**Chunk 1: Missing Value Handling**

**Title:** Handling Missing Values  
**Content:** Missing data is addressed using techniques like mean/median imputation for numeric data and mode for categorical data. Advanced techniques such as K-Nearest Neighbors (KNN) imputation predict missing values using patterns from similar data points. When data is too sparse, rows or columns may be removed.

**Chunk 2: Outlier Detection Techniques**

**Title:** Outlier Detection and Correction  
**Content:** Outliers are identified using Median Absolute Deviation (MAD), visualization methods (e.g., boxplots), and cluster-based algorithms. Correction techniques include Winsorization and robust regression, which reduce the influence of extreme values without full removal.

**Chunk 3: Deduplication**

**Title:** Identifying and Removing Duplicates  
**Content:** Duplicates can be identified using field comparisons, fuzzy matching, and profiling. Their removal improves storage, consistency, and performance. Regular cleansing with automated scripts or data quality rules is recommended.

**Chunk 4: Scaling and Standardization**

**Title:** Standardization and Scaling  
**Content:** Standardization (z-score) and min-max scaling align feature ranges to prevent model bias due to varying units or magnitudes. These preprocessing steps are essential for stable training and convergence.

**Chunk 5: Encoding Categorical Variables**

**Title:** Categorical Encoding  
**Content:** One-hot encoding converts categories into binary columns; label encoding assigns unique integers. Proper encoding allows models to handle categorical variables effectively.

**Chunk 6: Dealing with Imbalanced Data**

**Title:** Handling Imbalanced Data  
**Content:** Use SMOTE for oversampling minority classes and random under-sampling to reduce bias from majority classes. These methods help models learn better decision boundaries in class-imbalanced datasets.

**Chunk 7: Evaluation Metrics**

**Title:** Evaluation Metrics for Classification  
**Content:** Use accuracy with caution on imbalanced datasets. Prefer metrics like Precision, Recall, F1-score, MCC, and AUC-ROC. Each provides a unique view: Recall is critical when missing a condition is risky; Specificity is important for ruling out false positives.