Nº au catalogue 89-20-0006

Cadre pour l'utilisation des processus d'apprentissage automatique de façon responsable à Statistique Canada
Juillet 2020

Date de diffusion : le 3 mai 2021

Statistique Canada Statistics Canada

Canada

Comment obtenir d'autres renseignements

Pour toute demande de renseignements au sujet de ce produit ou sur l'ensemble des données et des services de Statistique Canada, visiter notre site Web à www.statcan.gc.ca.

Vous pouvez également communiquer avec nous par :

Courriel à STATCAN.infostats-infostats.STATCAN@canada.ca

Téléphone entre 8 h 30 et 16 h 30 du lundi au vendredi aux numéros suivants :

- Service de renseignements statistiques 1-800-263-1136
- Service national d'appareils de télécommunications pour les malentendants 1-800-363-7629
- Télécopieur 1-514-283-9350

Programme des services de dépôt

- Service de renseignements 1-800-635-7943
- Télécopieur 1-800-565-7757

Normes de service à la clientèle

Statistique Canada s'engage à fournir à ses clients des services rapides, fiables et courtois. À cet égard, notre organisme s'est doté de normes de service à la clientèle que les employés observent. Pour obtenir une copie de ces normes de service, veuillez communiquer avec Statistique Canada au numéro sans frais 1-800-263-1136. Les normes de service sont aussi publiées sur le site www.statcan.gc.ca sous « Contactez-nous » > « Normes de service à la clientèle ».

Note de reconnaissance

Le succès du système statistique du Canada repose sur un partenariat bien établi entre Statistique Canada et la population du Canada, les entreprises, les administrations et les autres organismes. Sans cette collaboration et cette bonne volonté, il serait impossible de produire des statistiques exactes et actuelles.

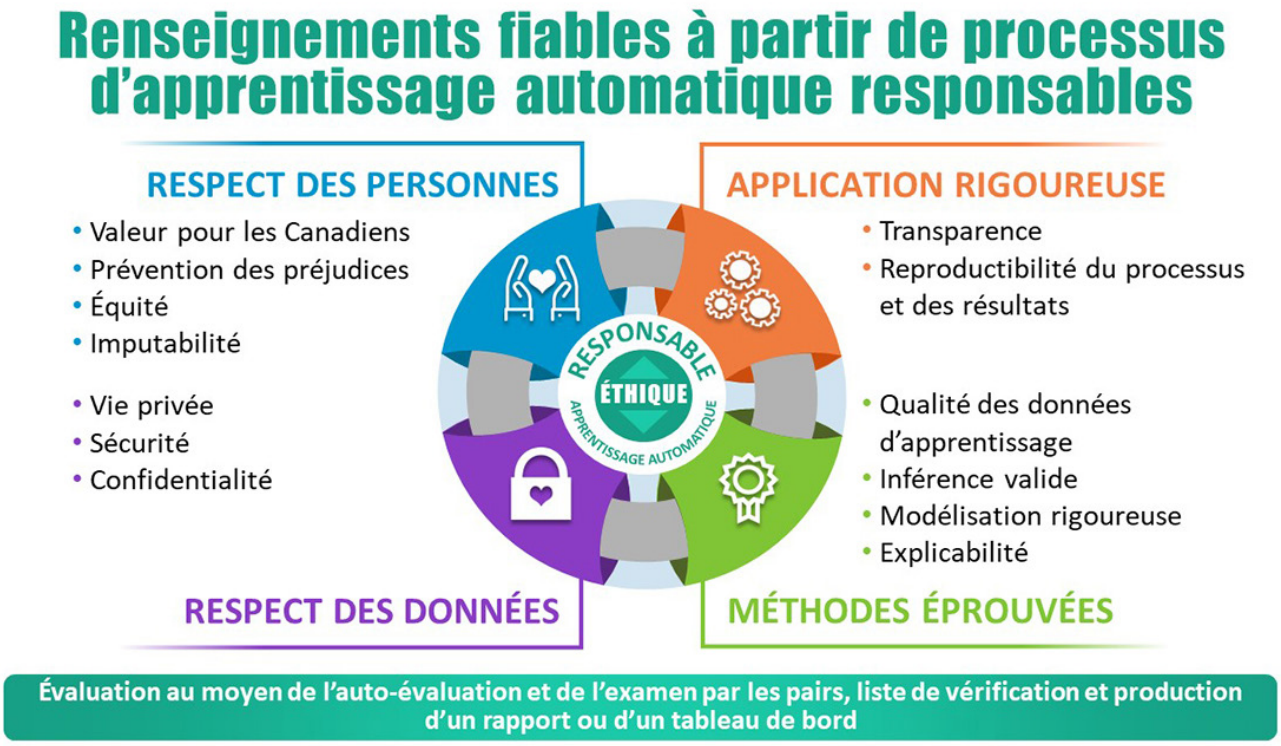
Publication autorisée par le ministre responsable de Statistique Canada

