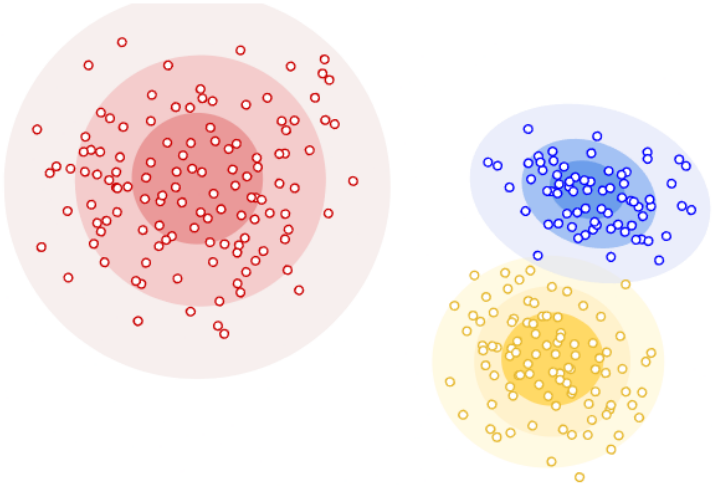
Université Claude Bernard Lyon 1

MIF11 - Machine Learning pour la création de devis



Master 1 Informatique – Université Claude Bernard Lyon 1
2022-2023

Doctorante	Enseignant	Étudiants
Amira BEN HADID	Hamamache KHEDDOUCI	Kévin TANG p1501263 David TRAN p1911682 Anh-Kiet VO p1907921

Table des matières

I. Introduction

Dans le cadre de notre projet d'Ouverture à la Recherche sous la supervision de M. Hamamache KHEDDOUCI à l'Université Claude Bernard Lyon 1, nous avons travaillé sur l'analyse d'un ensemble de données clients afin d'en extraire des informations pertinentes et exploitables par une entreprise.

Ces données nous ont été fournies, anonymisées, par l'entreprise lyonnaise Béton Direct qui est la première plateforme de vente en ligne, spécialisée dans la livraison de bétons et mortiers frais, pour les particuliers et micro-entreprises du bâtiment. Nous les remercions pour cela, pour leur confiance et de nous permettre de pouvoir nous exercer avec des données issues d'une base de données professionnelle.

Notre objectif était de comprendre la structure sous-jacente des données et d'identifier des groupes ou des clusters au sein de l'ensemble de données.

Pour atteindre cet objectif, nous avons suivi une méthodologie en plusieurs étapes comprenant l'extraction des données, une analyse exploratoire préliminaire, la sélection des caractéristiques importantes, un pré-traitement des données, la détermination du nombre optimal de clusters, l'application de l'algorithme K-means et l'analyse des résultats obtenus.

Ce rapport présentera en détail ces différentes étapes, les résultats obtenus et les principales conclusions de notre travail.

II. Analyse initiale des données

1. Extraction des données

Comme indiqué précédemment, nous avons reçu un ensemble de données de l'entreprise Béton Direct par l'intermédiaire de notre encadrant sous la forme d'un fichier CSV. Nous avons utilisé des outils et des bibliothèques spécifiques, tels que **pandas** et **numpy**, pour importer et charger les données dans un format adapté pour l'analyse.

Une fois les données extraites, nous avons procédé à une première exploration pour mieux les comprendre. Nous avons examiné les différentes variables disponibles, leur type, leur format et leur signification. Cette étape nous a permis de se familiariser avec les données et d'identifier d'éventuels problèmes tels que des valeurs manquantes, des erreurs de format ou des valeurs aberrantes.

L'extraction des données constitue une étape cruciale pour toute analyse, car elle garantit que nous disposons d'un ensemble de données complet et approprié pour les étapes suivantes de notre travail.

