Derived and Base Measure for Veracity

M(Ver): Weighted sum of accuracy, completeness, currentness and availability

Deri	Derived measure or indicator: M(ver)				
# 1	Derived measure or indicator	For	rmula		
	M(Ver): Weighted sum of accuracy, completeness,	М (т	ver) = Accuracy(MDS)*Wacc + Co	mnleteness(MDS)*W +	
	currentness and availability		rrentness(MDS)*W _{curr} + Availab	-	
	_				
		Ead	ch weight is set to ¼ by defa	ult.	
	with the measurement goal (whi	ich	Responsible (who analyzes)	Stakeholder (who uses)	Frequency (when)
goal	.)				
			Developer	Project Manager	Veracity of
Vera	city		_		dataset can be
			Data Analyst	Data Scientist	calculated on
			Data Engineer	Senior Management	monthly, quarterly or
			Data Engineer	Jenier nanagemene	yearly basis.
			Data Scientist		
Data	source (where the measurement		Storage of the result	Data interpretation rule	<u> </u> !s
data	will be extracted from)		(where data will be stored		
			after the extraction)	Veracity can range betwe	n on 0 and 1
Cred	dit Card classification -			higher veracity means be	
	os://www.kaggle.com/datasets/sam	nue	The data will be stored in	trustworthiness of data	
	tinhas/credit-card- sification-clean-data		excel file or database.	veracity means unreliabl	e data.
Стаз	SSIIICACIOII-CIEAII-UACA		In our case we will be	Veracity >= 0.8 means th	nat the data
			storing the result in	quality is good and it of	
			jupyter notebook for reporting purpose.	machine learning model.	
				Veracity < 0.8 means th	
				quality is good and it o	can be used for
				machine learning model.	
			I .	j	

The weights for each of the sub measures defines the importance of each sub measure in the calculation of veracity.

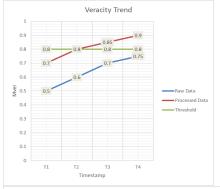
Analysis procedure

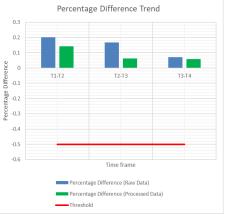
- 1. Retrieve recent accuracy value available at that timestamp from database.
- 2. Retrieve recent completeness value available at that timestamp from database.
- 3. Retrieve recent currentness value available at that timestamp from database.
- 4. Retrieve recent availability value available at that timestamp from database.
- **5.** Use the weighted sum formula to calculate veracity
- Analyze and interpret the results and make decisions

Potential decision making depending on the results

Veracity of data can be crucial for decision making process, for machine learning model to perform well veracity of data should be good enough, poor data from unreliable source can lead to failure of machine learning model. In business big data is used to study customer behaviour and if the veracity of data is low then this could lead to wrong decision making, however good quality data can be beneficial for the growth of the company.

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





Accuracy:

#1	Derived measure or indicator	For	Formula			
	Degree to which data attributes represent the true value in a specified context of use.	Availability (MDS) = $\frac{Hacc}{Hmax}$				
Link	with the measurement goal (whi	ich	Responsible (who analyzes)	Stakeholder (who uses)	Frequency (when)	
goal	L)					
Veracity			Developer	Senior management	The accuracy of data set can be	
			Data Analyst	Project manager	measured on monthly,	
			Data Engineer	Data scientist	quarterly or yearly basis.	
			Data Scientist	Data analyst		
	a source (where the measurement		Storage of the result	Data interpretation rule	es	
data	a will be extracted from)		(where data will be stored after the extraction)			
Credit Card classification - https://www.kaggle.com/datasets/samue		The data will be stored in	Successful request is carrequest which returns the	2		
lcortinhas/credit-card-classification-clean-data		excel file or database.	Every query to a databas as a request.	se is considered		
		In our case we will be storing the result in jupyter notebook for reporting purpose.	Accuracy = 1 - means the attributes always represents is a desired value implementation of a succelearning model.	sent truth value. for		

40163636 40193024 40194579 40201472

Accuracy = 0 means that data attributes do not hold true value .

Accuracy >= 0.90 means that 90% of the data attribute holds true value which can be useful to train our machine learning algorithm.

