BOULAINE Thomas
BOUZIANE Mehdi
BIRADEPAN Brunthaban

A3 – Groupe TD A

PROJET DIA : Optimisation de l'allocation de la bande passante par apprentissage automatique

DATE DE RENDU : LE JEUDI 16 JANVIER 2024

CHARGE DE TD : Mr FAKIR ZACKARY

SOMMAIRE :

I. INTRODUCTION

1) Contexte et objectifs du projet

Avec l'explosion de l'utilisation des données mobiles et l'émergence d'applications nécessitant une large bande passante (streaming, jeux en ligne, réalité augmentée), les fournisseurs de services de télécommunications font face à un défi majeur : optimiser l'allocation de la bande passante en temps réel.

Dans ce contexte, ce projet vise à développer un modèle d'apprentissage automatique capable de prédire la demande en bande passante en temps réel, tout en proposant des ajustements automatiques pour répartir efficacement les ressources réseau. Pour se faire, nous travaillerons avec les données disponibles sur la plateforme de l'OCDE (fournis dans le sujet). L'objectif est de maximiser la qualité de service et d'offrir une solution adaptée à la demande actuelle.

Les étapes clés incluent l'exploration des données, leur nettoyage, et l'application de différents algorithmes vu en cours.

2) Présentation des caractéristiques sélectionnées

Le jeu de données utilisé dans ce projet contient des variables pour analyser et prédire la demande en bande passante. Parmi toutes les caractéristiques présentes dans le jeu de donnée initial, notre groupe a décidé de retenir les suivantes :

- **Pays** : Permet de contextualiser les données par région ou territoire.
- **Abonnements au téléphone cellulaire mobile utilisant des cartes prépayées** : Indicateur de la popularité des services prépayés, souvent utilisés dans des pays émergents.
- **Investissements totaux dans les télécommunications (USD)** : Révèle l'effort financier des opérateurs pour améliorer leurs infrastructures, dans une monnaie universelle.
- **Total des abonnements au téléphone cellulaire mobile** : Montre la densité d'utilisateurs de services mobiles.
- **Total des lignes d'accès téléphoniques** : Sert d'indicateur pour les infrastructures filaires disponibles.
- **Total des recettes des télécommunications (USD)** : Reflète les revenus générés par les opérateurs et donne une indication sur la charge financière des clients.
- **Total des voies d'accès de communication** : Mesure la capacité maximale d'accès au réseau.

Nous n'avons pas retenu les autres caractéristiques pour des raisons qui seront évoqué dans la partie II de ce rapport.

3) Justification de leur pertinence

Les caractéristiques sélectionnées sont essentielles pour prédire la demande en bande passante et comprendre les facteurs qui influencent son utilisation :

- **Pays** : Les habitudes de consommation et les infrastructures varient fortement entre les régions. Ce critère permet d'intégrer une perspective géographique à l'analyse.
- **Abonnements prépayés et totaux** : Ces variables reflètent la taille et le comportement des bases d'utilisateurs, influençant directement la demande en bande passante.
- **Investissements et infrastructures (lignes d'accès et voies de communication)** : Ces indicateurs mesurent la capacité et la disponibilité des réseaux pour répondre à la demande croissante.
- **Recettes des télécommunications** : Permet de relier la demande à la rentabilité économique des opérateurs, offrant une perspective sur l'équilibre entre capacité et revenus.

II. PREPARATION DES DONNEES

1) Transformation et nettoyage des données

Concernant la préparation et le nettoyage des données, il est important de savoir que les étapes ci-dessous constitue une trame générale mais que plusieurs dataframe ont été construit par la suite pour chacune des méthodes vues lors de cette étude.

Nous avons tout d'abord procédé à un import du dataframe dans python, suite à quoi nous avons utilisé les méthodes .info() et .describe() pour comprendre notre dataframe. En parallèle, nous avons regardé les caractéristiques de ce dataframe sous Excel pour une meilleure visualisation et une compréhension plus rapide. Nous avons constaté que le dataframe contenait environ 8000 lignes et que ces dernières sont sous la forme Pays/Année/Series/valeur observé. Nous avons donc décidé de passer la colonne « Series » en plusieurs colonnes car les valeurs de « Series » correspondent en réalité aux variables auxquelles nous nous intéresserons dans la suite de notre étude. Nous avons donc utilisé la fonction .pivot() de la librairie pandas sur les variables Pays et Années (TIME_PERIOD). Le dataframe se retrouve donc sous la forme d'une observation (ligne) par Pays/Année avec les variables suivantes :

- Pays
- TIME_PERIOD
- Abonnements au téléphone cellulaire mobile utilisant des cartes prépayées
- Investissements totaux dans les télécommunications (pour lignes fixes et réseau mobile cellulaire)
- Investissements totaux dans les télécommunications (pour lignes fixes et réseau mobile cellulaire) USD
- Nombre d'abonnés à la télévision par câble
- Total Internet Protocol (IP) telephone subscriptions
- Total des abonnements au téléphone cellulaire mobile
- Total des abonnements au téléphone cellulaire mobile pour 100 habitants
- Total des lignes d'accès téléphoniques
- Total des recettes des télécommunications
- Total des recettes des télécommunications USD
- Total des voies d'accès de communication
- Total des voies d'accès de communication pour 100 habitants

On constate donc que les variables ont des similitudes et qu'elles comportent dans certains cas des totaux qui sont eux même des totaux des autres variables.

Par la suite, nous avons regardé le nombre de valeur vides par colonne pour en constater l'exploitabilité. On observe que les lignes ne sont que rarement complètes, et ce de manière

aléatoire. Certains pays n'ont pas de valeur pour certaines variables, d'autre pour certaines années, etc… Il faudra donc prendre plusieurs décisions pour pouvoir utiliser nos données.

Nous avons remarqué que les variables « Nombre d'abonnés à la télévision par câble » et « Total Internet Protocol (IP) telephone subscriptions » ont un taux de valeurs vide d'environ 50% ce qui pourrait être exploitable en restreignant notre dataframe sur uniquement certains pays ou certaines années, ou alors en remplissant les valeurs manquantes par des moyennes ou des médianes. Or dans les faits, nous avons décidé de ne pas exploiter ces variables car nous nous concentrons uniquement sur le réseau de communication de type téléphonique. De même pour les variables total des recettes et total des investissements, vous avons deux valeurs à chaque fois, une valeur en usd et une valeur dans la monnaie locale. Nous avons décidé de sélectionner la variable avec la monnaie en valeur usd, car cela nous permet d'avoir un socle commun pour nos analyses.

Par la suite nous avons également constaté que les années inférieur à 2005 ont plus de valeurs vides que les autres années, ce qui est compréhensible car le recensement était plus compliqué par le passé, nous avons donc décidé de restreindre notre jeu de données aux années supérieures à 2004.

On se retrouve par la suite avec un dataframe avec en moyenne 5% de valeurs vides sur lesquelles il n'y a pas de patern particulier. Il est donc assez complexe de remplir nos valeurs vides.

