Projet : ADD et Classification

Hicham Kaffaf

Table des matières

1 Introduction

Dans le cadre de ce projet, nous cherchons à déterminer des classes de LGD hétérogènes entre elles en mobilisant des méthodes de classification supervisées et non supervisées. L'objectif à terme est d'être en mesure de quantifier le niveau de pertes estimé en cas de défaut du client.

2 Traitement de données

Dans cette section, nous présentons les différentes étapes de traitement des variables que nous avons effectuées sur notre jeu de données. La base de données initiale est composée de 71 variables et de 184 739 lignes. Voici les modifications apportées à la base :

Type de Traitement	**Nombre de Variables**
Variables supprimées à cause des valeurs manquantes	26
Variables créées	10
Variables regroupées	6

TABLE 1 – Résumé du nombre de variables supprimées, créées et regroupées

Pour réduire le nombre de modalités et donner plus de sens aux variables, nous avons effectué les regroupements suivants :

- **purpose :**
  - **"crédit_consommation" :** éducation (educational), vacances (vacation), mariage (wedding), médical (medical), déménagement (moving)
  - **"achat_biens" :** achat important (major_purchase), voiture (car), maison (house)
  - **"amélioration_entretien" :** amélioration de l'habitat (home_improvement), énergie renouvelable (renewable_energy)
  - **"gestion_financière" :** consolidation de dettes (debt_consolidation), carte de crédit (credit_card)
  - **"autres" :** petite entreprise (small_business), autres (other)
- **emp_length :**
  - **"moins_de_3_ans" :** moins d'un an , 1 an, 2 ans.
  - **"3_à_5_ans" :** 3 ans, 4 ans , 5 ans
  - **"6_à_10_ans" :** 6 ans , 7 ans , 8 ans, 9 ans, plus de 10 ans
- **"addr_state" :**
  - **"Ouest" :** AZ, CA, CO, HI, ID, MT, NV, NM, OR, UT, WA, WY
  - **"Midwest" :** IA, IL, IN, KS, MI, MN, MO, NE, ND, OH, SD, WI
  - **"Sud" :** AL, AR, DC, DE, FL, GA, KY, LA, MD, MS, NC, OK, SC, TN, TX, VA, WV
  - **"Nord-Est" :** CT, MA, ME, NH, NJ, NY, PA, RI, VT
- **home_ownership :**
  - **"AUTRE" :** AUCUN, TOUT, AUTRE
  - Conserve les autres catégories telles quelles
- **verification_status :**
  - **"Vérifié" :** Vérifié, Source Vérifiée
  - **"Non Vérifié" :** Non Vérifié

Ces regroupements nous permettent de simplifier les analyses tout en conservant la pertinence des informations.

Pour garantir la qualité des données et minimiser l'impact des valeurs extrêmes, nous avons effectué les traitements suivants sur certaines variables : :

- **"revol_util"** :
  - **Traitement :** Les valeurs de `revol_util` supérieures à 100 ont été capées à 100.
- **"revol_bal"** :
  - **Traitement :** Les outliers ont été remplacés par le quantile 99%.
- **"annual_inc"** :
  - **Traitement :** Les outliers spécifiques ont été remplacés par le quantile 99%, avec certaines valeurs extrêmes fixées à 2,000,000.

Ces traitements nous permettent de minimiser l'impact des valeurs extrêmes et d'améliorer la qualité des analyses. Pour enrichir notre base de données et faciliter les analyses, nous avons créé les variables suivantes :

- **"principal_repaid_proportion"** :
  - **Description :** Proportion du principal remboursé calculée comme `total_rec_prncp / total_pymnt`.
  - **Raison :** Quantifier la part du principal remboursé.
- **"total_negative_incidents"** :
  - **Description :** Total des incidents négatifs calculé comme `collections_12_mths_ex_med + pub_rec`.
  - **Raison :** Quantifier les incidents négatifs dans l'historique de crédit.
- **"open_acc_proportion"** :
  - **Description :** Proportion de comptes ouverts calculée comme `open_acc / total_acc`.
  - **Raison :** Évaluer la gestion du crédit par le demandeur.
- **"credit_used"** :
  - **Description :** Crédit utilisé calculé comme `revol_bal * revol_util / 100`.
  - **Raison :** Trouver le crédit utilisé par le demandeur.
- **"total_rec"** :
  - **Description :** Montant total reçu, calculé comme `total_rec_int + total_rec_prncp`.
  - **Raison :** Fusionner les montants reçus et les intérêts reçus pour obtenir une vue globale des paiements reçus.

Après ces différents traitements, nous nous retrouvons avec 24 variables et 178,178 lignes.

3 Calcul de la LGD (Loss Given Default)

Le LGD (Loss Given Default) est calculé à l'aide de la formule suivante :

$$LGD = \frac{(\text{installment} \times \text{term}) - (\text{recoveries} + \text{total\_pymnt} - \text{collection\_recovery\_fee})}{\text{installment} \times \text{term}}$$

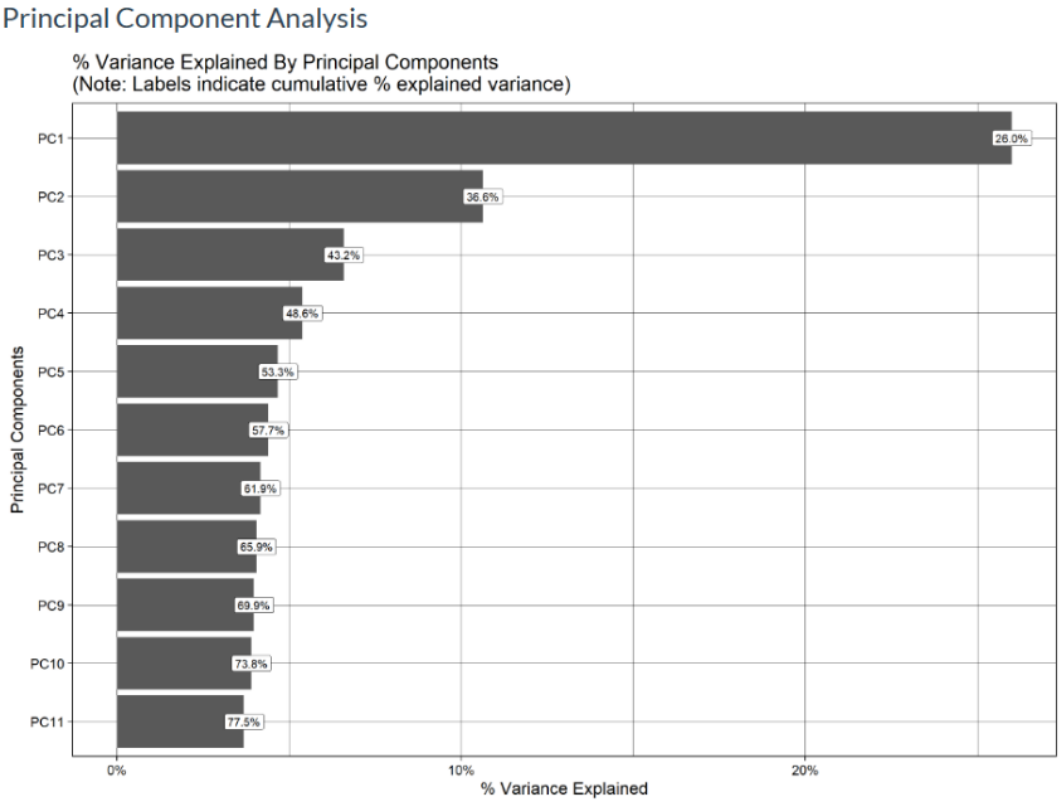
Signification des Variables :

- **"collection_recovery_fee"** : Frais de recouvrement post-défaut.
- **"installment"** : Le paiement mensuel dû par l'emprunteur si le prêt est accordé.
- **"recoveries"** : Recouvrement brut post-défaut.
- **"term"** : Le nombre de paiements sur le prêt. Les valeurs sont en mois et peuvent être soit 36, soit 60.
- **"total_pymnt"** : Paiements reçus à ce jour pour le montant total financé.

