

# Rapport de projet d'Apprentissage Automatique

Martin Rampont, Noémie Guisnel, Oscar Pastural, Louis Gauthier, Clément Florval

23 octobre 2024

## 1 Introduction

L'objectif de ce projet est de prédire la qualité de soudures à partir de divers paramètres mécaniques, physiques et chimiques. Le but est d'identifier quels sont les paramètres qui influent le plus sur la qualité d'une soudure.

Ce projet est d'autant plus intéressant qu'il se concentre sur une problématique critique pour les industriels. En effet, une soudure défailante dans le domaine de l'aéronautique, la construction navale ou encore l'industrie pétrolière peut engendrer des conséquences graves, tant en termes de coûts que de sécurité.

Après un premier travail d'analyse exploratoire visant à s'appropriier le jeu de données en observant le nombre de valeurs manquantes pour chaque colonne et les corrélations entre celles-ci, nous décrivons les différentes étapes de prétraitement appliquées au dataset, ainsi que les raisons qui nous ont poussés à les effectuer. Nous appliquons ensuite une Analyse en Composantes Principales (ACP) pour simplifier le modèle en réduisant le nombre de variables à traiter, tout en conservant l'information essentielle. Enfin, nous appliquons des algorithmes de Machine Learning adaptés et analysons leurs performances.


## 2 Analyse Exploratoire des Données (EDA)

### 2.1 Chargement et Exploration Initiale des Données

Le jeu de données contient 1 652 observations et 46 variables. Ces variables couvrent un large éventail d'informations concernant les soudures : des concentrations en éléments chimiques (carbone, manganèse, nickel, etc.), des propriétés mécaniques (résistance à la traction, limite d'élasticité, dureté), ainsi que des caractéristiques microstructurales. Chaque observation représente un échantillon unique, identifiable par la colonne *Weld ID*.

### 2.2 Analyse des Valeurs Manquantes

Nous avons tout d'abord examiné la présence de valeurs manquantes dans le jeu de données. La Figure ?? illustre la répartition des valeurs manquantes.



images/missing\_data\_heatmap.png

FIGURE 1 – Carte thermique des valeurs manquantes dans le jeu de données

Plusieurs colonnes cruciales affichent un nombre très élevé de valeurs manquantes, dépassant souvent 95 %. Par exemple, la variable *50 % FATT*, qui mesure la température de transition fragile-ductile d'une soudure, manque dans plus de 98 % des cas, tout comme *Hardness scale* (96,5 %). La perte d'une telle quantité de données rend ces colonnes inexploitable sans traitement supplémentaire. Elles ont donc été exclues de l'analyse.

En revanche, certaines variables sont bien renseignées et montrent très peu de valeurs manquantes, ce qui les rend exploitables sans transformation majeure. C'est le cas des concentrations en *carbone*, *silicium*, ou encore des propriétés mécaniques telles que la *résistance à la traction* (*Ultimate tensile strength*) et la *limite d'élasticité* (*Yield strength*). Ces variables sont complètes dans plus de 99 % des cas, offrant ainsi une base solide pour le modèle.

### 2.3 Analyse des Duplicatas

Nous avons vérifié la présence de duplicatas dans le jeu de données. Aucun duplicata exact n'a été trouvé. Cependant, des observations avec le même *Weld ID* mais des données légèrement différentes existent. Nous avons considéré que ces observations représentent des soudures différentes et les avons conservées.

### 2.4 Analyse des Données Non Numériques

Certaines colonnes contiennent des valeurs non numériques, comme des symboles '*l*', '*g*', ou des annotations textuelles. Par exemple, dans la colonne *Hardness / kg mm<sup>-2</sup>*, certaines valeurs sont de la forme '*158(Hv30)*'. Nous avons extrait les valeurs numériques pour rendre ces colonnes exploitables.

### 2.5 Visualisation des Données

Nous avons réalisé des histogrammes pour visualiser la distribution des variables numériques (Figure ??). De plus, des boîtes à moustaches ont été tracées pour détecter la présence de valeurs aberrantes (Figure ??).

