

Filière : ERP - BI

RAPPORT DE STAGE

Amélioration de la Prise de Décision Basée sur l'Analyse
de Données d'Appels grâce à la Segmentation , la
Classification et la Visualisation chez Tunisie Telecom.



Réalisé PAR : Mabrouk Ayoub
Encadré PAR : Rafik Amri

DU 26-06-2023 JUSQU'AU 26-08-2023

AU SEIN DE : TUNISIE TELECOM

Remerciement

Je tiens à prendre quelques instants pour exprimer ma profonde gratitude envers tous ceux qui ont été à mes côtés et m'ont fourni leur soutien moral inconditionnel tout au long de mon parcours universitaire.

Je tiens également à exprimer ma reconnaissance envers mon encadrant, Mr Rafik Amri, qui a joué un rôle majeur dans mon expérience professionnelle. Son expertise, sa patience et son engagement indéfectible ont été une source d'inspiration pour moi.

Je remercie Tunisie Telecom pour m'avoir offert l'opportunité enrichissante de réaliser mon stage au sein de leur organisation.

Je tiens à exprimer ma profonde gratitude envers tous mes professeurs à Esprit pour la formation en ERP-BI, qui a été une expérience transformative. Cette formation m'a permis d'acquérir des compétences essentielles en gestion d'entreprise et en analyse de données, ouvrant ainsi de nouvelles perspectives professionnelles. Les enseignements de qualité et les connaissances partagées ont été d'une valeur inestimable pour mon développement professionnel.

Sommaire

Remerciement.....	1
Introduction Générale.....	4
Contexte du projet	5
Etude de l'existant.....	6
Problématique.....	7
Besoins Fonctionnels.....	8
Besoins Non Fonctionnels.....	9
Choix des plateformes.....	10
CRISP	11
a- Phase 1 : Compréhension des Objectifs Métier	11
b- Phase 2 : Compréhension des Données :	12
c- Phase 3 : Préparation des Données :	14
d- Phase 4 : Modélisation.....	15
e- Phase 5 : Evaluation :	24
Visualisation avec Power Bi.....	25
a- Page 1 : General	25
b- Page 2 : Revenus	26
c- Page 3 : Questions.....	28
Conclusion.....	29

Introduction Générale

Au sein du secteur en constante évolution des télécommunications, Tunisie Telecom occupe une place prépondérante en tant que principal fournisseur de services de communication en Tunisie. En tant qu'ingénieur informatique en business intelligence au sein de cette entreprise, j'ai eu l'opportunité de travailler sur un projet passionnant visant à explorer les données des appels clients. Ces données riches et variées offrent une opportunité exceptionnelle d'optimiser les stratégies de gestion de la clientèle et d'améliorer la prise de décision grâce à l'analyse approfondie et à la modélisation prédictive.

Le principal objectif de ce stage était d'effectuer une analyse approfondie des données d'appels des clients de Tunisie Telecom. Cette analyse avait pour but de segmenter la clientèle en groupes pertinents, permettant ainsi une personnalisation plus précise des services et une meilleure compréhension des besoins spécifiques de chaque segment. De plus, nous avons cherché à mettre en place une classification prédictive en utilisant la régression logistique, afin d'anticiper les comportements futurs des clients et d'ajuster nos stratégies en conséquence. Et finalement créer un tableau de bord interactif en utilisant Power Bi . Cette initiative revêt une importance capitale, car elle contribue à renforcer la compétitivité de Tunisie Telecom sur le marché dynamique des télécommunications.

Contexte du projet

L'industrie des télécommunications a connu une transformation radicale au cours des dernières décennies, passant de simples services de communication à des plateformes intégrées offrant des données, de l'Internet et des services numériques avancés. Dans cet environnement en constante évolution, les entreprises de télécommunications doivent s'appuyer sur des données pour prendre des décisions éclairées et rester compétitives.

Tunisie Telecom occupe une place de choix en tant que principal fournisseur de services de communication en Tunisie. En tant qu'entreprise emblématique du secteur, elle fait face à des défis uniques liés à la concurrence accrue, aux demandes diversifiées des clients et à l'évolution rapide des technologies. Dans ce contexte, l'exploitation intelligente des données est devenue un impératif pour répondre aux attentes des clients et façonner l'avenir de l'entreprise.

La segmentation clientèle est une stratégie essentielle dans le domaine des télécommunications. Elle consiste à diviser la base de clients en groupes homogènes en fonction de critères tels que le comportement d'utilisation, la valeur client, la démographie, etc. Cette approche permet aux entreprises de mieux comprendre les besoins spécifiques de chaque groupe et de concevoir des offres sur mesure pour maximiser la satisfaction et la fidélité des clients.

Mon stage au sein de cette initiative a été une opportunité exceptionnelle d'acquérir des compétences pratiques dans le domaine de l'analyse de données . J'ai eu la chance de participer à des phases du projet, comme la segmentation de clients , la classification prédictive et la visualisation dans un table de bord interactif en travaillant avec des technologies telles que Python et Power Bi.

Ce stage m'a permis d'approfondir ma compréhension du domaine Business Intelligence .

Etude de l'existant

L'étude de l'existant a commencé par la collecte exhaustive des données d'appels provenant de différentes sources au sein de Tunisie Telecom. Ces données comprenaient des informations telles que les offres des clients, les durées d'appel, les différents revenus etc. Une fois les données rassemblées, une étape cruciale de préparation a été entreprise pour nettoyer les incohérences, traiter les valeurs manquantes et normaliser les formats, garantissant ainsi la qualité des données avant toute analyse.

