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

Table des figures

Chapitre 1

Traitement des Données des Alliages

1.1 Introduction

Avant d'appliquer des techniques de clustering, il est essentiel de préparer les données afin d'assurer la qualité et la fiabilité des résultats. Le prétraitement des données comprend plusieurs étapes, telles que le chargement des fichiers, la normalisation des formats, la gestion des valeurs manquantes et la transformation des chaînes de composition chimique en données numériques exploitables.

Dans notre cas, les données proviennent de plusieurs feuilles d'un fichier tableur contenant des informations sur différents alliages métalliques. Chaque entrée inclut la composition chimique sous forme textuelle ainsi que certaines propriétés physiques comme la température de transition vitreuse (Tg), la température de cristallisation (Tx), la température de liquidus (Tl) et le diamètre maximal ($Dmax$). Afin de rendre ces informations utilisables pour les algorithmes d'apprentissage automatique, un nettoyage et une structuration minutieuse ont été effectués.

1.2 Importation des bibliothèques

Tout d'abord, nous importons les bibliothèques nécessaires pour effectuer l'analyse :

- `pandas` pour la manipulation des données.
- `numpy` pour les opérations numériques.
- `matplotlib` pour tracer les résultats.
- `sklearn` pour le clustering et la réduction de dimensionnalité.
- `re` pour les expressions régulières afin d'extraire les composants des alliages.

1.3 Données initiales

Les données sont chargées à partir d'un fichier Open Document Spreadsheet (.ods) contenant les compositions et propriétés des alliages contenant deux tableaux distincts. Chaque tableau présente des informations sur des alliages métalliques, incluant leur composition chimique ainsi que certaines propriétés physiques. Ces tableaux ont été fusionnés en un seul `DataFrame` pour permettre une analyse cohérente et complète dans les étapes suivantes.

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

FIG. 1.1 : Fichier avant fusion des tableaux.

Alloy	Tg(K)	Tx(K)	Tl(K)	Dmax(mm)
Ag30.8 Ca30.8 Mg23.1 Cu15.4	413	432	803	2,5
Ag38.4 Mg30.8 Ca30.8	394	426	805	0,5
Ag38.5 Ca30.8 Mg23 Cu7.7	384	416	854	2
Ag38.5 Mg38.5 Ca15.4 Cu7.7	405	436	842	0,5
Ag46.2 Ca30.7 Mg23.1	399	426	765	0,7
Ag46.2 Mg23.2 Ca23 Cu7.7	398	430	825	2
Ag50 Ca30.8 Mg11.5 Cu7.7	452	487	809	1
Ag53.8 Ca30.5 Mg7.7 Cu7.7	428	488	843	0,3
Ag53.8 Mg15.4 Ca30.8	444	498	812	0,8
Ag53.8 Mg23.1 Ca23.1	451	488	887	0,7
Ag61.5 Mg23.1 Ca15.4	440	485	919	0,5
Au49 Cu26.9 Si16.3 Ag5.5 Pd2.3	401	459	644	5

FIG. 1.2 : Fichier aprés fusion des tableaux.

1.4 Nettoyage Initial

Les données sont nettoyées en supprimant les doublons et en gérant les valeurs manquantes.

1.5 Exploration initiale des données

Avant d'appliquer les techniques de clustering, une exploration préliminaire des données nettoyées a été réalisée afin de mieux comprendre leur structure et leur contenu. Les étapes suivantes ont été effectuées :

- Affichage des informations générales sur le `DataFrame`, y compris le nombre d'entrées non nulles, les types de données et la mémoire utilisée.
- Comptage du nombre total de lignes et de colonnes du jeu de données nettoyé.
- Extraction de la liste complète des noms de colonnes.
- Calcul de statistiques descriptives pour les colonnes numériques telles que la moyenne, l'écart-type, les valeurs minimales et maximales, ainsi que les quartiles.

Ces informations permettent de valider l'intégrité des données, de repérer d'éventuelles valeurs aberrantes et de s'assurer que les colonnes nécessaires à l'analyse sont bien présentes et correctement formatées.

1.6 Extraction des composants des alliages

Nous définissons une fonction `extract_alloy_components` qui extrait les métaux individuels et leurs pourcentages respectifs à partir de la chaîne d'alloi. La fonction utilise des expressions régulières pour gérer différents formats tels que les accolades {}, les crochets [], et les parenthèses (). Elle renvoie ensuite un dictionnaire des métaux et de leurs pourcentages correspondants, en nettoyant les entrées mal formatées.

1.6.1 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 $[\ldots]96$ multiplie tout à l'intérieur par 96%, et un groupe $\{\ldots\}97$ multiplie tout par 97%.

