CIGI QUALITA MOSIM 2023
Intégration de connaissances du domaine et de l'apprentissage automatique pour l'estimation des paramètres de fabrication

ABDOUL RAHIME DIALLO[1], ABDOURAHIM SYLLA[2]

[1] ARTS ET METIERS INSTITUTE OF TECHNOLOGY, UNIVERSITE DE LORRAINE, LCFC, HESAM UNIVERSITE
F-57070 Metz, France
abdoul-rahime.diallo@ensam.eu

[2] UNIV. GRENOBLE ALPES, CNRS, GRENOBLE INP, G-SCOP
46 Av. Félix Viallet, 38000 Grenoble, France
abdourahim.sylla@grenoble-inp.fr

***Résumé* – Dans le marché concurrentiel actuel, les clients demandent des produits personnalisés avec des délais de livraison courts. Dans un tel contexte où de nombreuses entreprises sont en concurrence pour les mêmes opportunités, une estimation rapide, précise et fiable des paramètres de fabrication (coût et temps) est indispensable pour gagner des appels d'offre. Cet article propose donc une nouvelle approche pour l'estimation des paramètres de fabrication. Elle est basée sur l'exploitation de données structurées, de données textuelles non structurées et des connaissances du domaine. Les premières expérimentations réalisées à l'aide d'un cas industriel montrent le grand potentiel de l'approche.**

***Abstract* – In today's competitive market, customers demand highly customized products with short lead times. In such a context where many companies are competing for the same opportunities, a quick, accurate, and reliable estimation of manufacturing parameters (cost and time) is essential to win tenders. This article therefore proposes a new approach for the estimation of manufacturing parameters. The proposed approach is based on the exploitation of structured data, unstructured textual data and domain knowledge. The first experiments realized thanks to an industrial use case show the great potential of this approach.**

***Mots clés* - Fabrication à la commande, Estimation de paramètres de fabrication, Fouille de texte, Apprentissage automatique, Connaissances du domaine**
***Keywords* – Make to order, Manufacturing parameters estimation, Text mining, Machine learning, Domain knowledge**

1 INTRODUCTION

Dans les situations industrielles de type « Make-To-Order (MTO)» ou « Engineer-To-Order (ETO)», ce sont les demandes des clients qui déclenchent la fabrication des produits (Hicks et al., 2000). Ces demandes sont de plus en plus personnalisées et diversifiées (Aldanondo & Vareilles, 2008). Ainsi, dans de nombreux secteurs industriels, afin d'augmenter leur chiffre d'affaires et de rester compétitives, les entreprises (fournisseurs) proposent une grande variété de produits (Pallant et al., 2020). De plus, pour être performant, un produit doit être de bonne qualité, proposé à un prix raisonnable, et livré dans un délai court (B. Kingsman et al., 1996). Un prix raisonnable étant considéré comme celui qui n'est pas trop élevé afin d'être accepté par les clients et pas trop bas afin de permettre à l'entreprise de faire des bénéfices. Confrontées à une augmentation sans précédent de sollicitations, les fournisseurs se doivent de réaliser une estimation rapide et fiable de la qualité, du coût et du temps de fabrication de leurs produits (García-Crespo et al., 2009).

Cependant, dans ces contextes MTO et ETO, généralement, afin de transmettre une offre à un client, le fournisseur doit estimer ces paramètres sans avoir réalisé la fabrication du produit (Karaoglan & Karademir, 2017). En fonction de la nature des besoins du client, deux approches d'estimation sont souvent utilisées. Lorsque le produit répondant aux besoins exprimés par le client est un « produit standard », c'est-à-dire un produit qui a déjà été fabriqué et livré à un client dans le passé, son estimation existante est réutilisée et, éventuellement, mise à jour par un expert. Lorsque le produit répondant aux besoins du client est un « produit non standard », c'est-à-dire un produit qui n'a pas encore été fabriqué par le fournisseur, les opérations et les ressources de fabrication nécessaires sont d'abord définies. Ensuite, sur la base de ses expériences, un expert effectue l'estimation des paramètres. Cependant, dans un contexte industriel caractérisé par un grand nombre de sollicitations et des processus de fabrication complexes, cette approche est chronophage, génère une charge cognitive élevée pour les experts, et peut conduire à des estimations imprécises et peu fiables (Sylla et al., 2021). Il est important de mentionner que cette problématique est plus fréquente dans les petites et moyennes entreprises (PME) qui manquent d'outils d'estimation performants. Dans la plupart de ces PME, les connaissances en matière d'estimation sont détenues par un expert (García-Crespo et al., 2009; Serrat et al., 2013). L'absence ou le départ de ce dernier pénalisent fortement l'entreprise. Il est donc primordial pour ces entreprises de se doter d'un outil d'estimation pertinent qui leur

permette de réaliser rapidement des estimations précises et fiables.

Fort heureusement, ces dernières années, l'introduction des nouvelles technologies de l'information et de la communication (capteurs, puces RFID, Enterprise Resource Planning (ERP) et Manufacturing Execution System (MES)) dans les PME offre de nouvelles opportunités (Murphy et al., 2019). De grandes quantités de données sont collectées sur les besoins des clients, les produits associés et leurs processus de fabrication. Ces données sont souvent stockées dans des systèmes d'information de type ERP et MES. Elles peuvent prendre différentes formes, notamment tabulaire structurée, image, mais aussi du texte non structuré. Elles représentent un gisement pouvant être exploité pour développer un outil d'estimation rapide, précis et fiable (Wang et al., 2016). De plus, comme nous l'avons dit plus haut, dans les PME, les experts détiennent des connaissances utiles à l'estimation des paramètres de fabrication. Ces connaissances, lorsqu'elles sont bien exploitées, permettent de tirer le meilleur profit des masses de données pour développer un outil d'estimation performant. Or, dans la littérature scientifique, de nombreux travaux reportés sur l'estimation des paramètres de fabrication exploitent uniquement des données structurées. Des données textuelles non structurées sont très rarement exploitées. Par ailleurs, la plupart des travaux n'intègrent pas les connaissances du domaine dans le développement des modèles d'estimation. Par connaissances du domaine, nous entendons les connaissances expertes implicites et les connaissances explicites générales telles que des formules physiques.

