Paper for the UNECE Machine Learning for Official Statistics Workshop 2023

Session 2: Quality Aspects of Machine Learning in Official Statistics

Title: Lessons learned when applying Machine Learning in Official Statistics: Why it helps to be a survey statistician and a data scientist!

Authors: Piet J.H. Daas [1,2] and Marco J.H. Puts [1]

[1] Statistics Netherlands, [2] Eindhoven University of Technology

Abstract:

At Statistics Netherlands the use of Machine Learning (ML) to extract information from texts and images has been studied since 2016. During this period, many so-called Big Data based statistics have been developed that are either in production (online platform economy detection), very close to production (innovative companies, cybercrime, land use identification, skills extraction from job ads, and other Natural Language Processing based applications) or have ended in the experimental phase (quite a lot). As a result of these studies, we learned important lessons on the quality issues that arise when applying ML in an official statistics production environment. The most important ones are: i) begin with a thorough preliminary investigation ii) create an as good as possible (preferably representative) training and test set, iii) examine the effect of various metrics during the model's training phase, iv) prefer 'transparent' ML-algorithms, v) perform extensive manual checks, vi) focus on both the internal and external validity of the model developed, vii) include the statistics production department when results start to look promising, and viii) anticipate answering all kinds of questions raised by traditional statisticians (non-data scientists).

In the presentation and paper, these issues will be discussed in the context of the work we have conducted regarding the development of an ML model focused on the identification of online platform companies by using website texts. Online platforms are defined by the OECD as "a digital service that facilitates interactions between two or more distinct but interdependent sets of users (whether firms or individuals) who interact through the service via the Internet." The model was trained on a set of positive and negative examples provided by experts and was used to identify the subpopulation of all (potential) online platform companies in the Business Register of Statistics Netherlands. The companies identified were subsequently checked by experts after which they received the Dutch Platform Economy survey. The first two questions in the latter survey focused on a) checking if the correct website was found for the particular company and b) checking the findings of the model regarding the correct identification of the platform status of the company. This enabled a proper external validation of the model developed. It also gave us an indication of the number of online platform companies in the Netherlands and revealed a remarkable conflict between some of the Statistics Netherlands experts' opinions and those of the companies themselves.

Introduction

Machine Learning (ML) based approaches can extract patterns from the data they are trained on. In this paper, we focus on so-called supervised learning methods, i.e., methods that have the answer included for the task on which the algorithms (or models) are trained [Murphy, 2022]. As such, ML methods are clear examples of working in a data-driven way; by doing this ML practitioners apply an "inductive way of working" [Adler and Rips, 2008]. The ultimate goal is to find patterns that are applicable to the whole population [UNECE, 2021]. The latter generalization is essential in the context of official statistics. And this is precisely the major concern when applying ML, or any other data-driven approach, within an official statistical context [Puts and Daas, 2021a].

In this paper, the most important questions regarding the use of ML are discussed in the context of the work we perform at Statistics Netherlands. Every question is first discussed in a general way and subsequently described for a specific case study: the detection of online platform businesses via their website text [Daas et al., 2023b]. For all clarity, online platforms are defined by the OECD [2019] as:

> "*a digital service that facilitates interactions between two or more distinct but interdependent sets of users (whether firms or individuals) who interact through the service via the Internet.*"

From this, it is obvious that an online platform is expected to have a website. The study resulted in the selection of a subpopulation of potential online platforms, which - after manual checking and removal of clearly not relevant businesses - all received a questionnaire. This approach has resulted in an official publication on this topic [Klijs, 2022] and has been applied for a number of years [Daas, 2023; Table 1].

The most important lessons learned during the ML-based studies performed at Statistics Netherlands, in our opinion, are:

1. Perform a preliminary investigation
2. Create a good training and test set
3. Examine the effect of various metrics during the model's training phase
4. Prefer 'transparent' ML algorithms
5. Perform extensive manual checks
6. Focus on both the internal and external validity of the model developed
7. Include the statistics production department when results start to look promising
8. Anticipate answering all kinds of questions raised by traditional statisticians (non-data scientists).

These topics are discussed in this paper followed by general recommendations and a suggestion of future work.

