DescribeML:

Language Reference Guide

(Version 0.0.5)

DescribeML is a VSCode language plugin to describe machine-learning datasets.

Full examples of the language can be found in the public open repository here

General Structure:

- Metadata

    - Dates
    - Citation
    - Description
    - Applications
    - Distribution
    - Authoring

- Composition

    - Instances
    - Attributes
    - Statistics
    - Consistency Rules

- Provenance

    - Gathering processes
    - Labeling processes
    - Data preprocesses

- Social Concerns

Metadata:

- **Title:** STRING: The public title of the dataset

- **Unique-identifier:** ID Machine-readable unique identifier of the dataset

- **Version:** ID The version of the dataset

- **Date:** The date of the dataset

  - **Created:** DATE The date where the dataset was initially created:
  - **Modified:** DATE The date where the dataset was last modified:
  - **Published:** DATE The publication date of the dataset:

    > Example

    ```
    Dates:
        Release Date: 10-08-20
        Modified Date: 10-08-20
        Published Date: 10-08-20
    ```

- **Citation:** The citation of the dataset, between chose between a raw citation and a structured format

  - **Raw Citation:** STRING Raw citation as text, or as Bibtex or equivalent format, of the dataset
  - **OR:**
    - **Title:** STRING The title of the dataset
    - **Authors:** STRING The authors of the dataset
    - **Year:** DATE The year of the dataset
    - **Journal/Conference:** STRING The publisher of the dataset
    - **Publisher:** STRING The publisher of the dataset:
    - **URL:** URL The URL of the dataset
    - **DOI:** ID The DOI of the dataset
    - **ISBN:** ID The ISBN of the dataset

      > Example:

      ```
      Citation:
              Title: "SIIM-ISIC 2020 Challenge Dataset. International
      Skin Imaging Collaboration"
              Year: 2020
              Publisher: "International Skin Imaging Collaboration"
              DOI: "doi.org/10.34970/2020-ds01"
              Url: "https://www.kaggle.com/c/siim-isic-melanoma-
      classification"
      ```

- **Description:** The description of the dataset

  - **Description:** STRING Textual description of the dataset **OR:**

- **Purposes** `STRING` For what purposes was the dataset created?
- **Tasks:** `TASKS ENUMERATE` List of ML tasks the dataset is intended for: `Autocomplete feature will guide you through the options`
- **Gaps:** `STRING` Which gaps does the dataset aims to fill
  - **Areas:** `ID` Set a list of areas of the dataset
  - **Tags:** `ID, ...` Set a list of Tags of the dataset

> Example:

```
Description:
    Purposes:
        Purposes: "The 2020 SIIM-ISIC Melanoma"
        Tasks:    [classification]
        Gaps:     "As the leading healthcare organization for
informatics in medical imaging..."
        Areas:    HealthCare
        Tags:     Images Melanoma diagnosis SkinImage
```

- **Applications** Summerize the applications of the dataset

  - **Past Uses:** `STRING` Summerize the past uses of the dataset
  - **Recommended uses:** `STRING` Summerize the recommended uses of the dataset
  - **Non-recommended uses:** `STRING` Summerize the non-recommended uses of the dataset.
  - **Benchmarking:** Benchmarking of the dataset
    - **Task:** `TASKS ENUMERATE` Task to benchmark `Autocomplete feature will guide you through the options`
    - **Metric:** Metric to benchmark
      - **F1:** `NUMBER` F1 score
      - **Accuracy:** `NUMBER` Accuracy score
      - **Precision:** `NUMBER` Precision score
      - **Recall:** `NUMBER` Recall score
    - **Reference:** `STRING` Source of the benchmark

> Example

```
Applications:
    Past Uses: "The 2020 SIIM-ISIC Melanoma Classification... "
    Recommended:
        "Identify melanoma in lesion images."
        "Predict incidence of melanoma in a population."
    Non-recommended: "Due to low population prevalence and
challenges with access."
    Benchmarking:
        Task: Language-model
        [
            Model: "ModelExample"
            Metrics:[
```

```
                        F1: 81
                        Accuracy: 81
                        Precision: 81
                        Recall: 81
                ]
                Reference: "https://www.kaggle.com/c/siim-isic-
        melanoma-classification/leaderboard"
            ]
```

