République Algérienne Démocratique et Populaire
Ministère de l'enseignement Supérieur et de la Recherche Scientifique
Université Benyoucef Benkhedda Alger 1

Faculté de Sciences

Département Informatique

Master 1 Ingénierie des systèmes d'information intelligents ISII

Rapport : Prédiction de la Capacité de Formation des Verres Métalliques

Module : Data Mining

Enseignant : M. Boufenar

Réalisé par :

- BELABBAS Rania
- ABAD Lysa Manal
- MOULAY ABDELLAH Asma

Année universitaire : 2024/2025

Table des matières

1 Traitement des Données des Alliages 4
1.1 Introduction 4
1.2 Lecture et Fusion des Tableaux Excel 4
1.3 Nettoyage Initial de la Colonne Alloys 5
1.4 Traitement Avancé de la Colonne Alloys 5
1.4.1 Principe de Calcul 5
1.4.2 Exemple de Traitement 6
1.5 Analyse des Formules d'Alliages 6
1.5.1 Formule 1 : Fe58MoB111:L14514C15B6Cr5Er2 6
1.5.2 Formule 2 : Ti0.450000 Cu0.378000 Zr0.100000 Ni0.072000 Sn . . 7
1.6 Organisation des Données des Alliages 7
1.7 Validation des Éléments Chimiques Extraits 8
1.8 Vérification des Pourcentages des Alliages 8

2 Clustering des Alliages 9
2.1 Introduction 9
2.2 Clustering avec la Méthode du Coude et KMeans 9
2.2.1 Étapes du Processus 9
2.2.2 Résultats et Visualisation 10
2.3 Analyse Post-Clustering 10
2.3.1 Méthodologie 10
2.3.2 Résultats 11
2.3.3 Interprétation 11
2.4 Évaluation du Clustering 11
2.5 Analyse des Propriétés Physiques par Cluster 12
2.6 Conclusion 12

Table des figures

1.1 Fichier avant fusion des tableaux. 4
1.2 Fichier aprés fusion des tableaux. 5
1.3 Fichier avec espaces. 5
1.4 Fichier sans espaces. 5
1.5 Fichier avant calculs. 6
1.6 Fichier aprés calculs. 6
1.7 Erreur dans les compositions. 6
1.8 Erreur dans les compositions. 7

2.1 Méthode du coude 9
2.2 Visualisation des clusters en 2D après réduction avec PCA 10
2.3 Pourcentages moyens des éléments chimiques par cluster 11

Chapitre 1

Traitement des Données des Alliages

1.1 Introduction

Ce chapitre présente les différentes étapes de traitement des données d'alliages extraites à partir du fichier `BMGs-2024.odt` initialement sous format Open Document. Cependant, pour faciliter la manipulation des données, ce fichier a été téléchargé et converti en format Excel. L'objectif est de nettoyer, fusionner, et effectuer des calculs de pourcentages sur les composants des alliages en tenant compte de diverses priorités de traitement des parenthèses, crochets et accolades.

1.2 Lecture et Fusion des Tableaux Excel

Le fichier `BMGs-2024.xlsx` contient deux tableaux côte à côte, dont les colonnes d'intérêt sont `Unnamed: 1` à `Unnamed: 5` et `Unnamed: 7` à `Unnamed: 11`. Les étapes suivantes ont été réalisées pour préparer les données :

- Suppression des lignes vides.
- Extraction des deux tableaux (df1 et df2).
- Renommage des colonnes extraites en `['Alloys', 'Tg', 'Tx', 'Tl', 'Dmax']`.
- Filtrage des lignes valides, en éliminant les titres et les valeurs `NaN` dans la colonne `Alloys`.
- Fusion des deux tableaux en un seul (df_combined).

