partner	Dependents	PhoneService	OnlineSecurity	OnlineBackup	De
0	0	0	0	0	0	0	0
1	1	0	1	0	1	1	1
2	1	0	1	0	1	1	0
3	1	0	1	0	0	1	1
4	0	0	1	0	1	0	1

5 rows × 29 columns

Figure 466 : Les variables catégorielles encodées

Afin de visualiser la corrélation entre les attributs de notre Dataset, on a travaillé sur l'affichage d'une matrice de corrélation globale et qui contient tous les types des données en question. Toujours dans le but de minimiser nos données et éviter la redondance.

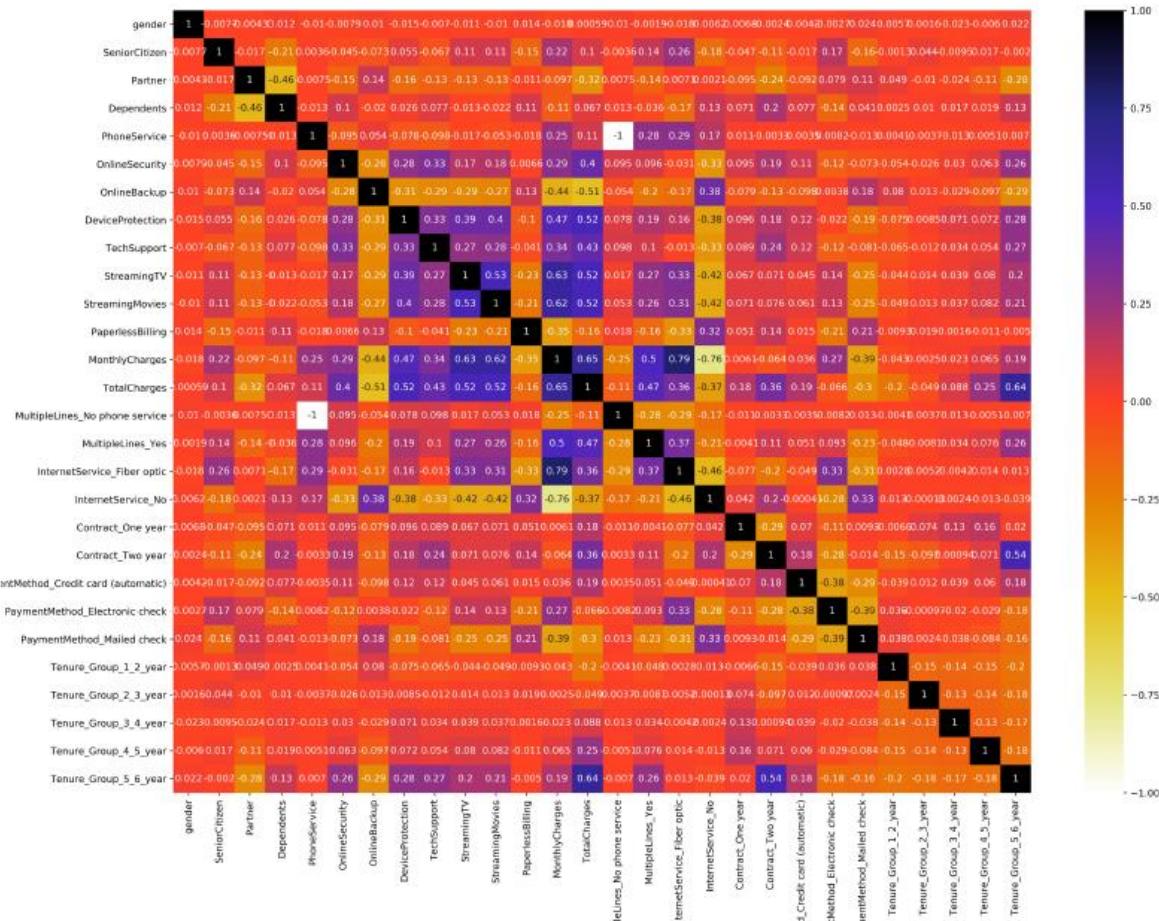


Figure 29 : Matrice de corrélation des attributs

On constate une forte corrélation entre les variables « MultipleLines_No phone service » et « PhoneService » donc pour éviter la redondance, on doit supprimer l'une des variables.

Pour l'article 2 on a dû ajouter quelques étapes de préparation des données pour le cas des caractéristiques continues pour s'assurer si elles suivent une distribution normale .

On a effectué une étude de la densité ainsi que de la droite d'Henry des deux variables « Tenure » et « Monthly Charges » :

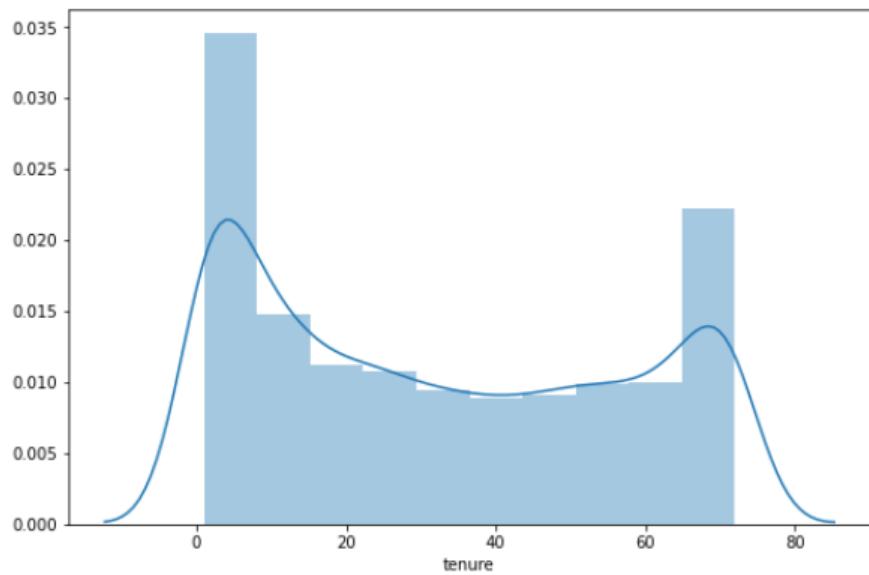


Figure 30 : Densité de "Tenure"

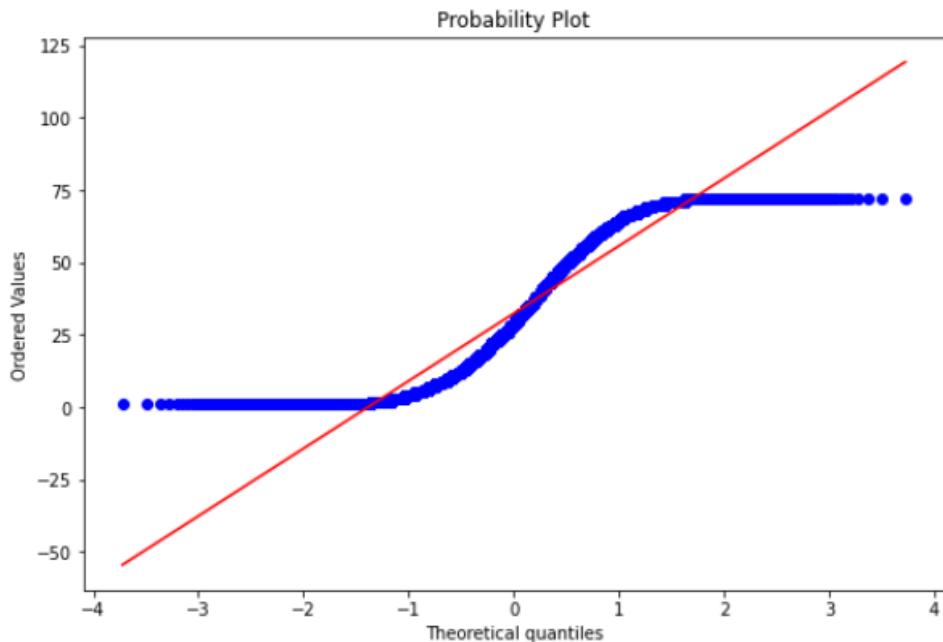


Figure 31 : Droite d'Henry de "Tenure"