Une version HTML est aussi disponible.

This publication is also available in English.

Table des matières

Cadre pour l'utilisation des processus d'apprentissage automatique de façon responsable à Statistique Canada Juillet 2020



Introduction

Le présent document est un guide pour les utilisateurs qui désirent élaborer et mettre en œuvre des projets d'apprentissage automatique. Il fournit une orientation et des conseils pratiques sur la façon responsable d'élaborer ces processus automatisés au sein de Statistique Canada, mais qui pourrait aussi être adopté par toute autre organisation. Le guide peut s'appliquer au travail de production tout aussi bien qu'au travail de recherche menant à des publications.

Il incombe au gestionnaire de s'assurer (en partenariat avec les fournisseurs de services tels que les scientifiques des données, les méthodologistes ainsi que les responsables des systèmes) que les lignes directrices sont suivies et que la liste de contrôle est remplie au moment opportun tout au long du cycle de vie d'un projet d'apprentissage automatique.

Contexte

L'apprentissage automatique est une science qui consiste à concevoir des algorithmes et des modèles statistiques mis en œuvre à l'aide de systèmes informatiques pour réaliser efficacement certaines tâches sans avoir recours à des instructions explicites, mais en s'appuyant plutôt sur des configurations et sur l'inférence. Les algorithmes d'apprentissage automatique servent à construire un modèle mathématique afin de faire des prévisions ou à prendre des décisions sans être explicitement programmés pour exécuter cette tâche en déduisant des comportements à adopter à partir de données.

L'apprentissage automatique peut améliorer l'efficacité des systèmes existants, fournir une efficacité accrue des opérations et améliorer la prise de décisions. La croissance rapide de la science des données et la présence accrue de processus d'apprentissage automatique dans notre travail à Statistique Canada soulèvent des questions

pressantes à propos des effets de ces technologies et de la gouvernance, de l'éthique et de notre responsabilité envers elles. Comme cela est indiqué dans les principes du gouvernement du Canada à propos de l'intelligence artificielle responsable (IA – ce qui inclut l'apprentissage automatique), le gouvernement s'efforcera :

1. de comprendre et de mesurer les effets de l'utilisation de l'IA (y compris l'apprentissage automatique) en élaborant et en partageant des outils et des approches;
2. d'être transparent lorsque ces processus sont utilisés et sur la façon de les utiliser, en commençant par expliquer clairement au public le besoin et l'avantage d'utiliser l'apprentissage automatique;
3. de fournir des explications valables au sujet des décisions prises à l'aide de l'IA, tout en offrant la possibilité d'examiner les résultats et de contester ces décisions;
4. d'être aussi ouvert que possible en partageant les codes sources, les données d'apprentissage et d'autres renseignements pertinents, tout en protégeant les renseignements personnels, l'intégrité du système, la sécurité nationale et la défense.

Afin de mettre cela en place, la Directive sur la prise de décision automatisée a été rédigée. Cette dernière est fondée sur les résultats de l'outil d'évaluation de l'incidence algorithmique qui est utilisé afin de déterminer si l'IA est acceptable d'un point de vue humain et éthique.