Alloys (compostion)	Tg(K)	Tx(K)	Tl(K)	Dmax (mm)	Alloys	Tg(K)	Tx(K)	Tl(K)	Dmax (mm)
Ag30.8 Ca30.8 Mg23.1 Cu15.4	413	432	803	2,5	Ag30.8 Mg30.8 Ca30.8 Cu7.7	407	427	809	2
Ag38.4 Mg30.8 Ca30.8	394	426	805	0,5	Ag38.4 Mg38.4 Ca23.2	391	425	796	1,1
Ag38.5 Ca30.8 Mg23 Cu7.7	384	416	854	2	Ag38.5 Mg30.8 Ca23.1 Cu7.7	387	420	833	3
Ag38.5 Mg38.5 Ca15.4 Cu7.7	405	436	842	0,5	Ag46.2 Ca30.5 Mg15.4 Cu7.7	414	445	805	0,8
Ag46.2 Ca30.7 Mg23.1	399	426	765	0,7	Ag46.2 Ca38.4 Mg15.3	407	439	809	0,3
Ag46.2 Mg23.2 Ca23 Cu7.7	398	430	825	2	Ag46.2 Mg30.7 Ca23.1	393	427	880	0,5
Ag50 Ca30.8 Mg11.5 Cu7.7	452	487	809	1	Ag50 Ca30.8 Mg19.2 Cu7.7	426	466	797	1,2
Ag53.8 Ca30.5 Mg7.7 Cu7.7	428	488	843	0,3	Ag53.8 Mg15.4 Ca23.1 Cu7.7	433	463	831	0,5
Ag53.8 Mg15.4 Ca30.8	444	498	812	0,8	Ag53.8 Mg23.1 Ca15.4 Cu7.7	407	463	877	0,5
Ag53.8 Mg23.1 Ca23.1	451	488	887	0,7	Ag 61.5 Ca23.1 Mg15.4	486	526	920	0,7
Ag61.5 Mg23.1 Ca15.4	440	485	919	0,5	Au46 Cu29 Si20 Ag5	395	420	664	1

FIG. 1.1 : Fichier avant fusion des tableaux.

Alloys	Tg	Tx	Tl	Dmax
Ag30.8 Ca	413	432	803	2,5
Ag38.4 Mg	394	426	805	0,5
Ag38.5 Ca	384	416	854	2
Ag38.5 Mg	405	436	842	0,5
Ag46.2 Ca	399	426	765	0,7
Ag46.2 Mg	398	420	825	2

FIG. 1.2 : Fichier aprés fusion des tableaux.

1.3 Nettoyage Initial de la Colonne Alloys

Le nettoyage initial a consisté à supprimer les espaces inutiles dans la colonne `Alloys`, tout en conservant les parenthèses, crochets et accolades qui sont nécessaires pour les calculs ultérieurs. Cette étape est cruciale afin de préparer les données pour les traitements de pourcentage qui suivent.

Alloys
Ag30.8 Ca30.8 Mg23.1 Cu15.4
Ag38.4 Mg30.8 Ca30.8
Ag38.5 Ca30.8 Mg23 Cu7.7

FIG. 1.3 : Fichier avec espaces.

Alloys
Ag30.8Ca30.8Mg23.1Cu15.4
Ag38.4Mg30.8Ca30.8
Ag38.5Ca30.8Mg23Cu7.7

FIG. 1.4 : Fichier sans espaces.

1.4 Traitement Avancé de la Colonne Alloys

L'objectif de cette étape était de réaliser des calculs de pourcentages pour chaque composant de l'alliage en appliquant une priorité de traitement aux groupes de pourcentages dans l'ordre suivant : accolades { }, crochets [ ], puis parenthèses ( ).

1.4.1 Principe de Calcul

Lorsqu'un groupe a un coefficient (par exemple $(Fe60Co40)75$), le calcul consiste à :

- Appliquer 60% de 75, soit 45, et 40% de 75, soit 30.
- Ensuite, appliquer les coefficients extérieurs successivement. Par exemple, un groupe [...]96 multiplie tout à l'intérieur par 96%, et un groupe {...}97 multiplie tout par 97%.

1.4.2 Exemple de Traitement

Prenons l'exemple suivant :

$$\{[(Fe60Co40)75B20Si5]96Nb4\}97Cr3$$

Le calcul des pourcentages donne les résultats suivants :

$$\text{Fe} = 41.904, \quad \text{Co} = 27.936, \quad \text{B} = 18.624, \quad \text{Si} = 4.656, \quad \text{Nb} = 3.88, \quad \text{Cr} = 3$$

	[(Fe60Co40)72B24Mo4]94Dy6
2	[(Fe60Co40)75B20Si5]96Nb4
3	{[(Fe60Co40)75B20Si5]96Nb4}97Cr3
4	{[(Fe60Co40)75B20Si5]96Nb4}99Cr1
5	[(Fe60Co40)75B20Si5]95Nb4Mo1
5	[(Fe60Co40)75B20Si5]95Nb4Zr1

FIG. 1.5 : Fichier avant calculs.

