Catalogue no. 89-20-0006

Framework for Responsible Machine Learning Processes at Statistics Canada July 2020

Release date: May 3, 2021

Statistics Canada Statistique Canada

Canada

How to obtain more information

For information about this product or the wide range of services and data available from Statistics Canada, visit our website, www.statcan.gc.ca.

You can also contact us by

Email at STATCAN.infostats-infostats.STATCAN@canada.ca

Telephone, from Monday to Friday, 8:30 a.m. to 4:30 p.m., at the following numbers:

• Statistical Information Service	1-800-263-1136
• National telecommunications device for the hearing impaired	1-800-363-7629
• Fax line	1-514-283-9350

Depository Services Program

• Inquiries line	1-800-635-7943
• Fax line	1-800-565-7757

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner. To this end, Statistics Canada has developed standards of service that its employees observe. To obtain a copy of these service standards, please contact Statistics Canada toll-free at 1-800-263-1136. The service standards are also published on www.statcan.gc.ca under “Contact us” > “Standards of service to the public”.

Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued co-operation and goodwill.

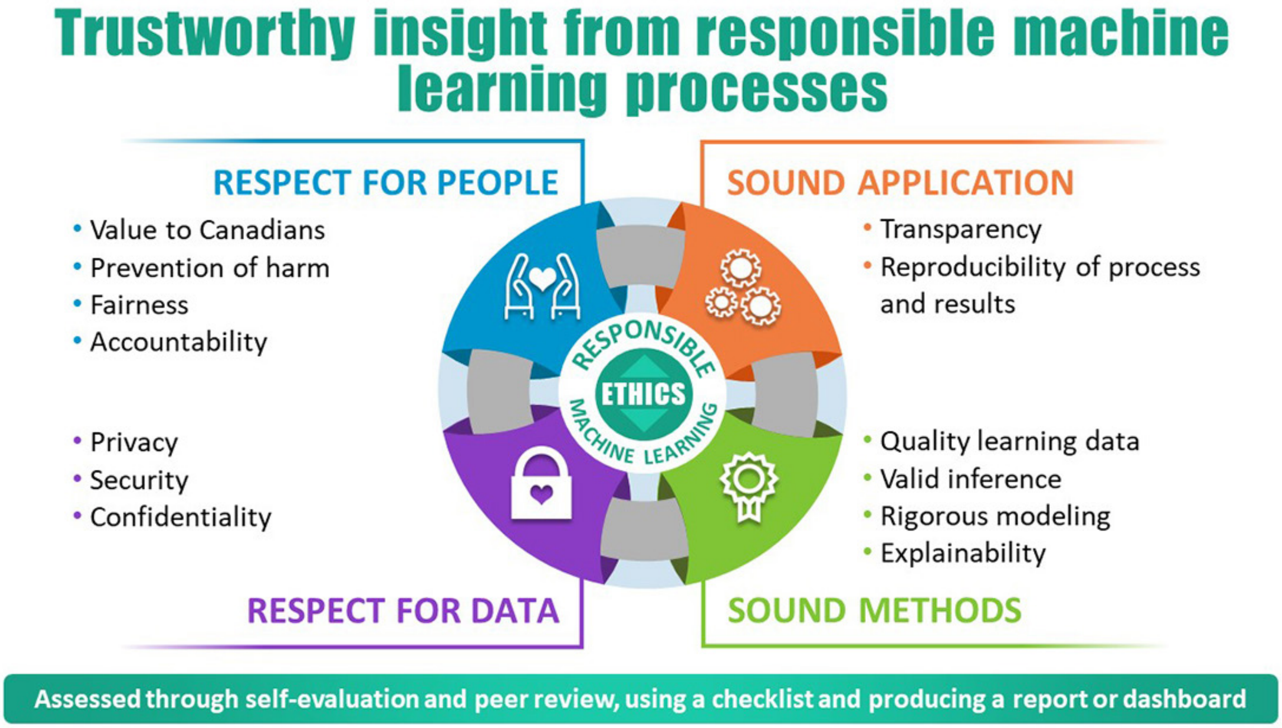
Published by authority of the Minister responsible for Statistics Canada

An HTML version is also available.

Cette publication est aussi disponible en français.