- **Distribution** Summerize the distribution of the dataset

  - **Is public?:** `BOOL` Indicate if the dataset is publicly available
  - **Licenses:** `LICENCES ENUMERATE` List of standard licenses, use others if not fit your case: `The Montreal data license , Creative Commons, CC0: Public Domain ...`
  - **Rights(stand-alone)** `ENUMERATE` Montreal data licence enumerate of stand-alone rights: Access | Tagging |'Distribute | Re-Represent
  - **Rights(with models):** `ENUMERATE` Montreal data licence enumerate of model related rights: `Benchmark | Research | Publish' | Internal Use | 'Output Commercialization' | Model Commercialization`
  - **Credits/Attribution Notice:** `STRING` Who needs to be credited when using the dataset
  - **Designated Third Parties:** `STRING` Third parties in charge of licensing and distribution issues
  - **Additional Conditions:** `STRING` Other issues specified by the authors

    > Example

    ```
    Distribution:
          Licences: CC BY 3.0 (Attribution 3.0 Unported)
          Rights(stand-alone): Access
          Rights(with models): Benchmark
          Additional Conditions "In addition to the CC-BY-NC license, the
    dataset is governed by the ISIC Terms of Use ... "
    ```

- **Authoring** Authoring of the dataset

  - **Authors** Authors of the dataset
    - **Name:** `STRING` Name of the author
    - **Email:** `EMAIL` Email of the author
  - **Founders** Founders of the dataset
    - **Name:** `STRING` Name of the founder
    - **Type:** `ENUMERATE` Type of the founder `private | public | mixed;`
    - **Grantor** `STRING` Grantor of the dataset
    - **Grant ID:** `ID` Machine-readable name of the grant id
  - **Maintainers** Maintainers of the dataset
    - **Name:** `STRING` Name of the maintainer
    - **Email:** `EMAIL` Email of the maintainer
  - **Erratum?:** `STRING` Is there any erratum?

- ○ **Data retention:** `STRING` Please indicate any data retention policy
- ○ **Version lifecycle:** `STRING` Describe the planned version lifecycle
- ○ **Contribution guidelines** `STRING` Is there any contribution guideline?

> Example:

```
Authoring:
    Authors:
        Name Skin_Imaging_Collaboration_ISIC  email emailo@emailo.com
        [...]
    Funders:
        Name The_University_of_Queensland  type mixed
            grantor "National Health and Medical Research Council
(NHMRC) – Centre of Research Excellence Scheme"
            grantId: APP1099021
        [...]
        Erratum?: "There is no erratum known"
        Contribution guidelines: "No contribution guidelines provided"
```

Composition:

- **Rationale** `STRING` Provide a composition rationale
- **Total Size** `NUMBER` Total size of tuples of the dataset
- **Instances** A composition description of each instance of the dataset

    - ○ **Instance:** `ID` Machine-readable name of the instance

    - ○ **Size:** `NUMBER` Size of the instance

    - ○ **Description:** `STRING` Description of the instance

    - ○ **Type:** `ENUMERATE` Type of the instance `Record-Data | Time-Series | Ordered | Graph | Other`

    - ○ **Attribute Number:** `NUMBER` Number of attributes

    - ○ **Attributes:** Description of each attribute of the instance

        - ▪ **attribute:** `ID` Machine-readable name of the attribute

        - ▪ **Description:** `STRING` Description of the attribute

        - ▪ **Associated label:** `Labels` Reference to a declared label in a labeling process (first you should complete the provenance part)

        - ▪ **unique values:** `NUMBER` Type of the attribute

        - ▪ **ofType:** `ENUMERATE` Type of the attribute `Categorical | Nominal` **If** `ofType` is `Categorical`

- **Statistics:** Statistic of the attribute

  - **Unique:** `NUMBER` Unique tuples (without duplications)

