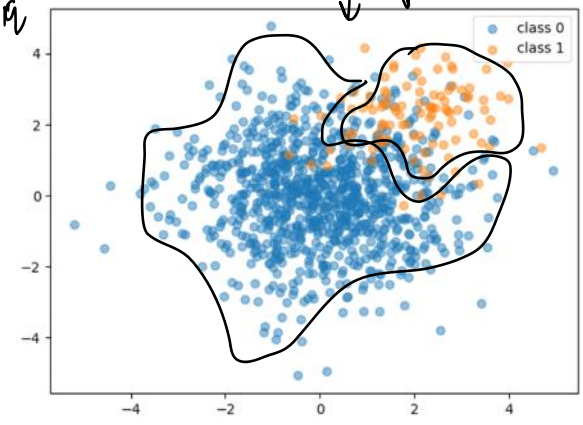
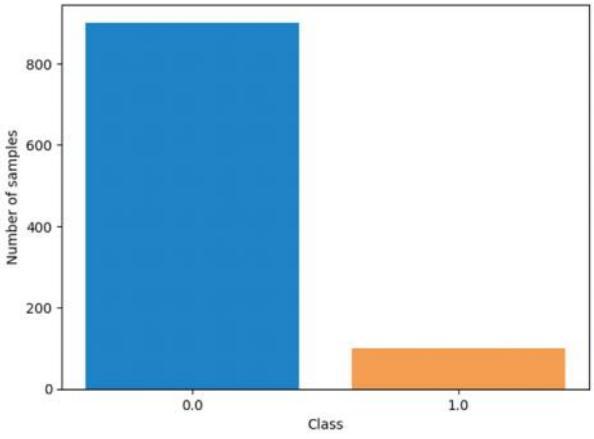
What is Imbalanced Data

24 April 2024 07:32

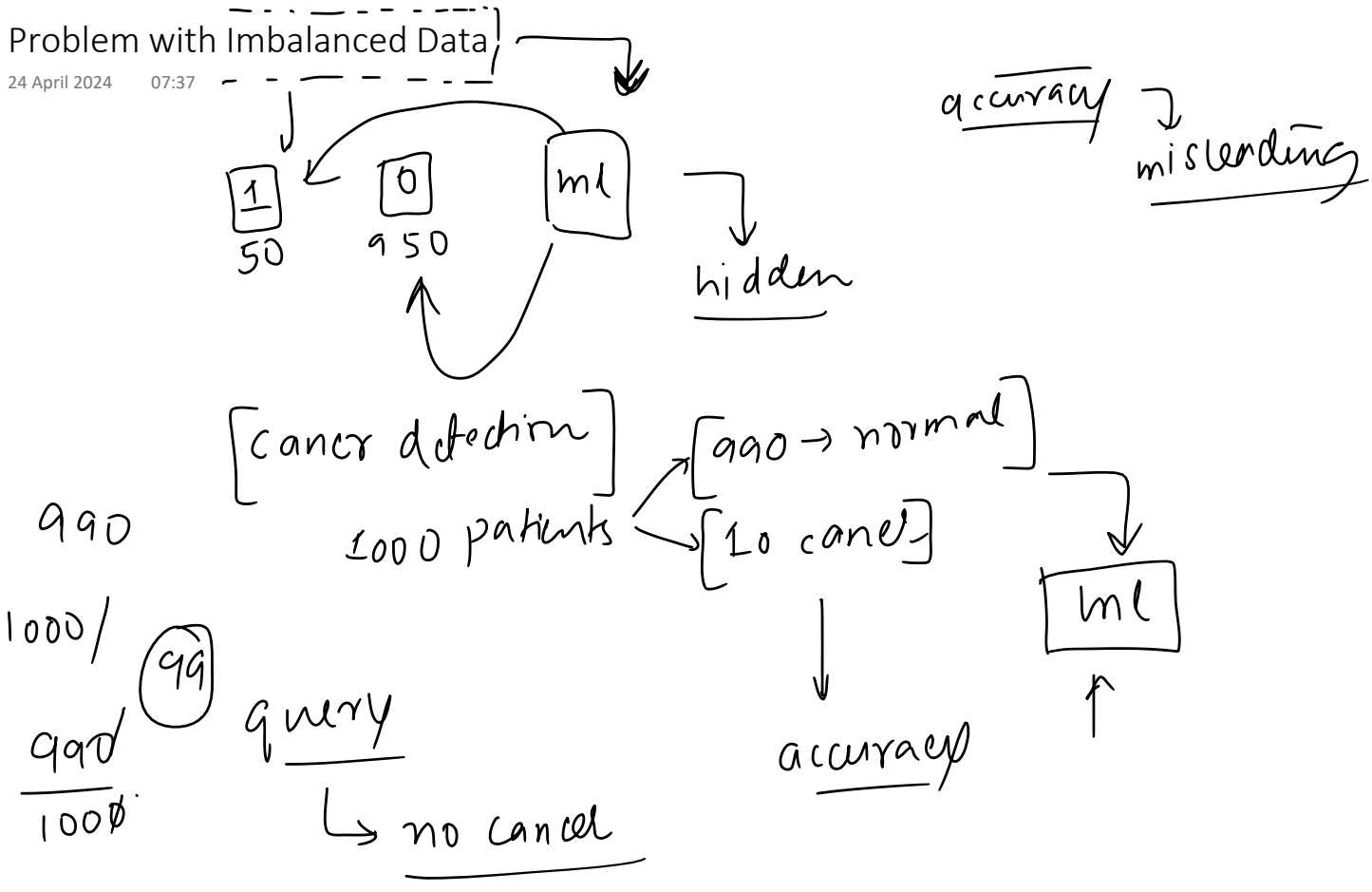
Imbalanced data refers to a situation in a dataset where the classes or categories are not equally represented. It can happen in both binary classification setup as well as multi-class classification setup.





Problem with Imbalanced Data

24 April 2024 07:37

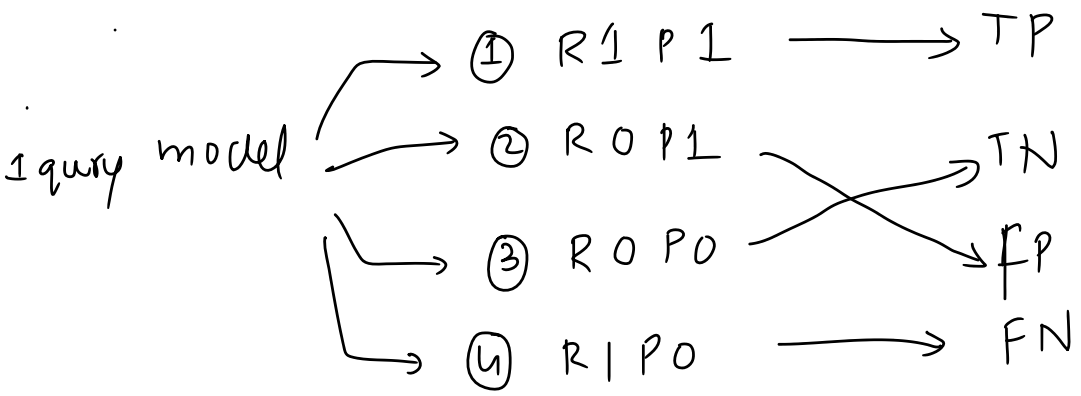


A Quick Detour!

