UQAC
Université du Québec
à Chicoutimi

DÉVELOPPEMENT D'UN MODÈLE POUR LE CHOIX D'UNE PLATEFORME DE MACHINE LEARNING POUR L'APPRENTISSAGE SUPERVISÉ DANS LES PETITES ENTREPRISES

PAR VINCENT GAGNON

MÉMOIRE PRÉSENTÉ À L'UNIVERSITÉ DU QUÉBEC À CHICOUTIMI EN VUE DE L'OBTENTION DU GRADE DE MAÎTRE ÈS SCIENCES (M. SC.) EN INFORMATIQUE

QUÉBEC, CANADA

RÉSUMÉ

La croissance rapide des technologies d'intelligence artificielle (IA) et d'apprentissage automatique (ML) offre des occasions significatives pour les entreprises de toutes tailles. Cependant, les petites entreprises rencontrent souvent des défis uniques lors de l'adoption de ces technologies en raison de ressources et d'expertise limitées. De plus, la multitude d'outils et de plateformes disponibles pour l'implémentation de l'intelligence artificielle complique la tâche lorsque celle-ci doit faire un choix. Cette étude vise à développer un modèle de décision pour aider les petites entreprises à sélectionner la plateforme d'apprentissage automatique (MLaaS) la plus adaptée pour l'apprentissage supervisé. Le modèle proposé est conçu pour être pratique et facile à utiliser, en tenant compte des besoins et des contraintes spécifiques des petites entreprises.

Nous commençons par examiner la littérature existante et identifier les critères clés qui influencent la sélection des plateformes MLaaS. Ces critères incluent la réputation du fournisseur de services, les coûts d'abonnement périodiques, la disponibilité des services et la portabilité de la plateforme. En utilisant le processus hiérarchique analytique (AHP) et sa version simplifiée, AHP-express, nous développons un cadre de prise de décision structuré que les petites entreprises peuvent utiliser pour évaluer et comparer efficacement différentes plateformes MLaaS.

Le modèle est appliqué dans une étude de cas impliquant Econochef, une petite entreprise en démarrage développant une application mobile de recommandation de recettes. En mettant en œuvre le modèle de décision, Econochef sélectionne avec succès une plateforme d'apprentissage automatique répondant à ses besoins et contraintes financières, démontrant l'efficacité du modèle.

Dans la discussion, les avantages et les limitations du modèle proposé sont soulignés et des idées pour de futures recherches sont proposées. Celles-ci incluent l'amélioration de la collecte et de la qualité des données, l'automatisation du processus de prise de décision, la validation du modèle à travers d'autres études de cas et l'assurance de la scalabilité et de l'adaptabilité du modèle aux nouvelles technologies.

Cette étude fournit un outil précieux pour les petites entreprises cherchant à tirer parti des technologies d'apprentissage automatique, offrant une approche claire et structurée pour sélectionner des plateformes MLaaS appropriées et facilitant l'intégration de l'IA dans leurs opérations.

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

LISTE DES ABRÉVIATIONS

IA	Intelligence Artificielle
ML	Machine Learning
PME	Petite et moyenne entreprise
MLaaS	Machine Learning as a Service

REMERCIEMENTS

Je tiens à exprimer ma gratitude envers toutes les personnes qui ont contribué à la réalisation de ce mémoire. Tout d'abord, je remercie Fabio Petrillo et Sylvain Hallé, de m'avoir initié à la recherche et pour leur soutien constant, leurs conseils éclairés et leur patience tout au long de ce projet. Leur rigueur et leur expertise ont été essentielles à l'avancement de mon travail.

Un grand merci à ma famille, en particulier à mes parents, pour leur soutien tout au long de mes études. Leur encouragement m'a permis d'affronter les défis avec sérénité et de me dépasser dans mes recherches.

Un immense merci à ma conjointe Jessica pour son amour, sa compréhension et son soutien inconditionnel tout au long de ce parcours. Sa présence m'a donné la force de persévérer, et son soutien m'a permis de garder le cap, même dans les moments les plus exigeants.

Enfin, je tiens à remercier tous ceux qui, de près ou de loin, ont contribué à l'accomplissement de ce travail.

CHAPITRE I
INTRODUCTION

1.1 CONTEXTE

L'intelligence artificielle (IA) gagne de plus en plus d'attention dans le monde actuel. En effet, on peut constater un "boom" dans le secteur de l'IA, notamment en raison de l'accès facile aux outils comme ChatGPT [2], Midjourney [3] ou DALL·E [4]. On voit un passage du monde académique au monde appliqué. Selon un sondage réalisé par McKinsey [5], l'apprentissage automatique est adopté de manière grandissante avec près de 25% de croissance annuelle. Les outils permettent au grand public d'expérimenter avec l'IA et débloquent de nouvelles possibilités aux entreprises qui font la course au développement de nouveaux produits, soit en intégrant les outils mentionnés précédemment ou en créant leur propre version. Cependant, ces produits, ou nouveaux outils, sont majoritairement développés par de très grandes entreprises comme Microsoft [6], Amazon [7] ou Google [8] en raison de la complexité et aux ressources nécessaires afin de développer de nouveaux modèles d'apprentissage automatique (ML).

On retrouve également de plus en plus d'outils et de services d'apprentissage automatique automatisés (AutoML) ou «low-code» qui permettent aux développeurs avec peu d'expérience avec le ML de développer des modèles avec des performances acceptables [9, 10]. Ces outils permettent d'automatiser la majorité des étapes que l'on peut retrouver dans ce que l'on appel un pipeline d'apprentissage automatique. Celui-ci regroupe toutes les étapes nécessaires afin d'implémenter l'apprentissage automatique dans un système de façon robuste et fiable.

La majorité de ces outils existants sont conçus pour être utilisés par de grandes entreprises disposant de multiples équipes de programmeurs spécialisés dans divers domaines.

Cela pose un défi particulier pour les petites et moyennes entreprises (PME) au Canada, qui manquent souvent des ressources et de l'expertise nécessaires pour tirer pleinement parti de ces technologies avancées. Or, en décembre 2021, les petites entreprises représentaient 97.9% de toutes les entreprises et employaient 67.7% de tous les employés [11]. On peut tout de même noter l'apparition de quelques propositions open-source pour les petites entreprises [12, 13].

Dans ce contexte il peut être difficile de faire un choix éclairé de la plateforme à utiliser considérant le nombre grandissant de celles-ci. On peut donc se poser les questions : est-il possible pour les petites entreprises de tirer avantage de cette explosion dans le monde de l'IA afin d'intégrer le ML dans leurs activités ? Quel est le niveau de connaissance minimal que les employés d'une petite entreprise doivent avoir afin d'implémenter un système d'apprentissage automatique dans leur entreprise ? Comment une petite entreprise peut-elle sélectionner la bonne plateforme selon ses caractéristiques ? On suppose que l'implémentation du ML dans ces petites entreprises pourrait leur amener un grand avantage compétitif et accélérer le développement de ces entreprises.

1.2 PROBLÉMATIQUE

Les plus grandes difficultés lors de l'adoption de l'IA en entreprise selon un sondage réalisé en 2022 par O'Reilly [11] sont le manque d'employés compétents dans le domaine de l'IA et le manque de données sur lesquels entraîner les modèles. Ces défis sont suivis de difficultés à cerner des cas d'utilisation appropriés, des défis techniques dans l'infrastructure, de l'entreprise qui ne voit pas le besoin d'implémenter l'IA, des préoccupations quant à l'aspect légal et finalement des difficultés à définir les réglages des hyperparamètres des modèles. De multiples défis lors de l'implémentation et du déploiement de systèmes ML sont également identifiés dans six sondages [14, 15, 16, 17, 18, 19]. Pour les défis reliés aux

données, on retrouve le manque de données étiquetées, les ensembles de données déséquilibrés, la complexité des données, le manque de diversification et de cas limites dans les données, des difficultés dans le nettoyage de données, la visualisation des données et la gestion des données ainsi que le versionnage de celle-ci. Des défis concernant la création des modèles sont également mentionnés, on retrouve la difficulté à avoir une bonne interprétabilité des modèles, des difficultés à construire des pipelines évolutifs, difficulté à construire des modèles plus complexes et la gestion du versionnage des modèles. Du côté de l'entraînement et de l'évaluation des modèles, ces sondages mentionnent également une dizaine de défis comme le manque de données permettant de vérifier la performance des modèles créés, la difficulté de reproduction des modèles et des résultats, la difficulté à déboguer les modèles, le coût en calcul de l'entraînement et la difficulté à trouver les bonnes métriques pour tester la performance des modèles. Finalement, on relève aussi des défis pour l'étape du déploiement, ceux-ci sont la déviation des performances entre les résultats du modèle à l'entraînement et lors du déploiement, la déviation des données à travers le temps, le phénomène de boucle de rétroaction [20], la détection de valeurs aberrantes et la mise à jour et le réentraînement des modèles automatiquement. La gestion de produits liés à l'apprentissage automatique s'étend sur plusieurs facteurs et thèmes qui ne sont pas tous étudiés [21].

Un sondage réalisé par McKinsey en décembre 2022 [22] démontre que l'adoption de l'intelligence artificielle a atteint un plateau en 2018 et que les entreprises qui sont déjà investies dans l'IA investissent davantage tandis qu'il est difficile pour les entreprises voulant se lancer dans ce domaine de faire l'implémentation de tels systèmes complexes. Cela pourrait être expliqué par le fait que les petites entreprises font face à des défis que les plus grandes entreprises n'ont pas. Par exemple, un nombre d'employés grandement inférieur et beaucoup moins spécialisés et compétents dans le domaine du ML, une moins grande quantité de données disponibles, ainsi que le manque de ressources, que ce soit du temps ou de l'argent

[23, 24]. Ces défis pourraient être mitigés si ces entreprises avaient accès à une marche à suivre simple et bien définie pour l'implémentation d'un système d'apprentissage automatique dans celles-ci.

Une multitude de grandes entreprises offrent des services infonuagiques, comme [7, 6, 8], qui sont parfois adaptées aux plus petites entreprises. Cependant, il est difficile pour celles-ci de déterminer si le coût de ces services en vaut la peine étant donné les moyens limités de ces entreprises. On doit aussi prendre en compte que l'utilisation de services «cloud» d'un fournisseur peut verrouiller l'entreprise à ce service. On peut suivre le même raisonnement pour les outils «open-source» mis en ligne sur place : est-ce que l'investissement en vaut la peine ?

D'ailleurs, des articles scientifiques mentionnent le manque d'études de cas et de recommandations de meilleures pratiques pour l'adoption de produits spécialisés en apprentissage automatique [21], surtout concentrés sur les PME [23].

1.3 OBJECTIF

Cette étude vise à proposer une marche à suivre pratique pour la sélection d'un service d'apprentissage automatique dans le cloud pour les petites entreprises qui envisagent d'implémenter cette fonctionnalité dans leur système.

Pour réaliser cet objectif, les plateformes qui offrent des services d'apprentissage automatique seront étudiées et des caractéristiques qui définissent ces plateformes seront établies. Un modèle de sélection de plateformes d'apprentissage automatique sera proposé selon les caractéristiques de l'entreprise qui veut faire un choix de service.

Ce modèle se veut un outil pratique et facile à utiliser pour les développeurs ayant peu d'expérience avec l'intelligence artificielle. La marche à suivre devra être claire et facile à comprendre et devra donner des résultats fiables que les entreprises pourront utiliser pour guider leur choix de plateforme. Les caractéristiques qui seront mesurées pour faire le choix devront être bien définies et faciles à tester.

L'approche proposée sera ensuite appliquée dans un cas réel d'une petite entreprise développant une application mobile afin de valider l'hypothèse. L'hypothèse est la suivante : le modèle de décision de plateforme ML proposé permet aux petites entreprises de faire un choix éclairé et de déployer un pipeline ML robuste et fiable rapidement et efficacement.

1.4 STRUCTURE

Ce mémoire aura la forme suivante : le chapitre 2 portera sur l'état de l'art dans le domaine, plus spécifiquement les méthodes actuelles et les travaux connexes à ce mémoire, le chapitre 3 proposera une approche qui permettra de valider l'hypothèse posée, dans le chapitre 4 nous appliquerons l'approche proposée dans un cas d'utilisation réel et finalement le chapitre 5 apportera des conclusions.

CHAPITRE II

ÉTAT DE L'ART

Cette section définit les termes pertinents à ce mémoire ainsi que les méthodes utilisées actuellement en pratique et les travaux connexes à cette étude.

2.1 DÉFINITIONS

2.1.1 INTELLIGENCE ARTIFICIELLE ET APPRENTISSAGE AUTOMATIQUE

L'intelligence artificielle est un terme général qui peut être défini comme la science et l'ingénierie qui permet aux machines de devenir intelligentes, plus particulièrement les logiciels [25]. Par «intelligence» on parle de l'aptitude à percevoir, synthétiser et inférer de l'information. [26]

L'apprentissage automatique est une branche de l'intelligence artificielle qui se concentre sur l'utilisation de données et d'algorithmes pour imiter la façon dont les humains apprennent, c'est-à-dire en améliorant la précision de ses prédictions graduellement à mesure de voir des nouveaux cas dans les données [27, 28]. Ce type d'IA utilise de larges ensembles de données afin de reconnaître des tendances dans ces données, ce qui permet à la machine de prendre des décisions ou de donner des recommandations aux utilisateurs par exemple.

Plusieurs concepts clés doivent être compris afin de bien saisir le fonctionnement de l'apprentissage automatique. Ces concepts sont bien décrits dans le livre de Burkov [29]. Premièrement, on retrouve les étiquettes. L'étiquette est l'information que l'on souhaite prédire à partir des données. Dans un contexte supervisé, chaque entrée dans l'ensemble de données est associée à une étiquette qui représente le résultat attendu. Par exemple, dans un

ensemble de données de reconnaissance d'images, l'étiquette pourrait indiquer la classe de l'objet présent dans chaque image (chat, chien, voiture, etc.).

Ensuite, les caractéristiques sont les variables ou les attributs mesurés dans les données d'entrée qui servent à construire le modèle. Ce sont les informations utilisées par l'algorithme pour prendre des décisions. Par exemple, dans une application de classification de fleurs, les caractéristiques pourraient inclure la longueur et la couleur des pétales. Il est également possible de déduire des nouvelles caractéristiques à partir d'existantes pour obtenir des caractéristiques plus pertinentes qui amélioreront la performance du modèle, par exemple un ratio entre deux autres caractéristiques. Ce processus se nomme ingénierie des caractéristiques.

Un modèle est une représentation mathématique ou statistique de la relation entre les caractéristiques et les étiquettes dans un ensemble de données. Le modèle est construit à partir des données d'entraînement et est ensuite utilisé pour faire des prédictions sur de nouvelles données. Il possède plusieurs paramètres que l'on peut ajuster selon les données à prédire. Il peut prendre plusieurs formes, selon l'algorithme utilisé, comme un réseau de neurones, un arbre de décision, ou encore une régression linéaire. L'entraînement du modèle est le processus par lequel le modèle apprend à partir des données. Cela implique de présenter un ensemble de données étiqueté au modèle, de manière à ajuster les paramètres internes pour minimiser l'erreur entre les prédictions et les résultats réels. L'objectif est que le modèle généralise bien ce qui lui permettra de faire des prédictions sur des données qu'il n'a jamais vues.

Une fois que le modèle a été entraîné, il faut s'assurer de sa performance en procédant à son évaluation. On utilise un ensemble de validation, un sous-ensemble des données initiales, pour ajuster les hyperparamètres, des paramètres définis avant l'entraînement, comme le taux d'apprentissage, et un ensemble de test pour estimer la capacité du modèle à bien généraliser sur de nouvelles données. Pour évaluer la qualité d'un modèle d'apprentissage automatique,

on utilise des métriques qui dépendent du type du modèle. Pour un problème de classification, on peut utiliser l'exactitude qui est définie comme la proportion de prédictions correctes sur l'ensemble des prédictions faites par le modèle. On retrouve également la précision et le rappel. Ces dernières permettent de mesurer les faux positifs et les faux négatifs, ce qui est utile pour des tâches de détection de fraude, par exemple. On peut également construire un graphique contenant la courbe décrivant la fonction d'efficacité du récepteur. Elle trace le taux de vrais positifs contre le taux de faux positifs. L'aire sous cette courbe est une valeur qui résume cette performance : une aire sous la courbe de 1 indique un modèle parfait, tandis qu'une valeur de 0,5 correspond à une prédiction aléatoire. Pour les tâches de régression, on peut mesurer l'erreur quadratique moyenne. Elle mesure la différence moyenne entre les prédictions du modèle et les valeurs réelles, en élevant la valeur au carré pour accentuer les erreurs importantes.

La mesure des performances du modèle peut dévoiler plusieurs problèmes avec le modèle. Le surapprentissage se produit lorsqu'un modèle apprend trop bien les détails de l'ensemble de données d'entraînement, au point qu'il devient moins performant sur des données non vues. Cela arrive lorsque le modèle est trop complexe par rapport aux données disponibles, ce qui entraîne une mauvaise généralisation. D'autre part, le sous-apprentissage se produit lorsque le modèle est trop simple pour capturer les tendances ou la structure des données. Dans ce cas, le modèle ne parvient pas à apprendre correctement même à partir des données d'entraînement, ce qui se traduit par des erreurs importantes à la fois sur les ensembles d'entraînement et de test.

Voici un exemple détaillé d'un système de classification de courriels qui doit distinguer entre des messages désirables ou indésirables, cela permet de mieux comprendre où s'appliquent les concepts décrits précédemment. Pour chaque courriel, l'étiquette sera soit désirable ou indésirable. Les caractéristiques peuvent inclure des éléments comme le nombre

de mots, la présence de certains mots-clés, ou même des informations sur l'expéditeur. Un modèle de classification, comme un arbre de décision ou un réseau de neurones, est entraîné à partir d'un grand nombre d'exemples de courriels, chacun étiqueté comme désirable ou indésirable. Le modèle utilise les courriels et leurs caractéristiques pour ajuster ses paramètres et minimiser l'erreur de classification. Une partie des données est réservée pour valider la performance du modèle et ajuster ses hyperparamètres, tandis qu'un autre ensemble est utilisé pour tester sa capacité à prédire correctement sur des courriels qu'il n'a jamais vu. Si, 90% des courriels testés sont correctement classés, alors l'exactitude du modèle serait de 0,90. Si, sur 100 courriels marqués comme indésirable par le modèle, 85 étaient véritablement des indésirables, alors la précision est de 85%. Si, sur 100 courriels indésirables, le modèle en a identifié correctement 80, alors le rappel est de 80%. Si l'aire sous la courbe de la fonction d'efficacité du récepteur est de 0,92, cela indiquerait un bon équilibre entre les vrais positifs et les faux positifs à différents seuils. Dans ce cas-ci, un surapprentissage pourrait survenir si le modèle apprend trop les spécificités des courriels d'entraînement, comme des phrases précises présentes uniquement dans cet ensemble. Il pourrait mal classifier des courriels qui ne partagent pas ces phrases, mais qui sont toutefois indésirables. Le sous-apprentissage pourrait survenir si le modèle est trop simple et est incapable de capturer les relations complexes entre les caractéristiques, comme la présence de mots précis, qui déterminent si un courriel est indésirable. Le modèle pourrait alors ne pas bien séparer les courriels indésirables des désirables.