À partir de l'histogramme de « Tenure », nous pouvons voir qu'un grand nombre de clients ne sont dans l'entreprise que depuis moins de 5 mois. Cela signifie qu'ils sont des clients relativement nouveaux. Il y a également un nombre important de clients qui font partie de l'entreprise depuis plus de 65 mois.

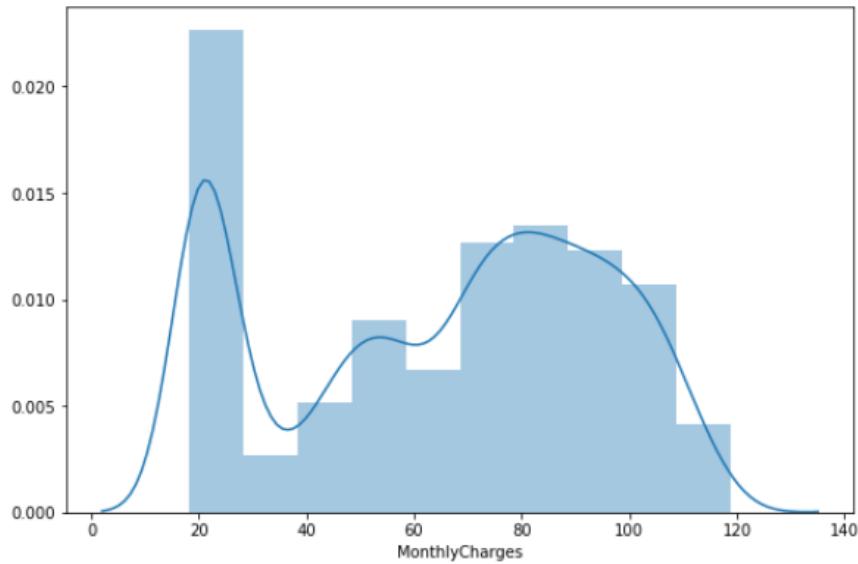


Figure 32 : Densité de "Monthly Charges"

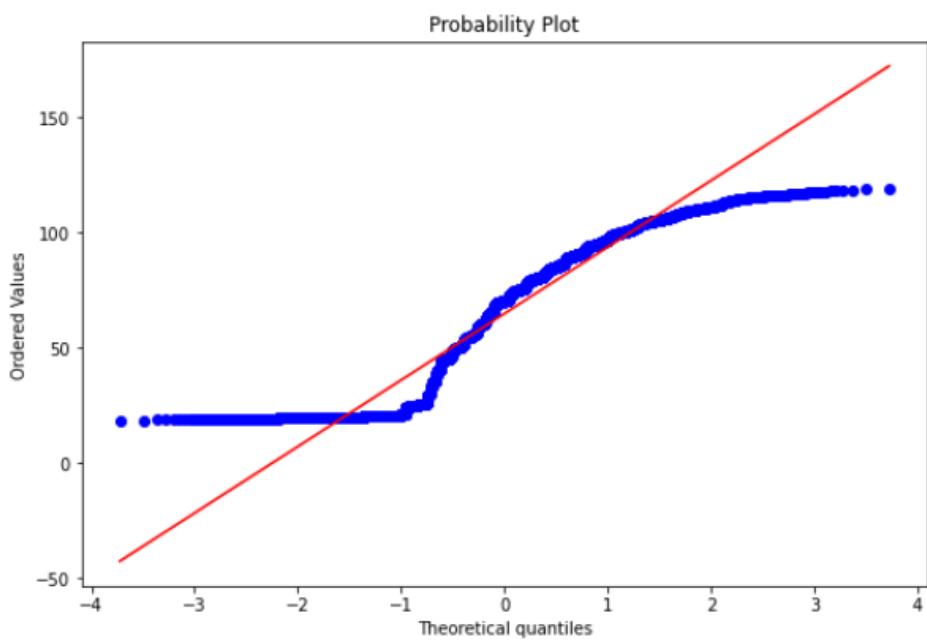


Figure 33 : Droite d'Henry de "Monthly Charges"

D'après l'histogramme de « MonthlyCharges », il semble avoir deux groupes de clients. Le premier groupe a des frais mensuels d'environ 20 à 25 tandis que le second groupe a des frais mensuels d'environ 75 à 95.

D'après les figures que nous avons obtenues nous pouvons clairement conclure que les deux variables étudiées ne suivent pas une distribution normale du coup on applique une conversion catégorielle.

b- Article 3

- On a commencé par la suppression des colonnes non nécessaires pour notre prédition « CustomerID » , « Count » , « Country » , « State » , « Churn Label »
- On a étudié la corrélation des différentes variables avec la variable cible :

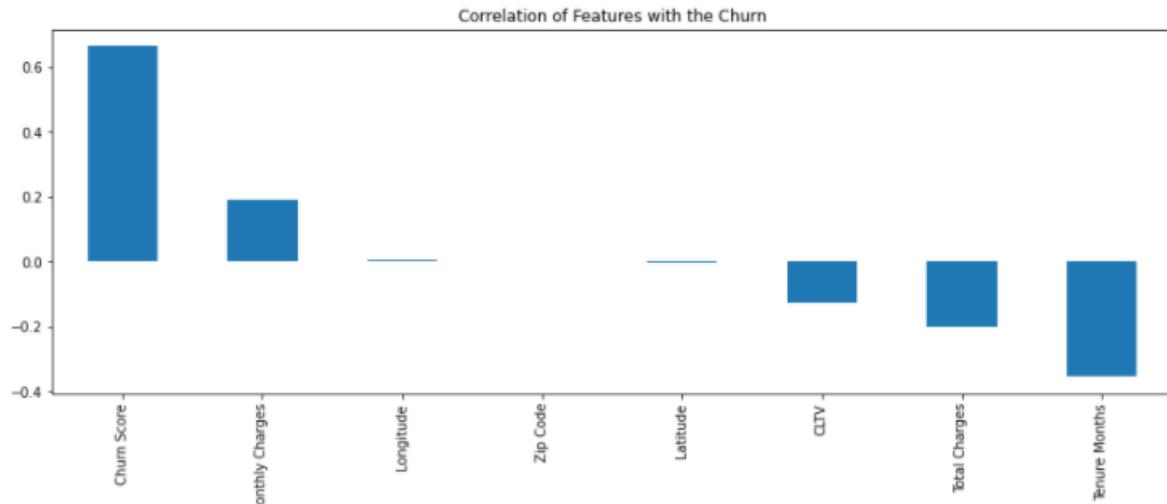


Figure 34: Corrélation avec la variable cible

- On a supprimé les variables qui sont faiblement corrélées avec « Churn » : Zip Code', 'Lat Long', 'Latitude', 'Longitude', 'City'.
- Ensuite on effectué un encodage des variables catégorielles et diviser nos données en données d'apprentissage et données de test.
- Pour le clustering on a effectué une transformation de centrage réduction (MinmaxScaler).