4 ACP et K-means

4.1 ACP

Dans un premier temps, nous avons choisi de mettre en œuvre une ACP afin de réduire la dimensionnalité des données, améliorer la visualisation, et identifier les variables les plus significatives pour notre analyse. À terme, cette méthode vise à faciliter l'interprétation et l'application des méthodes de clustering.



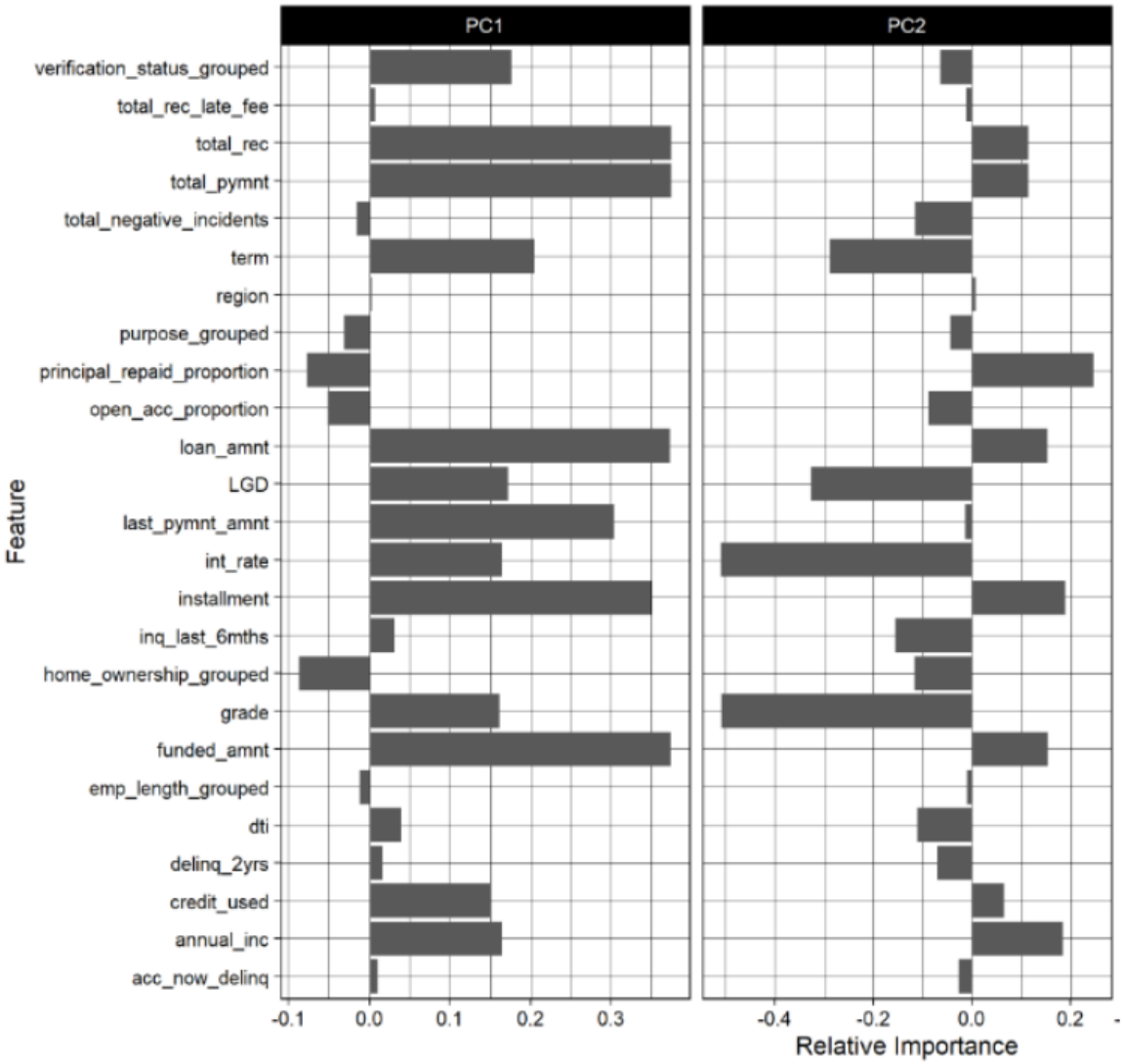
Comme indiqué précédemment, notre **PC1** représente 26 % de la variance totale, ce qui signifie que la première composante principale explique un quart de la variance observée. Elle constitue également la composante qui capture le maximum de variance, soulignant son importance dans notre analyse.

Notre deuxième composante principale explique quant à elle 10.6% de variance supplémentaire. Enfin, la composante principale 3 ajoute 6.6% de la variance.

Les composantes ultérieures, de la **PC4** à la **PC11**, expliquent chacune des portions décroissantes de la variance. La somme de la variance expliquée atteint 77.5 % avec la **PC11**, couvrant ainsi une majorité de la variance totale. Toutefois, une proportion restante de 22.5 % demeure inexpliquée par les composantes principales identifiées par notre modèle.

Toutefois, En utilisant les quatre premières composantes principales (**PC1** à **PC4**), nous parvenons à expliquer environ 48.6 % de la variance totale. Cela indique que près de la moitié des informations contenues dans les données initiales peut être capturée par seulement quatre composantes.

Analyse des variables contributrices



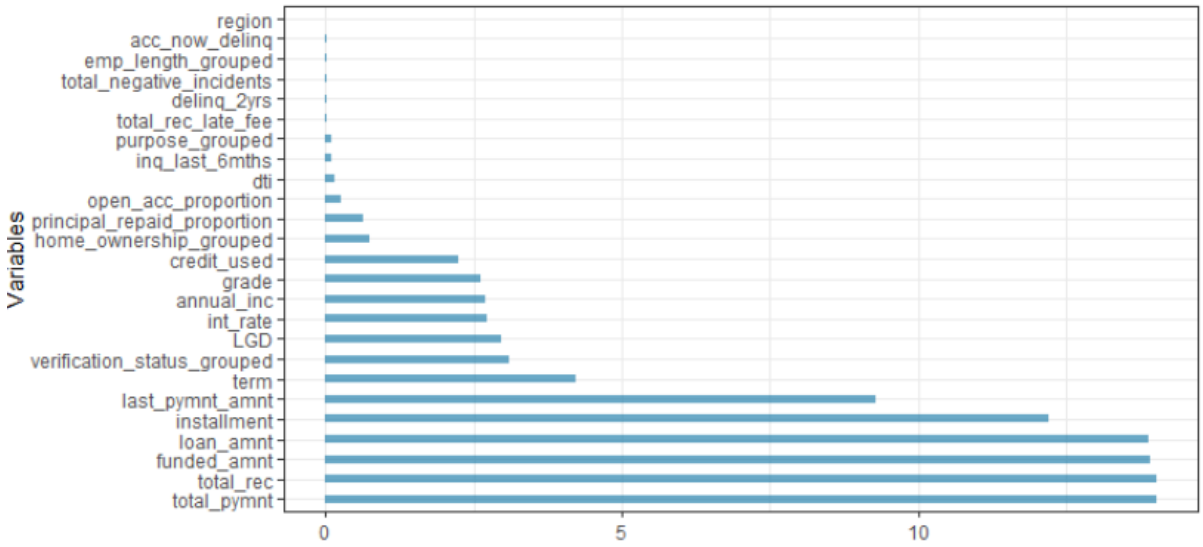
Les variables **"total_pymnt"** et **"total_rec"** contribuent fortement positivement à la **PC1**, indiquant ainsi qu'elles jouent un rôle prépondérant dans la variance capturée par cette composante.

