unicef
for every child

EVALUATION

Methodological guidance

May 2024

An operational framework for Machine Learning in evaluation

An operational framework for Machine Learning in evaluation

United Nations Children's Fund
Three United Nations Plaza
New York, New York 10017

May 2024

The report was drafted by Nabamallika Dehingia, with inputs from Eduard Bonet Porqueras, Uyen Huynh, Miguel Almanzar, Francesco Iacoella, and Zlata Bruckauf.

For further information, please contact:

Evaluation Office
United Nations Children's Fund
3 United Nations Plaza
New York, NY 10017, USA
evalhelp@unicef.org

Contents

Summary

Machine learning (ML) analytical approaches have significantly advanced our capacity to process large amounts of data, allowing for the systematic analysis of both quantitative and qualitative data in a timely and cost-effective manner. Over the past decade, these methodologies have found widespread utilization in development research, including in evaluations. In this paper, we present an operational framework for applying ML approaches in evaluation studies. The framework is based on a literature review and UNICEF Evaluation Office's (EO) experience with using ML in a few recent global evaluations. The framework identifies three broad categories of data types, and corresponding ML methods that can be used to answer evaluative questions. This report offers practical guidance to evaluation managers who are interested in applying these innovative approaches to their projects.

Key terminology

Structured data: A dataset that is defined, formatted, and can be analysed, such as survey data and other quantitative data.

Unstructured data: A dataset difficult to organize, search, and analyse, such as text data in reports and statements, photos, videos, etc.

Big data: Data that is large, complex, and expensive for traditional database systems to manage and analyse. These are continually generated digital data, such as financial transactions, internet searches, social media, and satellite data – sometimes referred to as 'frontier data.'

Artificial intelligence: AI is the umbrella term used to describe machines or models that can mimic human intelligence. In general, AI systems work by consuming a lot of data and using these data to make predictions (e.g., chatbots like ChatGPT, Gemini, other AI intelligence platforms and products).

Machine learning: ML broadly refers to the algorithms that can train a machine to learn from provided inputs, or mimic the human brain; they are the statistical methods and models instrumental in constructing AI systems.

Natural language processing: NLP pertains to the 'text analytics' arm of ML, covering statistical models used to analyse text data.

Supervised ML model: Supervised models are algorithms that learn from labelled data and make accurate predictions or decisions for unlabelled data. More details provided in the pilot project section.

Unsupervised ML model: Unsupervised models do not require any human input or labelled data to make decisions. They automatically identify patterns in the input data to group them into different categories. More details provided in the pilot project section.

Note: The paper focuses on ML methods (including natural language processing; NLP), instead of AI systems in general. Popular generative AI tools such as ChatGPT are not discussed. Generative AI tools often come with the 'black box' issue: a lack of information around the back-end processing and handling of data fed into the tools. As such, recommendations, and guidance on the use of these AI tools warrant a separate discussion with a greater emphasis on ethics and safety.

Introduction

The past decade has seen many applications of ML methods in the field of development research, offering new avenues for faster and systematic analysis of data. Responding to an ever evolving and increasingly complex development landscape, researchers and policy makers in the past two decades have demonstrated successful use of ML approaches to answer a variety of complex research and policy questions. Noteworthy applications include poverty estimation, where statistical models leverage publicly available geospatial satellite data to predict poverty levels in communities (Jean et al, 2016). Prediction and estimation of climate shocks, slum area boundaries and human movement using satellite imagery and mobile phone data are few other successful demonstrations of ML approaches (Hemphill et al, 2022; Lai et al, 2019; Luo et al, 2022).

With its capacity to swiftly process diverse and complex datasets, these methods have a lot to offer to the field of evaluation as well (Bravo et al, 2023). This is particularly relevant for multi-country evaluations, and evaluations in fragile settings, where traditional approaches such as survey data collection and qualitative interviews, can prove to be inadequate and extremely resource intensive. Rapid analysis of secondary and non-traditional sources of data using ML analytics can strengthen the evidence derived from traditional evaluation methods. For example, the large amount of publicly available big data such as mobile phone data and satellite imagery presents a remarkable opportunity for cost-effective learning. For international organizations, including UNICEF, country offices generate an abundance of textual monitoring and reporting data. ML models can analyse such quantitative as well as qualitative data, generating meaningful and timely evidence for evaluations.

In addition to producing new and insightful evidence, ML methods can be used to improve efficiencies within evaluation teams. These approaches can be adopted for automating analytical tasks that are often repeated across evaluations. For example, as part of the formative stage in most evaluations, previous reports are reviewed to gather insights. Review of programme documents and monitoring and expenditure data is also routinely carried out as part of the evaluation process. By developing ML solutions that can be replicated across thematic areas and data types, the number of resources and time spent in conducting these regular tasks can be reduced significantly.

In this current context of growing need for cost-effective and robust alternative evaluative methods, and rapid advancements in AI technology, this paper aims to explore the diverse applications of ML in various evaluation tasks. It seeks to contribute to the larger discourse surrounding the potential of ML-driven approaches in shaping effective and efficient evaluations. Specifically, the paper offers practical guidance on operationalizing these methods. Guidance is based on available literature as well as UNICEF Evaluation Office's (EO) learnings from implementing multiple pilot projects related to ML analytics. The paper also highlights key challenges, including technical capacity of evaluation teams, considerations regarding data privacy and confidentiality, sustainable adoption of these methods, and convergence of methodological innovations and evaluation objectives.

The paper is structured as follows: the next section presents three case studies, or pilot projects, that demonstrate UNICEF EO's work on application of ML, and corresponding learnings. This is followed by an introduction to an operational framework, developed to guide the use of ML for answering evaluation questions. The framework builds on current literature on ML and learnings from our pilot projects. It explores diverse ways in which these advanced methods can be used for evaluation, alongside related limitations, and concerns. The final section presents a few key recommendations for future applications.

Operationalizing Machine Learning: Pilot projects

UNICEF's Evaluation Function has experimented with the application of ML analytics across different projects over the past few years.[1] In this section, three pilot projects are described, all of which are focused on using ML for analysing text data from internal UNICEF documents.