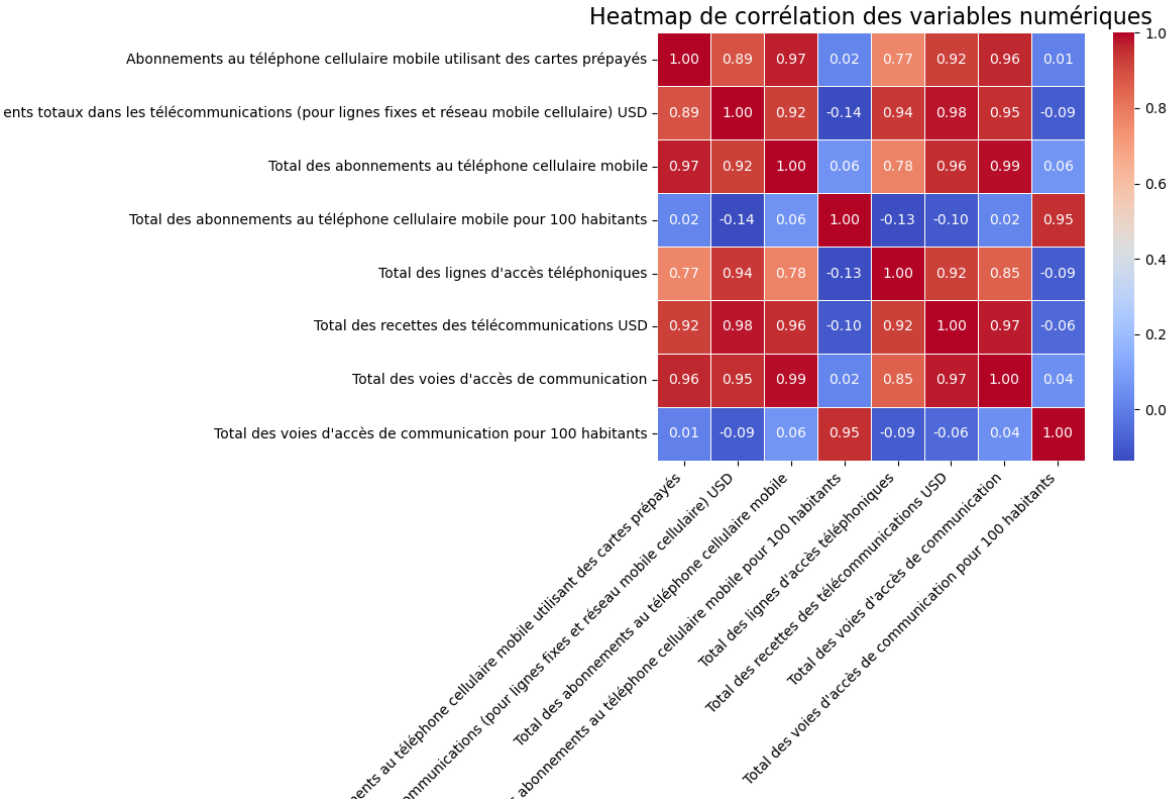
Résumé statistique des variables

	TIME_PERIOD	Abonnements au téléphone cellulaire mobile utilisant des cartes prépayés	totaux dans les télécommunications (pour lignes fixes et réseau mobile cellulaire) USD	Total des abonnements au téléphone cellulaire mobile	Total des abonnements au téléphone cellulaire mobile pour 100 habitants	Total des lignes d'accès téléphoniques	Total des recettes des télécommunications USD	Total des voies d'accès de communication	Total des voies d'accès de communication pour 100 habitants
count	577.00	577.00	577.00	577.00	577.00	577.00	577.00	577.00	577.00
mean	2009.19	13300026.41	5743544563.71	34503448.14	108.68	12751415.01	35459429402.08	54300598.13	163.38
std	5.37	20436038.13	13822198087.43	53232046.77	32.80	24454194.52	85752200727.15	85017888.12	38.78
min	2000.00	63000.00	15324074.00	215000.00	14.25	75716.00	0.00	375869.00	26.74
25%	2005.00	1636272.00	652815042.00	5600000.00	87.39	1235310.00	4240113111.00	8256092.00	138.60
50%	2009.00	4192525.00	1494581333.00	12952605.00	110.10	3646544.00	9065269444.00	19174389.00	169.85
75%	2014.00	12066059.00	5939791667.00	48422470.00	130.97	18708000.00	35463667876.00	77101276.00	189.27
max	2018.00	100099672.00	113301000000.00	421800000.00	187.04	184709483.00	630683000000.00	575565000.00	246.46

En retournant sur le tableau du résumé statistique, on comprend que certaines valeurs sont aberrantes, notamment le 0 minimum dans le total des recettes de télécommunication. Cette valeur est issue de la ligne Japon 2018 pour laquelle la valeur devait surement être vide ou manquante. Nous avons décidé, après avoir constaté les valeurs précédentes des recettes du Japon, de remplacer par la valeur de l'année précédente. Ce choix est mieux qu'une moyenne ou qu'une médiane car les valeurs des données augmentent au fur et à mesure du temps.

On constate que les données ont des valeurs très haute car on parle ici de valeur en millions voire milliards. On y voit de grand changement entre les minimums et maximums car comme dit précédemment, les services liés aux télécommunications ont augmenté de manière exponentielle entre les années 2000 et 2018.

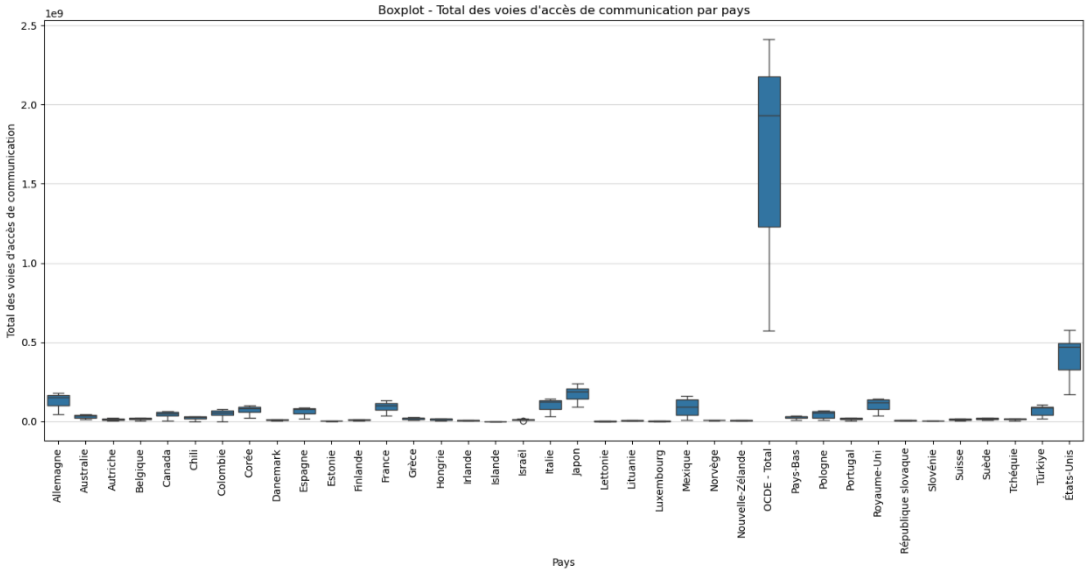
Par la suite nous faisons une heatmap de nos données pour constater la corrélation entre les différentes variables.



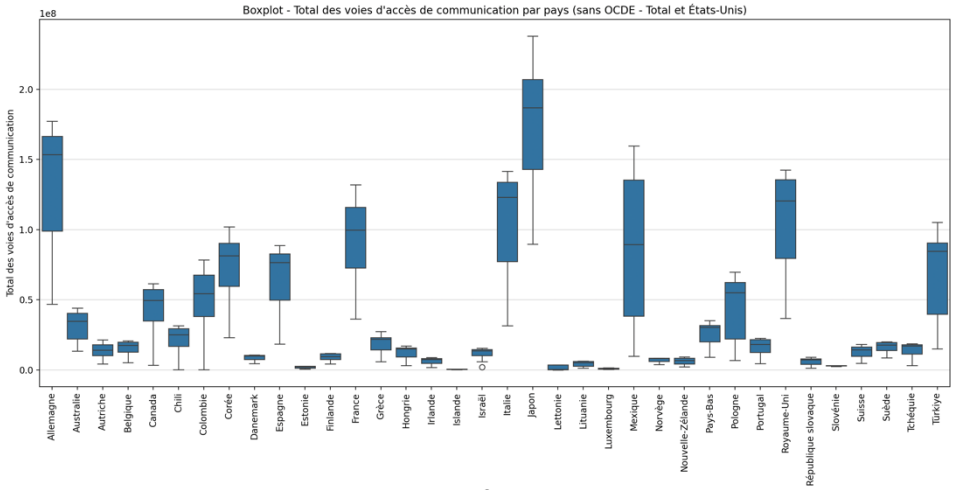
On constate que l'ensemble de nos variables sont fortement corrélés excepté les variables de ratio pour 100 habitants. Toutefois, ce n'est pas pour autant qu'elles ne sont pas exploitables. En restreignant notre dataframe elles peuvent le devenir par la suite.

Etant donné que toutes nos variables sont fortement corrélées (sauf exception), nous nous intéresserons à deux variables en particulier qui feront sens lors de nos analyses, notamment concernant la prédiction de ces dernières. On parle ici du total des voies d'accès de communication et du total des recettes de télécommunication.

Dans un premier temps nous avons fait un boxplot pour chaque pays du total des voies d'accès de communication.



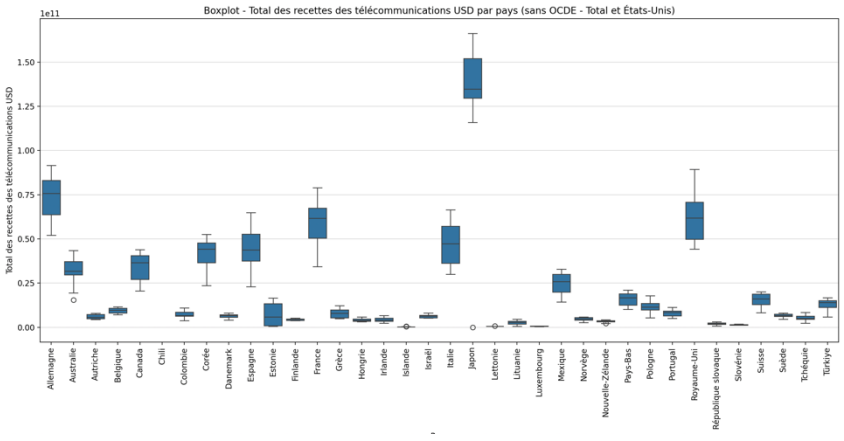
On constate que les valeurs pour ocde total se démarquent des autres, ce qui est logique car cela correspond aux totaux des pays. Par ailleurs les valeurs des Etats-Unis sont très hautes mais cela fait sens car les Etats-Unis regroupent énormément de personnes et sont très développés par rapport aux autres pays du dataframe. Pour la suite des visualisations, nous allons les omettre tout en sachant qu'ils ont des valeurs beaucoup plus élevées que les autres pays. Nous obtenons donc les graphiques suivants pour le total des voies d'accès de communication et pour les recettes :



On constate que les pays développés on des valeurs plus hautes que les pays en voie de développement Par exemple les pays comme Italie Allemagne et France se démarquent. On notera aussi la présence du Japon qui est le pays avec le plus de voies d'accès aux télécommunications. Par ailleurs on constate que certains pays se développent assez rapidement. Par exemple le Mexique a un boxplot très étendu ce qui veut dire que les valeurs étaient assez basses durant les années 2000 mais qu'elles ont fortement augmenté au fil du temps jusqu'à rattraper les pays développés. Attention toutefois à la surinterprétation. En effet on voit sur la Suisse par exemple se trouve dans le bas du panier, or cela fait sens car

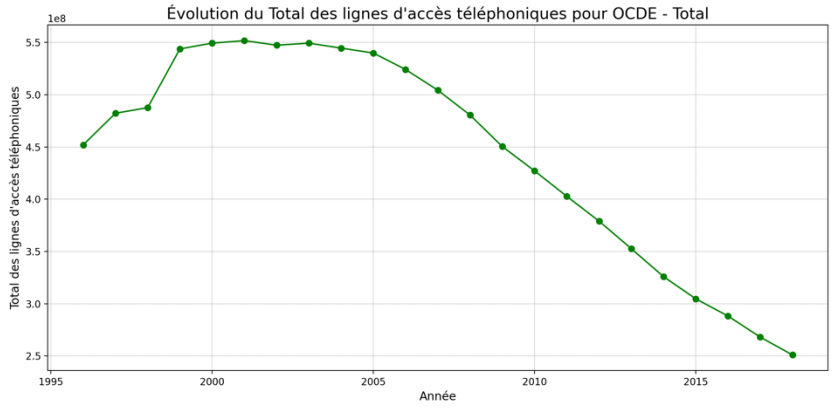
c’est un petit pays. Plus le pays est grand, plus il est probable que le total des voies d’accès de communication soit élevé étant donné que cela est proportionnel.