Accuracy 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.
- Lbd is counted using COUNT function to get number of records
- 3. P_j is calculated to get the total number of duplicate items and their specific count in each dataset using the function like COUNT()
- 4. H_{acc} and H_{max} are calculated using the formula.
- 5. Accuracy of the dataset will be calculated using the formula.
- 6. 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):

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

	40194579 402014
Potential decision making depending on the results	
Accuracy of the data attributes can give the overview about truthfulness of the data. This is an important measure in order to get the Machine Learning model trained with the correct data. If the accuracy value is more it will give the confidence to stakeholders in order to trust the results produced by the machine learning algorithms.	

Base	Measure: p _j					
#2	Measure (what: entity, attribute) Measures the total number of duplicate items and their specific count in each dataset			Scale type	Applicability	
	Entity: Dataset Attribute: Total number of duplicate items			Absolute	duplicate reco the dataset ar count in each	nderstand how many ords are there in and also their dataset. Better of data quality.
Who	measures?	Source of measurement	Wher	re to store the	Tool	Time (when to
l _	_		resu	ılt	Excel	measure)
Deve	loper	Credit Card classification - https://www.kaggle.com/datas	CSV	File	Jupyter	This metric
Data	Analyst	ets/samuelcortinhas/credit- card-classification-clean-	Data	abase	Notebook	could be measured on a
Data	Engineer	data			Python libraries	monthly, quarterly or
Data	Scientist				for data analysis like pandas , numpy etc.	yearly basis to calculate the accuracy trend of the database.

Collection procedure (how to collect the data)	Notes or comments:	
This can be calculated by using the excel built in function or python various data processing libraries.		

Base	Base measure: Lbd(MDS)						
#3	Measure (what: entity, attribute)			Scale type	Applicability	Applicability	
	Measures the total number of records in multiple datasets.			Absolute	sets acts as a	of records in data a fundamental unit t which can be	
	Entity: Dataset			late other derived also gives the			
	Attribute: Number of records				idea about the dataset.	e sizeof the	
Who	measures?	Source of measurement	_	re to store the	Tool	Time (when to	
Deve	eloper		resu	11.0		measure)	
Data	a Analyst	Credit Card classification - https://www.kaggle.com/datasets/samuelcortinhas/credit-	CSV	File	Excel Jupyter	This metric could be	
Data	Engineer	card-classification-clean-	Data	abase	Notebook	measured after creation and	
Data	a Scientist				Python libraries for data analysis like pandas , numpy etc.	after each update of datasets.	

	40194579 402017
Collection procedure (how to collect the data)	Notes or comments:
The data is loaded into excel sheet or database and the total number of records can be retrieved from query the database or using inbuilt functions of excel.	The number of records will be counted for each dataset for all time periods. I.e Length of records will be counted for a dataset as a whole and for each time period separately
	E.g if we have dataset D1,D2 for time T1,T2 then number of records will be lbd(D1) = lbd(D1T1) + lbd(D2T2)

Deri	ved measure or indicator: H _{acc}						
#1	Derived measure or indicator	Formula	rmula				
	Entropy of the given dataset	$H_{acc} = log_2(Lbd) - ((1/Lbd) * \sum_{j=1}^{k}$	$_{\text{c}} = \log_2(\text{Lbd}) - ((1/\text{Lbd}) * \sum_{j=1}^k p(j) * log2(p(j))$				
	with the measurement goal (which	h Responsible (who analyzes)	Stakeholder (who uses)	Frequency (when)			
goal Vera	city	Developer Data Analyst Data Engineer Data Scientist	Developer	The Hacc of data set can be measured on monthly, quarterly or yearly basis.			
Data	source (where the measurement	Storage of the result	Data interpretation rule	es			
data will be extracted from) Credit Card classification - https://www.kaggle.com/datasets/samue		<pre>(where data will be stored after the extraction) e The data will be stored in excel file or database.</pre>	This measure will be use the accuracy measure. If Hacc is greater than Hma will be more and hence t has truth value.	the value of ax, accuracy value			

40163636 40193024 40194579 40201472

lcortinhas/credit-card-	In our case we will be	
classification-clean-data	storing the result in	
	jupyter notebook for	
	reporting purpose.	

Analysis procedure

- 1. Dataset is loaded using the analyses tool, excel file or jupyter notebook.
- Lbd is counted using COUNT function to get number of records
- 3. P_j is calculated to get the total number of duplicate items and their specific count in each dataset using the function like COUNT()
- 4. $\ensuremath{\text{H}_{\text{acc}}}$ is calculated using the formula mentioned above.

