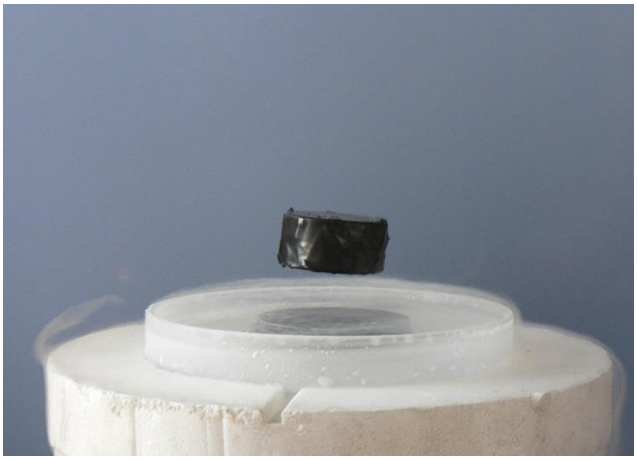
Intelligence Artificielle et applications en industrie 4.0

Rapport de Projet



Membre du groupe :

BADOLO Christian Thomas

NABI Daniel

Encadré par :

M. Sabeur ELKOSANTINI

Sommaire

1. Introduction

L'industrie 4.0 a apporté des avancées significatives dans les processus industriels en intégrant des technologies comme l'intelligence artificielle (IA) et l'apprentissage automatique (ML). Ces avancées ont permis une amélioration de la qualité des produits et des processus dans les industries, conduisant ainsi à l'émergence du concept de qualité 4.0. Dans ce contexte, l'industrie des super conducteurs n'est pas en reste. En raison de l'impact technologique et politique majeur que représentent les super conducteurs, l'utilisation des nouvelles technologies dans ce domaine est particulièrement pertinente. Ce rapport se concentre sur l'application de techniques d'IA et de ML à un ensemble de données sur la supraconductivité. Les données ont été généreusement mises à disposition le 10 novembre 2018, et comprennent deux fichiers contenant des informations sur 21263 supraconducteurs

La dataset se compose de deux fichiers principaux :

1. data.csv : Contient 81 caractéristiques extraites des 21263 supraconducteurs, ainsi que la température critique dans la 82ème colonne.

2. unique_m.csv : Comprend les formules chimiques décomposées pour tous les supraconducteurs du fichier train.csv. Les deux dernières colonnes contiennent la température critique et la formule chimique.

Ces données ont été collectées à partir de sources publiques (source en annexe). L'objectif principal de cette étude est de prédire la température critique des supraconducteurs en utilisant les caractéristiques extraites, ce qui représente un défi pertinent pour l'application de techniques avancées d'IA et de ML.

Le rapport suivant détaillera le processus d'analyse des données, la modélisation, l'évaluation des performances des modèles, et fournira des recommandations pour l'intégration de ces modèles dans des applications industrielles.

2. Compréhension du problème

2.1. Définition

Un superconducteur est un matériau qui, lorsqu'il est refroidi en dessous de sa température critique (ou de transition), perd toute résistance électrique et expulse entièrement le champ magnétique. En d'autres termes, un superconducteur permet à un courant électrique de circuler à travers lui sans aucune perte d'énergie due à la résistance électrique. Prédire la température critique d'un superconducteur est essentiel pour les industriels de garantir son comportement en termes de supraconductivité, sa qualité.

2.2. Caractéristiques essentielles

L'analyse des données sur la supraconductivité se concentre sur un ensemble de caractéristiques essentielles pour comprendre les propriétés des matériaux supraconducteurs. Parmi ces caractéristiques figurent le nombre d'éléments, la masse atomique moyenne, l'entropie atomique, la densité, la conductivité thermique, et d'autres mesures pertinentes. Ces données fournissent un aperçu détaillé des propriétés physiques et chimiques des matériaux, offrant ainsi une base solide pour étudier leur comportement thermodynamique et leur capacité à devenir des supraconducteurs.

2.3. Objectif

L'objectif principal de la modélisation est de prédire la température critique, une variable cruciale dans le domaine de la supraconductivité. Cette température critique est étroitement liée aux caractéristiques physiques et chimiques des matériaux, ce qui en fait une mesure clé pour évaluer leur potentiel supraconducteur. En utilisant les données disponibles sur les caractéristiques des matériaux, le modèle visera à établir des relations significatives entre ces caractéristiques et la température critique, contribuant ainsi à une meilleure compréhension des propriétés des supraconducteurs et ouvrant la voie à des applications industrielles plus avancées dans le domaine de la qualité 4.0.

3. Préparation des données

Les deux ensembles de données dont nous disposons ne comportent pas de valeurs manquantes. Nous avons décidé de ne pas utiliser les données relatives aux formules chimiques du fichier unique_m.csv, car ces données définissent les différentes caractéristiques présentes dans le fichier data.csv, telles que le nombre d'éléments, la masse atomique moyenne, l'entropie atomique, la densité, la conductivité thermique, etc. Ainsi, nous pouvons nous intéresser directement à ces caractéristiques sans avoir besoin de nous pencher sur la formule chimique elle-même.