4-Modélisation :

a-Article 1

Dans le but d'avoir des meilleures résultats, dans cette partie on s'intéressera à étudier la précision de ces différents algorithmes suivants :

-KNN :

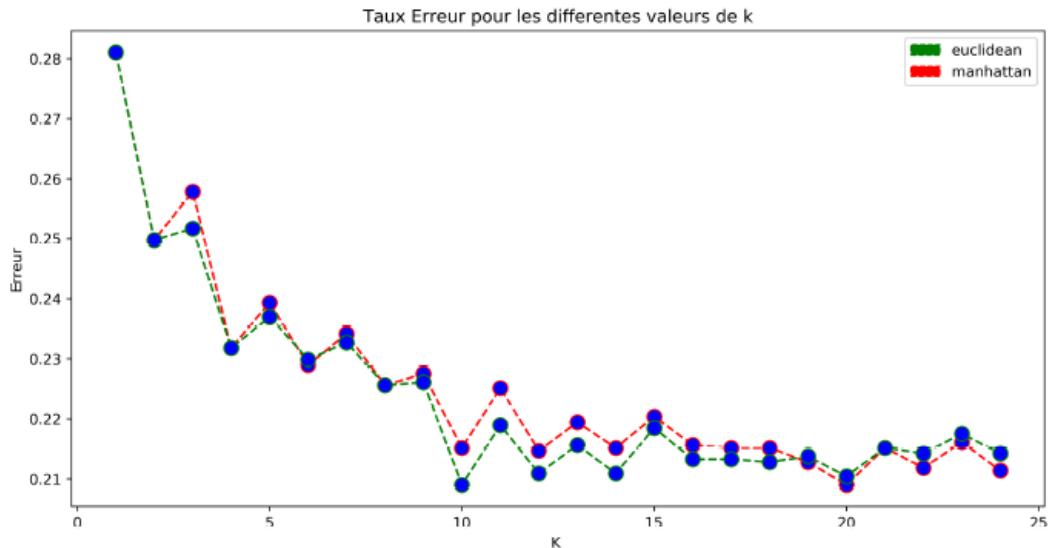


Figure 35 : Choix de la valeur k et la métrique

D'après la figure ci-dessus, on peut constater que la meilleure valeur pour k est égale à 20 avec une métrique de « Manhattan ».

Accuracy of K-NN classifier on training set: 0.8119

Accuracy of K-NN classifier on test set: 0.7910

-Random Forrest:

Accuracy of RandomForest classifier on training set: 0.9984

Accuracy of RandomForest classifier on test set: 0.7886

-SVM:

Accuracy of SVM classifier on training set: 0.8176

Accuracy of SVM classifier on test set: 0.7976

-Logistic Regression:

Accuracy of LogisticRegression classifier on training set: 0.8058

Accuracy of LogisticRegression classifier on test set: 0.7995

b-Article 2

On a commencé par une division des données en 70% données d'apprentissage et 30% données de test et puis on a appliqué les trois modèles « Gaussien », « Bernoulli » et « Multinomial » sur les 27 variables que nous avons et on a obtenu les résultats suivants :

- 'Gaussian': 0.7200315711278941
- 'Bernoulli': 0.7649312863687012
- 'multinomial': 0.7767176344352276

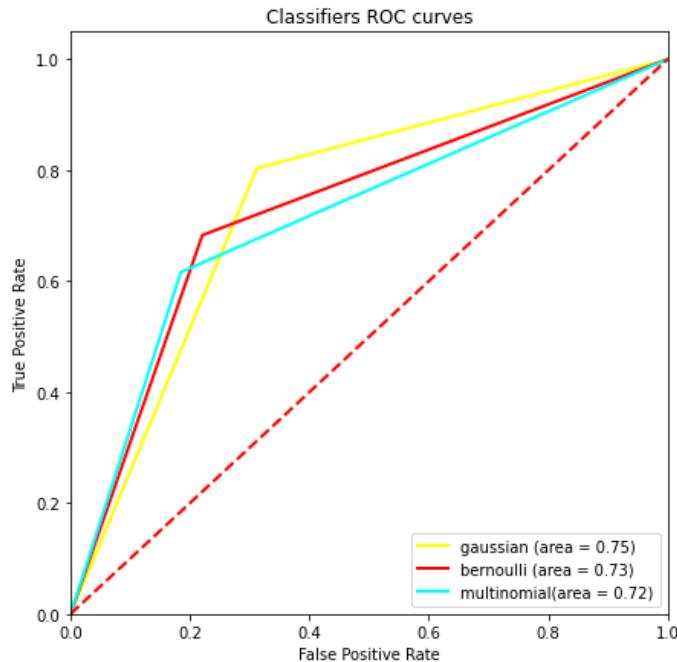


Figure 36 : Courbe Roc Naives Bayes

-D'après la Courbe ROC illustrée dans la figure 25 on peut clairement conclure qu e le modèle à retenir est « GaussianNB » puisque il atteint la plus grandes valeurs des vrais positifs en le comparant aux modèles de « BernoulliNB » et « Mutinomial NB »

-Puis on a passé à l'étape de vérification des données d'apprentissage de test et d'apprentissage avec lesquels on a commencé :

- Accuracy of Nb classifier on training set: 0.72
- Accuracy of Nb classifier on test set: 0.72

-Ensuite on calculait la valeur du « roc_auc » et « Accuracy » du départ : **0.71848341**

	precision	recall	f1-score	support
0	0.91	0.69	0.78	1555
1	0.48	0.80	0.60	555
accuracy			0.72	2110
macro avg	0.69	0.75	0.69	2110
weighted avg	0.79	0.72	0.73	2110

Figure 37 : f1 score

-D'après la figure on constate que la valeur de f1-score pour les données de test (1) est de 60% , pour les données d'apprentissage (0) elle est de 78%

-Matrice de confusion :

Cette matrice permet de comprendre de quelle façon le modèle de classification est confus lorsqu'il effectue des prédictions. Ceci permet non seulement de savoir quelles sont les erreurs commises, mais surtout le type d'erreurs commises .