Au cours de l'étude de l'existant, nous avons également examiné des études de cas spécifiques pour illustrer comment les données d'appels ont été utilisées pour résoudre des problèmes réels. Par exemple, nous avons étudié des cas où l'analyse des données d'appels a permis de détecter des problèmes de réseau, d'optimiser les plans tarifaires ou de cibler des campagnes promotionnelles. De plus, certaines tendances émergentes ont été identifiées, comme une augmentation de l'utilisation des données mobiles pendant certaines périodes.

il était essentiel de considérer l'environnement concurrentiel dans lequel évolue Tunisie Telecom. Les concurrents majeurs tels qu'Ooredoo et Orange jouent un rôle significatif sur le marché des télécommunications en Tunisie. L'analyse comparative avec ces acteurs a permis de contextualiser les performances et les stratégies de Tunisie Telecom au sein du paysage concurrentiel.

Une analyse attentive des concurrents a révélé les évolutions technologiques et les innovations qu'ils ont adoptées. Cela inclut l'introduction de nouveaux services, tels que la 5G, ainsi que les initiatives visant à améliorer l'expérience client via des applications mobiles et des services en ligne. Cette compréhension des tendances technologiques a éclairé les opportunités et les défis auxquels Tunisie Telecom est confrontée dans un environnement en constante évolution.

Problématique

Au cœur de notre projet réside une problématique complexe et cruciale dans le domaine de Business Intelligence : La problématique fondamentale de ce stage réside dans la manière dont Tunisie Telecom peut optimiser la segmentation de sa clientèle à partir des données d'appels. Comment pouvons-nous diviser de manière pertinente la vaste base de clients en segments distincts, chacun ayant des caractéristiques et des besoins spécifiques ?

Un autre défi majeur est d'utiliser la classification prédictive pour anticiper les comportements des clients, notamment en ce qui concerne la rétention. Comment pouvons-nous prédire avec précision quel groupe de clients est plus susceptible de résilier leur contrat ou de chercher d'autres fournisseurs ? La régression logistique est une approche prometteuse, mais comment pouvons-nous appliquer cette technique pour obtenir des résultats fiables et exploitables ?

La création d'un dashboard interactif via Power BI introduit une autre dimension à notre problématique. Comment pouvons-nous concevoir un tableau de bord visuellement attrayant et intuitif qui présente les résultats de manière convaincante ? Comment pouvons-nous garantir que les graphiques et les visualisations fournissent des informations pertinentes pour les parties prenantes et soutiennent les décisions stratégiques ?

Besoins Fonctionnels

Segmentation Pertinente de la Clientèle

L'un des besoins fonctionnels essentiels de ce projet est la capacité de segmenter la clientèle de manière pertinente. Il est primordial de développer des critères de segmentation solides et adaptés à l'environnement de Tunisie Telecom. Ces critères devraient permettre d'identifier des groupes de clients ayant des comportements similaires, des besoins communs et des préférences spécifiques en matière de services.

Modélisation Prédicative à l'aide de la Régression Logistique

Un besoin fondamental est la mise en place d'un modèle de régression logistique pour la classification prédictive des clients. Ce modèle devrait être capable de prendre en compte des variables pertinentes pour prédire les comportements futurs des clients, tels que la résiliation de contrat. Il devrait être conçu de manière à maximiser la précision et la fiabilité des prédictions.

Création d'un Dashboard Interactif avec Power BI

La création d'un dashboard interactif joue un rôle central dans ce projet. Le besoin est de concevoir un tableau de bord dynamique et visuellement attractif à l'aide de Power BI. Ce tableau de bord devrait regrouper les visualisations pertinentes, permettant aux parties prenantes de suivre les tendances, d'explorer les données et de prendre des décisions éclairées en temps réel.

Besoins Non Fonctionnels

Performance et Temps de Réponse

Un besoin non fonctionnel clé est la performance du système. Les analyses, les prédictions et les visualisations doivent être traitées rapidement pour garantir une expérience utilisateur fluide et réactive.

Sécurité et Confidentialité des Données

La sécurité des données est un besoin non fonctionnel critique. Les informations sensibles des clients doivent être protégées contre les accès non autorisés.

Convivialité et Accessibilité du Tableau de Bord

Un besoin non fonctionnel important est la convivialité et l'accessibilité du tableau de bord interactif. Il doit être conçu de manière intuitive, facilitant la navigation et la compréhension des utilisateurs, même sans connaissances techniques avancées.

Évolutivité et Maintenance Simplifiée

L'évolutivité du système est un besoin non fonctionnel crucial. À mesure que les données et les besoins de l'entreprise évoluent, le système doit être capable de gérer des volumes de données croissants sans compromettre les performances.

Choix des plateformes

Pour l'analyse approfondie des données d'appels, le choix s'est porté sur **Python** pour sa polyvalence en matière d'analyse statistique et de modélisation prédictive. Ce langage de programmation offre la flexibilité nécessaire pour manipuler les données, développer des modèles prédictifs et effectuer des analyses exploratoires.

Pour la création du dashboard interactif, Power BI s'est avéré être la plateforme idéale. **Power BI** offre une interface conviviale pour concevoir des visualisations interactives et des tableaux de bord dynamiques. Sa capacité à se connecter à diverses sources de données, à transformer les données en insights visuels et à permettre une exploration interactive a répondu aux besoins de présentation et de communication des résultats à travers des visualisations convaincantes.