À cette fin, un cadre pour l'utilisation des processus d'apprentissage automatique de façon responsable a été élaboré à Statistique Canada. Le cadre comprend des lignes directrices pour l'usage responsable de l'apprentissage automatique et une liste de contrôle connexe, qui sont organisées en quatre thèmes : le respect des personnes; le respect des données; des méthodes éprouvées; une application rigoureuse. Les quatre thèmes mis en commun assurent l'utilisation éthique des algorithmes et des résultats de l'apprentissage automatique.

Portée

Ces lignes directrices s'appliquent à tous les programmes et projets statistiques menés par Statistique Canada qui utilisent des algorithmes d'apprentissage automatique. Cela comprend les algorithmes d'apprentissage supervisés et non supervisés. Ces programmes comprennent : les programmes de données administratives, les enquêtes, les comptes macroéconomiques, les recensements, les études analytiques ou même des projets expérimentaux. L'apprentissage automatique peut être interne aussi bien qu'externe par rapport au programme. Ces lignes directrices offrent des directives unifiées avec des principes clés et des thèmes, ainsi qu'une liste de contrôle afin d'aider les scientifiques des données dans l'évaluation de leur travail. Le présent document est ancré dans une vision qui cherche à créer un milieu de travail moderne et à fournir une orientation et un soutien à ceux qui utilisent les techniques d'apprentissage automatique.

La validation des processus d'apprentissage automatique sera effectuée par une autoévaluation, un examen par les pairs, un comité, ou une combinaison de ceux-ci.

Ces lignes directrices sont conformes à la Politique de Statistique Canada sur l'intégrité scientifique (qui stipule que toutes activités scientifiques et de recherche doivent être effectuées d'une manière conforme à toutes les normes adéquates et applicables en matière d'excellence scientifique, d'éthique de la recherche et de conduite responsable de la recherche) et reflètent les valeurs fondamentales de Statistique Canada (inférences statistiques valides; qualité; rigueur). Elles doivent être utilisées à titre de complément aux Lignes directrices concernant la qualité et à l'ensemble des politiques de Statistique Canada. Il est entendu que les bonnes pratiques en matière de documentation, d'assurance de la qualité et de rapports sur la mesure du rendement seront également suivies, même si elles ne sont pas toujours mentionnées de façon explicite dans le présent document.

Thème : Respect des personnes

À Statistique Canada, nous visons à faire une utilisation efficace des ressources du gouvernement tout en produisant des renseignements qui aident les Canadiens à mieux comprendre leur pays. Ce thème inclut quatre attributs : la valeur pour les Canadiens; la prévention des préjudices; l'équité; l'imputabilité.

Le concept de **valeur pour les Canadiens** dans un contexte d'apprentissage automatique implique que son utilisation doit avoir une valeur ajoutée, que ce soit dans les produits eux-mêmes ou par une plus grande efficacité dans le processus de production.

Les lignes directrices pour la valeur pour les Canadiens

- S'assurer que l'algorithme d'apprentissage automatique a un avantage clair pour les utilisateurs des données.
- S'assurer que l'algorithme d'apprentissage automatique donne lieu à l'adéquation des produits statistiques à leur utilisation.
- S'assurer de la pertinence des algorithmes d'apprentissage automatique utilisés.

Dans le contexte de l'utilisation de l'apprentissage automatique à Statistique Canada, un préjudice pourrait être causé envers des populations vulnérables si des renseignements de nature délicate à leur sujet étaient rendus publics. La **prévention des préjudices** nécessite d'être au courant des dangers potentiels et d'avoir un dialogue constructif avec les intervenants et les porte-parole du milieu avant la mise en œuvre d'un projet d'apprentissage automatique.

Les lignes directrices pour la prévention des préjudices

Faire preuve de délicatesse envers les populations vulnérables.

L'équité implique que le principe de la proportionnalité entre les moyens et les fins soit respecté, et qu'un équilibre soit maintenu entre des intérêts et des objectifs différents. L'équité veille à ce que les personnes et les groupes ne soient pas victimes de préjugés injustes, de discrimination ou de stigmatisation.