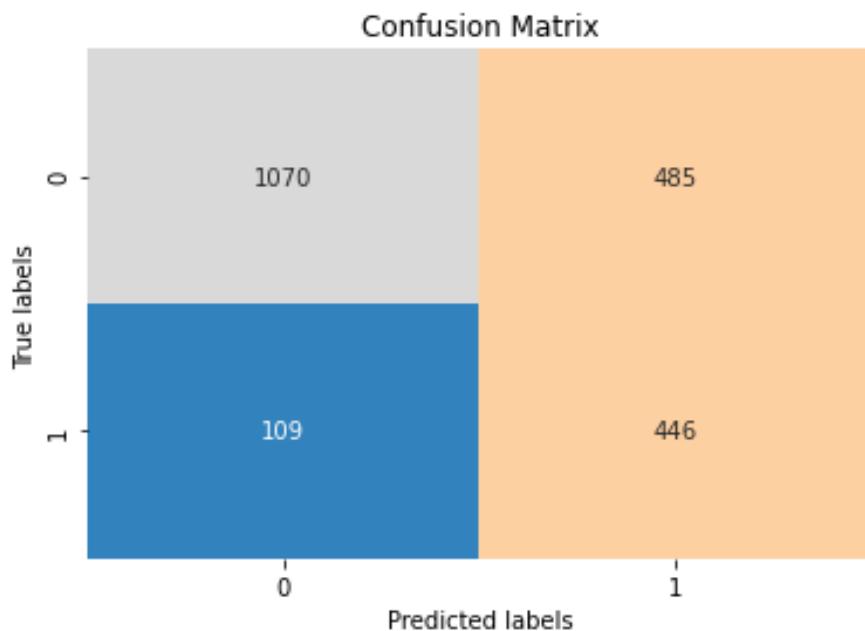


Figure 792 : Matrice de confusion

Dans notre cas la matrice affiche clairement des valeurs de vrais positifs et vrais négatifs largement supérieurs à celles des faux positifs et faux négatifs ce qui indique que nous sommes en train d'effectuer une bonne prédiction.

-Ensuite on a fait une “ feature selection” en utilisant SBS (Sequential Backward Selection) , SFS (Sequential Forward Selection) , SFFS (Sequential Floating Forward Selection) , SFBS (Sequential Floating Backward Selection) tout en calculant la valeur de “roc_auc” et “Accuracy”

- On va maintenant essayer d'appliquer une autre façon dont laquelle on normalise les variables quantitatives et on reverra les résultats de chaque algorithme , on a importé la bibliothèque minMaxScaler et on a fait une transformation centrage réduction , on a supprimé les ancienne colonnes "MonthlyCharges","TotalCharges","tenure" du la dataframe tout en les remplaçant par les nouvelles variables centrées réduites , ensuite on a répété le même processus de feature selection.

c-Article 3 :

Arbre de décision :

Un arbre de décision est un schéma représentant les résultats possibles d'une série de choix interconnectés , afin d'arriver au résultat on a commencé par une division de données (70% apprentissage 30% test) tout en initiant le DecisionTreeClassifier et entraînant le modèle à l'aide des données d'apprentissage .

Clustering :

On a commencé par une sélection des variables les plus importantes ("Tenure Months","Monthly Charges","Churn Score")

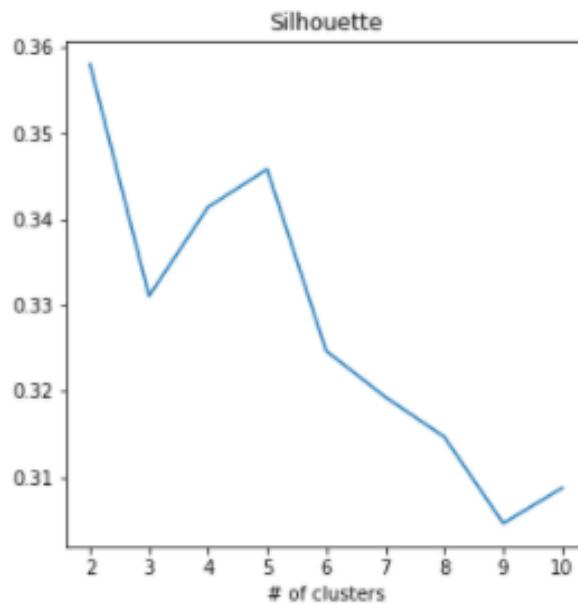


Figure 39 : Nombre de clusters

Afin de déterminer le nombre de clusters à réaliser on a affiché la courbe ci-dessus , qui a comme pic la valeur 3 (notre choix comme nombre de clusters) et puis on appliquer l'algorithme kmeans sur les clients .

5-Evaluation :

a- Article 1

Après une vérification faite sur les résultats trouvés dans l'article , on s'est demandé si on pourrait les améliorer en effectuer d'autres algorithmes sur nos données.

-Matrices de confusion :

Algorithme 1 (KNN) :

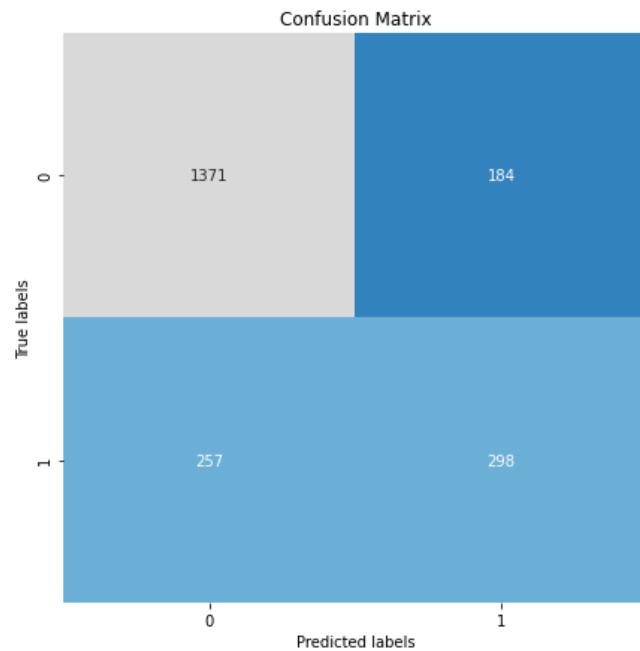


Figure 40 : Matrice de confusion KNN

Algorithme 2 (Random Forrest) :

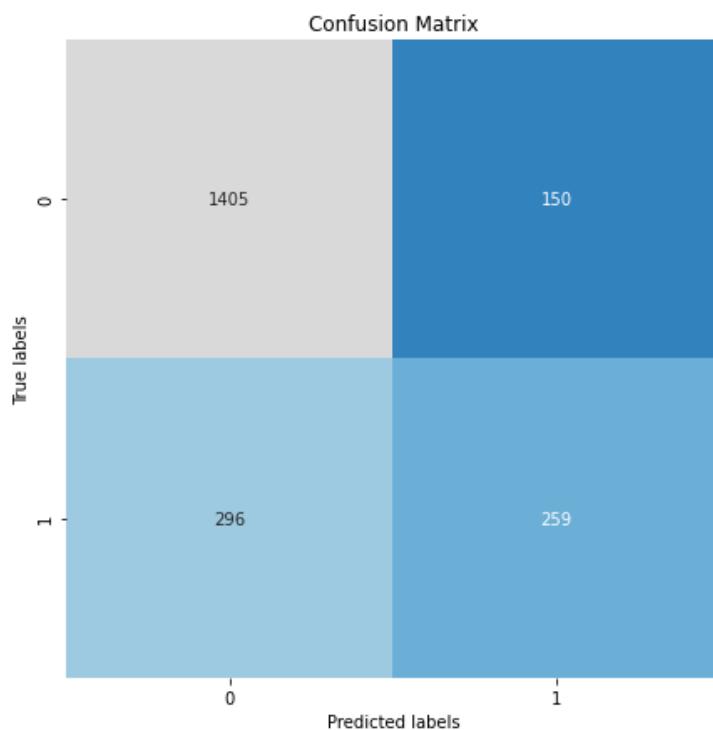


Figure 41 : Matrice de confusion Random Forrest

Algorithme 3 (Logistic Regression):