L'évolution des algorithmes d'apprentissage automatique a commencé par de simples modèles linéaires et a progressé vers des réseaux de neurones complexes capables d'effectuer des tâches que l'on croyait réservées aux humains, comme la reconnaissance vocale et l'interprétation d'images complexes. On retrouve cinq types d'apprentissage auto-

matique [30, 31, 32] : apprentissage supervisé, apprentissage non-supervisé, apprentissage auto-supervisé, apprentissage par renforcement et apprentissage semi-supervisé.

L'apprentissage supervisé est un type d'apprentissage automatique où le modèle est entrainé sur des données étiquetées, c'est-à-dire que la variable de sortie est connue. Par exemple, dans un modèle qui cherche à prédire le prix d'une voiture, les variables d'entrée pourraient être la puissance du moteur, la consommation d'essence et le nombre de passagers et la variable de sortie serait le prix. Les algorithmes d'apprentissage supervisés sont souvent utilisés pour la détection de fraude, l'analyse prédictive, la reconnaissance d'image et d'autres tâches. Notamment, les réseaux de neurones font partie de cette catégorie de techniques.

L'apprentissage non-supervisé est un type d'apprentissage automatique où l'inférence se fait sur des ensembles de données non étiquetées. Ceci a pour avantage de faciliter l'analyse exploratoire des données et permet la reconnaissance de modèles dans les données et la modélisation prédictive. Les algorithmes de regroupement sont les algorithmes les plus communs qui utilisent l'apprentissage non-supervisé. Ils permettent de regrouper les points de données selon leurs similarités. Ce type d'apprentissage automatique est souvent utilisé dans les systèmes de recommandation.

L'apprentissage auto-supervisé permet aux modèles de s'entraîner d'eux-mêmes sur des données non étiquetées. Ce type d'algorithme apprend une partie de l'entrée à partir d'une autre partie, générant les étiquettes automatiquement et transformant le problème en apprentissage supervisé. Ce type d'apprentissage automatique est utile lors de problèmes où le volume de données à étiqueter est énorme, par exemple le traitement automatique des langues ou la vision par ordinateur.

L'apprentissage par renforcement est un type d'algorithme qui apprend à l'aide d'un système de récompenses et punitions. Lors de l'entraînement, un agent pose des actions dans

un environnement spécifique pour atteindre un objectif précis. L'agent est récompensé ou puni selon ses actions à l'aide d'un système de points, qui encourage l'agent à effectuer les actions qui le rapprochent le plus de son objectif. L'apprentissage par renforcement est fréquemment utilisé dans les jeux vidéos et pour apprendre aux robots à répliquer des tâches humaines.

L'apprentissage semi-supervisé est une combinaison de l'apprentissage supervisé et non-supervisé. Ce type de modèle est entraîné à l'aide d'un petit ensemble de données étiquetées et d'un grand ensemble de données sans étiquettes. L'algorithme peut ensuite utiliser le regroupement pour identifier les groupes et étiqueter ces groupes à l'aide des données étiquetées qui se retrouvent dans ces groupes.

La construction d'un modèle implique six étapes [33]. La première se trouve à être la définition du problème. Celle-ci consiste en un accord entre les parties prenantes sur la façon dont le modèle devrait fonctionner. Ensuite, les données sont collectées, nettoyées et étiquetées lorsque nécessaire. On peut également modifier ces données en effectuant de l'ingénierie de caractéristiques afin d'ajouter des caractéristiques aux données existantes.

Le modèle est ensuite entraîné. Une multitude d'algorithmes et paramètres peuvent être testés afin de trouver le modèle qui offre les meilleurs résultats pour le cas d'utilisation. On évalue également celui-ci pour vérifier qu'il convient aux objectifs des parties prenantes. Le modèle est finalement déployé afin que des requêtes d'inférences puissent être effectuées sur celui-ci. la figure 2.1 illustre ces étapes.

Ces étapes semblent relativement simples, mais des défis émergent lors de l'implémentation d'un système d'apprentissage automatique. Malgré le fait que l'algorithme est au cœur d'un modèle d'apprentissage automatique, celui-ci n'est qu'une petite partie des éléments nécessaires à la construction d'un système complet. [34] explique bien ce principe et une

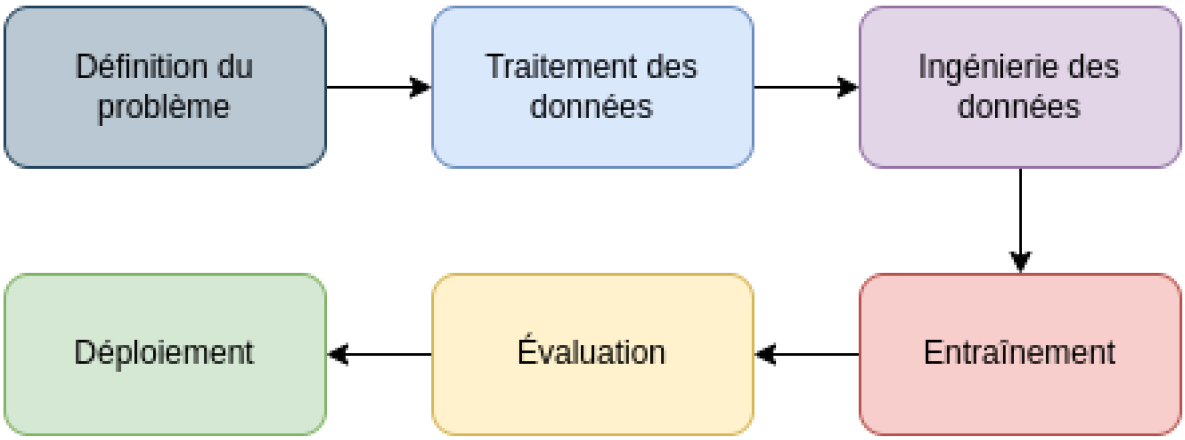


FIGURE 2.1 : Étapes pour la construction d'un modèle d'apprentissage automatique

vingtaine de défis liés au déploiement d'un modèle. Bien qu'il soit possible d'effectuer ces étapes manuellement, cette méthodologie devient rapidement trop lourde et coûteuse pour être soutenue à long terme. En effet, l'implémentation du modèle n'est qu'une partie du problème, le modèle doit également être mis à jour régulièrement avec des données récentes afin de ne pas perdre de sa pertinence. On retrouve des défis liés à l'implémentation dans la section 1.2.

Pour relever ces défis, il est essentiel d'automatiser le plus possible les étapes nécessaires à la création d'un système d'apprentissage automatique. Pour cela, il est nécessaire de faire l'utilisation du concept de MLOps, que nous décrivons dans la prochaine section.

2.1.2 DEVOPS ET MLOPS

La méthodologie DevOps permet d'intégrer le développement logiciel et les opérations reliées à celui-ci en utilisant le développement et le déploiement automatique ainsi que le monitoring de l'infrastructure d'un système. Ceci permet d'accélérer la vitesse de développement des logiciels de manière robuste, diminue les malentendus entre les développeurs et permet de résoudre les problèmes plus rapidement [35]. Cette façon de faire permet aux

acteurs d'équipes distinctes de mieux communiquer et évite des problèmes de répartition des tâches et des responsabilités.

On peut séparer le concept de DevOps en trois grandes étapes [36]. La construction inclut tout ce qui se rapporte à la compilation, la gestion des dépendances, la génération de documentation, l'exécution des tests ou le déploiement d'une application dans différents environnements. La phase de déploiement permet de partager, tester et faire le contrôle de version de l'infrastructure, réduisant les bogues causés par les différentes configurations de celle-ci. La phase des opérations consiste à en la mise en place des outils visant à s'assurer que le système ne se dégrade pas au fil du temps. On peut ainsi mettre en place des outils de monitoring et de journalisation. Le DevOps est de plus en plus utilisé en entreprise et prend une place majeure dans le développement de logiciels ; en effet, selon un sondage effectué par GitLab en 2023, 56% des entreprises utilisaient le DevOps comparativement à 47% en 2022 [37].

Le concept de MLOps est similaire à celui de DevOps, sauf qu'il est appliqué aux systèmes qui utilisent l'apprentissage automatique : on utilise donc les pratiques DevOps pour résoudre des problèmes liés à l'apprentissage automatique [38]. Cela combine les pratiques DevOps, l'apprentissage automatique et l'ingénierie des données. Alors que le mouvement DevOps a commencé dans les alentours des années 2007-2008 [39], la première mention du concept de MLOps est arrivée en 2015 dans un article écrit par Google [34]. Malgré le fait que des outils DevOps existaient déjà lors de l'arrivée du MLOps, la nature de l'apprentissage automatique ne permet pas toujours aux développeurs d'utiliser tous les outils DevOps dans les systèmes ML[40]. En effet, les logiciels traditionnels sont guidés par le code tandis que l'apprentissage automatique est guidé par les données et le code. Cela ajoute donc des aspects auparavant inconnus dans le développement logiciel. L'article de Google décrit pour la première fois les niveaux de dettes techniques élevés associés avec le déploiement manuel

de systèmes ML, raison de plus pour l'utilisation de MLOps. Celui-ci décrit également les différentes composantes qui font partie d'un système ML ; de plus quatre compagnies, Facebook [41], Nvidia [42], Spotify [43] et Google [44], documentent leur implémentation d'un pipeline MLOps à partir de ces composantes.

On peut diviser un pipeline MLOps en sept étapes en généralisant l'implémentation de ces différentes compagnies : le traitement des données, l'extraction des caractéristiques des données, la création des modèles, l'entraînement des modèles, l'évaluation des modèles, le déploiement et le monitoring et le logging. La figure 2.2 représente un pipeline MLOps typique :

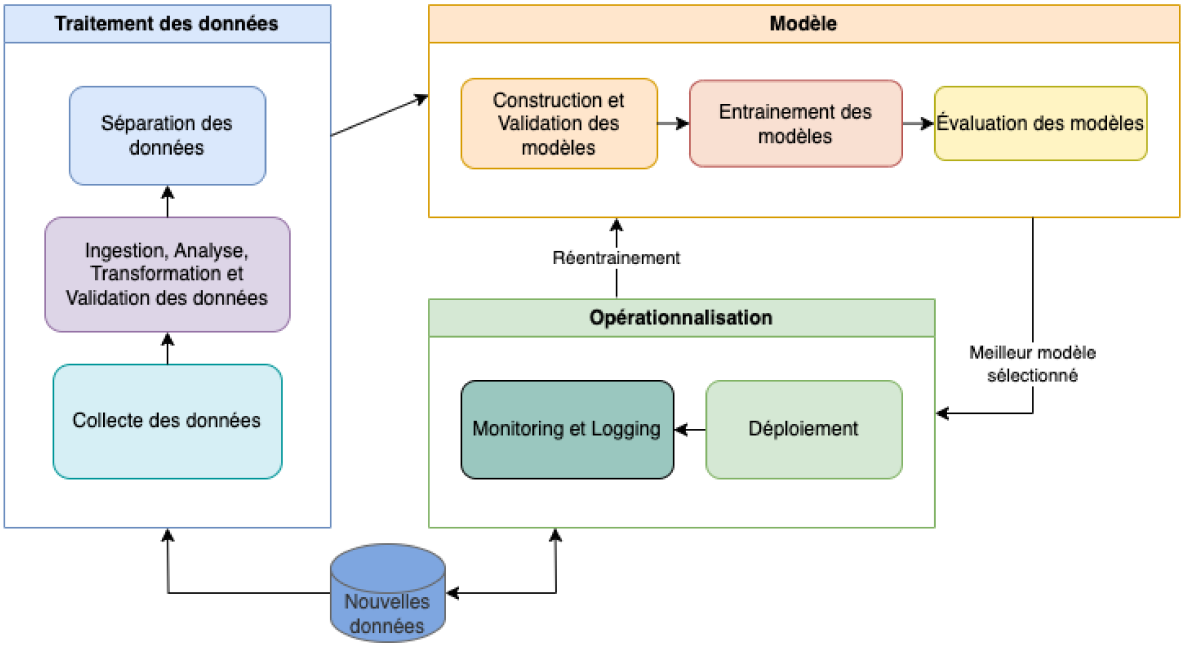


FIGURE 2.2 : Étapes d'un pipeline MLOps

Les étapes de construction d'un modèle décrites plus haut se retrouve tous dans le pipeline MLOps sauf celle du la définition du problème. Celle-ci doit se faire avant la construction du pipeline MLOps.

La méthode MLOps a été testée dans ces trois études de cas avec succès [45, 46, 47]. On considère généralement MLOps comme une bonne pratique a adopter. [48, 49, 50]

L'implémentation de chacune des étapes présentées dans le diagramme nécessite des connaissances spécialisées, ce qui demande souvent de la main-d'œuvre avec des connaissances dans chaque domaine respectif.

2.1.3 PETITE ENTREPRISE

Pour ce qui est de la définition d'une petite entreprise dans le contexte de ce mémoire, trois sources définissent les tailles d'entreprises selon certains critères. Selon le gouvernement du Canada [11], une petite entreprise est définie comme étant une entreprise employant de 1 à 99 employés payés ; il existe aussi des sous-catégories comme les micro-entreprises qui comptent de 1 à 4 employés. Selon les standards de la Banque mondiale [51], on peut définir la grandeur d'une entreprise selon 3 critères : le nombre d'employés, la valeur des actifs et le chiffre d'affaires annuel. Pour appartenir à une catégorie, une entreprise doit au minimum remplir la propriété du nombre d'employés et l'une des deux autres conditions. La figure 2.3 catégorise ces entreprises.

On trouve aussi des manières qualitatives de classer les entreprises, comme décrit par [24], selon le type de gestion de l'entreprise, le type d'employés de l'entreprise, la communication interne, les ventes, les relations avec les clients, la production, la recherche et le développement et le mode de financement. Le tableau 2.1 décrit les différences qualitatives entre les PME et les grandes entreprises selon cet article.

Ces sources sont un bon début afin de définir le terme «petite entreprise». Cependant elles ne prennent pas en compte le département dans lequel chacun des employés œuvre,

Catégorie	PME	Grandes entreprises
Gestion	— Propriétaire entrepreneur — Fonctions liées aux personnalités	— Gestionnaire entrepreneur — Division du travail selon le sujet
Personnel	— Manque d'employés gradués de l'université — Connaissances générales	— Majoritairement des gradués universitaires — Spécialisation
Organisation	— Contrats hautement personnalisés	— Communications hautement formelles
Ventes	— Position concurrentielle non définie et incertaine	— Position concurrentielle forte
Relations client	— Incertain	— Basée sur des contrats à long terme
Production	— Hautement demandant en main d'oeuvre	— Hautement demandant en capital, économies d'échelle
Recherche et développement	— Suit le marché, approche basée sur l'intuition	— Institutionnalisé
Finances	— Auto-financement	— Accès à des capitaux anonymes

TABLEAU 2.1 : Indicateurs qualitatifs discernant les PME des grandes entreprises

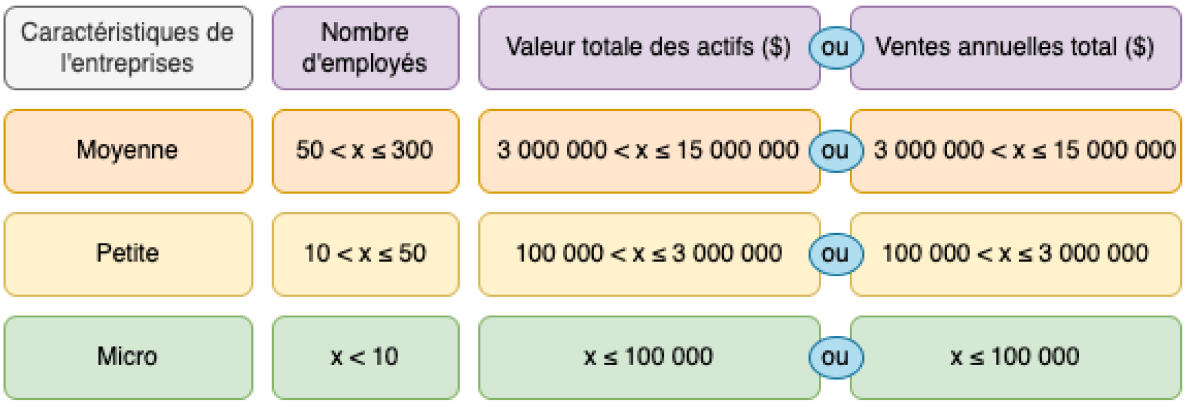


FIGURE 2.3 : Définition de micro, petite et moyenne entreprise selon les standards de la banque mondiale.

c'est-à-dire, le nombre d'employés étant des gestionnaires, le nombre d'employés dans le département des ressources humaines, le nombre d'employés dans le département technique, etc. En effet, ces informations sont également importantes pour ce qui a trait à la définition de petite entreprise dans ce travail de recherche. Une entreprise de dix employés dont tous les employés sont des développeurs est très différente d'une entreprise de dix employés dont seulement un est un développeur. Les ressources disponibles en matière de main-d'œuvre dans ces deux types d'entreprises seront très différentes : les petites entreprises auront de la difficulté à engager du personnel hautement qualifié dans des domaines spécialisés alors que les plus grandes entreprises pourront se permettre de la main-d'oeuvre plus spécialisé comme mentionné dans la ligne *personnel* du tableau 2.1. Cet aspect devrait être considéré lors de la définition de la grandeur d'une entreprise, surtout si celle-ci œuvre dans le département du développement logiciel et veut implémenter un système d'apprentissage automatique.

Avec les différentes informations recueillies, nous définissons une petite entreprise comme étant une entreprise possédant moins de dix employés auxquels moins de cinq d'entre eux sont dans le domaine du développement logiciel et aucun d'entre eux ne soit spécialisé dans

le domaine de l'apprentissage automatique, en plus de posséder en majorité les caractéristiques des PME définies dans le tableau 2.1.

2.2 MÉTHODES ACTUELLES EN MLOPS

On peut regrouper les différentes solutions d'implémentation de pipeline MLOps en trois grandes catégories [38], soit les solutions sur mesure, les solutions open-source et les services d'apprentissage automatique (MLaaS).

2.2.1 SOLUTIONS SUR MESURE

Les plus grandes entreprises ont souvent un système entièrement personnalisé qui varie grandement dans sa structure et sa complexité. Ces solutions sont développées par des experts dans différents domaines relatifs à apprentissage automatique, ce qui permet une performance optimale du système. Cependant, ce type de solution n'est pas adapté aux petites entreprises, car celui-ci nécessite des employés avec des connaissances spécialisées, beaucoup de temps et d'argent.