Les lignes directrices pour l'équité

- S'assurer que toutes les données d'apprentissage, les codes informatiques et les outils utilisés dans un processus d'apprentissage automatique sont acquis, entretenus et utilisés conformément aux protocoles existants.
- S'assurer que toutes les variables du modèle sont pertinentes, et qu'aucune ne pourrait causer d'injustices, de discrimination ou de stigmatisation.
- Les processus d'apprentissage automatique doivent protéger l'intégrité des données et leur confidentialité.
- S'assurer que l'équipe de développement n'intègre pas, par inadvertance, d'iniquités dans les processus d'apprentissage automatique. Cela pourrait être accompli grâce à des activités indépendantes, durant lesquelles les membres de l'équipe réfléchissent et mettent de côté leurs préjugés personnels, leurs idées préconçues et leurs expériences. Cela garantira que le développement et l'utilisation de l'apprentissage automatique sont compatibles avec le maintien de la diversité sociale et culturelle et ne restreignent pas la portée des choix de style de vie et des expériences personnelles, et que les membres de l'équipe de développement saisissent l'occasion d'anticiper les conséquences négatives potentielles à la suite de l'utilisation de l'apprentissage automatique.

L'imputabilité est l'obligation juridique et éthique d'une personne ou d'une organisation d'être responsable de son travail et de communiquer les résultats du travail de façon transparente. Les algorithmes ne sont pas responsables; quelqu'un est responsable des algorithmes.

Les lignes directrices pour l'imputabilité

- S'assurer que quelqu'un a été désigné comme étant responsable du projet d'apprentissage automatique et des résultats durant toutes les phases du projet (l'élaboration, le déploiement et la production).
- S'assurer qu'un plan est en place pour la surveillance du rendement et la maintenance des logiciels tout au long du cycle de vie du projet.
- S'assurer que les recommandations et les décisions sont consignées.

Thème : Respect des données

À Statistique Canada, nous prenons les données au sérieux. Ce thème a trois attributs : la protection de la vie privée des personnes auxquelles les données appartiennent; la sécurité des renseignements tout au long du cycle de vie des données; la confidentialité de renseignements identifiables.

La **vie privée** est le droit de se retirer et de ne pas être sujet à une quelconque forme de surveillance ou d'intrusion. Lors de l'acquisition de renseignements de nature délicate, les gouvernements ont des obligations relativement à la collecte, à l'utilisation, à la divulgation et à la conservation des renseignements personnels. Le terme vie privée réfère généralement à des renseignements concernant des particuliers (définition tirée de la Politique sur la protection des renseignements personnels et la confidentialité).

Les lignes directrices pour la protection des renseignements personnels

- S'assurer que les risques d'entrave à la vie privée sont réduits au minimum en utilisant le moins de variables d'identificateurs personnels possible tout au long du processus.
- S'assurer que les instruments pertinents de l'ensemble des politiques sont respectés, en particulier la Directive sur les évaluations des facteurs relatifs à la vie privée.

La **sécurité** représente les dispositions fondées sur l'évaluation de la menace et des risques qu'utilisent les organisations pour empêcher l'obtention ou la divulgation inadéquate de renseignements confidentiels. Les mesures de sécurité protègent aussi l'intégrité, la disponibilité et la valeur des fonds de renseignements. Cela englobe les protections matérielles, comme l'accès restreint aux zones où les renseignements sont entreposés et utilisés ou les autorisations de sécurité des employés, ainsi que les protections technologiques utilisées pour empêcher l'accès électronique non autorisé (définition tirée de la Politique sur la protection des renseignements personnels et la confidentialité).

Les lignes directrices pour la sécurité

- S'assurer que les renseignements de nature délicate sont sécurisés en conformité avec la Directive sur la sécurité des renseignements statistiques de nature délicate.

La **confidentialité** fait référence à la protection contre la divulgation de renseignements personnels identifiables concernant une personne, une entreprise ou une organisation. La confidentialité suppose une relation de « confiance » entre le fournisseur de renseignements et l'organisation qui les recueille; cette relation s'appuie sur l'assurance que ces renseignements ne seront pas divulgués sans l'autorisation de la personne ou sans l'autorité législative appropriée (définition tirée de la Politique sur la protection des renseignements personnels et la confidentialité).