2. Etat initial

La base de données possède 26 features différentes et est composée de 55117 devis réalisés par des clients de Béton Direct dont 9.74% ont été finalisés.
On peut remarquer qu'une grande proportion des devis n'ont pas été finalisée, ce qui peut être un problème pour les algorithmes de machine learning qui ont besoin d'une grande quantité de cas à la fois positifs et négatifs pour leur apprentissage, leur fonctionnement et pour produire des résultats précis.

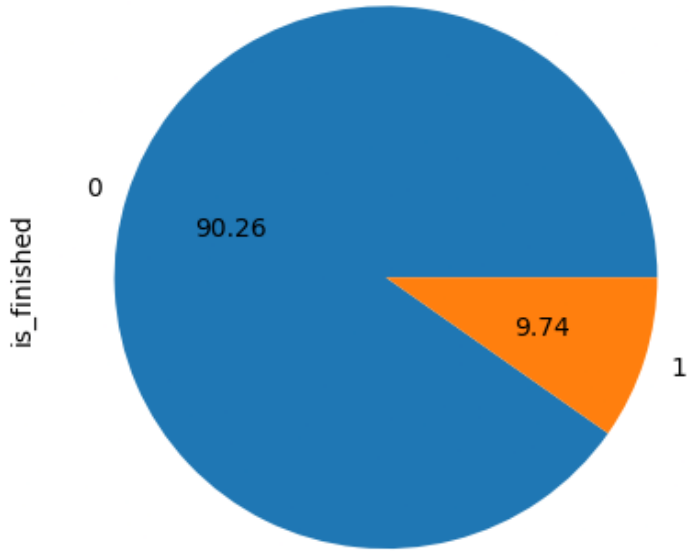


Figure 1 : diagramme des proportions de la variable "is_finished"

III. Sélection des features

Avec l'aide de notre encadrante Mme. Amira BEN HADID, nous avons entrepris de sélectionner les caractéristiques les plus pertinentes et informatives pour notre analyse. L'objectif était de réduire la dimensionnalité des données en éliminant les caractéristiques redondantes ou peu utiles, afin de simplifier le modèle et d'améliorer la performance de notre analyse. Nous avons donc parcouru une à une les différentes features pour comprendre de manière précise le rôle de chacune pour décider lesquelles sélectionner ou éliminer.
La sélection des features permet de concentrer l'analyse sur les caractéristiques les plus importantes, améliorant ainsi la qualité des résultats, la compréhension des phénomènes étudiés et l'efficacité de l'analyse des données. Elle constitue une étape essentielle pour maximiser l'information extraite des données et prendre des décisions éclairées.

Voici les features que nous avons sélectionnées à la fin de cette étape :

Features
is_pro
acquisition_id
work_type_id
volume
handling
is_finished
plant_id
previous_bill_count
client_maturity

Figure 2: liste des variables sélectionnées

IV. Pré-traitement des données

1. Nettoyage des valeurs manquantes ou aberrantes

Le nettoyage des valeurs manquantes ou aberrantes est une étape essentielle de notre analyse, visant à garantir la qualité des données utilisées. En éliminant les valeurs incohérentes ou manquantes, nous réduisons les biais potentiels et obtenons des résultats plus fiables et précis. Cela renforce la crédibilité de notre analyse en nous permettant de travailler avec des données de qualité. Les conclusions que nous tirons sont ainsi plus robustes et nos interprétations plus précises.

Pour aborder cette étape, nous avons tout d'abord identifié les valeurs manquantes dans notre jeu de données en utilisant des techniques de détection telles que la fonction **isna**(). Cela nous a permis de repérer les lignes ou les colonnes contenant des valeurs nulles.
Ensuite, nous avons pris des mesures pour traiter ces valeurs manquantes de manière appropriée. Nous avons décidé de supprimer les devis entiers contenant des valeurs manquantes car dans notre cas cela n'impacte pas de manière significative la qualité de la base de données et notre analyse. Dans d'autres cas, on aurait pu utiliser un **Imputer** pour remplacer les valeurs manquantes par une valeur moyenne, médiane ou la valeur la plus fréquente afin de préserver l'intégrité des données.