Table of contents

Framework for Responsible Machine Learning Processes at Statistics Canada July 2020



Introduction

This document is a handbook for practitioners developing and implementing Machine Learning (ML) processes. It provides guidance and practical advice on how to responsibly develop these automated processes within Statistics Canada but could be adopted by any organization. They can be applied to processes that are put in production or that are dealing with research.

It is the Manager's responsibility to ensure (in partnership with the appropriate service providers such as data scientists, methodologists or system's staff) that the Guidelines are followed, and the checklist is completed at the appropriate times throughout the lifecycle of the ML project.

Background

Machine learning is the science of developing algorithms and statistical models that computer systems use in order to perform a specific task effectively, without using explicit instructions, relying on patterns and inference instead. ML algorithms build a mathematical model based on learning from data, in order to make predictions or decisions without being explicitly programmed to perform that specific task.

Machine learning can improve the effectiveness of existing systems, provide greater efficiencies in operations, and enhance decision making. The rapidly growing capabilities and increasing presence of ML processes in our work at Statistics Canada raise pressing questions about the impact, governance, ethics, and accountability of these technologies. As indicated in the Government of Canada's Principles of Responsible Artificial Intelligence, (AI- which is inclusive of ML), the government will endeavor to:

1. understand and measure the impact of using AI (including ML) by developing and sharing tools and approaches
2. be transparent about how and when we are using these processes, starting with a clear user need and public benefit
3. provide meaningful explanations about decision making, while also offering opportunities to review results and challenge these decisions
4. be as open as we can by sharing source code, training data, and other relevant information, all while protecting personal information, system integrity, and national security and defense.

In order to put this in place, a Directive on Automated Decision-Making has been written and is based on the results from the Algorithmic Impact Assessment tool used to determined how acceptable are the AI solutions put in place from an ethical and human perspective.

To this end, a Framework for Responsible ML Processes has begun to be developed at Statistics Canada. The Framework consists of Guidelines for responsible ML and an accompanying checklist, which are organized into four themes: Respect for People; Respect for Data; Sound Methods; and Sound Application. All four themes work together to ensure ethical use of both the algorithms and the results.

Scope

These Guidelines apply to all statistical programs and projects conducted by Statistics Canada that use ML algorithms, including supervised and unsupervised learning algorithms. These can be administrative data programs, surveys, macroeconomic accounts, censuses, analytical studies or even experimental projects. The ML can be outward facing or internal to the program. These Guidelines offer unified guidance with key principles and themes, and a checklist to assist the ML practitioner in assessing their process. It is anchored in the vision to create a modern workplace culture and to provide direction and support to those using machine learning techniques.

The validation of ML processes will be completed by either self-assessment, peer review, committee, or a combination thereof.

These guidelines are consistent with the Statistics Canada Policy on Scientific Integrity (which states that all scientific and research activities are carried out in a manner that is consistent with all relevant and applicable standards of scientific excellence, research ethics, and responsible research conduct) and reflect the core values of Statistics Canada (valid statistical inference; quality; rigor). They are to be used as a complement to the Quality Guidelines, the Proportionality Framework and other instruments in the Statistics Canada Policy Suite. It is assumed that good practices for documentation, quality assurance and performance measurement reporting will also be followed, without specific instruction from these Guidelines.

Theme: Respect for People

At Statistics Canada we aim to make efficient use of government resources while producing information that helps Canadians better understand their country. This theme has four attributes: value to Canadians; prevention of harm; fairness; and accountability.

> The concept of **value to Canadians** in the context of ML is that the use of ML algorithms creates added value, whether in the products themselves or through greater efficiency in the production process.

Guidelines for Value to Canadians

- Ensure the ML algorithm has a clear benefit for the data users.
- Ensure the ML algorithm results in fit-for-use statistical products.
- Ensure the relevance of the ML algorithm.

In the context of machine learning processes at Statistics Canada, harm could come to vulnerable populations if sensitive information about them is released to the public. **Prevention of harm** in this context is being aware of the potential harm and having meaningful dialogue with stakeholders, spokespersons and advocates prior to development or release.

Guidelines for Prevention of harm

- Maintain sensitivity toward vulnerable populations.