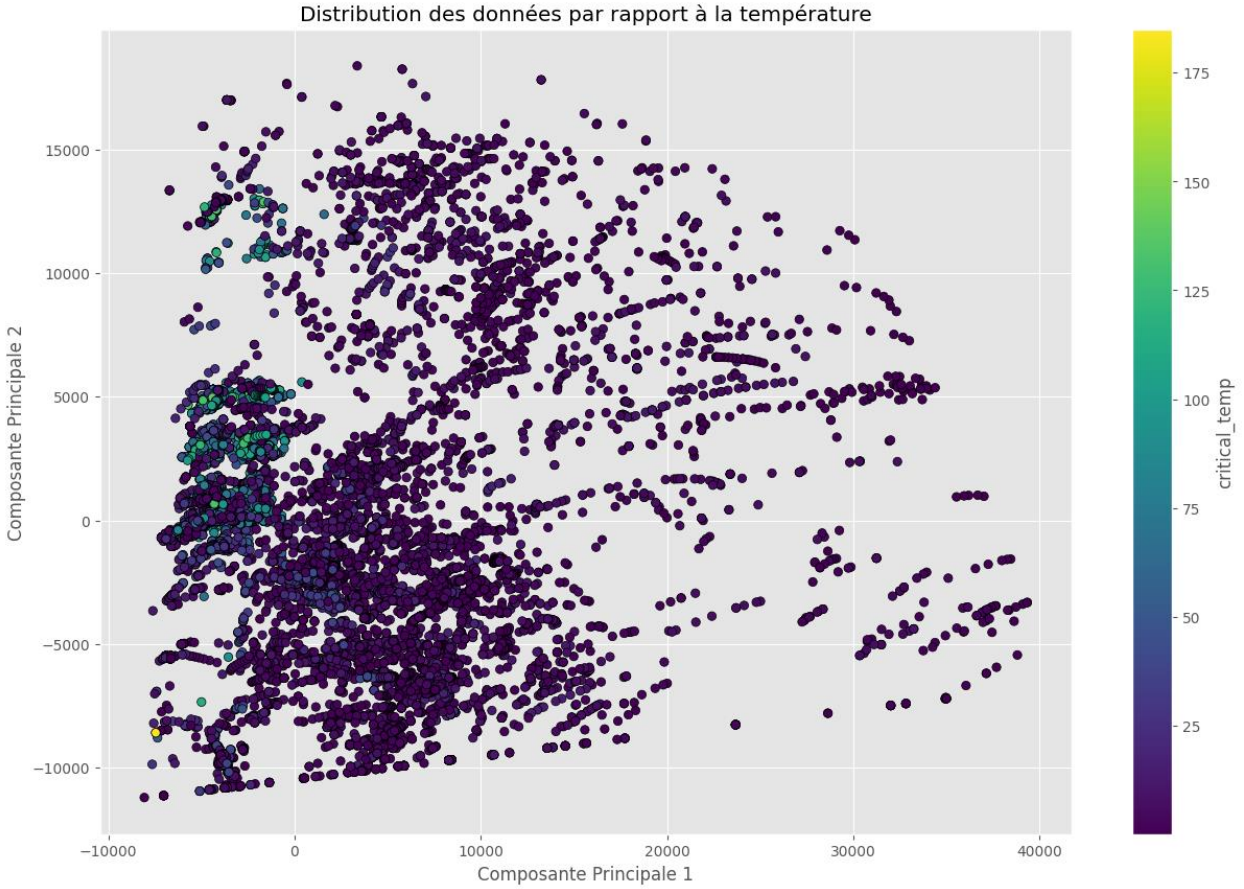
3.1. Visualisation des données

3.1.1. Matrice de corrélation

La matrice de corrélation (voir annexe) révèle des relations entre certaines features, telles que l'entropy_atomic_radius et l'entropy_fie qui sont fortement corrélées (corrélation de 1), ainsi que d'autres features comme mean_fie et mean_atomic_radius qui présentent des corrélations faibles avec la variable cible critical_temp (0.1 et 0.11 respectivement). Face à ces corrélations variées, il devient crucial d'utiliser des techniques de réduction de dimensionnalité pour sélectionner les features les plus appropriées. Une forte corrélation entre les features indique une certaine redondance, introduisant potentiellement de la multicollinéarité et affectant la stabilité et l'interprétation du modèle. D'un autre côté, une faible corrélation des features avec la variable cible critical_temp indique un manque de pertinence pour la prédiction de cette variable, ce qui nécessite une sélection soignée des features les plus informatives pour la tâche de prédiction. Dans ce contexte, l'utilisation de techniques de réduction de dimensionnalité telles que la sélection basée sur la corrélation de Pearson et l'Analyse en Composantes Principales (ACP) s'avère essentielle pour

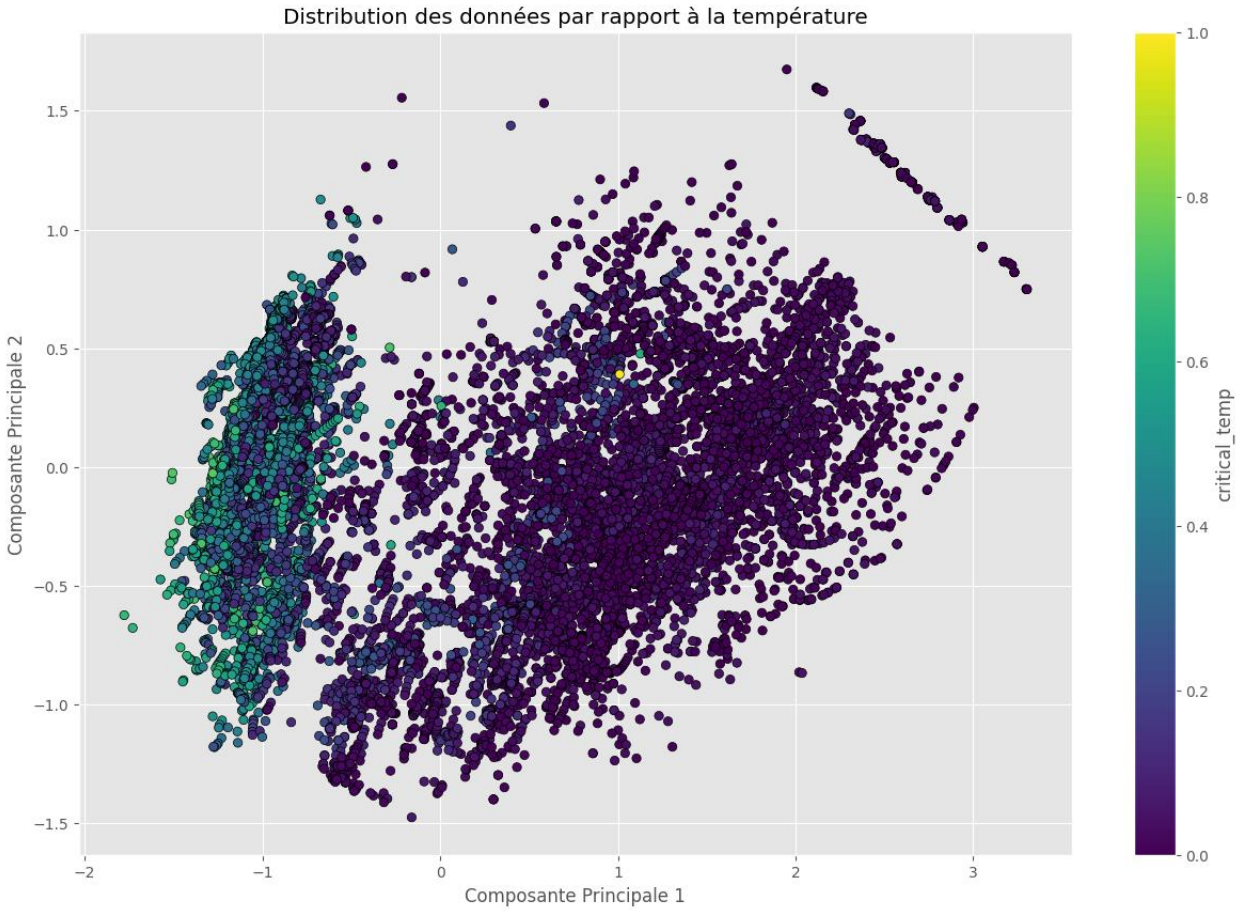
identifier les features les plus significatifs et améliorer ainsi la performance du modèle de prédiction de la température critique des supraconducteurs.

3.1.2. Visualisation de toutes les données non normalisées



La dispersion des données en fonction des deux composantes principales est significative, ce qui rend difficile la détection d'une relation linéaire des features avec la variable cible.

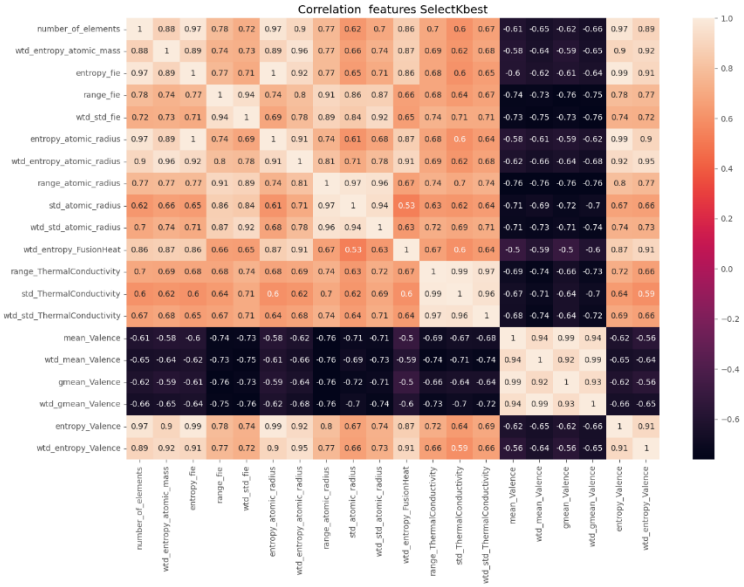
3.1.3. Visualisation après normalisation



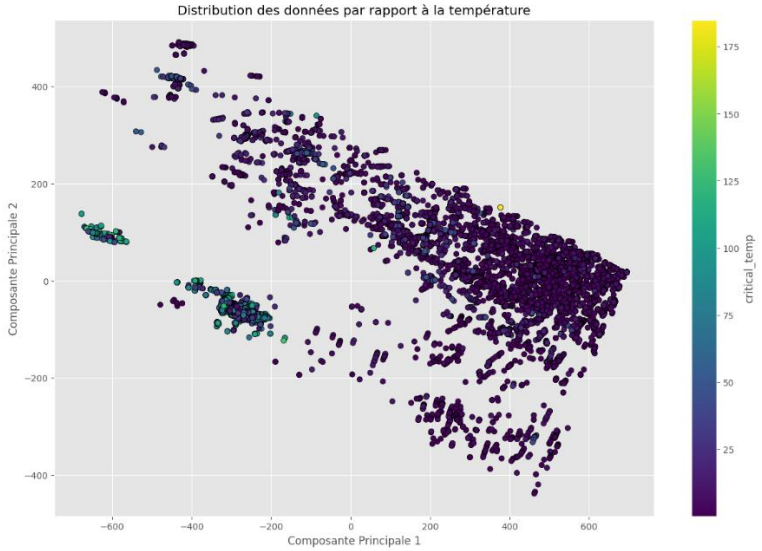
La faible dispersion des données suggère des résultats prometteurs en utilisant des modèles linéaires.