CRISP

a- Phase 1 : Compréhension des Objectifs Métier

Objectif Métier 1 (BO1) :

Amélioration de la Personnalisation des Offres

Data Mining Objective 1 (DMO1.1) : Identifier les segments de clients en fonction de leurs habitudes d'appels et de leurs préférences.

Data Mining Objective 2 (DMO1.2) : Analyser les modèles d'utilisation des services pour chaque segment afin de personnaliser les offres.

Objectif Métier 2 (BO2) :

Réduction de l'Attrition Client

Data Mining Objective 1 (DMO2.1) : Identifier les signaux précurseurs d'une possible résiliation de contrat par l'analyse des comportements d'appels et Construire un modèle de régression logistique pour prédire les clients à risque de résiliation, en se basant sur les données d'appels et d'autres variables pertinentes.

b- Phase 2 : Compréhension des Données :

La première étape de cette phase implique l'identification des différentes variables contenues dans les données d'appels , telles que 'DUREE_APPEL_TOT', 'nb_sms_tot', 'revenu_cdr_c', 'MNT_RECH' .

Ensuite, une analyse descriptive sera menée pour comprendre la distribution des variables clés. Cela peut inclure des statistiques descriptives telles que la moyenne, la médiane, l'écart type et les valeurs minimales et maximales des variables.

```
dataTelecomSelected.describe()
```

Executed in 106ms, 24 Aug at 00:47:38

	DUREE_APPEL_TOT	nb_sms_tot	revenu_cdr_c	MNT_RECH	FREQ_ACT_OUT	NB_JOUR_APPEL_TAXE	NB_SMS_TAXE	NB
count	92142.000000	92142.000000	92142.000000	92142.000000	92142.000000	92142.000000	92142.000000	
mean	106.280219	8.925385	4.974875	10.075597	22.262461	5.893715	5.061497	
std	192.534484	61.173818	8.049118	15.409211	55.523636	7.657599	21.126158	
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	
25%	4.230000	0.000000	0.260000	1.330000	0.200000	0.330000	0.000000	
50%	34.180000	0.670000	2.000000	4.330000	0.570000	2.670000	0.330000	
75%	125.037500	5.330000	6.380000	13.330000	3.100000	8.000000	3.670000	
max	4202.350000	4387.000000	217.530000	410.000000	106.000000	70.330000	1011.330000	

Fig 1 : Analyse Descriptive Des Variables

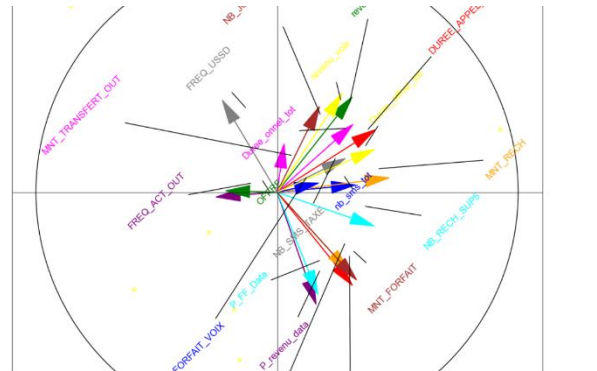
Parallèlement, des techniques de visualisation seront employées pour représenter graphiquement les données. Des graphiques tels que des histogrammes, des diagrammes circulaires et des diagrammes de dispersion permettront de saisir visuellement les tendances et les concentrations des données :

- Par exemple le Scatter Plot (voir Fig 2 ci-dessous) montre bien que les offres a fort revenu_cdr_c (revenu appel + sms) sont principalement : PRE – Trankil TT , PRE – 1=11 , PRE – Classic et PRE – TT 1500%
- Le Pie Chart (voir Fig 3 ci-dessous) montre aussi que ces 4 Offres sont les 3 principaux de point de vu nombre d'utilisateurs .

⇒ **La découverte de ces 4 principaux types d'offres a des implications directes sur les activités de communication et de marketing de Tunisie Telecom. Les efforts de marketing peuvent être ciblés de manière plus précise en fonction des caractéristiques de chaque offre et de chaque segment.**

c- Phase 3 : Préparation des Données :

Dans cette phase, les données d'appels seront nettoyées, transformées et préparées pour être utilisées dans les modèles. Les variables pertinentes seront sélectionnées (Selon PCA) et les données manquantes seront traitées.



```
# selection of the most important variables for each clustering
selected_variables = [
    'FREQ_USSD', 'revenu_cdr_c', 'DUREE_APPEL_TOT', 'MNT_RECH', 'VOLUME_SESSION', 'MNT_FORFAIT_DATA', 'FREQ_ACT_OUT'
]

# Select the desired variables from the dataframe
dataTelecomSelected = dataTelecom[selected_variables].copy()
dataTelecomSelected.head(2000)
Executed in 61ms, 22 Aug at 14:44:00
```

	FREQ_USSD	revenu_cdr_c	DUREE_APPEL_TOT	MNT_RECH	VOLUME_SESSION	MNT_FORFAIT_DATA	FREQ_ACT_OUT
0	1.18	8.94	666.64	33.67	0.02	0.00	0.03
1	184.00	7.31	50.09	7.33	0.00	0.00	0.30
2	184.00	3.23	42.72	3.67	0.00	0.00	0.60
3	184.00	4.55	80.43	4.33	0.21	0.00	0.60
4	17.96	0.53	3.10	2.67	0.00	1.50	2.65
...
1995	15.00	2.90	7.30	3.67	135.62	0.72	2.14
1996	184.00	4.54	24.52	4.33	0.00	0.00	1.29
1997	22.56	59.62	435.24	66.67	2381.06	6.00	0.05

```
dataTelecomSelectedSample=dataTelecomSelected.sample(n=1000)
```

Fig 4 : Selection Des Variables Pertinentes + Simplification De nb lignes

```
dataTelecomSelected.dropna(inplace=True)
```

Fig 5 : Nettoyage Des Données

d- Phase 4 : Modélisation

Pour l'Objectif Métier 1 (BO1) :

Data Mining Objective 1 (DMO1.1) :

a/ On va appliquer le clustering par **k-means** pour segmenter les clients en groupes homogènes basés sur les caractéristiques d'appels.