Traditionally, most evaluations involve the review of a large amount of text data, such as previous evaluation reports, planning and programme documents, and performance narratives. Analysing a large amount of text data is, however, extremely resource intensive. ML analytics can make such tasks quicker, more efficient, and robust. This is demonstrated by the following three pilot projects. Each example analyses text data regularly produced by UNICEF country offices: evaluation reports, country office annual reports/statements of progress, and country programme documents (CPDs) from UNICEF and UNFPA.

Pilot 1: Scoping and syntheses of evaluation reports

Project objective: To identify UNICEF evaluation reports that focus on the topic 'child access to justice', and synthesise findings from the identified reports

Data used: UNICEF evaluation reports

Contribution to evaluation: Scoping and inception phase; generate evidence necessary to identify evaluation scope and questions

Scoping and synthesis of previously published reports can be resource-intensive and time-consuming, especially given UNICEF's extensive repository of over 2,000 evaluation reports spanning the past two decades. In this pilot project, NLP models were used to predict which UNICEF evaluation reports focused on child access to justice. By replacing manual scoping and identification with NLP models, this task, which was implemented by one evaluation manager and one data scientist, proved to be cost-effective.

A supervised NLP model was used for the specific task of identifying evaluation reports on a single topic (see Figure 1). Supervised models are a foundational technique in AI analytics, enabling algorithms to learn from labelled data and make accurate predictions or decisions for unlabelled data. The term 'supervised' refers to the fact that during the training or building process, the algorithm is provided with labelled examples which serve as a guide for the algorithm, allowing it to understand the relationship between the inputs and the correct outputs. In essence, the algorithm generalizes from this labelled data, identifying patterns and correlations that enable it to make predictions or classifications on new, unseen data.

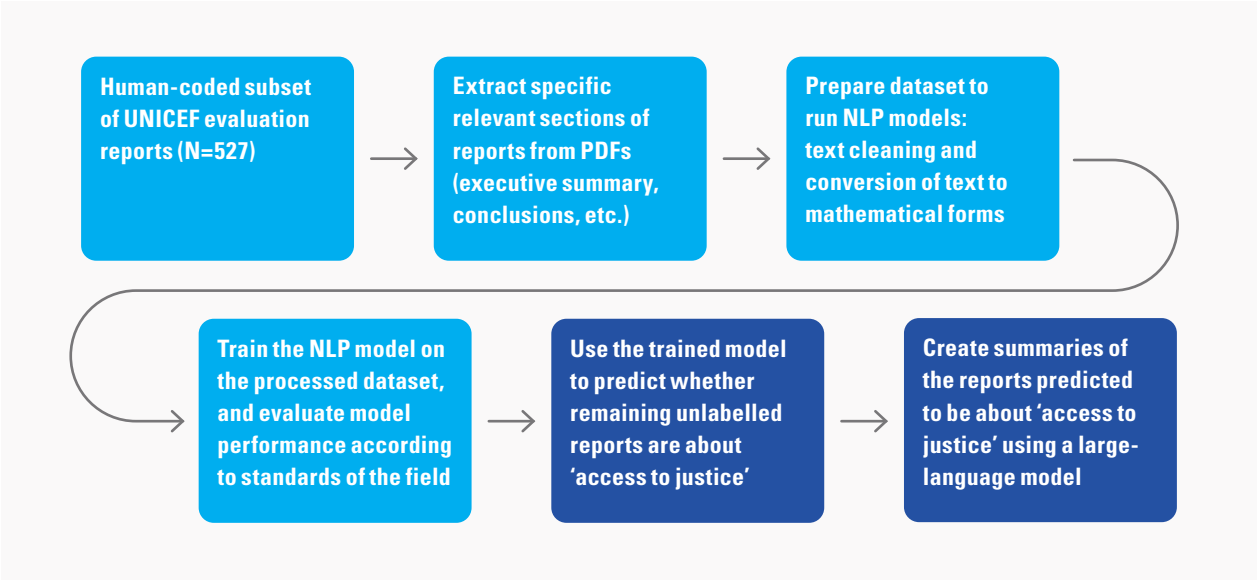
To illustrate, consider an example with a set of emails, with each email labelled as 'spam' and 'non-spam' by researchers/humans. Through supervised methods, a model is trained on this dataset, analysing features like keywords, sender information, and email structure, to find the specific features or word patterns that separate 'spam' emails from 'non-spam' emails. The training process involves iteratively adjusting the algorithm's parameters to minimize the difference between its predictions and the actual labels (spam versus

1 Some examples include a real-time evaluation of UNICEF's response to the humanitarian crisis in Ukraine in 2022-2023, which used NLP approaches to analyse a wide variety of internal reports. The same evaluation also examined sentiments towards refugees using publicly available social media data. In another syntheses of UNICEF evaluations conducted in 2023, advanced text mining and machine learning methods were used to gather insights from many reports.

non-spam). This optimization process, often referred to as 'learning,' continues until the algorithm achieves a satisfactory level of accuracy in its predictions. Once trained, the model can accurately predict whether new emails are spam or not, based on the identified and learned features from the training process.

This pilot project accesses a dataset of 527 UNICEF evaluation reports labelled as 'access to justice' or 'not on access to justice' by experts, as a part of a separate exercise. This served as the basis for running the supervised classification model. The models showed around 98 per cent accuracy; these adequately accurate models were used to make predictions for the remaining unlabelled evaluation reports.

Figure 1. Steps involved in predicting and summarizing UNICEF evaluation reports on 'access to justice'



Text data extraction and processing

One of the critical steps in applying ML methods for synthesis of evidence from documents is the extraction and processing of the text data. The process of text extraction was automated as much as possible, by developing a code on Python that downloaded all the evaluation reports from UNICEF's internal database using APIs, and extracted separately, different sections of the reports (Introduction, Executive Summary, Conclusions, etc.). However, due to inconsistencies in the structure of evaluation reports, sections of 40 per cent of the documents had to be extracted manually. Sections from all non-English reports were also translated to English using the Google Translate API. All analysis was carried out in Python.

Although there are many paid AI tools that support data processing as well as analysis of large documents (e.g., Microsoft Azure AI services), owing to limited resources, no paid software was used for text data processing. Given the availability of funds, evaluators can explore the potential use of such platforms, which can further improve efficiencies and make analysis of text data easier.

