Derived and Base Measure for Validity

<table>
<tr><th colspan="4">Derived measure or indicator: Validity</th></tr>
<tr><td>#</td><td>Derived measure or indicator
Mval or Validity of Big Data is defined in terms of its accuracy and correctness for the purpose of usage.</td><td colspan="2">Formula
$$Mval\ (MDS) = Credability\ (MDS) * W_{Cred} + Compliance(MDS) * W_{Compli}$$</td></tr>
<tr><td colspan="2">Link with the measurement goal (which goal)

Validity</td><td>Responsible (who analyzes)

Developer

Data Analyst

Data Engineer

Data Scientist</td><td>Stakeholder (who uses)

Senior management

Project manager

Data scientist

Data analyst</td><td>Frequency (when)

The validity of data set can be measured on monthly, quarterly, or yearly basis.</td></tr>
<tr><td colspan="2">Data source (where the measurement data will be extracted from)

Credit Card classification - https://www.kaggle.com/datasets/samuelcortinhas/credit-card-classification-clean-data</td><td>Storage of the result (where data will be stored after the extraction)

The data will be stored in excel file or database.

In our case we will be storing the result in jupyter notebook for reporting purpose.</td><td colspan="2">Data interpretation rules

Successful request is categorized as a request which returns the correct result.

Every query to a database is considered as a request.

Validity = 1 - means that the subject data is accurate and correct for the purpose of usage. This is a desired value for implementation of a successful machine learning model.

Validity = 0 means that data attributes are not correct.

Validity >= 0.90 means that 90% of the data attribute are accurate which can be</td></tr>
</table>

		useful to train our machine learning algorithm. Validity could increase or decrease depending upon the dataset size increasing or decreasing.

Analysis procedure 1. Dataset is loaded using the analyses tool, excel file or jupyter notebook. 2. Compliance and Credibility will be calculated using the formula. 3. Validity of the dataset will be calculated using the formula. 4. The value will be interpreted according to the decision-making rules and appropriate decision will be taken.	**Presentation of the results (sketch illustrating what it looks like):** Validity of the data will be presented as a single numerical value.
Potential decision making depending on the results Validity of the data attributes can give the overview about accuracy and correctness of the data. This is an important measure to get the Machine Learning model trained with the correct data. If the completeness value is more, it will give the confidence to stakeholders to trust the results produced by the machine learning algorithms.	

<table>
<tr><th colspan="5">Derived measure or indicator: Compliance</th></tr>
<tr><td>#</td><td colspan="2">Derived measure or indicator

Degree to which data has attributes that adhere to standards, conventions or regulations in force and similar rules relating to data quality in a specific context of use</td><td colspan="2">Formula

$$Compliance\ (MDS) = \frac{\sum_{\forall\ DS \in MDS} Nrec_{comp}(DS)}{Nds(MDS)}$$</td></tr>
<tr><td colspan="2">Link with the measurement goal (which goal)

Validity</td><td>Responsible (who analyzes)

Developer

Data Analyst

Data Engineer

Data Scientist</td><td>Stakeholder (who uses)

Senior management

Project manager

Data scientist

Data analyst</td><td>Frequency (when)

The compliance of data set can be measured on monthly, quarterly, or yearly basis.</td></tr>
<tr><td colspan="2">Data source (where the measurement data will be extracted from)

Credit Card classification - https://www.kaggle.com/datasets/samuelcortinhas/credit-card-classification-clean-data</td><td>Storage of the result (where data will be stored after the extraction)

The data will be stored in excel file or database.

In our case we will be storing the result in jupyter notebook for reporting purpose.</td><td colspan="2">Data interpretation rules

Successful request is categorized as a request which returns the correct result.

Every query to a database is considered as a request.

Compliance = 1 - means that the subject data adheres to standards, conventions or regulations in force and similar rules relating to data quality for the purpose of usage. This is a desired value for implementation of a successful machine learning model.

Compliance = 0 means that data attributes are not compliant.

