

UNIVERSITÉ D'ALGER 1
BENYOUCEF BENKHEDDA



جامعة الجزائر 1
بن يوسف بن خدة

République Algérienne Démocratique et Populaire

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université d'Alger 1 BENYOUCEF Benkhedda

(Faculté des Sciences)

Département Informatique

Première année Master (ISII)

Module : DATA MINING

Machine Learning Project : Regression et
Classification

Réalisé par :

HASSAINE Nazim G1

ASMANE Abderrahim G2

Années universitaire 2024/2025

Introduction

Ce projet a pour objectif l'exploration et l'analyse d'un jeu de données portant sur des alliages métalliques amorphes, afin d'en extraire des structures sous-jacentes à travers des techniques de classification non supervisée.

Les matériaux métalliques amorphes présentent des propriétés physico-chimiques particulières qui dépendent fortement de leur composition élémentaire et de leur traitement thermique. À travers l'étude de ces données, l'objectif est d'identifier des regroupements naturels d'alliages possédant des caractéristiques similaires, en s'appuyant sur des méthodes d'analyse de données.

Sur le plan technique, ce projet repose sur plusieurs étapes fondamentales du processus de Data Mining : le nettoyage et la transformation des données brutes, la gestion des valeurs manquantes et des valeurs extrêmes (outliers), la normalisation des variables, l'extraction des composantes chimiques, puis l'application de deux techniques de clustering : K-Means et le clustering hiérarchique agglomératif vue en cours.

La méthodologie adoptée suit une approche progressive :

- une phase de prétraitement des données, avec extraction et mise en forme des éléments chimiques présents dans les alliages.
- une phase d'analyse descriptive et de préparation des variables numériques (T_g , T_x , T_l , D_{max}), incluant la détection et le traitement d'outliers.
- une phase de normalisation par Z-score sur les variables pertinentes.
- enfin, une phase d'analyse par clustering, permettant de regrouper les alliages selon leurs similitudes structurelles et thermiques.

Ce rapport présente l'ensemble de ces étapes, illustrées par des visualisations, des justifications méthodologiques, et une interprétation des résultats obtenus.

Introduction.....	2
1) Prétraitement des données.....	4
1.1 Description du fichier source.....	4
1.2 Nettoyage et restructuration du DataFrame.....	4
1.2.1 Traitement de la colonne "Composition".....	5
1.2.2 Suppression des élément chimique inexistant.....	5
1.2.3 Suppression des compositions dépassant 100 %.....	5
1.3 Exemple de transformation.....	5
1.4 Format final du DataFrame.....	6
2) Analyse exploratoire et gestion des outliers.....	6
2.1 Statistiques descriptives.....	6
2.2 Vérification des valeurs manquantes.....	6
2.3 Détection des valeurs extrêmes (outliers).....	7
2.4 Tableau récapitulatif.....	7
2.5 Analyse et traitement des outliers.....	8
3) Normalisation des données.....	8
3.1 Méthode utilisée : Z-score standardization.....	8
3.2 Variables concernées.....	9
4) Clustering avec K-Means.....	9
4.1 Objectif.....	9
4.2 Sélection des variables.....	9
4.3 Choix du nombre de clusters (k).....	10
4.3.1 Méthode du coude (Elbow method).....	10
4.3.2 Méthode du score silhouette.....	10
4.4 Application de K-Means.....	11
4.4.1 Répartition des clusters.....	12
4.4.2 Analyse chimique des clusters.....	13
4.4.2.1 Tg vs Tx, Tl.....	13
4.4.2.2 Dmax vs les températures.....	13
4.4.2.3 Comportement par cluster.....	14
4.4.2.4 Interprétation finale.....	14
4.4.3 Bilan.....	14
5) Clustering hiérarchique.....	14
5.1 Objectif.....	14
5.2 Application de clustering hiérarchique.....	15
6) Conclusion général.....	15

1) Prétraitement des données

L'étape de prétraitement est essentielle pour garantir la qualité et la cohérence des données avant toute analyse. Le fichier source initial se présentait sous forme d'un tableau CSV contenant des informations sur les propriétés thermiques et la composition chimique de différents alliages métalliques.