1. Perform a preliminary investigation

A preliminary investigation is essential when little is known about the topic studied. This is certainly the case when a new topic is studied and when a data-driven way of working is applied. The best option is to start with a simple exploratory investigation of the data available. Here, it is important to not assume anything (because ML might not even work) and accept the fact that the data used in the

study does - very likely - not perfectly represent the target population. It's just about demonstrating that applying ML might provide interesting insights for the topic investigated.

For example, when one of the authors was contacted regarding the possibility of detecting online platforms with ML, one of the first questions was: can website texts be used to identify online platforms in the Netherlands? The answer was: perhaps, but that needs to be determined in an exploratory study. And that is what we did. The first dataset obtained was composed of 1034 businesses including the URL of their website. These were all businesses that had received a (first version of the) platform economy questionnaire and had responded to it. Of the URLs included, 926 websites (90%) could be scraped. Based on their answers to the questions, 168 of these businesses were online platforms (18%), 197 (21%) were not, and for 561 (61%) this could not be concluded. It was decided to use the positive and negative cases as the training and test set in the exploratory ML study. The sole aim of this study was to find out if the texts on those websites differed for the two classes. The scraped pages were processed in various ways [Daas et al., 2023b] and a whole range of different classification algorithms, all part of the Python scikit-learn package [Pedregosa et al., 2011], were tested under various settings. In the end, a Support Vector Machine based model with an accuracy of 74% was the best result obtained [Daas, 2020]. This result indicated that the idea of using ML to identify online platforms based on website texts seemed worth the effort to be studied in more detail (and nothing more).

2. Create a good training and test set

Supervised learning is one of the most used types of ML. It indicates that the learning is performed based on independent variables (i.e., features) and a dependent variable, the target variable [Murphy, 2022]. For a classification study, a dataset consisting of different features with a known class (either positive or negative) must be available. On this data a model is trained. Having a set of high-quality examples is essential to prevent algorithmic bias [Puts and Daas, 2021b]. Ideally, the dataset used for training (and testing) should represent the target population studied as perfectly as possible. This is, however, a very challenging composition to obtain. In our office, a set of positive examples is usually provided by experts. Based on their expertise, certain cases may be missed or underrepresented, which may introduce a bias (expert bias). The negative cases on the other hand are usually not provided; occasionally a few cases, that may very much resemble the positive cases, are available [Daas et al., 2023b]. For a proper set of negative cases, we prefer to use a random sample from the target population followed by a manual check to remove any positive cases included by chance. When studying websites, the websites linked to businesses in the Business Register of Statistics Netherlands are used as the sampling frame. Depending on the percentage of positive cases in the target population, which is usually unknown at the beginning of the study, it could also be worth the effort to manually check a random sample from the frame to obtain (additional) positive cases. In such a way, a training (and test) set that is - very likely - representative of the target population can be obtained. However, when such an approach is unfeasible, difficult to perform, or when the positive cases occur at a (very) low percentage in the target population, the second best option would be to construct a dataset with 50% positive and 50% negative cases. This, will - at least - enable one to test if there is a differences between the two classes studied. However, one has to realize that 50% percent positive and 50% negative do, almost certainly, not resemble the occurrence of those classes in 'real world' data. One can also imagine that, certainly for a new topic, the first (or second) dataset created will not (yet) be perfect in every sense, but that subsequent

iterations will gradually improve it. In our opinion, training a model on a dataset with a percentage of positive cases similar to those included in 'real world' data is to be preferred. However, especially when that percentage is low - some even suggested when it is below 20% - it may not even be possible to obtain a well-performing model (see below). Various methodologies to deal with class imbalances have been suggested [Kuhn and Johnson, 2013; chap 16]. We also developed a very effective one: a Bayesian adjustment method, see Figure 1, Puts and Daas [2021b], and Puts [2023].

For the online platform detection case, we found that taking a random sample of the target population is not a good approach. Our current best estimate of online platforms in the Netherlands is close to 0.25% of all businesses with a website included in the Dutch Business Register [Daas et al., 2023b; Gubbels, 2023]. It is impossible to obtain a useful classification model on a sample with such a low number of positive cases. This is caused by the fact that a model that identifies all cases as a non-platform will be correct 99.75% of the time. So, obviously we needed to use a higher percentage of positive cases to obtain a properly trained model. But what percentage is best? While looking at that, one of the authors observed that a model trained on a particular percentage of positive items introduced a bias when applied to datasets with different (known) percentages of positive items [Puts and Daas, 2021b]. These findings are shown in Figure 1a.