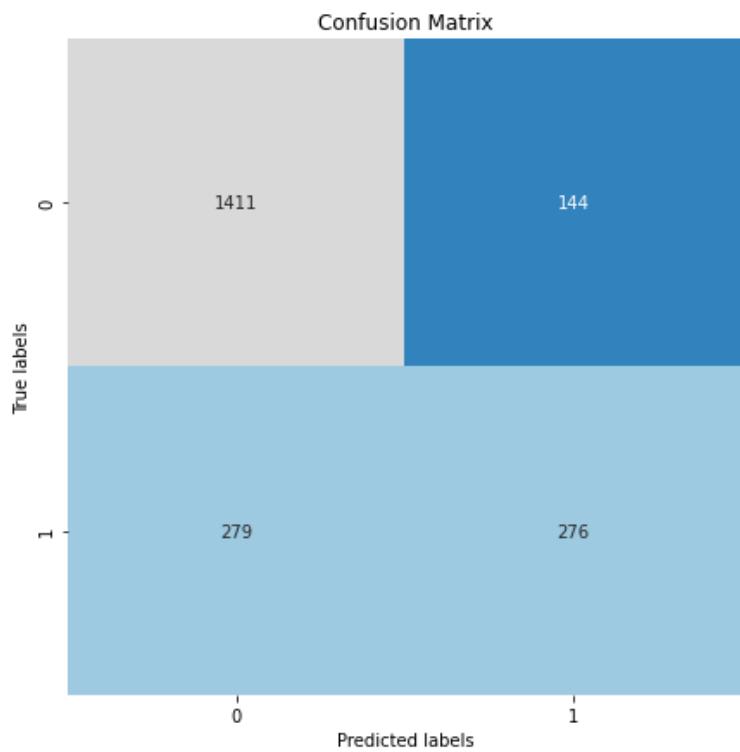


Figure 42 : Matrice de confusion Logistic Regression

Algorithme 4 (SVM):

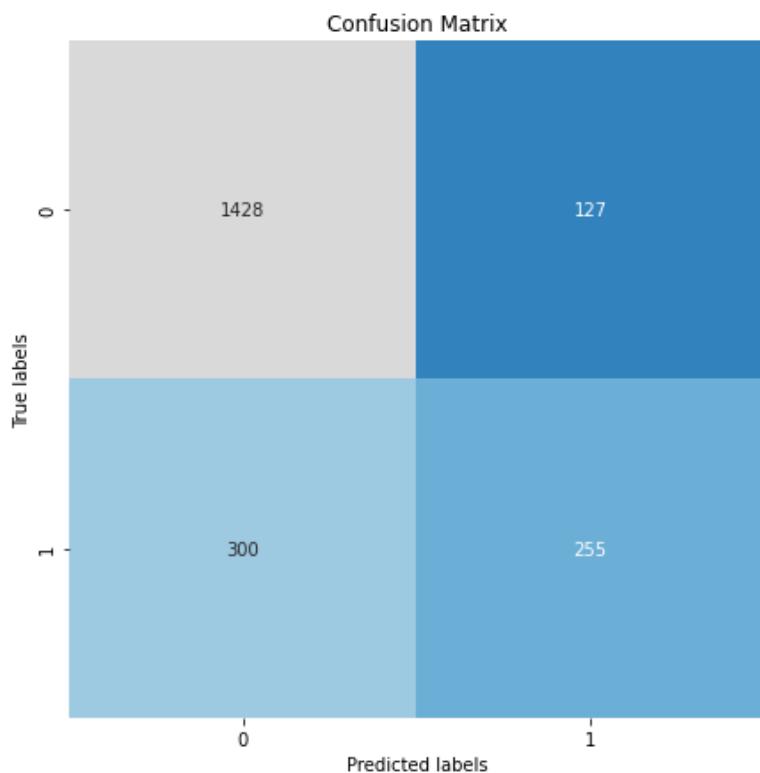


Figure 43 : Matrice de confusion SVM

-Table des scores:

Cette table affiche le score de chaque modèle tout en les ordonnant :

	Model	Score
2	Logistic Regression	0.799526
3	Support vector machine (SVM)	0.797630
0	k-nearest neighbors (KNN)	0.790995
1	Random forest	0.788626

Figure 44 : Tables des scores

-La courbe ROC (Receiver operating characteristic)

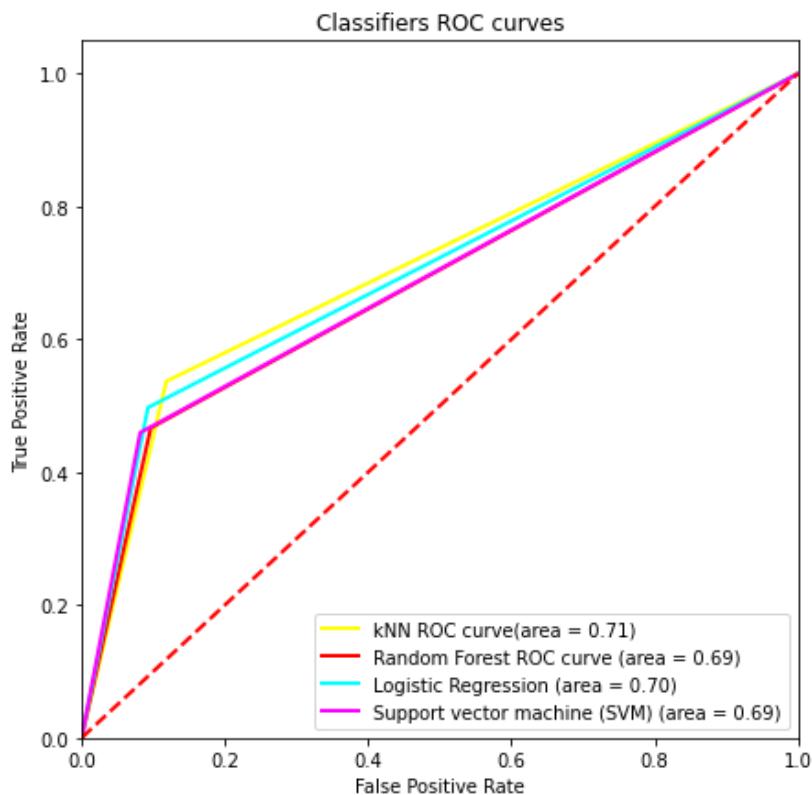


Figure 45 : Courbe Roc Article

-Dans cet article, nous avons comparé plusieurs algorithmes qui peuvent prédire si le client mettra fin au service et choisira une autre organisation ou non. La comparaison de plusieurs classificateurs nous aidera à prédire avec précision le désabonnement des clients. On a pu trouver des valeurs différentes et améliorées à celles trouvées dans l'article , ce qui indique que nous avons pu préparer nos données d'une façon meilleure. Afin de s'assurer que les meilleurs algorithmes ont été utilisés dans cet article , on a réalisé une comparaison avec « Logistic regression » qui d'après le tableau comparatif des scores, les tables de confusion et la courbe se manifeste comme étant plus performant que « KNN » « SVM » et « Random Forrest ».

b- Article 2

Pour mieux comparer les valeurs de « Accuracy » et de « Roc_auc » on a décidé de les afficher tous en pourcentage dans un tableau comparatif des différents modèles de performance qu'on vient d'utiliser (SFS – SBS – SFFS – SFBS) . Puis on affichera deux courbes représentatives des valeurs de « Accuracy » et de « Roc_auc » ensuite on ordonnera ces modèles suivant leurs scores.