3.2. Sélection de features sur les données non normalisées

3.2.1. Sélection de features avec SelectKbest

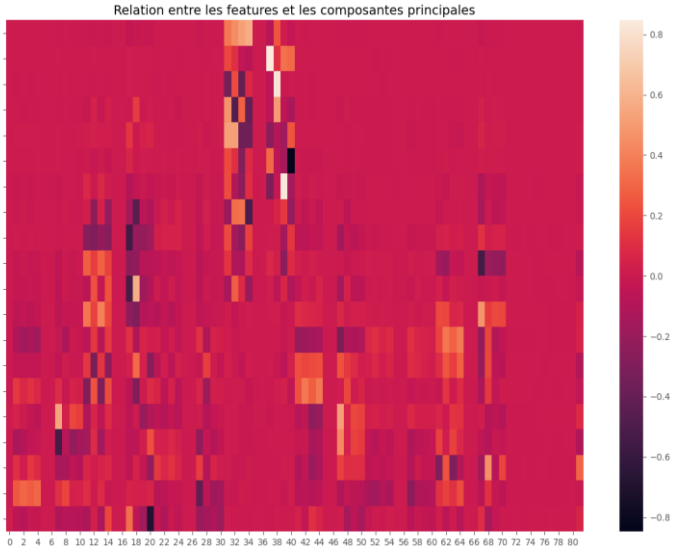


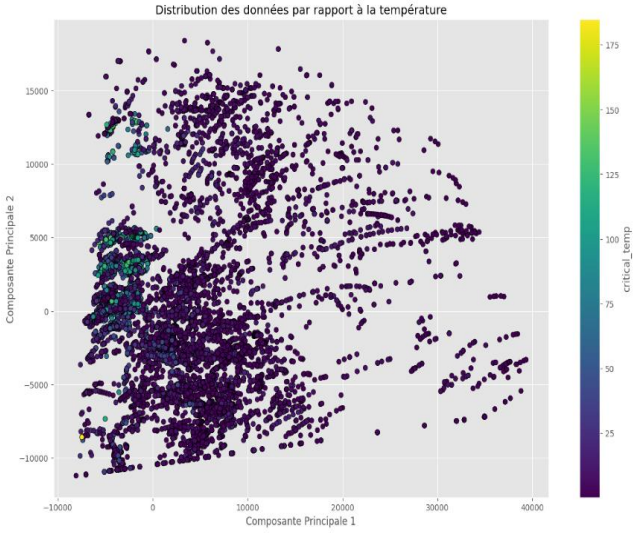
Les résultats obtenus ne sont pas satisfaisants en raison d'une forte corrélation entre la plupart des features.



Par contre la distribution des données présente une relation quasi linéaire

3.2.2. Sélection de features par l'ACP



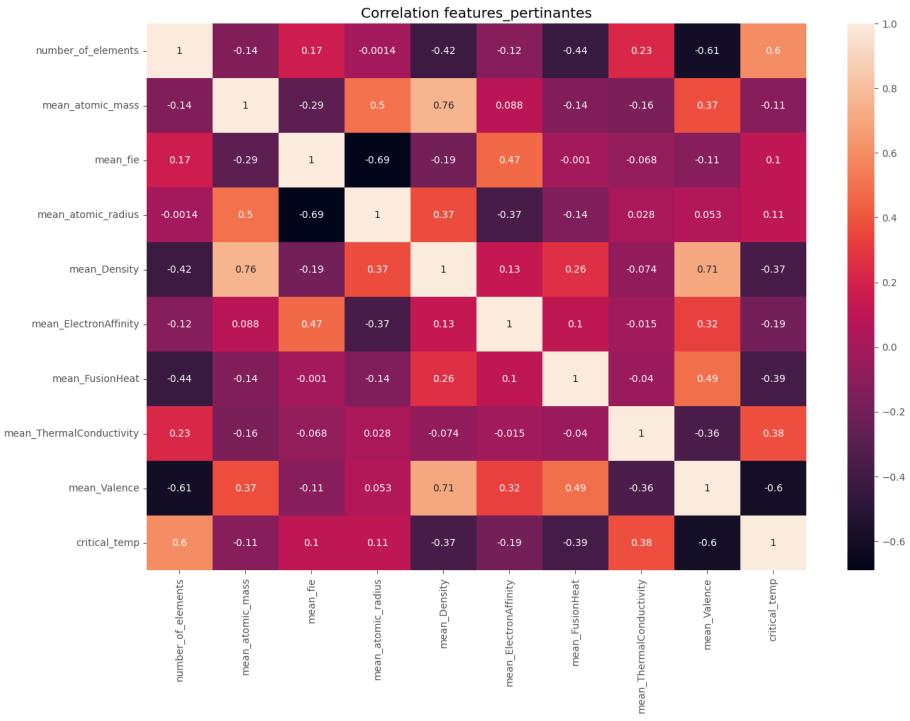


La matrice de corrélation est acceptable, mais les données sont très dispersées.

3.2.3. Sélection des features pertinentes basée sur la connaissance du domaine

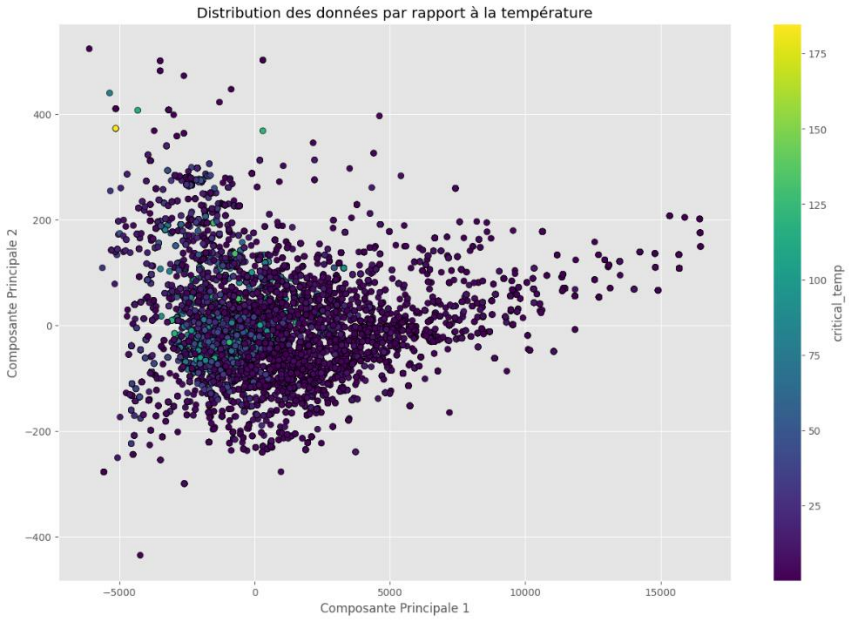
En analysant la dataset data.csv l'on a remarqué une répétition de features. Certaines features ont été obtenues par différentes mesures statistiques (moyenne, moyenne pondérée, moyenne géométrique, etc.) pour les mêmes propriétés sous-jacentes (masse atomique, la conductivité thermique, rayon atomique, etc.) Autrement dit plusieurs features mesurent différentes propriétés de manière similaire mais à travers des mesures statistiques différentes. Prenons l'exemple des caractéristiques liées à la masse atomique. Nous avons les features suivantes : 'mean_atomic_mass', 'wtd_mean_atomic_mass', 'gmean_atomic_mass', 'wtd_gmean_atomic_mass', 'entropy_atomic_mass', 'wtd_entropy_atomic_mass', 'range_atomic_mass', 'wtd_range_atomic_mass', 'std_atomic_mass', 'wtd_std_atomic_mass'.

Ces features expriment toutes des informations sur la masse atomique, mais chacune utilise une mesure statistique différente pour le faire (moyenne, moyenne pondérée, moyenne géométrique, entropie, etc.). Conserver toutes ces colonnes constitue probablement une redondance des données, car elles fournissent des informations similaires sur la même propriété sous-jacente. Ainsi, il est judicieux de supprimer certaines de ces features pour éviter la duplication des données et simplifier l'analyse tout en conservant les informations essentielles pour la modélisation. En supprimant, les features redondantes, l'on se retrouve avec 10 features pertinentes.



