**Key Data Cleaning Techniques**

* **Handling Missing Values**
  + **Imputation:** Use statistical methods (mean, median for numerical; mode for categorical) or advanced techniques like k-nearest neighbors (KNN) for more accurate filling of gaps[1](https://ijsrst.com/IJSRST52411130)[2](https://intelliarts.com/blog/data-preprocessing-in-machine-learning-best-practices/)[7](https://www.aimlstudies.co.uk/index.php/jaira/article/view/304).
  + **Removal:** Drop rows or columns with excessive missing data if imputation isn’t suitable[2](https://intelliarts.com/blog/data-preprocessing-in-machine-learning-best-practices/)[3](https://moldstud.com/articles/p-strategies-for-data-cleaning-and-preprocessing-in-healthcare-analysis).
* **Outlier Detection and Correction**
  + **Statistical Methods:** Use Median Absolute Deviation (MAD) or visualization (boxplots, scatterplots) to spot outliers[2](https://intelliarts.com/blog/data-preprocessing-in-machine-learning-best-practices/)[3](https://moldstud.com/articles/p-strategies-for-data-cleaning-and-preprocessing-in-healthcare-analysis).
  + **Clustering Algorithms:** Cluster-based methods can identify and isolate anomalous data points[1](https://ijsrst.com/IJSRST52411130)[7](https://www.aimlstudies.co.uk/index.php/jaira/article/view/304).
  + **Winsorization:** Cap extreme values to reduce their influence[3](https://moldstud.com/articles/p-strategies-for-data-cleaning-and-preprocessing-in-healthcare-analysis).
* **Deduplication**
  + Identify and remove duplicate records to ensure each patient or event is unique[4](https://github.com/Kaludii/CSV-Data-Cleaning-Tool)[6](https://www.4medica.com/blog_insights/healthcare-data-cleaning/).
* **Error Correction and Validation**
  + Check for logical inconsistencies, invalid values, or impossible dates/times[6](https://www.4medica.com/blog_insights/healthcare-data-cleaning/).
* **Standardization of Formats**
  + Ensure consistent date formats, units, and capitalization across the dataset[2](https://intelliarts.com/blog/data-preprocessing-in-machine-learning-best-practices/)[3](https://moldstud.com/articles/p-strategies-for-data-cleaning-and-preprocessing-in-healthcare-analysis)[6](https://www.4medica.com/blog_insights/healthcare-data-cleaning/).

**Data Transformation and Preprocessing Techniques**

* **Normalization and Scaling**
  + **Min-Max Scaling:** Rescales features to a[1](https://ijsrst.com/IJSRST52411130) range.
  + **Z-score Standardization:** Centers features to mean 0 and standard deviation 1[2](https://intelliarts.com/blog/data-preprocessing-in-machine-learning-best-practices/)[3](https://moldstud.com/articles/p-strategies-for-data-cleaning-and-preprocessing-in-healthcare-analysis)[5](https://pmc.ncbi.nlm.nih.gov/articles/PMC11470224/).
  + **Why:** Prevents features with larger scales from dominating model training.
* **Encoding Categorical Variables**
  + **One-Hot Encoding:** Converts categorical columns into binary columns.
  + **Label Encoding:** Assigns each category a unique integer[2](https://intelliarts.com/blog/data-preprocessing-in-machine-learning-best-practices/)[3](https://moldstud.com/articles/p-strategies-for-data-cleaning-and-preprocessing-in-healthcare-analysis).
* **Skewness Correction**
  + Apply log, square root, or Box-Cox transformations to reduce skewness and improve model performance[2](https://intelliarts.com/blog/data-preprocessing-in-machine-learning-best-practices/).
* **Feature Selection and Dimensionality Reduction**
  + Use PCA, autoencoders, or tree-based feature importance to reduce dataset size and focus on relevant variables[5](https://pmc.ncbi.nlm.nih.gov/articles/PMC11470224/)[7](https://www.aimlstudies.co.uk/index.php/jaira/article/view/304).