2.2.2 SOLUTIONS OPEN-SOURCE

On peut également également créer un pipeline MLOps en utilisant des outils open-source. Cela apporte des avantages [52] comme un coût moindre, de la flexibilité, la possibilité de personnaliser le code et généralement plus d'aide de la communauté. Cependant, ces outils seuls ne sont pas une solution tout-en-un, on doit combiner plusieurs outils afin de créer un système complet. Cela ajoute de la complexité au système et nécessite également un certain niveau de connaissance de la part des employés. L'entreprise doit aussi s'occuper de l'hébergement elle-même, ce qui peut compliquer les choses.

2.2.3 MACHINE LEARNING AS A SERVICE

Finalement, on retrouve les solutions de type *machine learning as a service* (MLaaS). Celles-ci consistent en des services d'apprentissage automatique offerts par des entreprises. Ces services sont basés dans le cloud, cela veut dire que le stockage des données et les calculs sont effectués sur les machines de l'entreprise qui fournit le service. La plupart des entreprises qui offrent ce service ont des modalités de paiement où l'on paie pour ce que l'on utilise, ce qui permet aux plus petites entreprises de ne pas se ruiner dans des installations de serveurs et de ne payer que pour ce dont elles ont réellement besoin.

Des plateformes offrent une expérience tout-en-un, par exemple AWS SageMaker [53], Microsoft Azure ML [54], Tensorflow TFX [55] et Google AI Platform [56]. Ces plateformes ont leur propre écosystème de composantes, ce qui permet de facilement créer un pipeline MLOps. Une plateforme typique inclurait, entre autres, un service de stockage de données, l'hébergement de modèle avec des APIs pour l'entraînement et l'inférence, un ensemble de métriques pour la surveillance des modèles et une interface qui accepte la personnalisation de l'utilisateur. En offrant des infrastructures gérées et une gamme d'implémentation pour des tâches communes, ce type de plateforme réduit grandement le fardeau opérationnel que demande l'entretien d'un modèle en production [34] et comporte de nombreux avantages pour les PME [57].

Dans cette catégorie, on retrouve également des plateformes dites low-code ou no-code, c'est-à-dire qu'elles ne nécessitent que peu ou pas de code. Ceci est un grand avantage pour les entreprises qui n'ont aucun employé spécialisé, car ces plateformes permettent l'implémentation de pipelines MLOps avec des connaissances minimales en apprentissage automatique. On retrouve même des plateformes qui utilisent le principe AutoML. Selon [58], AutoML est un remplaçant des humains pour l'identification de configurations, qui sont

exclusifs aux programmes d'apprentissage automatique, avec un budget informatique limité. Les configurations étant les différentes manières de paramétrer les outils nécessaires à la création d'un système ML. Un exemple d'AutoML pourrait être la recherche automatique des meilleurs hyperparamètres pour l'entraînement d'un modèle ou bien la préparation automatique des données d'entraînement. Le principe AutoML a tout de même ses limitations et ses problèmes [59] et les utilisateurs doivent tout de même être conscients de ceux-ci lors de l'implémentation du système [60, 61].

Toutefois, trois sources mentionnent que les plateformes utilisant AutoML peuvent arriver à de bons résultats et sont significativement plus faciles à gérer comparativement à l'implémentation personnalisée de systèmes ML [62, 59, 63]. Les entreprises peuvent utiliser ce type de plateforme à leur avantage pour implémenter des projets ML plus rapidement si elles possèdent des connaissances limitées dans les détails d'implémentation ML, (comme mentionné dans la section 2.1.3), et si leur problème peut être résolu avec des stratégies d'optimisation génériques [23]. AutoML étant un principe puissant pour les raisons mentionnées plus haut, et les services cloud comportant une multitude d'avantages importants pour les PME [64, 65], les services AutoML dans le cloud sont la solution idéale pour les PME qui veulent commencer dans le monde de l'apprentissage automatique.

2.2.4 INTÉGRATION DE L'INTELLIGENCE ARTIFICIELLE DANS LES PETITES VS. LES GRANDES ENTREPRISES

L'adoption de l'intelligence artificielle varie significativement selon la taille de l'entreprise, en raison des différences dans les ressources disponibles, la scalabilité et des besoins spécifiques des entreprises. Ces différences sont nommées dans cinq sondages et revues de littérature [66, 67, 68, 69, 70].

D'abord, les petites entreprises doivent souvent se fier à des solutions abordables et accessibles, telles que les services dans le cloud ou des modèles préentraînés. Cela permet de réduire le besoin de main-d'œuvre experte et d'infrastructure sur site. En comparaison, les grandes entreprises ont des ressources financières suffisantes pour engager des experts en IA et pour investir dans l'infrastructure nécessaire pour prendre en charge un système d'apprentissage automatique complet.

Ensuite, les petites entreprises ne disposent généralement que d'ensembles de données limités et peuvent avoir besoin d'utiliser des données synthétiques pour améliorer l'entraînement des modèles. À l'opposé, les grandes entreprises possèdent de vastes quantités de données, ce qui peut mener à des modèles plus performants.

L'implémentation de l'IA dans les petites entreprises se fait souvent de manière incrémentale, c'est-à-dire en commençant par des projets pilotes qui permettent de tester la viabilité d'un projet d'apprentissage automatique, et qui se transforme en système plus complet au fil du temps. En revanche, les grandes entreprises peuvent développer des solutions IA personnalisées adaptées aux besoins spécifiques de leurs cas d'utilisation.

Les petites entreprises utilisent aussi fréquemment des solutions IA basées sur des services cloud, puisque ceux-ci fournissent des modèles préconfigurés et des interfaces faciles à utiliser, abaissant ainsi la barrière d'entrée pour l'implémentation d'un système d'apprentissage automatique. Elles ont besoin de solutions IA flexibles et adaptables à leurs besoins sans nécessiter de modifications importantes ou de connaissances spécialisées. Les grandes entreprises, quant à elles, sont en mesure d'intégrer des plateformes d'analyse plus poussées qui combinent l'IA et les technologies de données massives afin d'obtenir des métriques en direct et des capacités de prédictions avancées.

Les outils AutoML sont souvent utilisés par les petites entreprises pour simplifier le processus de développement, ce qui permet aux utilisateurs sans expertise en apprentissage automatique de créer et déployer des modèles. Contrairement aux grandes entreprises qui peuvent se permettre de développer des modèles personnalisés optimisés à leurs cas d'utilisation spécifiques et qui peuvent également investir dans des technologies cloud et hybrides évolutives pour prendre en charge le déploiement à grande échelle des applications IA.

Le tableau 2.2 résume les différences mentionnées dans cette section.

2.3 IA ET ML POUR LES PETITES ENTREPRISES: OPPORTUNITÉS ET DÉFIS

L'implémentation de l'IA dans les petites entreprises offre de nombreuses opportunités [71, 72] qui viennent cependant avec des défis considérables. Ces technologies peuvent transformer les processus opérationnels, améliorer la productivité des employés, accélérer les tâches répétitives, planifier les horaires de travail de manière efficace, optimiser les processus manufacturiers et améliorer l'expérience client à travers des assistants virtuels, par exemple. Cependant, leur adoption présente des barrières uniques pour les petites entreprises par rapport aux grandes entreprises.

Du côté des opportunités, on note que les services cloud tels que Google Cloud AI [8], AWS Machine Learning [7] et Microsoft Azure AI [6] fournissent des modèles préconfigurés et des interfaces conviviales, ce qui rend l'implémentation de modèles d'apprentissage automatique plus accessible aux petites entreprises. Ces services permettent aux petites entreprises de démarrer des projets de ML sans nécessiter une infrastructure coûteuse ou des experts en IA internes.

Les outils AutoML simplifient le développement de modèles [10], ils permettent aux utilisateurs sans expertise en apprentissage automatique de créer et déployer des modèles

PME	Grandes entreprises
Doivent se fier aux solutions IA abordables et accessibles. Cela inclut les services IA dans le cloud et les modèles préentrainés qui réduisent le besoin de main-d'œuvre experte et d'infrastructure sur site.	Ont les ressources financières pour engager des experts en IA et en ML et investir dans l'infrastructure logiciel et matériel nécessaire.
Ont des ensembles de données de taille limitée. Les petites entreprises peuvent avoir besoin d'utiliser des données synthétiques pour améliorer l'entraînement des modèles.	Possèdent de grandes quantités de données, ce qui est crucial pour l'entraînement de modèles performants.
Privilégient les implémentations de l'IA d'une manière incrémentale en débutant par des projets pilotes qui résolvent un problème spécifique au lieu d'implémentations à grande échelle.	Peuvent développer des solutions IA personnalisées aux besoins spécifiques de l'entreprise. Cela inclut le développement de modèles personnalisés optimisés pour des cas d'utilisation spécifiques à l'entreprise.
Utilisent des solutions IA de services cloud, qui fournissent des modèles préconfigurés et des interfaces faciles à utiliser, ce qui abaisse la barrière d'entrée.	Ont plus de difficulté à gérer la complexité du projet. Cela est dû entre autres au traitement de données massives et à l'intégration du système ML dans plusieurs départements.
Nécessitent des solutions IA qui sont hautement flexibles et qui peuvent s'adapter à leurs besoins sans modifications importantes ou connaissances spécialisées.	Peuvent implémenter des plateformes d'analyse qui intègrent l'IA aux technologies de données massives, ce qui permet d'obtenir des métriques en direct et des capacités de prédiction avancées.
Peuvent utiliser les outils AutoML pour simplifier le processus de développement, ce qui permet aux utilisateurs sans expertise en IA de créer et déployer des modèles.	Peuvent investir dans des technologies cloud ou hybride évolutives pour prendre en charge le déploiement à grande échelle d'applications IA.

TABLEAU 2.2 : Différences spécifiques à l'intelligence artificielle entre les petites et les grandes entreprises

efficacement [9]. Cela est particulièrement bénéfique pour les petites entreprises qui peuvent utiliser ces outils pour automatiser des tâches répétitives et libérer du temps pour des activités plus stratégiques en plus d'accélérer le développement des projets d'apprentissage automatique [23]. Malgré les succès de AutoML, cette technologie est toujours en développement et qu'elle est parfois limitée en termes d'interprétabilité, de reproductibilité et de robustesse [73, 59].

La coopération avec des entreprises plus grandes qui possèdent de l'expérience en ML ou avec des universités ou des institutions de recherche peut également offrir une expertise en IA qui permet de contrer plusieurs problèmes que rencontrent les petites entreprises [23].

L'intégration de l'apprentissage automatique peut offrir un avantage concurrentiel en améliorant la prise de décision basée sur les données, en optimisant les opérations et en offrant des expériences personnalisées aux clients. Cela peut également accélérer l'innovation et le développement de nouveaux produits et services [5].

D'un autre côté, les petites entreprises disposent souvent de ressources limitées, tant du côté financier que du côté de la main-d'œuvre [24]. Le coût des services cloud peut être problématique pour certaines entreprises et le manque de personnel qualifié en IA peut gêner la mise en œuvre efficace des projets de ML.

La main-d'œuvre peu spécialisée en IA, en plus de rendre l'implémentation de système ML plus difficile, peut entraîner d'autres défis [23]. En effet, le manque de connaissance et d'expérience en IA en l'identification des cas d'utilisation possible pour 'apprentissage automatique dans l'entreprise difficile. Les cas d'utilisation identifiés se rendent rarement au stade de preuve de concept et l'entreprise se fie souvent sur une seule personne pour examiner les cas d'utilisation possibles. Les connaissances limitées en ML peuvent également préoccuper les gestionnaires qui ne verront pas l'intérêt du ML dans leur entreprise.

La collecte de données peut également devenir un défi. En effet, le traitement des données nécessitent des ressources qui, comme mentionné dans le paragraphe précédent, sont limitées. Dans certains cas l'entreprise peut utiliser des ensembles de données open-source ou des modèles préentraînés, cependant pour les cas où le domaine est spécifique à l'entreprise, cette étape est nécessaire et coûteuse [74] et souvent les entreprises n'emmagasinent pas assez de données ou bien des données de mauvaise qualité [75] . La récolte de données inclut jusqu'à neuf étapes telles que l'identification de sources de données, la collecte des données, l'intégration de plusieurs sources de données si nécessaire, le nettoyage et le prétraitement, l'étiquetage, la gestion du stockage, la documentation et la conformité des données.

Le manque de données est également un enjeu important qui peut causer des problèmes de performance lors de l'entraînement du modèle. Selon [23], l'insuffisance de données disponibles pour l'entraînement de modèles est l'un des défis principaux des petites entreprises.

2.4 ÉTUDES DE CAS ET APPLICATIONS PRATIQUES DANS LES PETITES ENTREPRISES

Trois études de cas ont déjà été réalisées dans le cadre de recherches. Faes et al. [61] effectue une étude de cas sur l'implémentation d'un modèle d'apprentissage profond par des professionnels de la santé sans expérience de codage en utilisant des outils autoML. Le manque de connaissance des professionnels de la santé de l'étude dans les domaines des mathématiques, des statistiques et de la programmation limite la performance des modèles. Les auteurs concluent que l'étude de cas a permis de créer des modèles avec des performances comparables aux modèles à la fine pointe dans la majorité des expériences. Cependant, ils mentionnent que ce type de modèle devrait être déployé avec prudence pour éviter la discrimination qui pourrait causer du tort. On peut faire un lien avec les petites entreprises, car les professionnels de la santé qui ont implémenté le modèle dans cette étude n'ont pas

de connaissance spécialisée en ML, ce qui ressemble au cas des petites entreprises. Cette étude démontre qu'il est possible de déployer des modèles de deep learning convenable sans spécialisation en ML à l'aide d'outils AutoML.

Bender et al. [76] fait le test de trois solutions AutoML Open-Source construites en se basant sur les données de PME dans le domaine manufacturier et les comparent avec des solutions construites manuellement. Plus précisément, dans la construction de modèles pour la prédiction des délais de mise en œuvre. Ces solutions AutoML ont pu surpasser certaines des solutions manuelles, mais pas tous. Les auteurs notent que l'utilisation d'AutoML est utile pour la sélection d'algorithmes, l'entraînement, l'optimisation d'hyperparamètres, l'analyse comparative, la sélection et l'encodage de données, elles ne prennent pas en charge pas les activités qui demandent beaucoup de main-d'œuvre comme la compréhension, la transformation, le filtrage et le prétraitement des données. Cette étude de cas démontre que l'utilisation d'outils AutoML est un facilitateur et aide dans plusieurs étapes nécessaires au développement d'un modèle d'apprentissage automatique. Cependant, aucun des outils testés n'a pu automatiser le développement en entier, ce qui démontre un certain minimum de connaissances nécessaire pour l'utilisation des outils testés.

Stühler et al. [77] examine l'efficacité d'approches AutoML pour automatiser le prétraitement des données et la création de modèles d'apprentissage automatique, en se concentrant sur leur utilité dans les PME avec des capacités limitées en science des données et en apprentissage automatique. L'étude de cas en question se penche sur la prédiction de la valeur résiduelle d'engins de chantier de construction et révèle que toutes les approches AutoML testées surpassent le pipeline ML de référence construit manuellement. Les méthodes AutoML utilisées sont adaptées aux personnes n'ayant que des connaissances du domaine et des compétences de base en traitement de données. L'approche proposée est centrée sur le développement de preuves de concept qui donne aux entreprises une méthode peu coûteuse et pratique pour développer

leur premier modèle ML et pour faire leur entrée dans le domaine de l'intelligence artificielle. Cette étude démontre l'efficacité des méthodes AutoML, surtout pour les entreprises ayant peu de connaissances en intelligence artificielle. Elle apporte un autre exemple de l'utilisation de AutoML pour faciliter l'implémentation d'un système d'apprentissage automatique.

2.5 TRAVAUX CONNEXES EN RAPPORT AVEC L'OBJECTIF

On retrouve une dizaine d'articles en lien avec l'objectif de ce mémoire. Ceux-ci sont présentés dans les prochaines sous-sections.

2.5.1 AUTOML

Le premier d'entre eux par Karmaker et al. [1] explique les différents avantages et défis présents dans l'implémentation d'un système d'apprentissage automatique et comment on peut se servir d'outils AutoML pour faire face à ces défis et augmenter l'efficacité d'un tel système. Cet article propose une échelle comportant sept niveaux qui permet de mesurer le niveau d'automatisation d'un système d'apprentissage automatique. Ces différents niveaux sont présentés brièvement dans le tableau 2.3.

Pour le niveau 0, on ne retrouve aucune automatisation et les algorithmes d'apprentissage automatique sont codés à la main. Les systèmes de niveau 1 fournissent des implémentations d'algorithmes d'apprentissage automatique. La plupart de ces outils sont des librairies qui sont utilisées par des scientifiques de données. Le niveau 2 décrit les systèmes qui utilisent les implémentations d'algorithmes d'apprentissage automatique en plus d'utiliser une composante qui permet d'explorer, tester et valider différents modèles potentiels. Cela nécessite encore beaucoup de labeurs manuels et est seulement accessible aux scientifiques de données.

Le niveau 3 automatise la recherche du meilleur modèle et le réglage des hyperparamètres dans un outil qui permet de lier ces tâches. Un système de niveau 4 peut faire l'étape de prétraitement des données automatiquement ainsi que toutes les autres fonctionnalités des niveaux précédents. Un expert du domaine peut interagir avec un tel système minimalement tandis qu'un expert en science des données doit encore faire une bonne partie du travail manuellement.

Un système de niveau 5 inclut l'automatisation des niveaux précédents en plus d'une composante qui permet aux experts de domaine de définir un objectif de prédiction selon les données fournies que le système pourra ensuite atteindre. Ce niveau d'automatisation comporte plusieurs défis non résolus comme comment bien interpréter les objectifs réels de l'utilisateur et comment sélectionner les données utiles à l'accomplissement de cet objectif. Finalement, un système de niveau 6 fournit un maximum d'automatisation et requiert un minimum d'efforts des scientifique des données. Ce niveau inclut l'automatisation des niveaux précédents ainsi que quelques composantes supplémentaires. Le système peut formuler des problèmes de prédictions presque de manière autonome avec les buts définis par les experts du domaine. Ce système peut recommander des tâches de prédiction pertinentes à l'utilisateur ainsi que présenter les résultats de manière à ce que les experts du domaine puissent être autonomes.

Les niveaux d'automatisation présentés par cet article permettent de mieux comprendre la portée et les besoins en main-d'œuvre des différentes options disponibles pour l'implémentation d'un système d'apprentissage automatique. Dans le cas d'une PME, on doit prioriser les outils et services qui favorisent un niveau d'automatisation plus élevé. Selon la définition des différents niveaux d'automatisation présentés, les différentes plateformes mentionnées dans la section 2.2.3 se situeraient dans les environs du niveau 5. En effet, ces plateformes permettent

une automatisation avancée et sont conçues pour permettre aux personnes non expertes en apprentissage automatique d'implémenter un système d'apprentissage automatique.