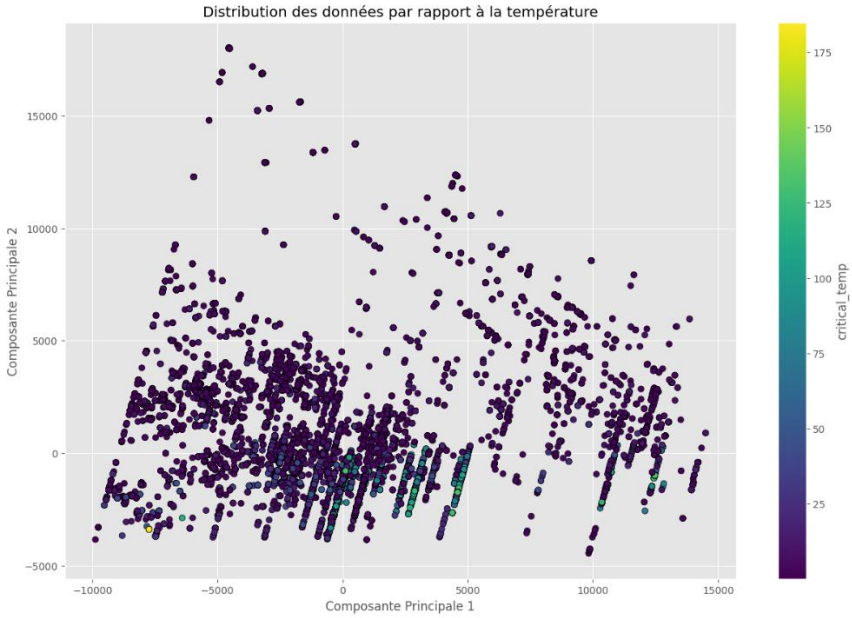
La faible corrélation entre les features est relativement acceptable, car chaque variable contribue de manière unique à la compréhension des données et peut potentiellement apporter des éléments importants pour la prédiction de la température critique des supraconducteurs. Cependant, certains features tels que le mean_atomic_mass et le mean_fie présentent une corrélation très faible avec la variable cible critical_temp. Il est crucial d'avoir des features fortement corrélés avec la variable

cible pour obtenir un modèle de prédiction efficace. Par conséquent, un modèle basé uniquement sur ces features risque de ne pas produire des résultats satisfaisants.



Cependant, nous pouvons observer une linéarité entre les features, ce qui indique que l'utilisation de modèles de régression linéaire pourrait être appropriée.

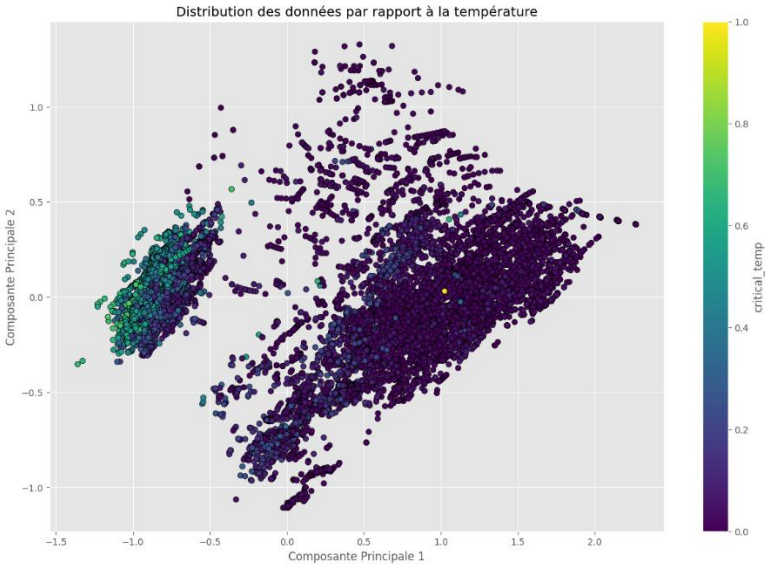
3.2.4. Sélection des features qui respectent un certain seuil de corrélation



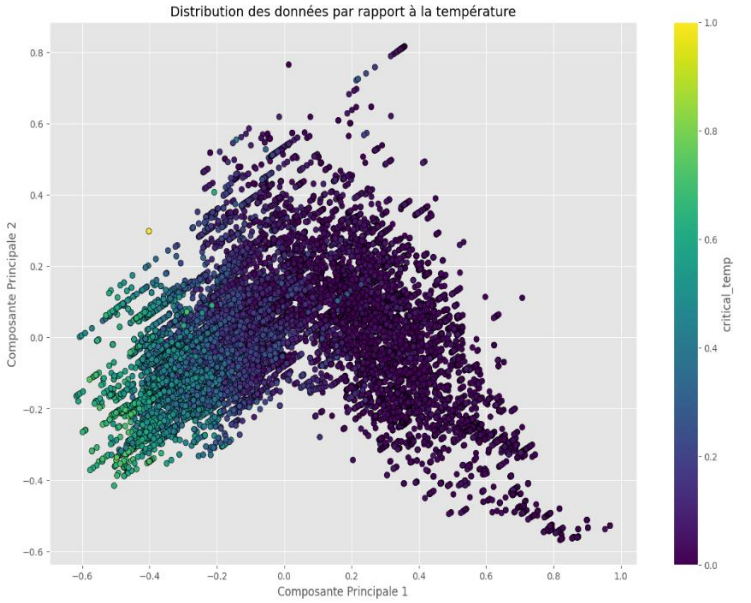
Dans ce cas également, la distinction d'une relation linéaire simple entre les données est complexe. Néanmoins, il est envisageable de tester des modèles de régression linéaire pour évaluer les résultats obtenus.

3.3. Sélection de features sur les données normalisées

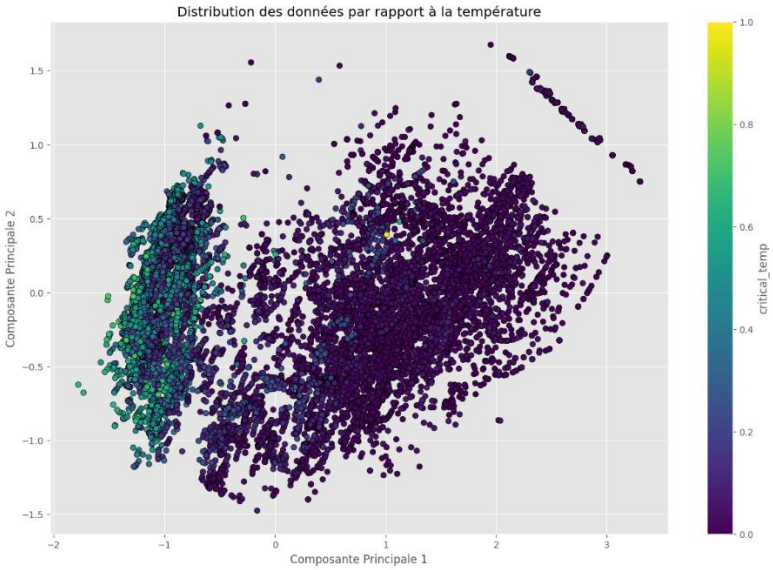
Sélection de features avec SelectKbest



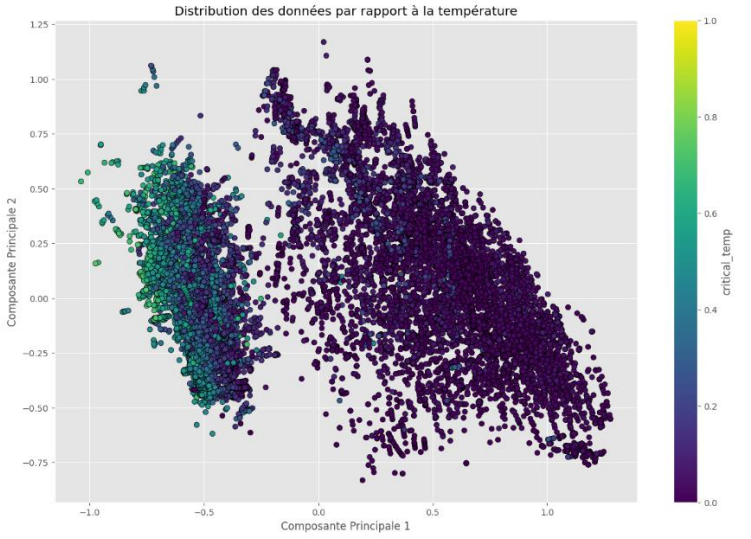
Sélection de features pertinentes basée sur la connaissance du domaine



Sélection de features par l'ACP



Sélection de features avec qui respectent un certain seuil de corrélation



D'après les graphiques ci-dessus, on remarque que la normalisation noie les relations linéaires. On a tendance à percevoir un problème de classification car les données semblent se diviser en classes distinctes. Cependant, dans notre cas, cette division n'est pas appropriée car l'objectif de l'étude est de prédire une valeur numérique (la température critique) plutôt que de classer par catégories.