Cet article s'intéresse à ce problème et présente une nouvelle approche de développement de modèles d'estimation de paramètres de fabrication à l'aide de modèles d'apprentissage automatique. L'approche proposée est basée sur la méthode « Cross Industry Standard Process for Data Mining (CRISP-DM) » (Chapman et al., 2000). C'est la méthode la plus couramment utilisée dans les projets d'apprentissage automatique (Plotnikova et al., 2020). L'originalité de la nouvelle approche proposée dans cet article repose, d'une part, sur l'exploitation simultanée de différentes formes de données (tabulaire structurée et textuelle non structurée) et, d'autre part, sur l'intégration effective de connaissances du domaine aux différentes étapes de la méthode CRISP-DM. Il faut préciser que pour la suite de l'article, le focus est placé sur l'estimation du coût de fabrication. Le reste de l'article est structuré comme suit. Dans la section 2, une revue de la littérature sur l'application des techniques d'apprentissage automatique à l'estimation de coût de fabrication est présentée. Ensuite, dans la section 3, l'approche proposée est décrite avant de présenter dans la section 4 les premiers résultats issus des expérimentations sur un cas industriel. Dans la section 5, les conclusions et les futures recherches sont présentées.

2 REVUE DE LA LITTERATURE

Dans la littérature, deux approches sont principalement utilisées pour l'estimation de coût de fabrication. La première utilise directement des modèles d'apprentissage automatique pour estimer le coût. Alors que la seconde considère le temps de fabrication comme le paramètre principal pour l'évaluation du coût de fabrication. Elle utilise l'apprentissage automatique pour estimer ce temps. Ensuite, une formule simple est utilisée pour calculer le coût sur la base du temps estimé. Dans la suite, les travaux sont regroupés selon les secteurs d'application.

Dans le secteur aéronautique, il apparaît que la capacité des sous-traitants pures (ne fabriquant que selon les besoins fournis par les clients) à rapidement faire des propositions avec des prix précis et fiables est un facteur clé qui leur permet de décrocher des commandes (de Cos et al., 2008). Ainsi, plusieurs travaux portent sur l'automatisation de l'estimation du coût de fabrication afin d'aider ces sous-traitants à rapidement réaliser des chiffrages précis et fiables. Dans (de Cos et al., 2008), les auteurs ont appliqué un modèle « Artificial Neural Network (ANN) » et montre que ce dernier est suffisamment précis pour prédire le coût de fabrication des composants de turbines pour l'aviation. Le modèle ANN s'est révélé plus performant que deux méthodes statistiques, à savoir la méthode « Projection Pursuit Regression » et la méthode « Local Polynomial Regression ». Dans (Deng & Yeh, 2011), le temps de fabrication est estimé afin de déterminer le coût de fabrication. Les auteurs ont comparé quatre modèles d'estimation de temps de fabrication des pièces de structures des avions. Les modèles ANN, « Support Vector Machines (SVM) », « Linear Regression (LR) et « Second-order Polynomial Regression » sont les modèles comparés. Le modèle SVM a donné les estimations les plus précises lors de la phase d'évaluation des modèles. De même, dans (Yu & Cai, 2015), les auteurs ont comparé les modèles SVM et ANN pour l'estimation de la durée d'assemblage des avions. Cette durée est utilisée pour calculer le coût de la main d'œuvre directe. Le modèle SVM est ressorti ici aussi comme celui qui a fourni les meilleures estimations. Enfin, dans (Loyer et al., 2016), les auteurs ont comparé cinq modèles pour l'estimation du coût de fabrication des composants de moteurs d'avions. Pour cette comparaison, plusieurs critères ont été considérés, à savoir : la précision du modèle, sa facilité d'entrainement, son interprétabilité et son coût de calcul. Les modèles SVM et « Gradient Boosted Trees (GBT) » ont obtenu les meilleurs résultats en termes de précision. En termes de facilité d'entraînement et de l'interprétabilité, le modèle LR est apparu comme le meilleur. Tandis que le modèle ANN s'est révélée plus complexe à entrainer et non interprétable. Ses résultats étaient par ailleurs instables.

La proposition de prix pour compétitionner pour des commandes est tout aussi cruciale pour les équipementiers automobiles. Dans (Cavalieri et al., 2004), afin d'estimer le coût de fabrication des freins à disque, les auteurs ont comparé un modèle paramétrique avec un modèle ANN. Le modèle ANN a fourni les meilleures estimations sur les données de test. Cependant, les auteurs soulignent l'impossibilité d'interpréter le modèle. Dans (Kumru & Kumru, 2014), les auteurs ont utilisé un modèle ANN pour la prédiction du temps de fabrication de pièces détachées automobiles. Ce modèle a été comparé avec un modèle LR ainsi qu'un modèle « Polynomial Regression (PR) ». Malgré des données d'entrainement limitées, le modèle ANN a surclassé les deux autres modèles. Dans (Özcan & Fiğlali, 2014), les auteurs ont développé un modèle ANN pour l'estimation du coût d'estampage pour les éléments de carrosserie et de structure des voitures. Le modèle ANN proposé a donné des estimations plus précises que l'approche analytique traditionnellement utilisée par l'entreprise. Dans (Bodendorf & Franke, 2021), les auteurs ont cherché à trouver la meilleure méthode pour estimer le coût de fabrication des roues pour un équipementier automobile. Ils ont conclu que les modèles d'apprentissage automatique étaient plus précis que les méthodes analytiques traditionnelles d'estimation du coût. Dans cette étude, les auteurs ont entrainé plusieurs modèles, notamment les modèles

« K-Nearest Neighbor (KNN) », LR, SVM, ANN, « Decision Tree (DT)» et « AdaBoost ». Les auteurs ont conclu que tous les modèles d'apprentissage automatique estimaient le coût avec une grande précision sauf pour certains produits faiblement représentés dans les données d'entrainements. Enfin, dans (Bodendorf et al., 2021), un modèle « Deep Learning (DL) » a été construit pour estimer le coût de fabrication de circuit imprimé pour l'automobile. Ce modèle est capable de prédire le coût de fabrication d'un nouveau circuit à fabriquer en utilisant son image.

