Notes

Saturday 11 May 2024 4:12 AM

→ Dataset where the classes / categories are not equally represented.
- → Majority class
- → Minority class

→ Applicable for both binary & multi-class problems.

→ The problem?
- → BIAS FOR MAJORITY CLASS
- → Misleading Metrics
- → Problem may go unnoticed

→ Importance
- Medical & healthcare
- Fraud detection
- Manufacturing & Quality control
- Customer Churn
- Cybersecurity

→ Techniques
- Data based
  - under sampling
  - over sampling
  - SMOTE
  - Hybrid sampling
- Algo based
  - Class Weighting
  - cost sensitive learning
  - Ensemble Methods
- Tuning based — Threshold tuning

```
sklearn.model_selection.train_test_split(*arrays, test_size=None, train_size=None, random_state=None,
shuffle=True, stratify=None)                                                                    [source]
```

https://imbalanced-learn.org/stable/common_pitfalls.html

Undersampling
- Random undersampling
- Tomek links
- Edited Nearest Neighbors
- Neighborhood cleaning Rule

→ Aim to declutter the decision boundary.

- Neighborhood cleaning Rule
- Cluster centroids

- oversampling
  - Random oversampling
  - SMOTE
  - Boderline SMOTE
  - ADASYN
  - SVM SMOTE
  - SMOTENC [When categorical data is present]