Data Preprocessing: Comprehensive Resource Guide

This guide covers Data Preprocessing fundamentals including: An Overview, Data Cleaning, Data Integration, Data Reduction, Data Transformation, and Data Discretization, with working resource links from SRT Telugu Lectures, Mahesh Huddar YouTube channels, and GeeksForGeeks.

---

1. Data Preprocessing: An Overview

Data preprocessing is the first and most crucial step in any data mining or machine learning pipeline. It involves preparing raw data for analysis by handling inconsistencies, cleaning noise, integrating data from multiple sources, reducing data volume, and transforming variables into suitable formats. Proper preprocessing ensures high-quality, reliable results and effective model training[1][2].

Why Data Preprocessing is Important:

- **Data Quality**: Real-world data is often incomplete, noisy, and inconsistent
- **Model Performance**: Clean, well-prepared data leads to better model accuracy
- **Computational Efficiency**: Reduced and transformed data speeds up processing
- **Feature Engineering**: Creates meaningful features from raw data
- **Handling Variability**: Normalizes different scales and formats

Major Steps in Data Preprocessing:

1. Data Cleaning - Handling missing values, noise, and inconsistencies
2. Data Integration - Combining data from multiple sources
3. Data Reduction - Reducing data volume while preserving information
4. Data Transformation - Converting data into appropriate formats
5. Data Discretization - Converting continuous data into discrete intervals

Video Resources:

- **Data Preprocessing Feature Engineering & Selection – Mahesh Huddar**
  https://www.youtube.com/watch?v=ifGJk2S3Y4U
  Covers the overall motivation, feature engineering, data cleaning, integration, transformation, and reduction with practical examples[3].
- **Read Data, Data Type, Describe - Mahesh Huddar**
  https://www.youtube.com/watch?v=0QJc6PrZiQs
  Introduction to data preprocessing with reading data and understanding data types[4].

Written Resources:

- **Data Preprocessing in Data Mining – GeeksForGeeks**
  https://www.geeksforgeeks.org/data-science/data-preprocessing-in-data-mining/

Comprehensive explanation of why preprocessing matters and details each major step[1].

- **What is Data Preprocessing in Data Science? – HCL GUVI**
  https://www.guvi.in/blog/what-is-data-preprocessing-in-data-science/
  Practical overview of preprocessing steps, techniques, and tools used in industry[2].

---

2. Data Cleaning

Data cleaning refers to the process of correcting errors, handling missing values, removing noise and outliers, and resolving inconsistencies in datasets. It is often the most time-consuming step in data preprocessing but is critical for ensuring data quality[5][6].

Key Data Cleaning Tasks

1. Handling Missing Values:

Missing data is a common problem in real-world datasets. Several strategies exist to handle missing values[5][7]:

- **Deletion Methods**:
  - Listwise deletion - Remove entire records with missing values
  - Pairwise deletion - Use available data for each analysis
  - Column deletion - Remove features with too many missing values
- **Imputation Methods**:
  - Mean/Median/Mode imputation - Replace with central tendency
  - Forward/Backward fill - Use previous/next valid value (time series)
  - Interpolation - Estimate values based on surrounding data points
  - Model-based imputation - Predict missing values using ML models
- **Indicator Variables**: Create binary flags indicating missingness

2. Handling Noisy Data:

Noise refers to random errors or variance in measured variables. Techniques include[5][8]:

- **Binning**: Smooth data by grouping into bins and replacing with mean/median
- **Regression**: Fit data to regression function to smooth noise
- **Clustering**: Detect and remove outliers using cluster analysis
- **Combined Methods**: Use human and computer inspection together

3. Handling Outliers:

Outliers are data points that deviate significantly from other observations[9][10]:

- **Detection Methods**:
  - Z-score method - Points beyond 3 standard deviations
  - IQR (Interquartile Range) method - Points outside 1.5 × IQR from quartiles
  - Visualization - Box plots, scatter plots
  - Statistical tests - Grubbs' test, Dixon's test
- **Treatment Options**:
  - Remove outliers if they are errors
  - Transform data to reduce outlier impact
  - Use robust methods less sensitive to outliers

– Keep outliers if they represent valid extreme cases

4. Removing Duplicates:

Identify and remove duplicate records that may arise from data integration or entry errors[5].