Fairness implies that the principle of proportionality between means and ends is respected, and that a balance is struck between competing interests and objectives. Fairness ensures that individuals and groups are free from unfair bias, discrimination and stigmatisation.

Guidelines for Fairness

- Ensure that all training data, code and tools used in the ML process are acquired, maintained and used in accordance with existing protocols.
- Ensure that all variables in the model are necessary to the outcome, and that none could cause unfair bias, discrimination or stigmatisation.
- ML processes must protect the integrity and confidentiality of data.
- Ensure that the development team does not inadvertently build unfairness into ML processes. This could be accomplished through “bracketing”, whereby team members reflect on and set aside their personal biases, preconceptions and experiences. This will ensure that the development and use of ML is compatible with maintaining social and cultural diversity and does not restrict the scope of lifestyle choices and personal experiences, and that development team members take the opportunity to anticipate potential adverse consequences of using the ML process.

Accountability is the legal and ethical obligation on an individual or organisation to be responsible for its activities and to disclose the results in a transparent manner. Algorithms are not accountable; somebody is accountable for the algorithms.

Guidelines for Accountability

- Ensure someone has been identified as being responsible and accountable for the ML process and its outcomes through all the phases (development, deployment and production).
- Ensure there is a plan for monitoring performance and maintenance of software throughout product lifecycle.
- Ensure an audit trail of recommendations and decisions exists.

Theme: Respect for Data

At Statistics Canada we take data seriously. This theme has three attributes: privacy of the people to whom the data pertain; security of information throughout the data lifecycle; and confidentiality of identifiable information.

Privacy is the right to be left alone, to be free from interference, from surveillance and from intrusions. When acquiring sensitive information, governments have obligations with respect to the collection, use, disclosure, and retention of personal information. Privacy generally refers to information about individual persons (definition from the Policy on Privacy and Confidentiality).

Guidelines for Privacy

- Ensure that privacy risks are minimized by carrying as few personal identifier variables as necessary through the process.
- Ensure that relevant instruments in the Policy Suite are followed, in particular The Directive on Conducting Privacy Impact Assessments.

Security is the arrangements organizations use to prevent confidential information from being obtained or disclosed inappropriately, based on assessed threats and risks. Security measures also protect the integrity, availability and value of the information assets. This includes both physical safeguards, such as restricted access to areas where the information is stored and used, and security clearances for employees, as well as technological safeguards to prevent unauthorized electronic access (definition from the Policy on Privacy and Confidentiality).

Guidelines for Security

Ensure that sensitive information is secure through compliance with The Directive on the security of sensitive statistical information.

Confidentiality refers to a protection not to release identifiable information about an individual (such as a person, business or organization). It implies a “trust” relationship between the supplier of the information and the organization collecting it; this relationship is built on the assurance that the information will not be disclosed without the individual’s permission or without due legal authority (definition from the Policy on Privacy and Confidentiality).

Guidelines for Confidentiality

- Ensure that confidentiality is protected through compliance with the policy on privacy and confidentiality.

Theme: Sound Application

Sound application refers to implementing, maintaining and documenting machine learning processes in such a way that the results are always reliable and the entire process can be understood and recreated. This theme has two attributes: transparency and reproducibility of process and results.

Transparency refers to having a clear justification for what makes this particular algorithm and learning data the most appropriate for the application under study. To achieve transparency, ML developers should be making comprehensive documentation, including making code accessible and available to others, without compromising confidentiality or privacy.

Guidelines for Transparency

- Clearly communicate to end users how, where and why Machine Learning was used in the process, including a description of the learning data, how the algorithm works and the model diagnostics that are used. Disclose any bias/limitations in the data. Share code when applicable.
- Ensure that all partners in Subject Matter, Data Science, Methodology and IT are involved as appropriate in the development of the Machine Learning models.

Reproducibility of process means that there is sufficient documentation and code sharing such that the ML process could be recreated "from scratch". **Reproducibility of results** means that the same results are reliably produced when all of the operating conditions are controlled. There are no ad hoc or human intervention steps that could alter the results.