-Tableau modèle de performance :

kfeatures	Accuracy NB	AUC NB	Accuracy SFS	AUC SFS	Accuracy SBS	AUC SBS	Accuracy SFFS	AUC SFFS	Accuracy SFBS	AUC SFBS
1	71.85%	0.745853	73.42%	0.672759	73.42%	0.672759	73.42%	0.672759	73.42%	0.672759
2	71.85%	0.745853	73.42%	0.746365	72.99%	0.746365	73.42%	0.746365	73.42%	0.746365
3	71.85%	0.745853	73.73%	0.794263	75.78%	0.794263	73.73%	0.794263	75.82%	0.794263
4	71.85%	0.745853	74.89%	0.803404	77.36%	0.803404	74.89%	0.803404	77.49%	0.803404
5	71.85%	0.745853	76.38%	0.810110	76.72%	0.807163	76.38%	0.810110	75.55%	0.810110
6	71.85%	0.745853	76.38%	0.814284	77.70%	0.813018	76.38%	0.814284	76.65%	0.813018
7	71.85%	0.745853	75.97%	0.817113	77.72%	0.817820	75.97%	0.817113	76.83%	0.817820
8	71.85%	0.745853	76.35%	0.819027	77.74%	0.820632	77.47%	0.819027	76.81%	0.821962
9	71.85%	0.745853	76.27%	0.822886	77.01%	0.822364	76.86%	0.822886	75.71%	0.823469
10	71.85%	0.745853	76.92%	0.824898	76.95%	0.823673	77.39%	0.824898	76.35%	0.824898
11	71.85%	0.745853	77.10%	0.826077	76.83%	0.825724	77.20%	0.826077	76.61%	0.825724
12	71.85%	0.745853	77.29%	0.826610	76.58%	0.827875	77.42%	0.826610	76.59%	0.827875
13	71.85%	0.745853	77.25%	0.827579	76.76%	0.828908	76.54%	0.827579	76.58%	0.828908
14	71.85%	0.745853	77.22%	0.828662	76.73%	0.829584	76.73%	0.829584	76.73%	0.829584
15	71.85%	0.745853	76.92%	0.829535	76.85%	0.830213	76.51%	0.830213	76.85%	0.830213
16	71.85%	0.745853	76.79%	0.830082	76.91%	0.830603	76.19%	0.830603	76.91%	0.830603
17	71.85%	0.745853	76.61%	0.830359	77.05%	0.830854	77.18%	0.830854	77.05%	0.830854
18	71.85%	0.745853	76.37%	0.830502	77.08%	0.830794	77.09%	0.830794	77.08%	0.830794
19	71.85%	0.745853	76.44%	0.830432	77.10%	0.830560	77.08%	0.830560	77.10%	0.830560
20	71.85%	0.745853	76.96%	0.830295	77.13%	0.830238	76.96%	0.830238	77.13%	0.830238
21	71.85%	0.745853	76.83%	0.829944	77.12%	0.829848	76.83%	0.829848	77.12%	0.829848
22	71.85%	0.745853	76.65%	0.829462	77.09%	0.829067	76.91%	0.829374	77.09%	0.829067
23	71.85%	0.745853	76.75%	0.828833	76.91%	0.828742	76.75%	0.828833	76.91%	0.828742
24	71.85%	0.745853	76.76%	0.828214	76.81%	0.828359	76.76%	0.828214	76.81%	0.828359
25	71.85%	0.745853	76.58%	0.827985	76.72%	0.827985	76.58%	0.827985	76.72%	0.827985
26	71.85%	0.745853	75.33%	0.827115	75.33%	0.827115	75.33%	0.827115	75.33%	0.827115
27	71.85%	0.745853	73.89%	0.824678	73.89%	0.824678	73.89%	0.824678	73.89%	0.824678

kfeatures	Accuracy NB	AUC NB	Accuracy SFS	AUC SFS	Accuracy SBS	AUC SBS	Accuracy SFFS	AUC SFFS	Accuracy SFBS	AUC SFBS
average	71.85%	0.745853	76.13%	0.81524	76.53%	0.815282	76.22%	0.815343	76.32%	0.815527

Figure 46: Tableau modèle de performance

-Courbe d'Accuracy des modèles :

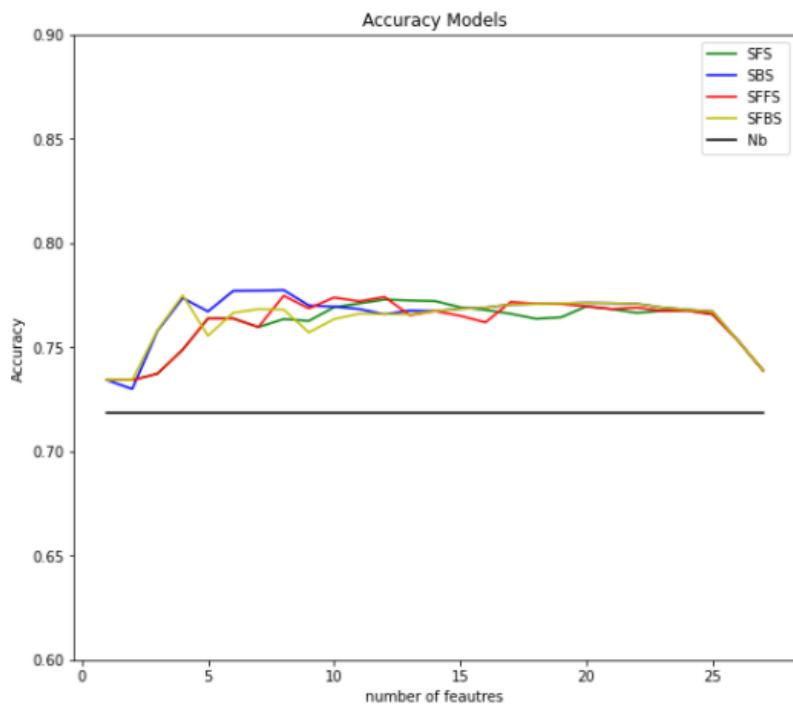


Figure 47: Accuracy des modèles

-Courbe d'AUC des modèles :