1.6.2 Exemple de Traitement

Prenons l'exemple suivant :

$$\{[(Fe60Co40)75B20Si5]96Nb4\}98Cr2$$

Le calcul des pourcentages donne les résultats suivants :

$$\text{Fe} = 42.768, \quad \text{Co} = 28.512, \quad \text{B} = 19.2, \quad \text{Si} = 4.8, \quad \text{Nb} = 3.92, \quad \text{Cr} = 2$$

131	[(Fe60 Co40)72 B24 Mo4]94 Dy6	847	927	1366	2
132	[(Fe60 Co40)75 B20 Si5]96 Nb4	825	875	1407	4
133	{[(Fe60 Co40)75 B20 Si5]96 Nb4}97 Cr3	831	874	1474	3,5
134	{[(Fe60 Co40)75 B20 Si5]96 Nb4}99 Cr1	827	871	1462	4
135	[(Fe60 Co40)75 B20 Si5]95 Nb4 Mo1	818	859	1414	2
136	[(Fe60 Co40)75 B20 Si5]95 Nb4 Zr1	825	866	1396	2

FIG. 1.3 : Fichier avant calculs.

131	{'Fe': 40.608, 'Co': 27.072, 'B': 22.56, 'Mo': 3.76, 'Dy': 6.0}	847	927	1366	2
132	{'Fe': 43.2, 'Co': 28.8, 'B': 19.2, 'Si': 4.8, 'Nb': 4.0}	825	875	1407	4
133	{'Fe': 41.904, 'Co': 27.936, 'B': 18.624, 'Si': 4.656, 'Nb': 3.88, 'Cr': 3.0}	831	874	1474	3,5
134	{'Fe': 42.768, 'Co': 28.512, 'B': 19.008, 'Si': 4.752, 'Nb': 3.96, 'Cr': 1.0}	827	871	1462	4
135	{'Fe': 42.75, 'Co': 28.5, 'B': 19.0, 'Si': 4.75, 'Nb': 4.0, 'Mo': 1.0}	818	859	1414	2
136	{'Fe': 42.75, 'Co': 28.5, 'B': 19.0, 'Si': 4.75, 'Nb': 4.0, 'Zr': 1.0}	825	866	1396	2

FIG. 1.4 : Fichier aprés calculs.

1.7 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.

Ag	Ca	Mg	Cu	Au	Si	Pd	Al	Ni
30,769	30,769	23,077	15,385	0	0	0	0	0
38,4	30,8	30,8	0	0	0	0	0	0
38,5	30,8	23	7,7	0	0	0	0	0
38,462	15,385	38,462	7,692	0	0	0	0	0
46,2	30,7	23,1	0	0	0	0	0	0
46,154	22,977	23,177	7,692	0	0	0	0	0

FIG. 1.5 : Extraction des compositions d'alliages.

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.

1.8 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.

	AC	AD	AE	AF
1	**W**	**Er**	**Mn**	**L**
2	0	0	0	0
3	0	0	0	0
4	0	0	0	0
5	0	0	0	0
6	0	0	0	0
7	0	0	0	0
8	0	0	0	0
9	0	0	0	0
10	0	0	0	0

FIG. 1.6 : Data set avec l'élément non valide L

	AC	AD	AE	AF
1	**W**	**Er**	**Mn**	**Dy**
2	0	0	0	0
3	0	0	0	0
4	0	0	0	0
5	0	0	0	0
6	0	0	0	0
7	0	0	0	0
8	0	0	0	0
9	0	0	0	0
10	0	0	0	0

FIG. 1.7 : Data set sans l'élément non valide L

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.

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.

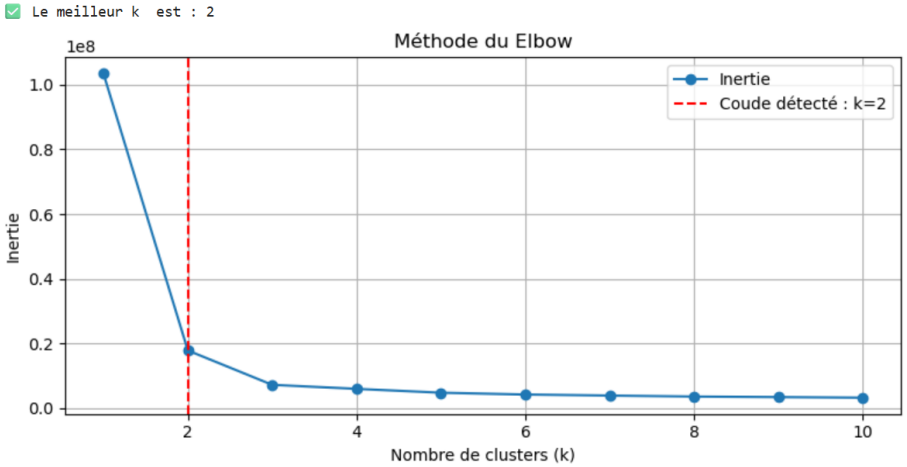

Fig. 2.1 : Méthode du coude

2. **Clustering final :** Avec le nombre optimal de clusters déterminé (k = 2), 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.