  - **Unique Percentage:** `NUMBER` Percentage of unique tuples

  - **Missing Values:** `NUMBER` Number of missing values

  - **Completeness:** `NUMBER` Completeness of the attribute

  - **Mode:** `STRING` Mode of the attribute

  - **First Rows:** `[0: ROW1, ...]` Percentage of the mode

  - **Min-leght:** `NUMBER` Min of the attribute

  - **Max-lenght:** `NUMBER` Max of the attribute

  - **Median-lenght:** `NUMER` Median lengths of the attribute

  - **Lenght-histogram:** `STRING` Histogram of the attribute

  - **Chi-Squared:** Chi-Squared of the attribute

    - **statistic:** Statistic of the chi-sqaure analysis
    - **p-value:** p-value of the chi-sqaure analysis

  - **Binary attribute:** `BOOL` Is a binary attribute?

    - **Symmetry:** `ENUMERATE Symmetryc | Asymmetryc`
    - **Attribute Sparsity:** `NUMBER` How sparse is the binary attribute?

  - **Categoric Distribution:** `["CATEGORY": "NUMBER"%, ...]` Categoric distribution of the attribute

    > Example:

    ```
    attribute: beningnant_malignant
        description: 'Type of the melanoma'
        label: skinLabel
        count: 33126
        ofType: Categorical
        Statistics:
            Missing Values: 0
            Completeness: 100
            Chi-Squared:
                p-value: 0
            Categoric Distribution:
                [
                    "beningnant": 80%,
                    "malignant": 20%
                ]
    ```

Else `ofType` is `Nominal`

- **Statistics:** Statistics of the attribute
  - **Mean:** `NUMBER` Unique tuples (without duplications)
  - **Median:** `NUMBER` Percentage of unique tuples
  - **Mode:** `NUMBER` Mode of the attribute
  - **Minimmum:** `NUMBER` Min of the attribute
  - **Maximmum:** `NUMBER` Max of the attribute
  - **Quartiles:** `[Q1:NUMBER, ...]` Median lengths of the attribute
  - **IQR:** `NUMBER` Histogram of the attribute

> Example

```
attribute: acidity
    description: 'wine acidity mesure'
    count: 33126
    ofType: Numerical
    Statistics:
            Mean: 4
            Median: 4.1
            Standard Desviation: 0.2
            Minimmum: 5
            Maximmum: 87
            Quartiles:  Q1:17 Q2:27 Q3:30 Q4:30
            IQR: 1.2
```

- **Statistics:** (instance) Statistic of the instance

  - **Correlations:** Correlation of the instance, choose one calculation type
    - **Pearson:** `[INDEX:"NUMBER", ...]` Pearson correlation of the instance
    - **Spearman:** `[INDEX:"NUMBER", ...]` Spearman correlation of the instance
    - **Kendall:** `[INDEX:"NUMBER", ...]` Kendall correlation of the instance
    - **Cramers:** `[INDEX:"NUMBER", ...]` Cramers correlation of the instance
    - **Phi-k** `[INDEX:"NUMBER", ...]` Phi-k correlation of the instance
  - **Pair Correlation** `Between [ATTRIBUTE], and [ATTRIBUTE]` Points the relevant pair-correlation between two instances of declared attributes.
  - **Quality Metrics:** General quality metrics of the instance
    - **Sparsity**: `NUMBER` Sparsity of the instance
    - **Completeness**: `NUMBER` Completeness of the instance
    - **Class balance**: `STRING` Class balance of the instance
    - **Noisy labels**: `STRING` Noisy labels of the instance

> Example:

```
Statistics:
    Correlations: Spearman: ['1': 0.2, '2':0.3, '3':0.4, '4':0.5,
'5':0.6, '6':0.7, '7':0.8, '8':0.9]
```

```
      Pair Correlation:
          between ImageId and diagnosis
          between age and external source
              From: "National statistical office"
              Rationale: "The age average is similar to the Nevada
  state age average due to
                          national statistical office average of
  2022 of Nevada"
      Quality Metrics:
          Completeness: 100
```

- **Consistency Rules:** Set the consistency rules of your dataset

    - **Rule:** `OCLExpression` OCL expression of the rule

> Example:

```
    Consistency rules:
    inv: skinImages : (age >= 0)
```

- **Dependencies:** Dependencies of the rule
    - **Description:** `STRING` Description of the dependencies
    - **Links:** `URL` Link to the dependency artifact
- **Instances relation:** `Relation: ID attribute: [ATTRIBUTE] is related to [INSTANCE]` Relation between instances