Guidelines for Reproducibility of process and results

- Ensure that a corporately supported code version control system (for example GitLab) is used to version-control all code written for the development, testing and implementation of the machine learning algorithm.
- Ensure that the following information regarding the development, testing and implementation of the machine learning algorithm is "bundled" together so that (current or future) stakeholders will have sufficient information to fully re-produce results whenever needed:
    - ▶ All pertinent code
    - ▶ Versions of all software tools used
    - ▶ Exact version of input data
    - ▶ All relevant intermediate and final results, diagnostics, log files
- Ensure that the final processing/analytical pipeline related to the development, testing and implementation of the machine learning algorithm – one that fully automatically (re-)generates all intermediate and final results, all diagnostics, and all log files, with no human intervention – can be executed by a stakeholder via the execution of a single "master" script/programme.
- Ensure the instructions (which should be nearly trivial) for how to execute the master script are well documented and included in the bundle mentioned above.
- Ensure that the development and use of ML is carried out in an energy efficient and environmentally sustainable manner, for example by minimizing the printing of documentation and outputs or by optimizing computer processing.

Theme: Sound Methods

Sound methods are those that can be relied on to efficiently and effectively produce the expected results. Typically we follow recognized protocols involving consultation with peers and experts, documentation and testing in developing sound methods. This theme has four attributes: quality of learning data; valid inference; rigorous modeling; and explainability.

In the context of ML processes, the **quality** of **learning data** is measured by both the consistency and the accuracy of labeled data. Coverage, meaning that the labels and descriptions cover the entire span of what the ML algorithm will encounter in production is also important to reduce the risk of bias or discrimination (fairness), and representativity in terms of the distribution of the input or feature variables is important for realistic measures of performance.

Guidelines for Quality learning data

- In our context, often the challenge is not bias or discrimination in the learning data, but rather a lack of labeled data to use for training, validation and testing. If resources for labeling data are tight, this could be a significant hurdle to overcome. Ensure early in the development phase that sufficient and representative labeled data exists, and if not, that adequate funding has been allocated to the labeling activity.
- Another real-life problem with labeled data is its quality. Poor quality labeled data has some or many labels erroneously assigned. If poor quality training data is used in the ML algorithm, the result could be poor convergence, high prediction error, and even bias. Use manual review of a representative sample of labeled data to measure rates of accuracy and consistency. Describe any known under coverage in the learning data relative to the target population. Report the distribution of labels and key independent variables (features) in the training data and relative to the target population to give an indication of representativity.
- Monitor consistency, accuracy, coverage and representativity of learning data over time to detect any drift or slippage in quality.
- It is important to have a good understanding of the learning data, to be able to assess to what extent it represents the entire population you expect to process. If the learning data comes from a survey, ensure you know the target population, the sample design, how the sample was selected, how the data was collected, and the nonresponse patterns across domains of the population. If the learning data comes from an administrative source, ensure you know the intended target population, coverage, and what pre-processing, editing and imputation has already been done. Consider how any aspect of the learning data could be a potential source of bias, for example given the sample design and nonresponse patterns, should weights be used in the ML algorithm? Document fully the source of the learning data and your analysis of potential bias.

Valid inference refers to the ability to extrapolate based on a sample to arrive at correct conclusions with a known precision about the population from which the sample was drawn. In the ML context, valid inference means that predictions made on never-before-seen data using the trained model are reasonably close to their respective true values in a high proportion, or in the case of categorical data, predictions are correct in a high proportion.

Guidelines for Valid inference

- Choose a validation or diagnostic method and associated metrics that are appropriate given the algorithm you are using and the context in which it will be applied. Consider the quality requirements of the ML process itself and of the statistical process in which it is used. If the ML algorithm is a prediction (regression or classification) algorithm, then evaluation metrics would be some form of prediction error as well as an assessment of stability (the extent of agreement between predictions from one execution to another of the same algorithm). In the case of clustering, stability measures are important. Instability is a sign of clusters not well separated. Generic validation protocols and metrics may or may not be appropriate for the problem at hand. Establish appropriate performance benchmarks respecting the quality requirements for end products and the amount of uncertainty the ML process contributes.
- Execute the chosen validation or diagnostic method and examine whether the model fulfills the prescribed benchmarks. When trying to achieve the benchmarks, it may often involve hyper parameter tuning. It should be done using a systematic approach to cover a wide range of values. The choice of hyper parameters to tune should be based on knowledge and with some intent in mind. If the metrics indicate poor performance consider the possibility that you need additional learning data. Consult with an experienced machine learning practitioner as necessary.