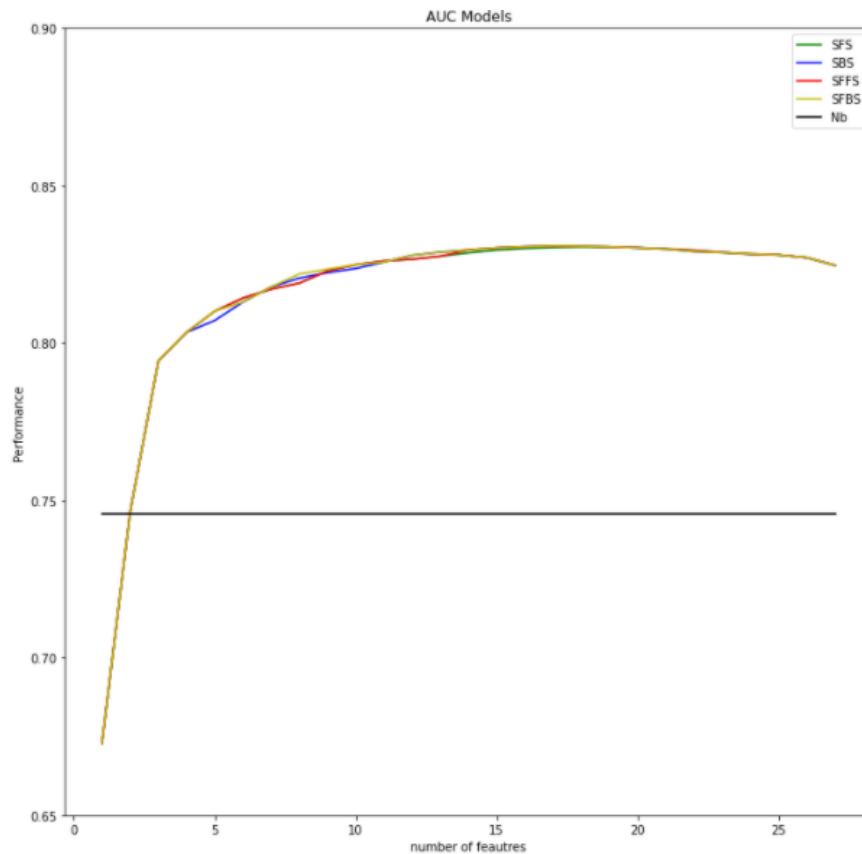


Figure 48 : AUC des modèles

-Tableaux d'ordonnancement de performance :

	Model	roc_auc
2	Aaccuracy SBS	0.830854
3	Aaccuracy SFFS	0.830854
4	Aaccuracy SFBS	0.830854
1	Aaccuracy SFS	0.830502
0	Aaccuracy NB	0.745853

Figure 999 : Ordonnancement de « roc_auc »

	Model	Score
2	Aaccuracy SBS	77.74%
4	Aaccuracy SFBS	77.49%
3	Aaccuracy SFFS	77.47%
1	Aaccuracy SFS	77.29%
0	Aaccuracy NB	71.85%

Figure 50: Ordonnancement de « Accuracy »

- Dans cette étude, il a été proposé d'appliquer la sélectionne des caractéristiques pertinentes (feature selection) . Les résultats les plus élevés sont un modèle avec sélection de caractéristiques SBS et SBFS ce qui conforme aux résultats de l'article. Néanmoins , si on compare notre tableau de performance dans les deux cas (normalisation des données continues ou pas) et celui de l'article on peut clairement voir une différence ce qui indique une meilleure application des modèles de notre part.

c-Article3 :

-Arbre de décision :

Résultats des scores des données d'apprentissage et de test :

Score sur les des données d'apprentissage : 1.0

Score sur les des données de test : 0.9781990521327014

	precision	recall	f1-score	support
0	0.98522800	0.98522800	0.98522800	1557
1	0.95840868	0.95840868	0.95840868	553
accuracy			0.97819905	2110
macro avg	0.97181834	0.97181834	0.97181834	2110
weighted avg	0.97819905	0.97819905	0.97819905	2110

Figure 51 : f1 score article 3

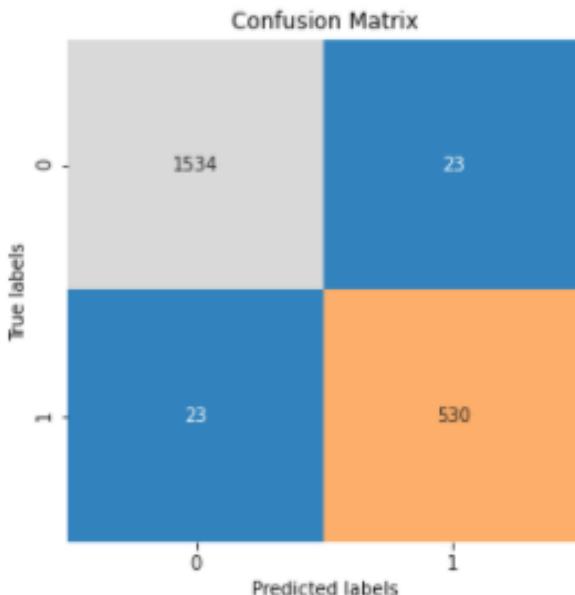


Figure 52 : Matrice de confusion arbre de décision

On remarque que ce modèle est assez performant vu les résultats de la matrice de confusion (vrais positifs et négatifs sont largement supérieurs aux faux positifs et faux négatifs) , aussi la valeur de f1 score pour (0) apprentissage et (1) test sont assez élevée .

-Clustering :

	Tenure Months	Monthly Charges	Churn Score	Cluster
0	0.014085	0.354229	0.852632	1
1	0.014085	0.521891	0.652632	1
2	0.098592	0.809950	0.852632	1
3	0.380282	0.861194	0.831579	1
4	0.676056	0.850249	0.884211	2

Figure 53 : Affichage de 5 premières lignes

-On remarque l'ajout d'une colonne cluster ayant les valeurs :

- 0 : cluster 1
- 1 : cluster 2
- 2 : cluster 3

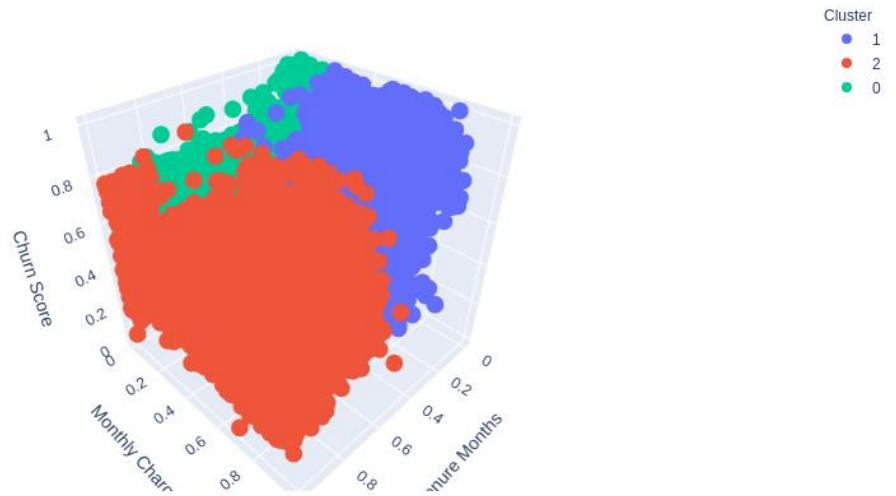


Figure 54 : Représentation 3D des clusters

-La figure nous montre la condensation des points de chaque cluster et aussi la clarté de la distinction de chaque cluster individuellement.