Potential decision making depending on the results

Higher accuracy of data means that decisions taken will be accurate and machine learning model will perform well.

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

 \mathbf{H}_{acc} of the data will be presented as a single numerical value.

Deri	Derived measure or indicator: H _{max}						
#3	Derived measure or indicator 1 Maximum Entropy of the given dataset	Formula $H_{max}(MDS) = \log_2(Lbd)$					
	with the measurement goal (which	h Responsible (who analyzes)	Stakeholder (who uses)	Frequency (when)			
goal Vera	city	Developer Data Analyst Data Engineer Data Scientist	Developer	The Hmax of data set can be measured on monthly, quarterly or yearly basis.			
D - 1 -		Standard Standard	Bata intermediation will	_			
	source (where the measurement will be extracted from)	Storage of the result (where data will be stored	Data interpretation rules				
Cred http lcor	it Card classification - s://www.kaggle.com/datasets/samu tinhas/credit-card- sification-clean-data	after the extraction)	This measure will be use the accuracy measure. If Hmax is less than Haccw, will be more and hence thas truth value.	the value of accuracy value			

40163636 40193024 40194579 40201472

Analysis procedure

- 1. Dataset is loaded using the analyses tool, excel file or jupyter notebook.
- 2. Lbd is counted using COUNT function to get number of records
- 3. H_{max} is calculated using the formula mentioned above.

Potential decision making depending on the results

This value is being used to calculate accuracy.

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

 $\ensuremath{H_{\text{max}}}$ of the data will be presented as a single numerical value.

Completeness:

Derived measure or indicator: Completeness

2 Derived measure or indicator

Completeness: Degree to which subject data associated with an entity has values for all expected attributes and related entity instances in a specific context of use.

Formula

$$Com_m (MDS) = \frac{[rec_no_null (MDS)]}{Lbd(MDS)}$$

Link with the measurement goal (which	Responsible (who analyzes)	Stakeholder (who uses)	Frequency (when)
goal)			
		Project Manager	Completeness of
	Developer	Data Scientist	the data can be
			calculated at
	Data Analyst		the start of the
Veracity			project,
	Data Engineer		periodically at
			certain time
	Data Scientist		intervals or it could be
			calculated eack
			time a new data
			is loaded into
			the system.
Data source (where the measurement	Storage of the result	Data interpretation rule	<u> </u>
data will be extracted from)	(where data will be stored	-	
	after the extraction)	Successful request is ca	tegorized as a
Credit Card classification -		request which returns th	e correct result.
https://www.kaggle.com/datasets/samue			
lcortinhas/credit-card-			
icoleimas/cicale cala	The data will be stored in	Every query to a databas	se is considered
classification-clean-data	The data will be stored in excel file or database.	Every query to a databas as a request.	se is considered
			se is considered
	excel file or database.	as a request. Completeness = 1 - means data associated with an	that the subject entity has values
	excel file or database. In our case we will be	as a request. Completeness = 1 - means data associated with an for all expected attributes.	that the subject entity has values ites and related
	excel file or database. In our case we will be storing the result in	as a request. Completeness = 1 - means data associated with an for all expected attribuentity instances. This is	that the subject entity has values ites and related s a desired value
	excel file or database. In our case we will be storing the result in jupyter notebook for	as a request. Completeness = 1 - means data associated with an for all expected attribuentity instances. This if or implementation of a	that the subject entity has values ites and related s a desired value
	excel file or database. In our case we will be storing the result in jupyter notebook for	as a request. Completeness = 1 - means data associated with an for all expected attribuentity instances. This is	that the subject entity has values ites and related s a desired value

40163636 40193024 40194579 40201472

Completeness = 0 means that data attributes hold null value .

Completeness >= 0.90 means that 90% of the data attribute holds non null value which can be useful to train our machine learning algorithm.

Completeness 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.
- Lbd is counted using COUNT function to get number of records
- 3. Rec_no_null is calculated to get the total number of no null items and their specific count in each dataset using the function like COUNT()
- 4. Completeness 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.

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

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

	40194579 402014
Potential decision making depending on the results	
Completeness of the data attributes can give the	
overview about absoluteness of the data. This is an	
important measure in order to get the Machine	
Learning model trained with the correct data. If	
the completeness value is more it will give the	
confidence to stakeholders in order to trust the	
results produced by the machine learning	
algorithms.	