The models predicted 59 evaluation reports to be focussed on access to justice for children. Following the identification of the relevant reports on access to justice by the supervised model, the text from these reports were synthesized using a large-language model - Bi-directional Encoder Transformation Model (BART). This model has been pre-trained on English language and is often used to summarize documents. BART performs abstractive summarization for large texts, by learning the entire document and generating paraphrased text to summarize the main points. This task of creating summaries for reports using ML proved to be efficient; two experts reviewed the summaries and the corresponding reports to check for inconsistencies. Around 6 or 7 of the 59 summaries missed some critical information. However, for a summarizer model to be used at scale, it must be of adequate accuracy, to avoid any errors in decision-making by evaluators. It was thus decided not to replicate this method of summarizing evaluation reports with a large-language model for other topics or evaluations. Given the current rapid improvements in AI tools such as ChatGPT, Claude, Gemini, etc., evaluation groups should consider testing the efficiency of these tools in summarizing reports.

Pilot 2: Mapping of UNICEF programme activities on violence against girls, boys and women (VAGBaW) across regions and countries

Project objective: To identify UNICEF country-level programmes and activities on VAGBaW using internal monitoring text data

Data used: UNICEF country statements of progress

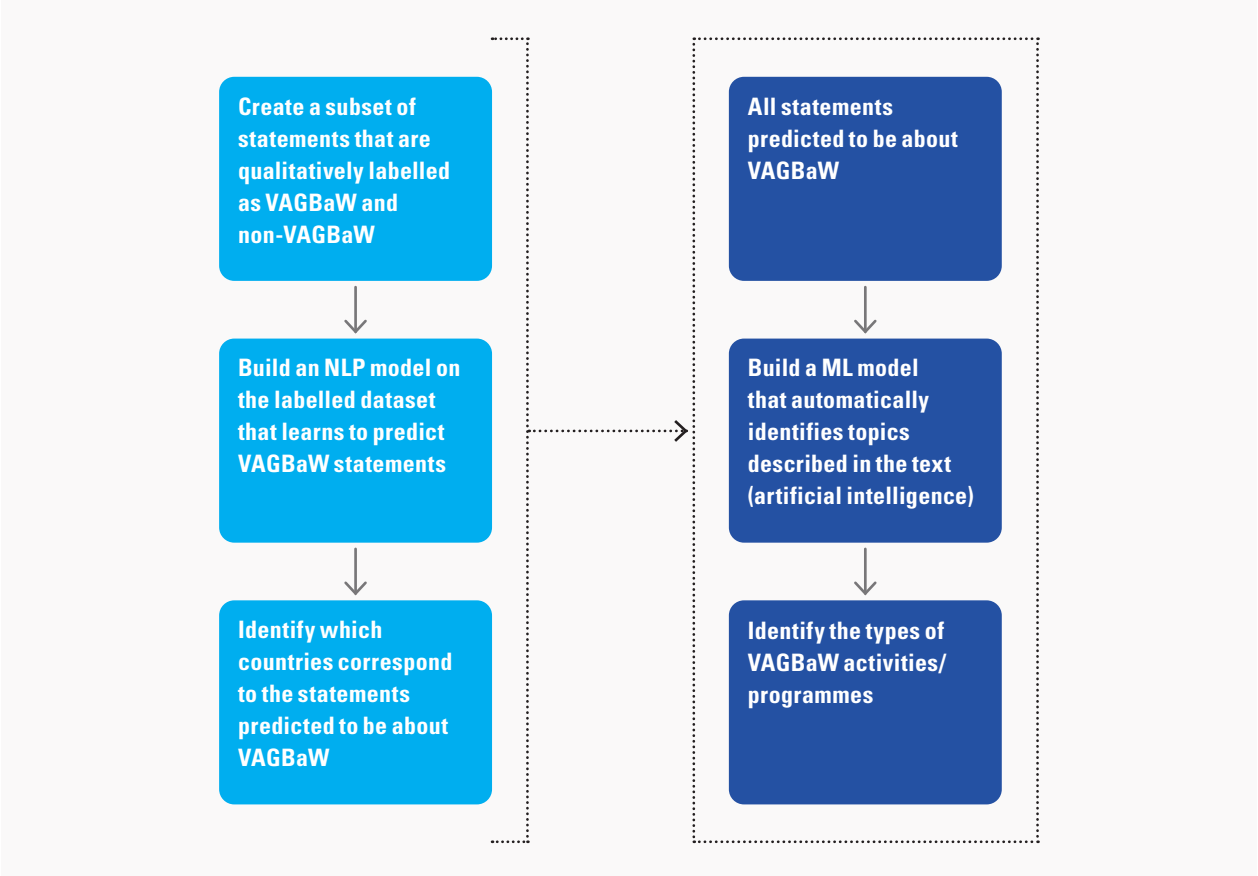
Contribution to evaluation: Scoping phase; generate evidence related to implemented programme activities

This pilot was implemented as part of the formative phase of a global evaluation of UNICEF's efforts on addressing VAGBaW, which aimed to map different activities related to VAGBaW implemented by UNICEF country offices.

The data used for this analysis was 'country output statements of progress.' All UNICEF country offices are required to report on their annual progress, based on the institutional framework of outputs and outcomes. For each planned output, a narrative is provided describing the activities implemented and progress achieved at the end of the year. These represent a large amount of text with rich information about the country office's programmes and activities. For the reporting year 2022, 13,708 progress statements were reported by UNICEF country offices. Each statement included an average of 408 words. The objectives were to 1) identify which statements described any VAGBaW activity, and 2) understand the specific type of VAGBaW programme activity described by the identified statements. To that end, two separate NLP models were used. A supervised model was used for the first task, followed by an unsupervised model for the second (Figure 2). A supervised model, as previously described, requires a pre-labelled dataset; two experts coded a set of 1,589 statements as 'VAGBaW' or 'non-VAGBaW.' The supervised model trained on this dataset was then used to predict the label of the remaining statements of progress. The results allowed for the identification of UNICEF country office implementing any activity on VAGBaW.

