Rapport de projet d'Apprentissage Automatique

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

24 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 défaillante 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'approprier 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. Plusieurs colonnes cruciales affichent un nombre très élevé de valeurs manquantes, dépassant parfois 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* (91,6 %). La perte d'une telle quantité de données rend ces colonnes inexploitables, elles ont donc été exclues de l'analyse.

En revanche, certaines variables sont bien renseignées et ne montrent aucune valeur manquante, ce qui les rend exploitables sans transformation majeure. C'est le cas des concentrations en *carbone*, *silicium*, ou encore de *Heat input* ou *Type of weld*.

D'autres variables présentent un taux de valeurs manquantes plus modéré, notamment des indicateurs clés de la qualité des soudures, tels que la *Yield strength*, l'*Ultimate tensile strength*, l'*Elongation*, la *Reduction of Area* et la *Charpy impact toughness*. Ces variables proviennent de tests destructifs ; par exemple, le test Charpy, qui mesure la *Charpy impact toughness*, détruit la soudure, ce qui empêche de réaliser d'autres mesures comme la *tensile strength* sur le même échantillon. De plus, nous avons remarqué que plusieurs tests étaient souvent réalisés sur des soudures identiques (ayant exactement les mêmes paramètres), avec un test de *Yield strength* et *UTS*, suivi de plusieurs tests Charpy à différentes températures. Pour traiter ces cas, nous avons regroupé les observations correspondant à une même soudure afin de combler les valeurs manquantes et d'éviter toute fuite de données (*data leakage*) entre l'ensemble d'entraînement et les ensembles de test/validation.

Nous avons également comparé la distribution des cibles en fonction de la présence ou de l'absence de données pour chaque variable. Lorsque les distributions étaient similaires, nous avons

considéré que les données manquaient de manière complètement aléatoire (MCAR) et avons imputé les valeurs manquantes avec la médiane. Pour les autres variables, où les données manquaient de manière conditionnelle (MAR), nous avons testé l'imputation par KNN, mais les meilleurs résultats ont été obtenus en imputant la médiane et en ajoutant une colonne binaire indiquant la présence ou non de données manquantes.

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 *'<'*, *'>'*, ou des annotations textuelles. Par exemple, dans la colonne *Hardness / kg mm*$^{-2}$, 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 créé des histogrammes et des boîtes à moustaches pour visualiser la distribution des variables numériques et détecter la présence de valeurs aberrantes.

2.6 Identification des Variables Cibles

Pour que notre modèle soit utile en production sur des soudures réelles, il est essentiel que les variables d'entrée soient disponibles avant tout test destructif. Cela signifie que les variables dérivées de tests destructifs (par exemple, *Yield strength*, *Ultimate tensile strength*, *Charpy impact toughness*) ne peuvent pas être utilisées comme features dans le modèle. Seules des variables mesurables avant la destruction de l'échantillon peuvent servir d'entrées dans le modèle.

Un cas particulier est la variable *Charpy temperature*, qui correspond à la température à laquelle le test Charpy est réalisé. Cette température est arbitraire et ne peut pas être prédite, car il existe souvent plusieurs tests pour un même échantillon réalisés à des températures différentes. Nous utilisons donc cette variable comme input pour fixer la température du test, ce qui permet de rendre les résultats comparables.

Les variables potentielles pour représenter la qualité des soudures sont donc :
— *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 1).

Variable	**Valeurs Présentes**	**Disponibilité (%)**
Yield strength / MPa	780	47,2 %
Ultimate tensile strength / MPa	738	44,6 %
Elongation / %	700	42,4 %
Reduction of Area / %	705	42,7 %
Charpy impact toughness / J	879	53,2 %
50 % FATT	31	1,9 %

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. Certaines corrélations intéressantes ont été observées :

- Forte corrélation entre *Yield strength* et *Ultimate tensile strength* (0,92).
- L'*Elongation*, la *Reduction of area* et la *Charpy impact toughness* sont inversement corrélées avec la *Yield strength* et l'*Ultimate tensile strength*.
- Corrélation entre *UTS* et *Chrome concentration* (0,49).

Bien que ces variables cibles soient des indicateurs positifs de la qualité des soudures, représentant des propriétés mécaniques telles que la résistance et la ductilité, et que des valeurs plus élevées sont généralement associées à une meilleure performance, certaines corrélations inverses sont observées. Cela s'explique par un compromis inhérent (*tradeoff*) entre ces propriétés, où une augmentation de la résistance, par exemple, peut se faire au détriment de la ductilité.