On retrouve plus ou moins le même phénomène pour les recettes de télécommunication

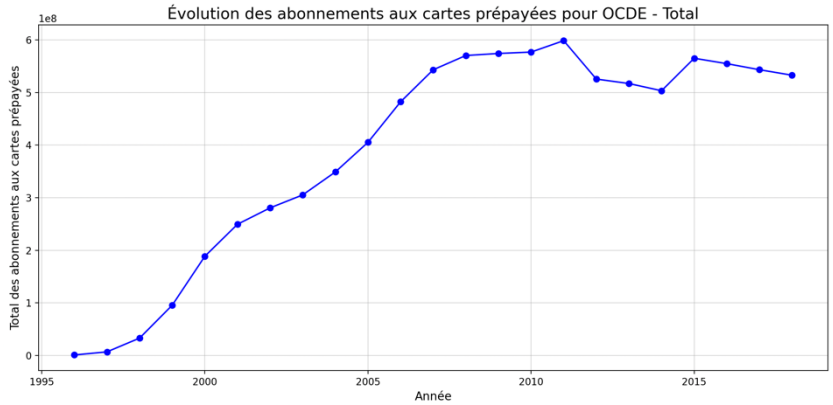


On constate toutefois que cette disparité est un peu moins présente et que certains pays se retrouvent avec des valeurs plus élevées. Cela peut être dû au prix effectif des abonnements dans les pays en question.

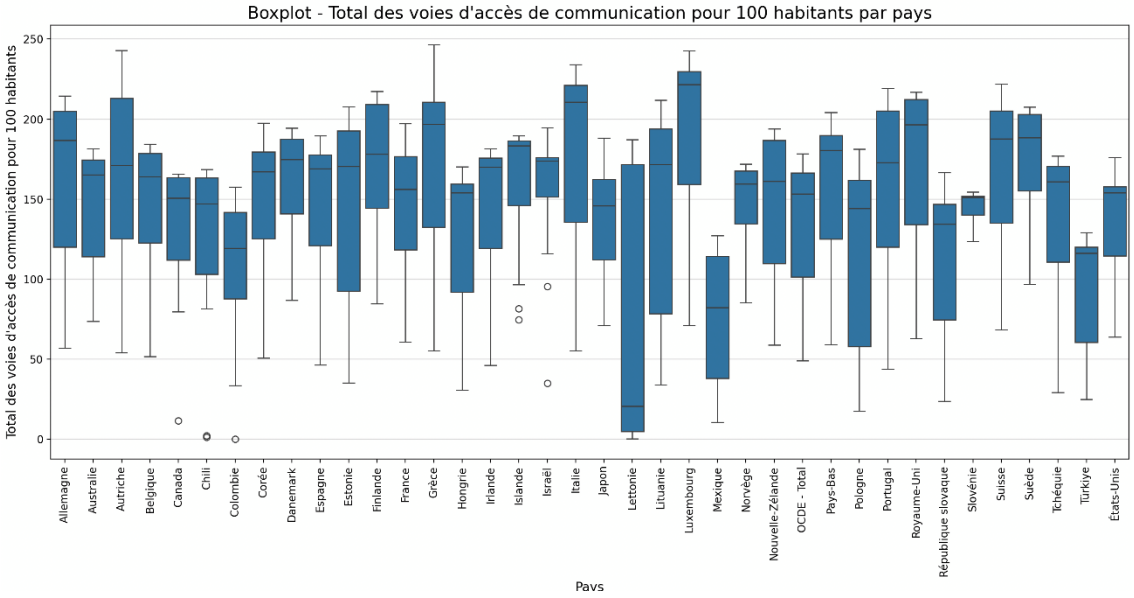
Concernant la variable du total d’accès des lignes téléphonique, nous obtenons le graphique suivant :



Nous nous concentrons ici sur « OCDE Total » car cela regroupe la tendance globale mondiale du nombre d’abonnements aux téléphones fixes. On constate qu’il y a une nette diminution des lignes fixes au fur et à mesure des années.



A l’inverse, pour le nombre de cartes prépayés, on constate que la tendance est globalement haussière mais que les abonnements stagnent et se voient même baisser vers les années 2011.



Cette analyse met en évidence des écarts notables dans l'accès aux télécommunications pour 100 habitants. Dans certains pays développés comme le Japon ou les États-Unis, le ratio approche les 2, ce qui traduit une infrastructure avancée et une large disponibilité des moyens de communication. À l'inverse, dans des pays moins développés comme le Chili ou la Colombie, ce ratio est souvent proche de 1, voire inférieur, reflétant des limitations structurelles et un accès plus restreint aux télécommunications. Ces résultats illustrent l'impact du développement économique sur la connectivité des populations.

2) Explication des différents DataFrames conçus

Le fichier Data_to_export.py nous sert à construire nos différents dataframes pour nos différents modèles.

- 01_02_Data.csv : dataframe moyenné avec les valeurs depuis 2005 utilisé pour Kmeans et HCA
- 03_04_05_06_Dataframe_reglin.csv : dataframe de toutes les données sans les valeurs vides (NA) pour les régressions linéaire simple, multiple et logistique ainsi que knn
- 04_Dataframe_reglin_1.csv : dataframe moyenné pour la seconde évaluation des modèle de régression et knn
- 07_08_Dataframe_integral.csv : dataframe utilisé pour ARIMA et LSTM

III. METHODOLOGIES APPLIQUEES ET RESULTATS

1) Algorithmes de regroupement

Problématique : Quels pays nécessitent une priorité en termes d'investissements pour éviter une saturation de leurs infrastructures de télécommunications ?

⇨ Identifier les clusters combinant un fort taux d'abonnements avec des investissements télécoms faibles.

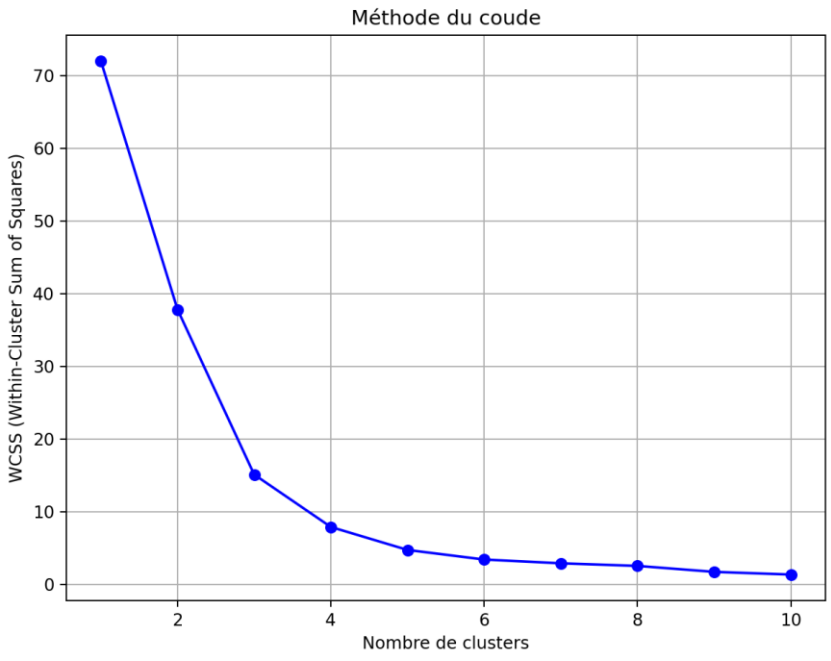
a) K-Means

K-Means est une méthode de clustering pour regrouper les données en fonction de leurs similarités. Dans ce projet, il permet de regrouper les pays ayant des profils proches en termes d'abonnements mobiles et d'investissements télécoms.