- On a ensuite affiché le pourcentage de chaque cluster par le camembert et le tableau ci-dessous :

Taille de chaque Cluster

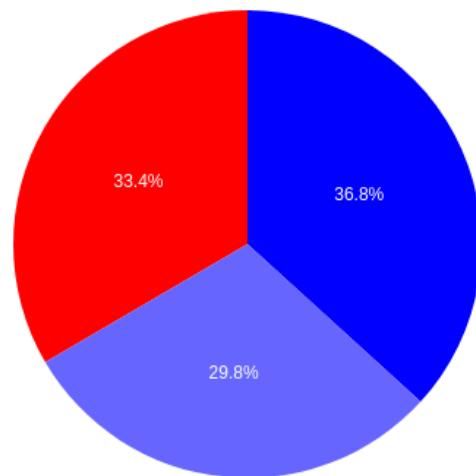


Figure 55 : La taille de chaque cluster 1

	size	Percentage
0	2097	0.298208
1	2350	0.334187
2	2585	0.367605

Figure 56 : La taille de chaque cluster 2

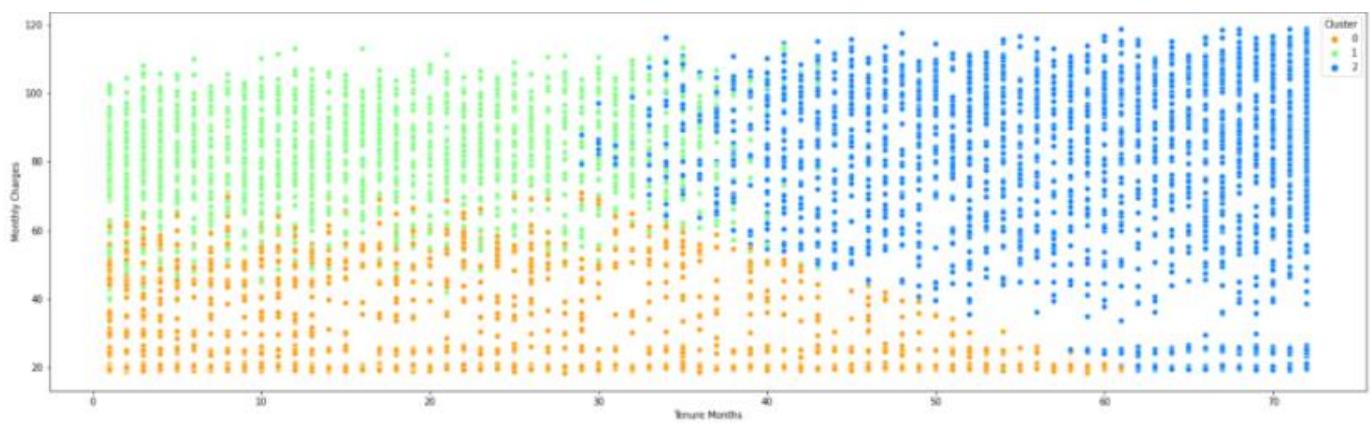


Figure 57: Tenure months en fonction de monthly charges

-On peut facilement visualiser des différents clusters d'après cette figure

(Exemple : la valeur de tenure Months pour le cluster 0 varie entre 0 et 60.)

-Analyse descriptive des différents clusters :

	Tenure Months	Monthly Charges	Churn Score
count	2097.000000	2097.000000	2097.000000
mean	19.714831	30.327372	52.539819
std	16.869383	13.644840	20.169332
min	1.000000	18.250000	7.000000
25%	4.000000	19.950000	35.000000
50%	15.000000	21.300000	52.000000
75%	32.000000	43.450000	69.000000
max	61.000000	70.900000	100.000000

Figure 58 : Cluster 0

	Tenure Months	Monthly Charges	Churn Score
count	2350.000000	2350.000000	2350.000000
mean	14.055319	79.507872	70.748511
std	11.434579	15.130307	19.344806
min	1.000000	39.500000	20.000000
25%	4.000000	70.200000	61.000000
50%	11.000000	79.000000	74.000000
75%	23.000000	90.500000	85.000000
max	44.000000	113.600000	100.000000

Figure 59: Cluster 1

	Tenure Months	Monthly Charges	Churn Score
count	2585.000000	2585.000000	2585.000000
mean	59.426692	79.389168	52.785687
std	11.056748	27.013960	19.835969
min	29.000000	19.100000	5.000000
25%	51.000000	64.050000	36.000000
50%	62.000000	85.250000	53.000000
75%	70.000000	100.650000	69.000000
max	72.000000	118.750000	100.000000

Figure 60 : Cluster 2

Cluster 0: Ce groupe de clients est de tenure months entre 1 et 61 mois , des charges mensuels entre 18.25 et 70.9 et le churn score varie entre 7 est 100 avec un score churn total de 2097

Cluster 1: Ce groupe de clients est de tenure months entre 1 et 44 mois , des charges mensuels entre 39.5 et 113.6 et le churn score varie entre 20 est 100 avec un churn score total de 2350

Cluster 2: Ce groupe de clients est de tenure months entre 29 et 72 mois , des charges mensuels entre 19.1 et 118.75 et le churn score varie entre 5 est 100 avec un score churn total de 2585

- ⇒ Les moyennes des churn scores indique que les clients qui appartiennent au cluster 1 sont ceux qui quitteront l'entreprise en premier lieu.

VIII - Outils de travail

Pour assurer la bonne réalisation de ce projet on a eu recours à plusieurs bibliothèques pour qu'on puisse utiliser les maintes ressources qu'elles offrent :

```
-import numpy  
-import pandas  
-import geopandas  
-import matplotlib.pyplot  
-from shapely.geometry import Point,Polygon  
-import seaborn  
-import pylab  
-import scipy.stats  
-from pandas_profiling import ProfileReport  
-from sklearn.naive_bayes import MultinomialNB, GaussianNB, BernoulliNB  
-from sklearn.model_selection  
-import cross_val_score,train_test_split  
-from sklearn.metrics import accuracy_score,roc_curve,  
auc,classification_report,confusion_matrix,roc_auc_score  
-from sklearn.preprocessing import MinMaxScaler  
-from mlxtend.feature_selection import SequentialFeatureSelector  
-from sklearn import metrics  
-from sklearn.preprocessing  
-import LabelEncoder  
-from sklearn.preprocessing import StandardScaler  
-from sklearn.model_selection import train_test_split  
-from sklearn.neighbors import KNeighborsClassifier  
-from sklearn.ensemble import RandomForestClassifier  
-from sklearn.linear_model import LogisticRegression  
-from sklearn.svm import SVC -import matplotlib.patches as mpatches  
-from sklearn.metrics import accuracy_score  
-from sklearn.cluster import KMeans  
-import plotly.express
```

```
-from sklearn.tree import DecisionTreeClassifier  
-from sklearn.preprocessing import MinMaxScaler
```

IX- Conclusion

Durant ce projet on a eu recours aux nouvelles connaissances que nous avons appris durant le module du Machine Learning et on a dû les mettre en évidence afin de pouvoir atteindre les résultats désirés. On est arrivé à voir les problèmes qui poussent le client à quitter cette entreprise et les avantages qui les aident à rester. Dans le cadre du gain en termes de clients et en termes d'argent, les méthodes effectuées nous étaient un moyen efficace pour atteindre cet objectif.