Dans la littérature, on retrouve également des articles portant sur l'injection plastique, la fabrication additive, la fabrication soustractive, la déformation plastique et l'assemblage par soudage. Dans (Florjanič et al., 2013), afin d'aider les experts dans la phase de proposition de prix aux clients, un modèle ANN est utilisé pour estimer le temps de fabrication de pièces moulées par injection plastique. Toujours dans la fabrication par déformation plastique, dans (Liu et al., 2018), un modèle ANN est utilisé pour estimer le coût d'une pièce sur la base de sa conception assistée par ordinateur (CAO). Un travail similaire concernant la fabrication additive a été reporté dans (Chan et al., 2018). Le coût d'un nouveau produit est estimé par des modèles de régression « Least Absolute Shrinkage and Selection Operator (LASSO) » et « Elastic-Net ». Dans (Sajadfar & Ma, 2015), une méthode hybride combinant un modèle non supervisé et des modèles supervisés est proposée pour l'estimation du coût des pièces à souder. Différents clusters sont formés à l'aide du modèle non supervisé. Pour chacun des clusters, des modèles de régression sont comparés à l'aide de l'outil « AutoML Weka ». Parmi les modèles comparés, un modèle ANN a fourni les prédictions les plus précises. Dans (Ning et al., 2020a), un modèle « Convolutional Neural Network (CNN) » est utilisé pour estimer les coûts de fabrication de pièces mécaniques à partir de leur image. Dans un autre article (Ning et al., 2020b), les mêmes auteurs comparent un modèle SVM et un modèle ANN pour estimer le coût de fabrication de pièces à partir d'images. Les estimations les plus précises ont été obtenues avec le modèle ANN. Dans le même sillage, dans (Yoo & Kang, 2021), un modèle CNN 3D est utilisé pour estimer le coût de fabrication de pièces mécaniques. Le modèle fournit aussi une visualisation des caractéristiques de la pièce qui augmentent le plus le coût de fabrication, offrant ainsi la possibilité d'optimiser la conception de la pièce à fabriquer.

Enfin, on retrouve des travaux portant sur d'autres secteurs industriels que l'aéronautique, l'automobile et la fabrication mécanique. Dans (Caputo & Pelagagge, 2008), un modèle ANN, un modèle paramétrique et des méthodes traditionnelles ont été comparés sur l'estimation du coût de fabrication d'équipements industriels pressurisés. Le modèle ANN est celui qui a réalisé les estimations les plus précises. Dans (Duran et al., 2012), deux types de modèle ANN (un perceptron multicouche et des réseaux de neurones à fonction de base radiale) ainsi qu'un modèle LR ont été comparés pour estimer le coût de fabrication des éléments de tuyauterie pour le transport de fluides. Les deux modèles de réseaux de neurones ont réalisé les estimations les plus précises. Dans (Serrat et al., 2013), une méthode basée sur un processus Gaussien est utilisée pour estimer le coût de fabrication de tuyaux à partir de leur CAO. Dans (Karaoglan & Karademir, 2017), un modèle ANN est utilisé pour estimer le temps de fabrication de transformateurs électriques en vue de définir un prix pour répondre à des appels d'offres. Dans (Leszczyński & Jasiński, 2020), un modèle ANN et un modèle paramétrique sont utilisés pour l'estimation du coût de fabrication de moteurs à induction. Les erreurs de prédiction du modèle ANN étaient nettement inférieures à celles du modèle paramétrique. Dans (Chou et al., 2010), plusieurs techniques sont utilisées pour estimer le coût de fabrication des écrans TFT-LCD. Celle impliquant un modèle ANN a été la plus performante.

En somme, l'analyse de la littérature a montré que les modèles de type Neural Network (ANN et CNN) sont les plus fréquemment utilisés pour l'estimation du coût de fabrication. Comparés à d'autres modèles, ils réalisent le plus souvent les estimations les plus précises, même si dans quelques cas ils ont été surclassés par des modèles SVM. Il faut noter que certains auteurs ont souligné la mise en œuvre complexe et le manque d'interprétabilité des modèles ANN et CNN. En termes de données exploitées pour la construction des modèles, nous avons remarqué que les spécifications techniques des produits (dimensions, matériau, géométrie, etc.) étaient les plus utilisées. Dans la majorité des travaux, uniquement des données tabulaires structurées sont exploitées. Quelques travaux exploitent des images et des dessins CAO. Par ailleurs, bien que les connaissances du domaine soient parfois exploitées dans la définition des attributs pertinents, elles sont rarement utilisées dans les autres étapes du processus de développement des modèles d'estimation. Or, pour de nombreux secteurs, les besoins des clients sont exprimés sous la forme de textes libres. Et comme mentionné dans l'introduction, l'exploitation des connaissances du domaine permettrait de tirer le meilleur profit des données. Par conséquent, l'objectif de notre article est de proposer une nouvelle approche de développement de modèles d'estimation de paramètres de fabrication. Cette approche est basée sur l'intégration de connaissances du domaine et de l'apprentissage automatique. Elle permet d'exploiter, en plus des données structurées, des données textuelles non structurées et les connaissances du domaine.

3 APPROCHE PROPOSEE

Comme mentionné dans les sections précédentes, l'approche proposée dans cet article est basée sur la méthode CRISP-DM. Simplement, comme montré dans la Figure 1, pour tirer le meilleur profit de l'ensemble des données disponibles, tabulaires structurées et textuelles non structurées, nous proposons la prise en compte des connaissances du domaine dans la réalisation des différentes tâches constituant la méthode CRISP-DM. Ces tâches et l'apport des connaissances du domaine à leur amélioration sont décrits dans la suite de cette section.

La première tâche **« Définition et compréhension du problème »** consiste à définir les objectifs du projet. Très souvent, l'ingénieur en charge du développement des modèles de prédiction n'est pas familier avec le domaine considéré. Par conséquent, il est important d'interagir avec des experts du domaine afin de recueillir les connaissances et les informations nécessaires à la définition des objectifs et des exigences du problème considéré. Une fois les données rassemblées, la deuxième tâche **« Compréhension des données »** consiste à étudier ces données, le plus souvent à l'aide d'une description statistique et de la visualisation de données, afin d'en avoir une vision claire. A ce niveau, même si les connaissances du domaine ne sont pas nécessairement implémentées dans un outil informatique pour traiter les données, elles peuvent aider à améliorer la définition les attributs pertinents à considérer et

