SNS COLLEGE OF TECHNOLOGY

Coimbatore-37.
An Autonomous Institution

COURSE NAME : 19CSE301 INTRODUCTION TO DATA SCIENCE

III YEAR/ V SEMESTER

UNIT – II
Topic: Data Quality

Ms.B.Sumathi
Assistant Professor
Department of Computer Science and Engineering

Introduction – Data Quality

- Measure a condition of data based on accuracy completeness the shared data is fit for the use of given purpose

  Eg: Duplicate data, Incomplete data

Three Measures:

- Data correctness: The collected data is in good manner without faulty data, outdated data or incorrect schema
- Data Freshness: Updated data is in up to date while sharing
- Data completeness: The collected data is in fully completed manner with effective content

Dimensions of data quality

- Accuracy: The data should reflect actual, real-world scenarios; the measure of accuracy can be confirmed with a verifiable source.
- Completeness: Completeness is a measure of the data's ability to effectively deliver all the required values that are available.
- Consistency: Data consistency refers to the uniformity of data as it moves across networks and applications. The same data values stored in difference locations should not conflict with one another.
- Validity: Data should be collected according to defined business rules and parameters, and should conform to the right format and fall within the right range.
- Uniqueness: Uniqueness ensures there are no duplications or overlapping of values across all data sets. Data cleansing and deduplication can help remedy a low uniqueness score.
- Timeliness: Timely data is data that is available when it is required. Data may be updated in real time to ensure that it is readily available and accessible.

Improvement of data Quality

- Data profiling - The first step in the data quality improvement process is understanding your data. Data profiling is the initial assessment of the current state of the data sets.
- Data Standardization - Disparate data sets are conformed to a common data format.
- Geocoding - The description of a location is transformed into coordinates that conform to U.S. and worldwide geographic standards
- Matching or Linking - Data matching identifies and merges matching pieces of information in big data sets.
- Data Quality Monitoring - Frequent data quality checks are essential. Data quality software in combination with machine learning can automatically detect, report, and correct data variations based on predefined business rules and parameters.
- Batch and Real time - Once the data is initially cleansed, an effective data quality framework should be able to deploy the same rules and processes across all applications and data types at scale.

References

- Tom M. Mitchell, “Machine Learning”, McGraw-Hill Education (India) Private Limited, 2013.
- 2Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, “An Introduction to Statistical Learning: with Applications in R”, Springer; First Edition 2013.


Thank
You