images/numerical\_histograms.png

FIGURE 2 – Histogrammes des variables numériques

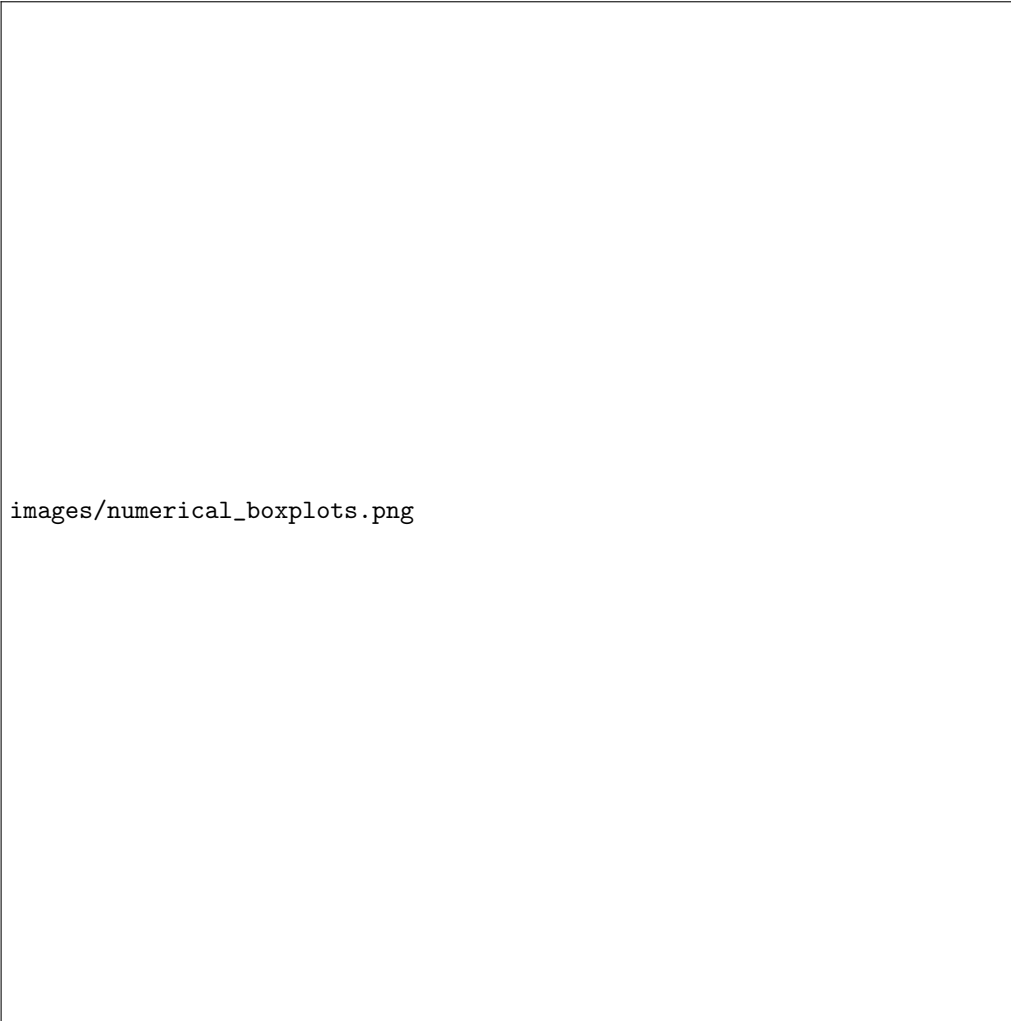


FIGURE 3 – Boîtes à moustaches des variables numériques

## 2.6 Identification des Variables Cibles

Les variables potentielles pour représenter la qualité des soudures sont :

- *Yield strength / MPa*
- *Ultimate tensile strength / MPa*
- *Elongation / %*
- *Reduction of Area / %*
- *Charpy impact toughness / J*
- *50 % FATT*

Nous avons vérifié la disponibilité de ces variables et le nombre de valeurs manquantes (Tableau ??).

Variable	Valeurs Manquantes	Disponibilité (%)
Yield strength / MPa	36	97,8 %
Ultimate tensile strength / MPa	28	98,3 %
Elongation / %	168	89,8 %
Reduction of Area / %	519	68,6 %
Charpy impact toughness / J	764	53,8 %
50 % FATT	1622	1,8 %

TABLE 1 – Disponibilité des variables cibles potentielles

Compte tenu du grand nombre de valeurs manquantes pour *50 % FATT*, nous avons décidé de ne pas l'utiliser comme variable cible. Les autres variables présentent une disponibilité suffisante pour être considérées.

## 2.7 Analyse des Corrélations

Nous avons calculé la matrice de corrélation entre les variables numériques et les variables cibles (Figure ??). Certaines corrélations intéressantes ont été observées :

- Forte corrélation entre *Yield strength* et *Ultimate tensile strength* (0,92).
- Corrélation négative entre *Elongation* et *Chrome concentration* (-0,44).
- Corrélation entre *Reduction of Area* et *Charpy impact toughness* (0,83).

