**Derived and Base Measure for Vincularity**

| **Derived measure or indicator: M(vin)** | | | |  | | |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **#** | Derived measure or indicator  M(vin): It represents the change in a quantity with respect to another quantity. Line graph of vincularity over time will provide the trend of vincularity. | Formula | | | | | | | |
|  |
| Link with the measurement goal (which goal)  Vincularity | | | Responsible (who analyzes)  Developer  Data Analyst  Data Engineer  Data Scientist | | | Stakeholder (who uses)  Project Manager  Data Scientist  Senior Management | | Frequency (when)  Vincularity can be calculated 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 scenario, it will be storing the result in jupyter notebook for reporting purposes. | | | Data interpretation rules  Vincularity can be calculated for different time periods. It’s value can be in the range [0, 1]  Vincularity(DS)Ti >= 0.6 can be inferred that 60% percent of the data should be traceable for the machine learning algorithm to give relevant results.  **Vincularity = 1.0 - data is completely traceable.**  Higher the Mvin, higher the connectivity and linkage of data and vice-versa | | | |
|  | | |  | | |  | | | |
| Analysis procedure   1. Calculate the base measure LDST 2. Calculate the base measure NDS 3. Calculate the base measure rec\_trace 4. Calculate the derived measure traceability 5. Calculate the average of traceability of all the datasets 6. Analyze and interpret the results and make decisions | | | | | Presentation of the results (sketch illustrating what it looks like): | | | | |
|  | | | | |
| Potential decision making depending on the results  Trend analysis will show us whether the vincularity is increasing or decreasing for both processed and raw extract. Increasing vincularity is a good sign and decreasing vincularity means data elements are not tracebale to source and could degrade the performance of machine learning model. | | | | |

| **Base measure: NDS** | | |  | | |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **#** | Measure (what: entity, attribute)  Measures the total number of datasets in Big Data  Entity: Dataset  Attribute: Number of total unique identifiers UIDDST of datasets in multiple datasets | | | | Scale type  Absolute | **Applicability**  **It helps in assessing the variety of datasets in terms of multiple datasets (MDS)** | | |
|  | | | |  |
| Who measures?  Data Analyst  Data Engineer  Data Scientist | | Source of measurement  Credit Card classification - https://www.kaggle.com/datasets/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. |  | |
|  | |  | |  | |  |  | |
| Collection procedure (how to collect the data) | | | | Notes or comments:  The number of multiple datasets will be counted for the entire data at a given time period.  E.g if we have dataset D1,D2 for time T1,T2 then number of data sets will be NDS(MDS) = 2 for T1 and T2 | | | | |
| This number should be given by the responsible person managing databases or excel files. | | | |

| **Derived measure or indicator: Traceability** | | | |  | | |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **#** | Derived measure or indicator  Traceability: It provides the degree to which data has attributes that provide an audit trail of access to the data and of any changes made to the data in a specific context of use. It is useful in calculating derived measure Mvin. | Formula | | | | | | | |
|  |
| Link with the measurement goal (which goal)  Vincularity | | | Responsible (who analyzes)  Developer  Data Analyst  Data Engineer  Data Scientist | | | Stakeholder (who uses)  Project Manager  Data Scientist  Senior Management | | Frequency (when)  Traceability can be calculated 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 scenario, it will be storing the result in jupyter notebook for reporting purposes. | | | Data interpretation rules  Traceability can be calculated for different datasets at a given time period. It’s value can be in the range [0, 1]  The average of traceabilities among all the datasets provide vincularity for the big data.  Higher the Traceability means higher the degree to which data has attributes that provide an audit trail of access to the data and of any changes made to the data in a specific context of use and vice-versa | | | |
|  | | |  | | |  | | | |
| Analysis procedure   1. Calculate the base measure Rec\_trace for a dataset 2. Calculate the base measure LDSTfor the dataset 3. Traceability can be calculated by dividing Rec\_trace of a DS by its LDST 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):  Traceability of the dataset will be presented as a single numerical value which will be used to calculate Mvin. | | | | |
|  | | | | |
| Potential decision making depending on the results  Traceability will provide the degree to which data has attributes that provide an audit trail of access to the data and of any changes made to the data in a specific context of use. This will allow us to easily follow our data all the way back to its original source. It will help us maintain clear and accurate insights, ability to track every transformation, dead-end, or link between the data points. | | | | |

| **Base measure: RecTrace** | | |  | | |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **#** | Measure (what: entity, attribute)  Measures the total number of records that are traceable in MDS  Entity: Dataset  Attribute: Number of total records that can be traced in multiple datasets | | | | Scale Type  Absolute | Applicability  Helps us to understand how many records are tracebale in multiple datasets.It helps in finding traceability of a dataset. | | |
|  | | | |  |
| Who measures?  Data Analyst  Data Engineer  Data Scientist | | Source of measurement  Credit Card classification - https://www.kaggle.com/datasets/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)  This metric could be measured on a monthly, quarterly or yearly basis to calculate the accuracy trend of the database. | |
|  | |  | |  | |  |  | |
| Collection procedure (how to collect the data) | | | | Notes or comments: | | | | |
| 1. Dataset is loaded using the analyses tool,  excel file or jupyter notebook.  2. RecTrace is counted using COUNT function to  get number of credible records in a dataset using metadata  3. The value will be interpreted according to  the decision-making rules and appropriate  decision will be taken. | | | |

| **Base measure: Ldst** | | |  | | |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **#** | Measure (what: entity, attribute)  Measures total number of occurrences of data elements in dataset (DS)  Entity: Dataset  Attribute: Number of occurrences of data elements in a DS | | | | Scale Type  Absolute | Applicability  Helps in finding traceability of a dataset. | | |
|  | | | |  |
| Who measures?  Data Analyst  Data Engineer  Data Scientist | | Source of measurement  Credit Card classification - https://www.kaggle.com/datasets/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)  This metric could be measured on a monthly, quarterly or yearly basis to calculate the accuracy trend of the database. | |
|  | |  | |  | |  |  | |
| Collection procedure (how to collect the data) | | | | Notes or comments: | | | | |
| 1. Dataset is loaded using the analyses tool,  excel file or jupyter notebook.  2. Ldst is counted using COUNT function to  get number of credible records in a dataset  3. The value will be interpreted according to  the decision-making rules and appropriate  decision will be taken. | | | |