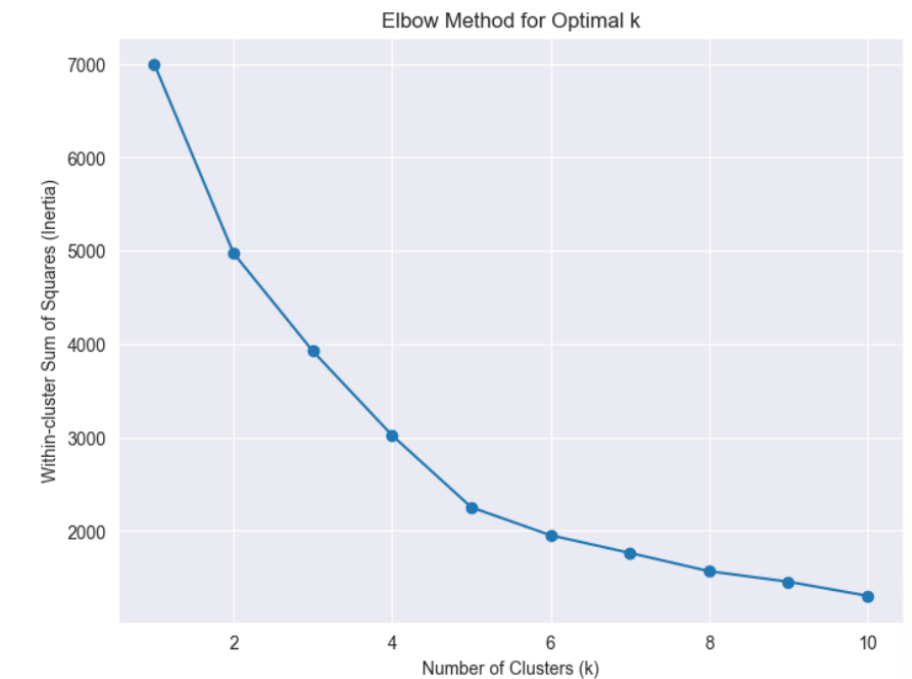


Fig 6 : Elbow Method

D'après Elbow Method , on cherche le point où la variation commence à s'aplatir, ressemblant à un coude. ➔ K = 6

	FREQ_USSD	revenu_cdr_c	DUREE_APPEL_TOT	MNT_RECH	VOLUME_SESSION	\
Cluster						
0	26.665098	2.688235	66.102026	7.659183	1168.690131	
1	184.000000	3.539544	46.483309	4.286691	208.072542	
2	108.760000	21.366707	436.830488	28.735732	846.444268	
3	16.853333	7.265238	320.034286	66.577143	34559.922857	
4	16.831029	4.977647	247.721912	29.628235	10597.473382	
5	184.000000	0.000283	0.004717	0.433868	2.909245	

	MNT_FORFAIT_DATA	FREQ_ACT_OUT
Cluster		
0	3.482680	1.698954
1	0.000000	4.090552
2	3.034268	0.091463
3	46.355714	0.513810
4	18.788971	0.956471
5	0.000000	173.461509

Fig 7 : Les Clusters De Kmeans

- **Cluster 0** : VOLUME_SESSION important , tous les autres variables sont faibles . ➔ utilisation fréquente des services de données mobiles .
Cependant, les autres variables étant faibles, il peut s'agir de clients qui s'engagent pas trop dans d'autres types d'activités ou de transactions.
- **Cluster 1** : FREQ_USSD maximal , les autres variables sont faibles voire minimales parfois . ➔ ce qui peut indiquer qu'ils sont actifs dans l'utilisation des services interactifs et des informations automatisées fournis par l'opérateur.
- **Cluster2** : FREQ_USSD moyen , revenu_cdr_c maximal, DUREE_APPEL_TOT maximale , les autres variables sont faibles
➔ Cela peut indiquer des clients engagés dans une diversité de services, notamment des appels prolongés, générant un revenu substantiel.

Cluster 3 : FREQ_USSD faible ,MNT_RECH , VOLUME_SESSION et MNT_FORFAIT_DATA maximaux, revenu_cdr_c moyen , FREQ_ACT_OUT faible , FREQ_ACT_OUT minimale . ➔ Cela pourrait représenter un groupe axé sur la connectivité via les données mobiles et les forfaits, et moins sur les appels et les interactions sortantes.

Cluster 4 : toutes les variables sont plus ou moins moyennes sauf VOLUME_SESSION très faibles . ➔ Ce qui peut suggérer une utilisation globalement limitée des services.

Cluster 5 : FREQ_ACT_OUT et FREQ_USSD maximales , le reste des variables sont pratiquement minimales ➔ Ils peuvent être des utilisateurs actifs, engagés dans des interactions sortantes fréquentes et des services basés sur USSD.