Provenance:

- **Curation Rationale** `STRING` Provide a provenance rationale
- **Gathering Processes:**
    - **Process:** `ID` Machine-readable name of the process
    - **Description:** `STRING` Description of the process
    - **When data was collected:** `STRING` Date where data the process was performed
    - **How data was collected** `STRING` How data was collected
    - **Is language data:** Set the speech situation
        - **Language:** `STRING` Language of the data
        - **Time and place:** `STRING`
        - **Modality:** `ENUMERATE` Modality of the speech `spoken/signed | written`
        - **Type:** `ENUMERATE` Type of the speech `scripted/edited | spontaneous`
        - **Syncrony:** `ENUMERATE` Synchrony of the speech `synchronous |asynchronous`
        - **Inteded Audience:** `STRING` Intended audience of the speech
    - **Social Issues:** `[SOCIAL ISSUES]` Relation of the gathering process with an already declared social issue instance
    - **Source:** Source of the data

- **Source:** `ID` machine-readable name of the source
- **Description:** `STRING` Description of the source
- **Noise:** `STRING` Description of the source's noise
- **Links:** `URL` Link to the source artifact
  - **Process Demographics:**
    - **Age:** `NUMBER` Median age of the participants
    - **Gender:** `STRING` Gender relation of the participants
    - **Country/Region** `STRING` Country/Region of the participants
    - **Race/Ethnicity** `STIRNG` Race or ethnicity of the participants
    - **Native Langugage** `STRING` Native language of the participants
    - **Socioeconomic status** `STRING` Socioeconomic status
    - **Number of speakers represented:** `NUMBER` Number of participants
    - **Precense of disorders in speech:** `STRING` Number of speakers
    - **Training in linguistics/other relevant disciplines** `STRING` Explain the training of the participants
  - **Gathering Team** Team in charge of gathering the data
    - **Who collects the data:** `STRING` Who collects the data
    - **Type** `ENUMERATE Internal | External | Contractors | Crowdsourcing`
    - **Demographics:** Demographics of the gathering team
      - **Age:** `NUMBER` Median age of the participants
      - **Gender:** `STRING` Gender relation of the participants
      - **Country/Region** `STRING` Country/Region of the participants
      - **Race/Ethnicity** `STIRNG` Race or ethnicity of the participants
      - **Native Langugage** `STRING` Native language of the participants
      - **Socioeconomic status** `STRING` Socioeconomic status
      - **Training in linguistics/other relevant disciplines** `STRING` Explain the training of the participants
  - **Gathering Requirements:** `Requirement: STRING, ...`

> Example:

```
Data Provenance:
    Curation Rationale:  "The curation process have been conducted by
several health institutions... "
    Gathering Processes:
        Process: GatheringProcess1
            Description:
                "The sources are: the Melanoma Institute Australia and
the ..."
            Source: GeneralHospital1
                Description: 'Source Description'
                Noise:
                    "Inconsistent lighting in images may alter skin
type"
                    "Duplicates:..."
            Related Instances: skinImages
            How data is collected: Manual Human Curator
            When data was collected:
```

```
                    Range: 1998 - 2019
                Process Demographics:
                    Country/Region: 'Australia'
                    [...]
                Gathering Team:
                    Who collects the data: "A team of dermatologists and
pathologists"
                        Type Internal
                Gather Requirements:
                    Requirement: "We queried clinical imaging databases across
the six centers to generate a ..."
```