2.5.2 SÉLECTION D'OUTILS ET DE PLATEFORMES GÉNÉRAL

L'article de Gyani et al. [78] procède à l'évaluation des méthodes de priorisation lors de l'évaluation des fournisseurs cloud par le biais d'une revue de littérature. Il propose également des critères et des sous-critères pour l'évaluation des services cloud. Les critères sont divisés dans six grandes catégories : sécurité, performance, migration, disponibilité, coût et responsabilité. Les auteurs présentent ensuite 19 techniques différentes retrouvées dans la littérature qui permettent de faire la sélection d'un fournisseur cloud selon divers critères. Ils expliquent aussi les différentes variations de la technique de sélection la plus populaire : le processus hiérarchique analytique (AHP), que l'on décrit dans la section 3.3. La revue de littérature se termine en concluant que la méthode préférée par les chercheurs des articles étudiés est AHP, suivi de la technique d'ordre de préférence par similarité avec la solution idéale, une méthode d'analyse décisionnelle à critères multiples.

Ce dernier article présente plusieurs points intéressants et pertinents à notre étude. Les critères présentés, même s'ils ne prennent pas en compte l'aspect apprentissage automatique, sont pour certains pertinents dans le cadre de notre étude. Cet article permet également d'avoir une idée des techniques utilisées par les chercheurs pour la sélection de fournisseurs cloud, ce qui est directement lié avec l'objectif de notre étude. Cependant, l'étude ne propose pas une seule méthode qui serait la meilleure et est une revue de littérature, contrairement à notre étude qui propose une méthode concrète.

L'objectif principal de Zdraveski et al. [79] est de présenter une liste de critères pouvant être utilisés lors de la comparaison de fournisseurs de services cloud. Les auteurs commencent

Niveau d'automatisation	Composantes automatisées
Niveau 0	Aucune
Niveau 1	Implémentation basique d'algorithmes d'apprentissage automatique
Niveau 2	Niveau 1 et l'exploration, les tests et la validation de modèles alternatifs
Niveau 3	Utilisation d'un framework pour AutoML qui inclut les composantes du niveau 2 et l'orchestration de ces composantes entre elles
Niveau 4	Niveau 3 et ajout d'une composante dans le framework AutoML qui fait l'ingénierie de données automatiquement
Niveau 5	Niveau 4 et une composante qui fait l'ingénierie de prédiction automatiquement
Niveau 6	Niveau 5 et une composante qui permet aux experts du domaine de formuler des tâches à un haut niveau d'abstraction ainsi qu'une composante qui fait automatiquement la présentation des résultats et des recommandations à l'utilisateur

TABLEAU 2.3 : Niveaux d'automatisation d'un système d'apprentissage automatique selon les différentes composantes automatisées présentées dans [1]

par présenter différents avantages et inconvénients de l'utilisation d'un fournisseur cloud pour combler des besoins informatiques. Certains des avantages sont la réduction des coûts, la facilité de déploiement d'applications, la scalabilité et plusieurs autres. Du côté des désavantages, on retrouve la dépendance à une connexion Internet rapide, le stockage de données de manière sécuritaire et le manque de standardisation des infrastructures. Ils proposent ensuite 12 fonctionnalités que l'on retrouve dans les services cloud ainsi que leurs métriques associées qui permettent de comparer différents fournisseurs. Ces fonctionnalités sont les suivantes : capacité, communication, calcul, mémoire, temps, coût, élasticité, fiabilité, scalabilité, disponibilité, sécurité des données et authentification.

L'article précédent présente certains critères intéressants pour notre étude et fait un survol utile des différentes caractéristiques qu'il est possible de mesurer dans les fournisseurs cloud, il faut toutefois mentionner plusieurs critères sont orientés vers la location de machine sur des serveurs cloud plutôt que des services qui font abstraction de la machine sous-jacente. C'est-à-dire que certaines des métriques proposées prennent des mesures à plus bas niveau. Par exemple, on retrouve la métrique «vitesse du processeur» dans la fonctionnalité «capacité». Ce type de mesure n'est pas pertinent lors du choix d'un service de type MLaaS car les fournisseurs font abstraction de ces choix. Contrairement à notre étude, l'article ne propose aucune méthode précise pour la sélection d'un service à partir de ces métriques.

Ensuite, on retrouve l'article de Repschlaeger et al. [80] qui propose une méthodologie pour la sélection d'un fournisseur cloud. Les auteurs commencent par faire une revue de 16 articles qui mentionnent des aspects clés que des fournisseurs cloud. Ils expliquent ensuite la méthodologie proposée, le modèle AHP, en expliquant que sa sélection est due au fait qu'il est populaire, efficace et facile à utiliser. Les critères de sélection sont identifiés à l'aide d'une revue systématique de la littérature de 55 articles pertinents, d'une évaluation de 793 fournisseurs de services cloud et de plusieurs entrevues avec différents acteurs concernés

par la sélection d'un service cloud. Au total, l'article divise 62 critères dans 21 catégories d'exigences abstraites qui sont à leur tour divisées dans 6 dimensions cibles. Ces dernières sont les suivantes : flexibilité, coût, sécurité et conformité, portée et performance, fiabilité et finalement gestion du service et du cloud. Ensuite, les auteurs ont conduit un sondage avec 7 responsables informatiques de 3 compagnies différentes afin d'obtenir des poids pour chacun des critères relevés précédemment. À noter qu'ils demandent des poids pour trois types de fournisseurs cloud : Software as a Service (SaaS), Platform as a Service (PaaS) et Infrastructure as a Service (IaaS).

Ce dernier article propose plusieurs critères pertinents qui seront considérés lors de notre étude tout en prenant compte du fait que ceux-ci ne sont pas orientés vers la sélection d'un service d'apprentissage automatique, contrairement à notre étude. Un autre point qui diffère est le fait que la méthode que nous proposons est plus flexible, car elle prend en compte les besoins de chaque entreprise individuellement, contrairement à celle de l'article précédent qui met en place des poids prédéterminés pour toutes les situations. Les auteurs simplifient la tâche des décideurs en prenant cette décision, mais leur méthode ne peut s'adapter aux besoins de ceux-ci et ne garantit pas d'être à l'épreuve du temps, car plusieurs paramètres pourraient varier. Par exemple, dans le futur les experts pourraient ne plus s'entendre sur le critère le plus important dans la sélection d'un fournisseur cloud car ces fournisseurs évoluent dans le temps. Pour finir, cet article ne propose pas de métriques ou de méthodes exactes pour mesurer les différents critères qu'il propose, ce que nous abordons.

L'article suivant par Garg et al. [81] propose un framework pour l'évaluation des services cloud. Les critères proposés se basent sur des attributs définit auparavant dans [82], ces derniers se nomment *Service Measurement Index* (SMI) et permettent de comparer des services aux entreprises d'une manière standard. Les attributs sont les suivants : responsabilité, agilité, coût, performance, fiabilité, sécurité et confidentialité et utilisabilité. Les attributs de

SMI étaient définis à haut niveau et les auteurs de [81] ont utilisé ces attributs pour définir précisément des critères pour la sélection d'un service cloud dans un framework nommé *Service Measurement Index Cloud framwork* (SMICloud). Les auteurs définissent ensuite 15 critères en se basant sur les attributs de SMI. Ces critères sont définis de manière précise avec des formules mathématiques lorsque ceux-ci sont quantifiables, et avec une description exhaustive de la manière dont on peut les mesurer lorsqu'ils sont qualifiables. On propose ensuite une méthodologie utilisant AHP pour la sélection d'un fournisseur cloud selon les critères définis. On y décrit les trois phases du processus : la décomposition du problème, le jugement des priorités et l'agrégation des priorités. La dernière section de l'article présente une étude de cas qui évalue trois différents fournisseurs cloud.

L'article précédent a plusieurs points en commun avec notre étude. Il propose plusieurs critères pertinents pour la sélection d'un service cloud et utilise encore une fois la méthode AHP. Cependant, il ne prend pas en compte l'aspect apprentissage automatique et certaines métriques sont orienté vers le côté matériel des serveurs, d'une manière comparable à [79]. On retrouve donc une méthode plus générale pour mesurer un fournisseur cloud contrairement à notre étude qui ne mesure que les fournisseurs de MLaaS.

Godse et Mulik [83] proposent une approche pour la sélection d'un service cloud de type SaaS. Ils y décrivent 16 attributs brièvement décrits divisés en cinq catégories : fonctionnalité, architecture, utilisabilité, réputation et coût. Cet article propose la même méthodologie que plusieurs autres articles décrivent dans cette section, soit AHP. L'étape de sélection d'un SaaS décrite dans l'article demande l'avis de cinq experts qui possèdent de l'expérience dans l'implémentation du type de SaaS qu'ils veulent sélectionner. Cette étude offre également une étude de cas pour la sélection d'une solution pour l'automatisation d'un système de vente en ligne.

Ce dernier propose plusieurs attributs intéressants dérivés des fournisseurs de solutions SaaS, qui possèdent des attributs similaires à ceux des solutions MLaaS, que notre étude prend en compte. Cependant, le fait que la méthodologie de l'étude fait appel à cinq experts du domaine fait en sorte que celle-ci n'est pas adaptée aux petites entreprises, qui ne possèdent que rarement autant de main-d'œuvre spécialisée.

L'article de Şener et al. [84] décrit un système d'aide à la décision dénommée ClouDSS. Les auteurs décrivent bien le fait que les services cloud peuvent être bénéfiques pour les PME et que ce type d'entreprise a de la difficulté à adopter des services cloud. L'une des raisons de cette adoption lente serait la complexité élevée des décisions à prendre lors de la sélection d'un fournisseur cloud. L'objectif de l'étude est de proposer un système d'aide à la décision qui pourra être utilisé par les PME. L'article propose 20 critères pour comparer les services cloud. Ceux-ci sont divisés en 5 catégories : fonctionnalité, sécurité et confidentialité, performance, utilisabilité et valeur économique. Le framework proposé est conçu pour être déployé dans le cloud afin que les décideurs puissent se connecter à celui-ci facilement sans avoir à installer de logiciels localement. Le framework stocke également des données à jour sur les critères des différents services cloud, les preneurs de décisions n'ont donc pas à faire des recherches des métriques de chaque fournisseur. Plusieurs tâches doivent être accomplies afin qu'un service soit recommandé : le décideur commence par se connecter à la plateforme et entre les informations pertinentes sur son entreprise comme le nombre d'employés et le secteur d'activité de celle-ci. L'utilisateur peut ensuite voir les métriques des fournisseurs cloud. Le système montre également les critiques d'autres utilisateurs concernant les fournisseurs cloud. Lorsqu'il veut avoir une recommandation, l'utilisateur choisit l'algorithme du modèle de décision, qui est AHP par défaut, les critères qui seront comparés et les fournisseurs qu'il veut comparer. Le décideur doit ensuite effectuer des comparaisons par paire pour chaque critère et pour chaque métrique de chaque fournisseur cloud. Le système fait ensuite le calcul des

meilleures alternatives pour le cas d'utilisation de l'entreprise avec l'algorithme sélectionné et montre les résultats au décideur qui peux finalement prendre une décision informée. L'article montre ensuite un exemple de son système en prenant comme cas d'utilisation une petite entreprise qui veut implémenter une solution de gestion de contenu d'entreprise dans le cloud.

Ce dernier article apporte une idée intéressante : l'implémentation de la base de données mise à jour régulièrement qui contient les différentes métriques des fournisseurs cloud. Cela faciliterait grandement la tâche des PME s'ils n'avaient pas à faire eux-mêmes la recherche de ces métriques dans les différents outils cloud. Par contre, il serait difficile de garder cette base de données à jour vu le nombre grandissant de services cloud. Le fait de s'appuyer sur des critiques d'autres utilisateurs peut être incertain, car chaque personne possède ses expériences personnelles. Les critères que cet article apporte sont pertinents pour notre étude et sont considérés pour notre modèle de décision.

2.5.3 SÉLECTION D'OUTILS ET DE PLATEFORMES D'APPRENTISSAGE AUTOMATIQUE

L'étude de Ruf et al. [85] fait l'analyse de 26 outils open-source utilisé pour au moins une des étapes d'un pipeline MLOps. Les auteurs décrivent en détail toutes les étapes d'un pipeline MLOps, qui sont similaires à celles que l'on retrouve dans la section 2.1.2. Cette étude analyse chaque outil individuellement et repère ce que chaque outil permet d'accomplir dans chacune des étapes du pipeline. Dans les caractéristiques qu'un outil peut avoir, on retrouve, entre autres, la capacité à gérer le prétraitement des données, la capacité à mesurer la qualité des données, la capacité à régler les hyperparamètres d'un modèle, la capacité à faire le contrôle de version du code, et plusieurs autres caractéristiques. L'étude présente ensuite un tableau comparatif de tous les outils étudiés qui permet d'avoir une vue d'ensemble de quelles combinaisons d'outils est possible pour la construction d'un pipeline MLOps. L'article ne

propose pas de méthode exacte pour la sélection du meilleur ensemble d'outils, mais suggère que de faire la sélection selon les besoins des développeurs, en favorisant une approche itérative qui ajoute un outil à la fois. L'étude continue avec un exemple étant l'automatisation partielle de la détection d'objet avec des outils MLOps en utilisant les suggestions soulignées.

Les caractéristiques proposées par ce dernier article sont de bas niveau, c'est-à-dire qu'il nécessite une connaissance avancée de l'apprentissage automatique et des différentes étapes d'un pipeline MLOps afin de pouvoir comparer les outils présentés. Contrairement à notre étude, on s'attend à ce que la personne qui fera la comparaison des outils soit un scientifique des données alors que nous supposons que le décideur aura des connaissances limitées du domaine de l'apprentissage automatique. Aussi, l'article ne propose pas de méthode précise pour faire la sélection d'un outil ou d'un ensemble d'outils, ce que proposons dans notre étude.

Ensuite, L'article de Kaymakci et al. [86] propose une méthode pour la sélection de services d'apprentissage automatique dans le cloud spécifiquement orienté vers les PME dans le domaine manufacturier. Les auteurs définissent 24 critères tirés de la littérature qu'ils séparent en six grandes catégories : sécurité, fiabilité, gestion dans le cloud, flexibilité, coût et performance. La première étape de leur méthode est d'évaluer les services selon les critères définis. L'algorithme AHP est utilisé avec une légère modification : une échelle de notation indépendante. Normalement, cet algorithme oblige l'utilisateur à faire une comparaison par paire de chaque service pour chaque critère, ce qui peut rapidement devenir encombrant avec plusieurs critères. La notation indépendante que l'article utilise permet d'éviter plusieurs de ces comparaisons, ce qui diminue la difficulté de la tâche et permet de rendre une décision plus rapidement. Après l'évaluation des services, les décideurs doivent comparer les critères en paires afin de poser l'importance de chacun de ces critères pour l'entreprise. Cette fois-ci, les comparaisons par paires sont effectuées comme il est décrit dans l'algorithme AHP originalement. Après toutes ces décisions, un résultat est obtenu pour chacun des services et

une décision finale peut être rendue. Les auteurs valident leur méthode à l'aide d'une étude de cas avec une entreprise de taille moyenne dans le domaine de la fabrication de pièces en métal.

L'article précédent est celui qui s'approche le plus de notre étude. Effectivement, il utilise une approche similaire et une variation de l'algorithme AHP pour la sélection du service qui essaie de simplifier le processus de décision. Les critères définis dans l'article sont également pris en compte pour notre étude. Par contre, l'étape où les décideurs doivent comparer chaque critère l'un avec l'autre peut être fastidieuse, car l'on retrouve 24 critères à comparer. Notre méthode prend soin de diminuer le nombre de critères à ceux que l'on considère comme essentiels afin de rendre le processus proposé rapide et facile. Une autre faiblesse de l'article est le manque de définitions précises des critères et de méthodes pour les mesurer, il peut donc être difficile de mesurer certains critères.

CHAPITRE III
APPROCHE PROPOSÉE

L'objectif de ce mémoire étant de proposer un modèle pour le choix d'une plateforme d'apprentissage automatique supervisé dans les petites entreprises, il y a plusieurs paramètres à établir afin de définir le modèle de sélection.

3.1 CRITÈRES DE SÉLECTION DES PLATEFORMES

Pour relever les caractéristiques des plateformes de type MLaaS, nous définissons certains critères de sélection déterminant les plateformes qui seront considérées pour le modèle de sélection. Ces critères ont été construits selon des besoins et défis pour l'implémentation d'un système de ML, spécifiques aux PME, et aux méthodes utilisées aujourd'hui mentionnés dans le chapitre 2. Les critères mentionnés sont une base afin de n'inclure que les plateformes qui sont viables pour les PME qui veulent se lancer dans le ML.

Premièrement, la plateforme doit offrir un service dans le cloud tel que défini par Mell et Grance [87] de l'institut national des normes et de la technologie (NIST). Ce critère est essentiel en raison du fait que notre approche est conçue pour les petites entreprises, qui n'ont souvent pas les moyens d'investir dans des infrastructures matérielles qui permettent d'héberger des plateformes ML sur des serveurs sur place. Les plateformes de type MLaaS permettent aux petites entreprises d'avoir la chance de se lancer dans de grands projets sans se heurter à une barrière d'entrée monétaire [38] que les PME ne peuvent franchir normalement en raison de leur budget limité [23]. L'utilisation d'un service dans le cloud a également une multitude d'avantages [88, 89, 57, 64] comme la simplicité d'administration et de maintenance, l'économie de ressources, un impact moins grand lors d'échecs ou d'améliorations dans le

service, une implémentation simplifiée, le paiement à l'utilisation, la scalabilité et l'élasticité du service. Toutes ces raisons rendent l'utilisation d'un service cloud pour les petites entreprises quasiment obligatoire pour avoir une chance de concurrencer avec les plus grandes entreprises.

Deuxièmement, on doit être capable d'implémenter un pipeline MLOps complet dans la plateforme sinon on doit avoir la possibilité d'ajouter des modules à la plateforme pour en être capable. Les étapes d'un pipeline MLOps ont été décrites dans la figure 2.2. Le déploiement manuel de systèmes ML entraîne des niveaux élevés de dette technique [34] en raison de l'enchevêtrement des données sur le code standard, MLOps est donc considéré comme la manière la plus efficace et fiable d'implémenter un système d'apprentissage automatique en production [17, 90] et permet d'automatiser toutes les étapes nécessaires au déploiement de tels systèmes. Cette automatisation est un grand avantage pour les petites entreprises qui n'ont pas beaucoup de ressources.

Troisièmement, la plateforme doit respecter les lois en vigueur sur la protection des données des utilisateurs et sur l'intelligence artificielle dans le dans lequel l'entreprise œuvre. Les lois évoluent constamment et rapidement dû au développement rapide de l'intelligence artificielle dans la vie quotidienne de la société [15]. L'exigence des plateformes à être conforme aux lois en vigueur retire une grande partie du poids qu'engendre le fait d'être toujours à jour face aux lois et permet aux PME de mettre leurs ressources sur les étapes fonctionnelles du cycle de vie ML.