Conclusion :

Les différents clusters reflètent des comportements d'utilisation variés, allant des utilisateurs intensifs d'appels aux clients centrés sur les données et les services USSD.

En comprenant les besoins et les comportements spécifiques de chaque cluster, Tunisie Telecom peut adapter ses offres et ses promotions pour répondre efficacement aux préférences de chaque groupe, augmentant ainsi la pertinence des interactions et la satisfaction des clients.

b/ Maintenant on va appliquer la methode CAH :

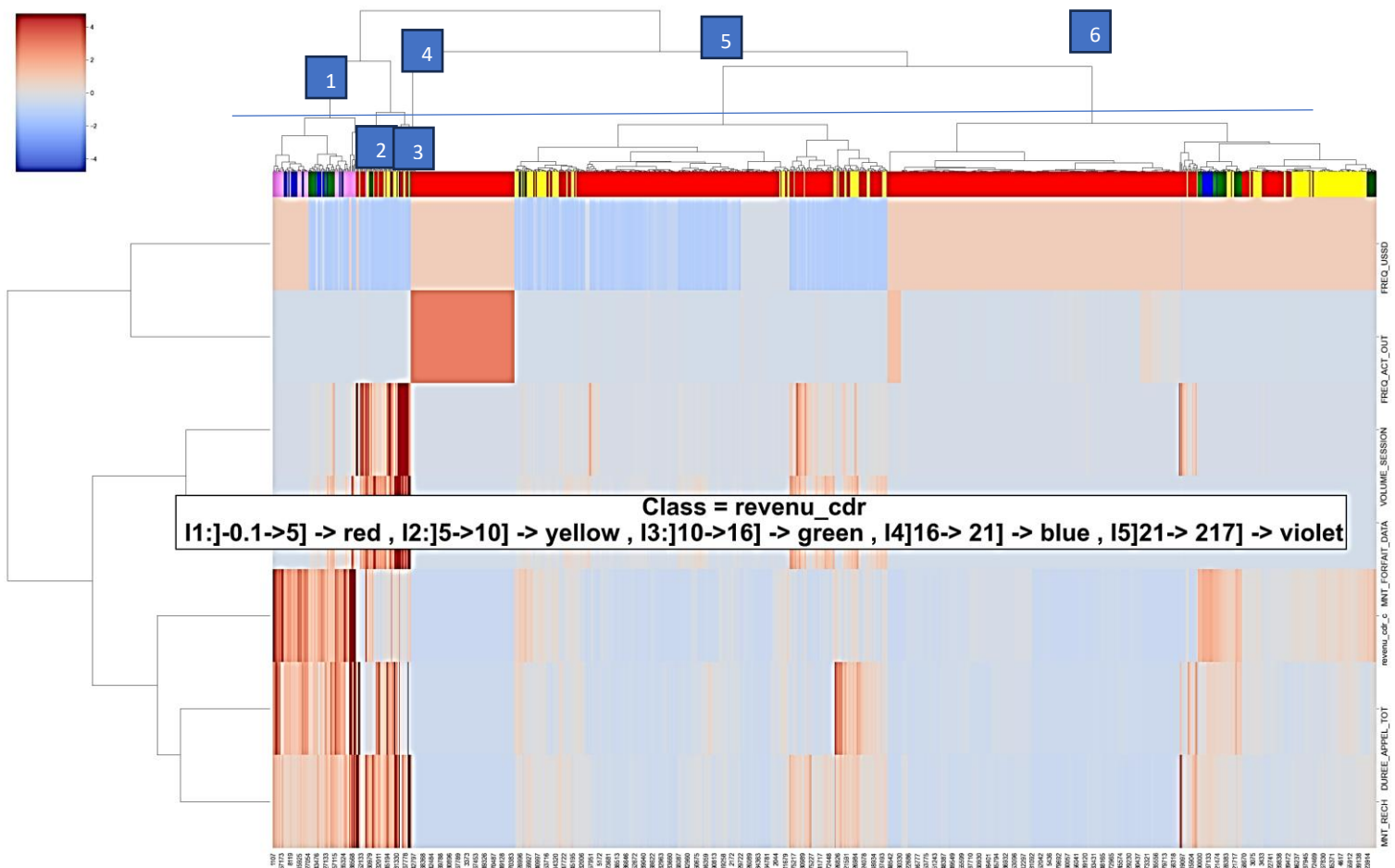


Fig 8 : Clustering - heatmap

Grp1 : revenue_cdr_c maximal , MNT_rech forte , DUREE_APPEL_TOT maximale les autres variables sont faibles ➔ Cela peut indiquer des clients à revenu élevé qui utilisent activement les services d'appels.

Grp2 : revenu_cdr_c faible , VOLUME_SESSION et MNT_FORFAIT_DATA fortes MNT_RECH maximale et le reste sont faible . ➔ C'est un groupe privilégiant la connectivité via les données mobiles malgré un faible revenu d'appels.

Grp3 : revenue_cdr_c moyen , VOLUME_SESSION et MNT_FORFAIT_DATA et MNT_RECH maximales le reste sont faibles

➔ Cela peut indiquer des clients avec une utilisation équilibrée des différents services proposés.

Grp4 et Grp 5: revenue_cdr_c faible , FREQ_USSD minimale les autres sont faibles. ➔ Cela suggère une utilisation minimale des services offerts.

Grp6 : revenue_cdr_c moyen , les autres sont faibles voire minimales

➔ Cela peut représenter des clients qui ne sont pas fortement engagés dans l'utilisation des divers services de Tunisie Telecom.