Next, to identify specific VAGBaW programme activity elements, an unsupervised model was used on the set of identified VAGBaW statements. Unsupervised models work without any pre-labelled or training dataset. This model learns, by itself, word patterns and occurrences, and groups the large number of statements into different categories. Each category represents a coherent topic. As model output, a list of keywords representing each category of statements is received. By reviewing the key phrases closely, one can infer the topic description. For instance, in the analysis, one of the reported categories included the keywords 'capacity development', 'trainings', 'train officers', etc. Given the broad understanding of VAGBaW programmes in UNICEF, it was possible to infer that this category of statement alludes to capacity development activities for VAGBaW reduction.

Figure 2. Steps involved in mapping different VAGBaW activities implemented by UNICEF country offices, using supervised and unsupervised machine learning models.



Pilot 3: Tracking country office priorities using intelligent search on country programme documents

Project objective: Analysis of changes in national priorities

Data used: UNICEF and UNFPA Country Programme Documents, from 72 countries, approved between the years 2010-2022

Contribution to evaluation: Generate evidence to answer one of the evaluation questions on programme coherence

This analysis was conducted across multiple evaluation projects covering thematic topics: child marriage, out-of-school children, and primary health care strengthening for improved child health outcomes. The steps included are described in one of the evaluations where it was first applied - the UNFPA-UNICEF Global Programme to End Child Marriage (GPECM) evaluation. The specific question to be answered by this analysis was whether the GPECM contributed to prioritize the elimination of child marriage across the programme countries. UNFPA and UNICEF from the 12 GPECM countries were analyzed and compared with those from 60 other countries, where child marriage also had a high prevalence. CPDs are strategic documents signed between each organization and the national government of the country, and describe key issues and strategic priorities to protect and promote child rights. Despite being unstructured data (text PDF documents), CPDs from those 72 countries, between 2010-2023, were used as the key data source for this analysis because they follow the same structure and indicate the programme priorities at national level. This allowed to assess whether the GPECM had any influence over the national agendas by making the elimination of child marriage a national priority. The semblance of structure of the CPDs allowed for analysing and generating insights from these documents.

Using Python, data from CPDs produced between 2010 and 2022 was processed and extracted. A simple keyword search was used to identify CPDs that referenced child marriage in the programme priorities section. A list of keywords was then developed by the evaluation expert, which was used in a Python code to search through all the extracted text (keywords include child marriage, early marriage, etc.). The keywords were reduced to their root or stem words before being used in the search (e.g., child marriage= child marr). An 'intelligent semantic search' model was also tested in addition to the basic 'keyword search' model. However, the latter returned more relevant results. Intelligent search usually refers to a search system that includes an ML analytical model that searches for words/phrases similar to the keywords, and also takes into account the context of the keywords. For example, upon feeding search words 'social protection' to an intelligent search system, it might search for 'social security', 'cash transfers,' and other relevant topics. However, this advanced search system was not observed to be necessary for this analysis. This is likely because the list of keywords generated by the expert was comprehensive, and the language used in CPDs is consistent, in that child marriage would not be referenced in any indirect way in the text. Once those CPDs referencing child marriage were identified, temporal trends were examined to answer the original question whether the GPECM had any influence introducing in national agendas the priority of elimination of child marriage.

UNICEF Evaluation Office's learnings from the pilot projects

One of our goals with implementing the pilot projects was to evaluate the feasibility of using ML text classification models for accurate replication of manual content identification, and to assess the speed and efficiency of these models. The pilots show promising results, suggesting potential for replicability and scalability to other thematic areas. NLP methods, when applied to relevant internal text data, can provide quick results, and contribute to the evaluation. At the same time, this exercise highlighted important challenges and considerations for the future:

Acquire basic ML related knowledge: A key operational learning from this work is the important role played by managers in facilitating such initiatives. Managers should possess a foundational understanding of ML analytics, enabling collaborative efforts with technical experts to effectively explore the application of these advanced methods. This collaboration ensures the development of final products that are not only technologically robust but also practical and valuable for the evaluation process and the end user.

Choose replicability over complexity: The third pilot project highlights that there is not always a need for advanced ML models to obtain impactful findings. Basic text mining or keyword searches returned relevant and accurate results when investigating country priorities for specific topics using CPDs. This solution was extremely quick to implement and was replicated across different thematic areas and evaluations. When applying ML and NLP methods, it is important not to lose sight of the evaluation objectives and find the most effective as well as resource-saving option.

Capitalize on successful applications: For application of NLP models, one of the most crucial as well as complex and time-consuming tasks is processing of the datasets, i.e., conversion of the text documents into an 'analysable' format. In this context, where applicable, analytical methods that prove to be successful on one type of document should be replicated for different thematic areas. For example, keyword search with CPDs was implemented for two different topics after its success with child marriage. Additionally, once a specific type of document has been processed, it is worthwhile to brainstorm and figure out different types of insights that can be generated by analysing that text, to make the most of the time invested in processing these documents.