On observe également que les variables**"loan_amnt"** et **"last_pymnt_amnt"** sont significatives pour la **PC1**, suggèrant que les montants des prêts et des derniers paiements sont des facteurs clés au sein de la première dimension. D'autre part, La variable **"total_rec_late_fee"** contribue positivement à **PC1**, ce qui indique que les frais de retard sont également significatifs. Par ailleurs, on constate que la durée de prêt est également un facteur important étant donné que la variable **"term"** montre une contribution plutôt notable.

Concernant la deuxième composante principale, la variable **"int_rate"** dispose d'une forte contribution positive, ce qui montre que la prise en compte des taux d'intérêt est primordiale au sein de cette dimension. De plus, les variables **"loan_amnt"** et **"funded_amnt"** contribuent positivement à **PC2**, indiquant que les montants des prêts et des financements sont également importants pour cette composante. Enfin, nous constatons également la contributions significatives des variables **"open_acc_proportion"** et **"principal_repaid_proportion"**.

La **PC1** est dominée principalement par des variables financières liées aux paiements et aux montants des prêts. Cela pourrait être interprété comme une dimension reflétant l'ampleur des transactions financières. La **PC2** quant à elle semble fortement influencée par le taux d'intérêt et les proportions de comptes ouverts et de principal remboursé. Cette dimension pourrait alors refléter des aspects liés à la gestion du crédit et des conditions d'octroi du prêt.

Analyse de la contribution des variables

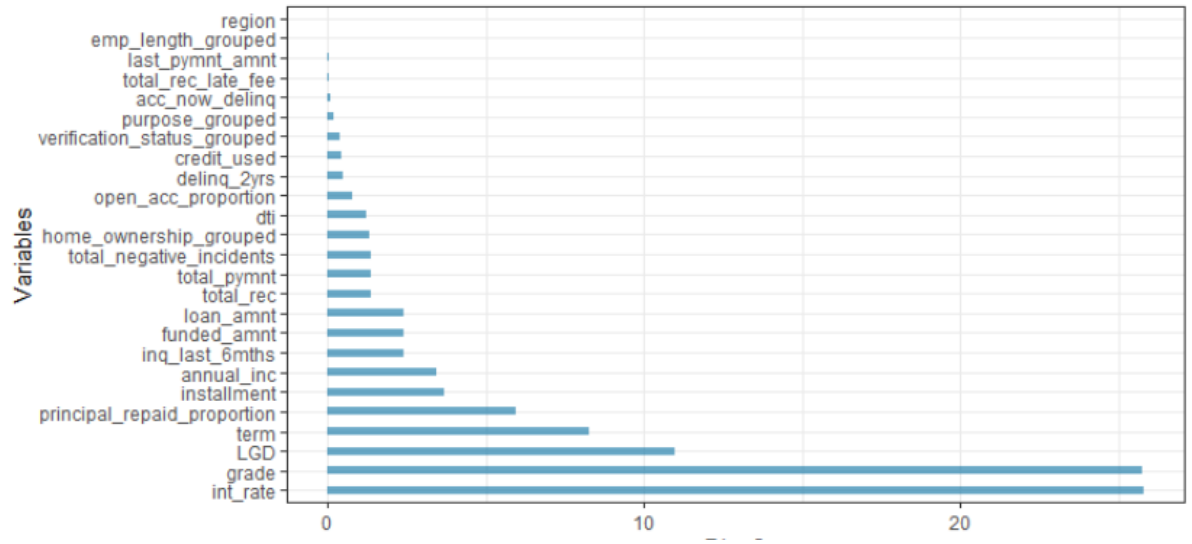


Ce graphique illustre la contribution des différentes variables aux composantes principales. Les variables ayant les plus fortes contributions jouent un rôle significatif dans la variance expliquée par les premières CP.

Ainsi, les variables **"total_pymnt"** et **"total_rec"** affichent les contributions les plus élevées. Cela démontre leur importance cruciale dans l'explication de la variance totale des données, car elles capturent une part importante de l'information. Par ailleurs, **"funded_amnt"** et **"loan_amnt"**, qui sont des indicateurs des montants des prêts, sont également très influentes et essentielles pour comprendre la structure des données.

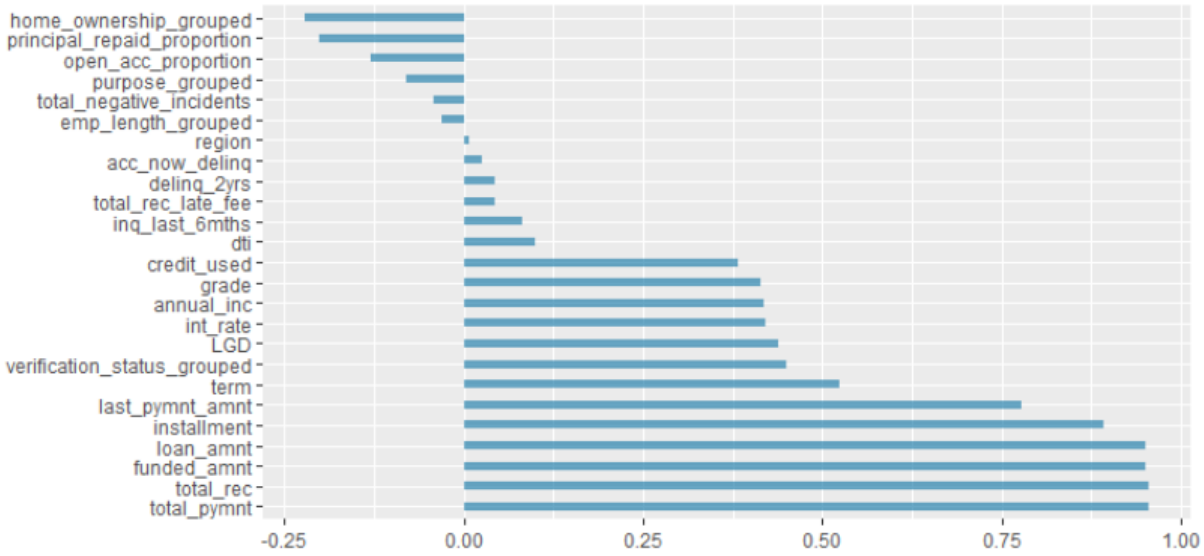
Les variables **"installment"** et **"last_pymnt_amnt"**, représentant respectivement les paiements réguliers et récents, montrent aussi des contributions significatives. Cela souligne leur rôle dans la capture de la variance des données. En outre, la durée du prêt, désignée par la variable term, et la variable de perte en cas de défaut (LGD), sont également importantes.

Les variables **"region", "acc_now_delinqu"**, et **"emp_length_grouped"** montrent les plus faibles contributions, ajoutant peu d'information supplémentaire dans le cadre de l'ACP. De même, les indicateurs de comportement de crédit passé, comme **"total_negative_incidents"** et **"delinq_2yrs"**, affichent des contributions minimes.



Ce deuxième graphique illustre la contribution de chaque variable à la deuxième composante principale. En effet, on observe que les contributions les plus élevées indiquent que la variable a un impact plus significatif sur la variance expliquée. À ce titre, les variables **"int_rate"**, **"grade"**, et **"LGD"** sont les plus contributives dans la deuxième composante principale, capturant une grande partie de la variance.