4. Modélisation

Les données que nous traitons sont des données continues, ce qui exclut d'emblée l'utilisation de modèles de classification. Pour obtenir de meilleurs résultats, nous nous intéressons aux méthodes de sélection de caractéristiques qui ont produit des features présentant une linéarité intéressante entre elles. Ainsi, nous appliquerons des modèles de régression sur les features obtenues à partir de :

- La sélection des features les plus pertinents à l'aide de SelectKBest sur toutes les données non normalisées.
- La réduction de dimension par l'Analyse en Composantes Principales (ACP) sur toutes les données non normalisées.
- La sélection de features basée sur la connaissance du domaine sur toutes les données non normalisées.
- La sélection de features respectant un certain seuil de corrélation sur toutes les données non normalisées.

Bien que la matrice de corrélation de l'ensemble de données non normalisées soit intéressante, la visualisation n'a pas révélé de forte relation linéaire. Par conséquent, nous testerons un réseau de neurones et un modèle de régression basé sur un arbre de décision à ces données.

Pour évaluer la performance de nos modèles de régression, nous utiliserons deux métriques couramment utilisées : le coefficient de détermination R^2 et la racine carrée de l'erreur quadratique moyenne RMSE (Root Mean Squared Error).

- Le coefficient de détermination R^2 est une mesure qui indique la proportion de la variance de la variable cible (Y) expliquée par les features du modèle. Il varie de 0 à 1, où 1 indique une excellente adéquation du modèle aux données, c'est-à-dire que toutes les variations de la variable cible sont expliquées par le modèle. Un R^2 proche de 0, en revanche, signifie que le modèle n'explique pas bien la variance de la variable cible et que ses prédictions sont peu fiables.
- Quant au RMSE, il mesure la moyenne des erreurs entre les valeurs prédites par le modèle et les valeurs réelles de la variable cible. Il est exprimé dans la même unité que la variable cible et permet d'évaluer la précision des prédictions du modèle. Un RMSE plus faible indique des prédictions plus précises et une meilleure adéquation du modèle aux données.

En plus d'utiliser les métriques de régression classiques telles que le coefficient de détermination R^2 et la racine carrée de l'erreur quadratique moyenne RMSE, nous visualisons également une courbe d'apprentissage pour évaluer la performance de nos modèles. La courbe d'apprentissage est un outil visuel puissant qui montre comment évolue la performance du modèle en fonction de la taille de l'ensemble d'entraînement. En traçant la courbe d'apprentissage, nous pourrons observer si nos modèles souffrent de surajustement (overfitting) ou de sous-ajustement (underfitting). Un modèle sur ajusté aura une différence significative entre ses performances sur les données d'entraînement et sur les données de test, tandis qu'un modèle sous-ajusté montrera des performances médiocres sur les deux ensembles de données. Une courbe d'apprentissage idéale montrera une convergence

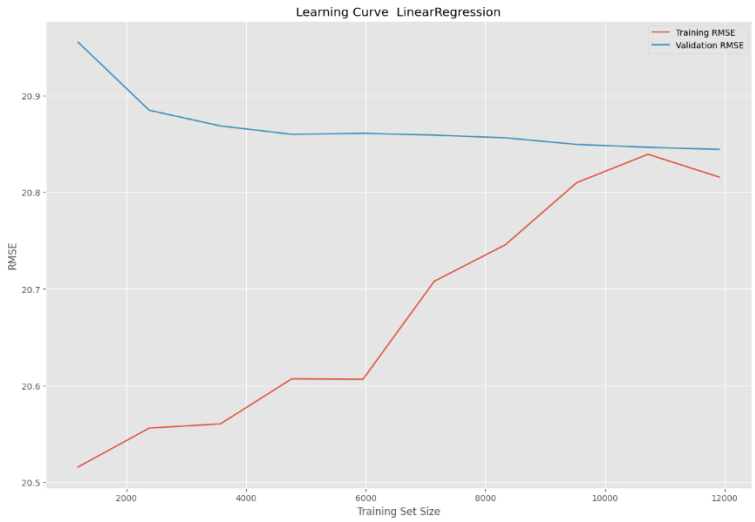
des performances entre les deux ensembles de données à mesure que la taille de l'ensemble d'entraînement augmente, indiquant ainsi une généralisation appropriée du modèle.

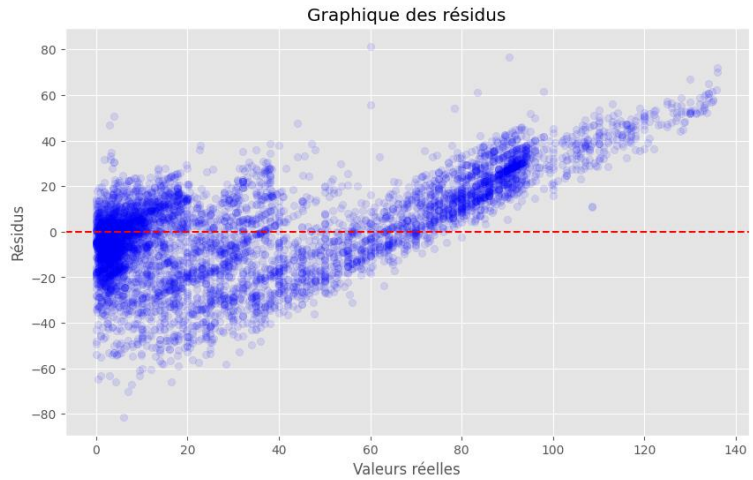
En combinant l'analyse des métriques R^2 et RMSE avec la visualisation de la courbe d'apprentissage, nous obtiendrons une évaluation approfondie de la performance de nos modèles de régression, ce qui nous permettra de prendre des décisions informées sur leur utilisation et leur amélioration.

5. Évaluation

Dans cette partie, nous examinerons les résultats obtenus incluant des courbes d'apprentissage, des graphiques de résidus, ainsi que les valeurs de R^2 et de RMSE pour les différents modèles utilisés.