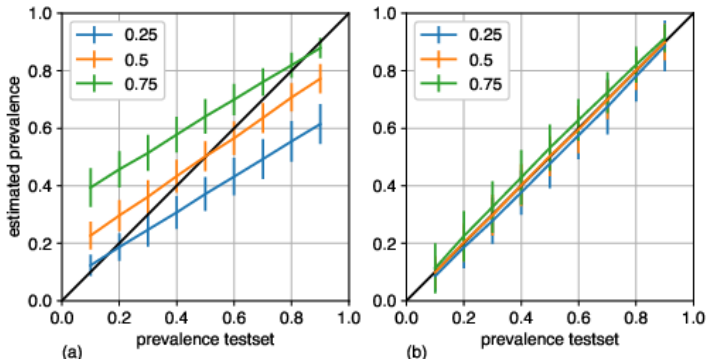

Figure 1. (a) Effects of training a model on a certain ratio of positive items on the classification of datasets with different ratios of positive items [Puts and Daas, 2021b]. (b) Results after applying a Bayesian correction method [Puts, 2023].

Figure 1a reveals that models tend to be biased towards the outcome of the percentage of positives on which they are trained. Along the x-axis, the true fraction of positive items can be observed, whereas the estimated fraction of positive items is shown along the y-axis. The gray line indicates the situation in which the true and estimated values are equal and the bias is thus zero. As Figure 1a illustrates, the estimated value is only correct at one particular point: the fraction on which the algorithm was actually trained. In all other cases, the estimate is biased. Since online platforms are rare, a model developed on a dataset with increased prevalence will likely overestimate the number of online platforms when applied to 'real-world' data. And this is actually what happened. However, this effect was greatly reduced by careful manual checking and validating of the outcome (of the model) by sending the companies selected a questionnaire [Daas et al., 2023b]. Currently, each step in this approach is being studied to improve it from the viewpoint of automating the selection process as much as possible. This work has resulted in the development of a new metric for training

the model which is less affected by a low percentage of positive examples [Gubbels, 2023]. It has also initiated the development of a Bayesian adjustment method [Puts, 2023] to correct for the bias of a specific group of ML classifiers. Classifiers that produce (distorted)probabilities as their outcome, such as Logistic regression and SVM-based models. The effectiveness of this correction method is shown in Figure 1b.

3. Examine the effect of various metrics during the model's training phase

Developing a model requires one to choose a certain evaluation metric by which the success of the performance of the model is determined. However, even when studying a binary classification problem, this is already a difficult choice. An obvious option is Accuracy, e.g., the proportion of correctly classified positive and negative cases of the total number of cases classified. But for a number of reasons, for instance, when the positive and negative cases have an imbalance, this may not be the optimal metric of choice. For such data, Balanced Accuracy, Area Under the Curve (AUC), and F1-score are other - often suggested - choices [Kuhn and Johnson, 2013]. But there are many other metrics available. One only has to look at the Wikipedia page on the Confusion Matrix [Wikipedia, 2023] to get an impression of the enormous range of (potential) metrics to choose from. But which one is the best? The latter highly depends on the ultimate goal of the study. When the goal is identifying all potential positive cases, focus on Recall as a metric. When the goal is getting an optimal separation between negative and positive cases, Balanced Accuracy, AUC, or Mathews Correlation Coefficient (MCC) are possible options. When the goal is to find the best compromise between Recall and Precision, one could use the F1 score, etc. Inside our office, almost every ML practitioner has its preferred metric and it is often unclear why. The best tip we can give is to choose the metric that best suits the needs of the topic studied (and not that of the researcher).

For online platform detection, we originally used Accuracy and Balanced Accuracy as evaluation metrics. However, this resulted in a model that detected many false positive cases when applied to 'real-world' data [Daas et al., 2023b], even when the percentage of positive cases was seriously reduced [Gubbels, 2023]. To determine what would be a better metric, the findings of many of the other most commonly used ones for imbalanced data, such as AUC and MCC, were compared. These all produced fairly and sometimes even very optimistic results. This initiated the search for a new metric, essentially a variant of MCC, which is more pessimistic [Gubbels, 2023].