Les lignes directrices pour la confidentialité

- Veiller à ce que la confidentialité soit protégée en conformité avec la Politique sur la protection des renseignements personnels et la confidentialité.

Thème : Application rigoureuse

Une **application rigoureuse** signifie de mettre en place, de maintenir et de documenter les processus d'apprentissage automatique de façon à ce que les résultats soient toujours fiables et que l'ensemble du processus puisse être compris et recréé. Ce thème a deux attributs : la transparence et la reproductibilité du processus et des résultats.

La **transparence** fait référence au fait d'avoir une justification claire de la raison pour laquelle cet algorithme et les données d'apprentissage sont les plus appropriés pour l'étude en cours. Pour être transparents, les développeurs devraient produire une documentation complète, y compris rendre accessible le code informatique à d'autres personnes, et ce, sans compromettre la confidentialité ou la protection des renseignements personnels.

Les lignes directrices pour la transparence

- Communiquer clairement aux utilisateurs comment, où et pourquoi l'apprentissage automatique a été utilisé dans le processus. Cela doit comprendre une description des données d'apprentissage, le fonctionnement de l'algorithme et les diagnostics de modèle qui sont utilisés. Divulguer tout biais ou les limites dans les données. Partager le code informatique s'il y a lieu.
- S'assurer que tous les partenaires, qu'ils soient experts sur le sujet, en science des données, en informatique et en méthodologie participent de manière appropriée au développement des modèles d'apprentissage automatique.

La **reproductibilité du processus** signifie qu'il y a suffisamment de documentation et que le code informatique a été suffisamment partagé pour faire en sorte que le processus soit reproduit, à partir de rien. La **reproductibilité des résultats** signifie que les mêmes résultats peuvent être reproduits de façon fiable lorsque toutes les conditions sont contrôlées. Il n'y a pas d'étapes qui modifient les résultats à la suite d'une intervention ponctuelle ou humaine.

Les lignes directrices pour la reproductibilité du processus et des résultats

- S'assurer qu'un système de contrôle des versions soutenu par l'organisme (par exemple GitLab) est utilisé pour gérer toutes les versions du code écrit pour l'élaboration, la mise à l'essai et la mise en œuvre de l'algorithme d'apprentissage automatique.

- S'assurer que les renseignements suivants concernant l'élaboration, la mise à l'essai et la mise en œuvre de l'algorithme d'apprentissage automatique sont « regroupés » dans un même paquet (« bundle ») afin que les intervenants (actuels ou futurs) aient suffisamment de renseignements pour bien reproduire les résultats lorsque cela est nécessaire :
    - tout le code informatique pertinent;
    - le numéro de version de tous les outils logiciels utilisés;
    - la version exacte des données d'entrée;
    - tous les résultats intermédiaires et finaux, les diagnostics et les fichiers journaux (« log files »).
- Veiller à ce que le pipeline final du traitement et de l'analyse lié à l'élaboration, à la mise à l'essai et à la mise en œuvre de l'algorithme d'apprentissage automatique puisse générer automatiquement tous les résultats intermédiaires et finaux, tous les diagnostics et tous les fichiers journaux, sans intervention humaine, et à ce qu'il soit exécuté par un intervenant au moyen d'un seul programme ou script de base.
- S'assurer que les instructions requises pour exécuter le script de base sont bien documentées et comprises dans le paquet mentionné plus haut.
- S'assurer que le développement et l'utilisation de l'apprentissage automatique sont effectués d'une manière éco énergétique et écologiquement durable, par exemple en minimisant l'impression de la documentation et des produits ou en optimisant le traitement informatique.