24 April 2024 08:43

Cancer detection

TP | TN | FP | FN



Spam detection

0 not spam 1 spam

TP → spam 1 m → 1

TN → spam 0 m → 0

FP → spam 0 m → 1

FN → spam 1 m → 0

5 points → model

1	0	
2	1	0
3	0	
4	1	0

① 5 → 1|1

③ ①

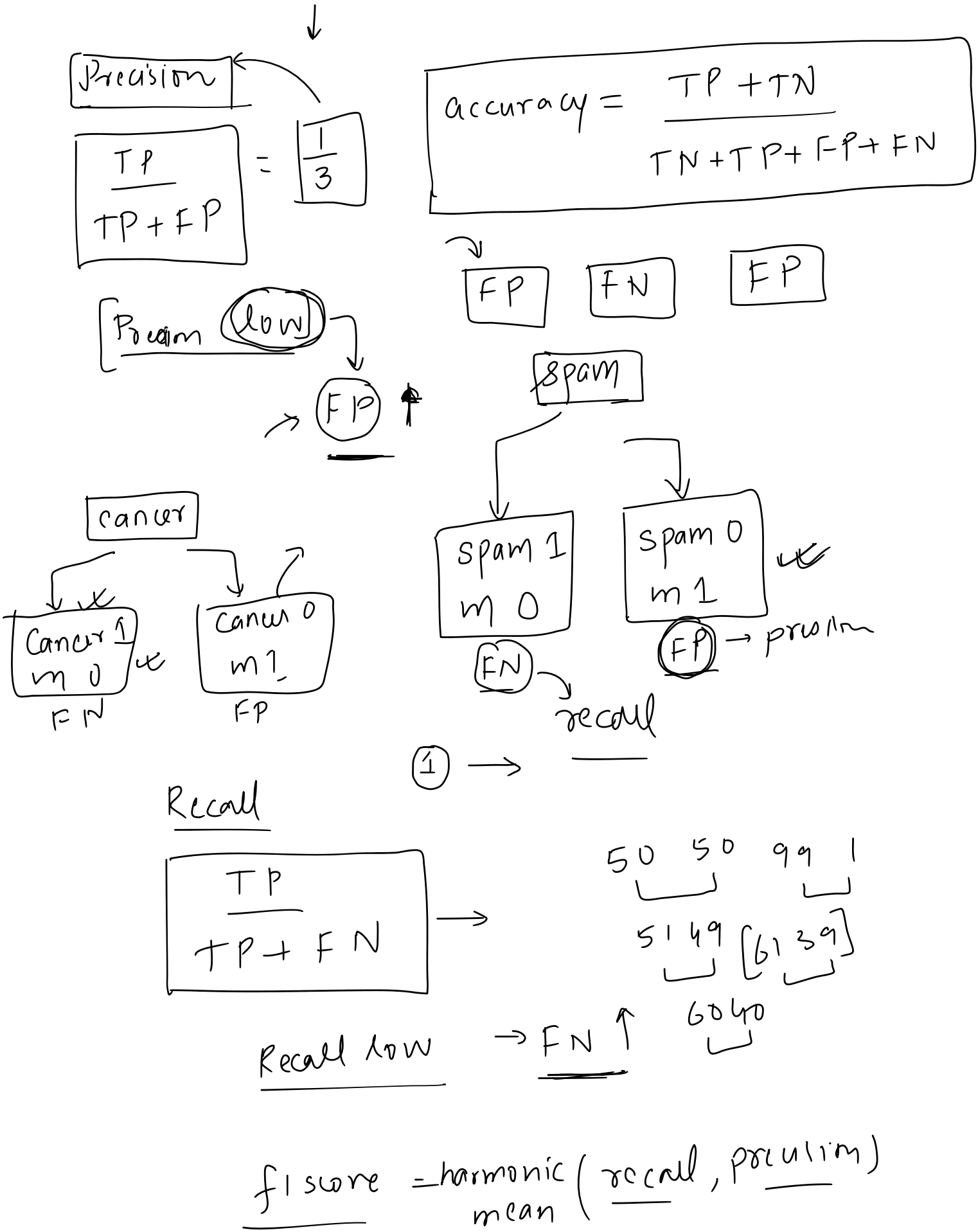
True 0

Pred

	0	1
True 0	TN	FP
True 1	FN	TP

confusion matrix

Accuracy


Precision
TP / TP+FP = 1/3
accuracy = TP+TN / TN+TP+FP+FN
FP
FN
FP
Spam low
FP
spam
cancer
Cancer 1
m 0
FN
cancer 0
m 1
FP
spam 1
m 0
FN
spam 0
m 1
FP
recall
1
Recall
TP / TP+FN
50 50 99 1
51 49 61 39
60 40
Recall low → FN ↑
f1 score = harmonic mean (recall, precision)

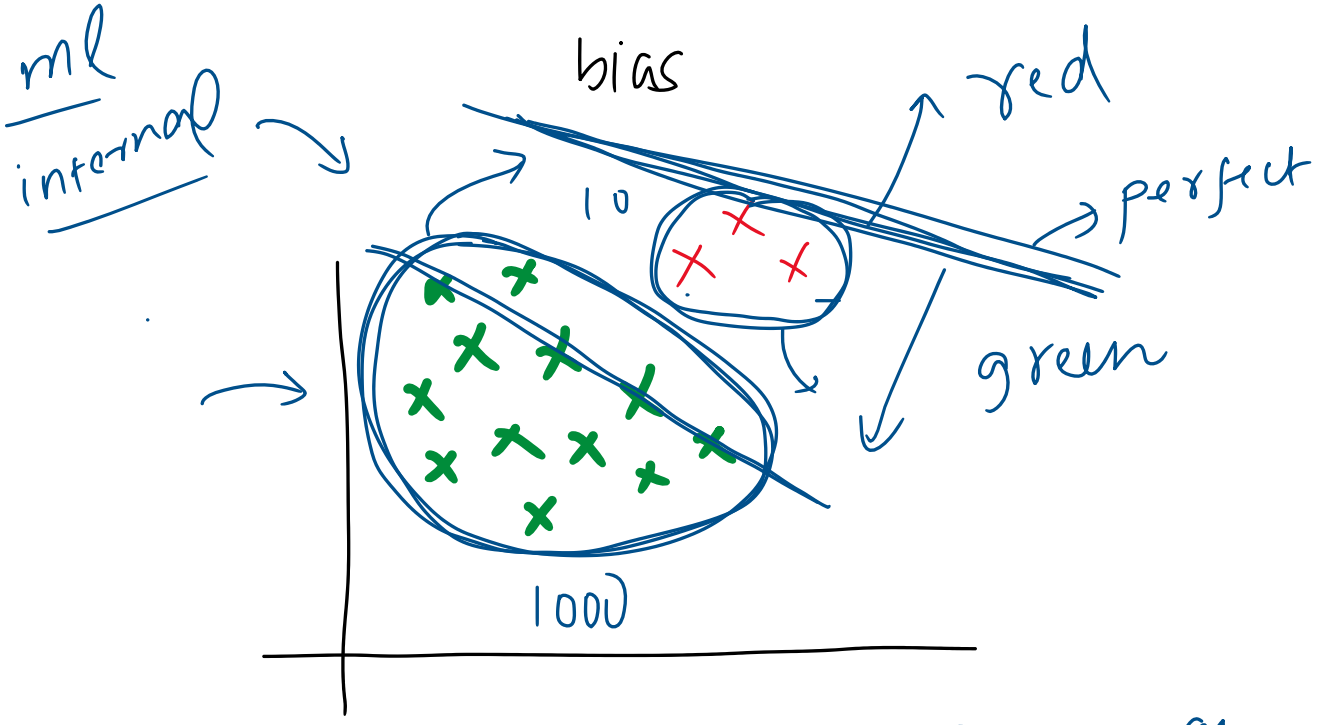
Why algos perform badly on Imbalanced Data?