1.1 Description du fichier source

Le fichier d'origine contenait environ 600 lignes et 11 colonnes. On pouvait distinguer deux sous-ensembles de colonnes :

- Colonnes 1 à 6 : regroupant des informations telles que le nom du matériau, les températures caractéristiques (Tg, Tx, Tl), la densité maximale (Dmax) ainsi que des remarques ou identifiants.
- Colonnes 7 à 11 : semblant correspondre à un second tableau (similaire au premier).

Certaines colonnes n'étaient pas nommées correctement ou contenaient des données incomplètes.

1.2 Nettoyage et restructuration du DataFrame

Nous avons effectué les opérations suivantes :

- Suppression de la première ligne (qui contenait des titres mal positionnés).
- Suppression de la première colonne (index dupliqué ou vide).
- Suppression des colonnes inutiles (comme les remarques ou les colonnes totalement vides).
- Regroupement des deux sous-tableaux en concaténant les lignes extraites des colonnes 7 à 11 sous celles des colonnes 1 à 6 (restructuration verticale).
- Réinitialisation de l'index pour obtenir un DataFrame cohérent.

À l'issue de cette étape, un DataFrame propre contenant les colonnes suivantes a été obtenu :

- Tg (Température de transition vitreuse)
- Tx (Température de cristallisation)
- Tl (Température de fusion)
- Dmax (Densité maximale ou diamètre max on reviendra plus bas)
- Composition (description textuelle des éléments chimiques présents et de leurs proportions %)

1.2.1 Traitement de la colonne "Composition"

La colonne "Composition" présentait des formats hétérogènes :

- Format simple : Ag30.8 Ca30.8 Mg23.1 Cu15.4
- Format imbriqué : {[Fe60 Co40]75 B20 Si5]96 Nb4}99 Cr1

Un parsing automatique a été mis en place pour extraire chaque élément chimique et son pourcentage, même dans le cas d'imbrications avec des coefficients d'application (par exemple : (Fe60 Co40)75 signifiant 75% d'un sous-alliage Fe/Co réparti à 60% et 40%).

Cette extraction a permis de transformer la colonne "Composition" en un ensemble de colonnes, une pour chaque élément chimique rencontré (ex. Ag, Ca, Mg, Fe, Co...), contenant le pourcentage réel de chaque élément dans l'alliage.

1.2.2 Suppression des élément chimique inexistant

Afin de garantir la cohérence chimique des données, nous avons vérifié que tous les éléments présents dans les compositions des verres appartenaient à la liste officielle des 118 éléments du tableau périodique. Pour ce faire, une liste de référence contenant les symboles normalisés (par exemple : "Si", "O", "Na", etc.) a été utilisée.

Chaque composition a été analysée individuellement. Si un élément présent ne figurait pas dans cette liste de référence, il a été considéré comme invalide.

1.2.3 Suppression des compositions dépassant 100 %

Afin de garantir la cohérence physico-chimique des données, nous avons vérifié que la somme des pourcentages massiques des éléments de chaque alliage ne dépasse pas 100 %. En effet, dans une composition réelle, la somme des éléments constitutifs d'un matériau doit idéalement être égale à 100 %.

Pour cela, une tolérance de 0.1 % a été appliquée pour tenir compte des arrondis. Les lignes dont la somme des éléments excédait 100.1 % ont été identifiées comme erronées et supprimées du jeu de données.

1.3 Exemple de transformation

Ligne brute (colonne "Composition") :

{[(Fe60 Co40)75 B20 Si5]96 Nb4}99 Cr1

Ligne transformée :

Fe: 43.2 Co: 28.8 B: 19.2 Si: 4.8 Nb: 3.96 Cr: 0.99

(les pourcentages ont été recalculés à partir des coefficients appliqués aux groupes imbriqués)

1.4 Format final du DataFrame

Le DataFrame final comporte désormais :

- 4 colonnes numériques : Tg, Tx, Tl, Dmax
- Environ 20 à 30 colonnes chimiques : Ag, Al, B, Ca, Cu, Fe, Ni, Nb, Zr, etc.
- Toutes les valeurs sont exprimées en unités physiques cohérentes et prêtes pour la normalisation et le clustering.

2) Analyse exploratoire et gestion des outliers

Avant d'appliquer des techniques de classification, il est indispensable d'examiner la qualité et la distribution des données. Cette étape permet de mieux comprendre la structure des variables numériques et d'identifier d'éventuelles anomalies ou irrégularités, notamment les valeurs manquantes et les outliers (valeurs extrêmes).