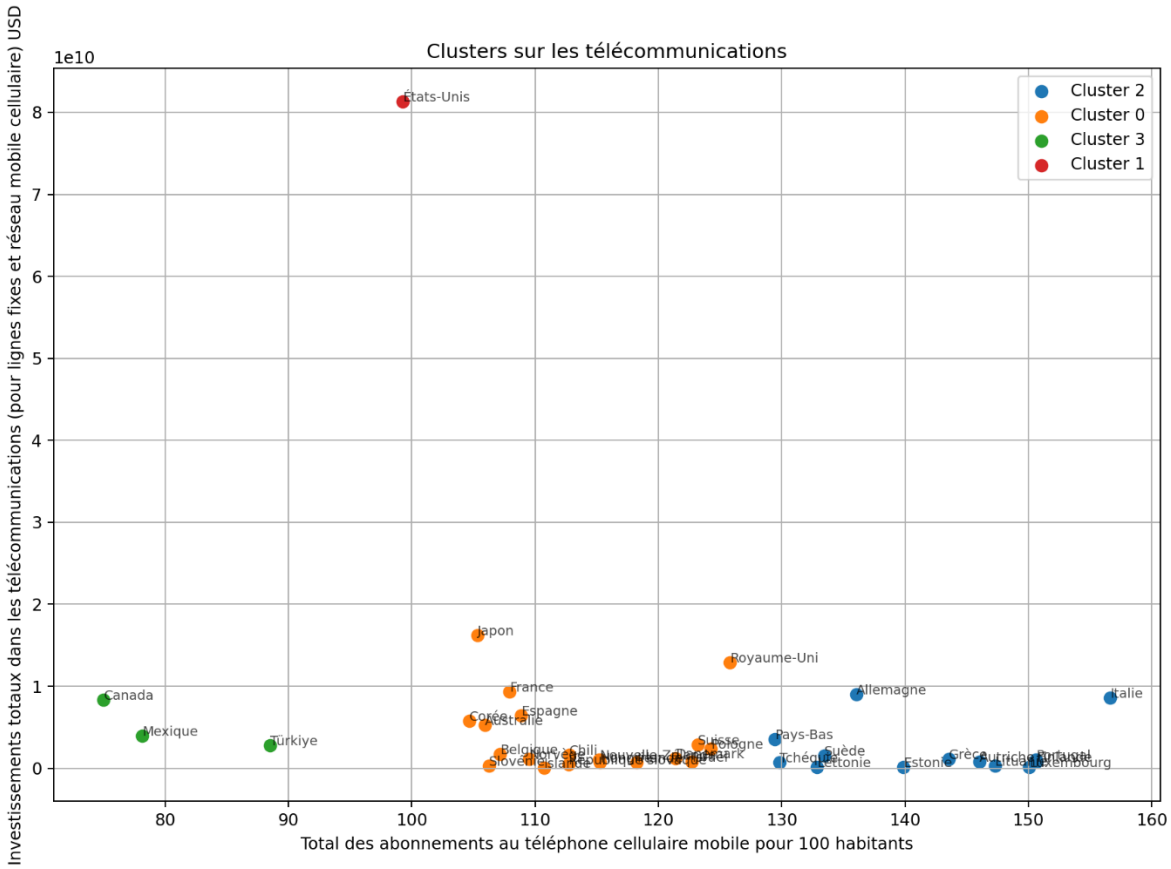
Le fait d'avoir des clusters va nous permettre d'identifier par des groupes pour qu'on puisse proposer des solutions d'optimisation de la bande passante plus adaptée à chaque cluster.

Les colonnes utilisées pour le Kmeans sont : *Total des abonnements au téléphone cellulaire mobile pour 100 habitants* et *Investissements totaux dans les télécommunications (pour lignes fixes et réseau mobile cellulaire) USD*.

Les données ont été standardisées afin de réduire les écarts d'échelle, et le nombre optimal de clusters a été fixé à 4 selon les résultats de la méthode du coude.



Nous avons appliqué la méthode du coude pour voir le nombre de cluster idéal à appliquer à notre algorithme (ici, 4).



Avec ce Kmeans, on peut voir qu'il y a 4 groupes qui se distinguent. Les Etats-Unis qui sont seule dans un cluster car ils font des investissements astronomiques. Ensuite, on retrouve 3 clusters distincts :

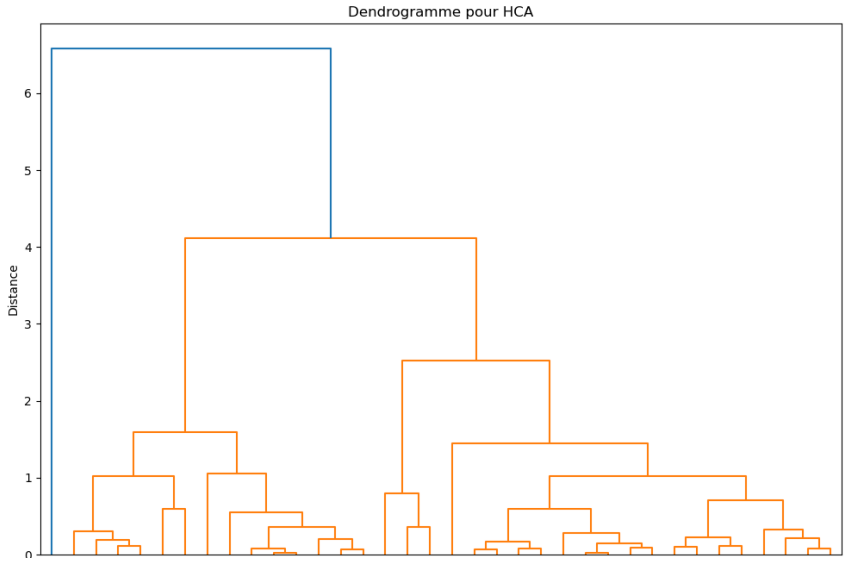
- Les pays où l'investissement est faible et le nombre d'abonnement est faible (Canada, Mexique, …)
- Les pays où l'investissement est un peu plus élevé et le nombre d'abonnement est modéré (Japon, Royaume-Uni, France, …)
- Les pays où l'investissement est faible mais par contre où le nombre d'abonnement est très importante. C'est dans les pays de ce cluster là qu'il faut faire des actions pour changer la tendance.

b) Analyse Hiérarchique des Clusters (HCA)

L'Analyse Hiérarchique des Clusters (HCA) est une méthode utilisée pour identifier les relations hiérarchiques entre les données des différents pays. Elle permet de compléter l'analyse non hiérarchique, notamment comme K-Means effectué dans la partie d'avant.

Les colonnes utilisées pour le HCA sont : *Total des abonnements au téléphone cellulaire mobile pour 100 habitants* et *Investissements totaux dans les télécommunications (pour lignes fixes et réseau mobile cellulaire) USD*.

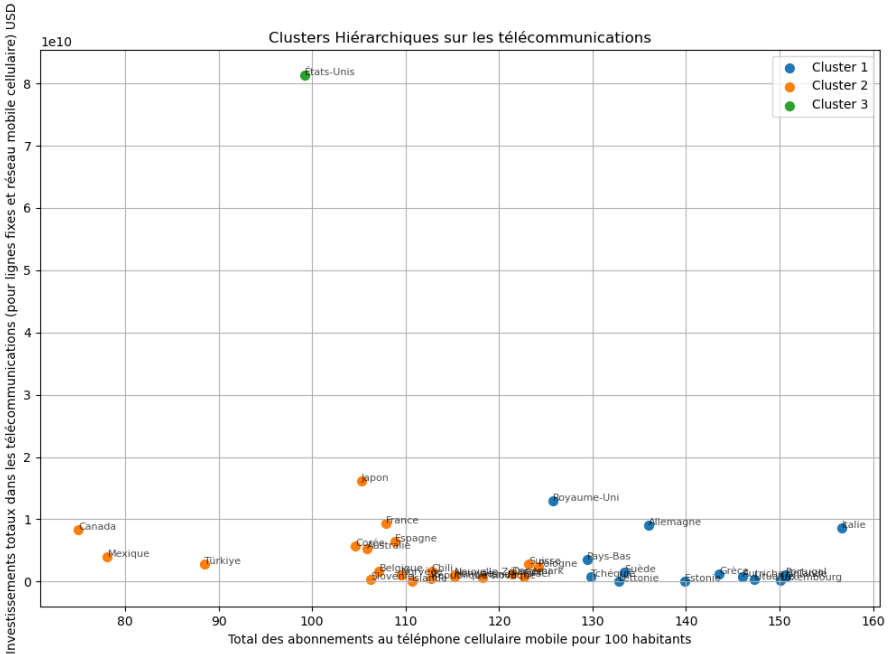
Les données ont été standardisées afin de réduire les écarts d'échelle, et le nombre optimal de clusters a été fixé à 3 selon les résultats du dendrogramme. Concernant les paramètres du HCA, nous avons choisi la méthode de liaison complète qui calcule les distances maximales entre les points d'un cluster.



Nous avons décidé ici de garder 3 comme paramètre « t » pour le cluster du HCA.

On se retrouve avec un résultat équivalent au Kmeans mais avec 3 clusters.

Pour répondre à la problématique, les pays contenus dans le cluster bleu nécessitent une priorité en investissement pour éviter de saturer leurs infrastructures.



2) Modèles de classification

Problématique : Comment prédire les recettes des télécommunications (faibles ou élevées) à partir des caractéristiques des infrastructures et abonnements, afin d'optimiser le budget des investissements futurs et garantir une croissance durable ?

a) K-Nearest Neighbors (KNN)

Le K-Nearest Neighbors (KNN) est un algorithme de classification supervisé qui attribue une classe à un point en fonction des classes majoritaires de ses voisins les plus proches, calculées selon une mesure de distance (euclidienne, ward,etc) .

Ici nous allons chercher à créer un modèle permettant de prédire si, en fonction des données du pays, la rentabilité se voit élevée ou faible. Cette information nous permettra par exemple de savoir si les investissements futurs seront amortis et le cas échéant de réduire les investissements pour ne pas avoir un modèle économique peu durable dans le temps. Dans le cas contraire, cela nous permettra d'investir davantage car les couts seront amortis plus rapidement.