Les plateformes seront recherchées dans la littérature grise récente à l'aide de recherches sur Google. On fait les recherches "mlaas platforms", "MLaaS", "Machine learning cloud" et "AutoML" sur Google et on garde tous les liens vers des blogues, articles ou sites web distincts qui comparent les plateformes MLaaS ou qui relèvent une ou des plateformes qui correspondent aux critères définis. Dans les trois premières pages de résultats, on retrouve

douze blogues et articles [91, 92, 93, 94, 95, 96, 97, 98, 99, 100, 101, 102] où sont cités des plateformes ML qui correspondent aux critères définis.

3.2 CARACTÉRISTIQUES DES PLATEFORMES MLAAS

Nous relevons ensuite les caractéristiques qui différencient et qui rassemblent les différentes plateformes MLaaS. On retrouve ces caractéristiques dans cinq articles scientifiques [80, 78, 83, 82, 84] trouvées de la manière suivante. Pour construire la requête, nous définissons 3 mots-clés généraux en rapport avec les définitions du chapitre 2 et les termes utilisés par les blogues et articles utilisés dans la section 3.1 : *cloud*, *service* et *evaluation*. Nous dérivons ensuite des synonymes et termes similaires afin de construire une requête qui sera utilisée dans les bases de données Google Scholar, ScienceDirect, SpringerLink, ResearchGate et IEEE Xplore. Pour tous les articles trouvés, on ne retient que les articles qui possèdent des caractéristiques «originales», à savoir les caractéristiques qui ne viennent pas directement d'un autre article. Cela évite de biaiser nos résultats en incluant des doublons. La figure 3.1 démontre nos résultats.

On retrouve deux types de caractéristiques : des caractéristiques mesurables et d'autres qui ne peuvent pas être mesurées, mais qu'on doit tout de même inclure, car elles ont un impact sur le choix de la plateforme. Les caractéristiques seront notées de 1 à 9 lors de l'évaluation des plateformes, selon les mesures et selon ce que les personnes qui jugent la plateforme croient être la meilleure évaluation. On ne retient que les caractéristiques qui sont mentionnées dans au moins quatre articles, ce qui nous laisse avec 4 caractéristiques pertinentes pour les petites entreprises qui veulent implémenter un système d'apprentissage automatique : réputation, coût périodique, disponibilité et portabilité. Le choix de ne garder que les caractéristiques qui font partie d'au moins 4 articles nous permet de ne garder que les caractéristiques pertinentes selon

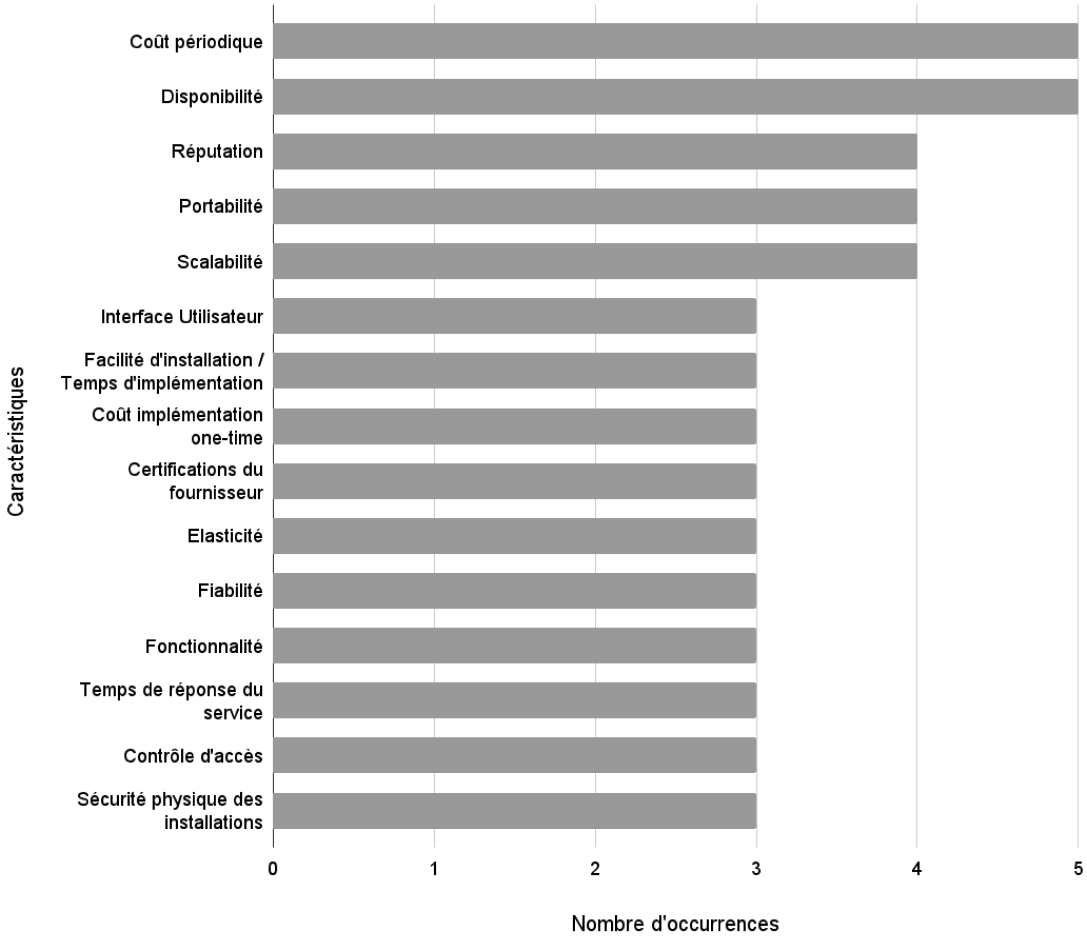


FIGURE 3.1 : Occurrence des caractéristiques avec plus de 3 occurrences retrouvées dans les articles scientifiques

la majorité des auteurs ce qui nous assure que les caractéristiques choisies sont viables. Cela a aussi pour effet de simplifier le modèle de sélection de plateformes ce qui est nécessaire pour les petites entreprises, qui comme mentionné dans la section 2.1.3, n'ont pas les moyens de tester tous les aspects des plateformes cloud offertes. Il est donc préférable pour ces entreprises de se concentrer sur les caractéristiques qui ont le plus d'importance. Ces caractéristiques seront maintenant expliquées plus en détail.

Réputation du fournisseur du service cloud La réputation de l'entreprise qui héberge et fournit le service dans le cloud. Cette caractéristique peut se mesurer de façon objective et/ou subjective. L'expérience passée de la personne qui fait l'évaluation des caractéristiques peut suffire pour l'évaluation de cette caractéristique. On peut également prendre en compte les parts de marché de chacun des fournisseurs, cela donne une bonne idée de la confiance que le public accorde aux entreprises et du nombre de clients que celle-ci fournit. La valeur de la marque est aussi importante, car certaines personnes pourraient préférer un nouveau produit d'un vendeur bien connu à un produit qui voit beaucoup d'utilisation provenant d'une compagnie moins connue [83].

Coût périodique pour l'abonnement au service Coût pour l'abonnement au service par période. Pour évaluer ce critère, on doit choisir un abonnement comparable pour tous les fournisseurs qui seront comparés. Pour les services qui facturent à l'utilisation, on estime au mieux le coût selon l'utilisation potentielle du système d'apprentissage automatique que l'entreprise veut mettre en place.

Disponibilité Disponibilité des serveurs dans le cloud donné par la formule suivante [81] :

$$\frac{\text{(temps de service) + (temps durant lequel le service n'était pas disponible)}}{\text{temps de service}}$$

On parle non seulement de la disponibilité du côté de temps processeur pour l'entraînement des modèles, mais aussi pour les endpoints d'inférence que les clients appellent. Cette mesure est souvent fournie par les fournisseurs cloud.

Portabilité Facilité avec laquelle l'entreprise peut migrer le système implémenté sur une autre plateforme ou environnement. Cela permet à l'entreprise d'éviter d'être vérouillé à un seul fournisseur, c'est-à-dire de ne plus être capable de changer de plateforme ou d'environnement dans le futur, car le coût que cela engendre est trop élevé comparativement aux bénéfices, ce qui essentiellement retient l'entreprise avec la plateforme initiale. On peut mesurer cette caractéristique de manière qualitative en plusieurs facettes : la facilité de trouver les fonctions nécessaires à la migration dans l'interface de la plateforme, le nombre de formats différents dans lequel on peut exporter les données ou n'importe quel autre fichier utile dans le système ou l'utilisation de formats standardisés dans les différents services du fournisseur cloud. D'autres mesures peuvent être ajoutées selon les besoins de l'entreprise qui teste les fournisseurs.

3.3 MÉTHODE POUR LA SÉLECTION D'UN SERVICE D'APPRENTISSAGE AUTOMATIQUE

Le problème de sélection de fournisseurs cloud a déjà été étudié par des chercheurs comme nous l'avons vu dans la section 2.5. Une critique des articles scientifiques existants sur le sujet [78] en vient à la conclusion que la méthode la plus utilisée et considérée comme la plus fiable par les chercheurs est la procédure hiérarchique d'analyse (AHP), qui est décrite en détail par son créateur Saaty [103].

Nous utiliserons une version simplifiée de AHP, nommé AHP-express par son auteur Leal [104]. Elle permet d'obtenir de bons résultats tout en diminuant substantiellement le

Description	Priorité
L'élément est aussi important que l'élément comparé	1
L'élément est modérément moins important que l'élément comparé	3
L'élément est fortement moins important que l'élément comparé	5
L'élément est très fortement moins important que l'élément comparé	7
L'élément est extrêmement moins important que l'élément comparé	9

TABLEAU 3.1 : Échelle d'importance pour la comparaison des éléments

nombre de comparaisons nécessaires entre les caractéristiques des plateformes. La méthode se résume en six étapes :

Étape 1 : Déterminer l'objectif du processus de décision et créer un arbre hiérarchique avec les différentes caractéristiques à comparer et leurs sous-caractéristiques si nécessaire.

Étape 2 : Pour chaque niveau intermédiaire, déterminer la caractéristique qui a le plus d'importance avec laquelle les autres caractéristiques seront comparées. Les autres caractéristiques sont ensuite comparées avec la plus importante à l'aide d'une échelle se retrouvant dans le tableau 3.1.

Étape 3 : Pour chaque caractéristique, faire les observations nécessaires sur les différentes alternatives et identifier la meilleure. Les autres alternatives sont ensuite comparées avec la meilleure et sont notées avec la même échelle que l'étape précédente.

Étape 4 : À la suite de l'étape 3, la formule 3.1 est appliquée pour calculer les priorités locales sur chacun des vecteurs obtenus lors des étapes précédentes. Les éléments de la formule sont définis comme suit : j est l'élément pour lequel on veut calculer la priorité, i est l'élément avec le plus d'importance qui est choisi comme base de comparaison, a_{ij} est la valeur de la comparaison de l'alternative i avec l'alternative j et pr_j est la priorité de l'alternative j par rapport au critère considéré.

$$pr_j = \frac{1}{a_{ij} * \sum_k \frac{1}{a_{ik}}} \tag{3.1}$$

Étape 5 : Calculer les priorités globales en remontant l'arbre hiérarchique jusqu'au premier niveau.

Étape 6 : Classer les alternatives à l'aide des résultats du vecteur de priorité du premier niveau de l'arbre.

3.4 EXEMPLE ILLUSTRATIF DE LA MÉTHODOLOGIE

Cette section présente un exemple des étapes de notre méthodologie et des calculs nécessaires afin d'arriver à une décision quant au choix d'une plateforme convenable.

3.4.1 DÉFINITION DE L'OBJECTIF

Pour illustrer la méthodologie de sélection des plateformes, nous considérons trois plateformes fictives (I, J, K) évaluées selon les critères suivants définis précédemment, soit : Réputation du fournisseur (*R*), Coût (*C*), Disponibilité (*D*) et Portabilité (*P*). L'arbre de décision de cet exemple est illustré dans la figure 3.2. L'objectif est de sélectionner une plateforme convenable pour le cas d'utilisation de l'entreprise fictive.

3.4.2 DÉFINITION DES PRIORITÉS DES CARACTÉRISTIQUES

Ensuite, il faut déterminer la caractéristique qui a le plus d'importance parmi toutes les caractéristiques. Dans notre exemple, le décideur choisit la disponibilité comme la plus importante. Il compare ensuite les autres caractéristiques une à une avec la plus importante à

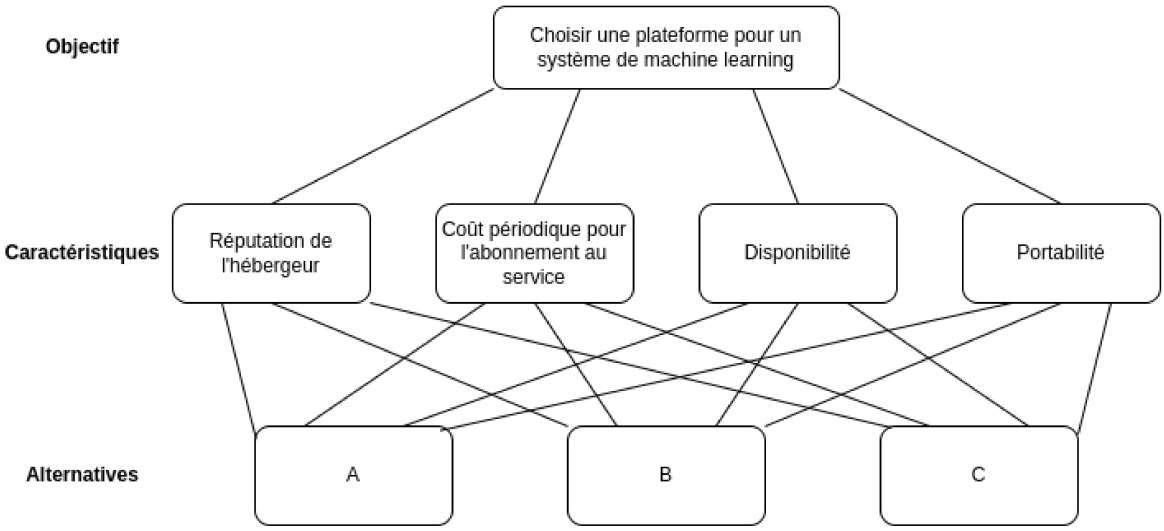


FIGURE 3.2 : Arbre pour la méthode AHP-express

l'aide de l'échelle du tableau 3.1. Ces choix se basent sur les définitions des caractéristiques fournies précédemment et selon les recherches effectuées par le décideur. Pour notre exemple, le décideur place les priorités 3, 4 et 5 sur les autres caractéristiques, soit la réputation du fournisseur (R), le coût (C) et la portabilité (P) respectivement. On attribue la priorité 1 à la caractéristique la plus importante, soit la disponibilité (D). On peut représenter ces priorités sous la forme d'un vecteur V :

$$V = \begin{bmatrix} R & C & D & P \end{bmatrix} = \begin{bmatrix} 3 & 4 & 1 & 5 \end{bmatrix}$$

3.4.3 COMPARAISON DES PLATEFORMES

On évalue ensuite chaque caractéristique de chaque plateforme. Par exemple, pour la réputation de l'hébergeur, le décideur évalue chaque plateforme sur cette caractéristique et choisit laquelle des plateformes est la meilleure pour ce critère. Les autres plateformes sont ensuite comparées avec la meilleure de la même manière que précédemment. Cela est fait

pour chaque caractéristique. On peut ensuite élaborer une matrice M, où les éléments C_p sont des priorités attribuées par le décideur aux caractéristiques des plateformes, où C représente la caractéristique et p est la plateforme. On remplit la matrice avec des valeurs arbitraires pour notre exemple :

$$M = \begin{bmatrix} R_A & R_B & R_C \\ C_A & C_B & C_C \\ R_A & R_B & R_C \\ P_A & P_B & P_C \end{bmatrix} = \begin{bmatrix} 1 & 5 & 3 \\ 8 & 1 & 1 \\ 5 & 1 & 6 \\ 1 & 2 & 3 \end{bmatrix}$$

3.4.4 CALCUL DES PRIORITÉS LOCALES

On doit ensuite appliquer la formule 3.1 à chaque élément de chaque vecteur. Par exemple, pour le vecteur des priorités des caractéristiques, la formule s'applique comme suit :

$$V = \begin{bmatrix} 3 & 4 & 1 & 5 \end{bmatrix}$$
$$V = \begin{bmatrix} \frac{1}{3 \cdot \sum_k \frac{1}{a_{ik}}} & \frac{1}{4 \cdot \sum_k \frac{1}{a_{ik}}} & \frac{1}{1 \cdot \sum_k \frac{1}{a_{ik}}} & \frac{1}{5 \cdot \sum_k \frac{1}{a_{ik}}} \end{bmatrix}$$
$$V = \begin{bmatrix} \frac{1}{3 \cdot 1,78\dot{3}} & \frac{1}{4 \cdot 1,78\dot{3}} & \frac{1}{1 \cdot 1,78\dot{3}} & \frac{1}{5 \cdot 1,78\dot{3}} \end{bmatrix}$$
$$V = \begin{bmatrix} 0,1869158879 & 0,1401869159 & 0,5607476636 & 0,1121495327 \end{bmatrix}$$

On répète l'opération sur chaque ligne de la matrice M :

$$M = \begin{bmatrix} 1 & 5 & 3 \\ 8 & 1 & 1 \\ 5 & 1 & 6 \\ 1 & 2 & 3 \end{bmatrix}$$

$$M = \begin{bmatrix} \frac{1}{1,5\dot{3}} & \frac{1}{5 \cdot 1,5\dot{3}} & \frac{1}{3 \cdot 1,5\dot{3}} \\ \frac{1}{8 \cdot 2,125} & \frac{1}{2,125} & \frac{1}{2,125} \\ \frac{1}{5 \cdot 1,3\dot{6}} & \frac{1}{1,3\dot{6}} & \frac{1}{6 \cdot 1,3\dot{6}} \\ \frac{1}{1,8\dot{3}} & \frac{1}{2 \cdot 1,8\dot{3}} & \frac{1}{3 \cdot 1,8\dot{3}} \end{bmatrix}$$

$$M = \begin{bmatrix} 0,65217391304 & 0,13043478263 & 0,21739130439 \\ 0,05882352941 & 0,47058823529 & 0,47058823529 \\ 0,14634146342 & 0,7317073171 & 0,12195121951 \\ 0,54545454545 & 0,27272727272 & 0,18181818181 \end{bmatrix}$$

On multiplie ensuite le vecteur V avec la matrice M pour obtenir le vecteur final F, avec des éléments S_p qui représentent le score de la plateforme p, qui permettent au décideur de prendre une décision sur une plateforme convenable pour l'objectif de l'entreprise :

$$F = \begin{bmatrix} S_A & S_B & S_C \end{bmatrix} = V \times M$$

$$F = \begin{bmatrix} 0,2733810613 & 0,5312400513 & 0,1953788876 \end{bmatrix}$$

3.4.5 RÉSULTATS

On interprète les résultats du vecteur F comme suit : plus le chiffre est grand, plus la plateforme est considérée comme convenable pour le cas de l'entreprise. Dans l'exemple, la plateforme considérée comme la plus convenable est la plateforme B, suivi de la plateforme A et finalement de la C. Le décideur pourrait s'appuyer sur ces résultats pour prendre une décision sur la plateforme à privilégier.