2.1 Statistiques descriptives

Une première analyse statistique a été réalisée sur les variables quantitatives principales : Tg, Tx, Tl et Dmax. Ces températures représentent respectivement la température de transition vitreuse, la température de cristallisation et la température de fusion des alliages. Dmax correspond à la densité maximale atteinte ou au diamètre max.

Les principales mesures (moyenne, écart-type, min, max, quartiles) ont été calculées à l'aide de la méthode `describe()` de pandas. Cela a permis de détecter des dispersions parfois élevées, en particulier sur la variable Dmax.

2.2 Vérification des valeurs manquantes

Un contrôle global du DataFrame a été effectué pour repérer les cellules contenant des valeurs manquantes (NaN). Grâce à `df.isnull().sum()`, nous avons pu constater que peu de lignes présentaient des valeurs manquantes, et uniquement sur des colonnes chimiques ou thermiques.

Les lignes incomplètes ont été supprimées pour garantir la fiabilité de l'analyse, car leur proportion était inférieure à 5 % du total.

2.3 Détection des valeurs extrêmes (outliers)

La détection des outliers s'est concentrée uniquement sur les quatre variables suivantes : Tg, Tx, Tl et Dmax. En effet, les colonnes représentant les pourcentages chimiques ne sont pas soumises à la même logique statistique : des pourcentages élevés ne constituent pas forcément des anomalies.

Deux méthodes ont été testées :

- Méthode IQR (InterQuartile Range) : on considère comme outliers les valeurs inférieures à $Q1 - 1.5 \times IQR$ ou supérieures à $Q3 + 1.5 \times IQR$.
- Méthode Z-score : on identifie les points dont le score normalisé (écart à la moyenne en unités d'écart-type) dépasse ± 3 .

Résultat : seul le champ Dmax contenait un nombre significatif de points extrêmes — environ 50 lignes sur 600 dépassaient les seuils classiques. En revanche, Tg, Tx et Tl ne présentaient pas d'outliers majeurs.

2.4 Tableau récapitulatif

Variable	Nombre d'outliers détectés (méthode IQR)
TG	0
TX	0
TL	0
Dmax	50

2.5 Analyse et traitement des outliers

Après discussion, il a été décidé de :

- Conserver tous les pourcentages chimiques tels quels, car ils sont physiquement plausibles.
- Deux interprétations sont possibles concernant la variable Dmax : si elle représente la densité maximale, alors les 50 valeurs extrêmes détectées, dont certaines atteignent 72, sont physiquement incohérentes, car la densité maximale connue pour les métaux denses ne dépasse généralement pas 23. Ces valeurs pourraient donc être considérées comme des anomalies, nécessitant soit leur suppression, soit l'abandon de ce jeu de données. En revanche, si Dmax fait référence au diamètre maximal, alors une valeur de 72 reste plausible selon le contexte industriel, et les points détectés comme 'outliers' par les méthodes statistiques ne seraient en réalité pas aberrants.
- Appliquer une normalisation Z-score par la suite afin de limiter leur influence lors du clustering.

3) Normalisation des données

Avant d'appliquer des algorithmes de clustering tels que K-Means ou les méthodes hiérarchiques, il est indispensable de normaliser les données numériques. Cette étape permet d'éviter que certaines variables à grande échelle n'éclipsent d'autres plus petites dans le calcul des distances.

3.1 Méthode utilisée : Z-score standardization

Nous avons choisi la méthode du Z-score, qui transforme chaque valeur x_i selon la formule :

$$Z = \frac{x - \mu}{\sigma}$$

où μ est la moyenne de la variable et σ son écart-type. Cette méthode ramène chaque variable à une distribution centrée en 0 et avec un écart-type de 1, ce qui est parfaitement adapté aux algorithmes de classification non supervisée.

3.2 Variables concernées

La normalisation a été appliquée uniquement sur les variables quantitatives suivantes :

- Tg : Température de transition vitreuse
- Tx : Température de cristallisation
- Tl : Température de fusion
- Dmax : Diamètre maximal

Les pourcentages chimiques n'ont pas été normalisés, car ils sont tous déjà exprimés dans la même unité (%), dans un intervalle connu [0, 100].