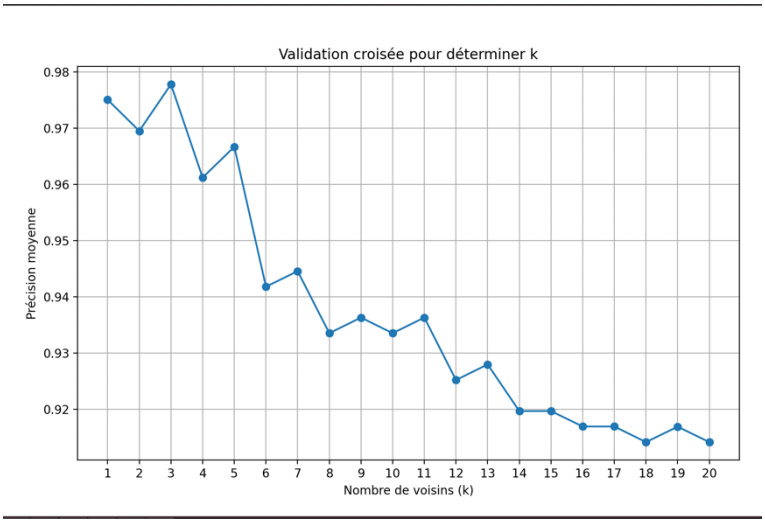
Nous allons donc entrainer notre modèle sur l'ensemble des données nettoyés disponible.

La variable d'intérêt ici est « Total des recettes des télécommunications USD » que nous allons diviser en 2 catégories, faibles ou élevées.

Étant donné que notre jeu de donnée est peu volumineux, nous nous accorderons le droit de sélectionner le plus de variables explicatives possibles tout en gardant une certaine pertinence logique. Nous allons donc sélectionner :

- "Investissements totaux dans les télécommunications (pour lignes fixes et réseau mobile cellulaire) USD"
- "Abonnements au téléphone cellulaire mobile utilisant des cartes prépayés"
- "Total des abonnements au téléphone cellulaire mobile"
- "Total des lignes d'accès téléphoniques"
- "Total des voies d'accès de communication"

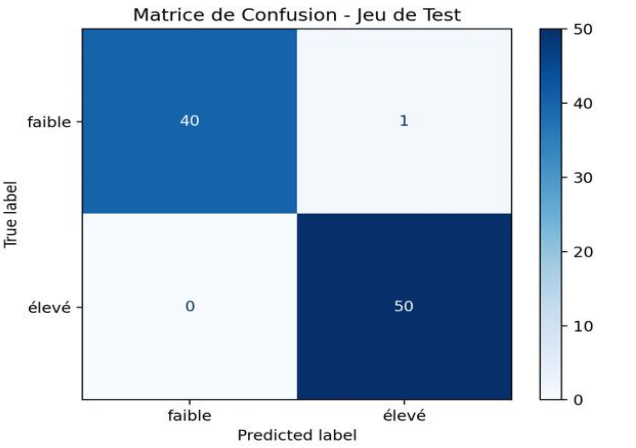
Pour appliquer notre algorithme aux données, nous créons un jeu de données train et un jeu de données test puis nous appliquons l'algorithme classique de knn. Toutefois, afin de trouver le nombre k de voisins les plus proches nous allons utiliser une cross-validation.



On constate premièrement que peu importe le nombre k, l'accuracy est globalement très bonne avec une valeur minimum de 90%.

Ici on constate que le k optimal serait k=3 car c'est le k ayant la plus grande accuracy (proche de 98%). Cette valeur de k nous arrange également car en termes de coût de calcul informatique, plus le k est faible plus le temps de calcul est diminué.

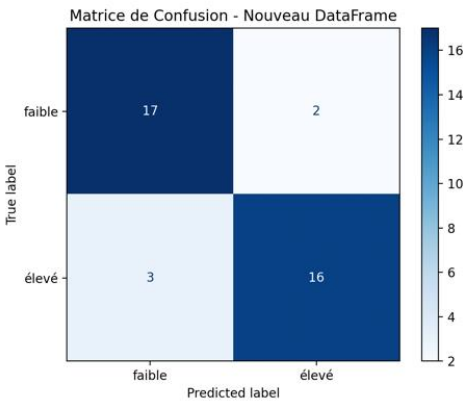
Après avoir créé notre modèle, nous l'utilisons sur le jeu de données test, nous obtenons la matrice de confusion suivante :



On constate que les résultats sont très satisfaisants, le modèle a réussi à bien prédire tous les pays ayant réellement un rendement élevé (50). Toutefois il s'est trompé sur un pays en le classifiant comme élevé alors qu'il était en réalité un pays avec faible rendement.

Pour effectuer un deuxième test, nous allons essayer de prédire le dataframe avec la moyenne des valeurs par pays (le dataframe utilisé pour les méthodes de clustering précédemment).

Nous obtenons la matrice de confusion suivante :



On constate que le modèle à fait légèrement plus d'erreurs qu'avec le set de test, mais l'accuracy reste très bonne avec une valeur de 87% de bonnes prédictions. Ces écarts peuvent s'expliquer par la manière dont les données ont été modifiés (moyenne) et le fait que le seuil de différentiation de notre target ne soit pas sensiblement la même.

b) Régression Logistique Binaire

La régression logistique est un algorithme de classification supervisé qui prédit la probabilité qu'une observation appartienne à une classe en modélisant une relation linéaire entre les variables explicatives et le logarithme du rapport des probabilités (log-odds). Elle attribue une classe en fonction d'un seuil (par défaut 0,5) appliqué aux probabilités prédites.

Ici dans le même principe que KNN, nous allons chercher à prédire la classe de rentabilité. Nous pourrons par la suite comparer les prédictions de KNN et de la régression logistique, ce qui nous permettra de choisir le modèle le plus adapté à nos données et à notre problématique.

Dans l'idéal, nous aurions pu utiliser la méthode stepwize, forward ou backward pour la sélection de nos variables. Toutefois, les données étant peu volumineuses que ce soit en nombre de variables ou d'observations, nous permettent de sélectionner les mêmes variables que précédemment avec KNN. Nous utiliserons donc exactement les mêmes données d'entrainement (observations et variables) ainsi que la même méthodologie d'évaluation de modèle à savoir une prédiction sur les données test et sur les données des algorithmes kmeans et hca (moyennes des observations par années).

Après avoir appliqué notre modèle sur nos données d'entrainement, nous obtenons les mesures suivantes :

Rapport de classification (jeu de test) :

	precision	recall	f1-score	support
faible	0.82	1.00	0.90	41
élevé	1.00	0.82	0.90	50
accuracy			0.90	91
macro avg	0.91	0.91	0.90	91
weighted avg	0.92	0.90	0.90	91

On constate que les mesures de précision, rappel et F1-score montrent que le modèle est globalement performant avec une accuracy de 90 %. Cependant, le rappel pour la classe "élevé" (0.82) indique que certaines valeurs "élevé" ne sont pas correctement détectées, ce qui pourrait être amélioré.

Concernant l'interprétation de notre modèle, nous obtenons les coefficients suivants :

=== Coefficients de la Régression Logistique ===

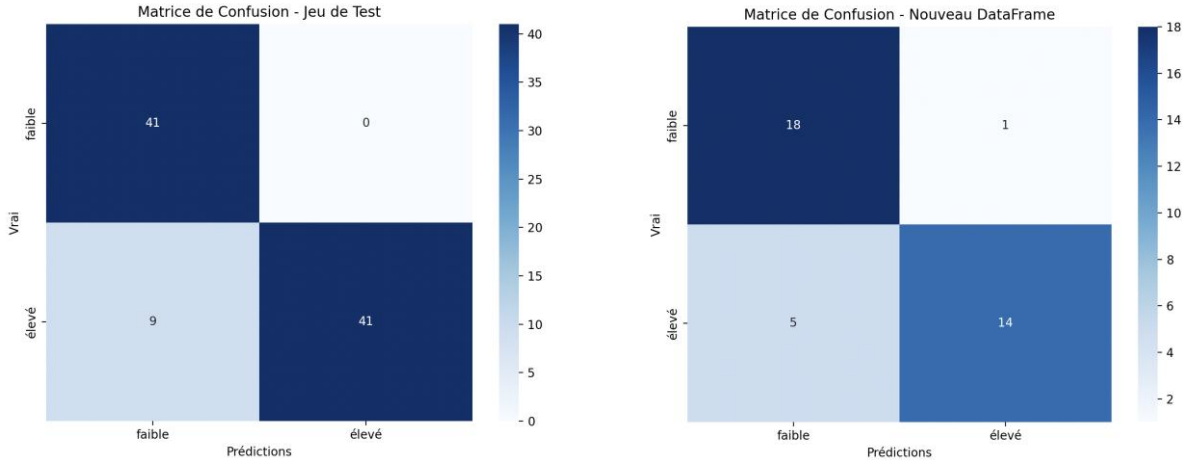
	Variable	Coefficient
0	Abonnements au téléphone cellulaire mobile uti...	2.829377
1	Investissements totaux dans les télécommunicat...	3.265713
2	Total des abonnements au téléphone cellulaire ...	2.381512
3	Total des lignes d'accès téléphoniques	3.385180
4	Total des voies d'accès de communication	2.670031

Par exemple le coefficient de 2.829 pour "Abonnements au téléphone cellulaire mobile utilisant des cartes prépayées" indique qu'une augmentation d'une unité de cette variable augmente significativement la probabilité d'appartenir à la classe "élevé", toutes choses étant égales par ailleurs.

On constate que la variable la plus influente ici est le total des lignes d'accès téléphonique avec le coefficient le plus élevé. Ce qui a du sens dans l'interprétation, plus il y a d'abonnements, plus les recettes sont élevées.