Analyse de la corrélation des variables

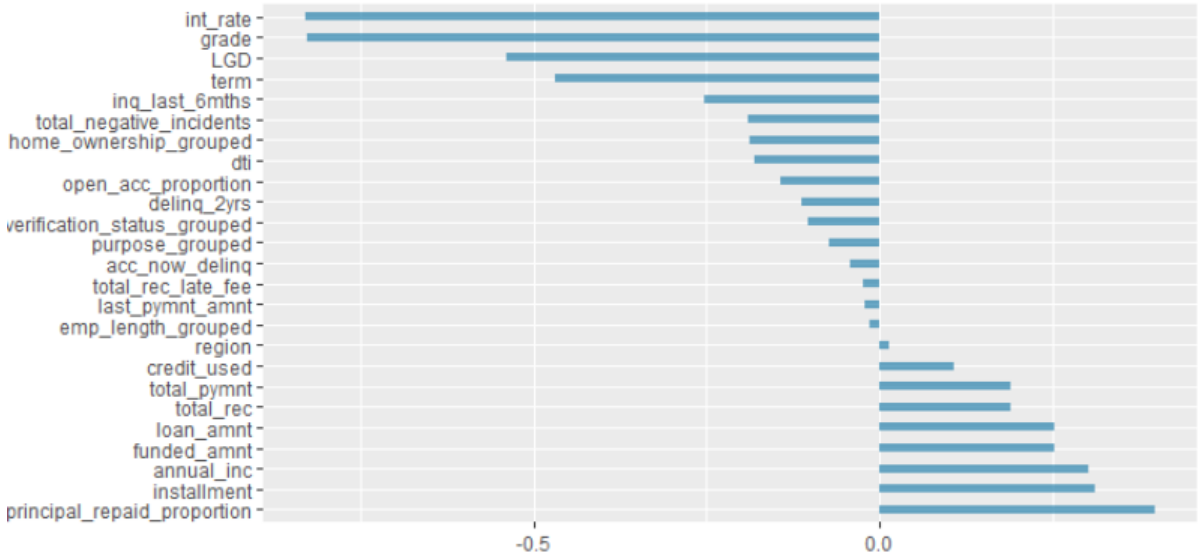


Ce graphique nous indique la corrélation de chaque variable avec la **PC1**. En effet, les corrélations plus élevées indiquent que la variable est fortement liée à la composante principale, ce qui signifie qu'elle contribue de manière significative à la variance expliquée par cette composante.

Les variables **"total_pymnt"** et **"total_rec"** affichent les corrélations les plus élevées avec la **PC1**, soulignant leur rôle déterminant dans l'explication de la variance et leur forte association avec cette composante. De même, **"funded_amnt"** et **"loan_amnt"**, qui représentent les montants des prêts, montrent également des corrélations élevées, indiquant leur importance essentielle pour comprendre la structure de la première composante.

Les variables **"installment"** et **"last_pymnt_amnt"**, représentant les paiements réguliers et récents, ont des corrélations significatives avec la **PC1**, soulignant leur pertinence. Par ailleurs, term et verification_status_grouped démontrent aussi des corrélations notables, affirmant leur importance pour la **PC1**.

Ainsi, il semblerais que les variables **"total_pymnt"**, **"total_rec"**, **"funded_amnt"**, et **"loan_amnt"** sont celles qui présentent les plus fortes corrélations avec la **PC1**.

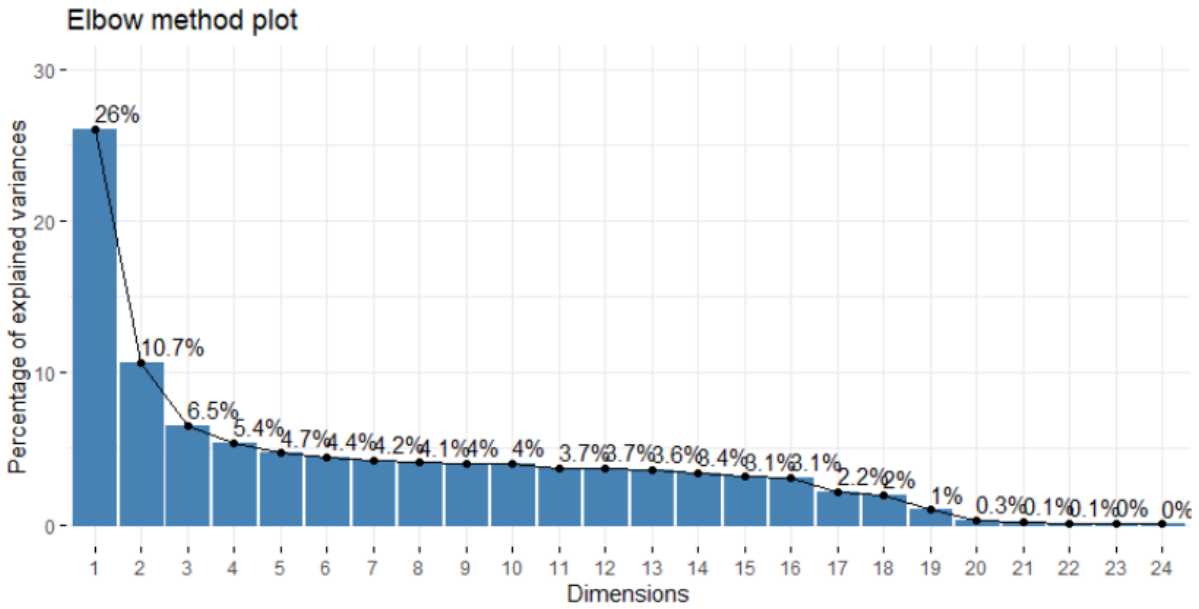


Ce graphique quant à lui illustre la corrélation de chaque variable avec la PC2.

Ici, les variables **"int_rate"** et **"grade"** présentent les corrélations les plus élevées avec la **PC2**. Le taux d'intérêt et la notation de crédit sont ainsi identifiés comme des facteurs dominants dans cette dimension. De plus, la perte en cas de défaut (LGD) montre également une forte corrélation.

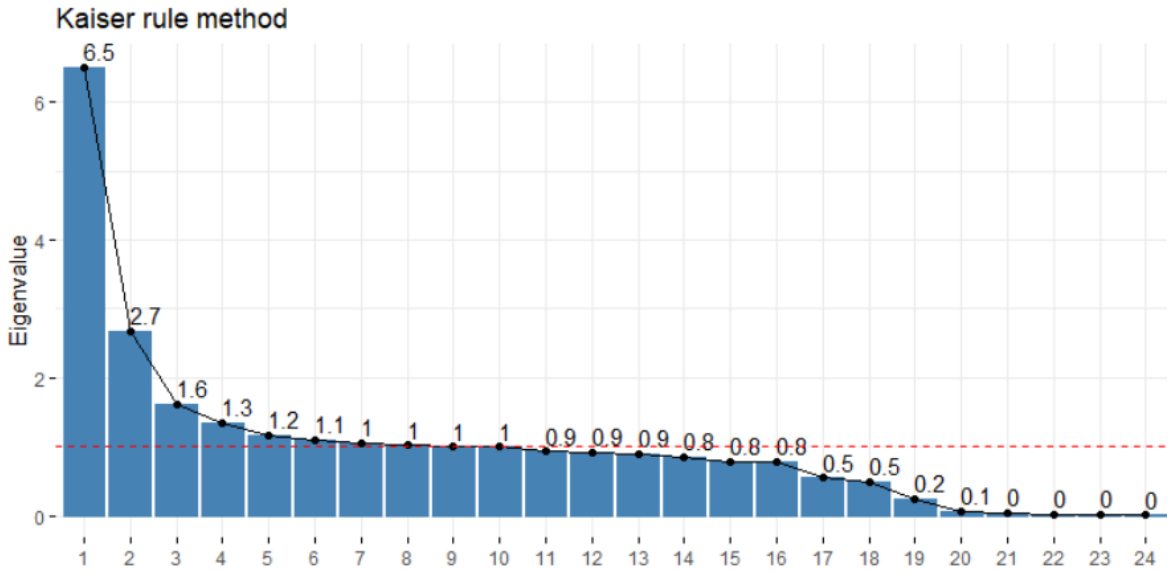
Les variables **"term"** et **"inq_last_6mths"** affichent également des corrélations significatives avec la **PC2**, ce qui souligne leur pertinence dans cette composante. De plus, les variables **"total_negative_incidents"** et **"home_ownership_grouped"** affichent aussi des corrélations notables.