4) Clustering avec K-Means

4.1 Objectif

Après la phase de préparation des données, nous avons appliqué la méthode de clustering K-Means afin de regrouper les alliages présentant des similarités thermiques et chimiques. L'objectif est d'identifier des familles homogènes de matériaux sans connaissance préalable des étiquettes de classes.

4.2 Sélection des variables

Le clustering a été effectué sur les variables suivantes :

- Tg, Tx, Tl, Dmax (normalisées)
- Les colonnes de pourcentages chimiques (Ag, Ca, Mg, Cu, etc.)

Ainsi, chaque alliage est représenté comme un vecteur dans un espace multidimensionnel.

4.3 Choix du nombre de clusters (k)

4.3.1 Méthode du coude (Elbow method)

Nous avons calculé l'inertie intra-cluster (somme des distances des points à leur centroïde) pour différentes valeurs de k (nombre de clusters), allant de 2 à 10. L'objectif est d'observer à partir de quel k l'inertie diminue moins significativement.

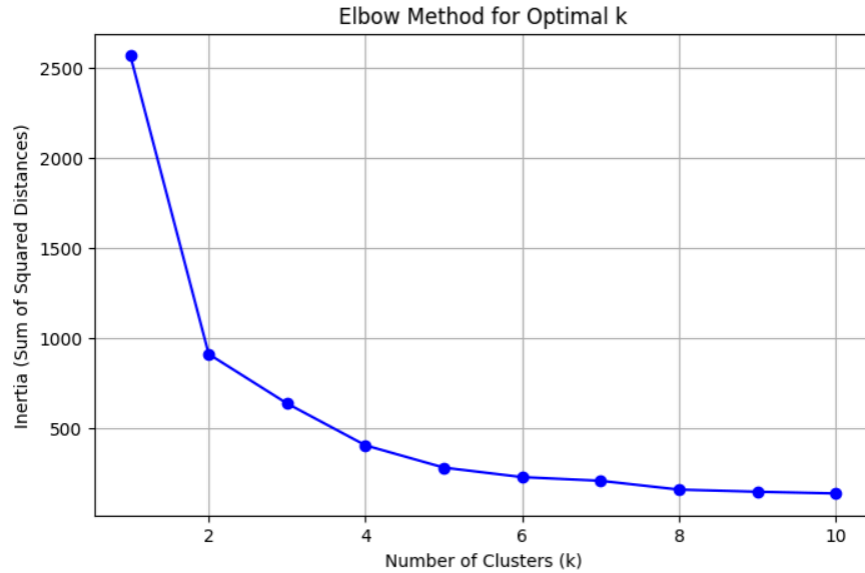


Figure 1: Illustration de la Courbe du coude – Inertie vs. k

Conclusion : la courbe ne montre pas de "coude" clairement défini, ce qui rend le choix de k difficile par simple observation. C'est pourquoi nous avons complété l'analyse avec une seconde méthode.

4.3.2 Méthode du score silhouette

Nous avons ensuite calculé le score silhouette pour chaque valeur de k. Ce score mesure la qualité du regroupement en comparant la cohésion interne d'un cluster à sa séparation avec les autres.

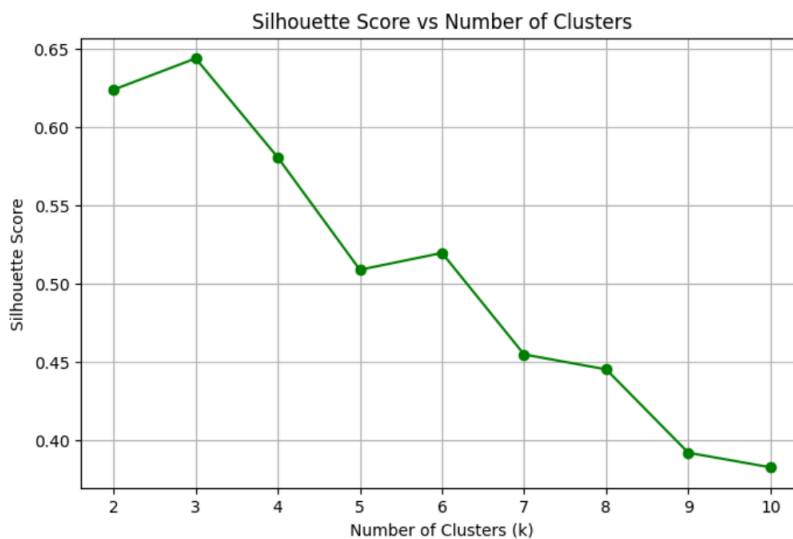


Figure 2 : Illustration de la courbe silhouette score

Conclusion : le score silhouette atteint un maximum à $k = 3$, ce qui suggère que la structure en 3 clusters est la plus cohérente. Ce critère a donc été retenu pour notre clustering final.