Compliance >= 0.90 means that 90% of the data attribute adhere to standards and</td></tr>
</table>

		regulations which can be useful to train our machine learning algorithm. Compliance could increase or decrease depending upon the dataset size increasing or decreasing.

Analysis procedure

1. Dataset is loaded using the analyses tool, excel file or jupyter notebook.
2. rec_comp is counted using COUNT function to get number of compliant records in a dataset
3. DS_comp is calculated by using the formula rec_comp/Nds
4. Compliance of the dataset will be calculated using the formula.
5. The value will be interpreted according to the decision-making rules and appropriate decision will be taken.

Potential decision making depending on the results

Compliance of the data attributes can give the overview about adherence to standards, conventions or regulations in force and similar rules of the data. This is an important measure to get the Machine Learning model trained with the correct data. If the completeness value is more, it will give the confidence to stakeholders to trust the results produced by the machine learning algorithms.

Presentation of the results (sketch illustrating what it looks like):

Compliance of the data will be presented as a single numerical value.

<table>
<tr><th colspan="6">Base measure: Compliant records (Nrec_comp)</th></tr>
<tr><td>#1</td><td colspan="2">Measure (what: entity, attribute)

Measures the number of compliant records in the dataset

Entity: Dataset

Attribute: Number of records</td><td>Scale type

Absolute</td><td colspan="2">Applicability

Total number of compliant records in data sets acts as a fundamental unit of measurement which can be used to calculate other derived measures.</td></tr>
<tr><td colspan="2">Who measures?

Data Analyst

Data Engineer

Data Scientist</td><td>Source of measurement

https://www.kaggle.com/samuelcortinhas/credit-card-classification-clean-data</td><td>Where to store the result

CSV File

Database</td><td>Tool

Excel

Jupyter Notebook

Python libraries for data analysis like pandas, NumPy etc.</td><td>Time (when to measure)

Compliant number of records can be measured each time new data is loaded into the database.</td></tr>
<tr><td colspan="3">Collection procedure (how to collect the data)

The data is loaded into excel sheet or database and the total number of compliant records can be retrieved from querying the database or using inbuilt functions of excel.</td><td colspan="3">Notes or comments:

None</td></tr>
</table>

<table>
<tr><td colspan="6">Base measure: Number of datasets (Nds)</td></tr>
<tr><td>#2</td><td colspan="2">Measure (what: entity, attribute)

Measures the number of records in the dataset

Entity: Dataset

Attribute: Number of records</td><td>Scale type

Absolute</td><td colspan="2">Applicability

Total number of records in data sets acts as a fundamental unit of measurement which can be used to calculate other derived measures. It also gives the idea about the size of the dataset.</td></tr>
<tr><td colspan="2">Who measures?

Data Engineer

Data Analyst

Business Analyst</td><td>Source of measurement

https://www.kaggle.com/samuelcortinhas/credit-card-classification-clean-data</td><td>Where to store the result

CSV File

Database</td><td>Tool

Excel

Jupyter Notebook

Python libraries for data analysis like pandas, NumPy etc.</td><td>Time (when to measure)

Number of records can be measured each time new data is loaded into the database.</td></tr>
<tr><td colspan="3">Collection procedure (how to collect the data)

This number should be given by the responsible person managing databases or excel files.</td><td colspan="3">Notes or comments:

None</td></tr>
</table>

<table>
<tr><th colspan="5">Derived measure or indicator: Credability</th></tr>
<tr><td>#2</td><td colspan="2">Derived measure or indicator

Degree to which data has attributes that are regarded as true and believable by users in a specific context of use</td><td colspan="2">Formula
$$Credability\ (MDS) = \frac{Nds_{cr}(MDS)}{Nds\ (MDS)}$$</td></tr>
<tr><td colspan="2">Link with the measurement goal (which goal)

Validity</td><td>Responsible (who analyzes)

Developer

Data Analyst

Data Engineer

Data Scientist</td><td>Stakeholder (who uses)

Senior management

Project manager

Data scientist

Data analyst</td><td>Frequency (when)

The credibility of data set can be measured on monthly, quarterly, or yearly basis.</td></tr>
<tr><td colspan="2">Data source (where the measurement data will be extracted from)

Credit Card classification - https://www.kaggle.com/datasets/samuelcortinhas/credit-card-classification-clean-data</td><td>Storage of the result (where data will be stored after the extraction)

The data will be stored in excel file or database.