5. Correcting Inconsistencies:

Resolve contradictory data, standardize formats, and fix data entry errors[5].

Video Resources:

- **How to Handle Missing Values – Mahesh Huddar**
  https://www.youtube.com/watch?v=jsTmzmPzvok
  Detailed tutorial on imputation methods and practical approaches for handling missing data with Python examples[7].
- **How to Detect and Handle Outlier – Mahesh Huddar**
  https://www.youtube.com/watch?v=jTJu2ZFonzI
  Complete guide on outlier detection and removal techniques in data preprocessing[9].
- **Z-Score based Outlier Detection – Mahesh Huddar**
  https://www.youtube.com/watch?v=tW4MMhkTzhs
  Step-by-step tutorial on using Z-score method for anomaly detection and removal[10].
- **IQR based Outlier Detection – Mahesh Huddar**
  https://www.youtube.com/watch?v=YZ2rkajHGZY
  Comprehensive explanation of Interquartile Range method for outlier detection with solved examples[11].
- **Missing Values | Data Cleaning in Telugu**
  https://www.youtube.com/watch?v=rIAbXi_-Z_s
  Data cleaning and missing value handling explained in Telugu[12].
- **Data Cleaning in Data Mining – SRT Telugu Lectures**
  https://www.youtube.com/watch?v=jioJO1qjSy8
  Complete tutorial on handling various missing value scenarios and cleaning techniques in Telugu[13].
- **Data Cleaning: Missing Values and Binning**
  https://www.youtube.com/watch?v=G-RI3G2PRUs
  Focused video on missing values treatment and binning methods[14].
- **Tasks in Data Preprocessing – SRT Telugu Lectures**
  https://www.youtube.com/watch?v=R-2j-hNkCII
  Overview of data preprocessing tasks including cleaning, integration, and transformation in Telugu[15].

Written Resources:

- **Data Preprocessing in Python – GeeksForGeeks**
  https://www.geeksforgeeks.org/machine-learning/data-preprocessing-machine-learning-python/
  Practical Python guide for data cleaning with code examples[6].

3. Data Integration

Data integration involves combining data from different sources (databases, files, web services) into a unified, consistent dataset. This process addresses schema mismatches, deduplication, and normalization across heterogeneous sources[16][17].

Key Challenges in Data Integration

1. Schema Integration:

Different sources may have different schemas, naming conventions, and data representations[16]:

- **Entity Identification Problem**: Same entity with different names across sources
- **Attribute Matching**: Identifying equivalent attributes across schemas
- **Structural Conflicts**: Different ways of representing same information

2. Redundancy and Correlation:

Identifying and handling redundant data that appears in multiple sources[16]:

- Correlation analysis to detect redundant attributes
- Covariance analysis for numeric attributes
- Chi-square test for categorical attributes

3. Data Value Conflicts:

Resolving conflicts when same real-world entity has different values across sources[16]:

- Different units of measurement
- Different levels of precision
- Different representations (e.g., date formats)

4. Tuple Duplication:

Detecting and merging duplicate records representing the same entity[16].

Integration Approaches:

- **Tight Coupling**: Physical integration into single database
- **Loose Coupling**: Virtual integration with query-time combination
- **ETL Process**: Extract, Transform, Load workflow
- **Data Warehousing**: Centralized repository for integrated data

Video Resources:

- **Data Integration in Data Preprocessing – SRT Telugu Lectures**
  https://www.youtube.com/watch?v=ExfDFieOd7I
  Methods and issues in merging disparate datasets explained in Telugu[17].
- **Data Mining and Data Warehousing Videos**
  https://www.youtube.com/watch?v=IqWLxC0cw68
  SRT Telugu Lectures compilation including data integration concepts[18].

Written Resources:

- **Data Integration in Data Mining – GeeksForGeeks** https://www.geeksforgeeks.org/machine-learning/data-integration-in-data-mining/ Comprehensive guide on data integration principles, challenges, and techniques[16].

---

4. Data Reduction

Data reduction techniques reduce dataset size while retaining important information. This is essential for efficient storage, faster computation, and improved model performance, especially with large-scale datasets[19][20].

