Professional Certificate in Data Quality Assurance using AI in Education

Data Validation Strategies

Data Validation Strategies: A set of techniques and approaches used to ensure the accuracy, quality, and reliability of data in the field of AI in Education. This glossary provides detailed explanations of key terms and concepts related to Data Validation Strategies.

Acronyms and Abbreviations:

* AI: Artificial Intelligence
* CRISP-DM: Cross-Industry Standard Process for Data Mining
* DQ: Data Quality
* DQA: Data Quality Assurance
* ML: Machine Learning

Challenge: A problem or task designed to test one's skills, knowledge, or ability to apply concepts in a practical setting.

Confidence Level: The probability that a given interval will contain the true value of a population parameter.

Data Cleaning: The process of identifying and correcting or removing errors, inconsistencies, and inaccuracies in a dataset to improve its quality.

Data Governance: The overall management and control of data-related activities, processes, and policies within an organization.

Data Mining: The process of discovering patterns, trends, and relationships in large datasets using statistical and machine learning techniques.

Data Profiling: The process of analyzing and understanding the characteristics, structure, and content of a dataset to identify potential quality issues and opportunities for improvement.

Data Quality (DQ): The degree to which data is accurate, complete, consistent, and fit for its intended use.

Data Quality Assurance (DQA): The process of ensuring that data is of high quality by implementing controls, checks, and validation procedures throughout the data lifecycle.

Data Quality Rules: Specific conditions or criteria that data must meet to be considered valid and reliable.

Data Validation: The process of checking and confirming that data meets specified quality rules and criteria.

Data Validation Framework: A structured approach to implementing data validation procedures, including the definition of quality rules, the selection of validation techniques, and the implementation of monitoring and reporting mechanisms.

Data Validation Techniques: Specific methods or tools used to validate data, such as data profiling, data cleaning, and machine learning algorithms.

Deep Learning: A subset of machine learning that uses artificial neural networks with multiple layers to model and analyze complex data patterns.

Machine Learning (ML): A subset of artificial intelligence that uses statistical and mathematical models to enable computers to learn and improve from data without being explicitly programmed.

Population: The entire group or universe of units about which inferences are to be drawn.

Probability Sampling: A sampling technique in which each unit in the population has a known, non-zero chance of being selected for the sample.

Random Sampling: A sampling technique in which each unit in the population has an equal chance of being selected for the sample.

Sample: A subset of units selected from a population for the purpose of estimation or testing.

Statistical Inference: The process of drawing conclusions about a population based on a sample of data.

Supervised Learning: A type of machine learning in which the algorithm is trained on labeled data, with both inputs and desired outputs provided.

Unsupervised Learning: A type of machine learning in which the algorithm is trained on unlabeled data, with only inputs provided.

Cross-Industry Standard Process for Data Mining (CRISP-DM): A widely adopted framework for planning and executing data mining projects, consisting of six phases: business understanding, data understanding, data preparation, modeling, evaluation, and deployment.

Data Cleaning Techniques: Methods used to identify and correct errors and inconsistencies in a dataset, such as:

* Data Imputation: The process of replacing missing or invalid data values with estimated or synthetic values.
* Data Normalization: The process of scaling numeric data values to a common range to improve comparability and reduce bias.
* Data Transformation: The process of converting data values from one format or structure to another to improve consistency and compatibility.
* Data Validation Rules: Specific criteria or conditions that data must meet to be considered valid and reliable, such as:
 + Format Validation: Checking that data values conform to specified formats, such as date or email address formats.
 + Range Validation: Checking that data values fall within specified ranges or limits.
 + Uniqueness Validation: Checking that data values are unique and do not duplicate other values in the

dataset.
* Data Profiling: The process of analyzing and understanding the characteristics, structure, and content of a dataset to identify potential quality issues and opportunities for improvement, such as:
 + Data Distribution: The distribution of data values across a range or interval.
 + Data Frequency: The number of occurrences of each unique data value in the dataset.
 + Data Nulls: The number and proportion of missing or null data values in the dataset.
 + Data Outliers: Data values that deviate significantly from other values in the dataset.
 + Data Redundancy: The presence of duplicate or unnecessary data values in the dataset.

Data Validation Framework: A structured approach to implementing data validation procedures, including:

* Data Quality Rules: Specific criteria or conditions that data must meet to be considered valid and reliable.
* Data Validation Techniques: Specific methods or tools used to validate data, such as:
 + Data Profiling: The process of analyzing and understanding the characteristics, structure, and content of a dataset to identify potential quality issues and opportunities for improvement.
 + Data Cleaning: The process of identifying and correcting or removing errors, inconsistencies, and inaccuracies in a dataset to improve its quality.
 + Machine Learning Algorithms: Statistical and mathematical models used to analyze and learn from data, such as decision trees, neural networks, and support vector machines.
* Monitoring and Reporting Mechanisms: Processes and tools used to track and report on data quality and validation performance, such as:
 + Data Quality Dashboards: Visual displays of key data quality metrics and indicators.
 + Data Quality Reports: Detailed reports on data quality issues, trends, and improvement opportunities.
 + Data Quality Alerts: Notifications of significant data quality issues or changes.

Data Validation Techniques: Specific methods or tools used to validate data, such as:

* Data Profiling: The process of analyzing and understanding the characteristics, structure, and content of a dataset to identify potential quality issues and opportunities for improvement.
* Data Cleaning: The process of identifying and correcting or removing errors, inconsistencies, and inaccuracies in a dataset to improve its quality.
* Machine Learning Algorithms: Statistical and mathematical models used to analyze and learn from data, such as decision trees, neural networks, and support vector machines.

Machine Learning Algorithms: Statistical and mathematical models used to analyze and learn from data, such as:

* Decision Trees: A type of machine learning algorithm that uses a tree-like model to represent decisions and their possible consequences.
* Neural Networks: A type of machine learning algorithm that uses artificial neural networks to model and analyze complex data patterns.
* Support Vector Machines: A type of machine learning algorithm that uses mathematical functions to classify and analyze data.

Data Quality Rules: Specific criteria or conditions that data must meet to be considered valid and reliable, such as:

* Format Validation: Checking that data values conform to specified formats, such as date or email address formats.