In our case we will be storing the result in jupyter notebook for reporting purpose.</td><td colspan="2">Data interpretation rules

Successful request is categorized as a request which returns the correct result.

Every query to a database is considered as a request.

Credability = 1 - means that the subject data is regarded as true and believable by users for the purpose of usage. This is a desired value for implementation of a successful machine learning model.

Credability = 0 means that data attributes are not truthful.

Credability >= 0.90 means that 90% of the data attribute are true and believable which can be useful to train our machine learning algorithm.</td></tr>
</table>

		Credability could increase or decrease depending upon the dataset size increasing or decreasing.

Analysis procedure

1. Dataset is loaded using the analyses tool, excel file or jupyter notebook.
2. cre_source is counted using COUNT function to get number of credible records in a dataset
3. Credibility of the dataset will be calculated using the formula.
4. The value will be interpreted according to the decision-making rules and appropriate decision will be taken.

Potential decision making depending on the results

Credibility of the data attributes can give the overview about truthfulness and reliability of the data. This is an important measure to get the Machine Learning model trained with the correct data. If the completeness value is more, it will give the confidence to stakeholders to trust the results produced by the machine learning algorithms.

Presentation of the results (sketch illustrating what it looks like):

Credibility of the data will be presented as a single numerical value.

<table>
<tr><td colspan="6">Base measure: Credible Datasets (Nds_cr)</td></tr>
<tr><td>#1</td><td colspan="2">Measure (what: entity, attribute)

Measures the number of credible records in the dataset

Entity: Dataset

Attribute: Number of records</td><td>Scale type

Absolute</td><td colspan="2">Applicability

Total number of credible records in data sets acts as a fundamental unit of measurement which can be used to calculate other derived measures.</td></tr>
<tr><td colspan="2">Who measures?

Data Analyst

Business Analyst</td><td>Source of measurement

https://www.kaggle.com/samuelcortinhas/credit-card-classification-clean-data</td><td>Where to store the result

CSV File

Database</td><td>Tool

Jupyter Notebook

Python libraries for data analysis like pandas, NumPy etc.</td><td>Time (when to measure)

Credible number of records can be measured each time new data is loaded into the database.</td></tr>
<tr><td colspan="3">Collection procedure (how to collect the data)

This number should be given by the responsible person managing databases or excel files.</td><td colspan="3">Notes or comments:

None</td></tr>
</table>

<table>
<tr><td colspan="6">Base measure: Number of datasets (Nds)</td></tr>
<tr><td>#2</td><td colspan="2">Measure (what: entity, attribute)

Measures the number of records in the dataset

Entity: Dataset

Attribute: Number of records</td><td>Scale type

Absolute</td><td colspan="2">Applicability

Total number of records in data sets acts as a fundamental unit of measurement which can be used to calculate other derived measures. It also gives the idea about the size of the dataset.</td></tr>
<tr><td colspan="2">Who measures?

Data Engineer

Data Analyst

Business Analyst</td><td>Source of measurement

https://www.kaggle.com/samuelcortinhas/credit-card-classification-clean-data</td><td>Where to store the result

CSV File

Database</td><td>Tool

Excel

Jupyter Notebook

Python libraries for data analysis like pandas, NumPy etc.</td><td>Time (when to measure)

Number of records can be measured each time new data is loaded into the database.</td></tr>
<tr><td colspan="3">Collection procedure (how to collect the data)

This number should be given by the responsible person managing databases or excel files.</td><td colspan="3">Notes or comments:

None</td></tr>
</table>