Data Reduction Strategies

1. Dimensionality Reduction:

Reducing the number of features (attributes) while preserving most information[19][21]:

- **Principal Component Analysis (PCA)**: Transform to uncorrelated principal components
- **Linear Discriminant Analysis (LDA)**: Maximize class separability
- **t-SNE**: Non-linear dimensionality reduction for visualization
- **Feature Selection**: Select subset of most relevant features
  - Filter methods - Statistical tests, correlation analysis
  - Wrapper methods - Use model performance for selection
  - Embedded methods - Feature selection during model training

2. Numerosity Reduction:

Reducing the number of data instances while maintaining data characteristics[19][22]:

- **Sampling**: Select representative subset of data
  - Random sampling - Simple random selection
  - Stratified sampling - Maintain class proportions
  - Cluster sampling - Sample from clusters
- **Histograms**: Approximate data distribution with bins
- **Clustering**: Replace cluster members with representative points
- **Regression**: Model data with regression equation

3. Data Compression:

Use encoding schemes to reduce data size[19]:

- **Lossless compression**: Original data can be reconstructed exactly
- **Lossy compression**: Some information loss but smaller size
- **Data cube aggregation**: Summarize data in data warehouses

Benefits of Data Reduction:

- Faster training and inference times
- Reduced storage requirements
- Mitigation of curse of dimensionality
- Improved model generalization
- Better data visualization

Video Resources:

- **Data Reduction Strategies in Data Mining – SRT Telugu Lectures**
https://www.youtube.com/watch?v=Jq32yJhdHns
Comprehensive tutorial on dimensionality reduction, numerosity reduction, and PCA explained in Telugu[19].
- **Data Reduction by Dr. Chiranjeevi Manike**
https://www.youtube.com/watch?v=OnkrqUxfpc4
Overview of different data reduction techniques and their applications[20].

Written Resources:

- **Dimensionality Reduction Techniques – GeeksForGeeks**
https://www.geeksforgeeks.org/data-science/dimensionality-reduction-techniques/
Detailed explanation of PCA, LDA, MDS, t-SNE, and other reduction methods[21].
- **Reduce Data Dimensionality using PCA - Python – GeeksForGeeks**
https://www.geeksforgeeks.org/machine-learning/reduce-data-dimentionality-using-pca-python/
Practical Python implementation of PCA for dimensionality reduction[23].
- **Model with Reduction Methods – GeeksForGeeks**
https://www.geeksforgeeks.org/machine-learning/ml-model-with-reduction-methods/
Explains feature selection and extraction methods with examples[22].
- **Curse of Dimensionality – GeeksForGeeks**
https://www.geeksforgeeks.org/machine-learning/curse-of-dimensionality-in-machine-learning/
Why dimensionality reduction is necessary and its impact on model performance[24].

5. Data Transformation

Data transformation involves changing data format or structure to make it suitable for analysis. This includes scaling, encoding, normalization, aggregation, and creating new features[25][26].

Key Transformation Techniques

1. Normalization and Standardization:

Scaling numerical features to similar ranges to prevent features with larger scales from dominating[27][28]:

- **Min-Max Normalization (Scaling)**:
$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$
Scales data to range [0, 1] or any [a, b]
- **Z-Score Normalization (Standardization)**:
$$x' = \frac{x - \mu}{\sigma}$$
Centers data around mean 0 with standard deviation 1

- **Decimal Scaling**:

  $$x' = \frac{x}{10^j}$$

  where j is smallest integer such that $\max(|x'|) < 1$
- **Robust Scaling**:
  Uses median and IQR, robust to outliers

2. Smoothing:

Removing noise from data using techniques like[29]:

- Binning (equal-width, equal-frequency)
- Moving averages
- Regression
- Clustering

3. Attribute Construction:

Creating new features from existing ones to capture more information[25]:

- Mathematical combinations (sum, difference, product, ratio)
- Polynomial features
- Domain-specific feature engineering

4. Aggregation:

Combining multiple attributes or records into summary values[30]:

- Sum, average, count, min, max
- Time-based aggregation (daily, weekly, monthly)
- Group-by operations

5. Discretization:

Converting continuous attributes to categorical (covered in detail in next section)[31].