31	Fe40.608000 Co27.072000 B22.560000 Mo3.760000Dy6
32	Fe43.200000 Co28.800000 B19.200000 Si4.800000Nb4
33	Fe41.904000 Co27.936000 B18.624000 Si4.656000 Nb3.880000Cr3
34	Fe42.768000 Co28.512000 B19.008000 Si4.752000 Nb3.960000Cr1
35	Fe42.750000 Co28.500000 B19.000000 Si4.750000Nb4Mo1
36	Fe42.750000 Co28.500000 B19.000000 Si4.750000Nb4Zr1

FIG. 1.6 : Fichier aprés calculs.

1.5 Analyse des Formules d'Alliages

Après avoir effectué les calculs nécessaires, des erreurs ont été rencontrées dans certaines des formules d'alliages. Ces erreurs sont dues à des valeurs qui ne respectent pas le format attendu, à savoir une séquence d'éléments chimiques suivis de quantités numériques, pouvant inclure des nombres à virgule flottante. Une validation des formats a été réalisée à l'aide d'expressions régulières, et les lignes contenant des formules incorrectes ont été extraites pour correction.

Dans cette section, nous analysons deux formules d'alliages spécifiques présentes dans les données :

	Alloys
128	Fe58MoB111:L14514C15B6Cr5Er2
346	Ti0.450000 Cu0.378000 Zr0.100000 Ni0.072000Sn

FIG. 1.7 : Erreur dans les compositions.

1.5.1 Formule 1 : Fe58MoB111:L14514C15B6Cr5Er2

Cette formule semble contenir plusieurs erreurs de format. En effet, bien qu'elle respecte partiellement le format d'éléments chimiques suivis de quantités, elle présente les problèmes suivants :

- La présence du caractère ":" entre les groupes `Fe58MoB111` et `L14514C15B6Cr5Er2`, ce qui n'est pas conforme à la structure attendue.
- Les quantités des éléments sont parfois exprimées avec un grand nombre de chiffres (par exemple, `14514`), ce qui pourrait ne pas être souhaité ou incorrect en fonction des exigences du format.

1.5.2 Formule 2 : Ti0.450000 Cu0.378000 Zr0.100000 Ni0.072000 Sn

Cette formule semble mieux respecter le format attendu, mais présente un autre probléme :

- L'élément `Sn` est présent sans quantité associée. Cela peut poser problème si la quantité doit toujours être spécifiée, même si elle est égale à 1 ou si une valeur par défaut est requise.

Ces deux formules d'alliages nécessitent des révisions pour se conformer aux standards de format et aux exigences de précision de notre analyse.

1.6 Organisation des Données des Alliages

Dans cette étape du traitement des données, un script Python a été utilisé pour extraire les éléments chimiques et leurs quantités à partir de la colonne des alliages présente dans le fichier Excel original. L'objectif était de transformer la représentation des alliages en un format structuré où chaque élément chimique est représenté par une colonne distincte, avec les quantités associées.

Le script fonctionne de la manière suivante :

1. **Chargement des données :** Le fichier Excel contenant les données des alliages est chargé dans un DataFrame à l'aide de la bibliothèque `pandas`.

2. **Extraction des éléments chimiques :** Une fonction a été définie pour analyser chaque formule d'alliage dans la colonne `Alloys`. Cette fonction utilise une expression régulière pour extraire les symboles des éléments chimiques et leurs quantités associées, qui peuvent être des nombres entiers ou décimaux.

3. **Création de colonnes par élément :** Après l'extraction, un dictionnaire contenant les éléments chimiques et leurs pourcentages a été créé pour chaque alliage. Ensuite, une colonne a été ajoutée pour chaque élément chimique unique, avec la quantité correspondante pour chaque ligne. Si un élément n'est pas présent dans une ligne donnée, sa quantité est définie à zéro.

4. **Sauvegarde des résultats :** Une fois la transformation terminée, le DataFrame modifié a été sauvegardé sous un nouveau fichier Excel, dans lequel chaque élément chimique possède sa propre colonne, facilitant ainsi l'analyse et la manipulation des données.