24 April 2024 08:35

imbalance → perform

ml → data → [majority]

↳ minory →



990 99 accuracy

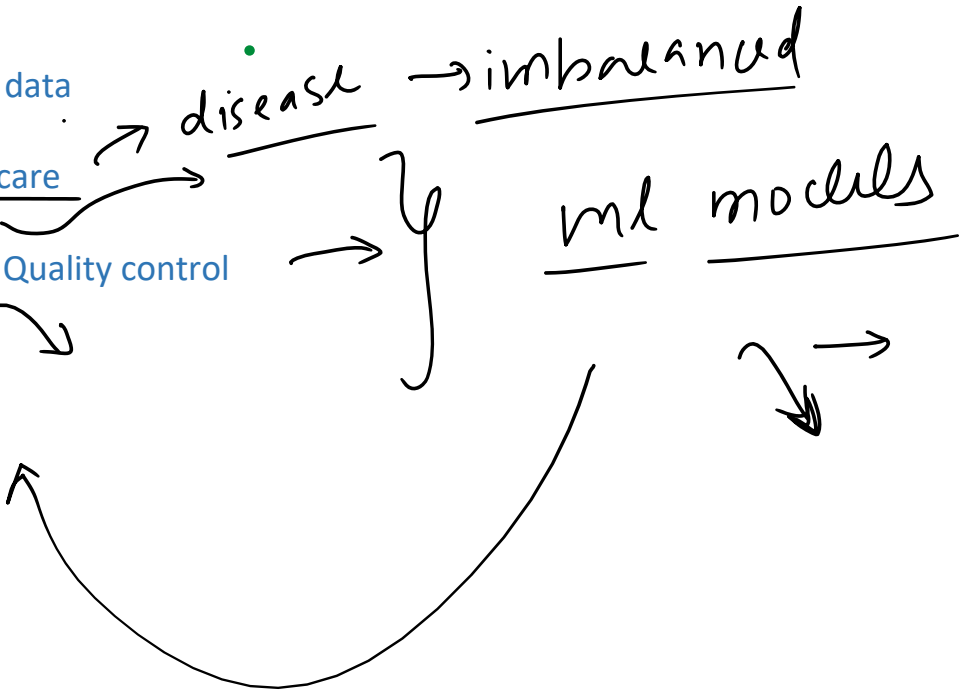
1000

Why Imbalanced Data is a huge deal? → imp

24 April 2024 08:35

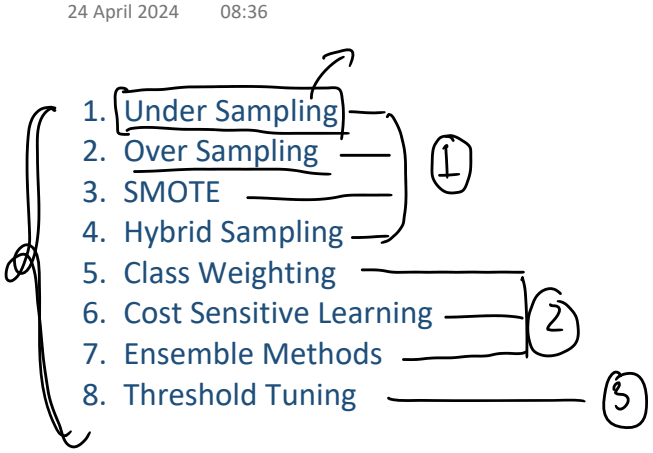
Where can we find such data

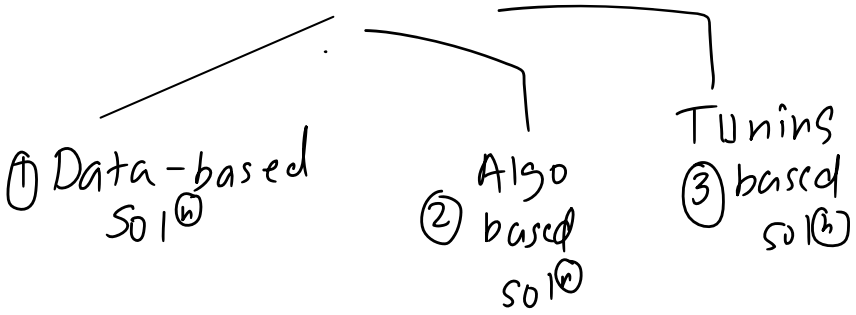
1. Medical and healthcare
2. Fraud Detection
3. Manufacturing and Quality control
4. Customer Churn
5. Cybersecurity



Techniques

24 April 2024 08:36

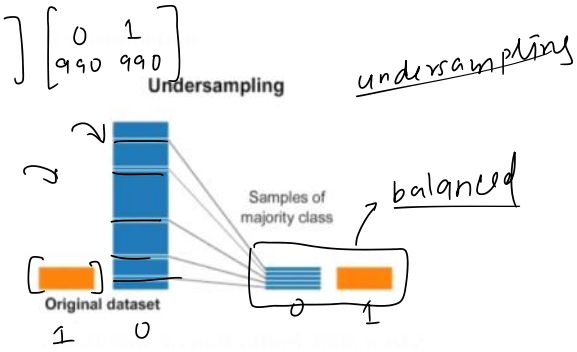




1 Under Sampling

24 April 2024 16:33

Under sampling is a technique used to balance imbalanced datasets. When a dataset has unequal distribution of classes, meaning one class (the majority class) has more samples and another class (the minority class) has fewer samples, under sampling can be used to address imbalanced data issues.



Concept of Under sampling

The concept of under sampling is straightforward: you randomly remove some samples from the majority class to create a balance between the classes. The objective is to reduce the size of the majority class to bring it closer to the minority class, allowing machine learning algorithms to work with balanced training data.

Why Is Under sampling Important?