3.5 CONCLUSION

Ce chapitre a présenté une méthodologie structurée pour évaluer et sélectionner des plateformes d'apprentissage automatique adaptées aux PME. Le prochain chapitre présente l'application de cette méthodologie dans le cadre d'un cas d'utilisation réel réalisé en coopération avec une petite entreprise.

CHAPITRE IV
APPLICATION DE L'APPROCHE PROPOSÉE

Dans ce chapitre, nous présentons une étude de cas qui utilise le modèle de décision défini dans le chapitre 3 pour implémenter un système d'apprentissage automatique. Dans la section 4.1, nous posons les paramètres de l'étude de cas et définissons le problème à résoudre. La section 4.2 présente les calculs effectués pour arriver à une décision quant au choix de la plateforme cloud d'apprentissage automatique à utiliser lors de l'implémentation du système ML. La section 4.3 présente les résultats et les détails de l'implémentation du système d'apprentissage automatique dans l'application de l'entreprise et finalement la section 4.4 discute des résultats obtenus.

4.1 CONFIGURATION ET PARAMÈTRES

L'étude de cas a été réalisée en coopération avec l'entreprise Econochef, une petite entreprise encore en phase de démarrage qui développe une application mobile visant les employés d'entreprises, et qui propose des recettes selon les rabais de la semaine dans les circulaires des épiceries. L'entreprise emploi environ 5 employés incluant deux développeurs logiciels. Le preneur de décision est un de ces développeur logiciel qui possède plus de connaissances dans les fournisseurs cloud. L'entreprise recherche principalement une solution abordable pour se lancer dans le monde de l'apprentissage automatique et valider son cas d'utilisation.

L'entreprise récupère des données sur les interactions des utilisateurs avec l'application à l'aide de Google Analytics. Dans notre cas d'utilisation, nous utiliserons les données concernant les évènements où l'utilisateur ajoute une recette dans ses favoris. On retrouve

Nom de l'attribut	Description
event_timestamp	Horodotage Unix de l'époque actuelle lorsque l'évènement à eu lieu
user_id	Numéro d'identification de l'utilisateur qui a déclenché l'évènement dans la base de données
recipe_id	Numéro d'identification de la recette dans la base de données

TABLEAU 4.1 : Description des attributs de l'ensemble de données des évènements utilisateur

un ensemble de données avec les évènements utilisateurs et leur horodatage et un autre ensemble de données contenant les informations des recettes. La description de ces ensembles de données se retrouve dans les tableaux 4.1 et 4.2.

Le cas d'utilisation est le suivant : l'entreprise veut implémenter un système de recommandation de recettes aux utilisateurs à l'aide des données décrites ci-dessus. Les étiquettes des données peuvent prendre plusieurs formes, comme le nombre de clics sur une recette dans l'application par les utilisateurs ou le nombre de secondes que les utilisateurs passent sur une recette. Les plateformes cloud considérées dans le modèle de sélection seront limitées à trois, identifiées grâce aux critères de parts de marché [105]. Cela réduit la complexité du modèle de sélection et simplifie la tâche pour les décideurs. Ces plateformes sont Amazon Cloud Services (AWS) [7], Microsoft Azure [6] et Google Cloud Platform (GCP) [8].

4.2 APPLICATION DU MODÈLE DE DÉCISION

Pour la sélection d'un service d'apprentissage automatique, nous suivons les étapes définies dans la section 3.3.

Nom de l'attribut	Description
id	Numéro d'identification de la recette dans la base de données
title	Nom de la recette
created_at	Horodotage Unix du moment où la recette a été ajoutée dans la base de données
notes	Annotations spéciales pour la recette
tips	Trucs et astuces pour la préparation de la recette
source	Provenance de la recette
fiber	Quantité de fibres présente dans la recette en grammes
protein	Quantité de protéines présente dans la recette en grammes
cooking_time	Temps de cuisson lors de la préparation de la recette
preparation_time	Temps de préparation de la recette
price	Prix des ingrédients de la recette
portion	Nombre de portions obtenues lors de la préparation de la recette
image	Lien vers l'image de la recette
difficulty	Difficulté avec laquelle la recette peut être préparée allant de 1 à 4
category	Liste de catégories dont la recette fait partie. Les valeurs possibles sont : Entrée et accompagnement, Facile et rapide, Brunchs et déjeuners, Boîte à lunch, Desserts, International / Cuisine du monde, Végétarien, Salades, Soupes et potages, Pâtes, Sandwichs et wraps, Collation, Plats principaux, Sauces et vinaigrettes et Évènement festif
foodtype	Liste de types d'alimentations dont la recette fait partie. Les valeurs possibles sont : Sans noix / arachides, Riche en protéines, Végétarien, Riche en fruits et légumes, Sans lactose et Riche en fibres
proteintype	Liste de protéines que l'on retrouve dans la recette. Les valeurs possibles sont : Poisson, Porc, Volaille, Veau, Boeuf, Fruits de mer, Légumineuses, Viande sauvage, Agneau, Viande chevaline, Canard et Gibier, Oeufs, Produits laitiers et substituts de viande, Aucun, Noix et graines et Tofu, Fèves de soya et Boissons de soya

TABLEAU 4.2 : Description des attributs de l'ensemble de données des recettes

4.2.1 DÉFINITION DE L'OBJECTIF ET CONSTRUCTION D'UN ARBRE HIÉRARCHIQUE

La première étape consiste en la définition de l'objectif et la création d'un arbre hiérarchique contenant les différentes caractéristiques et les choix de services possibles. La figure 4.1 montre cet arbre.

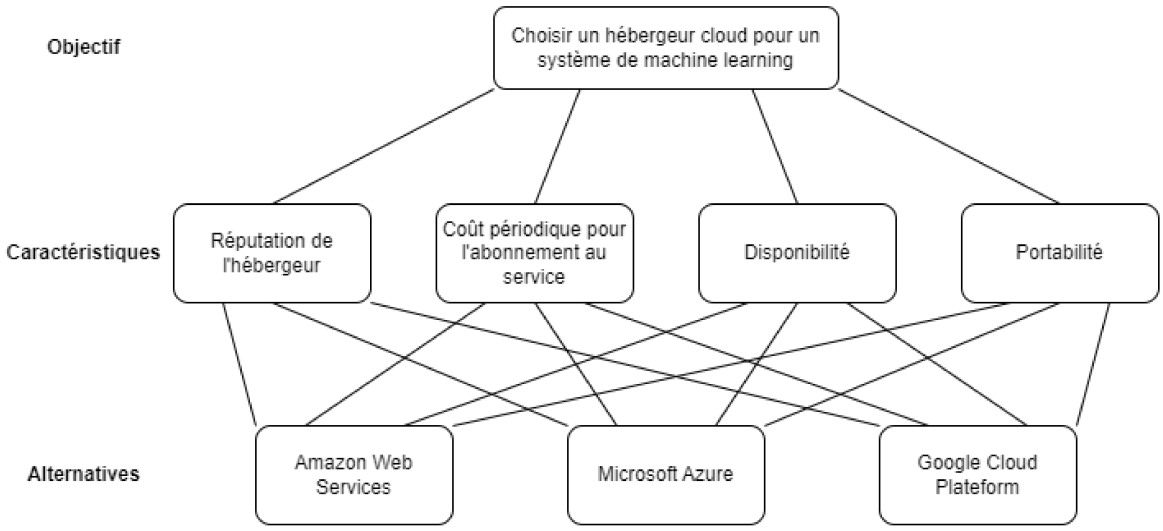


FIGURE 4.1 : Arbre pour le cas d'utilisation d'Econochef pour la méthode AHP-express

Le premier niveau de cet arbre consiste en l'objectif pour ce processus de décision. Dans le cas échéant, cet objectif est de faire le choix d'un hébergeur cloud où l'entreprise pourra implémenter son système d'apprentissage automatique.

Le deuxième niveau contient les caractéristiques définies dans la section 3.2 et le troisième niveau les alternatives, c'est-à-dire les fournisseurs cloud pris en compte lors des décisions pour atteindre l'objectif.

Caractéristique	Priorité
Réputation de l'hébergeur	4
Disponibilité	3
Portabilité	5

TABLEAU 4.3 : Priorités des caractéristiques en utilisant le coût pour l'abonnement au service comme point de comparaison

4.2.2 DÉFINITION DES PRIORITÉS DES CARACTÉRISTIQUES DES PLATEFORMES

La deuxième étape consiste à déterminer la caractéristique qui a le plus d'importance pour le niveau intermédiaire de l'arbre, soit celui comportant les caractéristiques des plateformes.

Le décideur a déterminé selon les besoins et ressources de l'entreprise que la caractéristique qui a le plus d'importance pour l'implémentation du système est le coût périodique pour l'abonnement au service.

Les autres caractéristiques sont ensuite comparées avec celle qui a le plus d'importance selon l'échelle du tableau 3.1. Tandis que le tableau 4.3 décrit les priorités définies par le preneur de décision.

Les priorités choisies sont les suivantes : le coût périodique pour l'abonnement au service est le plus important, la disponibilité est modérément moins importante que le coût, la réputation de l'hébergeur est modérément fortement moins importante que le coût et la portabilité est fortement moins importante que le coût de l'abonnement. Le coût est sélectionné comme le plus important car l'entreprise est en recherche de financement actuellement et ne peut se permettre une solution coûteuse. Le cas d'utilisation étant une preuve de concept, la disponibilité est choisie dans ce rang car le nombre d'utilisateurs d'Econochef est limité pour

l'instant mais l'entreprise veut tout de même pouvoir rester sur cette plateforme lorsqu'ils auront plus d'utilisateurs. Ensuite, le réputation de l'hébergeur est modérément importante pour l'entreprise pour le moment ainsi que la portabilité car l'entreprise ne souhaite pas déménager de plateforme à court terme.

4.2.3 ÉVALUATION DES CARACTÉRISTIQUES DES PLATEFORMES

La troisième étape du processus consiste en l'évaluation de chacune des caractéristiques pour chacune des plateformes selon les lignes directrices décrites dans la section 3.2.

COÛT PÉRIODIQUE POUR L'ABONNEMENT À LA PLATEFORME

En commençant par le coût, le preneur de décision relève l'information nécessaire sur cette caractéristique, soit le coût estimé pour l'implémentation d'un système ML dans le cas d'utilisation de l'entreprise. Pour obtenir une estimation du coût, la quantité de données, le nombre d'heures d'entraînement du modèle et le nombre de requêtes d'inférence doivent être pris en compte dans le calcul.

Le preneur de décision estime la quantité de données à moins de 100 Mo par mois, le nombre d'heures d'entraînement par mois à une et le nombre de requêtes par mois à 5000. On met à jour les données d'évènements utilisateurs une fois par jour et on entraîne le modèle une fois par semaine lorsque possible. On estime le coût périodique de chacune des plateformes selon ces paramètres à l'aide des informations sur les prix de chaque service et des calculateurs de prix que l'on retrouve sur le site web du fournisseur. Les résultats se trouvent dans le tableau 4.4.

Plateforme	Coût total par mois
Amazon Web Services	270,25$
Google Cloud Platform	3,85$
Microsoft Azure	144,48$

TABLEAU 4.4 : Coût total estimé en dollars américains par mois pour la mise en place d'un système d'apprentissage automatique pour le cas d'utilisation d'Econochef dans chacune des plateformes

Alternative	Priorité par rapport à GCP
Amazon Web Services	9
Microsoft Azure	7

TABLEAU 4.5 : Priorités du coût pour l'abonnement au service en utilisant Google Cloud Platform comme point de comparaison

Les grandes différences de prix entre les plateformes sont expliquées par le fait que AWS et Microsoft Azure utilisent un système de facturation à l'heure et on doit donc compter toutes les heures d'un mois, soit environ 720 heures, et ce même si aucun utilisateur n'utilise le service pendant plusieurs heures. Ces plateformes sont plutôt pensées pour les économies à grande échelle et plus avantageuses dans de plus grosses entreprises.

Inversement, GCP utilise plutôt une facturation qui charge au nombre de requêtes, ce qui est avantageux pour les plus petites entreprises qui ont peu de requêtes à servir. Cette plateforme est donc nettement plus avantageuse dans le cas d'Econochef.

Le preneur de décision pose ensuite la meilleure alternative comme étant Google Cloud Platform et compare les autres alternatives à celle-ci, de manière similaire à la définition des priorités de la section 4.2.2. Les résultats se retrouvent dans la figure 4.5.

Alternative	Priorité par rapport à AWS
Google Cloud Platform	5
Microsoft Azure	3

TABLEAU 4.6 : Priorités de la réputation de l'hébergeur du service en utilisant Amazon Web Services comme point de comparaison

RÉPUTATION DE L'HÉBERGEUR DU SERVICE

Pour la suite, le preneur de décision relève les informations concernant la réputation de l'hébergeur. Il prend en compte les parts de marché de chacune des plateformes pour établir les priorités de cette caractéristique. Les parts de marché de chacune des plateformes ont été sondées par Synergy Research Group en 2023 [105] et sont les suivantes : AWS détient environ 33%, GCP environ 11% et Microsoft Azure environ 23%.

Les services cloud d'Amazon sont donc les plus utilisés, et Amazon est donc choisi comme meilleure alternative pour la caractéristique de réputation de l'hébergeur du service. Les autres priorités posées par le preneur de décision sont représentées dans le tableau 4.6.

DISPONIBILITÉ

Pour la disponibilité, l'information se retrouve dans les détails des contrats de service. Les fournisseurs promettent un certain niveau de disponibilité et offrent un rabais lorsque celui-ci n'est pas respecté. La valeur de disponibilité promise dans le contrat sera retenue dans le cadre de l'évaluation de la plateforme.

Chacun des contrats des trois plateformes évaluées promet des disponibilités de 99,9% pour les services d'apprentissage automatique souhaités pour l'implémentation du système

Alternative	Priorité
Amazon Web Services	1
Google Cloud Platform	1
Microsoft Azure	1

TABLEAU 4.7 : Priorités pour la disponibilité

ML. En cas d'égalité, on pose la valeur pour la priorité à un pour chacune des plateformes, le tableau 4.7 présente les résultats.

PORTABILITÉ

Pour terminer, la portabilité est évaluée selon les critères définis dans la section 3.2.

Les données portant sur les utilisateurs sont hébergées ailleurs que sur les services des fournisseurs et cela n'est donc pas un problème pour la portabilité dans le cas d'Econochef. Seule la portabilité des modèles et des configurations est donc évaluée. Les informations sont tirées de la documentation et des tutoriels présents sur les sites web des fournisseurs.

Pour ce qui est d'AWS, l'utilisateur doit écrire du code dans un notebook Jupyter en Python, un format bien connu du domaine d'apprentissage automatique. Ce code est facilement exportable et exécutable à l'externe, mais l'entraînement du modèle se faisant dans les serveurs d'Amazon de manière opaque, il n'est pas possible d'entraîner un modèle sur une machine autre que celles des services d'Amazon. Les modèles sont des recettes prédéfinies qui ne possèdent que quelques hyperparamètres modifiables. La documentation ne mentionne pas la possibilité d'exporter des modèles pour l'utilisation externe.

Ensuite, Google Cloud Platform utilise une approche low-code. On ne retrouve aucun code pour la création, l'entraînement et le déploiement d'un modèle de recommandation. Le

Alternative	Priorité par rapport à GCP
Amazon Web Services	6
Microsoft Azure	7

TABLEAU 4.8 : Priorités de la portabilité en utilisant Google Cloud Platform comme point de comparaison

seul code nécessaire est celui pour l'importation de données dans le système. Cela apporte l'avantage qu'il est plus facile pour les personnes ayant peu d'expérience avec l'apprentissage automatique de créer des modèles. Par contre, il est difficile, voire impossible, de réentraîner le modèle ailleurs que dans la plateforme d'origine. Une section dans la documentation [106] décrit une méthode pour l'exportation d'un modèle dans cinq formats différents pour l'utilisation externe sur plusieurs plateformes.

Pour finir, Microsoft Azure utilise une approche no-code pour l'implémentation d'un système de recommandations. L'entraînement du modèle se fait de manière totalement automatique et opaque, sans possibilité de modifier les hyperparamètres. La documentation ne mentionne pas la possibilité d'exporter un modèle pour utilisation externe.

Après ces recherches, le preneur de décision définit les priorités pour la caractéristique de portabilité. Il choisit Google Cloud Platform comme la meilleure alternative qui sera utilisée comme point de comparaison avec les autres alternatives. Le tableau 4.8 affiche les résultats de ces comparaisons.

4.2.4 CALCUL DES PRIORITÉS LOCALES

La quatrième étape consiste en le calcul des valeurs des priorités locales de chaque niveau excepté celui à la racine. La formule 3.1 est utilisée pour ce calcul.

Réputation	Coût abonnement	Disponibilité	Portabilité
0,1401869159	0,5607476636	0,1869158879	0,1121495327

TABLEAU 4.9 : Valeurs de priorités pour les caractéristiques des plateformes

En utilisant les valeurs posées par le preneur de décision dans le tableau 4.3, on remplace les valeurs dans l'équation 3.1.

Pour simplifier, la formule peut être décortiquée en trouvant la valeur de $\sum_k \frac{1}{a_{ik}}$ sachant que a_{ij} est la valeur de la comparaison de l'alternative i avec l'alternative j :

$$\sum_k \frac{1}{a_{ik}} = \frac{1}{1} + \frac{1}{4} + \frac{1}{3} + \frac{1}{5} = \frac{107}{60} = 1,78\dot{3}$$

Les priorités locales peuvent ensuite être calculées à l'aide de la formule suivante :

$$pr_j = \frac{1}{a_{ij} \cdot \sum_k \frac{1}{a_{ik}}} = \frac{1}{a_{ij} \cdot \frac{107}{60}}$$

La formule précédente est appliquée pour chacune des caractéristiques et les résultats sont placés sous forme de vecteur dans le tableau 4.9. L'opération précédente est répétée pour calculer la valeur de priorité de chacune des alternatives dans chaque caractéristique. Les résultats se trouvent dans le tableau 4.10.