Alloys	Tg	Tx	Tl	Dmax	Mo	Cu	Au	Ti	Hf	Sm	B	P	V	Mg	L	Nd	Mm	Ni
Ag30.8Ca3	413	432	803	2,5	0	15,4	0	0	0	0	0	0	0	23,1	0	0	0	0
Ag38.4Mg	394	426	805	0,5	0	0	0	0	0	0	0	0	0	30,8	0	0	0	0

FIG. 1.8 : Erreur dans les compositions.

Le processus a permis de structurer les données des alliages de manière cohérente, facilitant l'analyse statistique et la comparaison entre les différents éléments présents dans les alliages étudiés.

1.7 Validation des Éléments Chimiques Extraits

Dans cette section, un processus de validation a été effectué pour vérifier que les éléments chimiques extraits d'un fichier de données étaient bien conformes à ceux présents dans la table périodique officielle. Cette validation permet de s'assurer que seules les entités chimiques reconnues par la communauté scientifique ont été extraites, et d'identifier toute donnée erronée ou incohérente.

Le processus a été réalisé de la manière suivante :

1. **Extraction des éléments :** Les éléments chimiques ont été extraits à partir du fichier de données, contenant des alliages et des formules chimiques.

2. **Comparaison :** Une comparaison a été effectuée entre les éléments extraits et ceux présents dans la table périodique officielle. Les éléments extraits qui ne figurent pas dans la table périodique ont été identifiés comme étant non valides.

Le processus a permis de valider les éléments extraits et de détecter les éléments non conformes, comme suit :

- **Éléments non valides détectés :** Les éléments suivants n'étaient pas présents dans la table périodique officielle : `L`, `Mm`.

Ces éléments doivent être retirés afin de poursuivre l'analyse des données.

Note : L'élément `L` pourrait être une erreur typographique pour un élément valide comme `La` (Lanthane), et `Mm` pourrait résulter d'une mauvaise interprétation ou d'un format incorrect.

1.8 Vérification des Pourcentages des Alliages

Dans cette étape du traitement des données, un script Python a été utilisé pour vérifier que la somme des pourcentages des éléments chimiques dans chaque alliage respectait une plage acceptable. En effet, pour chaque alliage, la somme des pourcentages des éléments doit idéalement être proche de 100%, dans la plage de [99.5%, 100.5%], en tenant compte des éventuelles imprécisions dans les mesures.

Résultat : Après l'exécution du script, il a été constaté qu'aucune ligne n'affichait de pourcentage anormal, ce qui garantit la cohérence des données pour cette étape de validation. Si des anomalies avaient été détectées, elles auraient été listées pour un examen plus approfondi.

Chapitre 2

Clustering des Alliages

2.1 Introduction

Une fois les données nettoyées et traitées, nous avons appliqué des méthodes de clustering afin de segmenter les alliages en fonction de leurs caractéristiques. Le but était de découvrir des groupes d'alliages présentant des similitudes dans leurs compositions et propriétés.

2.2 Clustering avec la Méthode du Coude et KMeans

Le clustering k-means est un processus itératif visant à minimiser la somme des distances entre les points de données et le centroïde des clusters.

Dans cette section, nous appliquons une technique de clustering pour regrouper les alliages en fonction de leurs compositions chimiques. L'objectif est de déterminer le nombre optimal de clusters en utilisant la méthode du coude, et de visualiser les résultats à l'aide de la réduction de dimensionnalité par l'analyse en composantes principales (PCA).

2.2.1 Étapes du Processus

1. **Détermination du nombre optimal de clusters :** La méthode du coude a été utilisée pour identifier le nombre optimal de clusters. Cette méthode consiste à rechercher la valeur de k où l'inertie cesse de diminuer de manière significative. Le point où cette diminution ralentit est considéré comme le "coude" et correspond au nombre optimal de clusters.