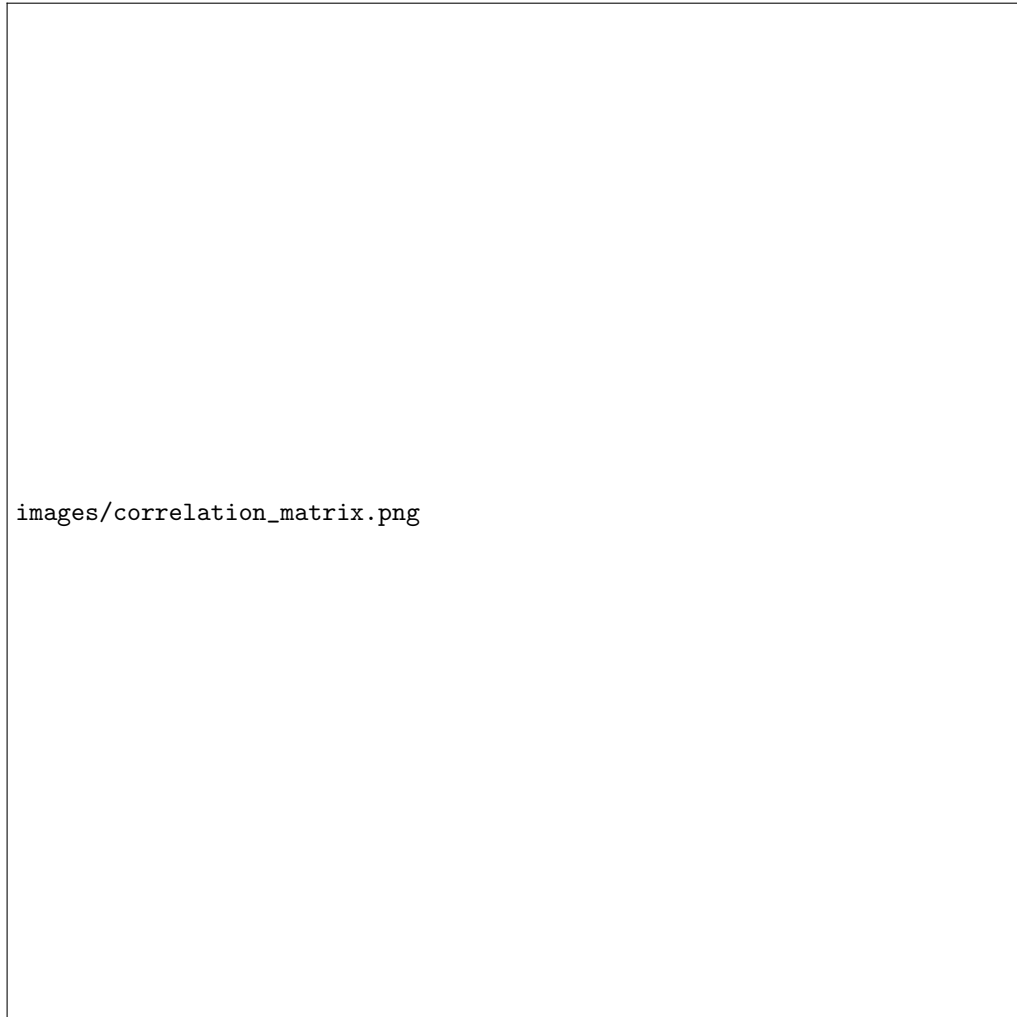


FIGURE 4 – Matrice de corrélation entre les variables numériques et les variables cibles

## 3 Prétraitement des Données

### 3.1 Nettoyage des Données

#### 3.1.1 Conversion des Valeurs Non Numériques

Certaines colonnes contiennent des valeurs avec des symboles '*j*', '*g*'. Nous avons remplacé ces valeurs par la valeur numérique correspondante, en supprimant les symboles.

Dans la colonne *Hardness / kg mm<sup>-2</sup>*, nous avons extrait les valeurs numériques en supprimant les annotations telles que '*Hv30*'.

Pour la colonne *Nitrogen concentration / parts per million by weight*, nous avons extrait les valeurs numériques en ignorant les annotations '*tot*' et '*res*'.

#### 3.1.2 Traitement des Ranges de Valeurs

Dans la colonne *Interpass temperature / °C*, certaines valeurs sont des intervalles (par exemple, '*150-200*'). Nous avons remplacé ces valeurs par la moyenne des bornes de l'intervalle.