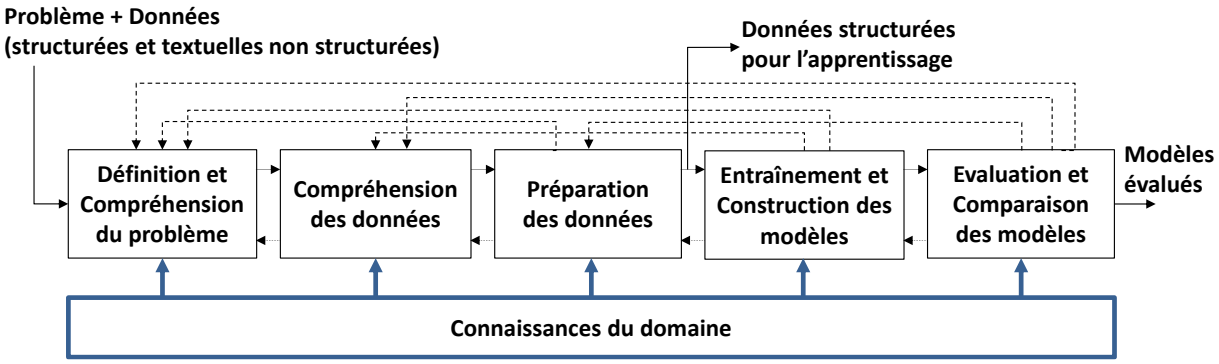

Figure 1 : Approche proposée

à supprimer ceux qui ne sont pas utiles. Les deux tâches suivantes **« Préparation des données »** et **« Entraînement et construction des modèles »** sont celles qui, le plus souvent, consomment énormément de temps. Il faut également noter que les contributions de cet article portent essentiellement sur ces deux tâches. Elles sont décrites dans les sections 3.1 et 3.2. La tâche **« Evaluation du modèle »** consiste à évaluer le modèle ou les modèles construits, et à les comparer entre eux ou à d'autres modèles. Les connaissances du domaine peuvent également servir à ce niveau pour définir les critères d'évaluation les plus pertinents et les plus appropriés pour le problème et le contexte de l'entreprise. La dernière tâche **« Déploiement ou intégration »** de la méthode CRISP-DM consiste à mettre l'outil d'estimation en exploitation. Cette tâche n'est pas considérée dans cet article. Il faut noter que, dans cette Figure 1, les flèches en traits continus représentent le sens direct d'exécution des tâches. Tandis que les flèches en traits discontinus représentent les retours d'information pouvant déclencher la réexécution des tâches afin d'améliorer le processus et les résultats.

3.1 Préparation des données

Comme le montre la Figure 2, nous considérons deux types de jeux de données pouvant être exploités pour la construction d'un modèle d'estimation. Ils sont décrits comme suit.

- Le premier type représente les données structurées non nettoyées. Ils contiennent, en plus des attributs et valeurs pertinents, des attributs et valeurs non pertinents pour l'apprentissage. Ces données peuvent également manquer d'attributs et valeurs pertinents pour l'apprentissage.
- Le deuxième type représente les données textuelles non structurées et non nettoyées. Sans doute le plus difficile à traiter et le plus important dans de nombreux projets. Ces données contiennent de nombreuses irrégularités, notamment des mots non pertinents, des mots manquants, des abréviations, différentes manières de désigner un même élément, différentes structurations des textes, des termes spécifiques à l'entreprise ou spécifiques à un expert de l'entreprise.

De ce fait, dans cette section, nous proposons une méthode permettant de construire un jeu de données pertinent pour l'apprentissage machine. Elle exploite les connaissances du domaine à toutes les étapes. Il est important de mentionner que la méthode proposée a été appliquée à un cas industriel (voir section 4). Cependant, il reste encore une phase d'évaluation avant de pouvoir la présenter de manière plus formalisée dans un article. C'est pourquoi, dans cet article, seulement les principes de la méthode sont décrits, sans fournir plus de détails sur les algorithmes développés. Ces principes sont structurés autour des quatre tâches suivantes.

3.1.1 Nettoyage des données

Les opérations appliquées dans cette tâche sont similaires pour les deux types de jeux de données. Il s'agit principalement de la suppression des attributs et valeur impertinents, la gestion des valeurs manquantes, la suppression des caractères qui entachent les valeurs pertinentes. La suppression des attributs impertinents est principalement basée sur les connaissances du domaine et la compréhension des besoins de l'entreprise. En ce qui concerne la gestion des valeurs manquantes, selon le contexte, les lignes contenant de nombreuses valeurs manquantes sont supprimées du jeu de données. Dans les autres cas, les valeurs manquantes sont complétées à l'aide de connaissances du domaine définies sous la forme de règles « IF …. THEN … ». La suppression des valeurs impertinentes est réalisée sur la base d'une analyse statistique complétée avec des connaissances du domaine définies sous la forme de règles « IF .... THEN ... ». Afin d'éliminer les caractères qui entachent les valeurs pertinentes, tout d'abord, à l'aide de connaissances du domaine, une liste de caractères « bruit » est définie pour chaque attribut ou groupe d'attributs. Ensuite, un algorithme pertinent est appliqué pour détecter et supprimer ces caractères « bruit ». Enfin, dans cette étape, pour certains attributs, il est parfois nécessaire de reformater les données en les convertissant dans la même unité et le même type.

3.1.2 Extraction de données à partir de textes

Comme le montre la Figure 2, cette tâche s'applique uniquement aux données textuelles non structurées. Le but étant de transformer l'ensemble de données textuelles en un ensemble de données tabulaires sous la forme d'attributs-valeurs, afin de récupérer les attributs et données pertinents non disponibles sous la forme de données structurées. Pour ce faire, sur la base des connaissances du domaine, une liste de « mots-clés » représentant des attributs pertinents pour le processus d'apprentissage est définie. Ensuite, en exploitant cette liste, des patterns identifiés dans les textes, et des connaissances du domaine, un algorithme est défini pour créer un nombre de colonnes correspondant au nombre de mots-clés et pour extraire des données utiles du texte afin de remplir les nouvelles colonnes définies.

3.1.3 Fusion de données

Cette tâche a pour but de rassembler dans un seul jeu de données des données structurées obtenues de différentes sources. Quelques opérations simples de fusion de données sont ainsi appliquées. La cohérence du nouveau jeu de données obtenu doit être vérifiée et les attributs redondants doivent être supprimées. Les connaissances du domaine servent à réaliser cette tâche de manière efficiente et fiable.