4.4 Application de K-Means

L'algorithme KMeans (scikit-learn) a été entraîné sur les données normalisées avec $k = 3$. Chaque observation a été assignée à un cluster.

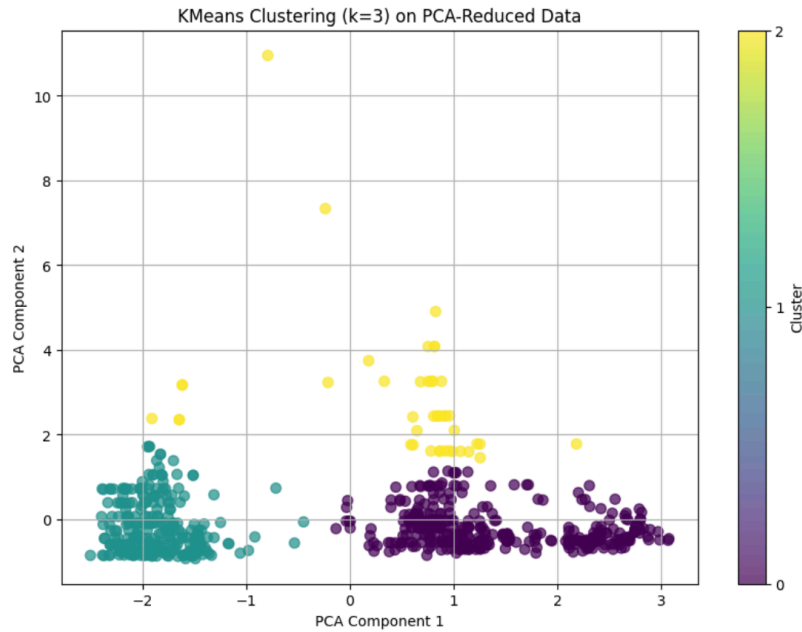


Figure 3 :Représentation des clusters K-Means

4.4.1 Répartition des clusters

Après l'assignation, nous avons analysé la taille de chaque groupe :

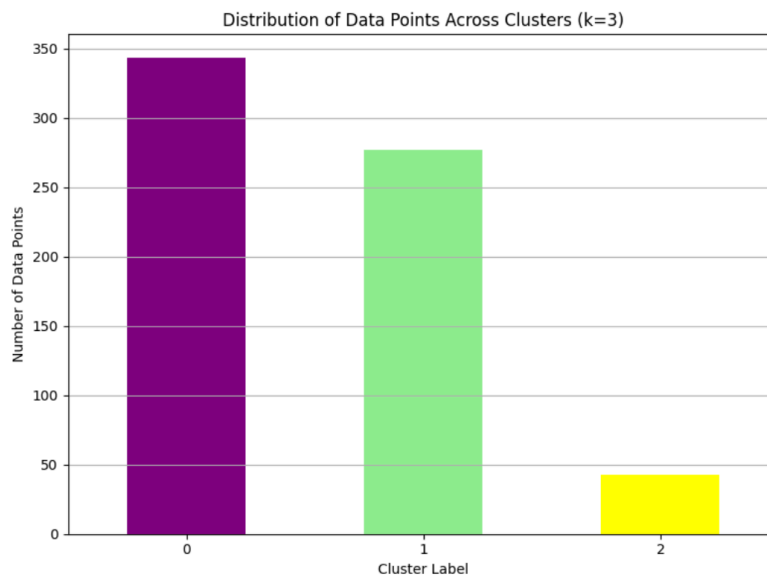


Figure 4: Diagramme à barres de la taille des clusters

Certains clusters sont très denses, d'autres plus dispersés. Cela suggère des familles d'alliages bien définies et d'autres plus variées.