Conclusion :

Cette segmentation fournit des informations précieuses pour la stratégie de tarification de Tunisie Telecom. Les groupes à fort revenu pourraient bénéficier d'offres haut de gamme, tandis que les groupes orientés vers les données pourraient être ciblés avec des promotions axées sur les forfaits de données. Cette approche peut aider l'entreprise à maximiser ses revenus en adaptant ses offres à différents segments de clients.

Correspondance :

-Groupe 1 (CAH) : Fort Revenu et Utilisation Élevée

Cluster Correspondant (K-means) : Cluster 2 (K-means)

Raison : Les deux groupes partagent des caractéristiques de revenu élevé et d'utilisation active des services.

-Groupe 2 (CAH) : Faible Revenu, Préférence pour les Données

Cluster Correspondant (K-means) : Cluster 3 (K-means)

Raison : Les deux groupes présentent des préférences pour les données (VOLUME_SESSION élevé et MNT_FORFAIT_DATA élevé) malgré un faible revenu.

-Groupe 3 (CAH) : Revenu Moyen, Utilisation Équilibrée

Cluster Correspondant (K-means) : Cluster 4 (K-means)

Raison : Les deux groupes affichent une utilisation moyenne et équilibrée des services, avec un revenu moyen.

-Groupe 4 (CAH) : Faible Revenu, Utilisation Minimale

Cluster Correspondant (K-means) : Cluster 0 (K-means)

Raison : Les deux groupes montrent un faible revenu et une utilisation minimale des services, ce qui est similaire au Cluster 0.

-Groupe 5 (CAH) : Cluster 1 (K-means) Faible Revenu, Utilisation Minimale
FREQ_USSD élevée voire maximale

-Groupe 6 (CAH) : Revenu Moyen, Utilisation Minimale

Cluster Correspondant (K-means) : Cluster 5 (K-means)

Raison : Les deux groupes ont un revenu moyen, mais une utilisation minimale des services, ce qui correspond au Cluster 5.

Data Mining Objective 2 (DMO1.2) :

personnaliser les offres pour chaque segment.

En s'appuyant sur les resultat de Scatter Plot et de CAH on peut correspondre chaque segment aux offres les plus adéquats :

Ordre de revenu_cdr_c decroissant selon heatmap :

1->6->2->3->4->5 (plus la couler violet est présnte plus le revenu est fort ensuite on passe au bleau puis vert ...)

-Grp1 → PRE – TANKIL TT , PRE – Classic, PRE –1=11 , PRE – Conso 500mil TT ,

PRE – EST 100%NEW, PRE – TM35mil/min, PRE – TT 1500% , HAYA , PRE – 900% , PRE – Double.

-Grp2,4,5 → PRE – BONUS sur recharge , PRE – JAWHARA 1=11 , PRE-OULIDHA ,

PRE – Messenger , PRE – Conso 500mil , PRE – carte jeune , PRE-css mobile , PRE-LOL

-Grp 3,6 → le reste

Data Mining Objective 1 (DMO2.1) : Identifier les signaux précurseurs d'une possible résiliation de contrat par l'analyse des comportements d'appels et Construire un modèle de régression logistique pour prédire les clients à risque de résiliation, en se basant sur les données d'appels et d'autres variables pertinentes.

```
# Sélectionnez les caractéristiques (features) et la variable cible
features = ['FREQ_USSD', 'revenu_cdr_c', 'DUREE_APPEL_TOT', 'MNT_RECH',
            'VOLUME_SESSION', 'MNT_FORFAIT_DATA', 'FREQ_ACT_OUT']
target = 'STATUT'

# Séparez les données en caractéristiques (X) et variable cible (y)
X = dataTelecomSelected[features]
y = dataTelecomSelected[target]

# Gérez les valeurs manquantes en utilisant SimpleImputer
imputer = SimpleImputer(strategy='mean')
X_imputed = imputer.fit_transform(X)

# Divisez les données en ensembles d'entraînement et de test
X_train, X_test, y_train, y_test = train_test_split(X_imputed, y, test_size=0.2, random_state=42)

# Créez et ajustez le modèle de régression logistique
model = LogisticRegression()
model.fit(X_train, y_train)

# Faites des prédictions sur l'ensemble de test
y_pred = model.predict(X_test)

# Affichez les résultats de la classification
print("Matrice de confusion:\n", confusion_matrix(y_test, y_pred))
print("\nRapport de classification:\n", classification_report(y_test, y_pred))
```

Fig 9 : Logistic Regression

```

new_data = pd.DataFrame({
    'FREQ_USSD': [184.00], # Remplacez par vos valeurs d'entrée
    'revenu_cdr_c': [0.3],
    'DUREE_APPEL_TOT': [7.82],
    'MNT_RECH': [0.33],
    'VOLUME_SESSION': [0],
    'MNT_FORFAIT_DATA': [0],
    'FREQ_ACT_OUT': [6.4]
})

# Gérez les valeurs manquantes dans les nouvelles données
new_data_imputed = imputer.transform(new_data)

# Faites des prédictions sur les nouvelles données
predictions = model.predict(new_data_imputed)

# Affichez les prédictions
print("Predicted Statut :", predictions)
Executed in 14ms, 24 Aug at 05:58:54

Predicted Statut : ['S']

```

Fig 10 : Prédire une possible résiliation

Lorsqu'on prédit le STATUT d'un client qui est soit A (active) soit S (suspendu)