4.2.5 CALCUL DES PRIORITÉS GLOBALES

La priorité globale d'une caractéristique d'une alternative est égale à la multiplication des priorités locales de la caractéristique et des alternatives. Cela revient à faire une multiplication

	Google Cloud Platform	Amazon Web Services	Microsoft Azure
Coût abonnement	0,7974683544	0,08860759494	0,1139240506
Réputation	0,1304347826	0,652173913	0,2173913043
Disponibilité	0,3333333333	0,3333333333	0,3333333333
Portabilité	0,7636363636	0,1272727273	0,1090909091

TABLEAU 4.10 : Tableau des valeurs de priorités pour les alternatives pour chaque caractéristique

Google Cloud Platform	Amazon Web Services	Microsoft Azure
0,6134105237	0,2176916242	0,1688978521

TABLEAU 4.11 : Valeurs de priorités globales des alternatives

entre le vecteur du tableau 4.9 et la matrice que l'on retrouve dans le tableau 4.10. Le vecteur résultant se trouve dans le tableau 4.11.

4.2.6 CLASSEMENT DES ALTERNATIVES ET INTERPRÉTATION DES RÉSULTATS

Les valeurs du vecteur des priorités globales sont ensuite utilisées afin de classer les alternatives. On peut observer que Google Cloud Platform est l'alternative dominante avec une valeur d'environ 0,61, Amazon Web Services est en deuxième position avec une valeur d'environ 0,22 et finalement Microsoft Azure avec une valeur d'environ 0,17.

La grande différence de valeurs entre GCP et les deux autres alternatives indique clairement que GCP est l'alternative la mieux adaptée selon les critères du preneur de décision.

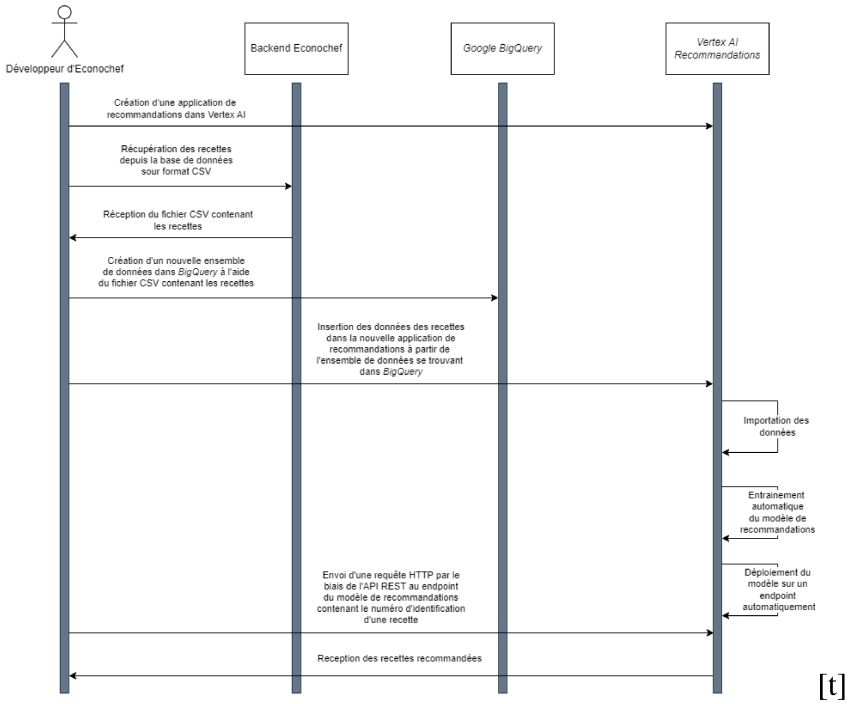
4.3 IMPLÉMENTATION DU SYSTÈME

La ressource de Google Cloud Platform qui est utilisée par l'entreprise est Vertex AI Search and Conversation. Cette ressource permet d'implémenter des IA conversationnelles, des suggestions de recherche personnalisées et des systèmes de recommandations. Le cas d'utilisation d'Econochef étant de construire un système de recommandations simple nécessitant peu de connaissances en apprentissage automatique, cette ressource est donc très bien adaptée.

Plusieurs types de modèles de recommandation sont disponibles, celui qui remplit les besoins de l'entreprise s'appelle *Generic Recommendations*. Cela permet d'obtenir des recommandations pour le contenu qui n'est pas des vidéos, des nouvelles ou de la musique. Le système implémenté fait des recommandations pour des recettes similaires et ne prend pas en compte les évènements utilisateur pour l'instant. Cela est dû au fait qu'il faut un historique d'au moins un million d'évènements, ce qui n'est pas encore le cas d'Econochef. Cependant, cette partie n'est pas difficile à ajouter une fois un nombre suffisant d'évènements atteint.

L'implémentation du système est assez simple. En premier lieu il faut importer les données dans des recettes dans le système. Cela peut se faire manuellement en important les données d'abord dans BigQuery, la plateforme d'entrepôts de données de Google, ou bien par le biais de l'API. On peut donc facilement automatiser l'importation de données en appelant l'API lors de l'insertion de nouvelles recettes dans la base de données d'Econochef.

L'importation de nouvelles données prend quelques minutes et ensuite l'entraînement du modèle commence. Cela peut prendre quelques heures avant de pouvoir faire des recommandations. Il est à noter que l'entraînement du modèle se fait de façon totalement opaque et qu'il est impossible de modifier les paramètres d'entraînement ou d'optimisation de celui-ci. Cela a pour effet de simplifier grandement le processus pour les personnes inexpérimentées



[t]

FIGURE 4.2 : Diagramme de séquence pour l'implémentation d'un système d'apprentissage automatique de recommandation de recettes dans Vertex AI pour Econochef

en apprentissage automatique mais pourrait avoir des impacts négatifs sur la performance du modèle. Le réentraînement se fait automatiquement et donc le modèle reste toujours à jour.

Une fois l'entraînement terminé, le modèle est automatiquement déployé sur un endpoint et il est ensuite possible d'inférer des recommandations. Une section *Preview* dans la console permet de faire de tests afin de voir des résultats préliminaires des recommandations. Le endpoint de recommandation nécessite le numéro d'identification de la recette dont on veut des résultats similaires et retourne ensuite les numéros d'identification de recettes recommandées, on peut ensuite aller chercher les informations nécessaires sur les recettes dans la base de données d'Econochef.

La figure 4.2 donne une vue d'ensemble du processus complet.

4.4 DISCUSSION

Dans cette section, nous discutons des résultats obtenus lors de la sélection d'un fournisseur de service dans la section 4.2 et de l'implémentation du système dans la section 4.3.

4.4.1 AVANTAGES DU MODÈLE PROPOSÉ

L'objectif défini dans la section 1.3 était de proposer un modèle de décision pour la sélection d'une plateforme supportant l'apprentissage automatique pour les petites entreprises afin de déployer un pipeline ML robuste. Il faut mentionner que notre étude est préliminaire et que d'autres travaux sont nécessaires pour confirmer notre hypothèse. Actuellement, nous évaluons notre approche comme l'a fait [85, 86, 80, 81, 83, 84], simplement en se fiant au fait que l'entreprise a pu implémenter un système d'apprentissage automatique et qu'il fonctionne.

Le modèle de décision proposé a pu permettre l'implémentation d'un système complet d'apprentissage automatique, du traitement des données jusqu'au déploiement et au réentrainement en continu du modèle. La méthode pour la sélection de la plateforme la mieux adaptée remplit l'objectif pour le cas d'utilisation de l'entreprise en question, tout en étant simple d'utilisation.

De plus, il est possible d'évaluer les plateformes à l'aide des caractéristiques choisies sans implémenter un système complet dans chacune de ces plateformes. Cela permet d'économiser du temps, qui est précieux dans les petites entreprises ayant des ressources limitées. Le fait de concentrer les efforts sur les caractéristiques principales, celles qui sont les plus citées dans les articles scientifiques, simplifie davantage la tâche du preneur de décisions. Tous les calculs qui doivent être effectués lors de l'utilisation du modèle de décision peuvent facilement être automatisés dans une feuille de calcul.

Le modèle de décision proposé diffère de ceux qui sont mentionnés dans la section 2.5. De fait, la plupart des articles qui proposent un modèle décisionnel le font de manière complexe, avec des dizaines de caractéristiques à évaluer en utilisant un algorithme complexe pour le calcul des priorités. Seulement dans [86] on retrouve une approche dédiée aux PME dans le domaine industriel, mais celle-ci utilise 24 caractéristiques différentes pour l'évaluation du modèle. De plus, l'approche proposée dans notre papier utilise le processus de sélection AHP-express, ce qui simplifie davantage la sélection d'une plateforme. Ce processus relativement nouveau n'avait pas encore été utilisé pour la sélection de fournisseurs cloud.

Le système de recommandation a pu être implémenté dans l'entreprise sans embûches. En effet, en raison de la nature automatisée des outils AutoML sur les plateformes considérées, les tâches à effectuer pour la création de l'application de recommandation ne sont pas liées à l'apprentissage automatique, autrement dit, la personne responsable n'a pas besoin de connaissances dans le domaine de l'intelligence artificielle. Cela est vrai pour toutes les plateformes considérées et devrait être un critère pour les petites entreprises sans expérience en apprentissage automatique. Le savoir nécessaire à la création du système de recommandation ne se limite qu'à certaines connaissances des bases de données SQL et des APIs REST. Ces notions sont nécessairement connues par au moins un développeur dans l'entreprise, car l'application mobile qu'Econochef développe utilise ces technologies.

4.4.2 LIMITATIONS DE L'ÉTUDE

Inévitablement, cette étude a des limitations. Premièrement, il est difficile de mesurer à quel point l'approche proposé est favorable comparativement à l'utilisation d'une autre approche. Les résultats présentés dans le cadre de cet étude sont préliminaires et d'autres travaux sont nécessaires afin de confirmer l'hypothèse que notre méthode permet la sélection d'une plateforme d'apprentissage automatique convenable pour le cas d'utilisation de la PME.

Il serait intéressant de suivre le cas de deux entreprises similaires où l'une d'entre elles utilise notre approche et l'autre non et d'ensuite comparer les résultats.

Ensuite, le fait que le système puisse être implémenté par des personnes avec peu de connaissances en apprentissage automatique pourrait poser des problèmes lors de l'évaluation des performances du modèle, car celles-ci n'auront peut-être pas les connaissances pour effectuer les évaluations nécessaires. Cependant, la plateforme étant majoritairement automatisée, cela est possiblement un faux problème, car celle-ci réentraine le modèle automatiquement périodiquement et assure des performances acceptables. Le cas d'utilisation simple de l'entreprise simplifie aussi la maintenance du système.

Ensuite, la méthode proposée aura aussi besoin d'être testée exhaustivement avec une multitude d'entreprises de différentes tailles ayant différents cas d'utilisation. Un seul cas d'utilisation n'est pas suffisant pour confirmer l'efficacité de cette méthode.

Il faut aussi mentionner que les priorités pour chaque caractéristique des plateformes sont définies de manière très arbitraire et à la discrétion du preneur de décision. En effet, il est très difficile de poser une importance objective à partir des options que l'on retrouve dans le tableau 3.1 à partir des mesures des caractéristiques que le preneur de décisions prend, même malgré le fait que les caractéristiques et la manière de les mesurer soient bien définies. Le résultat final pourrait donc varier d'un preneur de décision à l'autre pour le même cas d'utilisation. L'intégration de plusieurs personnes pour la prise de décision pourrait potentiellement rendre le processus plus fiable, mais pourrait également alourdir la tâche de sélection.

Dans le même ordre d'idées, bien que l'approche proposée soit abstraite et intemporelle, les résultats obtenus à un moment ne seront possiblement plus valables dans le futur. Il est donc important pour les entreprises de faire l'évaluation des caractéristiques sur des versions à jour des plateformes et de garder en tête qu'une décision prise avec l'approche proposée

deviendra inévitablement obsolète dans le futur. Les plateformes cloud, l'entreprise et le cas d'utilisation sont toujours en évolution, et il est donc indéniable que les valeurs de priorité des plateformes évolueront simultanément.

4.4.3 PERSPECTIVES FUTURES

Ces limitations ouvrent des perspectives de recherche intéressantes. L'une d'entre elles serait de tester le modèle de décision sur plusieurs entreprises et cas d'utilisations, ce qui augmenterait la fiabilité des résultats de l'étude. Des collaborations supplémentaires pourraient apporter de nouvelles idées quant aux caractéristiques à mesurer et permettre de confirmer notre modèle de décision. Des entreprises dans différents domaines et avec différents objectifs pourraient aussi confirmer la généralisation de notre modèle de décision.

Ensuite, la sélection de plus de caractéristiques ou l'utilisation d'autres algorithmes pour calculer les priorités pourrait constituer une étude intéressante et permettre de découvrir un modèle plus adapté que celui proposé par ce mémoire. Il serait même possible d'utiliser des techniques d'apprentissage automatique pour faire la sélection des plateformes. Cela nécessiterait cependant un ensemble de données de taille suffisante.

De plus, il serait intéressant de faire le suivi du cas d'Econochef pour confirmer le choix de la plateforme à plus long terme. Bien que celle-ci soit convenable actuellement, il faut s'assurer que le choix reste bon tant que l'entreprise reste une PME. La réévaluation des plateformes pourrait également être considérée afin de s'assurer que la plateforme sélectionnée soit toujours idéale pour les paramètres de l'entreprise dans le futur.

Un autre aspect intéressant serait d'automatiser le processus en créant un logiciel qui effectuerait les calculs des valeurs de priorités. Cela pourrait augmenter le niveau d'adoption

de l'approche proposée chez les entreprises qui doivent faire la sélection d'une plateforme cloud pour l'apprentissage automatique.

Étendre le modèle pour prendre en compte des plateformes Open-Source ouvrirait plus de possibilités aux entreprises qui veulent sélectionner une plateforme. Ces plateformes pourraient être plus convenables à certaines entreprises qui manipulent des données sensibles et dont la priorité est la confidentialité et la sécurité.

CHAPITRE V
CONCLUSIONS

Dans cette étude, nous avons développé un modèle de décision simple pour le choix d'une plateforme d'apprentissage automatique dans les petites et moyennes entreprises (PME). Nous avons défini les problématiques rencontrées lors de l'adoption de l'intelligence artificielle dans les petites entreprises et exploré les types de solutions disponibles pour ces entreprises. En nous basant sur les travaux existants, nous avons identifié et analysé les caractéristiques principales des plateformes cloud d'apprentissage automatique, que nous avons utilisées dans notre processus de décision.

La méthode de sélection proposée, fondée sur le processus AHP-express, permet aux PME de choisir une plateforme cloud en fonction de leurs objectifs spécifiques et des caractéristiques des plateformes. Nous avons démontré l'efficacité de cette méthode à travers une application pratique, apportant ainsi une contribution précieuse sous la forme d'une méthode simple et novatrice pour la sélection de fournisseurs cloud dans les PME.

Cette étude est un premier pas vers la simplification du travail des PME possédant des connaissances limitées en intelligence artificielle, mais souhaitant tout de même se lancer dans ce domaine. Les perspectives de recherches futures sont nombreuses : le suivi du cas d'utilisation d'Econochef, l'ajout d'un plus grand nombre d'entreprises pour confirmer les résultats de cette étude, ou encore l'extension du modèle pour inclure des plateformes open-source. Ces développements futurs pourraient enrichir et affiner le processus, offrant ainsi aux PME des outils encore plus robustes pour naviguer dans le paysage complexe de l'intelligence artificielle et de l'apprentissage automatique.

BIBLIOGRAPHIE

[1] S. K. Karmaker, M. M. Hassan, M. J. Smith, L. Xu, C. Zhai, et K. Veeramachaneni, "Automl to date and beyond : Challenges and opportunities," *ACM Computing Surveys (CSUR)*, vol. 54, n° 8, pp. 1–36, 2021.

[2] OpenAI, "Introducing ChatGPT." [En ligne]. Repéré à : https://openai.com/blog/chatgpt

[3] Midjourney, "Midjourney." [En ligne]. Repéré à : https://www.midjourney.com/home

[4] OpenAI, "DALL·E." [En ligne]. Repéré à : https://openai.com/research/dall-e

[5] A. Cam, M. Chui, et B. Hall, "Global ai survey : Ai proves its worth, but few scale impact," 2019. [En ligne]. Repéré à : https://www.mckinsey.com/featured-insights/artificial-intelligence/global-ai-survey-ai-proves-its-worth-but-few-scale-impact

[6] Microsoft Azure, "Azure AI : Solutions." [En ligne]. Repéré à : https://azure.microsoft.com/en-ca/solutions/ai/

[7] Amazon Web Services, "AWS Machine Learning : AI Services." [En ligne]. Repéré à : https://aws.amazon.com/machine-learning/ai-services/

[8] Google Cloud, "Google Cloud AI Products." [En ligne]. Repéré à : https://cloud.google.com/products/ai

[9] S. Lenkala, R. Marry, S. R. Gopovaram, T. C. Akinci, et O. Topsakal, "Comparison of automated machine learning (automl) tools for epileptic seizure detection using electroencephalograms (eeg)," *Computers*, vol. 12, n° 10, p. 197, 2023.

[10] L. M. Paladino, A. Hughes, A. Perera, O. Topsakal, et T. C. Akinci, "Evaluating the performance of automated machine learning (automl) tools for heart disease diagnosis and prediction," *AI*, vol. 4, n° 4, pp. 1036–1058, 2023.

[11] "Key small business statistics," Innovation, Sciences et Développement économique Canada, Rapport Technique Iu186-1E-PDF, 2022. [En ligne]. Repéré à : https:

//ised-isde.canada.ca/site/sme-research-statistics/en/key-small-business-statistics/key-small-business-statistics-2022

[12] J. Otterbach et T. Wollmann, “Chameleon : A semi-automl framework targeting quick and scalable development and deployment of production-ready ml systems for smes,” *arXiv preprint arXiv :2105.03669*, 2021.

[13] Z. Li, H. Guo, W. M. Wang, Y. Guan, A. V. Barenji, G. Q. Huang, K. S. McFall, et X. Chen, “A blockchain and automl approach for open and automated customer service,” *IEEE Transactions on Industrial Informatics*, vol. 15, n° 6, pp. 3642–3651, 2019.

[14] L. E. Lwakatare, A. Raj, J. Bosch, H. H. Olsson, et I. Crnkovic, “A taxonomy of software engineering challenges for machine learning systems : An empirical investigation,” dans *Agile Processes in Software Engineering and Extreme Programming : 20th International Conference, XP 2019, Montréal, QC, Canada, May 21–25, 2019, Proceedings 20*. Springer International Publishing, 2019, pp. 227–243.

[15] A. Paleyes, R.-G. Urma, et N. D. Lawrence, “Challenges in deploying machine learning : a survey of case studies,” *ACM computing surveys*, vol. 55, n° 6, pp. 1–29, 2022.

[16] L. Baier, F. Jöhren, et S. Seebacher, “Challenges in the deployment and operation of machine learning in practice.” dans *ECIS*, vol. 1, 2019.

[17] G. Symeonidis, E. Nerantzis, A. Kazakis, et G. A. Papakostas, “Mlops-definitions, tools and challenges,” dans *2022 IEEE 12th Annual Computing and Communication Workshop and Conference (CCWC)*. IEEE, 2022, pp. 0453–0460.

