### **Derived and Base Measure for Validity**

Deri	Derived measure or indicator: Validity								
#	relived measure or indicator value or Validity of Big Data s defined in terms of its couracy and correctness for he purpose of usage.  Formula $Mval (MDS) = Credability (MDS) * W_{Cred} + Compliance (MDS) * W_{Compli}$								
	with the measurement goal (which	h Responsible (who analyzes)	Stakeholder (who uses)	Frequency (when)					
goal	-)								
Vali	dity	Developer  Data Analyst	Senior management Project manager	The validity of data set can be measured on monthly,					
		Data Engineer	Data scientist	quarterly, or yearly basis.					
		Data Scientist	Data analyst						
data	a source (where the measurement will be extracted from)	Storage of the result (where data will be stored after the extraction)	Data interpretation rule Successful request is carequest which returns the						
http	os://www.kaggle.com/datasets/samu ctinhas/credit-card- ssification-clean-data	e The data will be stored in excel file or database.	Every query to a database is considered as a request.						
		In our case we will be storing the result in jupyter notebook for reporting purpose.	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 to data attribute are accus						

useful to train our machine learning algorithm.

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

#### Analysis procedure

- 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.

### 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.

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

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

Derived measure or indicator: Compliance							
# Derived measure of Degree to which do attributes that a standards, convening regulations in for similar rules reladate quality in a context of use	ata has dhere to tions or rce and ating to	rmula $Compliance (MDS) = \frac{\sum_{\forall DSS}}{}$	<sub>EMDS</sub> Nrec <sub>comp</sub> (DS) Nds(MDS)				
	ent goal (which	Responsible (who analyzes)	Stakeholder (who uses)	Frequency (when)			
<pre>goal) Validity</pre>		Developer  Data Analyst  Data Engineer  Data Scientist	Senior management Project manager Data scientist Data analyst	The compliance of data set can be measured on monthly, quarterly, or yearly basis.			
Data source (where the measurement data will be extracted from)  Credit Card classification - https://www.kaggle.com/datasets/samue lcortinhas/credit-card-classification-clean-data		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.	Successful request is converged to a database as a request.  Compliance = 1 - means data adheres to standar regulations in force and relating to data quality of usage. This is a desimplementation of a successful as a successful and the standard regulations of a successful and the successful and th	ategorized as a he correct result.  se is considered  that the subject ds, conventions or d similar rules y for the purpose ired value for cessful machine  at data attributes  s that 90% of the			

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

- 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
- 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.

#1	Measure (what: entity, attribute)			Scale type	Applicability	
	Measures the number of  Entity: Dataset  Attribute: Number of	of compliant records in the dat records	aset	Absolute	fundamental u	ta sets acts as a nit of measurement used to calculate
Who	measures?	Source of measurement	Wher	e to store the	Tool	Time (when to
Data	Analyst Engineer Scientist	https://www.kaggle.com/samue lcortinhas/credit-card- classification-clean-data		File Babase	Excel  Jupyter Notebook  Python libraries for data analysis like pandas, NumPy etc.	measure)  Compliant number of records can be measured each time new data is loaded into the database.
The tota	l number of compliant ying the database or u	to collect the data)  scel sheet or database and the records can be retrieved from asing inbuilt functions of	None	es or comments	•	

<pre>ures the number of  ty: Dataset ibute: Number of  res?</pre>			Absolute	sets acts as a of measurement	2
res?	1				
	Source of measurement	Where	to store the	Tool	Time (when to
neer yst Analyst	https://www.kaggle.com/samue lcortinhas/credit-card- classification-clean-data		File	Excel  Jupyter Notebook  Python libraries for data analysis like pandas, NumPy etc.	measure)  Number of records can be measured each time new data is loaded into the database.
Collection procedure (how to collect the data)  This number should be given by the responsible person			es or comments	•	
ī	Analyst  n procedure (how er should be give	classification-clean-data Analyst  procedure (how to collect the data)	Classification-clean-data  Datak  Analyst  The procedure (how to collect the data)  Per should be given by the responsible person	Classification-clean-data  Database  Analyst  An procedure (how to collect the data)  Procedure (how to collect the data)	Classification-clean-data  Analyst  Characteristic classification-clean-data  Database  Database  Database  Notebook  Python libraries for data analysis like pandas, NumPy etc.  Procedure (how to collect the data)  Notes or comments:

Derived measure or indicator: Credability							
#2 Derived measure or indicator	Formula	1					
Degree to which data has attributes that are regarded as true and believable by users in a specific context of use	$Credability(MDS) = \frac{N}{2}$	$\frac{Nds_{cr}(MDS)}{Nds (MDS)}$					
Link with the measurement goal (which	th Responsible (who analyzes)	Stakeholder (who uses)	Frequency (when)				
<pre>goal) Validity</pre>	Developer	Senior management	The credibility of data set can				
_	Data Analyst	Project manager	be measured on monthly,				
	Data Engineer	Data scientist	quarterly, or yearly basis.				
	Data Scientist	Data analyst					
Data source (where the measurement data will be extracted from)  Credit Card classification - https://www.kaggle.com/datasets/samulcortinhas/credit-card-classification-clean-data	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.	Successful request is carequest which returns the Every query to a database as a request.  Credability = 1 - means data is regarded as true by users for the purpose is a desired value for a successful machine least tributes are not truth.  Credability = 0 means the attributes are not truth.  Credability >= 0.90 meand data attribute are true which can be useful to the learning algorithm.	ategorized as a ne correct result. se is considered that the subject e and believable e of usage. This implementation of arning model.  nat data nful.  ns that 90% of the and believable				

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

#### Analysis procedure

- 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.

Base	Base measure: Credible Datasets (Nds_cr)						
#1	Measure (what: entity, attribute)  Measures the number of credible records in the dataset  Entity: Dataset  Attribute: Number of records			Scale type	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.		
			set	Absolute			
Who	measures?	Source of measurement	When	re to store the	Tool	Time (when to measure)	
	Analyst	https://www.kaggle.com/samue lcortinhas/credit-card-		File	Jupyter Notebook	Credible number	
Busi	ness Analyst	classification-clean-data	Data	abase	Python libraries for data analysis like pandas, NumPy etc.	of records can be measured each time new data is loaded into the database.	
Collection procedure (how to collect the data)  This number should be given by the responsible person				tes or comments:			
managing databases or excel files.			None				

Base	measure: Number of da	atasets (Nds)				40194579 4020
#2	#2 Measure (what: entity, attribute)  Measures the number of records in the dataset  Entity: Dataset  Attribute: Number of records		Scale type		Applicability	
				Absolute	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.	
Who	measures?	Source of measurement	When	e to store the	Tool	Time (when to
Data Busi	Engineer Analyst ness Analyst	https://www.kaggle.com/samue lcortinhas/credit-card- classification-clean-data		file Babase	Excel  Jupyter Notebook  Python libraries for data analysis like pandas, NumPy etc.	Number of records can be measured each time new data is loaded into the database.
Collection procedure (how to collect the data)  This number should be given by the responsible person managing databases or excel files.			None	es or comments:		