Features avec SelectKbest

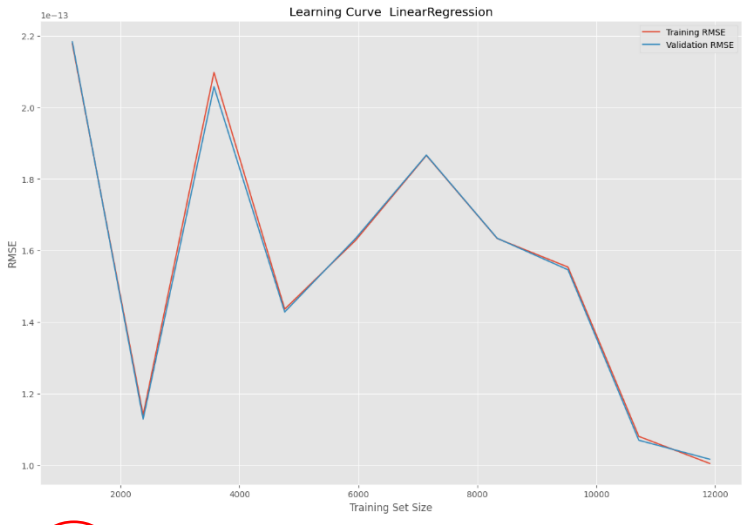


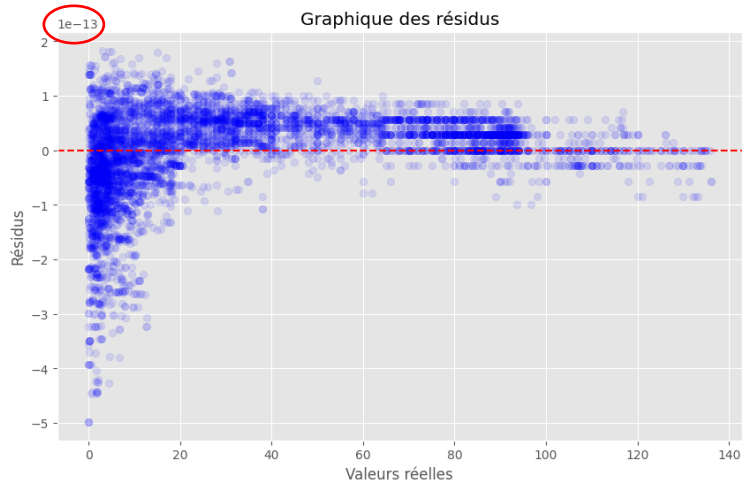


	Modèle	RMSE	R^2
0	LinearRegression	20.826736	0.631148

On observe une nette divergence entre les courbes d'apprentissage de l'entraînement et du test, suggérant ainsi un problème de surapprentissage (overfitting). La valeur élevée du RMSE du modèle confirme son inefficacité. De plus, la courbe des résidus révèle des erreurs trop grandes.

Features pertinentes basée sur la connaissance du domaine

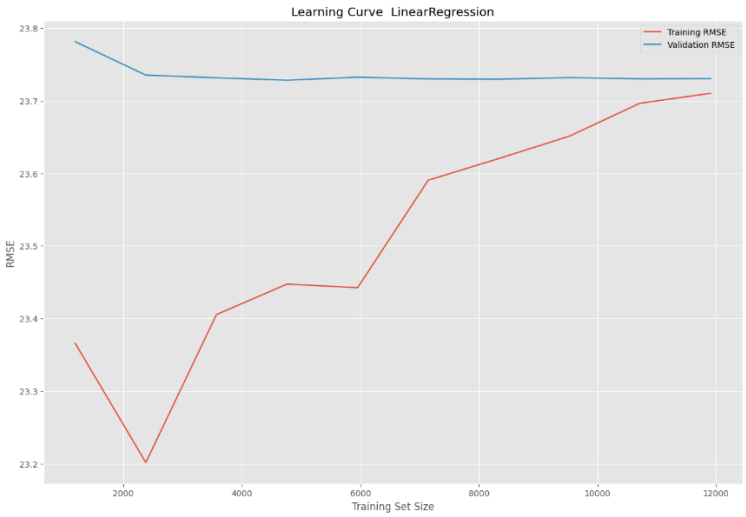


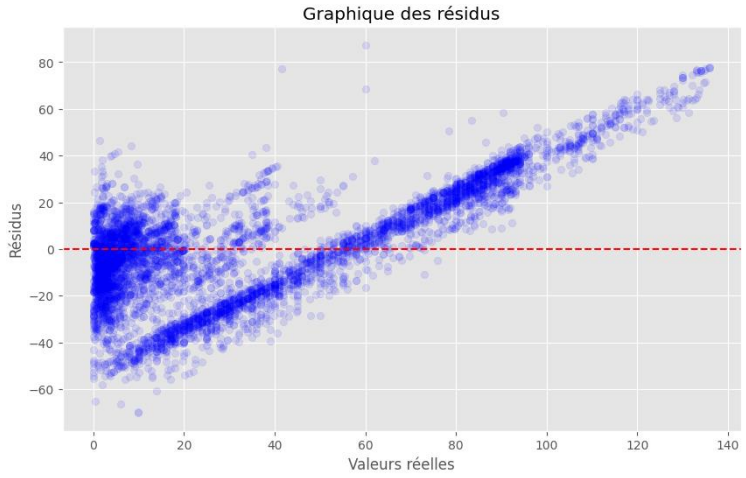


	Modèle	RMSE	R^2
0	LinearRegression	8.055894e-14	1.0

La courbe d'apprentissage de l'entraînement et du test se superpose quasi parfaitement, avec des valeurs acceptables du RMSE et du R^2. La dispersion des résidus de faibles valeurs confirme également ces résultats. En conclusion, le modèle est satisfaisant.

Features par l'ACP

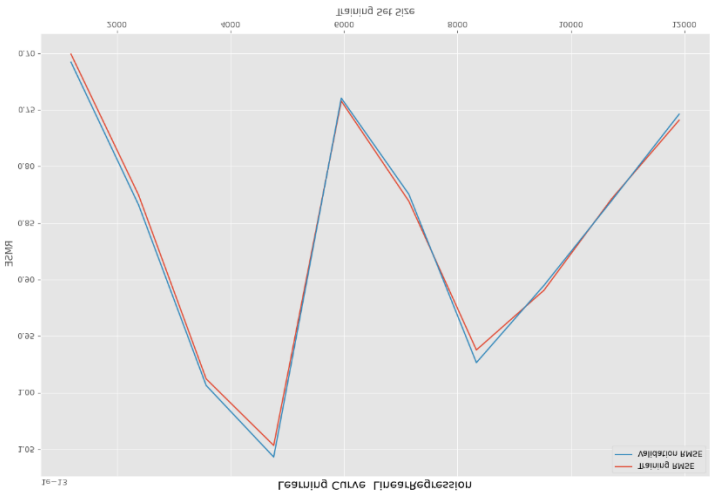


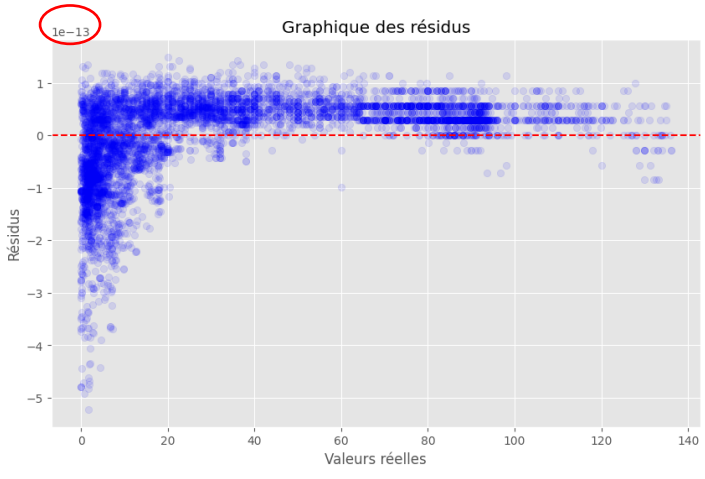


	Modèle	RMSE	R^2
0	LinearRegression	23.771202	0.519479

On observe une nette divergence entre les courbes d'apprentissage de l'entraînement et du test, suggérant ainsi un problème de surajustement. La valeur du RMSE et du R^2 du modèle confirme son inefficacité. De plus, la courbe des résidus révèle des erreurs importantes.