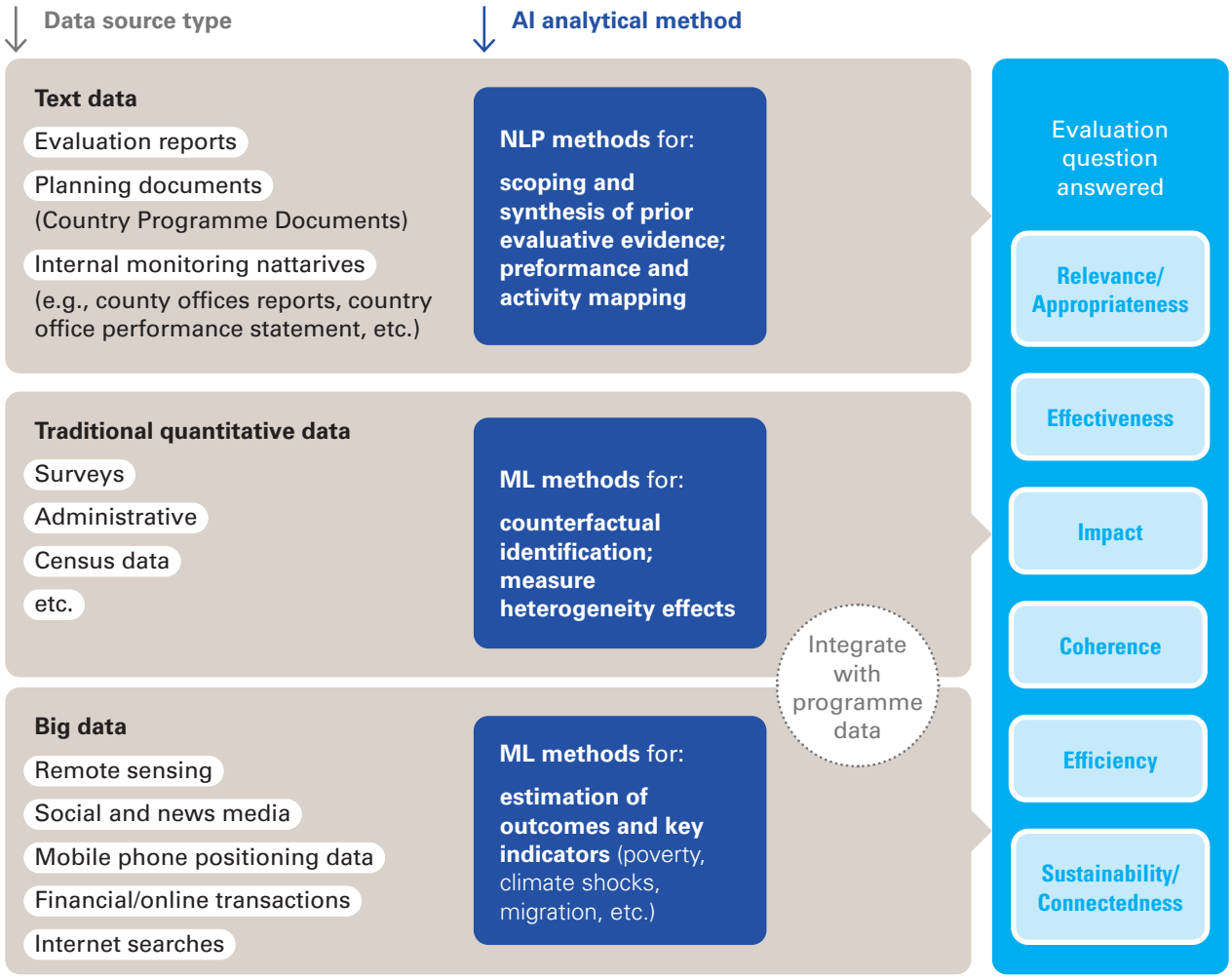
Explore relevant labelled dataset when using supervised machine learning models: The current method of using a supervised machine learning model requires substantial effort in creating a labelled dataset of reports classified by the topic area of interest. However, any such qualitative labelling work done in prior projects can be exploited to test more ML models. Alternatively, the presence of any open-source datasets where textual data has been classified by different topics should be explored. For example, Open-Source Approach to Classify Text Data by UN Sustainable Development Goals **(**OSDG) created a publicly available dataset of text that has been classified into the sustainable development goal (SDG) areas. Such datasets can be used to build distinct supervised NLP models to serve different evaluation-related needs.

A framework for use of Machine Learning in evaluation

Based on the learnings from the pilot projects described in the prior section, and existing relevant literature, we present in this section an operational framework for application of ML in evaluation studies (Figure 3).

While our pilot projects demonstrated different ways in which ML can be used to analyse reports and documents, similar methods can also be applied to a wide variety of structured quantitative data for evaluations. Broadly, the application of ML analytics in development research, and by extension, in evaluations, can be categorized by three types of data: a) big data, or frontier data, refers to large amounts of data produced very quickly and digitally such as satellite imagery, purchase transactions, news media, social media, digital text and imagery, etc.; b) traditional quantitative data used in research that includes household surveys, administrative data, census data, etc.; and c) text or unstructured data. The framework maps different uses of ML for evaluation, by the three types of data. These applications can answer relevant evaluation questions according to the OECD DAC criteria.[2]

Figure 3. Operational framework for application of ML analytics in evaluation



2 Most of the existing studies used ML to answer evaluation questions related to relevance, effectiveness, and impact of the programmes. No studies were found in literature that used these methods to assess coherence, efficiency, and sustainability. Although, synthesis of evidence from prior reports using NLP methods will likely cover these areas given the presence of such information.

Text data: As demonstrated by the pilot projects, NLP models can be used to scope and synthesize relevant documents. This can allow managers to cut costs and make evaluation tasks resource effective. Given the availability of relevant monitoring data such as performance narratives, NLP can also be used to identify and map programme activities and priorities across country offices. Many other organizations, including the World Bank, have used NLP to classify and synthesize evidence from project reports, showcasing effectiveness of these methods (Franzen et al, 2022). In a similar vein, the United Nations Development Programme (UNDP) has developed a search and analytics platform named Artificial Intelligence for Development Analytics (AIDA) that uses NLP methods to scope and synthesize findings from evaluation reports and other big data. Overall, this area of application indicates clear success in the context of evaluation projects, and evaluators should explore replication, or expansion of NLP methods described in the pilots.

Traditional quantitative data: ML methods have been used in several development research studies with traditional quantitative data such as household surveys, administrative data, and census data. Researchers used these models to identify risk factors for key outcomes, test causal relationships between variables, and more recently, to conduct small area estimations of different measures (Viljanen et al, 2022; Zivich et al, 2021). For evaluations (specifically impact evaluations), literature was found for two key categories of application: a) counterfactual identification, and b) measuring treatment heterogeneities (Özler 2022; McKenzie 2018).

There is a growing literature on 'causal machine learning', which refers to application of ML methods for answering causal inference related questions, i.e., impact related questions. Specifically, these methods focus on identification of well-matched samples, or counterfactuals, that can be used for impact evaluation studies (Tiffin, 2019). Counterfactual designs assess the impact of a programme by comparing a group that receives treatment with a control group over time, employing experimental or quasi-experimental methods to control for both observable and unobservable causal factors. However, this form of comparison may not always be practical or preferable, due to critical challenges in identifying a control group that is 'similar' to the programme group. Achieving a suitable balance between treatment and control groups is challenging in practice, especially when active samples, such as specific social groups or geographic areas, exhibit structural diversity. Matching techniques, including unsupervised machine learning, can be employed in such situations. There is evidence that suggests that machine learning algorithms of matching within a treatment effects framework can generate results that are robust (Linden and Yarnold, 2016). For example, Krief and DiazOrdaz (2019) used machine learning matching techniques to measure the impact of a national health insurance programme in Indonesia. The authors demonstrate a multivariate matching approach that uses ML to data-adaptively select balanced comparison groups.

