Data Quality Report

Dataset: Information of soil tests of cropland in Telangana state.

Data Type: Temporal

Number of Data Packets: 3225 | Start Time: 2022-06-11 11:49:03 | End Time: 2022-07-07 07:55:43

Overview

Metric	Score	Bar
Duplicate Presence	0.333	0.333 / 0.667
Adherence to Attribute Format	1.0	1 / 0
Absence of Unknown Attributes	1.0	1 / 0
Adherence to Mandatory Attributes	0.938	0.938 / 0.062

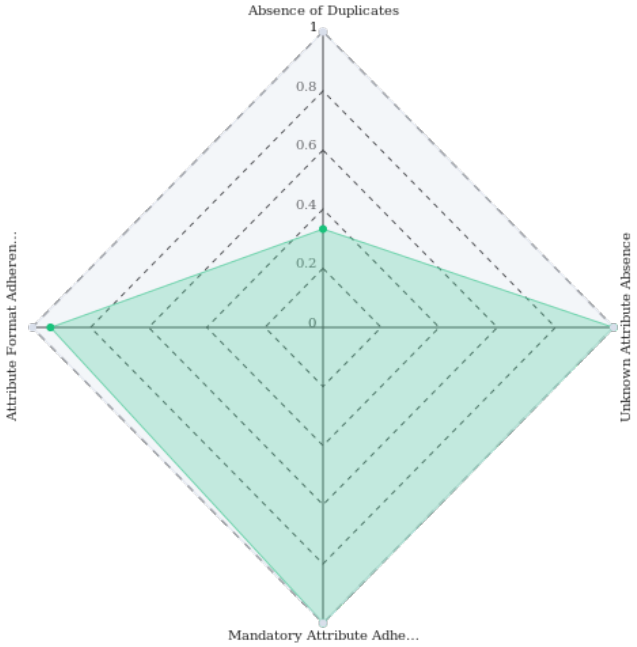
The Overall Data Quality Score of the dataset, computed by calculating an average of the above scores is:

0.818/1.00 or 81.8%

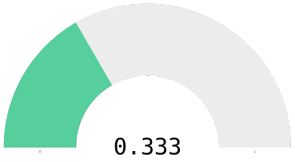
This data quality assessment report shows the score for four metrics that contribute to data quality.

The chart on the right shows an overview of the data quality of the dataset.

In the following pages you can find a detailed description and breakdown of each of these metrics.