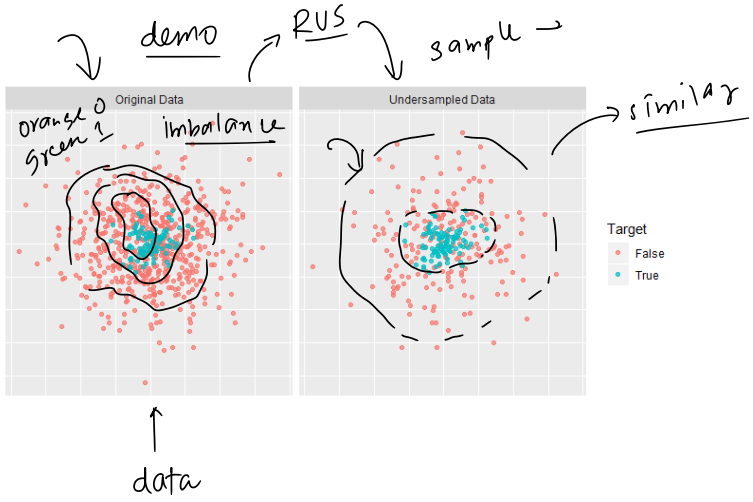
Imbalanced datasets often lead to models biased toward the majority class. By using under sampling, you can balance the learning, enabling models to properly learn from both classes. It offers several benefits:

- Reduction of Bias: Under sampling reduces the size of the majority class, thus reducing model bias and helping it learn from the minority class.
- Faster Training: With under sampling, the dataset becomes smaller, leading to faster training times.
- Balanced Learning: Models can learn from balanced data, allowing them to properly understand both classes.

Risks of Under sampling

However, there are some risks associated with under sampling:

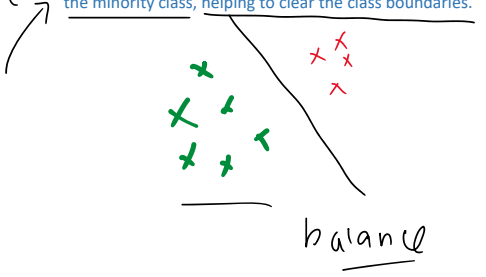
- Loss of Information: When you remove samples from the majority class, you might lose useful information, affecting model performance.
- Overfitting: A smaller dataset can lead to overfitting, especially if the minority class has limited examples.
- Sampling Bias: If under sampling is not done randomly, it can create a sampling bias, leading to a biased model.



Under sampling Techniques

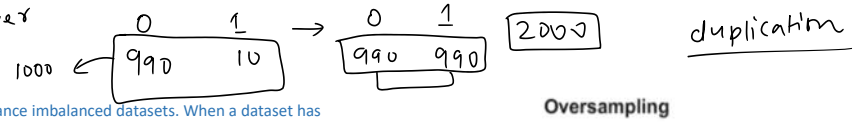
There are several techniques for under sampling to balance imbalanced datasets:

- Random Under sampling: This technique involves randomly removing examples from the majority class to create balance.
- Cluster-based Under sampling: In this approach, examples from the majority class are grouped into clusters, and then some examples are chosen from each cluster to be removed.
- Tomek Links: This technique removes examples from the majority class that are close to the minority class, helping to clear the class boundaries.



2. Over Sampling

24 April 2024 15:54



Oversampling is a technique used to balance imbalanced datasets. When a dataset has unequal class distribution, meaning one class (the majority class) has more samples and another class (the minority class) has fewer samples, oversampling is used to increase the size of the minority class to create a balance.

Concept of Oversampling

The concept of oversampling is to increase the number of examples in the minority class to bring it closer to the majority class in terms of size. The objective is to provide balanced training data to machine learning algorithms so that they can learn patterns from both classes.

Why Is Oversampling Important?

In imbalanced datasets, models often favour the majority class because it has more samples. Oversampling increases the size of the minority class, allowing models to learn from it properly. Here are some important reasons why oversampling is popular:

- Improved Learning: Oversampling provides more examples from the minority class, allowing models to learn more effectively.
- Reduced Bias: Oversampling creates a balanced dataset, reducing the model's bias toward the majority class.
- Enhanced Performance: A balanced dataset often improves model performance because it represents both classes equally.

Risks of Oversampling

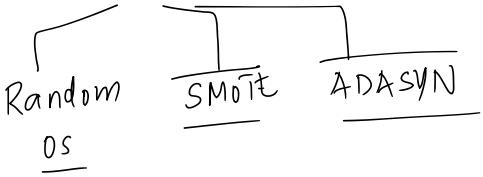
However, there are risks associated with oversampling:

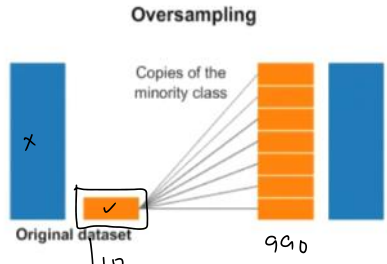
- Overfitting: When minority class examples are oversampled, there is a risk of overfitting because the model may learn too much from the specifics of these examples.
- Increased Data Size: Oversampling increases the dataset's size, requiring more computational resources and potentially slowing down training.
- Redundancy: Oversampling often involves duplicating some examples, leading to redundancy in the dataset.

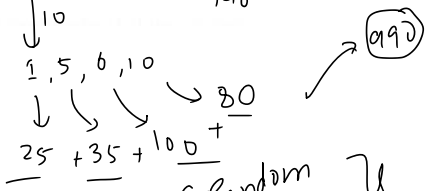
Oversampling Techniques

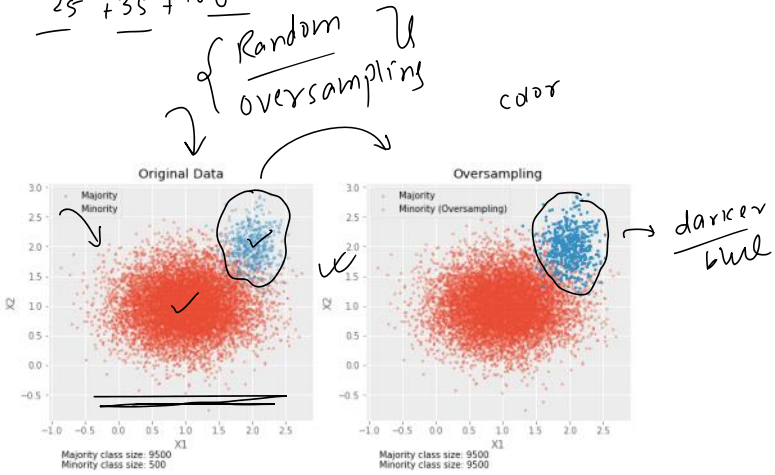
To balance imbalanced datasets, there are several oversampling techniques:

- Random Oversampling: This technique involves randomly duplicating examples from the minority class to match the size of the majority class.
- SMOTE (Synthetic Minority Over-sampling Technique): This is an advanced technique where new synthetic examples are created based on existing minority class examples. The goal is to avoid duplication and bring diversity to the dataset.
- ADASYN (Adaptive Synthetic Sampling): This is a variation of SMOTE that creates more synthetic examples where the model has more difficulty.