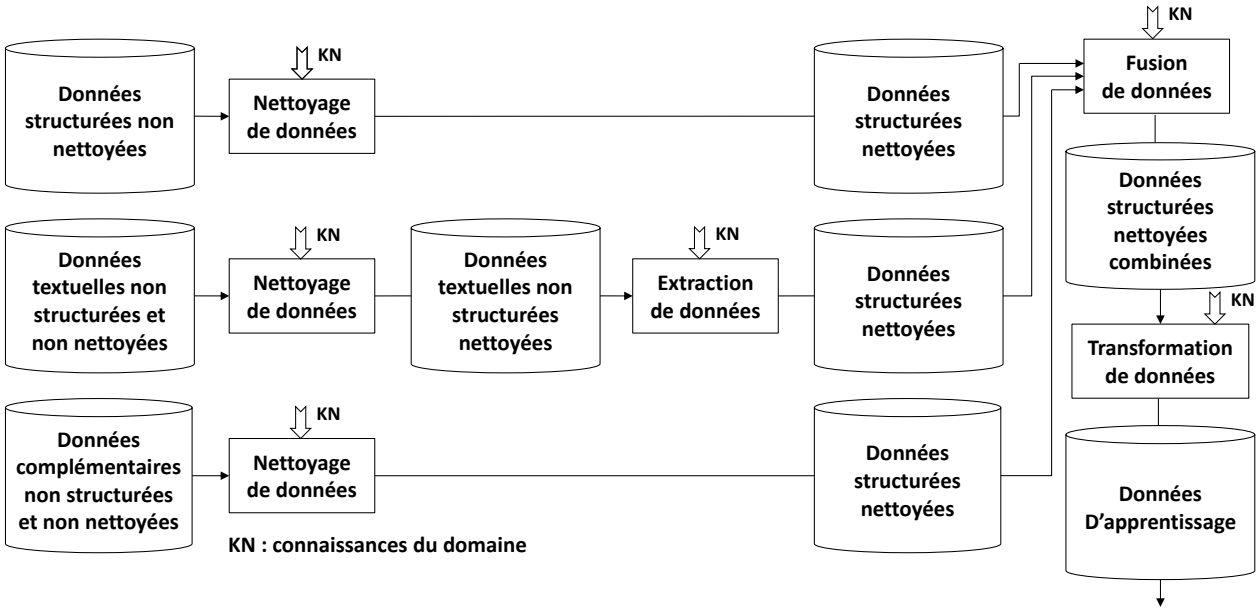

Figure 2 : Préparation des données

3.1.4 Transformation des données

Cette tâche est appliquée au jeu de données résultant de la fusion des différents jeux de données structurées provenant de différentes sources. Il arrive souvent qu'une agrégation d'attributs soit plus pertinente que les attributs pris individuellement. A cet effet, les connaissances du domaine sont utilisées pour transformer des données. En outre, de nombreux algorithmes d'apprentissage automatique fonctionnent mieux avec des valeurs numériques qu'avec des valeurs symboliques. Des techniques d'encodage peuvent être utilisées lorsque cela est nécessaire pour réaliser la transformation de valeurs symboliques en valeurs numériques. A la fin de cette étape, on dispose de données structurées dont les valeurs symboliques ont été encodées tandis que les valeurs numériques ont été normalisées. Nous pouvons donc passer à l'étape d'entrainement des modèles.

3.2 Entrainement et construction des modèles

Le domaine de l'apprentissage automatique offre une grande variété de modèles pour résoudre les problèmes de régression. Il faut rappeler que, dans ce domaine, un problème de régression consiste en la prédiction de valeurs numériques. Ainsi, pour construire un modèle de régression (ici un modèle d'estimation de temps de fabrication), il est nécessaire de tester et de comparer plusieurs modèles afin de trouver celui qui convient le mieux. Bien que la revue de la littérature puisse donner des idées sur les modèles qui fonctionnent le mieux pour un type de problème, la supériorité d'un modèle par rapport à un autre n'est pas asbsolue. En effet, un modèle peut être plus précis sur certains jeux de données mais moins précis sur d'autres, rendant la recherche du meilleur modèle une tâche fastidieuse. En outre, pour chaque modèle, il est nécessaire d'optimiser les hyperparamètres, ce qui peut également prendre beaucoup de temps. Pour accélérer ce processus, des solutions d'automatisation appelées « Automated Machine Learning (AutoML) » ont été proposées dans la littérature. Elles ont pour objectif d'aider à la construction rapide de modèles (He et al., 2021). Dans ce travail, nous avons utilisé trois solutions AutoML pour la construction des modèles d'estimation

4 EXPERIMENTATION SUR UN CAS INDUSTRIEL

Il est important de noter que l'expérimentation est présentée de sorte à ne pas divulguer certaines données de l'entreprise.

4.1 Présentation du cas d'étude

Cette étude porte sur l'estimation du coût de fabrication de nouveaux produits dans l'industrie de la métallurgie. L'entreprise considérée est une très petite entreprise qui fabrique sur commande selon les besoins des clients. Elle propose plusieurs familles de produits et chaque famille peut contenir d'innombrables produits en fonction des valeurs des différents paramètres de définition des produits tels que les dimensions géométriques, les matériaux utilisés, et les opérations de fabrication. Le temps de fabrication est le paramètre clé qui permet de déterminer le coût de fabrication. C'est-à-dire que suite à la demande d'un client et après avoir défini le processus de fabrication, un expert définit le temps de fabrication nécessaire. Ce temps est ensuite exploité dans une méthode de calcul simple pour déterminer le coût de fabrication. Ici, les modèles sont donc construits pour l'estimation de temps de fabrication.

Depuis plusieurs années, l'entreprise effectue plus de nombreuses estimations chaque année. La réalisation de ces estimations par un expert est chronophage et parfois source d'erreurs. Cela représente une charge cognitive importante pour l'expert. Aujourd'hui, l'objectif de l'entreprise est de réduire le temps de travail consacré à l'estimation du temps de fabrication et d'améliorer sa fiabilité. Ce qui leur permettrait de libérer du temps de travail pour des tâches à plus forte valeur ajoutée. Avec les années écoulées, une quantité importante de données portant sur les estimations déjà réalisées est disponible au sein de l'entreprise. Plus de 23000 estimations sont disponibles. Chaque estimation correspond à une demande d'un client. Il faut noter que pour chaque estimation, une petite partie des données décrivant les spécifications sont sous forme tabulaire structurée. Tandis que l'autre partie, la plus importante contenant les informations les plus utiles sur les besoins et le processus de fabrication, est disponible sous forme textuelle non structurée. Les deux types de données

complémentaires sont exploités à travers la méthode proposée dans cet article dans le but de développer un modèle d'estimation du temps de fabrication. Les résultats obtenus des premières expérimentations sont présentés dans la suite.