De plus, nous avons également identifié et traité les valeurs aberrantes dans nos données. Étudier les données en amont nous aura été bénéfique pour cela car avec les connaissances acquises du site internet de l'entreprise, nous avons compris par exemple que les volumes supérieurs à 50 m3 sont des valeurs aberrantes car il n'est pas possible de commander un tel volume au-delà de cette limite.

2. Normalisation des valeurs numériques

Après avoir séparé les features de type numérique et catégorielle, nous nous sommes concentrés ici sur les valeurs numériques de notre base de données.
Cette étape s'est concentrée sur la normalisation des variables numériques afin d'éliminer les biais potentiels et de faciliter une comparaison équitable entre les différentes caractéristiques.
Pour atteindre cet objectif, la méthode du **StandardScaler** a été utilisée, une technique couramment utilisée dans le domaine de l'analyse des données. Cette technique permet de centrer les valeurs autour de la moyenne et de les mettre à l'échelle en fonction de l'écart type. Ainsi, chaque feature est ajustée de manière à avoir une moyenne de 0 et un écart type de 1. Cette transformation standardise les variables, ce qui facilite leur comparaison indépendamment de leurs plages de valeurs initiales.

La normalisation des valeurs numériques revêt une importance capitale dans notre étude, car elle permet d'éliminer les distorsions potentielles dues aux différences de magnitude entre les variables. En standardisant les données, nous avons pu réduire l'influence de certaines caractéristiques ayant des plages de valeurs plus étendues sur les autres, évitant ainsi une domination injustifiée lors de l'analyse.
En éliminant les différences d'échelle, nous avons pu comparer et interpréter les coefficients, les poids et les contributions des différentes variables de manière plus juste.

3. Encodage des valeurs catégorielles

Les données que nous avons utilisées comprennent des variables catégorielles, telles que le genre, la région géographique, ou encore le type de produit. Cependant, la plupart des algorithmes de machine learning nécessitent que les données soient en format numérique pour pouvoir les traiter.
Il existe plusieurs techniques d'encodage, chacune adaptée à des situations spécifiques. Dans notre cas, nous avons utilisé **OneHotEncoder** de la bibliothèque scikit-learn, cette méthode consiste à créer de nouvelles variables binaires pour chaque catégorie présente dans la variable catégorielle. Ainsi, chaque observation est représentée par des 0 et des 1, indiquant si elle appartient ou non à une catégorie donnée.

Le processus de l'encodage one-hot consiste à créer de nouvelles variables binaires pour chaque valeur présente dans une variable catégorielle. Chaque observation est représentée par un vecteur binaire, où chaque élément indique si l'observation appartient ou non à une catégorie spécifique. Par exemple, si nous avons une variable "couleur" avec les catégories "rouge", "vert" et "bleu", nous créerons trois nouvelles variables : "rouge", "vert" et "bleu". Chaque observation sera marquée d'un "1" dans la variable correspondante à sa couleur et de "0" dans les autres variables.

Couleur		**Rouge**	**Vert**	**Bleu**
Rouge	->	1	0	0
Vert	->	0	1	0
Bleu	->	0	0	1

Figure 3 : exemple du fonctionnement de OneHotEncoder

L'encodage one-hot nous a permis de traiter efficacement les variables catégorielles avec plusieurs catégories distinctes. Il est important de noter que d'autres méthodes d'encodage existent, telles que l'encodage ordinal, qui attribue des valeurs numériques à chaque catégorie selon un ordre spécifié. Cependant, dans notre analyse, nous avons choisi d'utiliser exclusivement l'encodage one-hot en raison de la nature de nos variables catégorielles et de la nécessité de traiter chaque catégorie de manière indépendante.

V. Détermination du nombre de clusters optimaux

1. Méthodes utilisées

Une fois les données préparées correctement, nous pouvons maintenant commencer à analyser les données en profondeur.