3. SMOTE

24 April 2024 16:38

oversampling minory ↑ → ROS

SMOTE stands for "Synthetic Minority Over-sampling Technique." It's a technique used to increase the size of the minority class in imbalanced datasets. The goal of SMOTE is to increase the number of examples in the minority class by creating synthetic data points, providing more balanced data for machine learning models.

Concept of SMOTE

The main concept of SMOTE is to create synthetic examples among existing examples in the minority class. This means generating new samples artificially to balance the data, providing sufficient training data for the model to learn the patterns in the minority class effectively.

SMOTE operates in the following way:

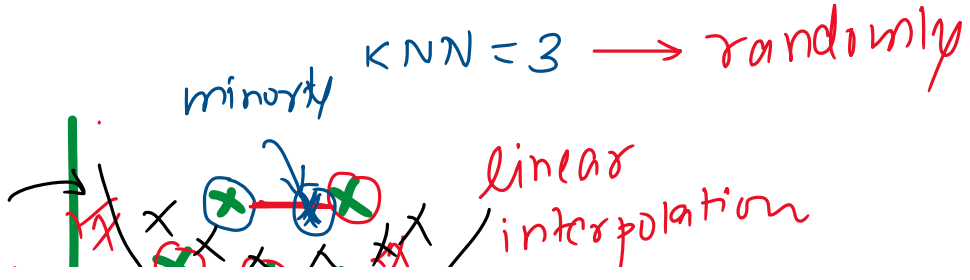
- Using Nearest Neighbours: SMOTE begins by finding the nearest neighbours for each example in the minority class. These nearest neighbours belong to the same class.
- Creating Synthetic Examples: For each example in the minority class, a few nearest neighbours are randomly selected. A synthetic example is then generated between the original example and the selected neighbour. Linear interpolation is used to create the synthetic example.
- Creating a Balanced Dataset: In this way, new synthetic examples are created for the minority class, bringing its size closer to or matching that of the majority class.

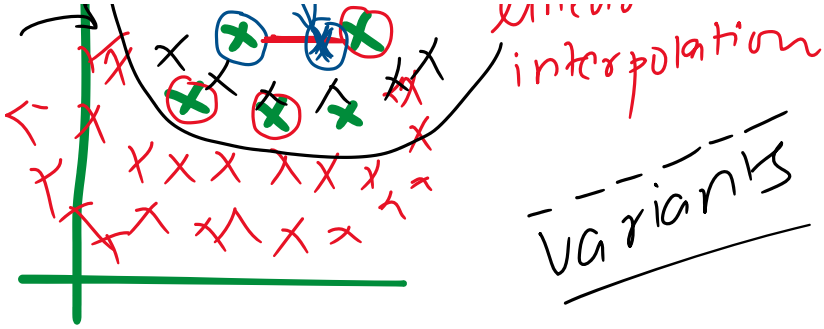
Benefits of SMOTE

- Reduces Overfitting: Because SMOTE creates synthetic examples, it avoids duplication, reducing the risk of overfitting.
- Improves Learning: The increase in the size of the minority class due to synthetic examples provides more training data, enabling the model to better learn the patterns in the minority class.
- Enhances Model Performance: Balanced datasets generally lead to better model performance, as they represent both classes more accurately.

Risks of SMOTE

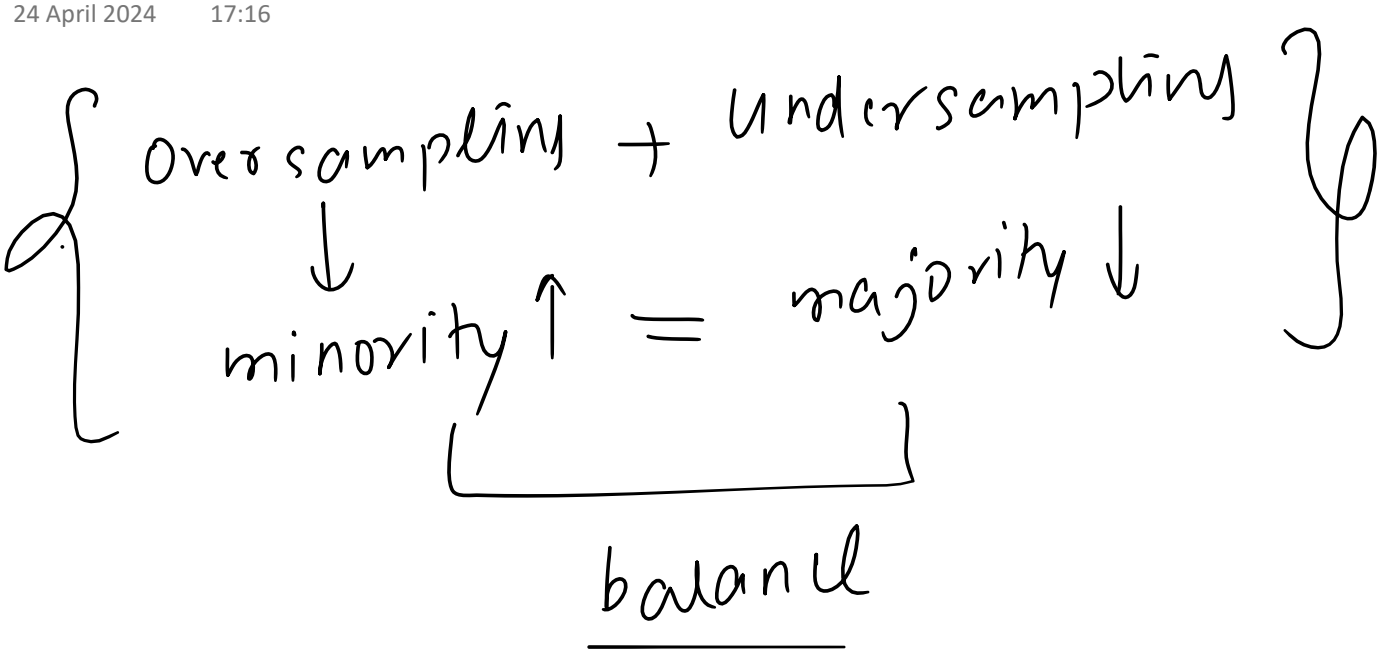
- Boundary Issues: Sometimes, SMOTE creates synthetic examples near decision boundaries, which can skew the boundary.
- Complexity: SMOTE's process is somewhat complex and can have high computational overhead.
- Class Separation: In some cases, over-creation of synthetic examples in the minority class can cause issues with class separation.