Heterogeneous treatment effect estimation is a class of impact estimation strategy that allows for estimation of impacts that vary, given the values of a set of confounding variables. The challenge here however, in many cases, is choosing the appropriate groups (for heterogeneity estimation) or subdividing the observed individuals by large numbers of confounding variables – this can result in increasingly inaccurate heterogeneous treatment effect estimations. ML models can be used to identify the groups for estimating heterogeneity in treatment effects, using causal trees, X-learners, etc. (Athey et al, 2021; Athey et al, 2019; McKenzie, 2018). In this non-parametric method, the data is allowed to say which groups are likely to benefit from treatment.

Given the availability of data and a study design suitable for impact assessment (experimental or non-experimental), ML methods show promise in terms of generating robust results by strengthening counterfactual identification and treatment heterogeneity effects estimation. Evaluators should consider testing these methods to assess their robustness over conventional techniques in their specific evaluation context.

Big or frontier data: ML analytics have been applied with a wide variety of big or frontier data types, particularly in research studies. ML models have been used to leverage geospatial satellite data to predict poverty levels in communities, albeit with a certain margin of error (Chi et al, 2021; Jean et al, 2016). The basis of this analysis is the theory-based and empirically verified relationship between household poverty and housing quality which can be captured by satellite imagery data (presence of electricity, roof material, etc.). Researchers have also used mobile phone data to estimate wealth of individuals in different geographic contexts (Blummenstock, 2018). Such analyses have been shown to be useful in humanitarian assistance targeting, given the limited availability of updated poverty data in many low-income countries (Aiken et al, 2022). Other applications of ML with big data involve estimation of climate shocks, prediction of slum area boundaries with satellite data, social media analysis for measuring people's perceptions, and analysing human movement and migration using cellular and satellite data (Hemphill et al, 2022; Lai et al, 2019; Luo et al, 2022).

In evaluations, such analyses can address a key challenge – the limited availability of quantitative data on outcomes or outputs at a population level. Lack of quantitative population-level data often leads to evaluations having to rely on key informant interviews and assessment of programme monitoring data to answer the evaluative questions. Such methods can be time consuming, expensive, and vulnerable to biased perceptions. In this context, publicly available big data can offer some solutions for strengthening an evaluation's empirical approach and overall findings.

For example, by using publicly available data from satellite or remote sensing sources, estimates of community level poverty (Hersh et al, 2021; Huang et al, 2021; Hu et al, 2022), agricultural productivity (Benos, 2021; Schwalbert et al, 2020), and slum dwellings (Müller et al, 2020) can be predicted. The fundamental logic behind these predictions is the proven relationship of household poverty, agricultural productivity, and slum dwellings with the quality of land cover (housing type, green cover, etc.), which can be captured by satellite imagery data. Impact evaluations of development programmes in climate-fragile contexts of Bangladesh, Nicaragua, and Niger have also used satellite imagery data for estimation of climate shocks (Macours et al, 2022; Pople et al, 2021; Premand & Stoeffler, 2022). In most cases, the relevant remote sensing datasets are publicly available and can be extracted using the Google Earth Engine.[3] These datasets offer an unprecedented opportunity for cost-effective socioeconomic evaluations. Inclusion of the satellite datasets and the relevant ML analysis in evaluations will, however, require advanced analytical expertise, particularly a strong knowledge of geospatial analysis. Evaluators should also be cognizant of critical limitations associated with applying these data and methods. For example, predictive modelling always comes with a certain margin of error. The predictive ML models should be evaluated in a robust manner, and the accuracy levels should be reported and considered while making recommendations (Blummenstock, 2018).

3 Google Earth Engine is a cloud-based service that combines satellite imagery and geospatial datasets with planetary-scale analysis capabilities.

Additionally, social media data, another form of big data, can provide estimates of public opinion of different issues, a measure that can be useful for evaluation of programmes about understanding public sentiment, communications, and behaviour change (Brosius and Cless, 2019; Gorodnichenko et al, 2021; Qazi et al, 2020). Evaluation managers can also review relevant existing research on the use of these big data sources for measuring community and population-level indicators, assess the advantages and challenges of these approaches, and decide whether their application is appropriate for their projects. Table 1 lists some examples of different big data sources, relevant indicators that can be measured using these data types, and the corresponding thematic areas covered by the indicators.

Table 1. Big data sources and corresponding information that can be extracted for evaluation

Data source	Information provided on (examples)	Relevant thematic areas covered (not exhaustive)	Level of ease in accessing data *
Remote sensing/ Satellite imagery	1. Agricultural land cover and quality 2. Urban cover 3. Night lights 4. Precipitation/ Rainfall 5. Human mobility	▸ Food and nutrition: Estimate agricultural productivity as proxy for consumption ▸ Poverty: Estimate poverty at community/ village levels ▸ Slums: Estimate slum boundaries ▸ Weather shocks: Estimate droughts, floods, etc. ▸ Migration: Measure people's movements and estimate migration patterns	✔
Social media and news data	1. Public opinion on any current topic	▸ Cross-cutting: Assess people's key needs during a crisis ▸ Mental health: Assess general mood, opinions and sentiments of populations	⊖
Mobile phone positioning data	1. Human mobility	▸ Migration: Measure people's movements and estimate migration patterns	✖
Financial/Online transactions	1. Financial services	▸ Financial inclusion: Estimate levels of financial inclusion ▸ Gender: Assess women's economic empowerment/financial inclusion	✖
Internet searches/ Google Trends data	1. Proxy for public needs	▸ Cross-cutting: Assess people's key needs during a crisis	✔

** ✔ publicly available data source; ⊖ few sources can be accessed after approvals; ✖ difficult to access direct data source (not publicly available)*

In published literature, we did not find cases where ML was used for the measurement of the programme itself. As is reflected in the operational framework, current research focuses on using these methods for estimating outcomes, outputs, or summarizing overall impact. However, with the growing implementation of large-scale digital interventions such as chatbots for mental health support, agricultural advice through mobile phones, ML analytics can likely be used to measure programme intensity as well.