6. Encoding Categorical Variables:

Converting categorical data to numerical format[25]:

- Label encoding - Assign integer labels
- One-hot encoding - Create binary columns for each category
- Target encoding - Use target variable statistics

Video Resources:

- **Feature Scaling: Min-Max, Z-Score – Mahesh Huddar**
  https://www.youtube.com/watch?v=XFBqx0icr-8
  Comprehensive tutorial on feature scaling including min-max normalization, z-score standardization, and robust scaling methods[27].
- **Min-Max Normalization & Z-Score – Mahesh Huddar**
  https://www.youtube.com/watch?v=jMvlyoegui4
  Step-by-step explanation of min-max and z-score normalization with solved numerical examples[28].

- **Min-Max Normalization in Data Mining**
  https://www.youtube.com/watch?v=8kVs2RV048c
  Detailed tutorial on min-max normalization technique[32].
- **Data Transformation Strategies – SRT Telugu Lectures**
  https://www.youtube.com/watch?v=A-1kdXLmmq8
  Covers smoothing, attribute construction, aggregation, and normalization in Telugu[29].
- **Smoothing, Aggregation, Normalization**
  https://www.youtube.com/watch?v=RQ0I1u-q8N8
  Complete tutorial on various transformation operations[30].
- **Aggregation Operators in DBMS – SRT Telugu Lectures**
  https://www.youtube.com/watch?v=R97rBH0L0mk
  Explains aggregation functions and their applications in Telugu[33].

Written Resources:

- **Data Transformation in Machine Learning – GeeksForGeeks**
  https://www.geeksforgeeks.org/machine-learning/data-transformation-in-machine-learning/
  Comprehensive guide on transformation methods and their importance[25].
- **Data Transformation in Data Mining – GeeksForGeeks**
  https://www.geeksforgeeks.org/dbms/data-transformation-in-data-mining/
  Detailed explanation of smoothing, aggregation, normalization, and discretization[26].
- **What is Data Transformation – GeeksForGeeks**
  https://www.geeksforgeeks.org/data-analysis/what-is-data-transformation/
  Overview of transformation techniques for data analysis workflows[31].

---

6. Data Discretization

Data discretization (also called binning or quantization) involves converting continuous numerical attributes into discrete intervals or categories. This technique is useful for improving algorithm performance, reducing computational complexity, and creating more interpretable models[34][35].

Why Discretization?

Benefits:

- **Reduced Complexity**: Fewer distinct values to process
- **Improved Interpretability**: Categories easier to understand than continuous values
- **Better Performance**: Some algorithms work better with discrete data
- **Noise Reduction**: Grouping can smooth out noise
- **Handling Outliers**: Extreme values grouped into bins

Discretization Methods

1. Equal-Width Binning (Uniform Binning):

Divides the range into k intervals of equal width[34][35]:

$$\text{Width} = \frac{\max(x) - \min(x)}{k}$$

- Simple and easy to implement
- May result in unbalanced bins if data is skewed
- Sensitive to outliers

2. Equal-Frequency Binning (Quantile Binning):

Creates bins with approximately equal number of data points in each bin[34][35]:

- More robust to outliers
- Ensures balanced bin sizes
- Bin boundaries adapt to data distribution

3. Binning by Cluster Analysis:

Uses clustering algorithms to identify natural groupings in data[34]:

- K-means clustering to find bin centers
- Assigns points to nearest cluster
- Adapts to data distribution

4. Decision Tree Discretization:

Uses decision tree algorithm to determine optimal split points based on class labels[34]:

- Supervised method (uses target variable)
- Finds splits that maximize information gain
- Creates bins that are meaningful for prediction

5. Histogram Analysis:

Analyzes data distribution to determine meaningful bin boundaries[35]:

- Identifies peaks and valleys in distribution
- Places boundaries at low-density regions
- Captures natural breaks in data

Smoothing by Bins

After discretization, smoothing techniques can be applied within bins[36][37]:

- **Smoothing by bin means**: Replace values with bin mean
- **Smoothing by bin medians**: Replace values with bin median
- **Smoothing by bin boundaries**: Replace with nearest boundary value

Video Resources:

- **Binning Methods – Mahesh Huddar**
  https://www.youtube.com/watch?v=K2fZYu2cX0c
  Complete tutorial on different types of binning and smoothing techniques[36].
- **Bin Boundary Data Mining – Mahesh Huddar**
  https://www.youtube.com/watch?v=dj11rv65L0c
  Detailed explanation of smoothing by bins and boundary detection methods[37].