Thème : Méthodes éprouvées

Les **méthodes éprouvées** sont celles qui peuvent être invoquées de manière efficace et efficiente afin de produire les résultats espérés. Nous suivons habituellement des protocoles reconnus qui comportent une consultation avec des pairs et des experts, de la documentation et des tests lorsque nous élaborons des méthodes éprouvées. Ce thème a quatre attributs : la qualité des données d'apprentissage; l'inférence valide; la modélisation rigoureuse; l'explicabilité.

Dans un contexte d'apprentissage automatique, la **qualité des données d'apprentissage** est mesurée par la cohérence et l'exactitude des données étiquetées. La couverture, ce qui signifie que les étiquettes et les descriptions couvrent tous les cas auxquels l'algorithme peut faire face dans la production, est également importante pour réduire le risque de partialité ou de discrimination (équité). La couverture est également importante pour assurer la représentativité des variables, ce qui est important lorsqu'on veut obtenir des mesures de rendement réalistes.

Lignes directrices concernant la qualité des données d'apprentissage

- Dans le présent contexte, le défi ne concerne pas la partialité ou la discrimination la plupart du temps dans les données d'apprentissage, mais plutôt un manque de données étiquetées que l'on peut utiliser pour l'entraînement, la validation et les tests. Si les ressources pour les données étiquetées sont limitées, cela pourrait être un obstacle important à surmonter. Il faut s'assurer au début de la phase de développement que suffisamment de données étiquetées et représentatives existent, sinon qu'un financement suffisant a été alloué à l'activité d'étiquetage.
- Un autre problème concret concerne la qualité des données étiquetées. Des données étiquetées de mauvaise qualité comportent quelques ou plusieurs étiquettes attribuées de façon erronée. Si des données de mauvaise qualité sont utilisées par un algorithme, on peut faire face à des problèmes de convergence, à d'importantes erreurs de prévision, et même de biais. Il faut faire un examen manuel d'un échantillon représentatif de données étiquetées pour mesurer les taux d'exactitude et de cohérence. Décrivez toute forme de sous-dénombrement dans les données d'apprentissage relatives à la population cible. Rapportez la distribution des étiquettes et des principales variables indépendantes (caractéristiques) dans les données d'apprentissage et faites une comparaison avec la population cible pour donner une idée de la représentativité.

- Surveiller l'uniformité, l'exactitude, la couverture et la représentativité des données d'apprentissage au fil du temps afin de détecter tout changement ou perte de qualité.
- Il est important d'avoir une bonne compréhension des données d'apprentissage, afin d'être capable d'évaluer dans quelle mesure elles représentent l'ensemble de la population cible. Si les données d'apprentissage viennent d'une enquête, assurez-vous de connaître la population cible, le plan d'échantillonnage, la façon dont l'échantillon a été sélectionné, la façon dont les données ont été recueillies, et les structures de non-réponse dans tous les domaines de la population. Si les données d'apprentissage proviennent d'une source administrative, assurez-vous de connaître la population cible, la couverture et quels prétraitement, vérification et techniques d'imputation ont déjà été effectués. Il faut aussi tenir compte de la façon dont les données d'apprentissage pourraient être une source de biais. Par exemple, étant donné le plan d'échantillonnage et les structures de non-réponse, faudrait-il inclure les poids de sondage dans notre utilisation d'un algorithme d'apprentissage automatique? Il faut documenter entièrement la source des données d'apprentissage et votre analyse de biais potentiel.

Une **inférence valide** désigne la capacité d'obtenir, à partir d'un échantillon, des conclusions plausibles et d'une précision connue de la population cible. Dans un contexte d'apprentissage automatique, une conclusion valable signifie que les prédictions à partir de données tests (jamais utilisées pour la modélisation) doivent être, dans une grande proportion, raisonnablement près de leurs vraies valeurs ou, dans le cas de données catégoriques, les prédictions sont exactes dans une grande proportion.