La détermination du nombre optimal de clusters est une étape préliminaire à l'utilisation de l'algorithme K-means. Cette étape nous permet de trouver le nombre de groupes ou de clusters qui maximise la cohésion intra-cluster tout en minimisant la similarité inter-cluster.
Plusieurs approches peuvent être utilisées pour déterminer le nombre de clusters optimal. Nous avons utilisé deux approches largement employées : la **méthode du coude** (elbow method) et la **méthode du coefficient de silhouette** (silhouette score).

La **méthode du coude** consiste à exécuter l'algorithme K-means avec un nombre de clusters croissant, puis à tracer la somme des distances au carré des échantillons par rapport à leur centre de cluster le plus proche en fonction du nombre de clusters. Le graphique ainsi obtenu présente une forme de courbe, ressemblant à un bras plié, d'où son nom. Le point du coude dans la courbe représente le nombre optimal de clusters où l'ajout de clusters supplémentaires ne conduit pas à une amélioration significative de la cohésion intra-cluster. Cependant, il est important de noter que l'interprétation de cette méthode peut parfois être subjective.

La **méthode du coefficient de silhouette** évalue la qualité d'un clustering en calculant la distance moyenne entre chaque échantillon et les autres échantillons du même cluster (cohésion intra-cluster) et en comparant cette distance à la distance moyenne entre l'échantillon et les échantillons des clusters voisins (similarité inter-cluster). Le coefficient de silhouette varie entre -1 et 1, où une valeur proche de 1 indique une bonne séparation entre les clusters, une valeur proche de 0 indique un chevauchement entre les clusters, et une valeur proche de -1 indique une mauvaise séparation entre les clusters. Le nombre optimal de clusters est celui qui maximise le coefficient de silhouette.

Dans notre analyse, nous avons utilisé ces deux approches pour déterminer le nombre optimal de clusters. Nous avons exécuté l'algorithme K-means avec différents nombres de clusters et avons tracé les courbes du coude et du coefficient de silhouette pour évaluer la cohésion intra-cluster et la similarité inter-cluster. Nous avons ainsi identifié le nombre optimal de clusters qui fournissait un équilibre entre la séparation des clusters et la cohésion intra-cluster.

2. Résultats obtenus

Nous avons donc obtenu des résultats sous la forme de graphiques pour ces deux méthodes.

La méthode du coude nous a permis d'observer l'évolution de l'inertie intra-cluster en fonction du nombre de clusters. Nous avons tracé une courbe représentant ces variations et identifié le point de coude, où l'augmentation de clusters n'entraîne plus une réduction significative de l'inertie intra-cluster. Ce point nous a suggéré un nombre de clusters optimal pour notre jeu de données. Comme nous pouvons le voir ci-dessous, le graphique nous indique un point de coude pour un nombre de 8 clusters.