⇒ On peut donc prévoir une possible de résiliation de contrat

e- Phase 5 : Evaluation :

```
from sklearn.metrics import calinski_harabasz_score

kmeans_labels = KmeansData['Cluster']
cah_cluster_labels = fcluster(Z, 6, criterion='distance')

calinski_kmeans = calinski_harabasz_score(X_scaled, kmeans_labels)
calinski_cah = calinski_harabasz_score(X_scaled, cah_cluster_labels)

print("Calinski-Harabasz Index - K-Means:", calinski_kmeans)
print("Calinski-Harabasz Index - CAH:", calinski_cah)
```

Executed in 31ms, 24 Aug at 06:42:05

Calinski-Harabasz Index - K-Means: 514.0948255978456

Calinski-Harabasz Index - CAH: 781.5807074368822

Si l'indice de Calinski-Harabasz (également connu sous le nom de critère du rapport de variance) est plus élevé, cela suggère que les clusters sont mieux définis et mieux séparés selon la méthode de regroupement choisie. Cet indice tient compte à la fois de la variance inter-cluster et de la variance intra-cluster, et une valeur plus élevée indique une meilleure qualité de regroupement.

⇒ **CAH était mieux définie que KMEANS**

Visualisation avec Power Bi

a- Page 1 : General

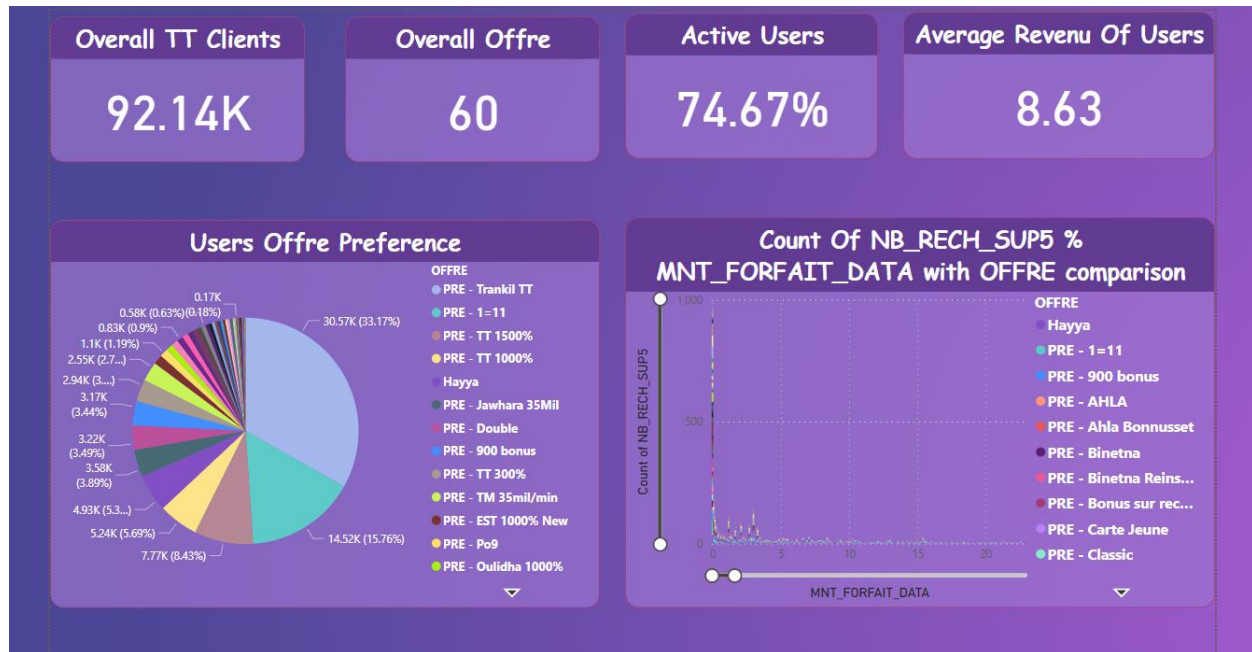


Fig 11 : Power Bi - Page General

Cette page Contient des visualisations générales sur les données Tunisie Telecom :

1. **Overall TT Clients** : Nombre Total des client TT
2. **Overall Offre** : Nombre Total des Offre TT disponibles
3. **Active Users** : Nombre Total des clients TT active (STATUT = 'A')
4. **Average Revenu Of Users** : Chaque client TT a une colonne ARPU ou se calcule le revenu en moyenne de ce client. Donc on a voulu calculer la moyenne de revenu pour tous les clients TT = moyenne des ARPU (moyenne des moyennes)

5. **Users Offre Preference** : Distributions des offres par nombre d'utilisateurs .

6. **Count Of NB_RECH_SUP5 % MNT_FORFAIT_DATA with OFFRE comparison** :

Le nombre total de NB_RECH_SUP5 par MNT_FORFAIT_DATA avec comparaison interactif des Offres. On remarque qu'ils sont inversement proportionnels . Les utilisateurs de forfait on tendance a charger leurs forfait directement par paiement en ligne ou par agence .