interpolation
variants

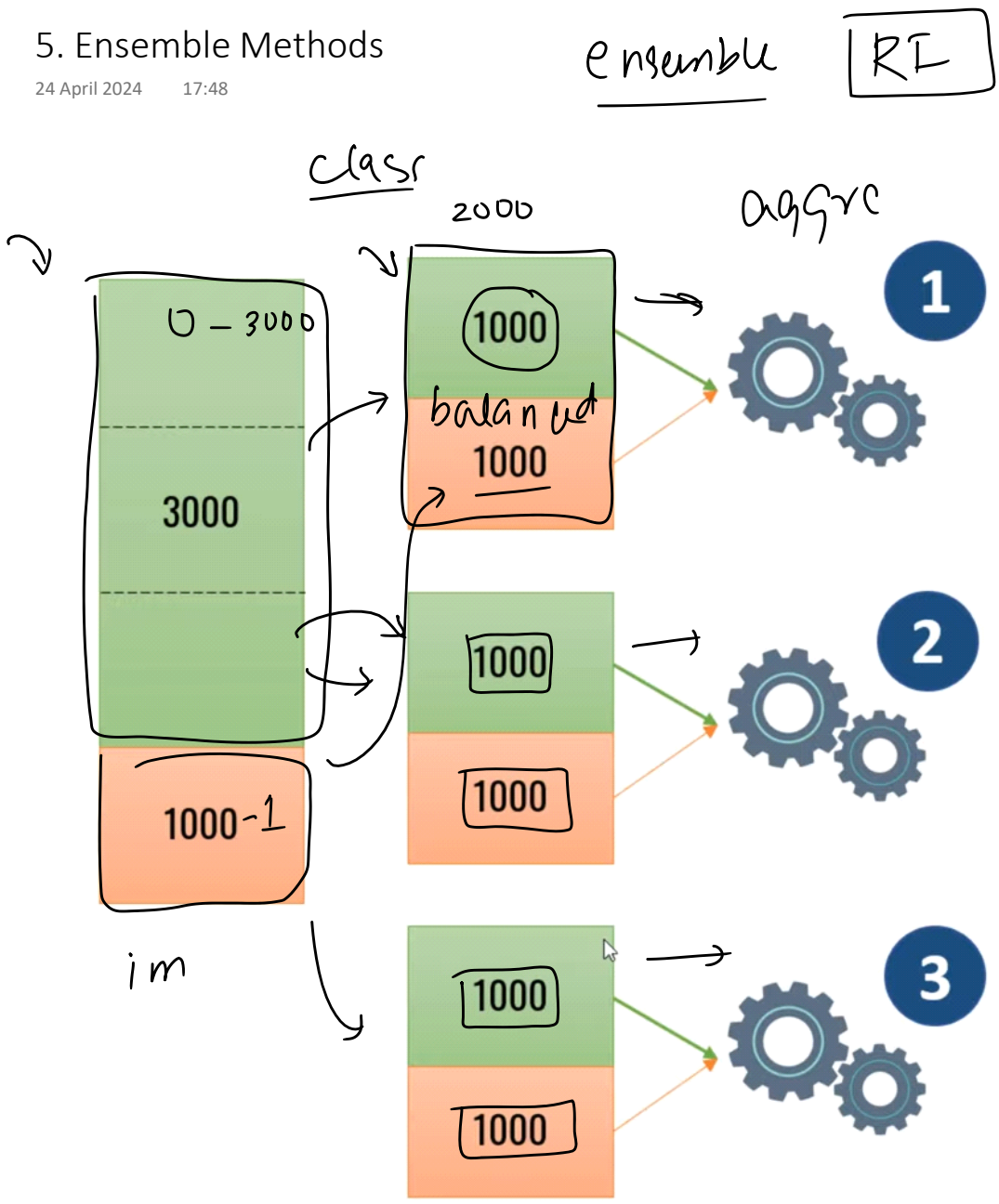
4. Hybrid Sampling

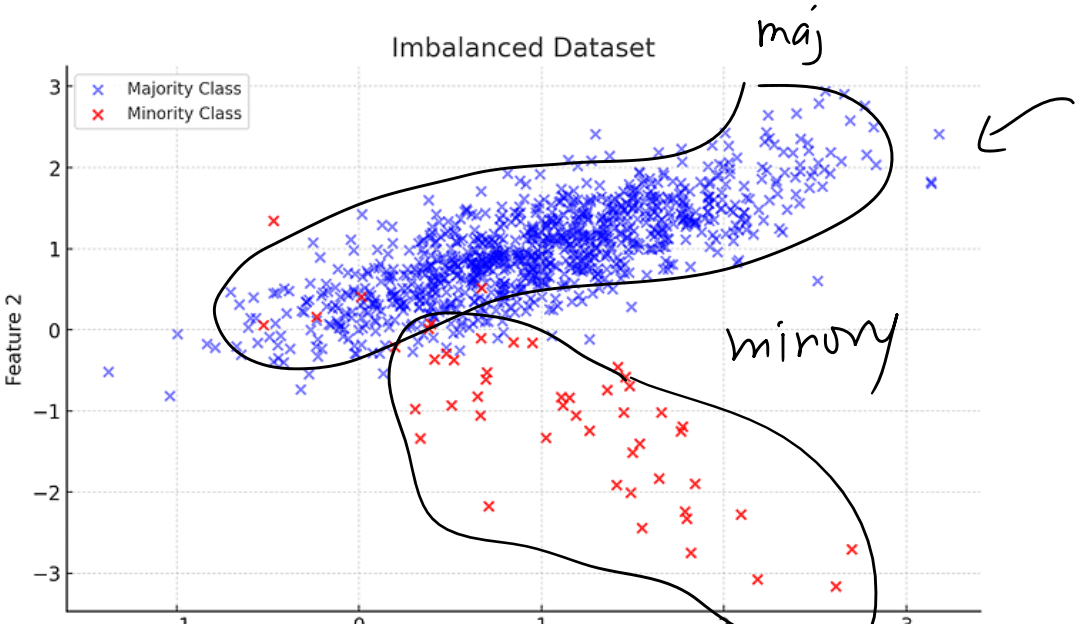
24 April 2024 17:16

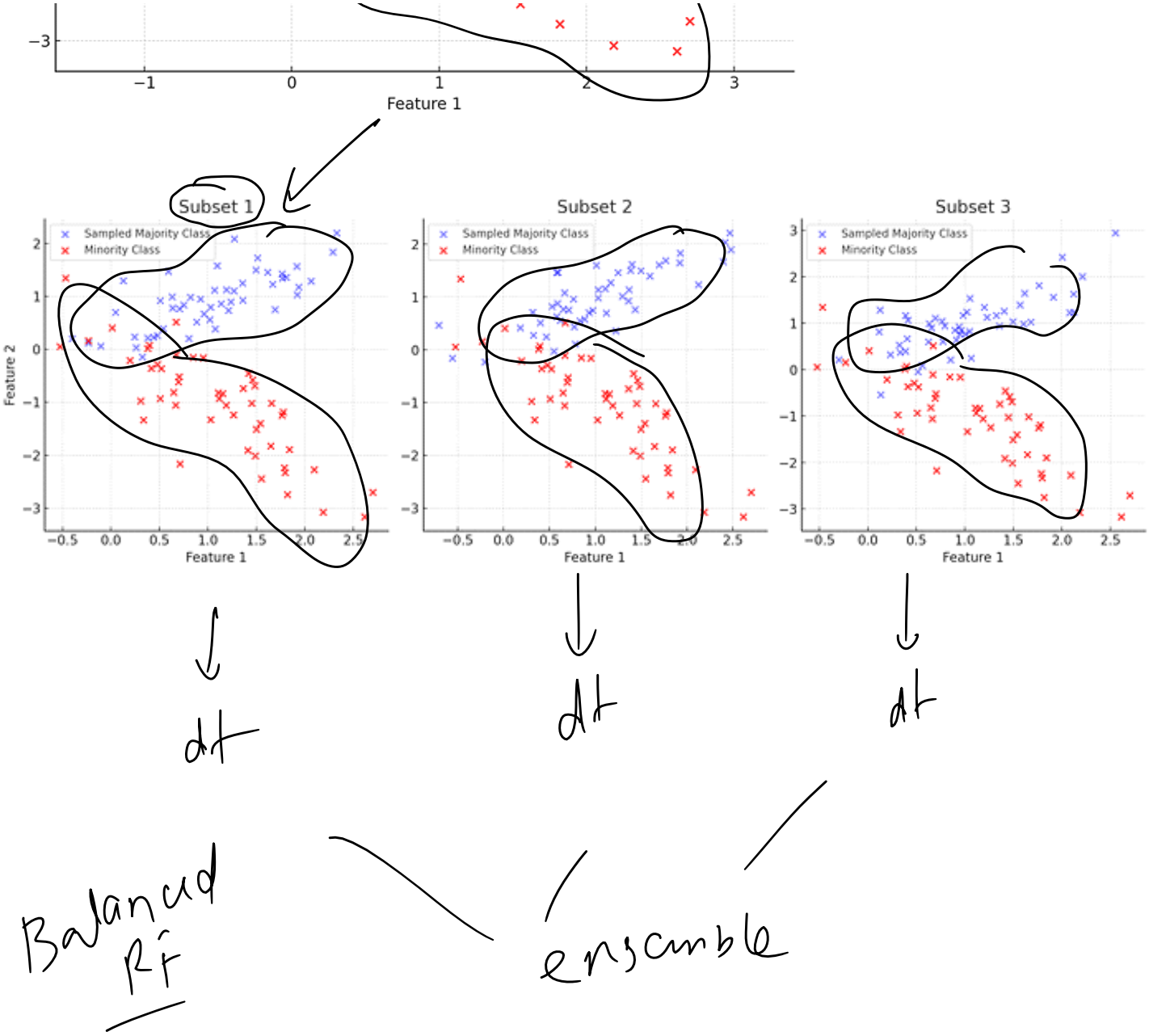


5. Ensemble Methods

24 April 2024 17:48