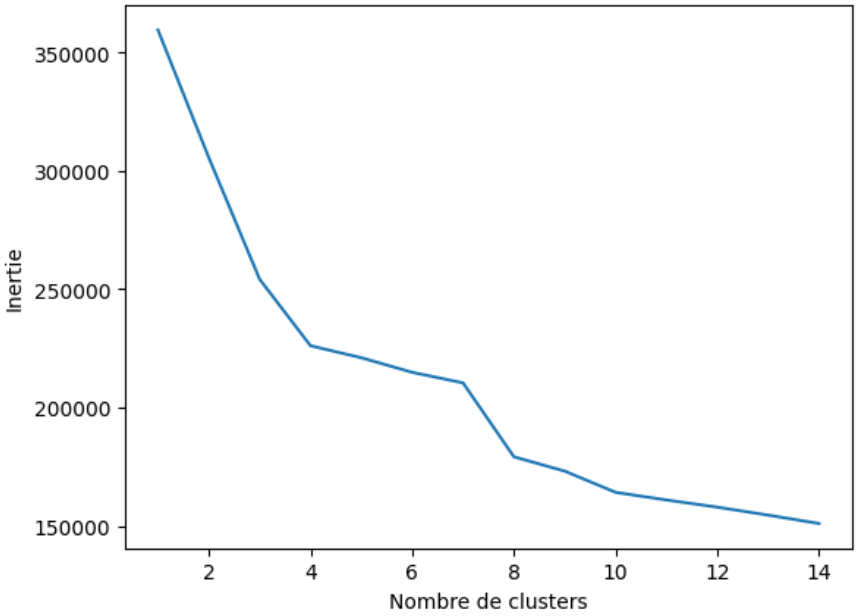


Figure 4 : graphique d'utilisation de la méthode du coude

La méthode du coefficient de silhouette nous a permis d'évaluer la qualité de la partition obtenue pour différents nombres de clusters. Nous avons calculé le coefficient de silhouette pour chaque échantillon, qui mesure à quel point il est bien classé dans son propre cluster par rapport aux autres clusters. Nous avons ensuite tracé un graphique montrant les valeurs moyennes du coefficient de silhouette pour chaque nombre de clusters. Le nombre de clusters associé au coefficient de silhouette le plus élevé indique une meilleure séparation des données.

Le silhouette score nous indique un score maximum obtenu pour un nombre de 4 clusters. En revanche, nous pouvons remarquer que le score obtenu est de 0.23, ce qui est très faible. Cela indique une séparation relativement faible entre les clusters et suggère qu'il y a une certaine superposition ou chevauchement entre les clusters.

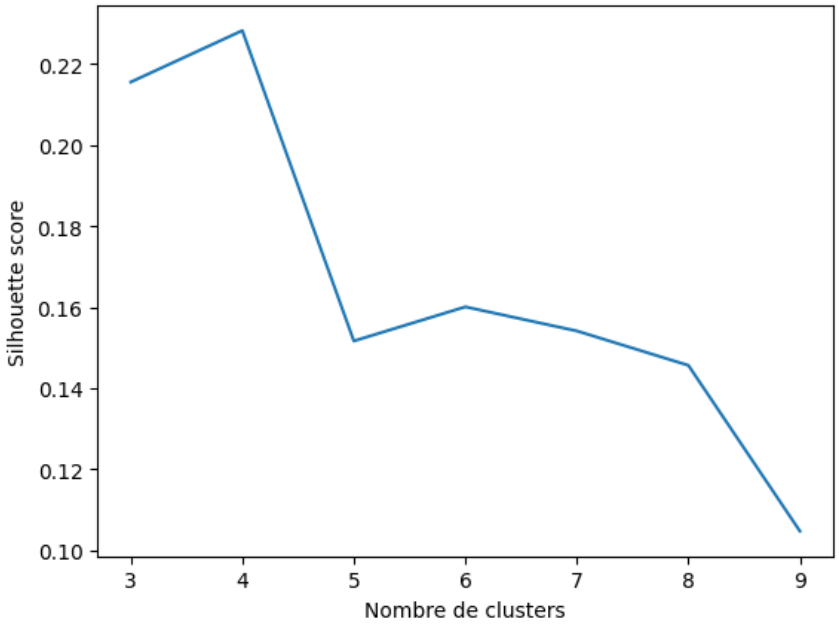


Figure 5 : graphique d'utilisation du coefficient de silhouette

Les deux méthodes que nous avons utilisées ont produit des résultats divergents. Malgré le fait que le coefficient de silhouette ait donné un résultat discutable, nous avons pris la décision d'appliquer l'algorithme de k-means en utilisant les deux nombres de clusters afin de déterminer de manière définitive le choix optimal. Cette approche nous permettra d'évaluer plus précisément la structure des données et d'identifier la configuration de clusters la plus appropriée.

VI. Algorithme K-Means

1. Description de l'algorithme

L'algorithme K-means est l'une des méthodes les plus couramment utilisées pour la segmentation de données. Il vise à diviser un ensemble de données en un nombre prédéfini de clusters en minimisant la somme des carrés des distances entre les points de données et les centres de cluster.
Dans notre étude, une fois que nous avons déterminé le nombre optimal de clusters à l'aide des méthodes précédentes, nous avons appliqué l'algorithme K-means sur nos données prétraitées.