Considerations for future application

We describe multiple ML methods that can be applied to evaluations in this paper. In certain applications, these can enhance efficiency, by reducing time and resources required to conduct qualitative reviews and assessment of reports and narratives. Alternatively, there are several ways to apply these methods in conjunction with relevant data sources, to directly answer evaluative questions. In this final section, key considerations are summarized for evaluators and managers contemplating or planning to apply ML analytical methods in their evaluations.

Building in-house skills on designing and managing data science efforts in evaluations: The most obvious prerequisite for evaluation teams that are considering the application of ML analytics is to build their human resource capacity with the relevant technical skills. While experts on ML analytics (analysts, data science specialists, etc.) can be hired externally for the duration of a project, internal evaluation teams should have an adequate and updated understanding of ML capabilities. This knowledge is necessary so that they can design effective ML solutions in collaboration with technology experts. For the successful application of ML analytics in evaluations, it is thus crucial to build internal capacities, at least in terms of the a) ability to imagine relevant ML solutions, b) choosing the right external experts for implementation, and c) collaboratively design the ML solutions/analysis to be included in the evaluations.

Starting with easy and low-cost solutions that are replicable and scalable: Broadly, there are two operational routes that evaluation teams can follow with regards to applying ML analytics in their projects: a) build complex ML systems that can house, process, and/or analyse relevant data and information; and b) focus on low-cost pilot projects that provide tailor-made solutions for specific projects. In a context with limited financial and human resources, building complex ML systems might not be feasible at first. UNICEF EO adopted the second approach of investing in small pilot projects, albeit with a perspective of possible replicability and scalability. In addition to being cost-effective, replication of a single approach across different evaluations allows for better learning in terms of 'what works' and 'what does not work.' It is also a resource-effective way of operationalizing NLP approaches, which usually require a lot of effort in terms of text cleaning and processing. Once a large amount of text has been prepared for ML analytics, this cleaned text should be used as much as possible to answer different evaluation questions. Successful application of low-cost pilot projects can also help in acquiring buy-in from senior leadership in investing in more complex ML systems.

Strengthening data management and processing systems for seamless and continued application of ML analytics: Application of any statistical model, including ML approaches, requires the availability of clean and error-free data; the analysis and models can only be as good as the data. In the context of ML, this can be especially important when applying models to analyse unstructured data such as reports and narratives (text data). Investing in strengthening data management and storage systems can facilitate better use of ML analytics. Data storage and engineering platforms such as Databricks can be useful in this regard. Stronger and easy-to-access data systems mean that teams spend less time manually extracting and processing large amounts of data. Data management systems should also allow for data interoperability (datasets connected to each other), so that diverse data can be merged and integrated for meaningful analysis.

Encouraging knowledge sharing and intellectual collaborations across teams and organizations: Even though there is a dearth of published literature on applications of ML in evaluations, most organizations and teams have been experimenting with these methods for a number of years. This is made apparent by the existence of multiple working groups on ML within UNICEF, and across the United Nations system and the broader not-for-profit community. By engaging in these formal or informal groups, evaluation teams can stay abreast of the latest ML analytics, learn from the experiences of others, and develop creative ways to apply ML methods in their own evaluations.

Taking limitations into account and learning from failures: While advancements in ML applications in the field of development practice are very promising, it is equally important to beware of the hype and pay attention to limitations. As with any innovation, failed experiments are often not as widely discussed as the successful ones. In this regard, engaging with knowledge sharing platforms or groups can again be very useful to share and learn from failed attempts. It is also important to be clear about the limitations of specific ML approaches. For example, it is necessary to be transparent about accuracy levels of the predictive models, and how that may affect the final insights. In certain cases, an ML approach that works on a particular data type might fail on a different dataset. From a learning perspective, it is important to think critically and share the different factors that led to the success or failure of any approach.

Recognizing the dynamic nature of this field: The broader field of ML analytics is rapidly expanding, with new ML applications and tools being released every few months. For teams designing ML solutions that are meant to be replicable or used over a long period of time (e.g., an automated system of scoping evaluation documents), they should build agile project systems, to allow constant improvements and changes, when necessary. This can allow for new ML tools or solutions to be easily integrated into old systems. On the other hand, ML initiatives may also need to adjust to evolving organizational needs and requirements, so having technical experts experienced in agile and experimental ML project delivery is thus advisable.

Critical considerations of ethics and safety: This paper has focused on the use of ML methods on publicly available data and internal reports. Although not within the purview of this paper, ethical and data protection concerns are relevant and critical when using generative AI tools such as ChatGPT with sensitive data like interviews and other identifiable data (Hogenhout, 2021). The application of such tools often requires the data to be inputted to the app, potentially leading to violation of data privacy rules. In such cases, users should bear in mind the ethical principles of 1) transparency and accountability; 2) fairness and mitigation, participation, and inclusiveness; 3) privacy and data protection; 4) accuracy and reliability; 5) upholding human rights; and, 6) adopting a human-centred approach. While leveraging AI's capabilities and constraints is important, fostering a dedication to fairness to encourage its responsible use is equally critical. The imperative of responsible and transparent ML usage also aligns with UNICEF's mission, guaranteeing that policy and decision-making processes are guided by impartial and more precise data analyses.

Bibliography

Aiken, E., Bellue, S., Karlan, D., Udry, C., & Blumenstock, J. E. (2022). Machine learning and phone data can improve targeting of humanitarian aid. Nature, 603(7903), 864-870.

Athey, S., Bergstrom, K., Hadad, V., Jamison, J. C., Ozler, B., Parisotto, L., & Sama, J. D. (2021). Shared Decision-Making: Can Improved Counseling Increase Willingness to Pay for Modern Contraceptives?

Athey, S., Tibshirani, J., & Wager, S. (2019). Generalized random forests.