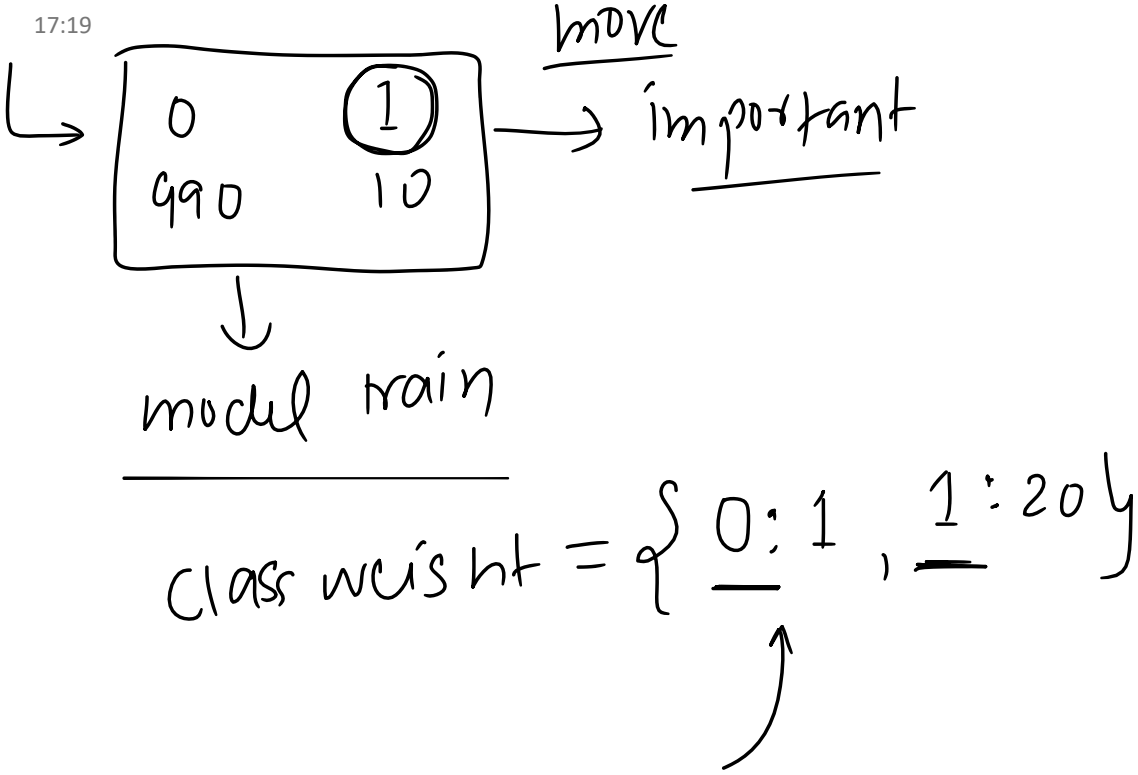




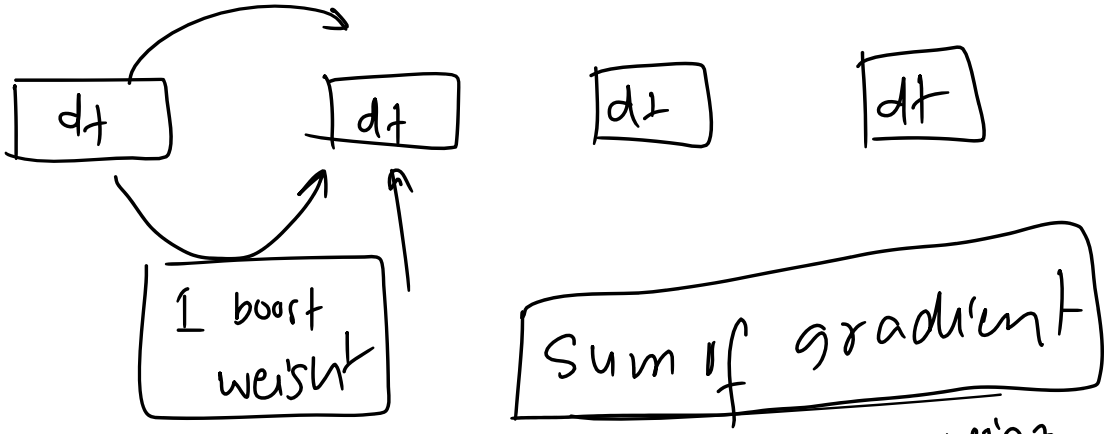

Feature 1
Subset 1
Subset 2
Subset 3
Sampled Majority Class
Minority Class
Feature 2
Feature 1
dt
dt
dt
Balanced RF
ensemble