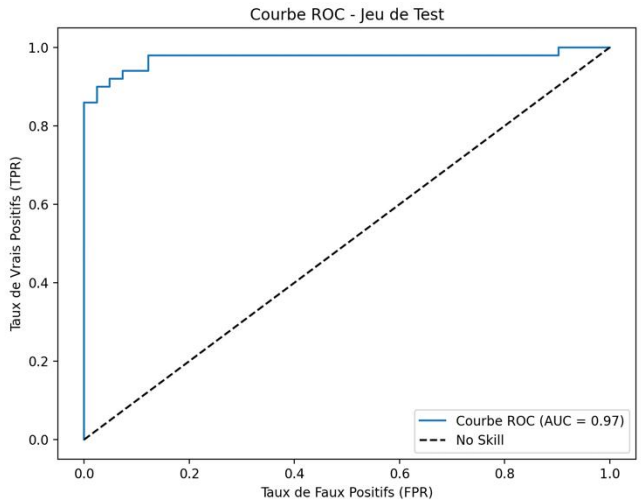
Pour améliorer notre modèle nous aurions pu utiliser une méthode de segmentation des variables continues en variable catégorielles. Nous avons par exemple la méthode des Weight of Evidence (WoE binning). Grâce à cela, nous aurions pu mieux capturer les relations non linéaires entre les variables explicatives continues et la cible, tout en réduisant l'effet des valeurs extrêmes et en améliorant l'interprétabilité des coefficients du modèle.

Concernant les matrices de confusions, nous obtenons respectivement pour le jeu de données de test et des moyennes les résultats suivants :



On constate que les résultats sont globalement similaires à la méthode KNN, toutefois, à peu de choses près, la régression logistique à des résultats moins bons dans l'ensemble.

Enfin nous utilisons une courbe de ROC pour mieux comprendre les prédictions :



La courbe ROC montre une excellente performance du modèle, avec une AUC de 0.97, ce qui indique que le modèle distingue très bien les classes "faible" et "élevé". La courbe s'approche du coin supérieur gauche, reflétant un bon compromis entre le taux de vrais positifs et le taux de faux positifs.

En conclusion, les deux modèles ont su parfaitement répondre à la problématique, avec des prédictions très fiable dans l'ensemble. Si l'on devait choisir une méthode, il faudrait opter en premier lieu pour les K plus proches voisins qui obtiennent de meilleurs résultats.

A l'avenir, les données d'un pays nous permettrons de savoir la rentabilité de ce dernier. Grâce à cela, nous pourrons savoir s'il est judicieux d'investir en s'endettant, tout en sachant que les revenus seront élevés et donc accélérer le processus de transformation des infrastructures de télécommunication.

3) Modèles de régression

Problématique : Comment prédire précisément le montant des recettes des télécommunications afin de mieux planifier les investissements et favoriser une gestion budgétaire efficace ?

a) Régression Linéaire Simple

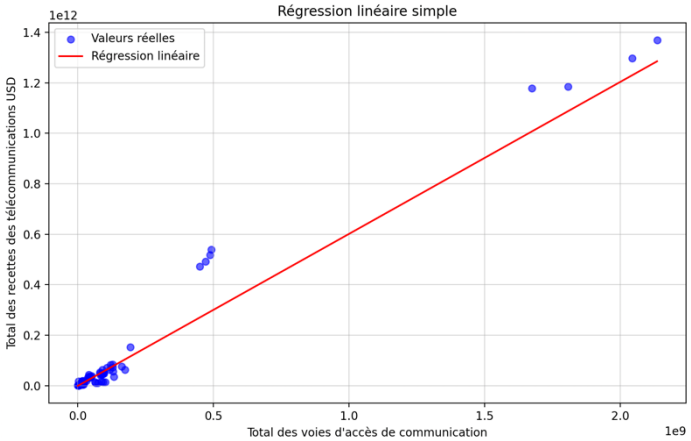
La régression linéaire simple modélise la relation entre une variable cible et une seule variable explicative à l'aide d'une équation linéaire (y = mx + b). Elle est utilisée pour prédire la valeur de la cible en fonction de la variable explicative, en supposant une relation linéaire entre les deux.

Après avoir effectué une classification pour prédire la recette (faible ou élevée), nous nous intéresserons au montant de cette recette. En effet, pour pouvoir être plus précis lors de notre gestion des flux d'argent, nous souhaitons avoir une estimation des recettes dans le but d'investir sans dépenser plus que le montant des recettes prévues

Pour la régression linéaire simple, nous allons tout d'abord essayer de prédire le montant des recettes (« Total des recettes des télécommunications USD ») en fonction de la variable "Total des voies d'accès de communication". Ces deux variables sont fortement corrélées positivement (plus de 0,90) ce qui veut dire que la régression linéaire simple devrait avoir de bons résultats.

Pour entrainer notre modèle, nous allons utiliser le jeu de données complet regroupant toutes les années et tous les pays. En effet, plus notre dataset est grand, plus le modèle sera précis.

Nous allons ensuite prédire le jeu de données test ainsi que les données moyennées pour vérifier si les prédictions sont fiables et si le R2 est convenable et cohérent avec les résultats du modèle. Après avoir appliqué l'algorithme sur notre set de train, nous l'appliquons sur les données de test, nous obtenons les métriques ainsi que la droite de régression suivante :



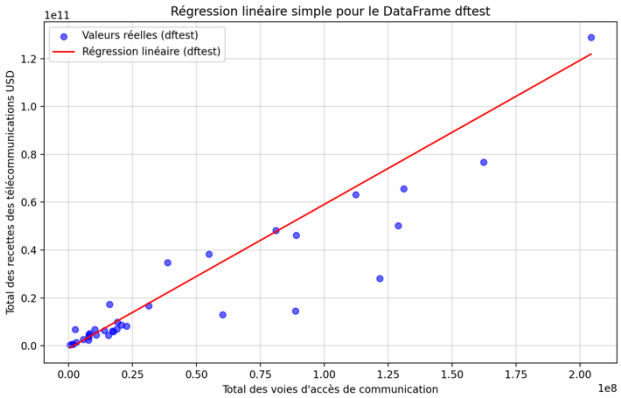
On constate que la droite de régression est assez pertinente pour les points situés en bas à gauche du graphique, mais qu'ils le sont moins pour les quelques valeurs situées au centre du graphique.

Cette disparité s'explique car certains pays ont des valeurs très élevés ce qui les rends plus complexes à prédire.

Concernant les métriques, nous pouvons simplement dire que le modèle présente une bonne performance globale avec un R^2 de 0.9605, indiquant que 96 % de la variance des recettes est expliquée par le modèle. Cependant, le RMSE élevé reflète l'impact des valeurs absolues importantes dans les données, ce qui pourrait être amélioré par une transformation des variables ou une normalisation des unités. Toutefois nous souhaitons obtenir un montant réel est concret des recettes, d'où le fait que nous n'avons pas normalisé ici.

```
Coefficient de détermination (R²) : 0.9605
Mean Squared Error (MSE) : 2.8635964389430463e+21
Root Mean Squared Error (RMSE) : 53512582062.0071
RMSE normalisé : 0.545
```

Sur le nouveau jeu de données (moyennes des valeurs par pays), nous obtenons les résultats suivants :




```
Évaluation sur le DataFrame de test (dftest) :
Coefficient de détermination (R²) : 0.7909
Mean Squared Error (MSE) : 1.64831709978683e+20
Root Mean Squared Error (RMSE) : 12838680227.293
```

On constate que la relation est assez bonne au départ mais qu'elle se dégrade légèrement sur les valeurs centrales. Toutefois il faut garder à l'esprit que l'échelle est en 1^e11, ce qui peut biaiser les écarts réels de prédiction.
Concernant les métriques, elles sont un peu moins convaincantes avec un R2 de 0,79, ce qui reste un bon modèle dans l'ensemble.

Une amélioration possible serait de chercher si une régression polynomiale pourrait convenir aux données.
Nous avons vu que la variables Total des voies d'accès de communication prédisait assez bien nos données. Toutefois, nous parlons ici d'une régression linéaire simple. L'ajout d'autres variables explicatives pourrait nous permettre de faire des prédictions encore plus précises.

b) Régression Linéaire Multiple

En suite de la régression linéaire simple et sur le même principe, nous allons ajouter des variables explicatives à notre modèle pour prédire avec plus d'efficacité notre modèle.

Nous allons, au même titre que la régression logistique ou les KNN, choisir les variables explicatives suivantes :

- "Investissements totaux dans les télécommunications (pour lignes fixes et réseau mobile cellulaire) USD"
- "Abonnements au téléphone cellulaire mobile utilisant des cartes prépayés"
- "Total des abonnements au téléphone cellulaire mobile"
- "Total des lignes d'accès téléphoniques"
- "Total des voies d'accès de communication"

Cela nous permettra de comparer les résultats. A savoir que toutes ces variables sont fortement corrélées avec notre variable cible « Total des recettes des télécommunications USD ».

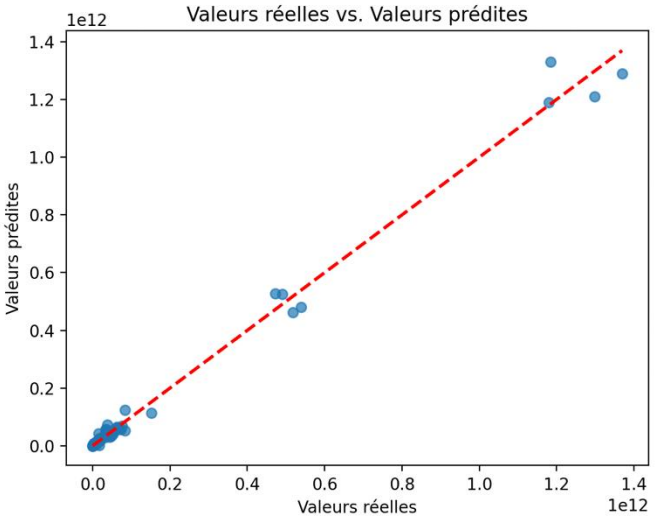
Après avoir créé notre modèle sur les données à l'aide de la méthode LinearRegression(). Nous obtenons les coefficients de modèle suivant :

```
=== Coefficients du modèle ===
                                              Variable  Coefficient
0  Investissements totaux dans les télécommunicat...     6.173388
1            Total des lignes d'accès téléphoniques   158.313626
2          Total des voies d'accès de communication    19.819791
Intercept : 59652222.6204
```

Le coefficient associé au "Total des lignes d'accès téléphoniques" (158.31) indique qu'une augmentation d'une unité de cette variable entraîne une augmentation moyenne de 158.31 unités des recettes des télécommunications, toutes choses égales par ailleurs. Ce résultat montre que cette variable a un impact majeur sur les recettes.