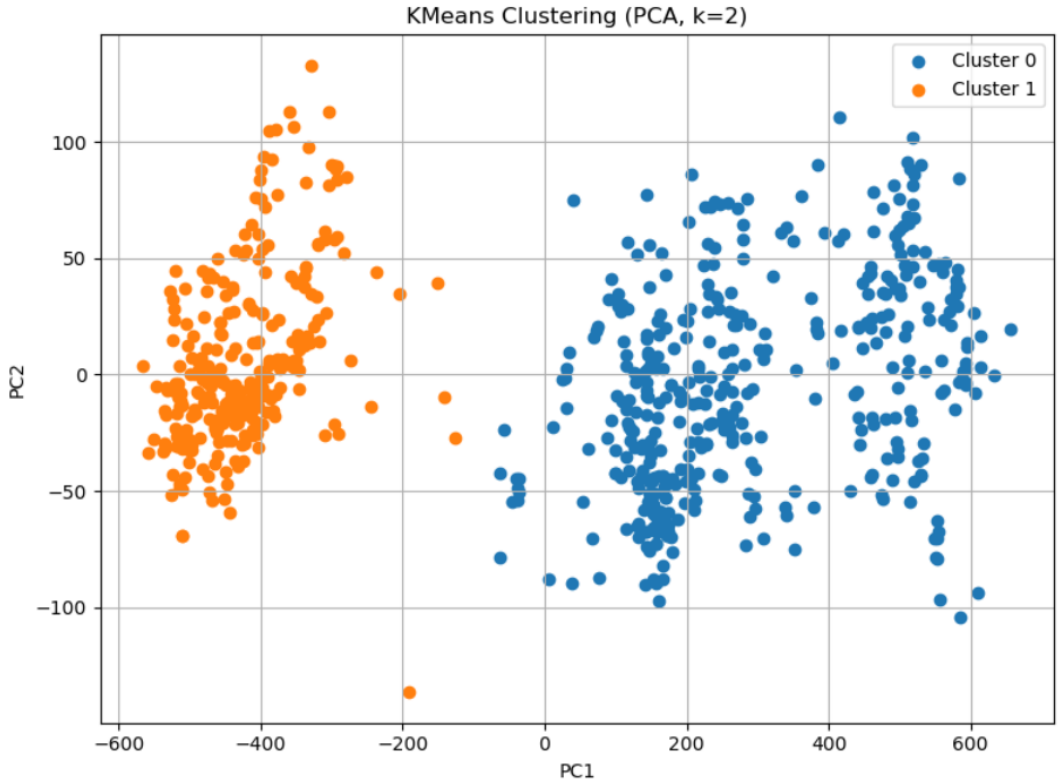

FIG. 2.2 : Visualisation des clusters en 2D après réduction avec PCA

Tl(K)	Dmax(mm)	Cluster
803	2,5	1
805	0,5	1
854	2	1

FIG. 2.3 : Nouvelle colonne "Cluster"

2.3 Évaluation du Clustering

Le score de silhouette est utilisé pour évaluer la qualité du clustering. Il mesure la similarité de chaque point avec son propre cluster par rapport aux autres clusters.

Un score proche de 1 indique des clusters bien séparés, tandis qu'un score proche de -1 suggère un mauvais clustering.

- 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 :

$$\text{Silhouette Score} = 0.723$$

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

2.4 Comparaison des Méthodes de Clustering sur Différents Sous-Ensembles de Données

Afin d'évaluer la qualité du clustering selon les caractéristiques utilisées, nous avons défini plusieurs fonctions d'analyse permettant de comparer les performances de l'algorithme `KMeans` appliqué à divers sous-ensembles des données.

Les données ont été divisées selon quatre configurations :

1. **Composition des alliages uniquement**
2. **Composition des alliages + température de transition vitreuse (Tg)**
3. **Composition des alliages + Tg + température de cristallisation (Tx)**
4. **Jeu de données complet (y compris *Tg*, *Tx*, *Tl*, *Dmax*)**

2.4.1 Comparaison des Scores de Silhouette

Pour évaluer la qualité des regroupements effectués par l'algorithme `KMeans`, nous avons utilisé le score de silhouette, qui mesure la cohésion et la séparation des clusters. Plus ce score est proche de 1, meilleure est la qualité du clustering.

Chaque sous-ensemble de données a été évalué en utilisant le nombre optimal de clusters déterminé par la méthode du coude (elbow) et la maximisation du score de silhouette. Les résultats sont présentés ci-dessous :

Sous-ensemble	**Nombre optimal de clusters (k)**	**Score de silhouette**
Composition uniquement	4	0.389
Composition + Tg	2	0.679
Composition + Tg + Tx	2	0.721
Ensemble complet	2	0.723

TAB. 2.1 : Scores de silhouette chaque sous-ensemble

Ces résultats montrent que :

- Le clustering basé uniquement sur la composition des alliages génère des clusters peu distincts (score de 0.389), même avec 4 groupes.

- L'ajout de la température de transition vitreuse (Tg) améliore significativement la qualité du clustering.
- L'ajout de la température de cristallisation (Tx) renforce encore cette amélioration.
- Le modèle le plus performant est obtenu avec l'ensemble des données, indiquant que les propriétés physiques comme Tg, Tx, Tl, et $Dmax$ contribuent à une segmentation plus cohérente des alliages.

Ces résultats indiquent que l'intégration de propriétés thermiques et physiques améliore considérablement la qualité du regroupement des alliages, et que les méthodes de clustering basées uniquement sur la composition sont insuffisantes pour capturer des structures complexes.

2.5 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.