L'algorithme commence par initialiser les centres de cluster de manière aléatoire, puis itère jusqu'à ce qu'une convergence soit atteinte. À chaque itération, les points de données sont assignés au cluster le plus proche en fonction de la distance euclidienne, et les centres des clusters sont mis à jour en calculant la moyenne des points assignés à chaque cluster.

Après avoir exécuté l'algorithme K-means, nous avons obtenu les résultats suivants : la position des centres de cluster finaux et l'assignation de chaque point de données à un cluster spécifique. Ces informations nous permettent de mieux comprendre la structure de nos données et d'effectuer des analyses plus approfondies sur chaque cluster.

2. Analyse en Composantes Principales (PCA)

À ce stade de l'analyse, nous avons réduit la dimensionnalité de nos données composées de 476 variables à 2 variables afin de construire une représentation graphique en 2 dimensions. Pour ce faire, nous avons utilisé l'**Analyse en Composantes Principales (PCA)**, une technique statistique largement utilisée dans la réduction de dimensionnalité des données tout en préservant au mieux les informations et les caractéristiques essentielles des données.

Le PCA permet de transformer un ensemble de variables initiales en un ensemble de nouvelles variables, appelées composantes principales qui sont des combinaisons linéaires des variables initiales. Son objectif est de sélectionner un nombre restreint de composantes principales qui expliquent la plupart de la variance totale. Ainsi, en projetant nos variables initiales sur les deux premières composantes principales, nous avons pu visualiser nos données dans un espace en deux dimensions.

Il est important de souligner que le PCA ne modifie pas les données initiales, mais offre une nouvelle représentation en réorganisant les axes dans l'espace. Cette approche trouve une large utilisation dans l'exploration et la visualisation des données pour faciliter l'analyse et l'interprétation des données.

3. Analyse graphique

- Avec 4 clusters :

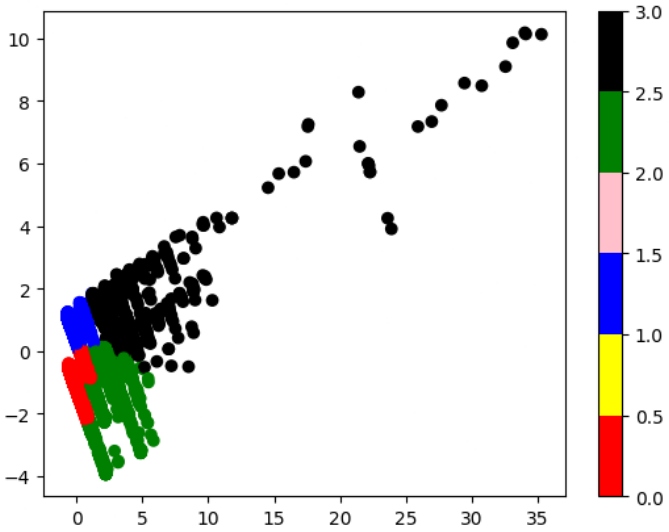


Figure 6 : représentation graphique de 4 clusters en 2 dimensions

Nous observons une variation dans la taille des clusters, certains étant compacts tandis que d'autres sont plus dispersés. En particulier, le cluster coloré en noir sur ce graphique possède en son sein à la fois une partie avec des données très proches les unes des autres et une partie avec des données très éparpillées. Cela montre très clairement un manque d'homogénéité de ce clusters et que ce nombre de clusters n'est pas un choix pertinent.