Base	Base measure: Rec_no_null (MDS)						
#1	Measure (what: entity	, attribute)		Scale type	Applicability		
	Measures the total number of records with no null values.			Absolute	Total number of no null record in data sets acts as a fundamental unit of measuremen		
	Entity: Dataset				which can be u	used to calculate measures.	
	Attribute: Number of records						
Who	measures?	Source of measurement		e to store the	Tool	Time (when to	
Deve	loper	Credit Card classification - https://www.kaggle.com/datas	resu	ılt	Excel	measure) Length of the	
Data	Analyst	ets/samuelcortinhas/credit- card-classification-clean-	CSV	File	Jupyter Notebook	data set can be measured each	
Data	Engineer	data	Data	abase	Python	time new data is loaded into the	
Data	Scientist				libraries for data analysis	database.	

				40194579 4020 <u>1</u> 4
			like pandas	
			, numpy etc.	
Collection procedure (how to	collect the data)	Notes or comments:		
The data is loaded into excel	sheet or database and the	None		
total number of no null record				
query the database or using i	ibulit functions of excer.			

Base	Base measure: Lbd(MDS)						
#2 Measure (what: entity,		, attribute)		Scale type	Applicability		
	Measures the total nu datasets. Entity: Dataset Attribute: Number of	mber of records in multiple records		Absolute	sets acts as a of measurement used to calcul	of records in data a fundamental unit twhich can be late other derived also gives the e sizeof the	
Who	measures?	Source of measurement	When	re to store the	Tool	Time (when to	
Deve	eloper		resu	ılt		measure)	
Data	a Analyst	Credit Card classification - https://www.kaggle.com/datas	CSV	File	Excel	Length of the	
Data	n Engineer	ets/samuelcortinhas/credit-	Data	abase	Jupyter Notebook	data set can be measured each time new data is	

40163636 40193024 40194579 40201472

Data Scientist	card-classification-clean- data		Python libraries for data analysis like pandas , numpy etc.	loaded into the database.
Collection procedure (how to collect the data)		Notes or comments:		
The data is loaded into excel sheet or database and the total number of records can be retrieved from query the database or using inbuilt functions of excel.		The number of records will be counted for each dataset for all time periods. I.e Length of records will be counted for a dataset as a whole and for each time period separately E.g if we have dataset D1,D2 for time T1,T2 then number of records will be lbd(D1) = lbd(D1T1) + lbd(D2T2)		

Currentness:

Deri	Derived measure or indicator: Currentness							
# 4	Derived measure or indicator	Formula						
	Currentness: Degree to which data has attributes that are of the right age in a specific context of use.	$Currentness (MDS) = \frac{[rec_acc_age(MDS)]}{Lbd(MDS)}$						
Link	with the measurement goal (whi	ch Responsible (who analyzes)	Stakeholder (who uses)	Frequency (when)				
goal)							
			Project Manager	Currentness of				
		Developer	Data Scientist	the data can be				
				calculated at				
		Data Analyst		the start of the				
Vera	city			project, periodically at				

40163636 40193024 40194579 4020<u>14</u>72

40194579 40201472					
	Data Engineer			certain time	
				intervals or it	
	Data Scientis	t		could be	
				calculated eack	
				time a new data	
				is loaded into	
				the system.	
Data source (where the measurement	Storage of the	e result	Data interpretation rule	_	
data will be extracted from)	_	ill be stored			
,	after the ext		For counting total number	er of records.	
Credit Card classification -			every record should be o		
https://www.kaggle.com/datasets/samue			counting without any fil		
lcortinhas/credit-card-	The data will	ho stored in	l councing without any iii	acces on data.	
classification-clean-data	excel file or		Data older than 10 years	will be	
orabbilitation ordan aata	excel life of	uatabase.	considered old data.	WIII SC	
	T		constacted of a data.		
	In our case w		Currentness of data will	he measured	
	storing the r		based on threshold value		
	jupyter noteb		currentness of the datas		
	reporting pur	pose.	above a certain value.	set should be	
			above a certain value.		
			Currentness(Dataset) >=	0.7 - rologant	
			for use in machine learn		
			Tor use in machine rearr	iiig modei	
			Currentness (Dataset) bet	ween 0.5 and 0.7	
			- relevant for use in ma		
			model with some caution.	_	
			Currentness(Dataset) <=	0.5 - can be used	
			for training machine lea		
			checking the relevancy of	3	
			one one rerevancy		
Analysis procedure	1	Presentation	 of the results (sketch il	lustrating what	
imarioro brocedure		it looks like		Tablia wilat	
		10 100mb like	, ·		
1 Detect to located and a significant		Currentness o	f the data will be presen	ted as a single	
1. Dataset is loaded using the anal	yses tool,	numerical val	_	ab a bringre	
excel file or jupyter notebook.		maniciicai vai	u		
2. Total number of records are calc					
inbuilt COUNT() function or its	equivalent				