Méthode du coude

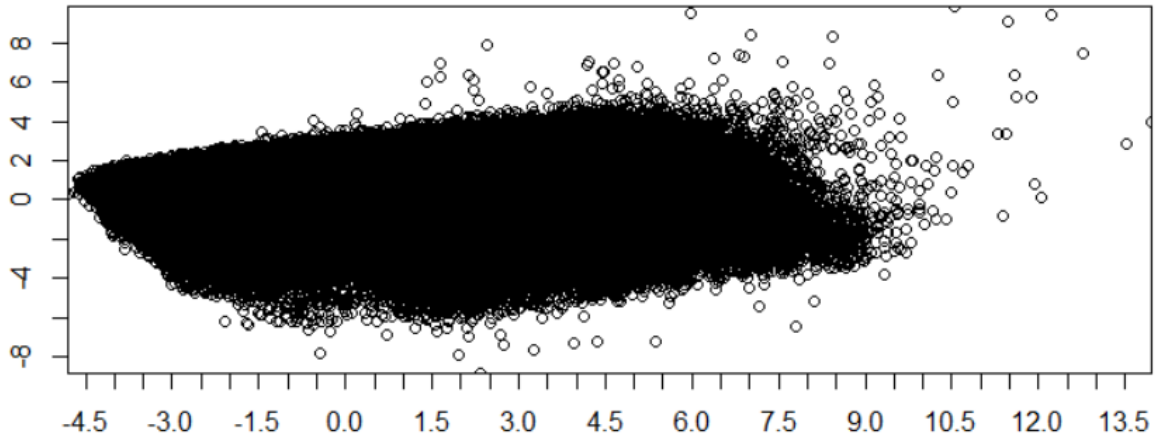


La méthode du coude suggère que conserver les composantes 3 et 4 pourrait être optimal, capturant ainsi environ 43,2% à 47,4% de la variance totale tout en simplifiant le modèle.

Règle de Kaiser



Selon la règle de Kaiser, seules les composantes avec des valeurs propres supérieures à 1 sont significatives. Dans ce graphique, cela inclut jusqu'à **PC6**, suggérant que ces sept composantes principales sont significatives et utiles pour capturer une part majeure de la variance tout en réduisant la dimensionnalité.

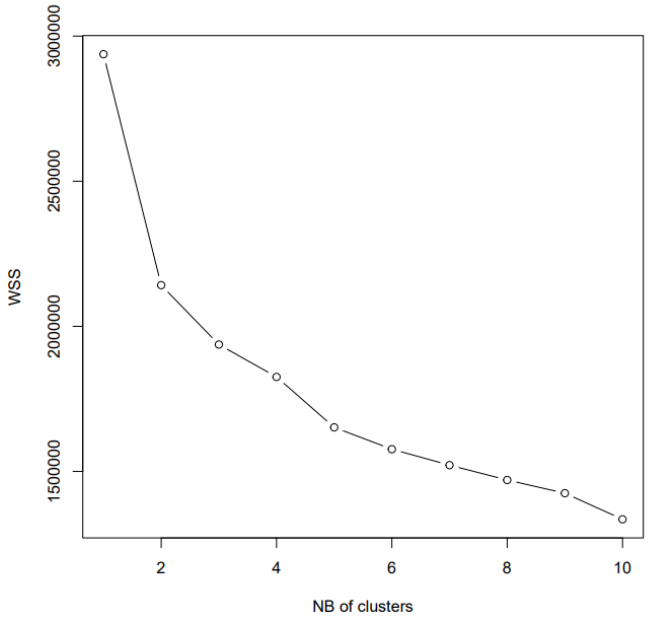


Ce graphique représente la répartition des données projetées sur les 2 premières CP, soient **CP1** en abscisses et **CP2** en ordonnées. La distribution des données montre une variance significative selon **PC1**, sans regroupements évidents à l'œil nu, et une variance plus modérée selon **PC2**. Quelques points éloignés peuvent potentiellement influencer l'analyse. Le nuage de points ne révèle pas de clusters évidents ni de séparations nettes, suggérant une homogénéité relative des données ou des clusters subtils nécessitant une analyse approfondie. En partant de ces postulat, nous allons utiliser la méthode des K-means pour partitionner les données en K-clusters, permettant d'identifier et d'analyser leur distribution.

4.2 K-Means

Initialement, nous avions prévu d'utiliser la Classification Ascendante Hiérarchique (CAH) pour créer les groupes. Cependant, en raison d'un problème d'allocation de mémoire, cette méthode n'a pas pu être réalisée. Nous avons donc décidé de nous orienter directement vers la méthode K-means.

Nous avons effectué une série d'analyses en faisant varier le nombre de clusters de 1 à 10 pour notre algorithme k-means. Pour chaque nombre de clusters, nous avons calculé la somme des carrés des distances intra-clusters (WSS). Ces calculs nous ont permis de représenter graphiquement la relation entre le nombre de clusters et la variance intra-cluster (WSS). L'objectif de cette analyse était de déterminer le nombre optimal de clusters en utilisant le critère du coude.



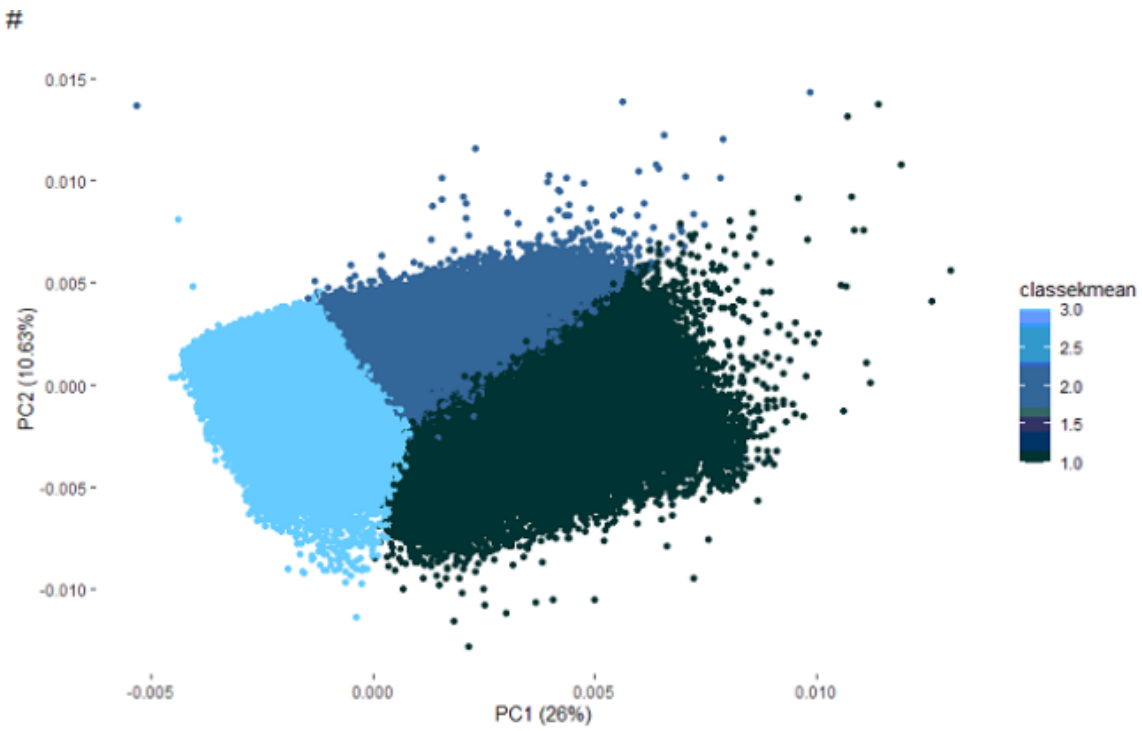
En regardant le graphique, on observe une forte diminution de la WSS entre 1 et 3 clusters. Après 3 clusters, la diminution de la WSS commence à ralentir de manière significative. Le point de coude se situe autour de 3 clusters. Cela signifie que 3 clusters représentent un bon équilibre entre la réduction de la variance intra-cluster et la simplicité du modèle.