Les lignes directrices pour l'inférence valide

- Choisir une méthode de validation ou de diagnostic et les mesures connexes qui sont appropriées compte tenu de l'algorithme que vous utilisez et du contexte dans lequel il sera appliqué. Examiner les exigences en matière de qualité du processus d'apprentissage automatique (AA) lui-même et du processus statistique pour lequel il est utilisé. Si l'algorithme d'AA est un algorithme de prédiction (lié à une classification ou à une régression), alors les mesures d'exactitude prendraient la forme d'une erreur de prédiction. Une mesure de stabilité (la mesure de l'accord entre les prédictions d'une exécution à l'autre d'un même algorithme) est également de mise. Par exemple, dans le cas de regroupement (« clustering »), les mesures de stabilité sont importantes. L'instabilité est un signe de grappes mal séparées. Les protocoles de validation génériques et les mesures de qualité peuvent être ou ne pas être appropriés pour le problème soumis. Il faut établir des cibles de rendement appropriées en fonction de la qualité requise des produits finaux et le niveau d'incertitude auquel le processus d'AA contribue.
- Exécuter la méthode de validation ou le diagnostic et examiner si le modèle atteint les cibles. Lorsqu'on essaie d'attendre nos cibles, il est souvent possible d'ajuster nos hyper paramètres. Le choix des hyper paramètres devrait se faire à l'aide d'une approche systématique pour couvrir une vaste gamme de valeurs et il devrait être fondé sur les connaissances tout en ayant une intention à l'esprit. Si l'évaluation de la méthode est mauvaise, il faut envisager la possibilité d'obtenir plus de données d'apprentissage. Consultez un expert en apprentissage automatique au besoin.
- Pour un algorithme d'apprentissage automatique supervisé, il faut s'assurer qu'un ensemble de données d'essai est mis de côté dès le début du cycle de développement et qu'il n'est jamais utilisé avant l'évaluation finale du modèle optimisé. Pour un algorithme d'apprentissage automatique non supervisé, les résultats sont plus subjectifs puisque la vraie valeur est inconnue; il faut communiquer ce fait.
- Lorsque les données d'apprentissage proviennent d'un échantillon tiré au hasard, prenez bien soin de considérer l'utilisation de poids lorsque l'algorithme le permet.
- Utiliser un système de gestion des ensembles de logiciels, qui comprend à la fois les logiciels R et Python. Il est recommandé d'utiliser uniquement les logiciels R et Python qui ont un nombre énorme d'utilisateurs, qui auraient relevé toute erreur. Si vous songez à utiliser un logiciel ou un module d'une autre source, des essais rigoureux doivent être effectués et il est recommandé de consulter les communautés d'utilisateurs et les experts en apprentissage automatique pour s'assurer que l'ensemble de logiciels est de bonne qualité et que les résultats sont exacts et tels qu'attendus. Inclure les résultats de vos recherches et les conclusions dans un registre central.
- Faire un suivi des mesures de rendement au fil du temps afin de déterminer à quel moment il faudra entraîner le modèle de nouveau.

Une **modélisation rigoureuse** en apprentissage automatique consiste à s'assurer que les algorithmes sont vérifiés et validés. Cela permettra aux utilisateurs et aux décideurs de faire confiance à l'algorithme à juste titre du point de vue de l'adaptation des données à leur utilisation, de la fiabilité et de la robustesse.

Les lignes directrices pour une modélisation rigoureuse

- Une modélisation rigoureuse garantira que les données d'apprentissage sont une représentation fidèle de la population visée par l'application d'apprentissage automatique. Cela garantit également que le modèle est appliqué dans les limites appropriées, telles que définies par les données d'apprentissage. L'erreur de généralisation est l'erreur obtenue lorsqu'on applique le modèle sur de nouvelles données. L'erreur de généralisation a deux sources; elle peut se manifester lorsqu'on a un sous-ajustement du modèle (parfois appelé biais), c.-à-d. lorsque le modèle n'est pas suffisamment complexe (a trop peu de paramètres). D'autre part, l'erreur de généralisation peut se manifester également lorsqu'on a un sur-ajustement du modèle (parfois appelé la variance), c'est-à-dire lorsque le modèle est trop complexe et ne permet pas de saisir la tendance générale dans les données. Dans les deux cas, l'erreur de mesure fondée sur les données d'apprentissage est beaucoup plus petite que l'erreur de mesure fondée sur les données de test. Il faut s'assurer qu'une modélisation rigoureuse est menée et que l'ensemble des données d'apprentissage est choisi de façon à bien représenter la population afin que la mesure d'erreur obtenue sur les données de test soit une source fiable ou produise une estimation raisonnable de l'erreur de généralisation du modèle optimisé, en particulier lorsque celui-ci est déployé dans l'environnement de production.