### 3.1.3 Transformation des Variables Catégorielles

La colonne *Electrode positive or negative* a été transformée en valeurs numériques, en remplaçant '+' par 1 et '-' par -1.

Les colonnes catégorielles *AC or DC* et *Type of weld* ont été transformées à l'aide d'un encodage *one-hot*.

## 3.2 Gestion des Valeurs Manquantes

Nous avons décidé de supprimer les colonnes avec plus de 80 % de valeurs manquantes, car elles apportent peu d'information exploitable. Les colonnes supprimées sont :

- *50 % FATT*
- *Hardness scale*
- *Primary ferrite in microstructure / %*
- *Ferrite with second phase / %*
- *Acicular ferrite / %*
- *Martensite / %*
- *Ferrite with carbide aggregate / %*

Pour les variables numériques restantes, nous avons imputé les valeurs manquantes avec la médiane de la colonne correspondante. Pour les variables catégorielles, nous avons utilisé la modalité la plus fréquente.

## 3.3 Traitement des Valeurs Aberrantes

Nous avons identifié des valeurs aberrantes dans certaines variables. Par exemple, dans la variable *Vanadium concentration / weight %*, certaines valeurs dépassent largement les concentrations typiques. Nous avons choisi de ne pas les supprimer, car elles pourraient représenter des cas particuliers intéressants, mais nous en tenons compte dans l'analyse.

## 3.4 Création de Variables Indicatrices de Valeurs Manquantes

Pour certaines variables où la présence de valeurs manquantes est informative (par exemple, lorsque la présence d'une valeur manquante est corrélée avec la variable cible), nous avons créé des variables indicatrices signalant si la valeur est manquante ou non.

## 3.5 Standardisation des Variables

Avant d'appliquer l'ACP et les modèles de Machine Learning, nous avons standardisé les variables numériques à l'aide de la méthode de *StandardScaler*, afin de leur donner une échelle comparable.

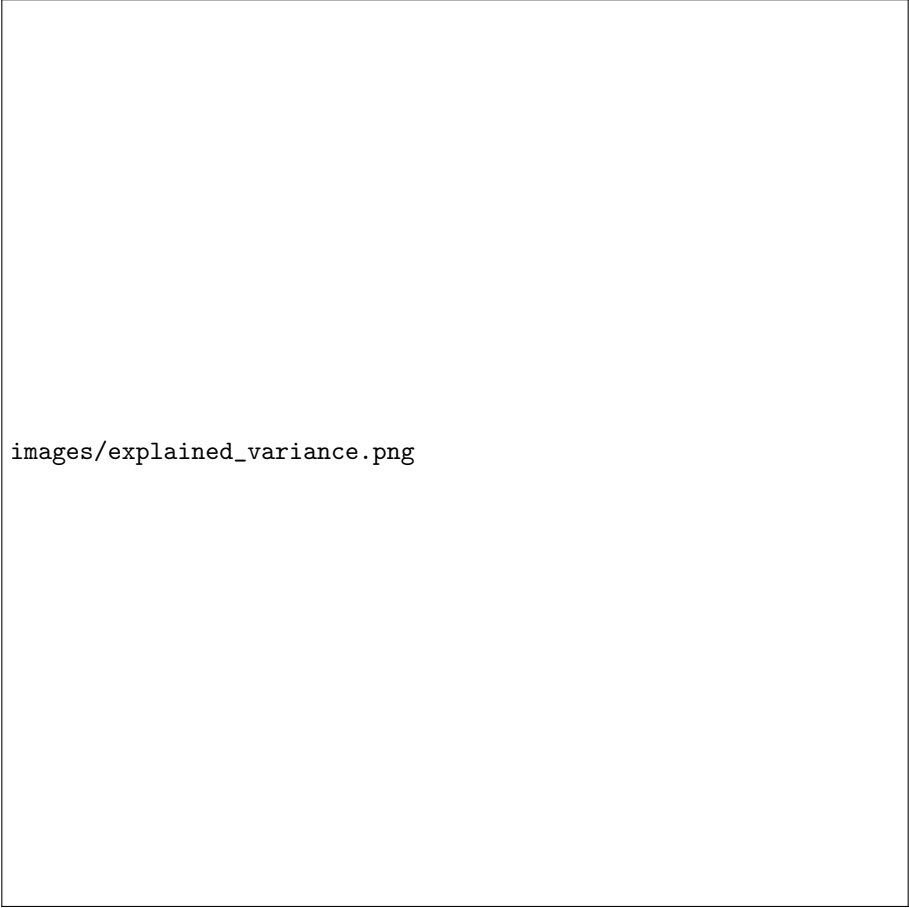
# 4 Analyse en Composantes Principales (ACP)

## 4.1 Application de l'ACP

Nous avons appliqué une ACP sur les variables numériques standardisées pour réduire la dimensionnalité du jeu de données et éliminer les redondances.