— Nombre d'individus dans chaque cluster :

  — **Cluster 1** : 30,924 individus
  — **Cluster 2** : 44,809 individus
  — **Cluster 3** : 102,445 individus

— Nombre d'observations total : 178,178

Ce graphique ci-dessous nous montre la répartition des individus dans les trois clusters sur les deux premières composantes principales (**PC1** et **PC2**). Chaque couleur représentant un cluster différent, permettant de distinguer la séparation des groupes dans l'espace des composantes principales. Les clusters sont séparés mais montrent des zones de chevauchement. Cela indique que, bien que distincts, les clusters ne sont pas complètement isolés les uns des autres.



5 Méthode supervisée

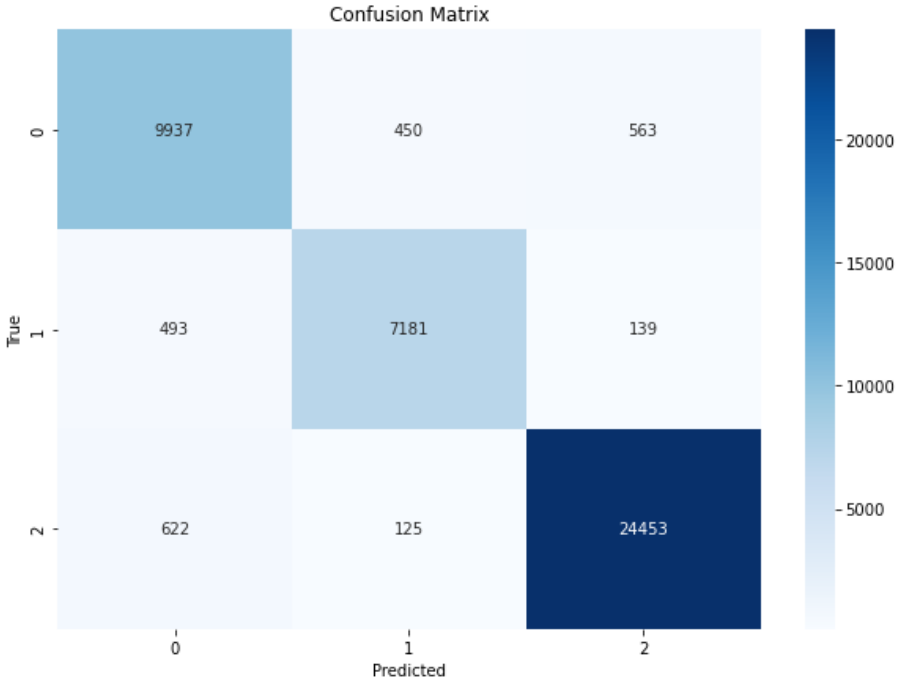
5.1 Zoom sur les variables conservées et non-séléctionnées

Nous n'avons pas sélectionné ces variables car elles ne sont connus avant le processus d'inititation du prêt.

Variables	Description
funded_amnt	C'est le montant effectivement financé pour le prêt, qui peut varier par rapport au montant initialement demandé. Ce montant est déterminé après l'approbation du prêt et n'est pas une donnée initiale.
total_pymnt	Le total des paiements reçus à ce jour. Cette variable dépend des paiements reçus après l'initiation du prêt.
total_rec_late_fee	Les frais de retard reçus à ce jour. Ces frais sont connus uniquement après que des paiements en retard aient été effectués.
last_pymnt_amnt	Le montant du dernier paiement reçu. Cela dépend des paiements effectués par le client après le début du prêt.
principal_repaid_proportion	Proportion du principal remboursé, calculée en fonction des paiements reçus après l'initiation du prêt.
total_rec	Total des paiements reçus. Cela reflète les paiements effectués après le début du prêt.
acc_now_delinq	Le nombre de comptes actuellement en souffrance. Cela peut être dynamique et changer avec le temps, mais généralement ce n'est pas connu avant le processus de recouvrement.

5.2 Arbre de descisions

L'objectif ici est de classifier les données par cluster. Les variables explicatives sont les 20 variables restantes (term, annuel income, ...). On divise les données en un ensemble de test et un ensemble d'entraînent (75%, 25%). Un modèle d'arbre de décisions est ensuite entraîné puis testé sur l'échantillon test.