Un modèle qui est **explicable** est un modèle qui est suffisamment documenté. Les documents doivent expliquer clairement de quelle façon les résultats devraient être utilisés et permettre de déterminer quelles conclusions on peut tirer ou encore ce qui devrait être exploré plus en profondeur. En d'autres mots, un modèle explicable n'est pas une boîte noire. Il permet d'établir une relation de confiance entre le développeur et l'utilisateur des produits de la modélisation. Quoique semblable, cette exigence est distincte de la transparence. Cette dernière exige qu'il y ait de la documentation expliquant les raisons pour lesquelles le processus d'apprentissage automatique a été utilisé et la façon dont il a été utilisé.

Les lignes directrices pour l'explicabilité

- Documenter les explications, les graphiques et donner des exemples précis pour expliquer les résultats du processus. Confirmer avec les utilisateurs que vous fournirez des explications compréhensibles et convaincantes.
- Utiliser des outils de source ouverte pour faciliter l'interprétation de votre modèle si nécessaire. Une approche par déduction à rebours (*reverse-engineering*) peut aider à voir les relations importantes entre la structure des données et les résultats. Documenter toutes les techniques et les outils utilisés pour démontrer l'explicabilité du modèle ou le processus de modélisation. Inclure une explication de la relation entre les variables indépendantes (caractéristiques) et les résultats.
- Inclure une description de toutes transformations des variables (variables dérivées en fonction d'autres variables par exemple) ou des paramètres du modèle et les raisons pour lesquelles elles ont été effectuées (par exemple, pour améliorer l'exactitude).

Évaluation du processus d'apprentissage automatique et répercussions positives des lignes directrices

La mesure dans laquelle les processus d'apprentissage automatique à Statistique Canada doivent satisfaire aux exigences du cadre est déterminée par une autoévaluation, un examen par les pairs, une liste de contrôle, un tableau de bord ou une combinaison de ceux-ci.

Les **lignes directrices** sont un ensemble de recommandations traitant de chaque attribut au sein de chaque thème. Les lignes directrices devraient avoir des répercussions positives (et idéalement mesurables) sur les processus ou leur produit. Une réponse honnête aux questions de la **liste de contrôle** indiquera si, ou dans quelle mesure, le but des lignes directrices a été atteint. Le scientifique des données ou les responsables du processus peuvent se servir de la liste de contrôle comme une **auto-evaluation** lors de la planification et de l'élaboration des différentes phases afin de s'assurer que rien n'a été oublié ou négligé. Un **examen par les pairs** est nécessaire pour passer du prototype d'apprentissage automatique en mode de production. La liste de contrôle sert à guider l'évaluateur, mais ne vise pas à le restreindre ou à limiter l'examen d'une quelconque façon. Lorsque de nouvelles méthodes ou techniques ou de nouveaux outils sont utilisés, ou lorsque les processus d'apprentissage automatique sont présentés aux directeurs des principaux programmes statistiques, nous recommandons que la méthodologie et la liste des réponses aux questions soient approuvées par un comité d'experts, similaire à l'un des comités de revue scientifique en méthodologie. La version automatisée de la liste de contrôle pourra être accessible au moyen d'un **portail** à partir duquel les scientifiques des données ou les responsables du processus pourront répondre aux questions. Les réponses seront saisies dans une base de données et elles seront regroupées et présentées dans un **tableau de bord** au sein du programme, de la Direction générale, de l'organisme, ou de tout niveau, à n'importe quelle fréquence, pour les besoins de la gestion interne des ressources et l'assurance de la qualité. La production de rapports pour des clients externes (produits de diffusion) devrait aussi pouvoir être analysée de la même façon, bien que d'autres documents puissent être requis.