- **Binning Data Mining | Equal Frequency Width Partitioning**
  https://www.youtube.com/watch?v=OPuJaNGGLeM
  Comparison of equal-width and equal-frequency binning with solved examples[38].
- **Data Cleaning/Binning – SRT Telugu Lectures**
  https://www.youtube.com/watch?v=G-RI3G2PRUs
  Shows binning as discretization method in Telugu[14].

Written Resources:

- **Discretization – GeeksForGeeks**
  https://www.geeksforgeeks.org/dbms/discretization/
  Comprehensive guide defining discretization approaches with worked examples and algorithms[34].
- **Discretization by Histogram Analysis – GeeksForGeeks**
  https://www.geeksforgeeks.org/data-science/discretization-by-histogram-analysis-in-data-mining/
  Explains histogram-based binning and cluster-based discretization methods[35].

Complete Playlists and Resources

Mahesh Huddar Playlists

- **Data Preprocessing for Data Mining Playlist**
  https://www.youtube.com/playlist?list=PL4gu8xQu0_5Le_OyCHx-fhTOIi-WDHjuy
  Complete playlist covering all aspects of data preprocessing including cleaning, integration, reduction, transformation, and discretization[39].
- **Data Science and Machine Learning Playlist**
  https://www.youtube.com/playlist?list=PL4gu8xQu0_5JfrwNOq9r1WtCmofDxzDEH
  Broader playlist including data preprocessing, feature engineering, and machine learning topics[40].
- **Machine Learning Playlist**
  https://www.youtube.com/playlist?list=PL4gu8xQu0_5JBO1FKRO5p20wc8DprlOgn
  Machine learning fundamentals including preprocessing as foundation[41].
- **Mahesh Huddar - YouTube Channel**
  https://www.youtube.com/@MaheshHuddar
  Main channel with comprehensive data mining and machine learning tutorials[42].

SRT Telugu Lectures Resources

- **SRT Telugu Lectures - YouTube Channel**
  https://www.youtube.com/@srttelugulectures
  Channel with data mining and preprocessing videos in Telugu[43].
- **Data Mining and Data Warehousing Lectures in Telugu Playlist**
  https://www.youtube.com/playlist?list=PL06g_pc9cPAjYqaLnLKmwDrZiiHtPp9FU
  Complete playlist covering data preprocessing and data mining concepts in Telugu[44].

Additional Resources

- **Data Preprocessing Notes From Mahesh Huddar - Scribd**
https://www.scribd.com/document/901162519/Data-Preprocessing-Notes-From-Mahesh-Huddar
Comprehensive notes covering all preprocessing topics[45].
- **Data Preprocessing: A Complete Guide - DataCamp**
https://www.datacamp.com/blog/data-preprocessing
Complete guide with Python examples for all preprocessing techniques[46].

Summary and Best Practices

Key Takeaways

1. **Data preprocessing is essential**: It directly impacts model quality and performance
2. **Understand your data first**: Use exploratory data analysis before preprocessing
3. **Handle missing values appropriately**: Choose method based on missingness pattern
4. **Scale features for distance-based algorithms**: Normalization/standardization crucial for many ML models
5. **Reduce dimensionality carefully**: Balance information retention with complexity reduction
6. **Document preprocessing steps**: Keep track of transformations for reproducibility

Best Practices

- **Split before preprocessing**: Fit preprocessing on training data only, apply to test data
- **Use pipelines**: Automate preprocessing steps to ensure consistency
- **Validate preprocessing impact**: Check how each step affects model performance
- **Consider domain knowledge**: Some transformations may not make sense in your domain
- **Handle outliers carefully**: Determine if they are errors or valid extreme values
- **Document decisions**: Record why certain preprocessing choices were made

Common Pitfalls to Avoid

- **Data leakage**: Using information from test set during preprocessing
- **Over-preprocessing**: Removing too much information from data
- **Ignoring data distribution**: Applying inappropriate transformations
- **One-size-fits-all approach**: Different algorithms need different preprocessing
- **Forgetting to save preprocessing objects**: Need to apply same transformations to new data