Features avec qui respectent un certain seuil de corrélation

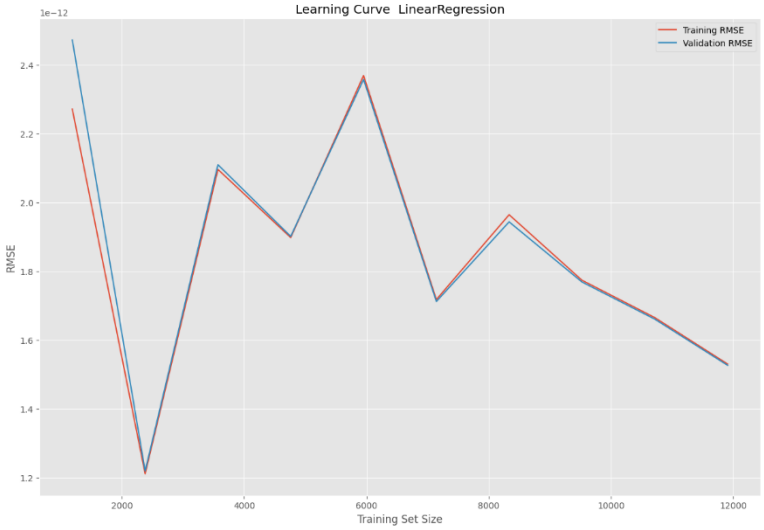


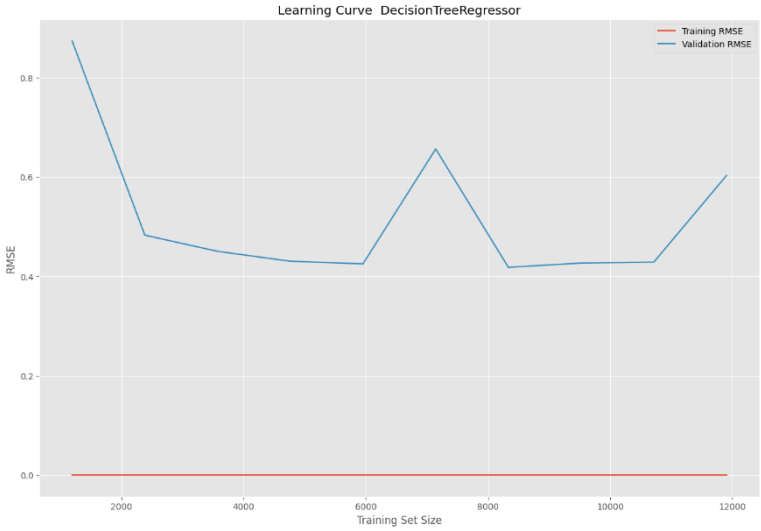


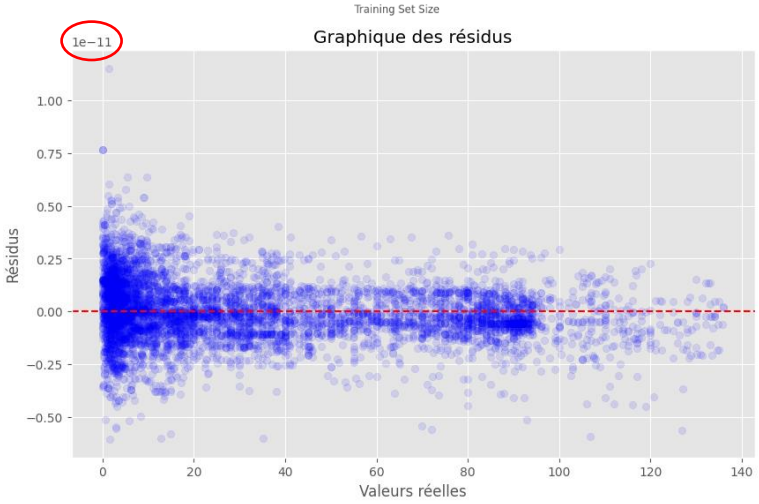
	Modèle	RMSE	R^2
0	LinearRegression	8.195134e-14	1.0

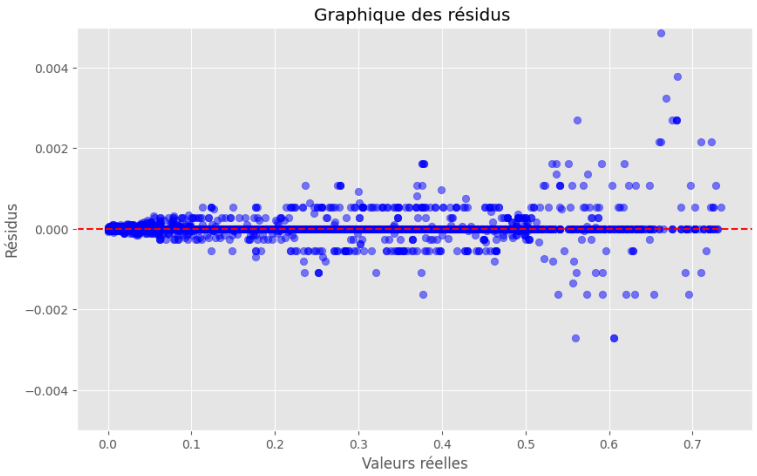
La courbe d'apprentissage de l'entraînement et du test se superpose parfaitement, avec des valeurs idéales de RMSE et de R^2. La dispersion des résidus de faibles valeurs confirme également ces résultats. En conclusion, ce modèle est aussi satisfaisant.

Toutes les features des données de base









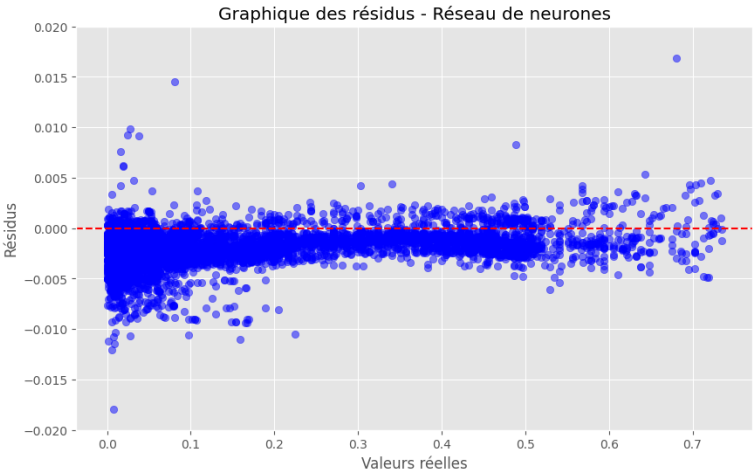
	Modèle	RMSE	R^2
0	LinearRegression	1.436632e-12	1.0

	Modèle	RMSE	R^2
0	DecisionTreeRegressor	0.000251	0.999998

La courbe d'apprentissage de l'entraînement et du test se superpose parfaitement, avec des valeurs idéales de RMSE et de R^2. La dispersion des résidus de faibles valeurs confirme également ces résultats. En conclusion, ce modèle est aussi satisfaisant.

On observe une nette divergence entre les courbes d'apprentissage de l'entraînement et du test, suggérant probablement un problème de surajustement. La valeur du RMSE et du R^2 du modèle confirme son inefficacité. De plus, la courbe des résidus révèle des erreurs importantes.

Réseau de neurones



	Modèle	RMSE	R^2
0	Reseau de neurone	0.001439	0.99994

La valeur du RMSE et du R^2 du modèle de réseau de neurone confirme son inefficacité. De plus, la courbe des résidus révèle des erreurs relativement faibles.

6. Déploiement

Pour déployer le modèle de prédiction de manière efficace et fiable dans une industrie 4.0 de fabrication de superconducteurs, il est nécessaire de garantir sa stabilité, sa performance et son intégration transparente avec les systèmes existants de l'usine. Pour cela, nous adapterons la démarche suivante :

- **Choix de l'infrastructure de déploiement** : Lorsque nous décidons de l'infrastructure de déploiement pour notre modèle de prédiction de température critique des superconducteurs, nous évaluons attentivement les différentes options en fonction des besoins spécifiques de l'usine. Nous considérons des facteurs tels que la latence, la sécurité et la capacité de traitement en temps réel. Après cette évaluation, nous sélectionnons l'infrastructure la mieux adaptée, qu'il s'agisse de serveurs locaux dans l'usine pour un contrôle total ou de services cloud pour une évolutivité maximale.
- **Configuration de l'infrastructure** : Une fois le choix d'infrastructure fait, nous procédons à sa configuration en provisionnant les ressources nécessaires, y compris les serveurs, les bases de données et les réseaux. Nous accordons une attention particulière à l'optimisation des performances, en utilisant des accélérateurs matériels si nécessaire et en mettant en place des mécanismes de mise en cache des données pour réduire les temps de latence.
- **Déploiement du modèle** : Le déploiement du modèle commence par son exportation dans un format compatible avec l'infrastructure choisie. Une fois exporté, nous le déployons sur l'infrastructure, en utilisant des conteneurs Docker pour assurer la portabilité ou en le déployant directement sur des instances de calcul dédiées.
- **Intégration avec les systèmes existants** : Une fois le modèle déployé, nous procédons à son intégration avec les systèmes existants de l'usine. Cela implique souvent le développement d'interfaces personnalisées pour permettre la communication bidirectionnelle entre le modèle et les systèmes de contrôle de processus, ainsi que l'intégration des données provenant des systèmes de gestion des données de l'usine.
- **Tests de déploiement** : Enfin, nous effectuons des tests de déploiement pour vérifier que le modèle fonctionne correctement dans l'environnement de production. Cela inclut des tests fonctionnels pour valider la précision des prédictions et des tests d'intégration pour assurer que le modèle s'intègre harmonieusement avec les systèmes existants de l'usine.