- Avec 8 clusters :

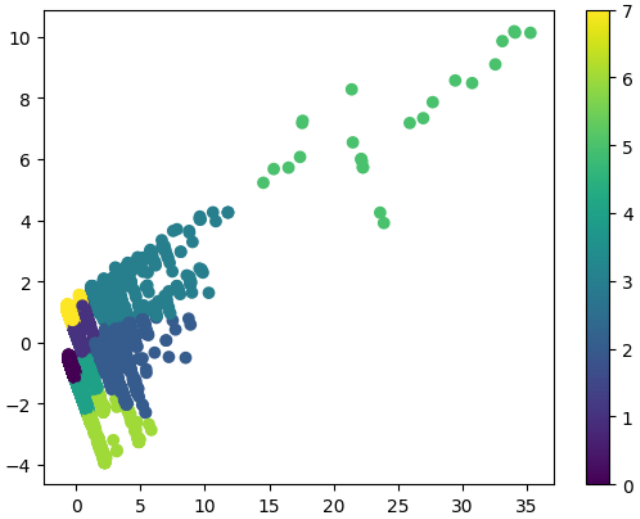


Figure 7 : représentation graphique de 8 clusters en 2 dimensions

Ici, le choix du nombre de clusters semble plus pertinent car les données isolées se retrouvent dans un même cluster, différent des données regroupées de manière plus compacte.

Pour la suite du projet, nous nous sommes donc concentrés sur l'analyse des 8 clusters. Il est maintenant nécessaire d'effectuer une analyse plus détaillée afin de comprendre les raisons pour lesquelles les clusters ont été regroupés de cette manière.

VII. Analyse des clusters

Cette étape nous a permis de mieux comprendre les facteurs qui contribuent à la formation des clusters et d'identifier les caractéristiques clés qui définissent chaque groupe.
En analysant les valeurs et les proportions des différentes caractéristiques, nous avons pu mettre en évidence les similarités et les différences entre les clusters, ce qui nous a aidés à interpréter les résultats de manière plus précise et à obtenir des informations plus significatives sur la structure des données.

L'analyse approfondie des clusters nous a permis de déterminer les variables ayant exercé la plus grande influence sur leur formation. Les résultats sont synthétisés dans le tableau ci-dessous, mettant en évidence les caractéristiques clés qui ont contribué à leur regroupement :

Cluster	**Features**	**Pourcentage d'influence**
1	handling	27.7 %
	acquisition_id	25.3 %
2	work_type_id	23.32 %
	handling	22.46 %
	acquisition_id	21.91 %

3	is_pro	58.71 %
	acquisition_id	14.21 %
4	volume	41 %
	acquisition_id	21.0 %
5	client_maturity	24.1 %
	handling	23.21 %
	acquisition_id	22.56 %
6	previous_bill_count	81.87 %
7	handling	26.36 %
	acquisition_id	26.14 %
8	volume	68.52 %
	acquisition_id	12.52 %

Figure 8 : tableau montrant les features les plus influentes dans chaque cluster

Ces pourcentages nous fournissent une vision globale des clusters et des caractéristiques qui les rassemblent. Cependant, ils ne nous indiquent pas les valeurs spécifiques des variables qui contribuent à leur regroupement. Par exemple, dans le dernier cluster, nous observons que le volume a une influence de 68,52% sur son regroupement, mais nous ne savons pas si ce sont des volumes faibles ou élevés qui jouent ce rôle.
Afin de clarifier cette information, nous avons procédé à l'identification et au comptage des valeurs distinctes de chaque variable dans chaque cluster. Cela nous permet de déterminer avec précision les caractéristiques spécifiques qui rassemblent les données au sein de chaque cluster.

Les résultats que nous avons obtenus nous permettent de caractériser chaque cluster par une phrase descriptive, résumant les principales caractéristiques des données regroupées. En analysant les valeurs et les proportions des différentes variables dans chaque cluster, nous pouvons formuler des observations significatives qui décrivent la nature et les traits distinctifs de chaque groupe de données. Ces descriptions fournissent des indications précieuses sur les comportements ou les profils qui se dégagent au sein de chaque cluster, permettant ainsi une meilleure compréhension des tendances et des structures présentes dans les données.