References

[1] GeeksForGeeks. (2019). Data Preprocessing in Data Mining. https://www.geeksforgeeks.org/data-science/data-preprocessing-in-data-mining/

[2] HCL GUVI. (2025). What is Data Preprocessing in Data Science? https://www.guvi.in/blog/what-is-data-preprocessing-in-data-science/

[3] Mahesh Huddar. (2021, August 6). Data Preprocessing Feature Engineering & Selection. *YouTube*. https://www.youtube.com/watch?v=ifGJk2S3Y4U

[4] Mahesh Huddar. (2021, August 6). Data Preprocessing Read Data Data Type Describe. *YouTube*. https://www.youtube.com/watch?v=0QJc6PrZiQs

[5] GeeksForGeeks. (2019). Data Preprocessing in Data Mining. https://www.geeksforgeeks.org/data-science/data-preprocessing-in-data-mining/

[6] GeeksForGeeks. (2017). Data Preprocessing in Python. https://www.geeksforgeeks.org/machine-learning/data-preprocessing-machine-learning-python/

[7] Mahesh Huddar. (2021, August 7). How to Handle Missing Values Data Preprocessing. *YouTube*. https://www.youtube.com/watch?v=jsTmzmPzvok

[8] GeeksForGeeks. (2019). Data Preprocessing in Data Mining. https://www.geeksforgeeks.org/data-science/data-preprocessing-in-data-mining/

[9] Mahesh Huddar. (2021, August 7). How to detect and handle Outlier Data Preprocessing. *YouTube*. https://www.youtube.com/watch?v=jTJu2ZFonzI

[10] Mahesh Huddar. (2023, February 24). Z-Score based Outlier or Anomaly detection and Removal. *YouTube*. https://www.youtube.com/watch?v=tW4MMhkTzhs

[11] Mahesh Huddar. (2023, February 22). Inter Quartile Range (IQR) based Outlier or anomaly detection. *YouTube*. https://www.youtube.com/watch?v=YZ2rkajHGZY

[12] Educational Resource. (2024, August 14). Missing Values | Data Cleaning | Data Preprocessing in Telugu. *YouTube*. https://www.youtube.com/watch?v=rIAbXi_-Z_s

[13] Educational Resource. (2024, January 25). Data cleaning in Data mining how to handle missing values. *YouTube*. https://www.youtube.com/watch?v=jioJO1qjSy8

[14] Educational Resource. DATA CLEANING-PART 1 (missing values and binning). *YouTube*. https://www.youtube.com/watch?v=G-RI3G2PRUs

[15] SRT Telugu Lectures. (2022, November 18). tasks in data preprocessing Telugu. *YouTube*. https://www.youtube.com/watch?v=R-2j-hNkCII

[16] GeeksForGeeks. (2019). Data Integration in Data Mining. https://www.geeksforgeeks.org/machine-learning/data-integration-in-data-mining/

[17] Educational Resource. (2022, December 2). data integration in data preprocessing. *YouTube*. https://www.youtube.com/watch?v=ExfDFieOd7I

[18] SRT Telugu Lectures. (2022, November 30). data mining and data warehouse videos. *YouTube*. https://www.youtube.com/watch?v=IqWLxC0cw68

[19] SRT Telugu Lectures. (2022, December 7). data reduction strategies in data mining and data warehousing. *YouTube*. https://www.youtube.com/watch?v=Jq32yJhdHns

[20] Dr. Chiranjeevi Manike. Data Reduction. *YouTube*. https://www.youtube.com/watch?v=OnkrqUxfpc4

[21] GeeksForGeeks. (2025). Dimensionality Reduction Techniques. https://www.geeksforgeeks.org/data-science/dimensionality-reduction-techniques/

[22] GeeksForGeeks. (2023). Model with Reduction Methods - Machine Learning. https://www.geeksforgeeks.org/machine-learning/ml-model-with-reduction-methods/

[23] GeeksForGeeks. (2022). Reduce Data Dimensionality using PCA - Python. https://www.geeksforgeeks.org/machine-learning/reduce-data-dimentionality-using-pca-python/