7. Conclusion

- Résumé des principales conclusions tirées de l'étude.
- Perspectives et suggestions pour des travaux de recherche ou des améliorations supplémentaires.

Ce projet s'est concentré sur l'application de techniques d'intelligence artificielle (IA) et d'apprentissage automatique (ML) à un ensemble de données sur la supraconductivité, dans le contexte de l'industrie 4.0. L'objectif principal était de prédire la température critique des supraconducteurs en utilisant des features extraites des données disponibles.

Dans le cadre de cette étude, plusieurs étapes ont été suivies. Tout d'abord, une compréhension approfondie du problème a été établie, définissant les caractéristiques essentielles des matériaux supraconducteurs et l'objectif de prédiction de la température critique. Ensuite, une préparation minutieuse des données a été effectuée, comprenant la visualisation, la matrice de corrélation, la normalisation des données et la sélection de features pertinentes.

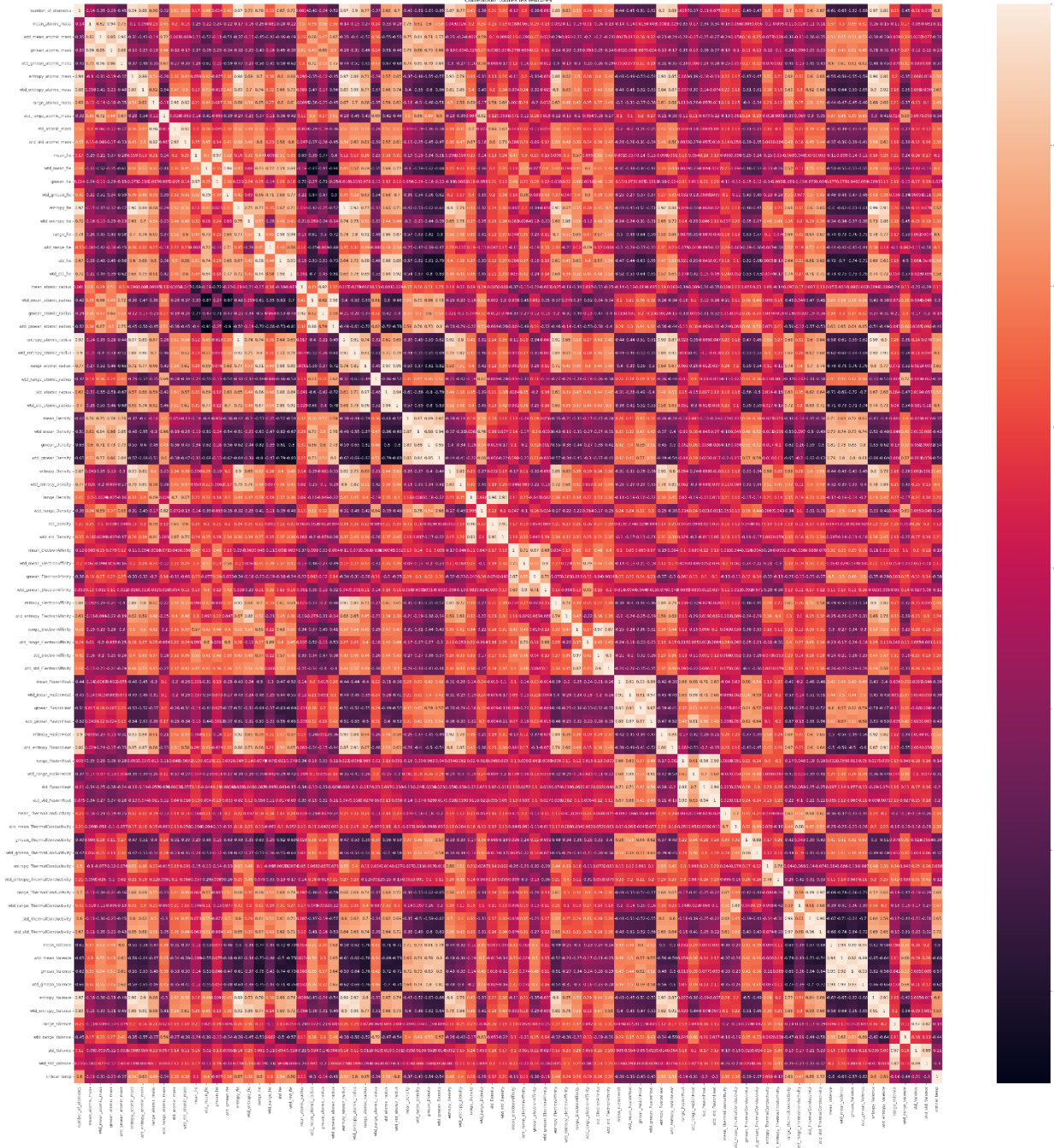
La modélisation a été abordée en utilisant différentes techniques, notamment la régression linéaire, les réseaux de neurones, et les arbres de décision. L'évaluation des modèles a été réalisée en utilisant des métriques telles que le coefficient de détermination R^2, la racine carrée de l'erreur quadratique moyenne RMSE, ainsi que des courbes d'apprentissage pour détecter le surajustement ou le sous-ajustement des modèles.

Les résultats obtenus ont montré que la sélection de features pertinentes basée sur la connaissance du domaine a conduit à des modèles plus performants. Le modèle de régression linéaire simple s'est avéré être le plus performant parmi les modèles testés, offrant les résultats les plus satisfaisants avec une **RMSE= 8,055894e-14** et un **R^2 = 1.**

En conclusion, ce travail démontre l'importance de la préparation des données, de la sélection de features et du choix judicieux des modèles pour obtenir des prédictions précises dans le domaine de la supraconductivité. Ces résultats peuvent avoir des implications significatives pour l'industrie 4.0 en améliorant la qualité des matériaux et des processus, tout en ouvrant la voie à de nouvelles avancées technologiques dans le domaine des supraconducteurs.

8. Annexes

Matrice de corrélation toutes les données



Bibliographie :

1. Hamidieh,Kam. (2018). Superconductivty Data. UCI Machine Learning Repository. https://doi.org/10.24432/C53P47
2. DataCamp. "Tutorial: Learning Curves". Available online: https://www.datacamp.com/tutorial/tutorial-learning-curves
3. Towards Data Science. "ANOVA for Feature Selection in Machine Learning". Available online: https://towardsdatascience.com/anova-for-feature-selection-in-machine-learning-d9305e228476
4. Connaissance des Énergies. "Supraconductivité : définition physique, applications dans l'énergie, enjeux et chiffres clés". Available online: https://www.connaissancedesenergies.org/
5. ITER - International Thermonuclear Experimental Reactor. "L'USINE CRYOGÉNIQUE". Available online: https://www.iter.org/
6. Statistics How To. "RMSE (Root Mean Square Error)". Available online: https://www.statisticshowto.com/probability-and-statistics/regression-analysis/rmse-root-mean-square-error/
7. Analytics Vidhya. "PCA: Practical Guide to Principal Component Analysis in Python". Available online: https://www.analyticsvidhya.com/blog/2016/03/pca-practical-guide-principal-component-analysis-python/
8. DataTab. "Tutorial: Pearson Correlation". Available online: https://datatab.net/tutorial/pearson-correlation
9. HAL - Hyper Articles en Ligne. "Les atomes. Caractéristiques et structure des atomes". Available online: https://theses.hal.science/tel-00009361/file/chapitre1.PDF
10. Énergie Nucléaire. "Supraconducteur : tout ce qu'il faut savoir sur ce matériau". Available online: https://www.energie-nucleaire.net/