Benos, L., Tagarakis, A. C., Dolias, G., Berruto, R., Kateris, D., & Bochtis, D. (2021). Machine learning in agriculture: A comprehensive updated review. Sensors, 21(11), 3758.

Blumenstock, J. E. (2018a). Estimating economic characteristics with phone data. In AEA papers and proceedings (Vol. 108, pp. 72-76). 2014 Broadway, Suite 305, Nashville, TN 37203: American Economic Association.

Blumenstock, J. (2018b). Don't forget people in the use of big data for development.

Bravo, L., Hagh, A., Joseph, R., Kambe, H., Xiang, Y., & Vaessen, J. (2023). Machine Learning in Evaluative Synthesis: Lessons from Private Sector Evaluation in the World Bank Group.

Brosius, L., & Cless, A. (2019). Utilizing Social Media Analytics to Demonstrate Program Impact.

Chi, G., Fang, H., Chatterjee, S., & Blumenstock, J. E. (2022). Microestimates of wealth for all low-and middle-income countries. Proceedings of the National Academy of Sciences, 119(3), e2113658119.

FAO. (2023). Using artificial intelligence to assess FAO's knowledge base on the technology accelerator. Rome. https://doi.org/10.4060/cc6724en

Franzen, S., Quang, C., Schweizer, L., Budzier, A., Hrstich, P., Reissfelder, S., ... & Raimondo, E. (2022). Advanced Content Analysis.

Gorodnichenko, Y., Pham, T., & Talavera, O. (2021). Social media, sentiment and public opinions: Evidence from# Brexit and# USElection. European Economic Review, 136, 103772.

Hemphill, L., Schöpke-Gonzalez, A., & Panda, A. (2022). Comparative sensitivity of social media data and their acceptable use in research. Scientific Data, 9(1), 643.

Hersh, J., Engstrom, R., & Mann, M. (2021). Open data for algorithms: mapping poverty in Belize using open satellite derived features and machine learning. Information Technology for Development, 27(2), 263-292.

Hogenhout, L. (2021). A framework for ethical AI at the United Nations. *arXiv preprint arXiv:2104.12547*.

Hu, S., Ge, Y., Liu, M., Ren, Z., & Zhang, X. (2022). Village-level poverty identification using machine learning, high-resolution images, and geospatial data. International Journal of Applied Earth Observation and Geoinformation, 107, 102694.

Huang, L. Y., Hsiang, S. M., & Gonzalez-Navarro, M. (2021). Using satellite imagery and deep learning to evaluate the impact of anti-poverty programs (No. w29105). National Bureau of Economic Research.

Jean, N., M. Burke, M. Xie, M. Davis, D. B. Lobell, and S. Ermon. 2016. "Combining Satellite Imagery and Machine Learning to Predict Poverty." Science 353 (6301): 790–94. https://doi.org/10.1126/science.aaf7894.

Kreif, N., & DiazOrdaz, K. (2019). Machine learning in policy evaluation: new tools for causal inference. arXiv preprint arXiv:1903.00402.

Lai, S., Erbach-Schoenberg, E. Z., Pezzulo, C., Ruktanonchai, N. W., Sorichetta, A., Steele, J., ... & Tatem, A. J. (2019). Exploring the use of mobile phone data for national migration statistics. Palgrave communications, 5(1), 1-10.

Linden, A., & Yarnold, P. R. (2016). Using machine learning to assess covariate balance in matching studies. Journal of Evaluation in Clinical Practice, 22(6), 848-854.

Luo, E., Kuffer, M., & Wang, J. (2022). Urban poverty maps-From characterising deprivation using geo-spatial data to capturing deprivation from space. Sustainable Cities and Society, 84, 104033.

Macours, K., Premand, P., & Vakis, R. (2022). Transfers, Diversification and Household Risk Strategies: Can productive safety nets help households manage climatic variability?. The Economic Journal, 132(647), 2438-2470.

McKenzie, D. (2018). How can machine learning and artificial intelligence be used in development interventions and impact evaluations?. World Bank Development Impact blog, 5.

Müller, I., Taubenböck, H., Kuffer, M., & Wurm, M. (2020). Misperceptions of predominant slum locations? Spatial analysis of slum locations in terms of topography based on earth observation data. Remote sensing, 12(15), 2474.

Özler, B. (2022). What's new in the analysis of heterogeneous treatment effects?. World Bank Blogs. Available at: https://blogs.worldbank.org/impactevaluations/whats-new-analysis-heterogeneous-treatment-effects

Pople, A., Hill, R., Dercon, S., & Brunckhorst, B. (2021). Anticipatory cash transfers in climate disaster response.

Premand, P., & Stoeffler, Q. (2022). Cash transfers, climatic shocks and resilience in the Sahel. Journal of Environmental Economics and Management, 116, 102744.

Qazi, A., Qazi, J., Naseer, K., Zeeshan, M., Hardaker, G., Maitama, J. Z., & Haruna, K. (2020). Analyzing situational awareness through public opinion to predict adoption of social distancing amid pandemic COVID-19. Journal of medical virology, 92(7), 849-855.

Schwalbert, R. A., Amado, T., Corassa, G., Pott, L. P., Prasad, P. V., & Ciampitti, I. A. (2020). Satellite-based soybean yield forecast: Integrating machine learning and weather data for improving crop yield prediction in southern Brazil. Agricultural and Forest Meteorology, 284, 107886.

Tiffin, M. A. J. (2019). Machine learning and causality: The impact of financial crises on growth. International Monetary Fund.

Viljanen, M., Meijerink, L., Zwakhals, L., & van de Kassteele, J. (2022). A machine learning approach to small area estimation: predicting the health, housing and well-being of the population of Netherlands. International Journal of Health Geographics, 21(1), 4.

Zivich, P. N., & Breskin, A. (2021). Machine learning for causal inference: on the use of cross-fit estimators. Epidemiology (Cambridge, Mass.), 32(3), 393.

unicef

for every child

For further information, please contact:

UNICEF
Evaluation Office
3 United Nations Plaza
New York, NY 10017
USA

www.unicef.org/evaluation

UNICEF-Evaluation

x.com/UNICEFEval

evalhelp@unicef.org

May 2024