[24] GeeksForGeeks. (2024). Curse of Dimensionality in Machine Learning. https://www.geeksforgeeks.org/machine-learning/curse-of-dimensionality-in-machine-learning/

[25] GeeksForGeeks. (2024). Data Transformation in Machine Learning. https://www.geeksforgeeks.org/machine-learning/data-transformation-in-machine-learning/

[26] GeeksForGeeks. (2020). Data Transformation in Data Mining. https://www.geeksforgeeks.org/dbms/data-transformation-in-data-mining/

[27] Mahesh Huddar. (2021, August 7). How to perform feature scaling z score min-max and robust. *YouTube*. https://www.youtube.com/watch?v=XFBqx0icr-8

[28] Mahesh Huddar. (2023, January 8). Min-Max Normalization | Z-Score by Mean Absolute Deviation. *YouTube*. https://www.youtube.com/watch?v=jMvlyoegui4

[29] SRT Telugu Lectures. (2022, December 7). strategies of data transformation in data mining. *YouTube*. https://www.youtube.com/watch?v=A-1kdXLmmq8

[30] Educational Resource. (2019, September 18). Smoothing, Aggregation, Generalization, Normalization. *YouTube*. https://www.youtube.com/watch?v=RQ0I1u-q8N8

[31] GeeksForGeeks. (2024). What is Data Transformation? https://www.geeksforgeeks.org/data-analysis/what-is-data-transformation/

[32] Mahesh Huddar. (2021, April 11). Min Max Normalization Normalisation Data. *YouTube*. https://www.youtube.com/watch?v=8kVs2RV048c

[33] SRT Telugu Lectures. (2022, April 9). aggregation operators in DBMS. *YouTube*. https://www.youtube.com/watch?v=R97rBH0L0mk

[34] GeeksForGeeks. (2025). Discretization. https://www.geeksforgeeks.org/dbms/discretization/

[35] GeeksForGeeks. (2022). Discretization By Histogram Analysis in Data Mining. https://www.geeksforgeeks.org/data-science/discretization-by-histogram-analysis-in-data-mining/

[36] Mahesh Huddar. (2025, February 22). Binning methods. *YouTube*. https://www.youtube.com/watch?v=K2fZYu2cX0c

[37] Mahesh Huddar. (2023, January 7). Bin Boundary Data Mining. *YouTube*. https://www.youtube.com/watch?v=dj11rv65L0c

[38] Mahesh Huddar. (2025, October 5). Binning Data Mining | Equal Frequency Width Partitioning. *YouTube*. https://www.youtube.com/watch?v=OPuJaNGGLeM

[39] Mahesh Huddar. (n.d.). Data Preprocessing for Data Mining [Playlist]. *YouTube*. https://www.youtube.com/playlist?list=PL4gu8xQu0_5Le_OyCHx-fhTOIi-WDHjuy

[40] Mahesh Huddar. (n.d.). Data Science and Machine Learning [Playlist]. *YouTube*. https://www.youtube.com/playlist?list=PL4gu8xQu0_5JfrwNOq9r1WtCmofDxzDEH

[41] Mahesh Huddar. (n.d.). Machine Learning [Playlist]. *YouTube*. https://www.youtube.com/playlist?list=PL4gu8xQu0_5JBO1FKRO5p20wc8DprlOgn

[42] Mahesh Huddar. (n.d.). Mahesh Huddar [YouTube Channel]. *YouTube*. https://www.youtube.com/@MaheshHuddar

[43] SRT Telugu Lectures. (n.d.). SRT Telugu Lectures [YouTube Channel]. *YouTube*. https://www.youtube.com/@srttelugulectures

[44] SRT Telugu Lectures. (n.d.). Data mining and Data warehousing Lectures in Telugu [Playlist]. *YouTube*. https://www.youtube.com/playlist?list=PL06g_pc9cPAjYqaLnLKmwDrZiiHtPp9FU

[45] Scribd. (2025). Data Preprocessing Notes From Mahesh Huddar. https://www.scribd.com/document/901162519/Data-Preprocessing-Notes-From-Mahesh-Huddar

[46] DataCamp. (2025). Data Preprocessing: A Complete Guide with Python Examples. https://www.datacamp.com/blog/data-preprocessing