[18] D. Research, “Artificial intelligence and machine learning projects are obstructed by data issues,” 2019.

[19] R. Nazir, A. Bucaioni, et P. Pelliccione, “Architecting ML-enabled systems : Challenges, best practices, and design decisions,” *Journal of Systems and Software*, vol. 207, p. 111860, 2024.

[20] R. Jiang, S. Chiappa, T. Lattimore, A. György, et P. Kohli, “Degenerate feedback loops in recommender systems,” dans *Proceedings of the 2019 AAAI/ACM Conference on AI,*

Ethics, and Society, 2019, pp. 383–390.

[21] L. Leite, P. R. M. Meirelles, F. Kon, C. Rocha *et al.*, “Practices for managing machine learning products : a multivocal literature review,” *Authorea Preprints*, 2023.

[22] M. Chui, B. Hall, H. Mayhew, A. Singla, et A. Sukharevsky, “The state of ai in 2022-and a half decade in review,” 12 2022. [En ligne]. Repéré à : https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai-in-2022-and-a-half-decade-in-review

[23] M. Bauer, C. van Dinther, et D. Kiefer, “Machine learning in sme : an empirical study on enablers and success factors,” 2020.

[24] A. Andriyanto et R. Doss, “Problems and solutions of service architecture in small and medium enterprise communities,” *arXiv preprint arXiv :2004.10660*, 2020.

[25] J. McCarthy *et al.*, “What is artificial intelligence,” 2007.

[26] Wikipedia, “Artificial intelligence,” 2023. [En ligne]. Repéré à : https://en.wikipedia.org/wiki/Artificial_intelligence

[27] IBM. (2023) What is machine learning ? [En ligne]. Repéré à : https://www.ibm.com/topics/machine-learning

[28] J. M. Helm, A. M. Swiergosz, H. S. Haeberle, J. M. Karnuta, J. L. Schaffer, V. E. Krebs, A. I. Spitzer, et P. N. Ramkumar, “Machine learning and artificial intelligence : definitions, applications, and future directions,” *Current reviews in musculoskeletal medicine*, vol. 13, pp. 69–76, 2020.

[29] A. Burkov, *Machine learning engineering*. True Positive Incorporated Montreal, QC, Canada, 2020, vol. 1.

[30] IBM. (2023) Five machine learning types to know. [En ligne]. Repéré à : https://www.ibm.com/blog/machine-learning-types/

[31] AWS. (2024) What is machine learning. [En ligne]. Repéré à : https://aws.amazon.com/what-is/machine-learning/

[32] G. Cloud. (2024) What is machine learning. [En ligne]. Repéré à : https://cloud.google.com/learn/what-is-machine-learning

[33] G. Lorenzoni, P. Alencar, N. Nascimento, et D. Cowan, "Machine learning model development from a software engineering perspective : A systematic literature review," *arXiv preprint arXiv :2102.07574*, 2021.

[34] D. Sculley, G. Holt, D. Golovin, E. Davydov, T. Phillips, D. Ebner, V. Chaudhary, M. Young, J.-F. Crespo, et D. Dennison, "Hidden technical debt in machine learning systems," *Advances in neural information processing systems*, vol. 28, 2015.

[35] M. Virmani, "Understanding devops & bridging the gap from continuous integration to continuous delivery," dans *Fifth international conference on the innovative computing technology (intech 2015)*. IEEE, 2015, pp. 78–82.

[36] C. Ebert, G. Gallardo, J. Hernantes, et N. Serrano, "Devops," *IEEE software*, vol. 33, n° 3, pp. 94–100, 2016.

[37] GitLab, "Security without sacrifices," Rapport Technique, 2023. [En ligne]. Repéré à : https://about.gitlab.com/developer-survey/

[38] A. F. Varón Maya, "The state of mlops," 2021.

[39] I. Buchanan, "History of devops," 2023. [En ligne]. Repéré à : https://www.atlassian.com/devops/what-is-devops/history-of-devops

[40] A. Barrak, E. E. Eghan, et B. Adams, "On the co-evolution of ml pipelines and source code-empirical study of dvc projects," dans *2021 IEEE International Conference on Software Analysis, Evolution and Reengineering (SANER)*. IEEE, 2021, pp. 422–433.

[41] K. Hazelwood, S. Bird, D. Brooks, S. Chintala, U. Diril, D. Dzhulgakov, M. Fawzy, B. Jia, Y. Jia, A. Kalro *et al.*, "Applied machine learning at facebook : A datacenter infrastructure

perspective," dans *2018 IEEE International Symposium on High Performance Computer Architecture (HPCA)*. IEEE, 2018, pp. 620–629.

[42] R. Merrit, "What is mlops ?" 2020. [En ligne]. Repéré à : https://blogs.nvidia.com/blog/2020/09/03/what-is-mlops/

[43] J. Baer et S. Ngahane, "The winding road to better machine learning infrastructure through tensorflow extended and kubeflow," 2019. [En ligne]. Repéré à : https://engineering.atspotify.com/2019/12/the-winding-road-to-better-machine-learning-infrastructure-through-tensorflow-extended-and-kubeflow

[44] Kubeflow, "Introduction to kubeflow," 2021. [En ligne]. Repéré à : https://www.kubeflow.org/docs/started/introduction/

[45] N. Chedid, J. Tabbal, A. Kabbara, S. Allouch, et M. Hassan, "The development of an automated machine learning pipeline for the detection of alzheimer's disease," *Scientific Reports*, vol. 12, n° 1, p. 18137, 2022.

[46] E. Zhang, G. Catania, et D. T. Trugman, "Autoterm : an automated pipeline for glacier terminus extraction using machine learning and a "big data" repository of greenland glacier termini," *The Cryosphere*, vol. 17, n° 8, pp. 3485–3503, 2023.

[47] A. R. Dakak, P. Bouvet, L. Gueye, N. T. Duy, A. Autret, B. Fayard, et V. Kaftandjian, "Automation of non-destructive evaluation of casting parts based on computed tomography and machine learning," *e-Journal of Nondestructive Testing*, 2023. [En ligne]. Repéré à : https://api.semanticscholar.org/CorpusID:257215509

[48] V. Kumar, D. Ghosh, et S. Srivastava, "Efficient mlops pipeline for transfer learning and reuse of pre-trained ml models," dans *2023 IEEE International Conference on Advanced Networks and Telecommunications Systems (ANTS)*. IEEE, 2023, pp. 1–6.

[49] M. Testi, M. Ballabio, E. Frontoni, G. Iannello, S. Moccia, P. Soda, et G. Vessio, "Mlops : A taxonomy and a methodology," *IEEE Access*, vol. 10, pp. 63 606–63 618, 2022.

[50] J. Diaz-De-Arcaya, A. I. Torre-Bastida, G. Zarate, R. Minon, et A. Almeida, "A joint study of the challenges, opportunities, and roadmap of mlops and aiops : A systematic

survey," *ACM Computing Surveys*, vol. 56, nº 4, pp. 1–30, 2023.

[51] *Financing Micro, Small, and Medium Enterprises*. The World Bank, 2008. [En ligne]. Repéré à : https://elibrary.worldbank.org/doi/abs/10.1596/978-0-8213-7417-7

[52] C. Rodriguez et O. Gürcay, "Open-source software in business and its advantages & disadvantages," 2020.

[53] K. Venkateswar, "Using amazon {SageMaker} to operationalize machine learning," 2019.

[54] A. Team, "Azureml : Anatomy of a machine learning service," dans *Conference on Predictive APIs and Apps*. PMLR, 2016, pp. 1–13.

[55] D. Baylor, K. Haas, K. Katsiapis, S. Leong, R. Liu, C. Menwald, H. Miao, N. Polyzotis, M. Trott, et M. Zinkevich, "Continuous training for production ML in the TensorFlow extended (TFX) platform," dans *2019 USENIX Conference on Operational Machine Learning (OpML 19)*. Santa Clara, CA : USENIX Association, mai 2019, pp. 51–53. [En ligne]. Repéré à : https://www.usenix.org/conference/opml19/presentation/baylor

[56] E. Bisong et E. Bisong, "An overview of google cloud platform services," *Building Machine Learning and Deep Learning Models on Google Cloud Platform : A Comprehensive Guide for Beginners*, pp. 7–10, 2019.

[57] P. Gupta, A. Seetharaman, et J. R. Raj, "The usage and adoption of cloud computing by small and medium businesses," *International journal of information management*, vol. 33, nº 5, pp. 861–874, 2013.

[58] Q. Yao, M. Wang, Y. Chen, W. Dai, Y.-F. Li, W.-W. Tu, Q. Yang, et Y. Yu, "Taking human out of learning applications : A survey on automated machine learning," *arXiv preprint arXiv :1810.13306*, 2018.

[59] X. He, K. Zhao, et X. Chu, "Automl : A survey of the state-of-the-art," *Knowledge-based systems*, vol. 212, p. 106622, 2021.

[60] A. Ebadi, Y. Gauthier, S. Tremblay, et P. Paul, "How can automated machine learning help business data science teams ?" dans *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*. Institute of Electrical and Electronics Engineers Inc., 12 2019, pp. 1186–1191.

[61] L. Faes, S. K. Wagner, D. J. Fu, X. Liu, E. Korot, J. R. Ledsam, T. Back, R. Chopra, N. Pontikos, C. Kern *et al.*, "Automated deep learning design for medical image classification by health-care professionals with no coding experience : a feasibility study," *The Lancet Digital Health*, vol. 1, n° 5, pp. e232–e242, 2019.

[62] I. Guyon, L. Sun-Hosoya, M. Boullé, H. J. Escalante, S. Escalera, Z. Liu, D. Jajetic, B. Ray, M. Saeed, M. Sebag *et al.*, "Analysis of the automl challenge series," *Automated Machine Learning*, vol. 177, pp. 177–219, 2019.

[63] A. Truong, A. Walters, J. Goodsitt, K. Hines, C. B. Bruss, et R. Farivar, "Towards automated machine learning : Evaluation and comparison of automl approaches and tools," dans *2019 IEEE 31st international conference on tools with artificial intelligence (ICTAI)*. IEEE, 2019, pp. 1471–1479.

[64] D. Widyastuti et I. Irwansyah, "Benefits and challenges of cloud computing technology adoption in small and medium enterprises (SMEs)," *Bandung Creative Movement (BCM)*, vol. 4, n° 1, 2018.

[65] N. A. Sultan, "Reaching for the "cloud" : How smes can manage," *International journal of information management*, vol. 31, n° 3, pp. 272–278, 2011.

[66] M. Rožman, D. Oreški, K. Crnogaj, et P. Tominc, "Agility and artificial intelligence adoption : Small vs. large enterprises," *Naše gospodarstvo/Our economy*, vol. 69, n° 4, pp. 26–37, 2023.

[67] R. Saracco, "Perspectives on ai adoption in italy, the role of the italian ai strategy," *Discover Artificial Intelligence*, vol. 2, n° 1, p. 9, 2022.

[68] A. K. POLZER, J. P. ZEIRINGER, et S. THALMANN, "Automl as facilitator of ai adoption in smes : An analysis of automl use cases," *36th Bled eConference Digital Economy and Society : The Balancing Act for Digital Innovation in Times of Instability*, p. 713, 2023.

[69] Y. Hermansyah, “Assessing the impact of communicative artificial intelligence based accounting information systems on small and medium enterprises,” *Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications*, vol. 14, n° 3, pp. 230–239, 2023.

[70] S. S. Ingalagi, R. Mutkekar, et P. Kulkarni, “Artificial intelligence (ai) adaptation : Analysis of determinants among small to medium-sized enterprises (sme’s),” dans *IOP Conference Series : Materials Science and Engineering*, vol. 1049, n° 1. IOP Publishing, 2021, p. 012017.

[71] A. W. Services, “Artificial intelligence for every small business,” 2024. [En ligne]. Repéré à : https://aws.amazon.com/fr/smart-business/solutions/artificial-intelligence-small-medium-business/

[72] Baseline, “Jeu de cartes d’ia,” Rapport Technique, 2024. [En ligne]. Repéré à : https://40509020.hs-sites.com/jeu-de-cartes-dia

[73] F. Hutter, L. Kotthoff, et J. Vanschoren, *Automated machine learning : methods, systems, challenges*. Springer Nature, 2019.

[74] Y. Roh, G. Heo, et S. E. Whang, “A survey on data collection for machine learning : a big data-ai integration perspective,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 33, n° 4, pp. 1328–1347, 2019.

[75] M. Bianchini et V. Michalkova, “Data analytics in smes : Trends and policies,” 2019.

[76] J. Bender, M. Trat, et J. Ovtcharova, “Benchmarking automl-supported lead time prediction,” *Procedia Computer Science*, vol. 200, pp. 482–494, 2022.

[77] H. Stühler, D. Klau, M.-A. Zöller, A. Beiderwellen-Bedrikow, et C. Tutschku, “End-to-end implementation of automated price forecasting applications,” *SN Computer Science*, vol. 5, n° 4, p. 402, 2024.

[78] J. Gyani, A. Ahmed, et M. A. Haq, “MCDM and various prioritization methods in AHP for CSS : A comprehensive review,” *IEEE Access*, vol. 10, pp. 33 492–33 511, 2022.

[79] D. Zdraveski, M. Janeska, et S. Taleska, “Evaluating cloud computing services,” 2020.

[80] J. Repschlaeger, S. Wind, R. Zarnekow, et K. Turowski, “Decision model for selecting a cloud provider : A study of service model decision priorities,” 2013.

[81] S. K. Garg, S. Versteeg, et R. Buyya, “A framework for ranking of cloud computing services,” *Future Generation Computer Systems*, vol. 29, n° 4, pp. 1012–1023, 2013.

[82] J. Siegel et J. Perdue, “Cloud services measures for global use : the service measurement index (SMI),” dans *2012 Annual SRII global conference*. IEEE, 2012, pp. 411–415.

[83] M. Godse et S. Mulik, “An approach for selecting software-as-a-service (SaaS) product,” dans *2009 IEEE International Conference on Cloud Computing*. IEEE, 2009, pp. 155–158.

[84] U. Şener, E. Gökalp, et P. E. Eren, “ClouDSS : A decision support system for cloud service selection,” dans *Economics of Grids, Clouds, Systems, and Services : 14th International Conference, GECON 2017, Biarritz, France, September 19-21, 2017, Proceedings 14*. Springer, 2017, pp. 249–261.

[85] P. Ruf, M. Madan, C. Reich, et D. Ould-Abdeslam, “Demystifying mlops and presenting a recipe for the selection of open-source tools,” *Applied Sciences*, vol. 11, n° 19, p. 8861, 2021.

[86] C. Kaymakci, S. Wenninger, P. Pelger, et A. Sauer, “A systematic selection process of machine learning cloud services for manufacturing smes,” *Computers*, vol. 11, n° 1, p. 14, 2022.

[87] P. Mell, T. Grance *et al.*, “The NIST definition of cloud computing,” 2011.

[88] P. A. Abdalla et A. Varol, “Advantages to disadvantages of cloud computing for small-sized business,” dans *2019 7th International Symposium on Digital Forensics and Security (ISDFS)*. IEEE, 2019, pp. 1–6.

[89] M.-G. Avram, “Advantages and challenges of adopting cloud computing from an enter-

prise perspective," *Procedia Technology*, vol. 12, pp. 529–534, 2014.

[90] D. Kreuzberger, N. Kühl, et S. Hirschl, "Machine learning operations (mlops) : Overview, definition, and architecture," *IEEE access*, 2023.

[91] AltexSoft, "Comparing machine learning as a service : Amazon, microsoft azure, google cloud ai, ibm watson," Feb 2020. [En ligne]. Repéré à : https://www.altexsoft.com/blog/datascience/comparing-machine-learning-as-a-service-amazon-microsoft-azure-google-cloud-ai-ibm-watson/

[92] B. Delovski, "What are the most popular machine learning as a service (mlaas) tools in 2023 ?" 2023. [En ligne]. Repéré à : https://www.edlitera.com/en/blog/posts/most-popular-ml-as-a-service-tools

[93] Zaveria, "Top 10 mlaas platforms that techies should be aware of in 2023," 2022. [En ligne]. Repéré à : https://www.analyticsinsight.net/top-10-mlaas-platforms-that-techies-should-be-aware-of-in-2023/

[94] A. Joby, "What is machine learning as a service (mlaas) ?" 2023. [En ligne]. Repéré à : https://www.g2.com/articles/machine-learning-as-a-service

[95] G. S. Panwar, "Best machine learning as a service platforms (mlaas) that you want to check as a data scientist," 2023. [En ligne]. Repéré à : https://neptune.ai/blog/best-machine-learning-as-a-service-platforms-mlaas

[96] M. A. Richardson, "Top 7 mlaas platforms you should consider in 2021," 2021. [En ligne]. Repéré à : https://www.spiceworks.com/tech/artificial-intelligence/articles/top-mlaas-platforms-to-consider/

[97] R. Wolff, "What is mlaas &amp ; what are the best platforms ?" 2020. [En ligne]. Repéré à : https://monkeylearn.com/blog/mlaas/

[98] P. Orza, "Mlaas platforms : The comparative guide," 2022. [En ligne]. Repéré à : https://levity.ai/blog/mlaas-platforms-comparative-guide

[99] V. Kuprenko, “Mlaas platforms for novices and pros : Opt the one you need,” 2021. [En ligne]. Repéré à : https://towardsdatascience.com/mlaas-platforms-for-novices-and-pros-opt-the-one-you-need-6b64c377a89c

[100] SlashDot, “Best machine learning as a service (mlaas) platforms.” [En ligne]. Repéré à : https://slashdot.org/software/machine-learning-as-a-service/

[101] S. Robinson, “Comparing mlaas providers by cost, ux and ease of use,” 2020. [En ligne]. Repéré à : https://www.techtarget.com/searchenterpriseai/feature/Comparing-MLaaS-providers-by-cost-UX-and-ease-of-use

[102] The LF AI & Data, “Lf ai & data foundation interactive landscape,” 2023. [En ligne]. Repéré à : https://landscape.lfai.foundation/

[103] T. L. Saaty, “How to make a decision : the analytic hierarchy process,” *European journal of operational research*, vol. 48, n° 1, pp. 9–26, 1990.

[104] J. E. Leal, “Ahp-express : A simplified version of the analytical hierarchy process method,” *MethodsX*, vol. 7, p. 100748, 2020.

[105] Synergy Research Group, “Cloud spending growth rate slows but q4 still up by $10 billion from 2021 ; microsoft gains market share,” 2023. [En ligne]. Repéré à : https://www.srgresearch.com/articles/cloud-spending-growth-rate-slows-but-q4-still-up-by-10-billion-from-2021-microsoft-gains-market-share

[106] Google Cloud Platform, “Export a model.” [En ligne]. Repéré à : https://cloud.google.com/vertex-ai/docs/samples/aiplatform-export-model-sample#aiplatform_export_model_sample-python