- For a supervised machine learning algorithm, ensure that a testing data set is set aside early on in the development cycle and is never used until the final evaluation of the optimized model. For an unsupervised learning algorithm, results are more subjective since the truth is unknown; one needs to report this information.
- When learning data is coming from a survey sample drawn randomly, carefully consider the use of weights for specific algorithm.
- Use a software package manager system, which includes both R and Python packages. It is recommended to only use R and Python packages that have a huge user base, who would have caught any errors. If you are considering using a software package or module from another source, do rigorous testing and consult with user communities and ML experts to ensure that the package is of good quality and that results are correct and as expected. Include the results of your research and findings in a central repository.
- Track performance metrics through time so as to notice when a model needs to be re-trained.

Rigorous modeling in ML means ensuring that the algorithms are verified and validated. This will enable users and decision makers to justifiably trust the algorithm in terms of fitness for use, reliability and robustness.

Guidelines for Rigorous modeling

Rigorous modeling will ensure that learning data is a valid representation of the machine learning application's population. It also ensures that the model is applied within the proper boundaries, as defined by the learning data. Generalization error is the expected error on new or never-before-seen data. Generalization error can go in two directions; there can be underfitting of the model (sometimes called bias by the ML community), i.e. if the model is not complex enough (has too few parameters). On the other hand, generalization error can also come from overfitting of the model (sometimes called variance by the ML community), if the model is too complex and does not capture the generic trend across the data, but instead wraps itself too tightly around the training data. In both cases the error metric based on the training data is much smaller than the error metric based on the testing data. Ensure rigorous modeling is done and that the testing data set is chosen to be representative of the population so that the resulting testing error should be a reliable/reasonable estimate of the *generalization error* of the optimized model, specifically with respect to how the optimized model is actually deployed in the production setting.

A model that is **explainable** is one with sufficient documentation that it is clear how the results should be used, and what sorts of conclusions or further investigations can be supported. In other words, an explainable model is not a black box. This builds trust both on the part of developers and users of the model's outputs. This is similar to yet distinct from transparency, which requires there to be documentation explaining why and how the machine learning process was used.

Guidelines for Explainability

- Use text explanations, visualizations and specific examples to explain the outputs of the ML process. Confirm with end users that you are providing explanations that they understand and find convincing.
- Use open source tools to aid with interpreting your model if necessary. A reverse-engineer approach can help to see important relationships between structure and output. Document all techniques and tools used to aid the interpretability of the model/modelling process. Include an explanation of the relationship between the independent variables (features) and the outcome.
- Include in the explanation a description of any transformation or engineering of features (variables) in the model and why they were done (for example to improve accuracy).

Assessment of the ML process and positive impact of the Guidelines

The degree to which machine learning processes at Statistics Canada meet the requirements of the Framework is assessed through self-evaluation, peer review, a checklist, a dashboard or a combination thereof.

The **Guidelines** are a set of recommendations addressing each attribute within each theme. Guidelines are formulated so as to have a tangible (and ideally measurable) positive impact on the process and/or its product. The **checklist** is questions whose honest response will indicate if or to what extent the intention of the guidelines has been realized. The ML developers and/or process owners can do the checklist as a **self-assessment** while in the planning and development phases to ensure that they did not forget or overlook anything. A **peer-review** is required for passing from the prototype to production version of the ML process. The checklist questions serve to guide the peer-reviewers, but are not meant to restrain or limit the review in any way. When new methods, techniques or tools are being used, or when ML processes are introduced to key statistical programs, we recommend that the methodology and checklist question responses be approved by a Committee of experts, such as one of the Methodology Scientific Review Committees. The automated version of the checklist will be a **portal** where ML developers and/or process owners answer the questions; the responses will be captured in a database; and responses will be aggregated and reported in a **dashboard** at the program, branch, corporate, or any level, at any frequency, for internal management of resources and quality assurance. Reporting for external audiences (dissemination products) should also be possible from the same system, although additional documentation may be required.