40163636 40193024 40194579 40201472

- 3. Number of records within acceptable range will be calculated by applying filter over timestamp of data record.
- 4. Currentness of the dataset will be calculated using the formula.
- 5. Currentness of each dataset will be added to get the total currentness of MDS at various stages of data processing.
- 6. The value will be interpreted according to the decision making rules and appropriate decisions will be taken.

Potential decision making depending on the results

If the currentness of the data is within the acceptable range then the data can be used to train machine learning models to identify recent trends in data. If the data is too old then the decision derived from the data would not be relevant to the current scenario. Data needs to be updated if the dataset currentness value is too low.

uacasecs.	Base	Base measure: Lbd(MDS)					
datasets. Total number of records in datasets.	#1	Measure (what: entity, attribute)	Scale type	Applicability			
Entity: Dataset of measurement which can be		datasets. Entity: Dataset	Absolute	used to calculate other derived measures. It also gives the idea about the sizeof the			

40163636 40193024 40194579 40<u>201</u>472

Who measures?	Source of measurement	Where to store the	Tool	40194579 40201 Time (when to		
who measures:	Source of measurement	result	1001	measure)		
Developer	Credit Card classification -	lesuic	Excel	measure)		
Developer	https://www.kaggle.com/datas	CSV File	DACCI			
Data Analyst	ets/samuelcortinhas/credit-		Jupyter	Length of the		
Data Analyst	card-classification-clean-	Database	Notebook	data set can be		
Data Engineer	data		Nocebook	measured each		
Data Engineer			Python	time new data is		
Data Scientist			libraries	loaded into the		
Data Berenerse			for data	database.		
			analysis			
			like pandas			
			, numpy etc.			
Collection procedure (how	to collect the data)	Notes or comments:				
•						
The data is loaded into ex	cel sheet or database and the	The number of records will be counted for each				
total number of records ca	an be retrieved from query the	dataset for all time periods. I.e Length of records				
database or using inbuilt	<u> </u>	will be counted for a dataset as a whole and for				
-	each time period separately					
		E.g if we have dataset D1,D2 for time T1,T2 then				
		number of records w	will be lbd(D1)	= lbd(D1T1) +		
		lbd(D2T2)				

Base	Base measure: Rec_acc_age (MDS)							
#2	Measure (what: entity, attribute)	Scale type	Applicability					
		Absolute	The number of records in acceptable range brackets is a relevant metric for calculating					

40163636 40193024 401<u>94579 40201</u>472

						40194579 4020
	Provides the total	number of records with ages that			currentness of data and it also	
	fall within the acc	ceptable range based on the upper	and		tells helps us analyze the age	
	lower quartiles of	the Box and Whisker.				d its relevancy of
	Entity: Dataset				future use.	
	Enercy. Dataset					
	Attribute: number of records within acceptable age					
	range					
Who	measures?	Source of measurement	Whe	re to store the	Tool	Time (when to
			res	ılt		measure)
Deve	loper	Credit Card classification -				
	-	https://www.kaggle.com/datas			Excel	
Data	Analyst	ets/samuelcortinhas/credit-	CSV	File		Total number of
	-	card-classification-clean-			Jupyter	records with
Data	Engineer	data	Data	abase	Notebook	acceptable age
2404			200		1.000001	range should be
Data	Scientist				Python	measured before
					libraries	calculating
					for data	currentness of
					analysis	data. This could
					like pandas	happen before
					, numpy etc.	the start of the
					, namp ₁ coc.	project,
						periodically at
						certain time
						intervals to
						keep the track
						of currentness
						of data or
						whenever a new
						data is loaded
						into system.
						Inco system.
Coll	ection procedure (ho	ow to collect the data)	No	tes or comments:	<u> </u>	1

40163636 40193024 40194579 40201472

The data can be collected by applying control limits	The number of records within acceptable ranges will
for the acceptable ranges over time attribute and	be calculated for each dataset D(i) and not for
filtering the number of records within acceptable	each time stamp.
ranges.	