Le modèle met en évidence que les lignes d'accès téléphoniques jouent un rôle prépondérant dans la génération des recettes. Bien que les investissements et les voies d'accès de communication aient un impact significatif, leur effet reste plus modéré.

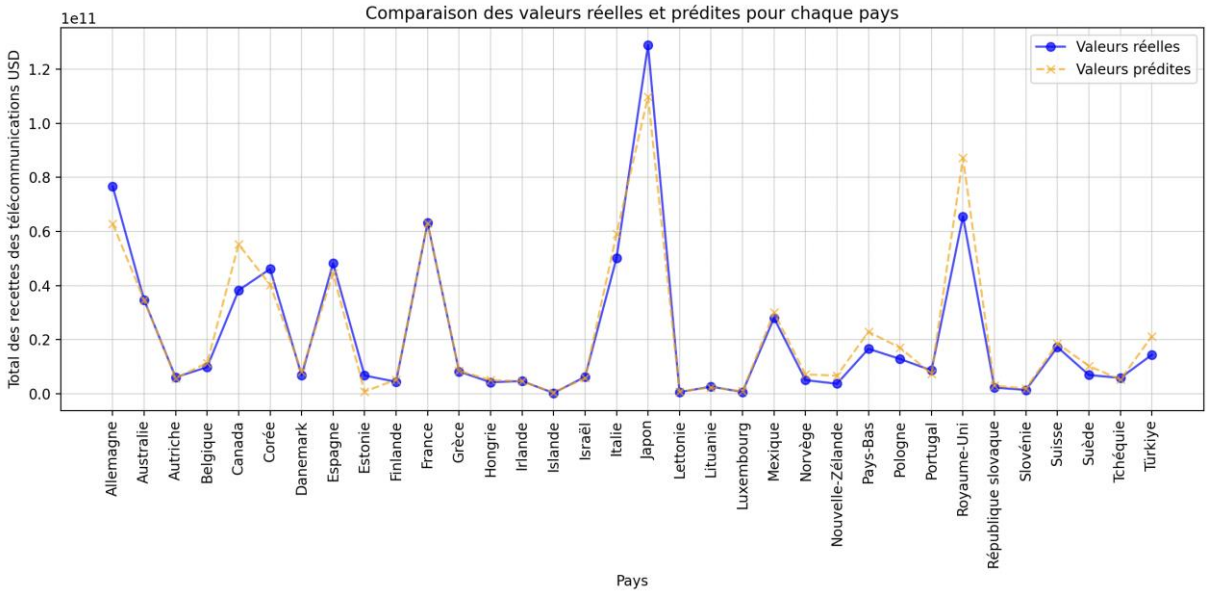
Par la suite nous obtenons les prédictions suivantes :



Le graphique ci-contre indique que le modèle de régression linéaire multiple offre des prédictions précises, bien qu'il y ait de légers écarts pour les valeurs les plus élevées. Le r2 du modèle est de 0.99 ce qui est proche de la perfection.

Nous allons par la suite essayer le modèle sur les données moyennées.

Concernant les données moyennées, nous obtenons le graphique suivant ainsi qu'un r2 de 0,96 :



On constate que les prédictions sont excellentes et que le R2 est très bon. La régression linéaire multiple est une réelle amélioration par rapport à la régression linéaire simple et nous permet de prédire avec très peu d'écart la valeur réelle de la recette future.

Concernant les améliorations possibles, nous aurions pu utiliser les méthodes de sélection de variable comme pour la régression logistique, notamment les méthodes stepwize, forward et backward, mais les modèles sont excellents et ne sont pas trop lourd donc cette étape est négligeable pour le moment.

En résumé, le modèle de régression linéaire multiple est un très bon complément aux méthodes de classification qui nous permettent d'avoir un premier aperçu des recettes prévues. Les acteurs économiques pourront mieux gérer les flux d'argent et donc accélérer les processus d'améliorations des infrastructures de télécommunication.

IV. ANALYSE DES SERIES TEMPORELLES

Problématique : Comment prédire l'évolution de la demande en bande passante et identifier les pics potentiels afin d'optimiser l'allocation des ressources réseau ?

⇨ Prédire le total des voies d'accès de communication en fonction du temps

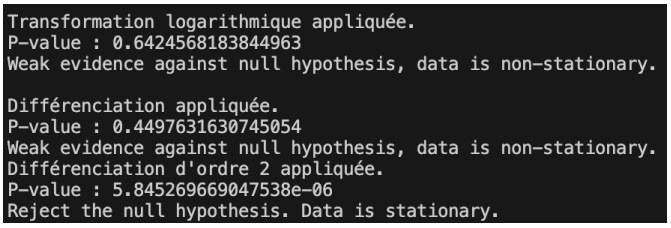
1) Recherche sur ARIMA

ARIMA est une méthode statistique largement utilisée pour modéliser et prédire les séries temporelles. Elle combine trois composants principaux :

- **Auto-Regressive (AR)** : Prend en compte la corrélation entre les observations actuelles et passées.
- **Integrated (I)** : S'occupe de rendre la série stationnaire en différenciant les données (soustraction entre observations successives).
- **Moving Average (MA)** : Modélise les erreurs de prévision comme une combinaison linéaire des erreurs passées.

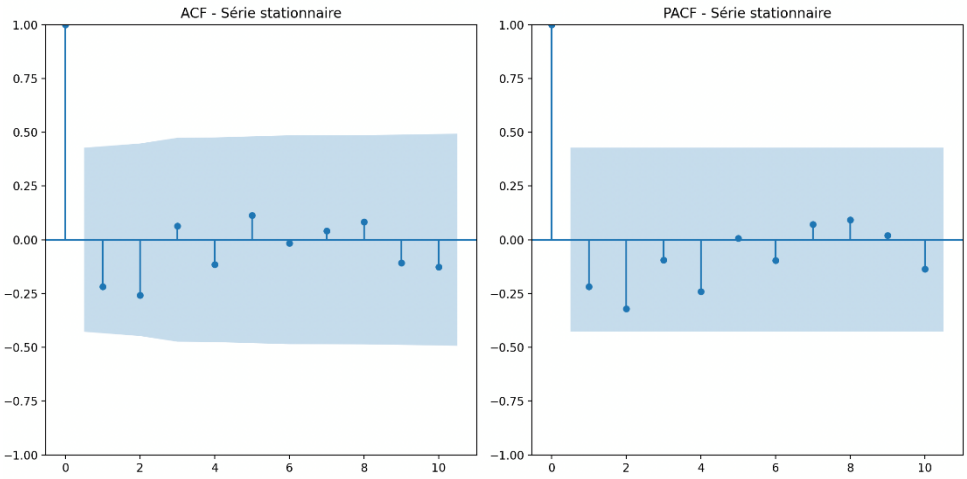
Dans ce projet, ARIMA a été appliqué pour modéliser et prédire le total des voies d'accès de communication des pays de l'OCDE. Les étapes principales ont inclus la transformation des données pour rendre la série stationnaire, la sélection des paramètres du modèle et l'évaluation de la qualité des résidus.

Nous avons d'abord évalué la stationnarité de la série temporelle en appliquant un test ADF.



La série a été transformée avec une échelle logarithmique et plusieurs différenciations successives jusqu'à ce qu'elle devienne stationnaire. On constate ici que notre paramètre d est de 3, car la pvaleur du test devient significative à partir de 3 itérations. C'est-à-dire que notre série est stationnaire après trois log-différenciation.

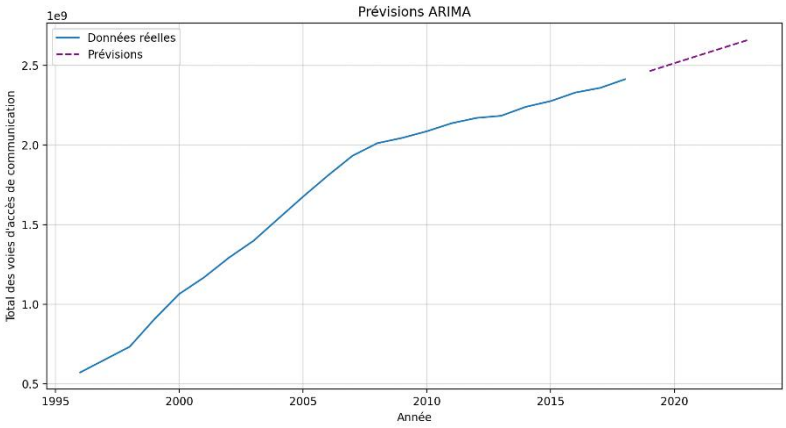
Ensuite, nous avons utilisé les fonctions ACF et PACF pour identifier les valeurs optimales des termes AR (p) et termes MA (q).