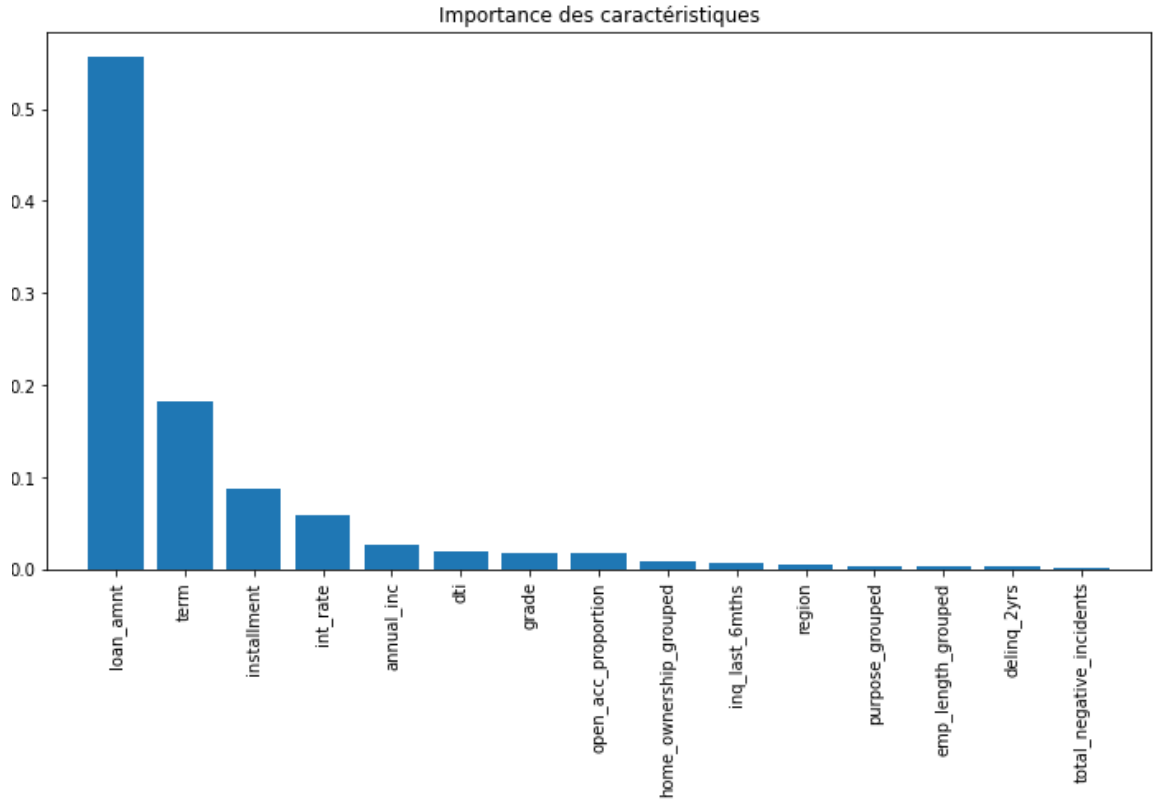


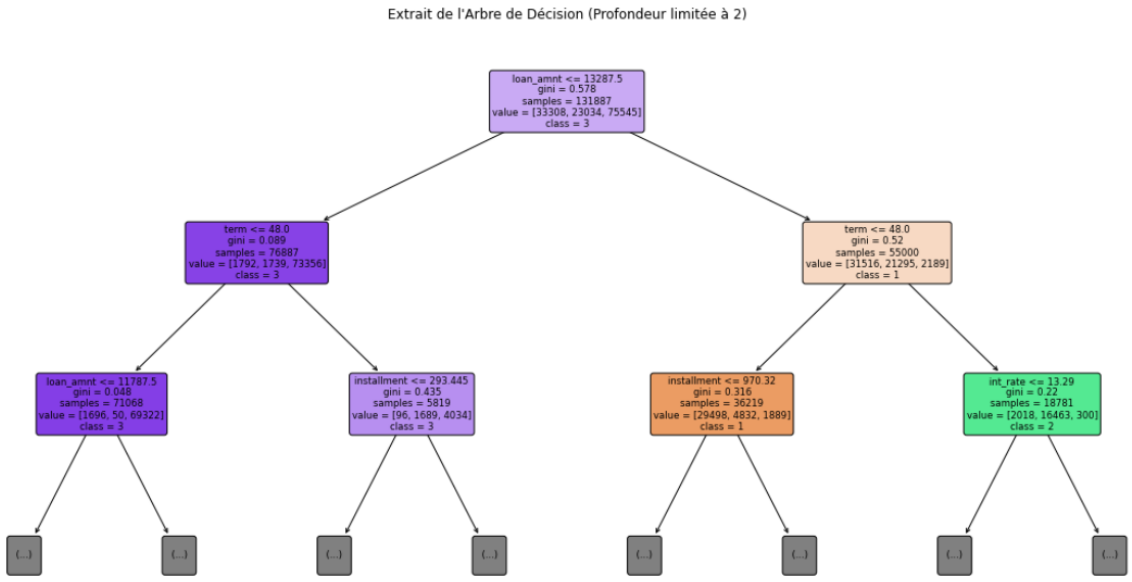
Cluster	Précision
1	0.90
2	0.93
3	0.97

Table 2 – Précision par clusters

La matrice de confusion indique que pour la classe 0, le modèle a fait 9937 prédictions correctes et a commis quelques erreurs en classant certains échantillons dans les classes 1 et 2. Pour la classe 1, le modèle a correctement prédit 7181 fois, et pour la classe 2, il a fait 24453 bonnes prédictions. Ces résultats sont également reflétés dans les scores de précision élevés pour chaque classe : 0.90 pour la classe 0, 0.93 pour la classe 1, et 0.97 pour la classe 2.

On détermine également ici que le montant du prêt **"loan_amnt"** est de loin la caractéristique la plus importante, suivi par la durée du prêt **"term"** et le montant des mensualités **"installment"**. D'autres caractéristiques comme le taux d'intérêt **"int_rate"**, le revenu annuel **"annual_inc"** et le grade **"grade"** ont également une certaine importance, mais moindre. Les caractéristiques restantes ont un impact négligeable sur les prédictions du modèle. En partant de ce postulat, on relance le modèle qu'avec les variables les plus importante pour que le modèle soit au final plus interpretable. Il se trouve que malgrès avoir enlevé 9 variables, le modèle n'a pas perdu son pouvoir discriminant, le score de précisions du modèle s'inscrit à 94%.





Enfin, la visualisation de l'arbre de décision montre que le nœud racine est basée sur le montant du prêt **"loan_amnt"**. Si ce montant est inférieur ou égal à 13287.5, la majorité des échantillons appartiennent à la classe 3. Sinon, la décision se poursuit alors vers les classes 1 ou 2. À contrario, sur la branche gauche, les échantillons avec un **"loan_amnt"** inférieur ou égal à 13287.5 sont majoritairement classés en classe 3, en particulier si la durée du prêt **"term"** est inférieure ou égale à 48.0. Sur la branche droite, les échantillons avec un **"loan_amnt"** supérieur à 13287.5 sont répartis entre les classes 1 et 2, avec le **"term"** et les mensualités **"installment"** influençant cette séparation. L'indice de Gini est utilisé pour évaluer la pureté des nœuds : les nœuds avec un Gini faible (proche de 0) sont plus purs, contenant principalement une seule classe, tandis que ceux avec un Gini élevé (proche de 0.5) sont plus mixtes.

6 Analyse des résultats obtenu

6.1 Tests statistiques

Dans cette analyse, nous avons effectué une série de tests statistiques pour évaluer l'hétérogénéité et les différences entre les groupes formés par l'algorithme de clustering K-means. Les variables analysées incluent des variables numériques et qualitatives, telles que *loan_amnt*, *int_rate*, *installment*, *annual_inc*, *LGD*, et *grade*.

Les tests suivants ont été réalisés :

- **Test de Bartlett** : pour vérifier l'homogénéité des variances entre les groupes pour les variables numériques (*loan_amnt*, *int_rate*, *installment*, *annual_inc*, *LGD*).
- **Test du chi-deux** : pour tester l'indépendance entre la variable catégorielle *grade* et les groupes formés.
- **Test de Kruskal-Wallis** : pour comparer les distributions entre les groupes pour les variables numériques (*loan_amnt*, *int_rate*, *installment*, *annual_inc*, *LGD*).
- **Test ANOVA de Welch** : pour comparer les moyennes entre les groupes lorsque les variances sont inégales pour les variables numériques (*loan_amnt*, *int_rate*, *installment*, *annual_inc*, *LGD*).
- **Test de Shapiro-Wilk** : pour tester la normalité des données pour les variables numériques (*loan_amnt*, *int_rate*, *installment*, *annual_inc*, *LGD*).
- **Test post-hoc de Dunn** : pour identifier les groupes spécifiques qui diffèrent après un test de Kruskal-Wallis pour les variables numériques.
- **Test post-hoc de Tukey-Kramer** : pour des comparaisons multiples après une ANOVA pour les variables numériques.

Les résultats de ces tests sont résumés dans le tableau ci-dessous.

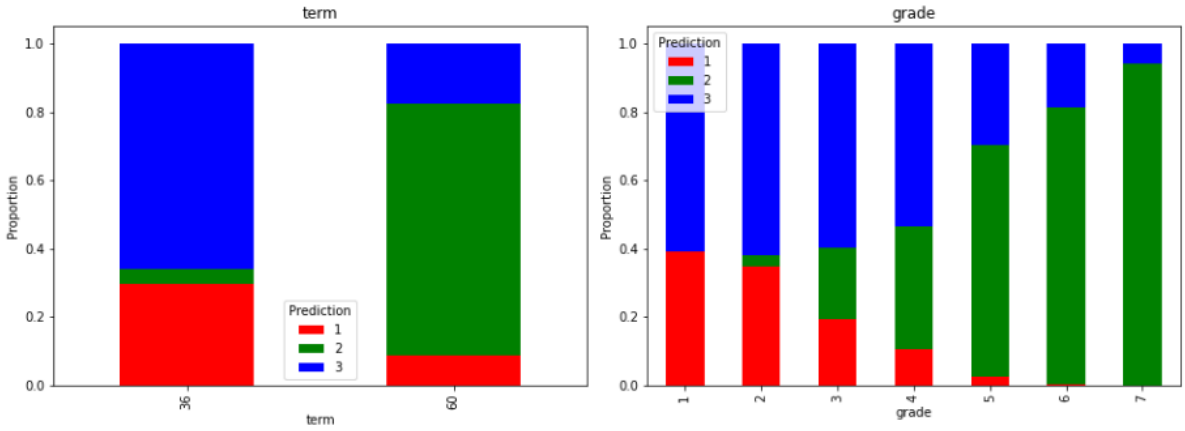
Test	**Non rejet H0**	**Rejet H0**
Bartlett test	-	19
Khi-deux test	-	1
Kruskal-Wallis test	-	19
One way Welch-ANOVA test	-	19
Shapiro-Wilk test	-	19
Dunn test	4	53
Tukey-Kramer test	3	54

TABLE 3 – Résultats des tests statistiques

Les résultats indiquent que les groupes formés par l'algorithme K-means sont bien distincts et présentent des différences significatives en termes de variances, distributions, et moyennes pour les variables analysées. Ces distinctions confirment l'efficacité du clustering dans la segmentation des données en groupes significativement différents.