**Best Practices for CSV Healthcare Data**

* **Automated Tools:** Use Python libraries like pandas and scikit-learn for cleaning and transforming CSV files.
* **Visualization:** Employ histograms, boxplots, and scatterplots to understand data distributions and identify issues[3](https://moldstud.com/articles/p-strategies-for-data-cleaning-and-preprocessing-in-healthcare-analysis).
* **Documentation:** Keep a record of all preprocessing steps for reproducibility and auditability.

**Practical Resources for CSV Data Cleaning**

* **CSV Data Cleaning Tool (Streamlit App):**
  + Upload, merge, remove duplicates/empty rows, and preview/download cleaned CSVs in a user-friendly interface[4](https://github.com/Kaludii/CSV-Data-Cleaning-Tool).
  + GitHub: Kaludii/CSV-Data-Cleaning-Tool
  + Requirements: Python, Streamlit, pandas.
* **Python Libraries:**
  + **pandas:** For reading, cleaning, and transforming CSV files.
  + **scikit-learn:** For advanced preprocessing (imputation, scaling, encoding).
  + **Streamlit:** For building interactive data cleaning apps

1. **Handling Missing Values:**

* Utilizes a **K-Nearest Neighbor (KNN)** algorithm-based approach to impute missing values. This method predicts the missing entries by leveraging the values of the nearest neighbors in the dataset, providing a more nuanced estimate than simpler techniques.

1. **Outlier Detection:**

* Employs a **cluster-based algorithm** for detecting outliers in medical data. Outliers, which can either represent noise or significant insights, are identified to improve the validity of the dataset.

1. **Handling Imbalanced Data:**

* Applies **Synthetic Minority Over-sampling Technique (SMOTE)** for over-sampling the minority class, thereby achieving a more balanced dataset.
* Implements **random resampling** for under-sampling the majority class when necessary, helping to reduce the bias that can occur in machine learning models due to class imbalance.

**Dealing with Outliers**

* Outliers can distort analysis and model performance. Identify and address them using robust methods:
  + Median Absolute Deviation (MAD): Detects outliers based on the median, reducing sensitivity to extreme values.
  + Winsorization: Caps extreme values to reduce their influence without removing data points.
  + Robust Regression: Applies models that downweight the effect of outliers on regression coefficients.
  + Data Segmentation: Divides data into subsets to minimize outlier impact within specific groups1.

**Standardizing and Scaling Variables**

* Standardization brings all variables to a common scale, eliminating unit inconsistencies and preventing bias in analysis.
  + Typically done by subtracting the mean and dividing by the standard deviation (z-score normalization).
* Scaling (e.g., min-max scaling, normalization) ensures variables are within a specific range, improving model stability and convergence speed1.

**Handling Missing Data**

* Missing data can bias results if ignored. Use imputation techniques to fill gaps:
  + Mean/Mode/Median Imputation: Simple but may introduce bias.
  + Regression Imputation: Predicts missing values using relationships between variables.
  + Multiple Imputation: Creates several imputed datasets to account for uncertainty.
  + K-Nearest Neighbor Imputation: Fills missing values based on similar data points1.

**Identifying and Removing Duplicate Records**

* Duplicate records can inflate storage, cause inconsistencies, and skew analysis.
  + Identify duplicates via field comparison, fuzzy matching, or data profiling.
  + Remove duplicates using automated scripts, data quality rules, database constraints, and regular data cleansing routines1.

**Key Takeaways**

* Robust outlier handling, standardization, and scaling are critical for accurate, unbiased analysis.
* Imputation techniques address missing data to ensure completeness and reliability.
* Regular identification and removal of duplicates maintain data integrity and efficiency.
* Adopting these preprocessing best practices enables RAG systems to generate more meaningful and trustworthy insights

| **Metric** | **What It Measures** | **When to Use/Notes** |
| --- | --- | --- |
| Accuracy | Overall correctness | Misleading if classes are imbalanced |
| Precision | Correctness of positive predictions | Important when FP are costly |
| Recall | Coverage of actual positives | Critical when FN are costly |
| Specificity | Coverage of actual negatives | Important for ruling out conditions |
| F1 Score | Balance between precision/recall | Good for imbalanced data |
| MCC | Overall correlation | Robust to class imbalance |
| AUC-ROC | Discrimination ability | Good overall performance indicator |