Nous avons opté pour p = 1 et q = 1 dans la construction du modèle ARIMA en nous appuyant sur l'analyse des graphiques ACF et PACF. Le lag 1, significatif dans les deux graphiques, indique une dépendance suffisante avec le premier retard. Ce choix permet de proposer un modèle simple tout en capturant les principales dynamiques de la série temporelle.

Nous aurions également pu opter pour les paramètres p = 3 et q = 3, car les lags 2 et 3 présentent encore une certaine significativité. Cependant, cela aurait moins de sens, car ces lags ont un impact beaucoup moins marqué, et un modèle plus complexe pourrait ne pas apporter d'amélioration significative à la qualité des prévisions.

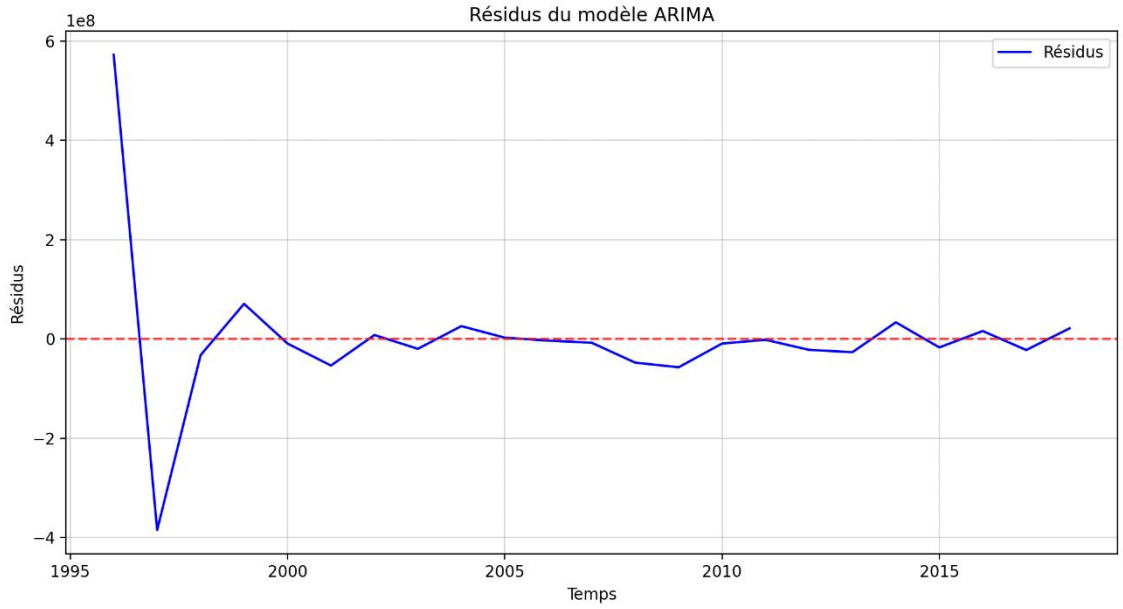
Après avoir construit notre modèle, nous obtenons donc le graphique des prévisions suivant :



On observe que les prévisions pour les 5 années suivant 2018 montrent une augmentation régulière. Cette évolution est cohérente avec les résultats obtenus précédemment à l'aide des méthodes de machine learning, qui suggéraient également une croissance soutenue des voies d'accès de communication.

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.4473	0.162	2.753	0.006	0.129	0.766
ma.L1	−0.6740	0.230	−2.931	0.003	−1.125	−0.223
sigma2	6.27e+14	nan	nan	nan	nan	nan

Concernant la validation du modèle on constate que les pvaleur de ar.l1 et de ma.l1 sont significatives ce qui veut dire que les termes autorégressifs (AR) et de moyenne mobile (MA) du modèle ARIMA contribuent significativement à expliquer la variance des données. Cela valide l'adéquation du modèle à la série temporelle.



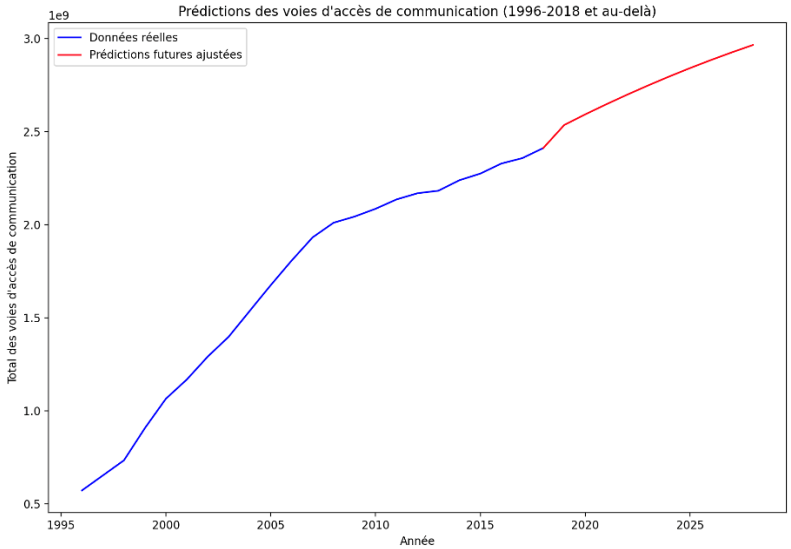
Les résidus du modèle ARIMA montrent une variance relativement constante après l'an 2000, respectant l'hypothèse d'homoscédasticité essentielle à la validité du modèle. Cependant, un problème est observé avant 2000, où les résidus affichent une forte amplitude et des variations importantes. Cela est peut-être dû au valeurs qui sont assez différentes pour ces périodes.

2) Recherche sur LSTM

LSTM (Long Short-Term Memory) est un type de réseau de neurones conçu pour travailler avec des données séquentielles, comme les séries temporelles. Elle est utilisé pour prédire des tendances futures. Dans notre cas par exemple, prévoir la demande de bande passante en fonction de son évolution dans le temps, ou encore prévoir des pics de consommation. Son point fort est de comprendre les relations à long terme dans les données. Ce modèle est capable de se souvenir des informations importantes sur une longue période afin de pouvoir les réutiliser pour faire des prédictions précises.

Dans notre cas par exemple, nous avons utilisé LSTM pour prévoir la demande de bande passante en fonction de son évolution dans le temps, ou encore prévoir des pics de consommation.

Pour développer le modèle, nous avons utilisé les concernant le total des voies d'accès de communication des pays membres de l'OCDE. Après avoir normalisé les données à l'aide de la méthode MinMaxScaler, des fenêtres temporelles de 20 pas ont été créées afin de structurer les données pour la prédiction. Ces fenêtres permettent au modèle de comprendre les variations sur une période définie, offrant ainsi une meilleure capacité de prédiction.



La courbe bleue représente les données réelles entre 1996 et 2018, tandis que la courbe rouge montre les prédictions futures ajustées générées par le modèle.

Les performances ont été évaluées à l'aide des métriques MAE et RMSE, avec des résultats respectifs de 50 041 512 et 50 041 512. Ces valeurs, bien qu'élevées, sont cohérentes avec l'échelle des données, où la dernière valeur réelle était de 2 412 618 090 et la première prédiction future de 2 536 548 864. Cela montre que l'erreur absolue est relativement faible en proportion des valeurs observées.

En conclusion, l'utilisation du LSTM a permis de répondre efficacement à la problématique de la prévision de la demande en bande passante. Ce modèle s'est avéré capable de capturer les dynamiques à long terme des séries temporelles et de fournir des prédictions précises, contribuant ainsi à une meilleure gestion des ressources réseau. Nous avons pu observer une tendance reflétant les besoins croissants en bande passante, en ligne avec l'évolution des infrastructures et l'augmentation des utilisateurs de services numériques.

V. SYNTHESE

1) Synthèse des leçons tirées

Ce projet a permis de mettre en pratique diverses méthodologies d'analyse de données et d'apprentissage automatique appliquées à un problème d'optimisation de la bande passante. Les différents algorithmes utilisés, qu'il s'agisse de clustering, de classification ou de régression, ont apporté des résultats pertinents pour répondre aux différentes problématiques pour chaque type d'algorithme.

Le clustering avec K-Means et HCA a mis en évidence des regroupements entre pays en fonction de leurs investissements et de leur nombre d'abonnements. Ces clusters ont fourni une base pour identifier les zones prioritaires en termes d'investissement.

Les modèles de classification, notamment KNN et la régression logistique, ont permis de prédire la rentabilité des télécommunications en fonction des données infrastructurelles et des abonnements. Ces outils sont utiles pour optimiser les retours potentiels sur investissement.

Les régressions linéaires simple et multiple ont démontré leur pertinence pour prédire les recettes des télécommunications, avec une très bonne précision globale. Elles ont fourni des estimations chiffrées pour planifier les ressources financières.

Enfin, les modèles de séries temporelles ont été particulièrement efficaces pour détecter les tendances et prédire les besoins futurs en infrastructures de communication, avec une interprétation des résultats.

2) Améliorations potentielles

Une limitation importante réside dans le nombre restreint de colonnes présentes dans le jeu de données. Bien que le jeu initial comprenait davantage de variables, la plupart d'entre elles étaient soit incomplètes, soit redondantes, ce qui a conduit à leur exclusion. Cette réduction limite la capacité des analyses à refléter toute la complexité du problème traité. Ajouter des colonnes pertinentes et exploitables permettrait d'améliorer la richesse et la précision des résultats.

Par ailleurs, l'exploration de modèles prédictifs supplémentaires pourrait apporter des avancées significatives. En complément des méthodes déjà appliquées, intégrer des modèles comme « Random Forest » offrirait la possibilité de mieux capturer les relations complexes entre les variables et de renforcer la fiabilité des prédictions.