**5. How do you handle imbalanced datasets? Would the approach differ when the imbalance ratio is 1:10 versus 1:10000?**

**(This table is converted to image format to fit the page limits)**

A screenshot of a document

Description automatically generated

**I. Handling Imbalanced Data**

**Data-Level Methods:**

1. **Data Collection:**
   * Collect more training samples for the minority subsets.
   * Generate synthetic data to balance the dataset.
2. **Resampling Methods:**
   * Increase the sampling weight of the minority data samples.

**Algorithm-Level Methods:**

1. **Cost-Sensitive Learning:**
   * Assign higher learning weights to the minority data samples.

**Combine with Changing Performance Metric:**

* + In regular classification, we evaluate results using metrics like Confusion Matrix, Recall, Precision, and F-measure/Accuracy.
  + For imbalanced data, consider using Kappa (or Cohen’s kappa), which normalizes classification accuracy by the imbalance of the classes in the data.

1. **Different Modeling Approaches:**
   * Decision trees can still perform well with imbalanced data.
   * Use anomaly detection/change detection methods designed to recognize rare and irregular cases.
   * Combine results from multiple models (ensemble methods).

**II. Does the Approach Differ When the Imbalance Ratio is 1:10 Versus 1:10,000?**

Yes, the approach would differ. If we heavily resample a highly imbalanced dataset (like 1:10,000), we might miss important cases from the majority group in our training samples. To handle this:

* Maintain resampling at an adequate level.
* Combine with cost-sensitive learning to give more weight to minority samples while preserving important cases from the majority samples.