4.4.2 Analyse chimique des clusters

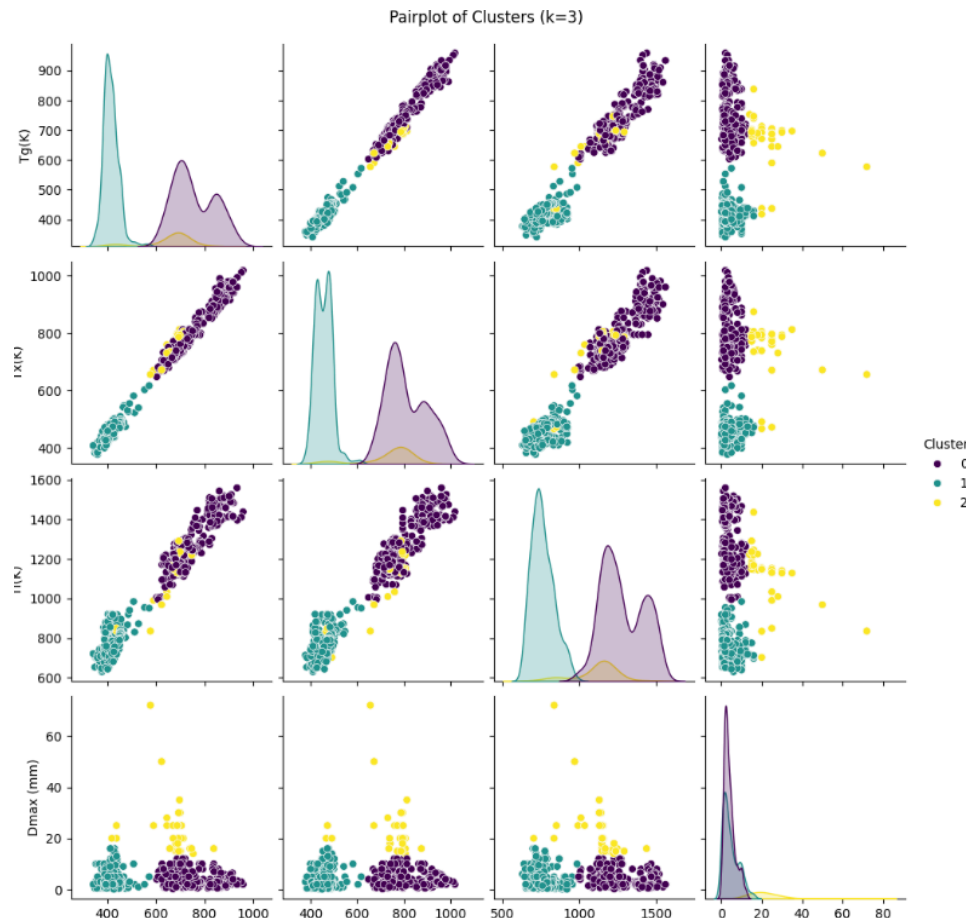


Figure 5 : grille de nuages de points et de distributions

4.4.2.1 T_g vs T_x , T_l

- Il existe une forte corrélation linéaire positive entre T_g , T_x et T_l . Cela suggère que les verres avec une température de transition vitreuse élevée ont aussi des températures de cristallisation et de liquidus plus élevées.
- Cela est attendu car ces températures sont physiquement liées à la structure du verre.

4.4.2.2 Dmax vs les températures

- Dmax n'est pas fortement corrélé avec Tg, Tx ou Tl : on observe une dispersion importante.
- Le cluster jaune (2) est nettement distinct ici : il présente des Dmax beaucoup plus élevés, malgré des températures variées. Cela peut suggérer une composition ou une structure très différente

4.4.2.3 Comportement par cluster

- Cluster 1 (cyan) : regroupe avec Tg, Tx, Tl faibles à modérés. Dmax est généralement faible (<10 mm).
- Cluster 0 (violet) : regroupe avec Tg, Tx, Tl élevés, mais Dmax reste faible. Ce groupe semble bien formé, mais peu épais.

4.4.2.4 Interprétation finale

- La classification en 3 clusters semble pertinente : chaque groupe présente une tendance thermique spécifique.
- Le lien fort entre Tg, Tx et Tl est confirmé.

4.4.3 Bilan

K-Means a permis de structurer les alliages selon leurs propriétés physiques et chimiques, en révélant des tendances de composition associées à des comportements thermiques spécifiques. Ces résultats peuvent servir de base à une classification plus fine ou à la prédiction de propriétés pour de nouveaux alliages.