- **LabelingProcesses:**
  - **Labeling process:** ID Machine-readable name of the labeling process
  - **Description:** STRING Description of the labeling process
  - **Type:** ENUMERATE 'Bounding boxes' | 'Lines and splines' | 'Semantinc Segmentation' | '3D cuboids' | 'Polygonal segmentation' | 'Landmark and key-point' | 'Image and video annotations' | 'Entity annotation' | 'Content and textual categorization
  - **Labels:** Labels of the labeling process
    - **Label:** ID Machine-readable name of the label
    - **Description:** STRING Description of the label
    - **Mapping:** [ATTRIBUTE,...] Relate a label with instances of attributes already declared in the documentation
  - **Labeling Team**:
    - **Who collects the data:** STRING Who collects the data
    - **Type** ENUMERATE Internal, External, Contractors, Crowdsourcing
    - **Demographics:** Demographics of the gathering team
      - **Age:** NUMBER Median age of the participants
      - **Gender:** STRING Gender relation of the participants
      - **Country/Region** STRING Country/Region of the participants
      - **Race/Ethnicity** STIRNG Race or ethnicity of the participants
      - **Native Langugage** STRING Native language of the participants
      - **Socioeconomic status** STRING Socioeconomic status
      - **Number of speakers represented:** NUMBER Number of participants
      - **Precense of disorders in speech:** STRING Number of speakers
      - **Training in linguistics/other relevant disciplines** STRING Explain the training of the participants
  - **Infrastructure:** Infrastructure used to annotate the data
    - **Tool:** STRING Tool used to annotate the data
    - **Platform:** STRING Platform where the tool works
    - **Version:** STRING Version of the tool and platform
    - **Language:** STRING Language of the tool
    - **Comments:** STRING Provide comments about the tool
  - **Validation:** Validation methods to ensure annotation quality
    - **Validation Methods:** STRING Validation method used
    - **Validation Dates:** STRING Dates where the validation where done annotations

- **Golden Questions:** Golden Question pass to the annotators
    - **Question:** `STRING` Textual question
    - **Inter-annotation agreement:** `NUMBER` Inter-annotation agreement for each question. Low values mean low confidence in the annotation
- **Validation Requirements:** `Requirement: STRING, ...` Provide comments about the validation tool
  - **Labeling Requirements:** `Requirement: STRING, ...`

> Example:

```
LabelingProcesses:
        Labeling process: skinLabeling
            Description: "Medical staff looking at the data and images
and annotating the diagnosis"
                Type: Image and video annotations
                Labels:
                    Label: skinLabel
                        Description: "marked as beningnant or malignant"
                        Mapping: beningnant_malignant
                Labeling Team:
                    Who collects the data: "Internal Medical staff"
                    Type Internal
                    Country/Region: "Australia"
                Label Requirements:
                    Requirement: "1) Images containing any potentially
identifying features, such as jewelry
```

- **Preprocesses:** Data preprocesses done over the data
  - **Preprocess:** `ID` machine-readable name of the preprocess
  - **Type:** `ENUMERATE` Type of preprocess applied `'Missing Values' | 'Data Augmentation' | 'Outlier Filtering' | 'Remove Duplicates' | 'Data reduction' | 'Sampling' | 'Data Normalization' | 'Others'`
  - **Description:** `STRING` Description of the preprocess
  - **Social Issues:** `[SOCIAL ISSUES]` Relation of the preprocess with an already declared social issue instance

---

Social Concerns

- **Social Concerns**
  - **Rationale:** `STRING` Rationale of the social concerns of the dataset
  - **Social Issues:** Social issues identified from the data
    - **Social Issue:** `ID` Machine-readable name of the social issue
    - **IssueType:** `ENUMERATE` Type of social concern `'Privacy' | 'Bias' | 'Sensitive Data' | 'Social Impact'`
    - **Description:** `STRING` Description of the social issue
    - **Related Attributes** `attribute: [ATTRIBUTE]` Attributes related to the social issue
    - **Instace belong to people:**

- **Have sensitive attributes?** `[Attribute], ...` List of sensitive attributes
- **Are there protected groups?** ENUMERATE (Yes, No, Unknown)
- **Might be offensive?** STRING Is there offensive content in the dataset

> Examples

```
Social Concerns:
    Rationale: 'Dataset may not be representative of the real
world data, and the cavenience sample is not representative of
general incidence of melanoma'
    Social Issue: raceRepresentative
        IssueType: Bias
        Description: "Dataset is not representative with respect
to darker skin types"
        Related Attributes:
            attribute: ImageId
```

For any related question, please contact the authors at: jginermi@uoc.edu