3 Prétraitement des Données

3.1 Résumé des transformations effectuées

Catégorie	**Changements Apportés**
Suppression de Colonnes	
Weld ID	Suppression de la colonne car elle ne correspond pas à une caractéristique physique, chimique ou mécanique des soudures, mais à un identifiant non pertinent pour l'analyse.
Colonnes avec plus de 80% de valeurs manquantes	Suppression des colonnes ayant plus de 80% de valeurs manquantes. Ces colonnes sont considérées comme inexploitables à cause du trop faible nombre de données disponibles, comme *Primary ferrite*, *Ferrite with second phase*, et *Acicular ferrite*.
Gestion des Valeurs Manquantes	
Imputation numérique	Les valeurs manquantes des colonnes numériques sont imputées par la médiane. Bien que cette méthode ne conserve pas la distribution d'origine des données, elle est robuste aux valeurs aberrantes et offre une solution simple et efficace, surtout lorsque les données manquantes ne sont pas trop nombreuses.
Imputation catégorielle	Les valeurs manquantes dans les colonnes catégorielles sont imputées par la valeur la plus fréquente (*mode*), ce qui permet de maintenir une cohérence dans les données, bien que cela puisse introduire un biais si les données manquantes sont *MAR* (Missing At Random).
Colonnes avec indicateurs de manquants	Pour des colonnes où il a été déterminé que les données manquantes sont *MAR* (Missing At Random), telles que *Chromium concentration*, *Tungsten concentration*, et *Copper concentration*, des indicateurs binaires (*0 = non manquant*, *1 = manquant*) ont été créés. Cela permet de capturer l'impact potentiel des valeurs manquantes et d'améliorer la modélisation.
Traitement des Colonnes Numériques	
Colonnes avec signes $\leq$	Les valeurs précédées du signe $\leq$ sont remplacées par la moitié de ces valeurs pour fournir une estimation utilisable tout en restant proche des données originales.
Hardness / kg mm^{-2}	Les annotations non pertinentes comme *'158(Hv30)'* sont supprimées pour ne garder que la valeur numérique de la dureté. La colonne est ensuite convertie en valeurs numériques exploitables.
Interpass temperature / °C	Les valeurs exprimées sous forme d'intervalles (*150-200°C*) sont remplacées par la moyenne des bornes de l'intervalle afin de simplifier l'analyse. Les valeurs sont ensuite converties en données numériques.
Traitement des Colonnes Catégorielles	

Encodage catégoriel	Un encodage *one-hot* est réalisé pour les colonnes *AC or DC* et *Type of weld* afin de transformer les catégories en variables numériques. Une catégorie est supprimée dans chaque cas pour éviter la multicollinéarité.
Electrode positive or negative	Les signes + et - sont transformés en valeurs numériques : *1* pour une électrode positive et *-1* pour une électrode négative, facilitant ainsi leur utilisation dans les modèles de machine learning.

TABLE 2 – Tableau récapitulatif des transformations de données correspondant au prétraitement effectué.

3.2 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. Au lieu de les supprimer, nous avons appliqué une winsorisation pour limiter l'impact de ces valeurs extrêmes tout en conservant leur information. Cette méthode permet de réduire l'influence des outliers tout en prenant en compte ces observations dans l'analyse.

3.3 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.

3.4 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

Nous avons examiné 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.

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*. Cette visualisation permet de détecter des structures ou des regroupements dans les données.

5 Application des Algorithmes de Machine Learning

5.1 Sélection des Variables Cible

Nous avons choisi de prédire plusieurs variables essentielles à la qualité des soudures : *Yield strength / MPa*, *Ultimate tensile strength / MPa*, *Elongation / %*, *Reduction of Area / %*, et *Charpy impact toughness / J*. Ces variables sont des indicateurs clés qui permettent d'évaluer les propriétés mécaniques et la résistance globale des soudures, et donc leur qualité.

5.2 Modèles Utilisés

Durant notre phase de test, nous avons utilisé plusieurs modèles de régression :
— Régression Linéaire
— Forêt d'arbres décisionnels (*Random Forest*)
— Gradient Boosting
— Machine à Vecteurs de Support (*Support Vector Machine*)
— XGBoost

Après une phase d'expérimentation et d'optimisation des hyperparamètres via une recherche par grille (*grid search*), nous avons retenu le modèle XGBoost. Celui-ci a montré des performances supérieures en termes de précision prédictive, tout en offrant une certaine explicabilité grâce à l'analyse de l'importance des caractéristiques (voir Figure 1).