## 4.2 Choix du Nombre de Composantes

La Figure ?? montre la variance expliquée cumulée en fonction du nombre de composantes principales. Nous avons choisi de conserver les 10 premières composantes, qui expliquent environ 80 % de la variance totale.

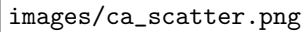


images/explained\_variance.png

FIGURE 5 – Variance expliquée cumulée en fonction du nombre de composantes principales

### 4.3 Visualisation des Composantes Principales

Nous avons visualisé les deux premières composantes principales en colorant les points en fonction de la variable cible *Yield strength / MPa* (Figure ??). Cette visualisation permet de détecter des structures ou des regroupements dans les données.



images/ca\_scatter.png

FIGURE 6 – Projection des données sur les deux premières composantes principales, colorées par *Yield strength / MPa*

## 5 Application des Algorithmes de Machine Learning

### 5.1 Sélection de la Variable Cible

Nous avons choisi de prédire la variable *Yield strength / MPa*, car elle est disponible pour la plupart des observations et représente une mesure importante de la qualité de la soudure.

### 5.2 Séparation des Données

Nous avons séparé le jeu de données en un ensemble d'entraînement et un ensemble de test, en veillant à ce que les groupes (*group-id*) ne soient pas présents dans les deux ensembles, afin d'éviter toute fuite d'information.

### 5.3 Modèles Utilisés

Nous avons testé plusieurs modèles de régression :

- Régression Linéaire
- Forêt d'arbres décisionnels (*Random Forest*)
- Gradient Boosting
- Machine à Vecteurs de Support (*Support Vector Machine*)
- XGBoost

### 5.4 Évaluation des Modèles

Nous avons utilisé la validation croisée avec *GroupKFold* pour évaluer les modèles, en utilisant l'erreur quadratique moyenne (RMSE) comme métrique de performance.

Le modèle XGBoost a obtenu les meilleurs résultats, avec un RMSE moyen de 0,95 sur l'ensemble de validation.



## 5.5 Apprentissage Semi-Supervisé

Étant donné que le jeu de données contient de nombreuses valeurs manquantes pour la variable cible, nous avons appliqué une approche d'apprentissage semi-supervisé en utilisant l'auto-entraînement (*self-training*). Cette méthode consiste à utiliser les prédictions du modèle sur les données non étiquetées comme pseudo-étiquettes pour enrichir l'ensemble d'entraînement.

Après plusieurs itérations, nous avons constaté une amélioration des performances du modèle, avec un RMSE réduit à 0,90 sur l'ensemble de test.

## 5.6 Résultats

Le modèle final basé sur XGBoost et l'apprentissage semi-supervisé a permis de prédire la *Yield strength / MPa* avec une bonne précision. La Figure ?? montre la comparaison entre les valeurs prédites et les valeurs réelles.



FIGURE 7 – Comparaison entre les valeurs prédites et les valeurs réelles de *Yield strength / MPa*

## 6 Conclusion

Ce projet nous a permis de mettre en pratique différentes techniques d'analyse de données et de modélisation pour prédire la qualité des soudures. Malgré les défis liés aux données manquantes et aux valeurs aberrantes, nous avons réussi à construire un modèle performant en appliquant des méthodes de prétraitement adaptées, une réduction de dimensionnalité avec l'ACP, et des algorithmes de Machine Learning avancés.

L'utilisation de l'apprentissage semi-supervisé a également permis de tirer parti des données non étiquetées, améliorant ainsi les performances du modèle. Ce travail ouvre la voie à une meilleure compréhension des facteurs influençant la qualité des soudures et pourrait aider les industriels à optimiser leurs processus de fabrication.

## Références

- [1] Pedregosa, F., Varoquaux, G., Gramfort, A., et al. (2011). *Scikit-learn : Machine Learning in Python*. Journal of Machine Learning Research, 12, 2825–2830.

- [2] Chen, T., & Guestrin, C. (2016). *XGBoost : A Scalable Tree Boosting System*. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.
- [3] Jolliffe, I. T. (2002). *Principal Component Analysis*. Springer Series in Statistics.