5) Clustering hiérarchique

5.1 Objectif

L'agglomération hiérarchique (ou classification hiérarchique ascendante) est une méthode de regroupement qui construit une hiérarchie de clusters en partant de chaque point comme un cluster individuel, puis en fusionnant itérativement les clusters les plus proches jusqu'à obtenir un seul cluster englobant toutes les données ou un nombre souhaité de clusters. Cette approche permet de visualiser la

structure des données sous forme d'un dendrogramme, facilitant l'analyse des regroupements à différents niveaux de similarité.

5.2 Application de clustering hiérarchique

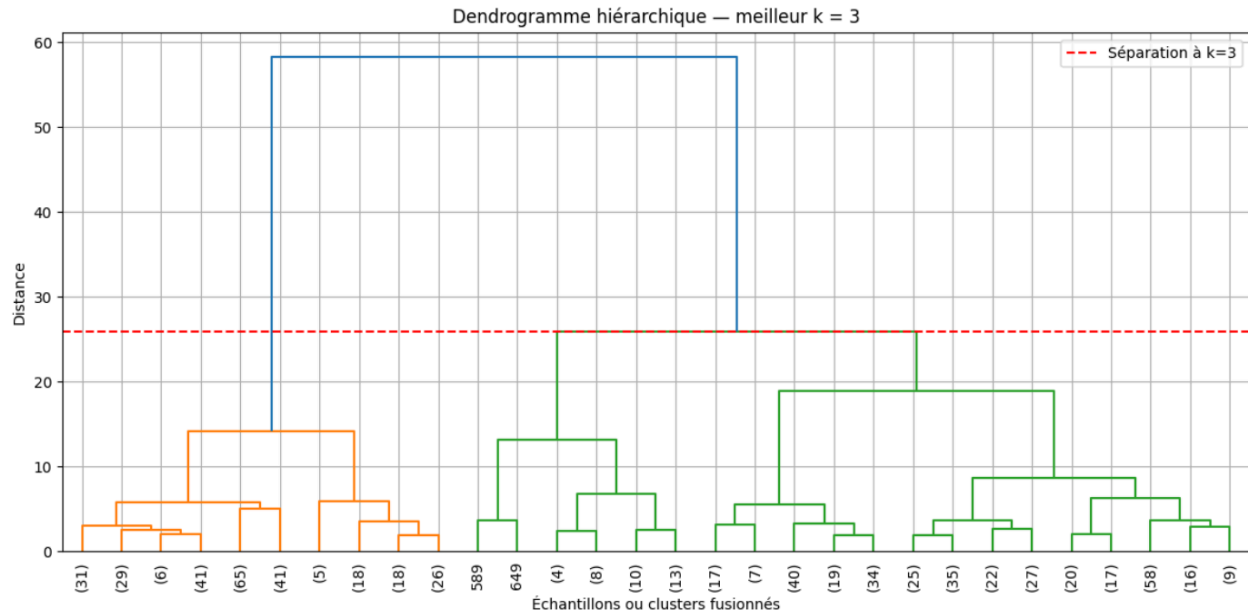


Figure 6 :Dendrogramme du clustering hiérarchique

6) Conclusion général




Dans ce projet, nous avons mené une analyse approfondie d'un ensemble de données expérimentales décrivant des métalliques à travers leurs compositions chimiques et leurs propriétés thermiques (Tg, Tx, Tl) ainsi que leur épaisseur maximale amorphe (Dmax).

Nous avons d'abord procédé à un prétraitement rigoureux des données incluant :

- la conversion des compositions chimiques textuelles en structures exploitables,
- la suppression des éléments chimiques non reconnus,
- l'élimination des lignes incohérentes où la somme des pourcentages d'éléments dépassait 100%,
- ainsi que le retrait des outliers les plus extrêmes sur la variable Dmax.

Ces étapes ont permis d'assurer l'intégrité et la fiabilité du jeu de données.

Ensuite, à l'aide de techniques de clustering (notamment KMeans avec $k=3$), nous avons pu identifier trois familles distinctes de verres métalliques :

-  Un second groupe à températures plus élevées mais avec également de faibles D_{max} .
-  Un premier groupe caractérisé par des températures faibles et une faible D_{max} .
-  Un troisième groupe plus hétérogène, marqué par des D_{max} très élevés, représentant potentiellement des compositions particulièrement stables ou atypiques.