4. Prefer 'transparent' ML algorithms

For official statistics, there is a need to fully understand the process by which results are obtained. This is a problem for some ML algorithms and it touches on the topic of explainable AI [Gunning et al., 2019]. Making clear how the results are obtained is, for instance, extremely challenging for Deep Learning and other Neural Network-based methods. These algorithms are essentially all black boxes; see Puts and Daas [2021a] for a more in-depth discussion. Here, it's enough to state that, in the context of official statistics, something in the trend of 'Occam's razor' has to be applied: simple algorithms are preferred over complex ones. Certainly when their performances are fairly similar.

For online platform detection, it was important to understand what features (words) were used by the model to identify those kinds of businesses. The best model obtained, an SVM model, enabled us to do that. Here, we found that the words 'platform', 'account', 'login', and 'register' all had high weights [Daas et al., 2023b; Table 1]. These words make sense in the context of online platforms and,

therefore, indicated that the model picked up the intended classification topic. The words with high negative coefficients were indicative of a heterogeneous group of websites which is not an unexpected finding as there is a whole range of non-platform websites. Manual checking the findings also contributes to a better understanding of the model developed (see next point).

5. Perform extensive manual checks

After an ML model has been developed the checking part starts. The most important question here is: is the model measuring what it is supposed to measure? This is all about the concept of interest [Daas, 2023] on which essential insights can be obtained by performing a number of checks. For example, by manual checking (a random sample of) the classified cases. This should preferably be done by multiple experts on the topic (see also the next point) [Daas et al., 2023a]. One can also look at the features and the weights included in the model (if that's possible), look at SHAP values [Lundberg and Lee, 2017], look at the probability distribution on the test set or another unseen dataset (if the model can produce these), create detailed location-based maps of the findings (if that information is available), etc. The discussion paper of Daas and van der Doef [2021] provides examples of this. Another option is to obtain more detailed information from another source for (some of) the cases classified or contact some of the cases studied directly (see below). By performing these checks and not finding any unexpected relations, the researcher will become more and more ensured that the model is detecting the phenomenon it is supposed to detect.

In the online platform case, the model-based probabilities were plotted and (samples of) web pages were manually checked by multiple experts. In addition, the questionnaire send to the companies identified included two important questions at the start. The first was about the website of the business and was used to determine if the website assigned was indeed the correct one for the business receiving the questionnaire. Of the response received, this was for a bit more than 90% correct. The second question specifically asked if the business was an online platform according to the OECD definition (which was additionally provided). To this question, a remarkably high number of businesses answered 'No' (~75%). Subsequent manual inspection of the responses revealed two main reasons for this. The first one was the result of the inclusion of a substantial number of false positive cases (~54%). These of course answered 'No'. The second reason resulted from businesses that answered 'No' even though they were, after additional checking by experts, extremely likely to be true online platforms. Of these online platforms, the false 'false' positives, 21% reported that they were not an online platform. This percentage seems to indicate that a considerable number of online platform businesses answered incorrectly because they either did not understand the OECD definition provided or did not want to answer the remaining questions. The questionnaire-based validation study, however, revealed a positive association between the probabilities produced by the model and the organization's response to the survey question on being an online platform [Daas et al., 2023b]. This means that the model can, indeed, be used to obtain a subpopulation of potential platform organizations from the target population. This subpopulation is a good starting point to study platform organizations in more detail [Daas et al., 2023b].

6. Focus on both the internal and external validity of the model developed

When one develops a model, in a supervised setting, usually a random sample (for instance, 80%) is drawn from the data on which the model is trained [Murphy, 2022]. During training, the algorithm "learns" the difference between the two cases (positives and negatives) in the best possible way. The

remaining part of the original dataset (i.e., 20%) is used as a test set. This test set is used to, independently, determine how well the model can discern between the two classes. It is independent because the test set contains (if all has been done well) entirely new examples - with known outcomes - for the model. We refer to this procedure as the internal validation of the model's performance [Puts and Daas, 2021b]. However, for official statistics, we are predominantly interested in the performance of the model on the target population; i.e., on all of the 'real world' data. In other words, for the online platform model, we want to know how well the model performs on a large dataset that includes many new, unseen, cases. We refer to this as the external validation of the model. This requires data, if possible with known outcomes for some cases, from a substantially larger dataset; ideally a representative part of or even the target population as a whole [Daas and van der Doef, 2021].