4.2 Préparation des données

Une étape importante de cette étude a consisté à extraire du jeu de données textuelles des données utiles et à les structurer. En effet, afin de favoriser leur compréhension par les clients, ces textes sont rédigés par l'expert selon le client considéré. Ce qui génère une panoplie de termes et de structures de textes. Par exemple, on peut identifier près dix manières différentes de définir une caractéristique d'un produit. Chaque produit pouvant être défini par une dizaine de caractéristiques. Les séquences des caractéristiques dans le texte varient et les mots utilisés pour faire les liens varient également. Pour deux familles de produits différentes, les différences sont encore plus importantes. Par ailleurs, outre les informations explicitement mentionnées dans le texte, il existe des informations qui doivent être déduites en fonction des besoins du client et des règles commerciales. Ainsi, l'absence de certaines informations dans certaines lignes ne doit pas être considérée comme des valeurs manquantes.

La méthode proposée à la section 3.1.2, « Extraction d'attributs pertinents à partir du texte », est appliquée à ces données textuelles pour en extraire des données utiles et de les structurer sous forme tabulaire. Tout d'abord, à l'aide des connaissances du domaine, une syntaxe normalisée a été définie pour chaque caractéristique ou attribut. Ensuite, tous les termes utilisés dans les textes pour désigner un même attribut normalisé sont remplacés par cet attribut normalisé. Ces attributs normalisés sont ensuite utilisés comme mot-clé pour l'extraction des données et leur structuration en colonnes. Chaque mot-clé étant un attribut, il forme une colonne dans le futur jeu de données en construction. Sans rentrer dans les détails et les particularités qui peuvent être gérés autrement dans l'algorithme de structuration des textes, pour un texte donné, la séquence de termes qui se trouve entre deux mots-clés (donc deux attributs) est placée dans la colonne de l'attribut précédent la séquence, immédiatement à gauche. Cette séquence contient très généralement la valeur de l'attribut. Mais elle contient aussi des « bruits » tels que des termes utilisés pour faire les liaisons mais aussi d'autres caractères impertinents. Parfois, la valeur de l'attribut est manquante, parfois mal écrite. Les bruits sont supprimés à l'aide d'un algorithme qui utilise une liste « stop words » ou en se basant sur le format standard des valeurs de l'attribut concerné. Il faut noter que cette opération de nettoyage des données est réalisée de la même manière pour les jeux de données structurées. La gestion de valeurs manquantes ou de valeurs mal écrites est également réalisée de la même manière, essentiellement à l'aide de connaissances du domaine sous la forme de règles « IF … THEN … ». Près de 23000 lignes d'estimation ont été structurées et nettoyées. L'extraction d'information à partir de textes a permis de compléter au minimum 10 attributs pertinents pour chaque ligne. Un algorithme de calcul de la précision et de l'efficacité de la structuration n'a pas encore été implémenté ni testé. Cependant, les premiers retours d'un expert du domaine montrent que les données sont correctement structurées.

A la suite des opérations d'extraction d'attributs pertinents à partir de données textuelles et de nettoyage de tous les jeux de données, un assemblage des différents jeux de données est réalisé à l'aide d'opérations simples de fusion de données. Une fois toutes les données rassemblées, il a fallu gérer les lignes présentant des valeurs manquantes et des lignes inexploitables à cause d'un certain nombre de facteurs tels que l'incohérence entre la spécification et la valeur estimée du temps de fabrication. Des règles simples basées sur les connaissances du domaine ont permis de corriger un certain nombre d'incohérences et de compléter certaines données. D'autres règles également fondées sur les connaissances du domaine ont permis de détecter les lignes inexploitables soit pour incohérence soit pour manque d'estimation du temps de fabrication ou de valeur pour certains attributs.

Enfin, il a été nécessaire de transformer le format initial de certains attributs. Par exemple, un attribut pour lequel un certain écart entre ses valeurs n'a pas d'influence sur le temps de fabrication a été redéfini. Des valeurs jugées similaires par l'expert ont été regroupées en catégorie et ces catégories constituent le nouveau domaine de l'attribut dans le jeu de données destiné à l'apprentissage. Un autre exemple concerne une caractéristique qui est défini à l'aide de plusieurs attributs différents selon les clients. Il y a donc des lignes pour lesquelles un attribut X est rempli et les autres attributs pas remplis. Grâce aux connaissances du domaine un seul attribut parmi tous les attributs possibles a été retenu et une règle a été appliquée pour transformer les valeurs des autres attributs afin de compléter l'attribut retenu pour toutes les lignes du jeu de données. Nous avons également l'exemple d'un attribut jugé insuffisant pour représenter une caractéristique importante ayant un impact sur le temps de fabrication. A l'aide de connaissances expertes, cet attribut a été transformé pour créer un autre attribut qui l'a remplacé. Enfin, deux nouveaux attributs ont été créés pour représenter deux caractéristiques importantes qui n'étaient pas explicitement présentes dans le jeu de données. Ils ont tous été créés en appliquant une formule fournie par un expert.

Le jeu de données obtenu à la fin des opérations de préparation de données est constitué de quatorze attributs. Ils ont tous été validés par un expert. Trois types de données existent : symboliques, booléennes et numériques continues. Chaque type nécessite une préparation avant d'alimenter les modèles d'apprentissage automatique. Le module de prétraitement de « Scikit-Learn » a été utilisé à cet effet.

4.3 Entrainement et évaluation des modèles

Comme mentionné dans section 1, l'entreprise propose plusieurs familles de produits et chaque famille contient différentes variantes de produit qui ont des attributs en commun mais également des attributs spécifiques qui leur sont propres. La question majeure dans une telle situation est de trouver le bon niveau auquel un modèle d'estimation sera plus performant, sachant que la priorité de l'entreprise est la précision du modèle. Nous avons donc décidé d'expérimenter trois stratégies et de les comparer. Ces stratégies sont : (i) un modèle par variante de produit, (ii) un modèle par famille de produits, (iii) un modèle pour l'ensemble des familles de produits. Pour le reste de cet article, seulement la première stratégie est considérée. Un modèle est construit pour chaque variante de produit. Ce qui implique que, pour chaque variante, dans le jeu de données exploité pour l'apprentissage, toutes les lignes sont définies par le même ensemble d'attributs. Les lignes se distinguent uniquement par les valeurs des attributs.

Il faut cependant noter que les résultats présentés dans la section suivante portent sur une seule variante de produit. Pour cette variante, 1036 lignes (estimations réalisées par le passé) ont été exploitées. Les données sont reparties en données d'entrainement (75%) et en données de test (25%), grâce à un module dédié de « Scikit-Learn ». La division a été faite avant l'encodage et la normalisation des données pour éviter les fuites d'information.