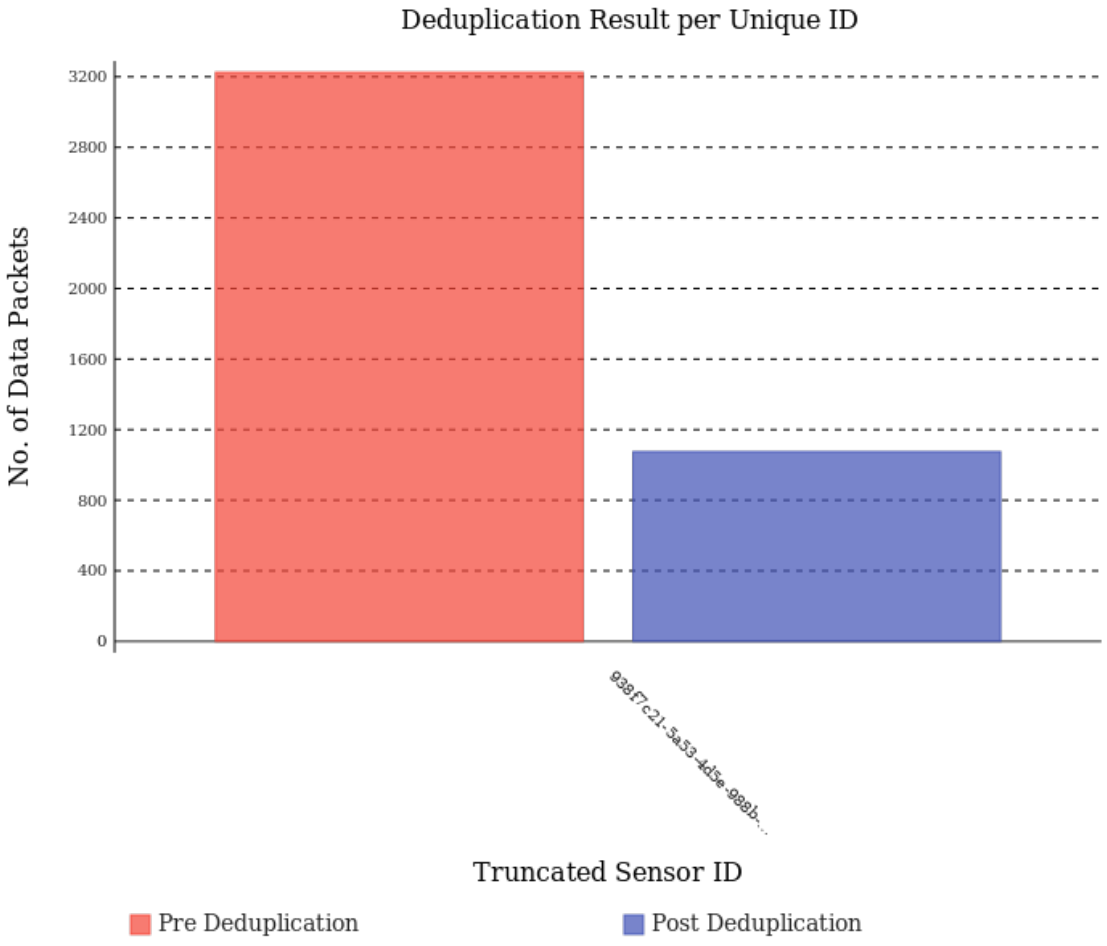
Duplicate Detection



This metric conveys how many duplicate data points are present in the dataset.

The duplicates in a dataset are identified as duplicates if any two data packets are received with exactly the same values for all the attributes within that data packet.

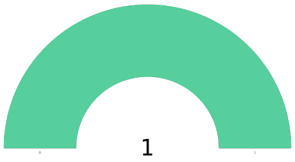
2150 duplicate data packets have been identified in the dataset.
This metric is calculated on a score from 0 to 1, where a score of 0 indicates that all the data packets are duplicates and a score of 1 indicates that none of the data packets are duplicates.
The chart below shows the number of data packets before and after deduplication on a per unique ID basis. If a unique ID is not represented in the chart, it means that there were no duplicate values received from that unique ID.



Metrics for Schema Analysis

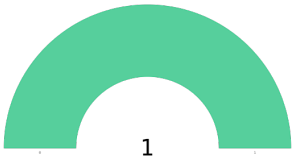
The remaining three metrics are an analysis of the metadata that is provided along with the dataset. This metadata is provided in the form of a schema, a document that delineates the different types of attributes, the data types of each attribute (integer, float, string, etc.) as well as the range of the observations under each attribute. This document also provides the mandatory attributes that the dataset must contain, as well as a list of all the expected attributes in the dataset.

Attribute Format Adherence



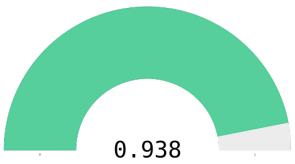
The attribute format metric checks whether the format of the data packets being evaluated matches the format defined in the data schema. The various possible formats include number, string, float, and object.
The format adherence metric is computed using the json schema validation method. The count of errors is incremented when the data type of an evaluated data packet does not match the data type specified in the data schema.
A higher score for the attribute format metric indicates a relatively lower proportion of data packets that contain attributes that do not adhere to the format defined in the schema, and a lower score for the attribute format metric indicates a relatively greater proportion of data packets with incorrect attribute formats.

Absence of Unknown Attributes



The unknown attributes A higher score for the attribute format metric indicates a relatively lower proportion of data packets that contain attributes that do not adhere to the format defined in the schema, and a lower score for the attribute format metric indicates a relatively greater proportion of data packets with incorrect attribute formats.metric computes the number of data packets with attributes that are present in the dataset but are not specified in the schema in any capacity.
This metric is computed by validating the data against the schema. A higher score for this metric indicates a relatively lower proportion of data packets that contain attributes that are not present in the data schema and a lower score indicates a relatively greater proportion of data packets with unknown attributes.
This metric represents the total number of unknown attributes in the dataset.

Adherence to Mandatory Attributes



The mandatory attributes metric checks whether the list of mandatory attributes defined in the data schema are all present in the dataset. This validation is performed for each data packet in the dataset.
A higher score for the mandatory attributes metric indicates that there is a relatively greater proportion of data packets with values present for all the mandatory attributes, and a lower score for the mandatory attributes metric indicates that there is a relatively lower proportion.
This metric is an indicator of the completeness of the dataset. Null values received under mandatory attributes are also included in the count of the number of missing attributes.