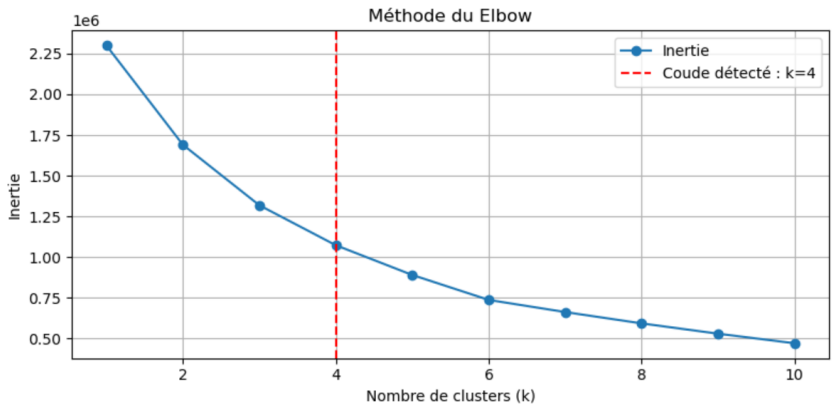

FIG. 2.1 : Méthode du coude

2. **Clustering final :** Avec le nombre optimal de clusters déterminé, nous avons exécuté l'algorithme `KMeans` sur l'ensemble des données pour assigner chaque alliage

à un cluster. Une nouvelle colonne, `Cluster`, a été ajoutée au jeu de données pour indiquer l'appartenance de chaque alliage à un cluster particulier.

3. **Réduction de dimensionnalité (PCA) pour la visualisation :** L'analyse en composantes principales (PCA) a été appliquée pour réduire les données de plusieurs dimensions à deux dimensions, facilitant ainsi la visualisation des clusters. Les résultats ont été visualisés dans un graphique 2D, où chaque point représente un alliage et est coloré en fonction de son cluster.

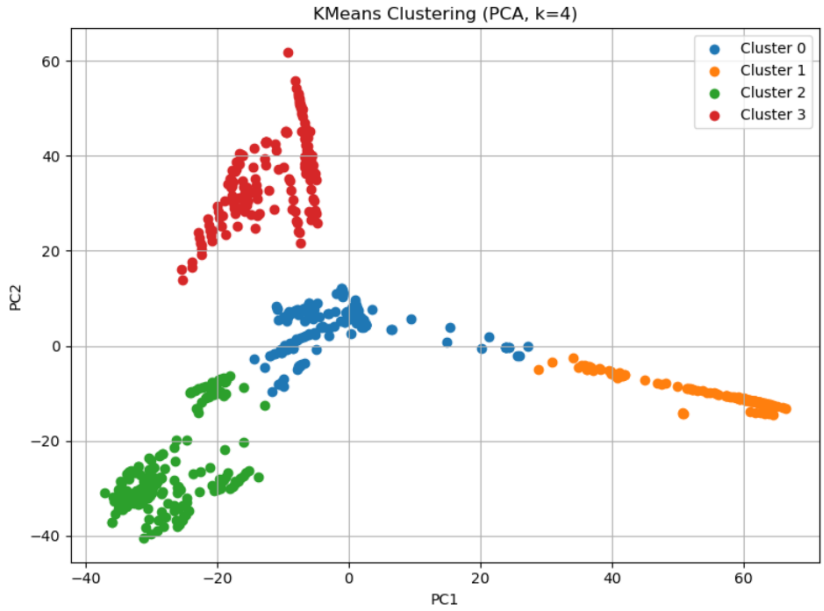

FIG. 2.2 : Visualisation des clusters en 2D après réduction avec PCA

2.2.2 Résultats et Visualisation

Le meilleur nombre de clusters, déterminé par la méthode du coude, est $k = 4$, comme indiqué par le "coude" dans le graphique de l'inertie.

Le graphique suivant montre les résultats du clustering, où chaque couleur représente un cluster distinct dans l'espace réduit par PCA.

Note : Ce processus permet de mieux comprendre la structure des données des alliages et peut être utile pour identifier des groupes d'alliages ayant des propriétés similaires.

2.3 Analyse Post-Clustering

Après avoir appliqué l'algorithme de clustering KMeans, une analyse des compositions chimiques moyennes de chaque cluster a été réalisée afin d'identifier les éléments dominants dans chacun des groupes.

2.3.1 Méthodologie

Les moyennes des pourcentages de chaque élément chimique ont été calculées pour chaque cluster à l'aide de la fonction `groupby()` sur la colonne `Cluster`. Ces moyennes ont ensuite été visualisées sous forme de diagramme en barres, où chaque barre représente la moyenne d'un élément donné pour un cluster spécifique.