For the online platform study, the external validity was determined via the questionnaire and by manually inspecting webpages of positively and negatively classified cases [Daas et al., 2023b]. Verifying the outcome of the test set by manual inspection is also a way to check the internal validity of a model. In one of our studies, manual checking was extensively done and it revealed that the manually determined internal validation percentages were usually 1 á 2% lower than the (test) percentages reported by the model [Daas et al., 2023a].

7. Include the statistics production department when results start to look promising

Apart from all technical issues, there is another important thing to one needs to pay attention to. We found that successful application of ML in statistics production is only possible when the department (or sector) responsible for statistical publications on the topic studied has confidence in the ML-based results. Because ML is fairly new in the world of official statistics, time is required to let others get acquainted with the new methods and new ways of working [De Broe et al., 2021]. Here, it is essential to involve people responsible for the statistical publications on the topic, at the moment that results start to look promising. It's even better when they contact you for help. In that respect, it is certainly beneficial that more and more young employees familiar with ML (and AI) start working at National Statistical Institutes (NSIs). Promoting Data Science and ML inside the office, and thereby creating a community, is another way to stimulate its use, as is creating a list of positive examples [Daas, 2023; Table1].

In that sense, the detection of online platforms started well. Statisticians working at the department responsible for publications on that topic contacted the authors. They required assistance in developing a method that could do that based on website texts. By joining forces, it was demonstrated that a ML-based model could produce a list of potential online platform businesses that were, subsequently, send a questionnaire [Daas et al., 2023b].

8. Anticipate answering all kinds of questions raised by traditional statisticians (non-data scientists)

Two types of questions are usually posed. The first are the somewhat annoying questions posed by many opponents of data science. The others are interesting, fundamental, questions. It's sometimes difficult to distinguish between both as the first type of question could also be the result of a lack of knowledge on the ML (data-driven) way of working [Daas, 2023]. However, on a positive note, there is a definitive need to critically look at the application of ML in statistics. This is needed as the ML way of working differs from the way (many) statisticians work with data at NSIs. ML practitioners

really work data-driven, e.g., they try to extract patterns from data with ML algorithms, while official statisticians tend to analyze data in a theory driven way [De Broe et al., 2020]. Regarding the ML part, it is essential that the 'patterns' found can be generalized. For example, when - in hindsight - a non-representative set of features is obtained from data and included in an ML model, it is highly likely that the model-based findings on new (unseen) data will be incorrect. This is a major risk when working data-driven, but should not be interpreted as fundamentally wrong. It is essential to keep generalizability in mind when applying ML. This is the reason why the external validity of ML models is so important. Another advantage of working data-driven is that one may run into unexpected findings – things not (yet) included in an existing theory; see Daas [2023] for an example.

In that sense, we were lucky with the online platform detection study. Because it was performed after a number of other approaches - also based on website texts – had been tried; i.e., Innovative company [Daas and van der Doef, 2021] and Artificial Intelligence company [Daas and de Wolf, 2021] detection. As a result, the approach used was not new and the discussions were predominantly constructive.

Conclusions and recommendations

From the above, it is clear that the topic of applying ML in an official statistical context benefits from understanding both worlds. Here, it really helps to be a statistician and a data scientist. It should also be clear that, since our 2021 paper [Puts and Daas, 2021b], considerable progress has been made regarding the application of ML for official statistics. However, there is still a need to study this topic further. Suggestions on the areas to investigate in more detail have been made before [Puts and Daas, 2021b] and are still valid. For completion, the topics identified are: i) Methodology concerning the human annotation of data, ii) Sampling the population to obtain representative training sets, iii) Using stratification in the context of Machine Learning, iv) Data structure engineering and selection to increase the transparency of models, v) Reducing spurious correlations, vi) Methodology for studying causation, vii) Correcting the bias caused by the ML model, and vii) Dealing with concept drift (representativity over time). The results described in this paper predominantly relate to obtaining representative training sets (point ii) and bias caused by ML models (point vii). They each reveal considerable progress in these areas. Since many questions are unanswered and researching them takes considerable effort and time, it would be great if those topics could be jointly studied in various international initiatives.