6.2 Analyse des résultats par visualisation graphique

Variables catégorielles



Les graphiques montrent que les termes de 36 mois sont principalement dans le cluster 3, tandis que ceux de 60 mois sont majoritairement dans le cluster 2. Les grades A et B sont dominés par le cluster 3, tandis que les grades C à G sont principalement dans le cluster 2. Les grades F et G sont exclusivement dans le cluster 2.

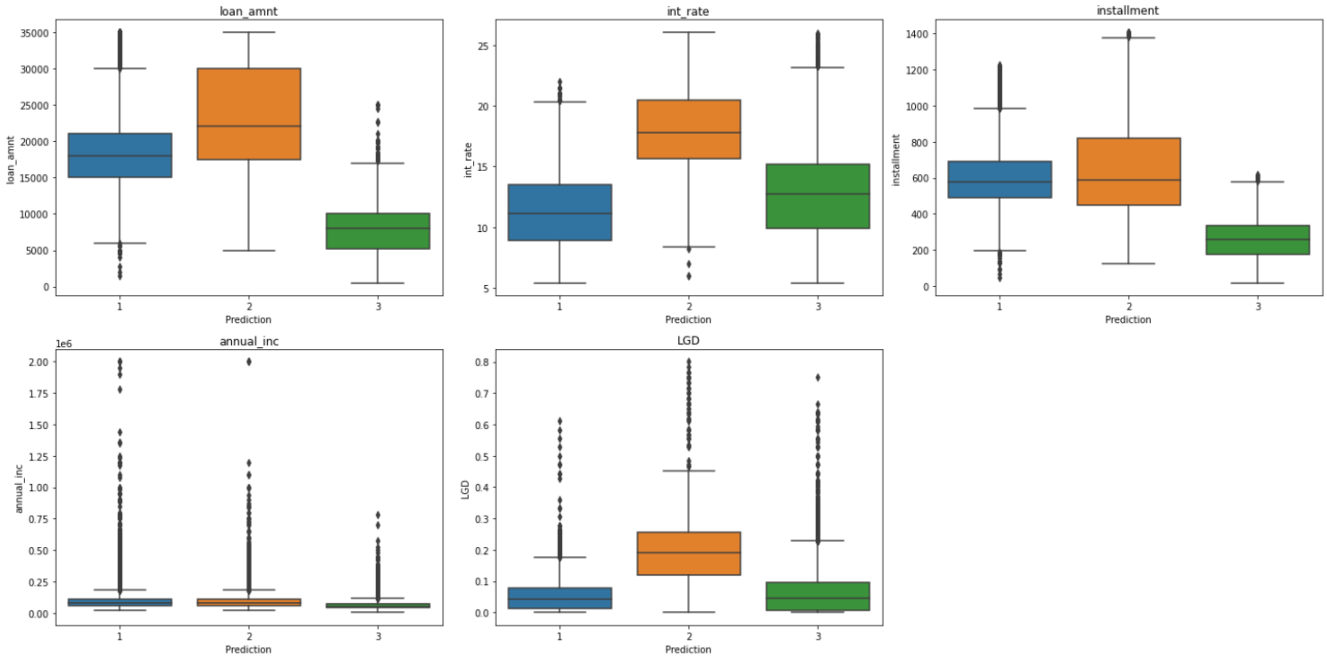
VARIABLE	CATEGORY	Number	1	2	3
term	36	%	21	94	94
term	60	%	79	6	6
grade	A	%	0	30	21
grade	B	%	6	45	36
grade	C	%	30	18	26
grade	D	%	29	6	13
grade	E	%	22	1	3
grade	F	%	10	0	1
grade	G	%	3	0	0

TABLE 4 – Répartition des clusters selon les variables *term* et *grade*

Variables quantitatives

Vname	Cluster	mean	SD
annual_inc	3	59,573	30,085
annual_inc	1	94,251	61,391
annual_inc	2	94,723	64,357
installment	3	257	107
installment	1	652	262
installment	2	609	166
LGD	3	0.060	0.06
LGD	1	0.190	0.10
LGD	2	0.050	0.05
loan_amnt	3	7,853	3,237
loan_amnt	1	23,233	7,293
loan_amnt	2	18,946	4,835

TABLE 5 – Statistiques descriptives par cluster



Cluster 1

Les emprunteurs de ce groupe semblent avoir des prêts de montants moyens, des taux d'intérêt bas et des versements moyens. Le revenu annuel médian est bas mais présente une grande variabilité, et la perte en cas de défaut est faible.

Cluster 2

Ce groupe a des prêts de montants élevés, des taux d'intérêt plus élevés et des versements plus élevés. Le revenu annuel est similaire aux autres clusters, mais la perte en cas de défaut est légèrement plus élevée.

Cluster 3

Ce groupe a des prêts de montants plus bas, avec des taux d'intérêt similaires à ceux du cluster 2 mais avec des versements plus bas. Le revenu annuel est bas et la perte en cas de défaut est similaire à celle du cluster 1.

7 Axes d'ouverture

Dans le cadre de ce projet, plusieurs axes d'ouverture intéressants n'ont pas pu être explorés en raison de contraintes de temps. Voici les principales pistes envisagées :

Une optimisation des paramètres du modèle K-Means, comme le choix du nombre optimal de clusters et l'utilisation de K-Means++, pourrait potentiellement améliorer la précision et la stabilité du clustering. De plus, l'exploration d'autres méthodes de clustering telles que les modèles de mélange gaussien (GMM), DBSCAN ou la classification hiérarchique agglomérative (CAH) que nous avons prévu d'utiliser, aurait pu offrir des perspectives complémentaires. Notons que la méthode CAH n'a pas pu être utilisée en raison de problèmes d'allocation de mémoire, et nous sommes donc passés directement à la méthode K-Means.

La mise en œuvre de techniques de validation croisée plus exhaustives et l'évaluation de modèles avec des métriques de performance variées (silhouette score, Davies-Bouldin index) auraient également permis une évaluation plus rigoureuse des performances des modèles. Le fait d'intégrer de nouvelles variables explicatives ou des sources externes, comme des données macroéconomiques, aurait pu affiner les analyses et offrir une vue plus complète du risque de défaut.

Par ailleurs, le développement de modèles prédictifs avancés (réseaux de neurones, méthodes d'ensemble, deep learning) pourrait également offrir des prédictions plus précises en capturant des relations complexes dans les données.

Enfin, la réalisation d'analyses de scénarios et de tests de résistance (stress testing) évaluerait la robustesse des clusters et des modèles sous différentes conditions économiques, fournissant des informations précieuses pour la gestion du risque.

En conclusion, bien que ces axes n'aient pas pu être explorés, ils représentent des pistes prometteuses pour des travaux futurs, enrichissant les analyses et améliorant la gestion du risque de crédit bancaire.

$$\text{LD1} = 0.09\times\text{age}+0.12\times\text{menopause}+0.09\times\text{tumor\_size}+0.12\times\text{inv\_nodes}+0.49\times\text{node\_caps}+1.03\times\text{deg\_malig}-0.43\times$$