Availability:

Deri	ved measure or indicator: Avail	ability		
#5	Derived measure or indicator	Formula		
	Currentness of the data will be presented as a single numerical value.	Availability (MDS) = $\frac{[n_succ_req(MDS)]}{n_req(MDS)}$		
Link goal	with the measurement goal (whi	.ch Responsible (who analyzes)	Stakeholder (who uses)	Frequency (when)
Vera	acity	Developer	Senior management	The availability of data set can be measured on

			<u>40194579 402014</u> 7	
	Data Analyst	Project manager	monthly,	
			quarterly or	
	Data Engineer	Data scientist	yearly basis.	
	Data Scientist	Data analyst		
		_		
Data source (where the measurement	Storage of the result	Data interpretation rule	es	
data will be extracted from)	(where data will be stored			
	after the extraction)			
		Successful request is ca	ategorised as	
Credit Card classification -		request which returns co	_	
https://www.kaggle.com/datasets/samue	The data will be stored in	_		
lcortinhas/credit-card-	excel file or database.	Every query to a databas	se is considered	
classification-clean-data		as request.		
	In our case we will be	_		
	storing the result in	Availability = 1 - means	s that the	
	jupyter notebook for	database is available at all times, for		
	reporting purpose.	every request a successf		
	31 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	been returned. This is a desired value		
		for implementation of su		
		learning model.		
		,		
		Availability = 0 means t	that database does	
		not return result for any request.		
		Availability >= 0.99 means that 99% of the request were successful. This is a		
		acceptable value for tra		
		learning and the model w	_	
		time.		
		Availability >=90 means	that 90% of the	
		request were successful		
		increase the training ti		
		learning model significa		
		might not be available f		
		cases. More number of re		
		required to fetch data.	-	
	I	I .	ļ	

40163636 40193024 40194579 40201472

Availability could increase or decrease depending upon the number of successful requests to database. If number of successful requests fall then the availability is expected to go down.

Analysis procedure

- Dataset is loaded using the analyses tool, excel file or jupyter notebook.
- 2. Total number of requests and successful requests are retrieved from the query log or issue log.
- 3. Availability 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

Availability of the dataset can give an overview of the resiliency of the system infrastructure, low availability could lead to decrease in confidence of stakeholders in the system leading to abandoning of the system whereas high availability could increase the confidence of stakeholders which is preferred by stakeholders for training machine learning model.

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

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

Base measure: N succ req (MDS)						
#1	Measure (what: entity	y, attribute)		Scale type	Applicability	
	Measures the number of successful request from an Asserver, database etc. Entity: Dataset Attribute: number of successful requests		ΡΙ	Absolute	The number of successful request gives us the metric to calculate availability and gives us the intuition about the likelihood of success of a request to an API or database.	
Who	measures?	Source of measurement	When	re to store the	Tool	Time (when to measure)
Data Data	eloper Analyst Engineer Scientist	Credit Card classification - https://www.kaggle.com/datas ets/samuelcortinhas/credit-card-classification-clean-data		File	Excel Jupyter Notebook Python libraries for data analysis like pandas , numpy etc.	This metric could be measured on a monthly, quarterly or yearly basis to calculate the availability trend of the database.
Collection procedure (how to collect the data) Generally API request or queries are logged for future/audit references therefore count the number of requests for which the correct responses have been returned from the database or dataset.		Notes or comments: In case of static dataset count the number of successful queries on the database.				

#2	Measure (what: entity, attribute)			Scale type	Applicability	
	Measures the total number of requests to a database within a given timeframe. Entity: Dataset Attribute: Total number of requests to dataset.			Absolute The number of request to a database could be considere the fundamental unit of database which gives us the idea about the frequency of usage an importance of the dataset. More number of requests means that the database is usage is high a it is important.		d be considered as al unit of a gives us the e frequency of thance of the number of s that the sage is high and
Who	measures?	Source of measurement	When	re to store the	Tool	Time (when to measure)
Data Data	eloper Analyst Engineer Scientist	Credit Card classification - https://www.kaggle.com/datas ets/samuelcortinhas/credit-card-classification-clean-data		File	Excel Jupyter Notebook Python libraries for data analysis like pandas , numpy etc.	This metric could be measured on a monthly, quarterly or yearly basis to calculate the availability trend of the database.
Collection procedure (how to collect the data) Generally, API request or queries are logged for future/audit references therefore count the number of requests/queries performed on database or dataset.		Notes or comments: In case of static dataset count the number of queries on the database.				