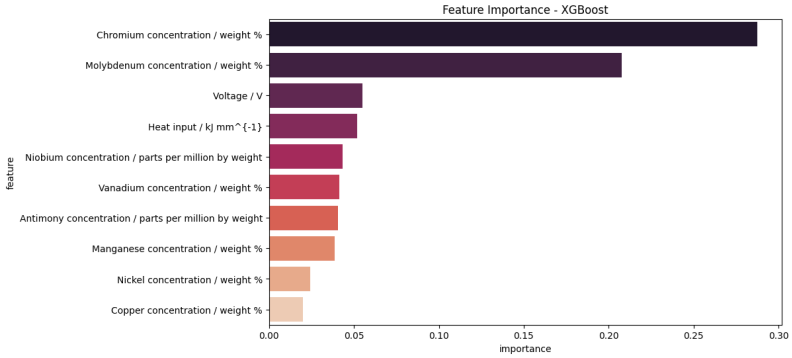

FIGURE 1 – Importance des caractéristiques du modèle XGBoost

5.3 Évaluation des Modèles

Nous avons évalué nos modèles à l'aide de la validation croisée en utilisant la méthode *GroupK-Fold*. Deux métriques de performance ont été retenues : l'erreur quadratique moyenne (RMSE) et le coefficient de détermination (R2). Le modèle XGBoost a montré les meilleures performances globales sur nos 5 variables cibles, avec une RMSE moyenne (standardisée) inférieure à 0,5 et un R2 élevé ($>$0.8), indiquant une bonne capacité de prédiction.

5.4 Apprentissage semi-supervisé

Initialement, nous avions envisagé l'utilisation d'une méthode d'apprentissage semi-supervisé basée sur l'auto-entraînement (*self-training*) pour pallier les valeurs manquantes des variables cibles. Cependant, une analyse plus approfondie du jeu de données a révélé que la majorité des valeurs manquantes étaient en réalité dues à un mauvais regroupement des données dans le dataset, et non à des absences réelles de mesures. En conséquence, nous avons abandonné cette approche, d'autant plus que les tentatives d'utilisation de l'auto-entraînement n'ont pas donné de bons résultats.

5.5 Résultats

Le modèle final basé sur XGBoost a montré une précision élevée pour toutes les variables cibles sélectionnées, avec une bonne capacité de généralisation. La Figure 2 illustre la comparaison entre les valeurs prédites et les valeurs réelles pour la variable *Yield strength / MPa*.

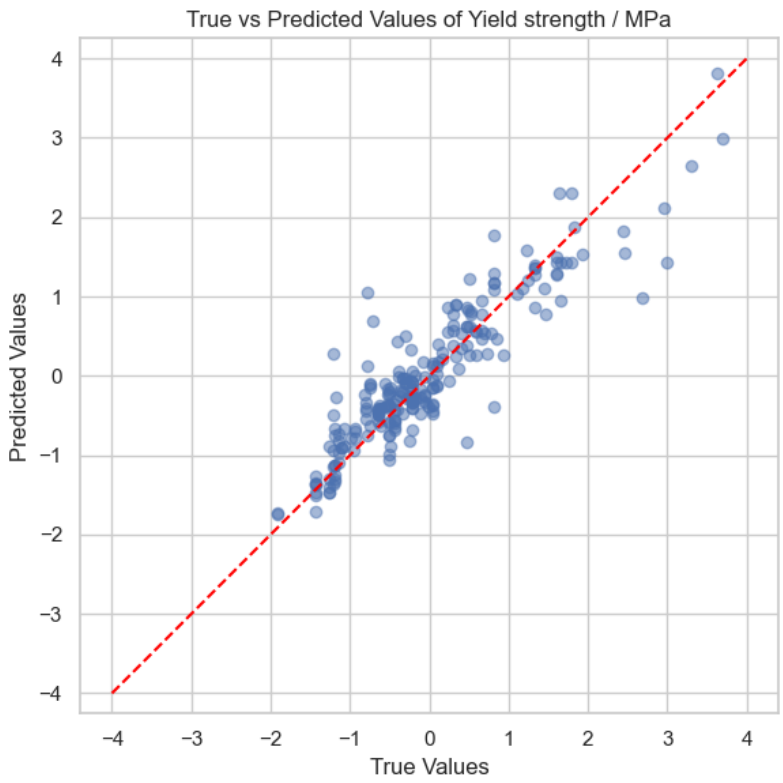

FIGURE 2 – Comparaison entre les valeurs prédites et les valeurs réelles de *Yield strength / MPa*

6 Développement d'une Interface Web

Afin de rendre notre modèle accessible et de démontrer son utilité pratique, nous avons développé une application web à l'aide du framework Flask. Cette interface permet aux utilisateurs d'entrer manuellement les paramètres de soudure et d'obtenir une estimation de la qualité pour chaque variable cible ainsi qu'un score global.

De plus, l'application offre la possibilité d'uploader un fichier CSV contenant plusieurs échantillons de soudures. Le modèle prédit alors la qualité des soudures pour chaque échantillon, et un fichier CSV avec les prédictions peut être téléchargé.

Cette application web facilite l'interaction avec le modèle et permet une évaluation rapide de la qualité des soudures en fonction de différents paramètres. Elle pourrait être particulièrement utile pour les industriels souhaitant estimer la qualité de nouvelles soudures sans avoir à effectuer de tests destructifs immédiats.

7 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.