**Derived and Base Measure for Validity**

| **Derived measure or indicator:Validity** | | | |  | | |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **#** | Derived measure or indicator | Formula | | | | | | | |
| Mval or Validity of Big Data is defined in terms of its accuracy and correctness for the purpose of usage. |
| Link with the measurement goal (which goal)  Validity | | | Responsible (who analyzes)  Developer  Data Analyst  Data Engineer  Data Scientist | | | Stakeholder (who uses)  Senior management  Project manager  Data scientist  Data analyst | | Frequency (when)  The validity 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/samuelcortinhas/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. | | | 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 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 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. | | | | |

| **Derived measure or indicator:Compliance** | | | |  | | |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **#** | Derived measure or indicator | Formula | | | | | | | |
| 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 |
| Link with the measurement goal (which goal) | | | Responsible (who analyzes) | | | Stakeholder (who uses)  Senior management  Project manager  Data scientist  Data analyst | | Frequency (when)  The compliance of data set can be measured on monthly, quarterly or yearly basis. | |
| Validity(evaluate accuracy and correctness of data from different data sources/time intervals)  Data source (where the measurement data will be extracted from) | | | Developer  Data Analyst  Data Engineer  Data Scientist  Storage of the result (where data will be stored after the extraction) | | | 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 regulations which can be useful to train our machine learning algorithm.  Compliance could increase or decrease depending upon the dataset size increasing or decreasing. | | | |
| https://www.kaggle.com/samuelcortinhas/credit-card-classification-clean-data | | | 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. | | |  | | | |
| 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. | | | | | Presentation of the results (sketch illustrating what it looks like):  Compliance of the data will be presented as a single numerical value. | | | | |
|  | | | | |
| 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 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 measure:Compliant records(Nrec\_comp)** | | |  | | |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| #1 | Measure (what: entity, attribute)  Measures the number of compliant records in the dataset | | | | Scale type  Absolute | 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. | | |
| Entity: Dataset  Attribute: Number of records | | | |  |
| Who measures?  Data Analyst  Data Engineer  Data Scientist | | Source of measurement  https://www.kaggle.com/samuelcortinhas/credit-card-classification-clean-data | | Where to store the result  CSV File  Database | | Tool  Excel  Jupyter Notebook  Python libraries for data analysis like pandas , numpy etc. | Time (when to measure)  Compliant number of records can be measured each time new data is loaded into the database. | |
|  | |  | |  | |  |  | |
| 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. | | | | Notes or comments:  None | | | | |
|  | | | |

| **Base measure:Number of datasets(Nds)** | | |  | | |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| #2 | Measure (what: entity, attribute)  Measures the number of records in the dataset | | | | Scale type  Absolute | 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 sizeof the dataset. | | |
| Entity: Dataset  Attribute: Number of records | | | |  |
| Who measures?  Data Analyst  Data Engineer  Data Scientist | | Source of measurement  https://www.kaggle.com/samuelcortinhas/credit-card-classification-clean-data | | Where to store the result  CSV File  Database | | Tool  Excel  Jupyter Notebook  Python libraries for data analysis like pandas , numpy etc. | Time (when to 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 managing databases or excel files.. | | | | Notes or comments:  None | | | | |
|  | | | |

| **Derived measure or indicator:Credability** | | | |  | | |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **#** | Derived measure or indicator | Formula | | | | | | | |
| Degree to which data has attributes that are regarded as true and believable by users in a specific context of use |
| Link with the measurement goal (which goal) | | | Responsible (who analyzes)  Developer  Data Analyst  Data Engineer  Data Scientist | | | Stakeholder (who uses)  Senior management  Project manager  Data scientist  Data analyst | | Frequency (when)  The credibility 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/samuelcortinhas/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. | | | 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.  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. | | | | | Presentation of the results (sketch illustrating what it looks like):  Credibility of the data will be presented as a single numerical value. | | | | |
|  | | | | |
| 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 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 measure:Credible Datasets(Nds\_cr)** | | |  | | |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **#1** | Measure (what: entity, attribute)  Measures the number of credible records in the dataset | | | | Scale type  Absolute | 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. | | |
| Entity: Dataset  Attribute: Number of records | | | |  |
| Who measures?  Data Analyst  Business Analyst | | Source of measurement  https://www.kaggle.com/samuelcortinhas/credit-card-classification-clean-data | | Where to store the result  CSV File  Database | | Tool  Jupyter Notebook  Python libraries for data analysis like pandas , numpy etc. | Time (when to measure)  Credible 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. | | | | Notes or comments:  None | | | | |
|  | | | |

| **Base measure:Number of datasets(Nds)** | | |  | | |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| #2 | Measure (what: entity, attribute)  Measures the number of records in the dataset | | | | Scale type  Absolute | 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 sizeof the dataset. | | |
| Entity: Dataset  Attribute: Number of records | | | |  |
| Who measures?  Data Engineer  Data Analyst  Business Analyst | | Source of measurement  https://www.kaggle.com/samuelcortinhas/credit-card-classification-clean-data | | Where to store the result  CSV File  Database | | Tool  Excel  Jupyter Notebook  Python libraries for data analysis like pandas , numpy etc. | Time (when to 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 managing databases or excel files. | | | | Notes or comments:  None | | | | |
|  | | | |