b- Page 2 : Revenus

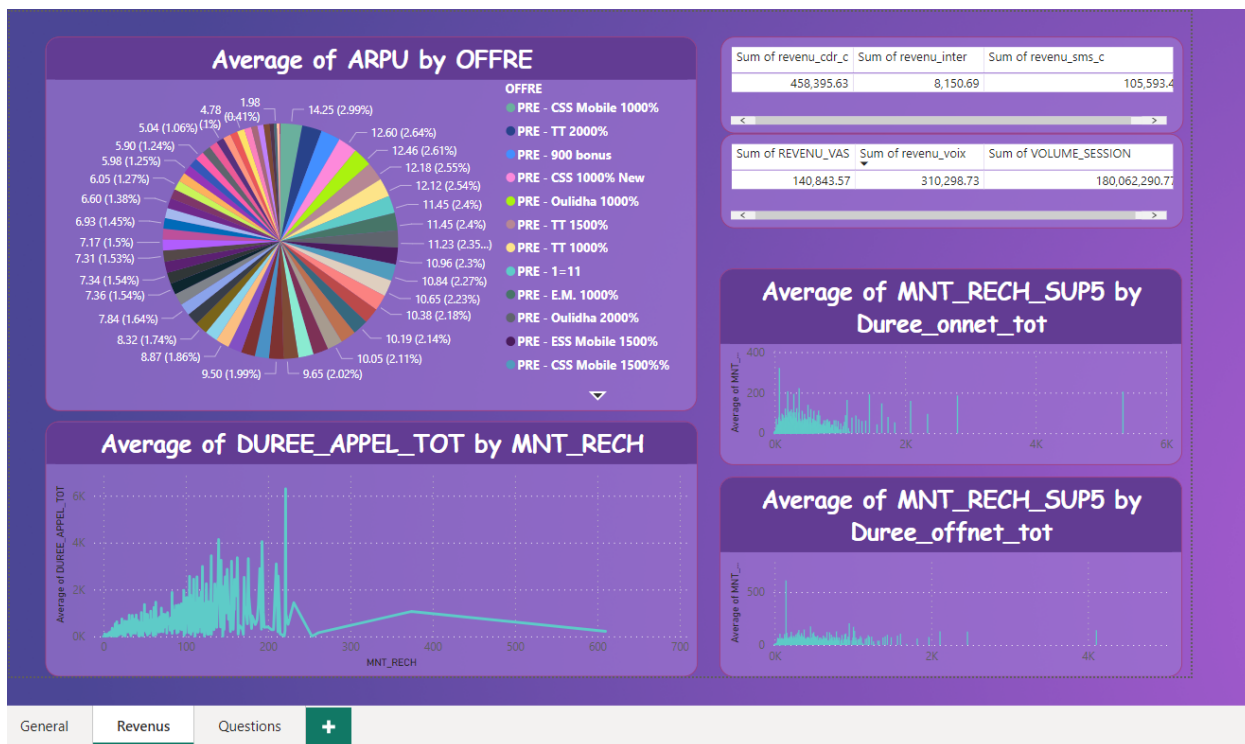


Fig 12 : Power Bi - Page Revenus

Cette page Contient des visualisations générales sur les Revenus Tunisie Telecom .

1. **Average Of ARPU By OFFRE :** Pour chaque client TT , on a une colonne ARPU (Average Revenu Per User) . Donc calculer la moyenne ARPU pour la totalité des client (Sum ARPU / Total des clients TT) et les classer par offre nous donne une idée sur la rentabilité de chaque offre TT
2. **Average of DUREE APPEL TOT by MNT RECH :** Ce plot montre l'évolution des DUREE_APPEL_TOT en fonction de MNT_RECH : la zone entre 100 et 250 correspond à une durée totale maximale . On peut conclure que ceux ces client sont les consommateurs les plus importants des services appel .
En dessous de cette zone , les clients semblent être ou bien suspendus (durée nulle) ou une faible utilisations des appel . En dessus de cette zone , les utilisateur n'utilisent pas beaucoup les services appels malgré la forte dépense comme MNT_RECH : ce sont les client abonnées en d'autres services comme forfait internet , jeux , sms taxes etc...
3. **Average of MNT RECH SUP5 by Duree onnet tot**
4. **Average of MNT RECH SUP5 by Duree offnet tot**
Ceux sont des plot qui illustres la moyenne de recharge sup5 en rapport avec la durée des appel onnet (TT vers TT) ou offnet (TT vers un autre opérateur) pour une durée entre 0 et 2k , l'utilisation des RECH_SUP5 est maximale . Avec quelques variations. En dehors de cette zone , elle est faible voire nulle .
5. **Les autres :** ceux sont des sommes des différentes revenus + somme de VOLUME_SESSION .

c- Page 3 : Questions



Fig 13 : Power Bi - Page Questions

Cette page ressemble un tout petit peu au fameux ChatGPT sauf qu'il se s'agit pas vraiment d'une chat. il y a des mots clés précises qu'on doit utiliser pour formuler une question dont Power BI est capable de répondre .

Conclusion

Le présent stage au sein de Tunisie Telecom, axé sur l'analyse des données des appels clients, la segmentation et la classification prédictive, a abouti à des découvertes et des insights précieux qui ouvrent la voie à une utilisation plus stratégique des données pour l'innovation et l'amélioration continue.

La segmentation des clients en clusters distincts, tant par la méthode K-means que par l'analyse de hiérarchie ascendante (CAH), a permis d'isoler des groupes spécifiques de clients partageant des comportements et des préférences similaires.

La modélisation prédictive, en particulier la classification par régression logistique, a permis de prédire avec une précision accrue les comportements futurs des clients, comme la résiliation de contrat.

En fin de compte, ce stage a permis de transformer des données brutes en informations exploitables, démontrant le potentiel de l'analyse de données pour stimuler la croissance et le succès d'une entreprise. L'expérience acquise tout au long de ce projet se révélera inestimable pour guider les futures initiatives de traitement des données et d'amélioration des services chez Tunisie Telecom.