Cluster	**Description**
1	Le **grand public non bricoleurs** qui découvre Béton Direct par hasard et qui achète de **petits volumes** de béton
2	Les clients **débrouillard** qui réalisent leurs travaux eux-mêmes et ne font qu'**un seul type de travaux**

3	Les **professionnels** qui ont besoin de produits **rapidement**, travaillent dans l'urgence
4	Les clients qui achètent un **volume moyen** de béton (15-35)
5	Les clients **débrouillards** qui réalisent leurs travaux eux-mêmes **sous 15 jours**
6	Les **professionnels** qui achètent très **régulièrement**, les plus **fidèles** à Béton Direct, achètent toujours le même **seul produit**
7	les clients **débrouillard** qui réalisent leurs travaux eux-mêmes et achètent souvent le **même produit**
8	Les acheteurs de **gros volume** de béton

Figure 9 : tableau descriptif des clusters

VIII. Conclusion

1. Résumé

En conclusion, notre étude a porté sur l'analyse et la classification de données à l'aide de l'algorithme K-means.
Nous avons suivi une méthodologie rigoureuse qui comprenait l'extraction des données, l'exploration des caractéristiques, la sélection des variables pertinentes, le nettoyage des valeurs manquantes ou aberrantes, le prétraitement des valeurs numériques et l'encodage des valeurs catégorielles. Grâce à l'utilisation de la méthode du coude, nous avons déterminé que le nombre optimal de clusters pour notre jeu de données était de 8. Nous avons ensuite appliqué l'algorithme K-means pour regrouper les données en fonction de leurs similarités.
L'analyse des résultats a révélé des clusters distincts, chacun présentant des caractéristiques spécifiques. Ces informations nous ont permis de formuler des descriptions concises et précises pour chaque cluster, mettant en évidence les facteurs clés qui les différencient.

Cependant, il convient de noter que notre étude comporte certaines limites. Les résultats et les conclusions obtenus sont spécifiques à notre jeu de données et peuvent ne pas être généralisables à d'autres ensembles de données.

2. Ouverture au projet suivant

Les connaissances acquises à travers cette analyse peuvent être utilisées pour prendre des décisions éclairées, améliorer les stratégies et les performances pour l'entreprise Béton Direct.

De plus, dans le cadre de notre étude, nous avons réalisé une analyse croisée avec les travaux du groupe de Yann et Iliesse, qui ont également exploré les mêmes données. Leur objectif principal était de prédire quels clients étaient les plus susceptibles de finaliser leur devis, afin de créer une classification des clients pour prioriser les rappels clients.

Dans cette perspective, nous avons développé notre propre approche de classification par ordre d'importance des clusters, en prenant en compte les caractéristiques et les regroupements que nous avons identifiés. Nous avons ensuite partagé ces résultats avec le groupe de Yann et Iliesse, afin de compléter leur travail et de fournir une vision plus globale de l'analyse des clients. Voici le classement que nous avons fourni :

Classement	Clusters
1	Cluster 5
2	Cluster 2
3	Cluster 4
4	Cluster 0
5	Cluster 6
6	Cluster 1
7	Cluster 3
8	Cluster 7

Figure 10 : classement des priorités des clusters

Cette collaboration entre nos deux équipes a permis d'enrichir nos résultats respectifs et d'ouvrir de nouvelles perspectives quant à l'optimisation des rappels clients. En combinant nos approches et en exploitant les synergies entre nos travaux, nous avons pu fournir des recommandations plus ciblées et pertinentes pour les actions de suivi à entreprendre.

IX. Table des illustrations

X. Annexes

Bibliothèque Scikit-learn :
https://scikit-learn.org/stable/index.html

Compte Youtube Machine Learnia :
https://www.youtube.com/@MachineLearnia

Mohamad KANAAN, Analyse des comportements des clients sur un site marchand en ligne, Lyon, Université Claude Bernard Lyon 1 (thèse de doctorat en informatique), 142 p.