**3GPP TSG-WG SA2 Meeting #157 *S2-2307076***

**Berlin, Germany, May 22 – 26, 2023**

**Source: Huawei, HiSilicon**

**Title: Discussion paper on accuracy of analytics**

**Document for: Discussion**

**Agenda Item: 9.23.2**

**Work Item / Release: eNA\_Ph3 / Rel-18**

*Abstract:* There is an EN about the definition of Analytics/ML Model Accuracy information. This discussion paper is intended to ensure that the definition of Analytics/ML Model Accuracy information sufficiently clear in eNA. To achieve this goal, the main task is to accurately identify the type of analysis ID and evaluate it case by case.

# 1. Introduction

This discussion paper has two parts.

**Part1**

Part1 tries to solve the following EN about the definition of Analytics/ML Model Accuracy information and classify the Analytics ID and introduce various evaluation methods in eNA to improve the evaluation ability of 5GC.

Editor’s Note: It is FFS whether Analytics/ML Model Accuracy information in addition includes recall, precision and balanced accuracy (i.e. (true positive ratio + true negative ratio)/2).

Editor's note: It is FFS to determine whether the deviation refers to a difference between accuracy values or difference of prediction from the actual network data.

**Part2**

Based on the approved CR that the Analysis feedback information contains an indication that the action taken by the consumer will affect the ground truth. However, the influence of action on the ground truth data has not studied yet in TR phase.

Part2 tries to study the potential relationship between action and ground truth and propose that ground truth should only be used when consumer does not take actions in this release.

# 2. Discussion

**Part1**

According to the research of AI/ML, Classification and Regression are two important kind of problems. Both Regression and Classification algorithms are known as Supervised Learning algorithms and are used to predict in Machine learning and work with labelled datasets. The difference of them is that regression finds correlations between dependent and independent variables, and classification is an algorithm that finds functions that help divide the dataset into classes based on various parameters.

Based on the above definition, it is possible to classify the analysis IDs and its prediction in eNA into classification or regression. For example, classification tasks include Abnormal behaviour related network data analytics, QoS Sustainability Analytics, etc. And Prediction of numerical values, such as the number of UEs (average and variance) or PDU sessions and the quality of data transmission (UL/DL data rate, package delay), the predicted CPU load, is a regression problem.

Even if the analytic ID is the same, the prediction type of its tasks may have different types. For example, abnormal behaviour prediction, the output of abnormal behaviour prediction includes the abnormality type, abnormality level, abnormality trend, and number of affected abnormal UEs. The prediction of an exception type and an exception level generally belongs to a classification task, however the number of affected UEs belongs to a regression task.

|  |  |  |
| --- | --- | --- |
| **Task type** | **Class** | **Example** |
| Regression | natural number | > Number of UE Registrations (NOTE 1) Number of predicted UE registrations at the Network Slice instance (average, variance). > Number of predicted PDU Session establishments of the Network Slice instance (average, variance). > NumberAverage represents average and variance information >Indicates the dispersion amount of the reported data volume or transaction dispersion type. This value is only applicable to DISPERSION event. |
| Classification | Enumeration type | > Exception ID the risk detected by NWDAF（*class=number of Exception IDs*）  > qosFlowRetThd RetainabilityThreshold C 0..1 The reporting QoS Flow Retainability Threshold that are met or crossed for 5QI of GBR resource type. *( class = 2)* (NOTE 1) > Exception trend (up/down/unknown/stable)*( class = 4)* > Enumeration DispersionClass > RANKING Data/transaction usage ranked high (i.e.value 1), medium (2) or low (3). This value is only applicable to DISPERSION event. *(class = 3)*  > Exception level: Scalar value indicating the severity of the abnormal behaviour *(class = number of exception level)* |
| Regression or Classification | finite number set | > Value of NF CPU Usage that triggers notification (NOTE 2) Minimum = 0. Maximum = 100. |
| Regression | real number | > Predicted UL/DL data rate, Traffic volume > Indicates average Packet Delay. This value is only applicable to DN\_PERFORMANCE event. |

Figure1. The task type and class of the Analytics/ML model

**Observation1：Analysis IDs and its predictions can be classified into classification tasks and regression tasks based on task types.**

For regression problem, the model can be evaluated using the loss on the specified dataset. Loss is the penalty for a bad prediction. That is, loss is a number indicating how bad the model's prediction was on a single example. If the model's prediction is perfect, the loss is zero; otherwise, the loss is greater. Besides, there is some hyper-parameters in the loss function, for example, the weight of each class or data, the type of distance which loss used such as root mean squared, cross entropy, etc.

For classification tasks, precision and recall are also widely used.

Thus it is not sufficient to assess the accuracy of the analysis ID with only accuracy itself. Because even if the accuracy is the same, it does not mean that the model has the same performance.

In release 18, it is noticed the model metric has defined in Federated Learning procedure[1]:

*The consumer (NWDAF containing AnLF) sends a subscription request to NWDAF containing MTLF to retrieve a ML model, using Nnwdaf\_MLModelProvision service as defined in TS 23.288 [5], including Analytics ID, model metric (e.g., accuracy, precision, recall, etc), pre-determined status (e.g. requested accuracy level and/or total training time) and/or periodically reporting condition (e.g. training round/time).*

In order to make the evaluation methods is align, it is proposed to reuse the Analytics/ML Model metric, which is to represent various evaluation methods for Analytics/ML Model, to support various evaluation methods on both Analytics and model in eNA. The following methods should be included:

* Accuracy: which are composed of the number of correct predictions out of all predictions and the corresponding number of samples.
* Precision (also called positive predictive value) is the fraction of relevant instances among the retrieved instances,
* Recall (also known as sensitivity) is the fraction of relevant instances that were retrieved.
* Loss is the penalty for a bad prediction. indicating how bad the model's prediction was on a single example.
* Classification report is to represent the classification result on the test dataset. For example, the set of classes in test dataset is C, the size of C is N, where Ci is an element in the set. The classification report contains the number of samples for which the model predicts samples of class Ci as class Cj. In other words, it may be understood that the classification report is an N\*N matrix.