6. Class Weighting

24 April 2024 17:19

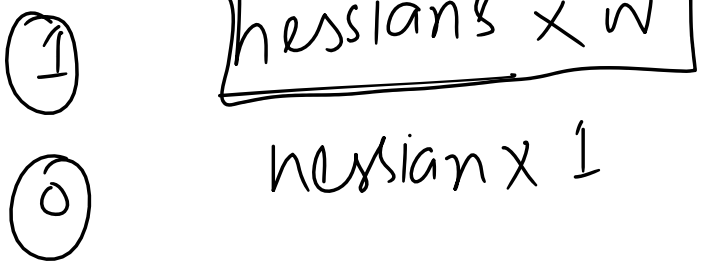


Boosting



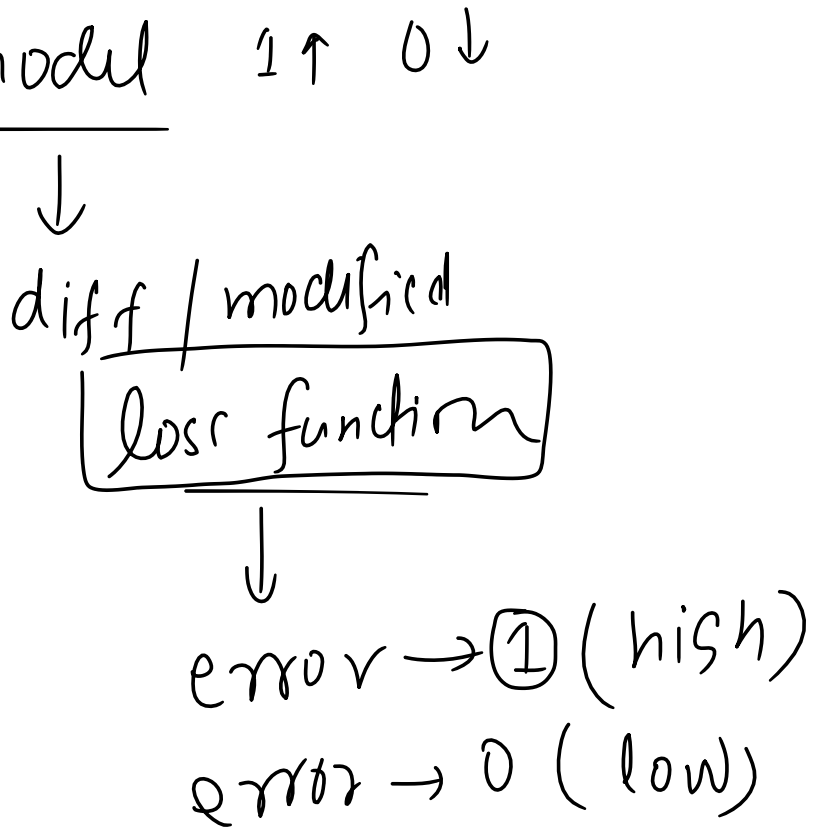
1 → gradient * w

0 → gradient * 1



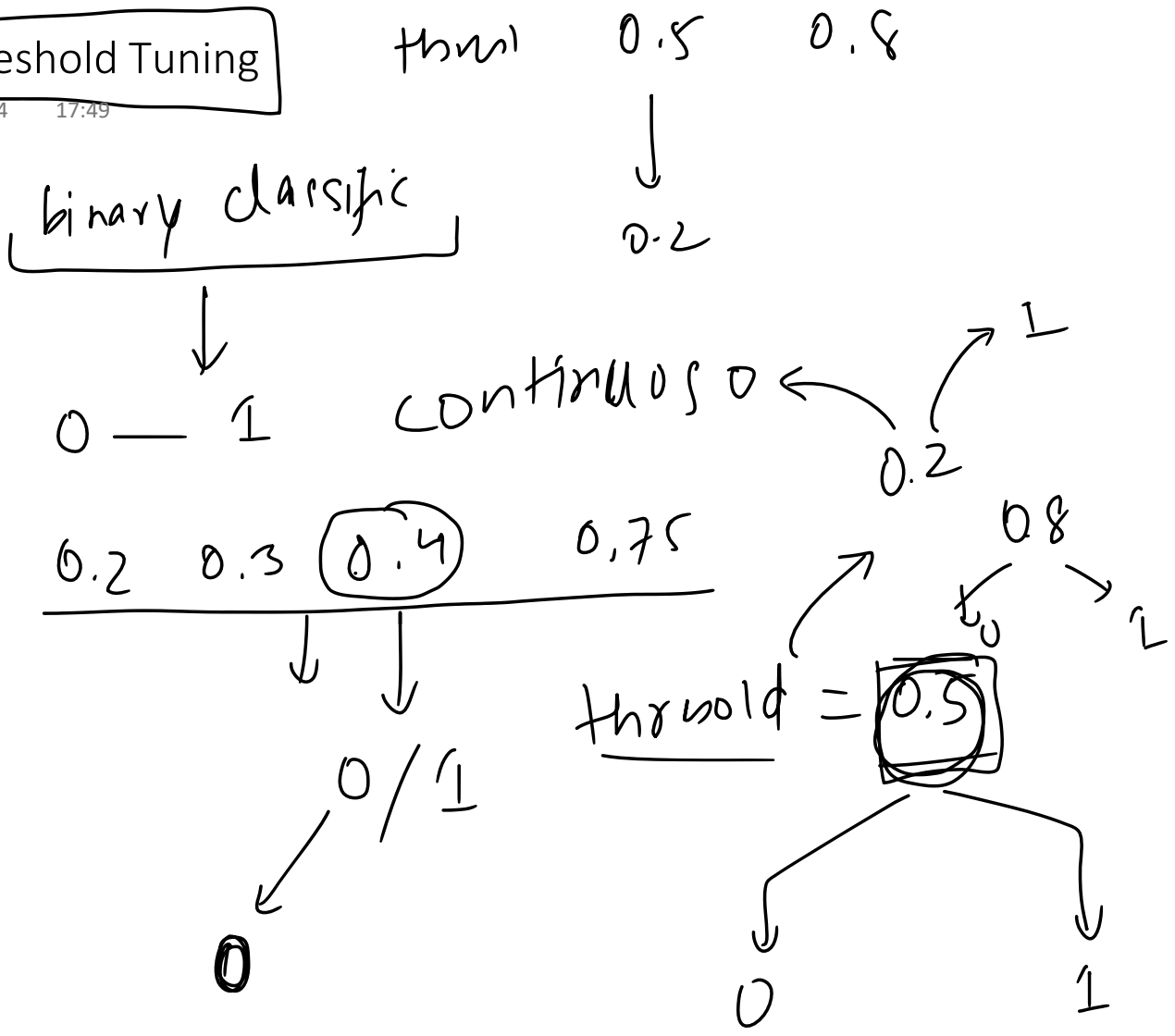
7. Cost Sensitive Learning

24 April 2024 17:48



8. Threshold Tuning

24 April 2024 17:49



List of All Techniques

24 April 2024 15:33