Pour la sélection des modèles et le réglage des hyperparamètres, nous avons utilisé trois solutions AutoML qui sont H2O.ai (Ledell & Poirier, 2020), TPOT (Olson et al., 2016), et mljar-supervisé (Płońska & Płoński, 2021). Pour les trois solutions AutoML, la procédure de validation croisée (avec k =10) a été choisie pour l'évaluation des modèles sur les données d'entrainement. Les meilleurs modèles fournis par chaque solution sont ensuite évalués sur les données test. Le tableau 1 montre les performances de chaque modèle suivant trois métriques les plus utilisées dans la littérature, à savoir RMSE, MAE et MAPE.

Tableau 1. Evaluation des modèles

	TPOT	H2O	Mljar-supervised
RMSE	4,675	4,909	4,445
MAE	2,388	2,401	2,049
MAPE (%)	31,84	33,61	25,93
Meilleure modèle	Gradient Boosting Machine	Extra-trees pipeline	Ensemble model

Nous constatons que les trois solutions ont des performances assez proches sur les données test, avec cependant un léger avantage pour « mljar-supervised ». Le meilleur modèle fourni par ce dernier est la combinaison de modèles NN, CatBoost et XGBoost. La prédiction du modèle « Ensemble model » est la moyenne pondérée des modèles de base.

5 Conclusion

Dans cet article nous avons proposé une nouvelle approche pour l'estimation des paramètres de fabrication en exploitant à la fois des données structurées et des données textuelles non structurées. Les connaissances du domaine sont intégrées aux différentes étapes de l'approche afin de tirer le meilleur profit des données. Une expérimentation a été réalisée sur un cas industriel pour l'estimation des temps de fabrication en vue du calcul des coûts des produits et de la définition des prix des offres à transmettre aux clients. Il faut noter que les premiers résultats sont prometteurs, notamment l'extraction des données utiles à partir des textes et leur structuration. Cependant, des travaux supplémentaires sont en cours, afin d'améliorer et de mieux formaliser les propositions.

6 Remerciements

This work has been partially supported by the MIAI Multidisci- plinary AI Institute at the Univ. Grenoble Alpes: (MIAI@Grenoble Alpes - ANR-19-P3IA-0003

7 References

Aldanondo, M., & Vareilles, E. (2008). Configuration for mass customization: How to extend product configuration towards requirements and process configuration. *Journal of Intelligent Manufacturing*, *19*(5), 521–535. https://doi.org/10.1007/s10845-008-0135-z

Ali, M., Ali, R., Khan, W. A., Han, S. C., Bang, J., Hur, T., Kim, D., Lee, S., & Kang, B. H. (2018). A Data-Driven Knowledge Acquisition System: An End-to-End Knowledge Engineering Process for Generating Production Rules. *IEEE Access*, *6*, 15587–15607. https://doi.org/10.1109/ACCESS.2018.2817022

Bodendorf, F., & Franke, J. (2021). A machine learning approach to estimate product costs in the early product design phase: A use case from the automotive industry. *Procedia CIRP*, *100*, 643–648. https://doi.org/10.1016/j.procir.2021.05.137

Bodendorf, F., Merbele, S., & Franke, J. (2021). Deep learning based cost estimation of circuit boards: a case study in the automotive industry. *International Journal of Production Research*, 1–22. https://doi.org/10.1080/00207543.2021.1998698

Caputo, A. C., & Pelagagge, P. M. (2008). Parametric and neural methods for cost estimation of process vessels. *International Journal of Production Economics*, *112*(2), 934–954. https://doi.org/10.1016/j.ijpe.2007.08.002

Cavalieri, S., Maccarrone, P., & Pinto, R. (2004). Parametric vs. neural network models for the estimation of production costs: A case study in the automotive industry. *International Journal of Production Economics*, *91*(2), 165–177. https://doi.org/10.1016/j.ijpe.2003.08.005

Chan, S. L., Lu, Y., & Wang, Y. (2018). Data-driven cost estimation for additive manufacturing in cybermanufacturing. *Journal of Manufacturing Systems*, *46*, 115–126. https://doi.org/10.1016/j.jmsy.2017.12.001

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). Step-by-step data mining guide. In *SPSS inc* (Vol. 78). DaimlerChrysler.

Chou, J.-S., Tai, Y., & Chang, L.-J. (2010). Predicting the development cost of TFT-LCD manufacturing equipment with artificial intelligence models. *International Journal of Production Economics*, *128*(1), 339–350. https://doi.org/10.1016/j.ijpe.2010.07.031

de Cos, J., Sanchez, F., Ortega, F., & Montequin, V. (2008). Rapid cost estimation of metallic components for the aerospace industry. *International Journal of Production Economics*, *112*(1), 470–482. https://doi.org/10.1016/j.ijpe.2007.05.016

Deng, S., & Yeh, T.-H. (2011). Using least squares support vector machines for the airframe structures manufacturing cost estimation. *International Journal of Production Economics*, *131*(2), 701–708. https://doi.org/10.1016/j.ijpe.2011.02.019

Duran, O., Maciel, J., & Rodriguez, N. (2012). Comparisons between two types of neural networks for manufacturing cost estimation of piping elements. *Expert Systems with Applications*, *39*(9), 7788–7795. https://doi.org/10.1016/j.eswa.2012.01.095

Florjanič, B., Govekar, E., & Kuzman, K. (2013). Neural Network-Based Model for Supporting the Expert Driven Project Estimation Process in Mold Manufacturing. *Strojniški Vestnik – Journal of Mechanical Engineering*, *59*(01), 3–13. https://doi.org/10.5545/sv-jme.2012.747

García-Crespo, Á., Ruiz-Mezcua, B., Luis López-Cuadrado, J.,

González-Carrasco, I., García-Crespo, Á., Ruiz-Mezcua, · B, López-Cuadrado, J. L., González-Carrasco, I., & Ruiz-Mezcua, B. (2009). A review of conventional and knowledge based systems for machining price quotation. *Journal of Intelligent Manufacturing 2009 22:6*, *22*(6), 823–841. https://doi.org/10.1007/S10845-009-0335-1

He, X., Zhao, K., & Chu, X. (2021). AutoML: A survey of the state-of-the-art. *Knowledge-Based Systems*, *212*, 106622. https://doi.org/10.1016/j.knosys.2020.106622