**Observation 2: Only “Accuracy” is not sufficient for evaluating Analytics ID and models.**

**Proposal 1: It is proposed to reuse model metric for accuracy information notification.**

For the subscription accuracy information for classification, e.g. Accuracy, Precision, Recall, there could be different thresholds for different methods. Besides, there could be other aspects impacting the thresholds. For example,

1. There is ground truth data including output data and a label, where the label is an exception level of an abnormal behaviour corresponding to each input data in Analytic ID “Abnormal behaviour”. It is assumed that there are three types of exception levels: levels 1, 2, and 3. Different levels indicate different exception levels. Consumers take actions based on the exception level. Now, consider the input data X, which is labelled with an exception of level 1, and for Model 1, its output is level 2, and for Model 2, its output is level 3. In this case, the accuracy information of the two models is the same, but the error degree of the model 1 is lower than that of the model 2. Thus, there is no unified meaning of accuracy considering the above case. Cost matrix was introduced (a matrix with the same number of rows and columns, indicating the costs for correct and incorrect predictions as described in [2]) to solve this issue in academia. Considering different type of Consumer NFs may treat the error degree differently, e.g. PCF should be more sensitive than the AMF as PCF may release the PDU Session but the AMF may just change the RA of the UE in this case.
2. Besides, in abnormal behaviour prediction, the class could be the set of abnormal behaviour which the Analytics ID support. The AMF focuses more on the accuracy of unexpected wakeup and unexpected radio link failures, however the SMF focuses more on the accuracy of unexpected long-live/large rate flows. Therefore, different consumers have different accuracy requirements for different exception ID (class). Thus the consumer may have different thresholds for different classes (or to say different weights for different classes). But there are different types of Analytics and different consumers may have different thresholds per class, it is not sufficient to provide accuracy information per class. It is proposed to provide the weight of each classes to the NWDAF to get a unified feedback of accuracy information no matter how many classes the Analytic ID has. For instance, the number of classes of model is 3, C1 (accuracy = 95%)、C2 (accuracy = 85%) and C3 (accuracy = 90%). If the weights are same, the accuracy of model is (95%+85%+90%)/3 = 90%; If the weights are C1=0.2, C2=0.3, C3=0.5, the Accuracy of model is 95\*0.2+85%\*0.3+90%\*0.5 = 89.5%

**Observation 3: There could be different thresholds for different methods (e.g. Accuracy, Precision, Recall). Besides, error degree and weights for different classes have impact on unified meaning of accuracy information.**

**Proposal 2: It is proposed to have different thresholds for each method and also take the Cost matrix and weight of class of the consumer into account.**

**Part2**

**Observation5: In this release, the ground truth data is only considered when Analytics Feedback Information shows there is no action triggered by the analytics output in the consumer.**

Ground truth is the value of the target variable for a particular item of labelled input data, according to definition of ISO/IEC 22989:2022. In other words, “Ground truth” refers to the correct or “true” output value of a specific analytics ID or model. It could be used to evaluate the accuracy of Analytic ID(s) or Model(s).

At the last meeting, we concluded that the Analysis feedback information contains an indication that the action taken by the consumer will affect the ground truth. As we all know, the ground truth is used to generate the accuracy information of the analytics ID and model. However, SA2 didn’t fully study the impact of actions on ground truth in the TR phase.

Define model as M1, where X indicates a feature set of the dataset, and Y indicates a label set of the dataset. {Xi, Yi} represents a sample data pair. To avoid ambiguity of Y, we further distinguish Y with Y\_pred ,Y\_true and Y\_obse, Y\_pred indicates the output of the model to the input X, Y\_true indicates the actual value corresponding to the output, the ground truth of Y, and the Y\_obse indicates the observed value of Y.

There are two opions.

* Option 1: There is no action taken by consumer, and the Y\_true is Y\_obse. If the Y\_pre is equal to Y\_obse, the prediction is correct.
* Option 2: The consumer may take actions based on the local policy and Y\_pred. The action may have an impact on Y\_obse. As a result, ground truth of Y for input X may no longer equal to observed Y\_obse, because the influence of action on Y\_Obse could be higher or lower than Y\_true. Due to it is no way to get the Y\_true anymore, the truth of influence is unknown.Even if the Y\_true is equal to Y\_obse, It is possible that the prediction is wrong, but the consumer's action based on the wrong prediction corrects the error.

It can be seen from above discussion, it is unsuitable to evaluate the model using ground truth data that affected by the action before figuring out the relationship between action and ground truth data. So it is recommended that ground truth should only be used when consumer does not take actions in the release.

**Proposal 3: Ground truth should only be used when consumer does not take actions in the release.**

# 3. Conclusion and proposal(s)

Proposal 1 and 2 are reflected in S2-2307077.

Proposal 3 is reflected in S2-2307075.

# 4. References

[1] 3GPP TS 23.288: "Architecture enhancements for 5G System (5GS) to support network data analytics services".

[2] Elkan, C. (2001, August). The foundations of cost-sensitive learning. In International joint conference on artificial intelligence (Vol. 17, No. 1, pp. 973-978). Lawrence Erlbaum Associates Ltd.