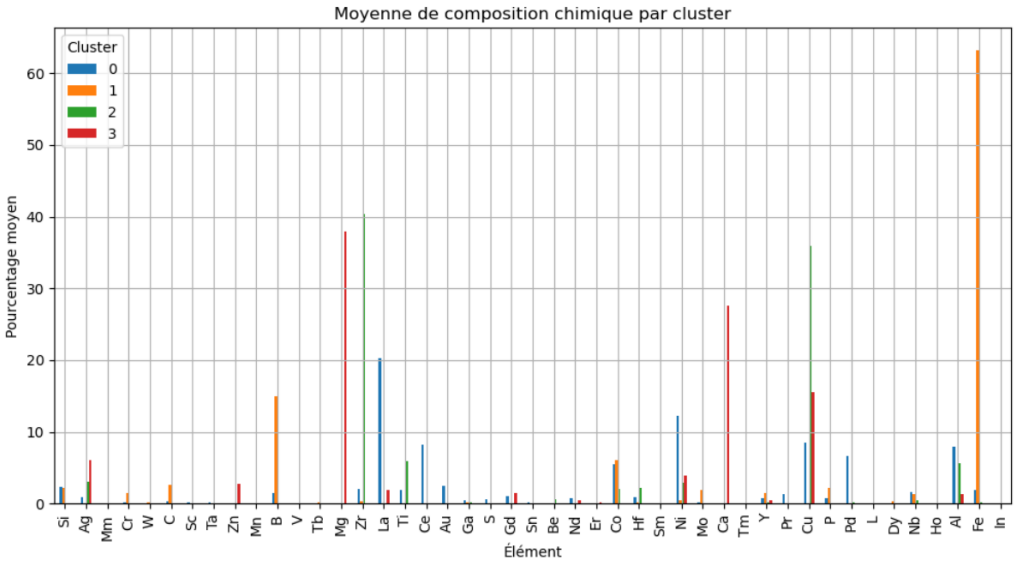

FIG. 2.3 : Pourcentages moyens des éléments chimiques par cluster

2.3.2 Résultats

Le graphique (Figure 2.3) met en évidence les tendances suivantes :

- Le **Fer (Fe)** est l'élément le plus dominant dans le **cluster 2**, avec une moyenne nettement supérieure à celle observée dans les autres clusters.
- Certains éléments comme le **Zirconium (Zr)** ou le **Cuivre (Cu)** peuvent également présenter des pics localisés dans des clusters spécifiques, suggérant des sous-groupes d'alliages caractérisés par une composition chimique particulière.

2.3.3 Interprétation

Cette analyse permet d'établir des profils chimiques typiques pour chaque cluster. Ces profils peuvent être utiles pour relier les compositions aux propriétés physiques des alliages, ou pour guider la conception de nouveaux alliages à partir des tendances identifiées dans les données.

2.4 Évaluation du Clustering

Pour évaluer la qualité du regroupement obtenu, le **Silhouette Score** a été calculé. Cet indice permet de mesurer la cohésion et la séparation des clusters. Il varie entre -1 et 1 :

- Un score proche de **1** indique que les échantillons sont bien regroupés au sein de leurs clusters respectifs.
- Un score proche de **0** indique un chevauchement important entre les clusters.
- Un score négatif suggère une mauvaise affectation des points.

Dans notre cas, le score obtenu est de :

$$\text{Silhouette Score} = 0.411$$

(cette valeur a été calculée automatiquement avec la fonction `silhouette_score` de `scikit-learn`).

2.5 Analyse des Propriétés Physiques par Cluster

Une moyenne des principales propriétés physiques (`Tg`, `Tx`, `Tl`, `Dmax`) a été calculée pour chaque cluster. Ces propriétés sont essentielles pour l'évaluation de la stabilité thermique et de la capacité de formage des alliages métalliques amorphes.

Cluster	**Tg**	**Tx**	**Tl**	**Dmax**
0	543.43	592.52	951.46	6.07
1	829.47	881.53	1409.44	3.01
2	685.21	748.54	1166.93	7.55
3	415.36	453.03	761.94	3.55

TAB. 2.1 : Valeurs moyennes des propriétés physiques par cluster

Cette classification permet d'identifier des tendances spécifiques par groupe, telles que des plages de températures de transition vitreuse ou de cristallisation plus élevées, potentiellement corrélées à certaines compositions chimiques.

2.6 Conclusion

Le traitement des données d'alliages et l'application du clustering ont permis de mieux comprendre les propriétés des alliages et de les segmenter en groupes significatifs. Ces résultats ouvrent la voie à des analyses plus poussées pour optimiser la sélection des matériaux dans des processus industriels.