Hicks, C., McGovern, T., & Earl, C. F. (2000). Supply chain management: A strategic issue in engineer to order manufacturing. *International Journal of Production Economics*, *65*(2), 179–190. https://doi.org/10.1016/S0925-5273(99)00026-2

Ioannou, G., & Dimitriou, S. (2012). Lead time estimation in MRP/ERP for make-to-order manufacturing systems. *International Journal of Production Economics*, *139*(2), 551–563. https://doi.org/10.1016/j.ijpe.2012.05.029

Karaoglan, A. D., & Karademir, O. (2017). Flow time and product cost estimation by using an artificial neural network (ANN): A case study for transformer orders. *Engineering Economist*, *62*(3), 272–292. https://doi.org/10.1080/0013791X.2016.1185808

Kingsman, B. G., & Mercer, A. (1997). Strike Rate Matrices for Integrating Marketing and Production During the Tendering Process in Make-to-Order Subcontractors. *International Transactions in Operational Research*, *4*(4), 251–257. https://doi.org/10.1111/J.1475-3995.1997.TB00081.X

Kingsman, B., Hendry, L., Mercer, A., & De Souza, A. (1996). Responding to customer enquiries in make-to-order companies Problems and solutions. *International Journal of Production Economics*, *46–47*, 219–231. https://doi.org/10.1016/0925-5273(95)00199-9

Kumru, M., & Kumru, P. Y. (2014). Using artificial neural networks to forecast operation times in metal industry. *International Journal of Computer Integrated Manufacturing*, *27*(1), 48–59. https://doi.org/10.1080/0951192X.2013.800231

Ledell, E., & Poirier, S. (2020). H2O AutoML: Scalable automatic machine learning. *7th ICML Workshop on Automated Machine Learning, 2020*, 1–16.

Leszczyński, Z., & Jasiński, T. (2020). Comparison of Product Life Cycle Cost Estimating Models Based on Neural Networks and Parametric Techniques—A Case Study for Induction Motors. *Sustainability*, *12*(20), 8353. https://doi.org/10.3390/su12208353

Liu, W., Yang, C., & Zhou, X. (2018). A network quotation framework for customised parts through rough requests. *International Journal of Computer Integrated Manufacturing*, *31*(12), 1220–1234. https://doi.org/10.1080/0951192X.2018.1529429

Loyer, J. L., Henriques, E., Fontul, M., & Wiseall, S. (2016). Comparison of Machine Learning methods applied to the estimation of manufacturing cost of jet engine components. *International Journal of Production Economics*, *178*, 109–119. https://doi.org/10.1016/j.ijpe.2016.05.006

Murphy, R., Newell, A., Hargaden, V., & Papakostas, N. (2019). Machine learning technologies for order flowtime estimation in manufacturing systems. *Procedia CIRP*, *81*, 701–706. https://doi.org/10.1016/J.PROCIR.2019.03.179

Ning, F., Shi, Y., Cai, M., Xu, W., & Zhang, X. (2020a). Manufacturing cost estimation based on a deep-learning method. *Journal of Manufacturing Systems*, *54*, 186–195. https://doi.org/10.1016/j.jmsy.2019.12.005

Ning, F., Shi, Y., Cai, M., Xu, W., & Zhang, X. (2020b). Manufacturing cost estimation based on the machining process and deep-learning method. *Journal of Manufacturing Systems*, *56*, 11–22. https://doi.org/10.1016/j.jmsy.2020.04.011

Olson, R. S., Bartley, N., Urbanowicz, R. J., & Moore, J. H. (2016). Evaluation of a tree-based pipeline optimization tool for automating data science. *GECCO 2016 - Proceedings of the 2016 Genetic and Evolutionary Computation Conference*, 485–492. https://doi.org/10.1145/2908812.2908918

Özcan, B., & Fiğlali, A. (2014). Artificial neural networks for the cost estimation of stamping dies. *Neural Computing and Applications 2014 25:3*, *25*(3), 717–726. https://doi.org/10.1007/S00521-014-1546-8

Pallant, J., Sands, S., & Karpen, I. (2020). Product customization: A profile of consumer demand. *Journal of Retailing and Consumer Services*, *54*, 102030. https://doi.org/10.1016/J.JRETCONSER.2019.102030

Płońska, A., & Płoński, P. (2021). *MLJAR: State-of-the-art Automated Machine Learning Framework for Tabular Data*. MLJAR.

Plotnikova, V., Dumas, M., & Milani, F. (2020). Adaptations of data mining methodologies: A systematic literature review. *PeerJ Computer Science*, *6*, 1–43. https://doi.org/10.7717/PEERJ-CS.267

Sajadfar, N., & Ma, Y. (2015). A hybrid cost estimation framework based on feature-oriented data mining approach. *Advanced Engineering Informatics*, *29*(3), 633–647. https://doi.org/10.1016/j.aei.2015.06.001

Serrat, J., Lumbreras, F., & López, A. M. (2013). Cost estimation of custom hoses from STL files and CAD drawings. *Computers in Industry*, *64*(3), 299–309. https://doi.org/10.1016/j.compind.2012.11.009

Sylla, A., Coudert, T., Vareilles, E., Geneste, L., & Aldanondo, M. (2021). Possibilistic Pareto-dominance approach to support technical bid selection under imprecision and uncertainty in engineer-to-order bidding process. *International Journal of Production Research*, *59*(21), 6361–6381. https://doi.org/10.1080/00207543.2020.1812754

Verlinden, B., Duflou, J. R., Collin, P., & Cattrysse, D. (2008). Cost estimation for sheet metal parts using multiple regression and artificial neural networks: A case study. *International Journal of Production Economics*, *111*(2), 484–492. https://doi.org/10.1016/J.IJPE.2007.02.004

Wang, G., Gunasekaran, A., Ngai, E. W. T., & Papadopoulos, T. (2016). Big data analytics in logistics and supply chain management : Certain investigations for research and applications. *International Journal of Production Economics*, *176*, 98–110. https://doi.org/10.1016/j.ijpe.2016.03.014

Yoo, S., & Kang, N. (2021). Explainable artificial intelligence for manufacturing cost estimation and machining feature visualization. *Expert Systems with Applications*, *183*, 115430. https://doi.org/10.1016/j.eswa.2021.115430

Yu, T., & Cai, H. (2015). The Prediction of the Man-Hour in Aircraft Assembly Based on Support Vector Machine Particle Swarm Optimization. *Journal of Aerospace Technology and Management*, *7*(1), 19–30. https://doi.org/10.5028/JATM.V7I1.409