References:

Adler, J.E. and Rips, L.J. (2008). *Reasoning: studies of human inference and its foundations*. Cambridge Univ. Press, Part II: Modes of reasoning.

Daas, P.J.H. (2020). Text classification of online platform websites. Internal report (in Dutch), Statistics Netherlands, the Netherlands.

Daas, P.J.H. (2023). Big Data and Official Statistics. Inaugural lecture, Eindhoven University of Technology, May 26. Link: https://www.tue.nl/en/our-university/calendar-and-events/26-05-2023-inaugural-lecture-profdr-piet-daas

Daas, P.J.H., De Miguel, B., De Miguel, M. (2023a). Identifying Drone Web Sites in Multiple Countries and Languages with a Single Model. *Journal of Data Science*, 1-14.

Daas, P.J.H., de Wolf, N.J. (2021). Identifying different types of companies via their website text. Abstract for the Symposium on Data Science and Statistics (SDSS) 2021, online, USA. Link: https://ww2.amstat.org/meetings/sdss/2021/onlineprogram/AbstractDetails.cfm?AbstractID=309790

Daas, P., Hassink, W., and Klijs, B. (2023b). On the Validity of Using Webpage Texts to Identify the Target Population of a Survey: An Application to Detect Online Platforms. *Submitted for publication.*

Daas, P.J.H., van der Doef, S. (2021). Using Website Texts to detect Innovative Companies. CBDS discussion paper 01-21, Statistics Netherlands, the Netherlands

De Broe, S., Struijs, P., Daas, P., van Delden, A., Burger, J., van den Brakel, J., ten Bosch, O., Zeelenberg, K, and Ypma, W. (2020). Updating the Paradigm of Official Statistics: New Quality Criteria for Integrating New Data and Methods in Official Statistics. *Statistical Journal of the IAOS* 37(1), 343-360

De Broe, S., ten Bosch, O., Daas, P., Buiten, G., Laevens, B., and Kroese, B. (2021). The need for timely official statistics: the pandemic as a driver for innovation. *Statistical Journal of the IAOS* 37(4), 1221-1227

Gubbels, L. (2023). The sample Pearson Correlation Coefficient for classification models and identifying platform economy businesses from web-scraped data. Master thesis Applied and Industrial Mathematics, Eindhoven University of Technology, The Netherlands.

Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., and Yang, G-Z. (2019). XAI-Explainable Artificial Intelligence. *Science Robotics*, 4(37). DOI: 10.1126/scirobotics.aay7120

Klijs, B. (2022). Monitor Online Platforms 2021. Webpage, Statistics Netherlands. Link: https://www.cbs.nl/nl-nl/longread/rapportages/2022/monitor-online-platformen-2021 (in Dutch).

Kuhn, M. and Johnson, K. (2013). *Applied Predictive Modelling*. Springer.

Lundberg, S.M. and Lee, S.I. (2017). A Unified Approach to Interpreting Model Predictions. NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems, pp. 4768-4777.

Murphy, K.P. (2022). *Probabilistic Machine Learning: An Introduction*. MIT Press.

OECD (2019). *Measuring the Digital transformation: A Roadmap for the Future*. Organization for Economic Co-operation and Development, Paris. Link: https://doi.org/10.1787/9789264311992-en.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12: 2825–2830. DOI: 10.5555/1953048.2078195

Puts, M. (2023). BayesCCal: Bayesian Calibration of classifiers, Code on Github. https://github.com/mputs/BayesCCal

Puts, M.J.H. and Daas, P.J.H. (2021a). Machine Learning from the perspective of Official Statistics. *The Survey Statistician* 84, 12-17.

Puts, M.J.H. and Daas, P.J.H. (2021b). Unbiased Estimations Based on Binary Classifiers: A Maximum Likelihood Approach. Symposium on Data Science and Statistics (SDSS) 2021, online. https://arxiv.org/abs/2102.08659

UNECE (2021). *Machine Learning for Official Statistics*. United Nations publication, Geneva. Link: https://unece.org/sites/default/files/2022-09/ECECESSTAT20216.pdf

Wikipedia (2023). Confusion Matrix. Web page, link: https://en.wikipedia.org/wiki/Confusion_matrix