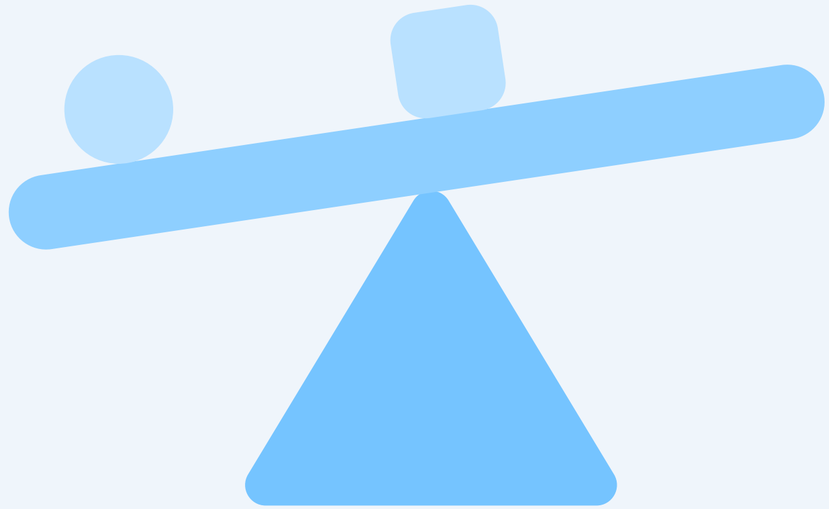
Imbalanced Data



Supervised By:
DR Doaa

Team

Abdulrahman Tarek

Mohamed Agag

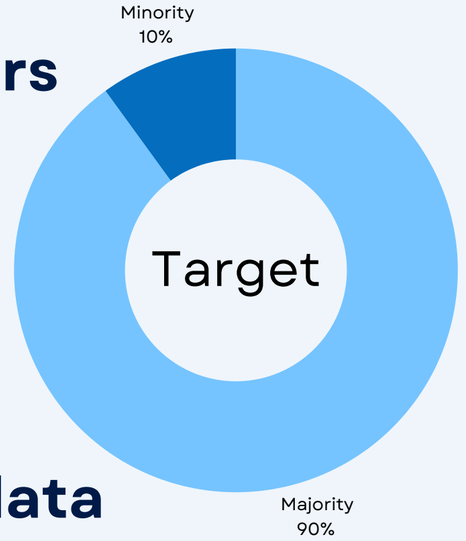
Overview

- Introduction
- Evaluation Metrics
- Techniques to Handle Imbalanced Data
- Conclusions
- References

Introduction

- **What is Imbalanced Data?**
- **Why Address It?**

A **Imbalanced data occurs when one class in a dataset significantly outnumbers others.**



B **Ignoring imbalanced data can lead to biased models that favor the majority class, missing critical minority cases.**

Evaluation Metrics

Many evaluation parameters can be measured to evaluate the classification model such as accuracy, error, precision, and Recall(sensitivity).

Thus,a confusion matrix is used to calculate values of these measurements.

confusion matrix

		Predicted: Positive	Predicted: Negative
Actual	Positive	TP	FN
Actual	Negative	FP	TN

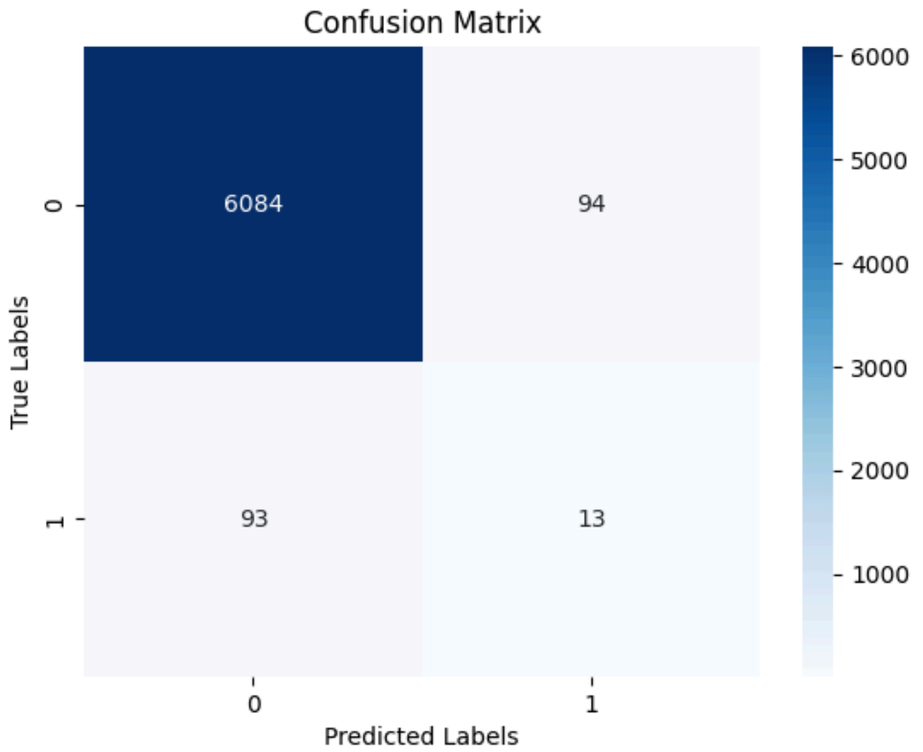
Evaluation Parameters

	Formula	Intuitive Meaning
Precision	TP / (TP + FP)	The percentage of positive predictions those are correct.
Recall	TP / (TP + FN)	The percentage of positive labeled instances that were predicted as positive.
Accuracy	(TP + TN) / (TP + TN + FP + FN)	The percentage of predictions those are correct.
Error	1-Accuracy	The percentage of predictions those are incorrect.

confusion matrix for our data

Target (Fraud)
Yes : 1(fraud)
No: 0 (Not_fraud)

- **recall = 0.12**
- **precision = 0.12**

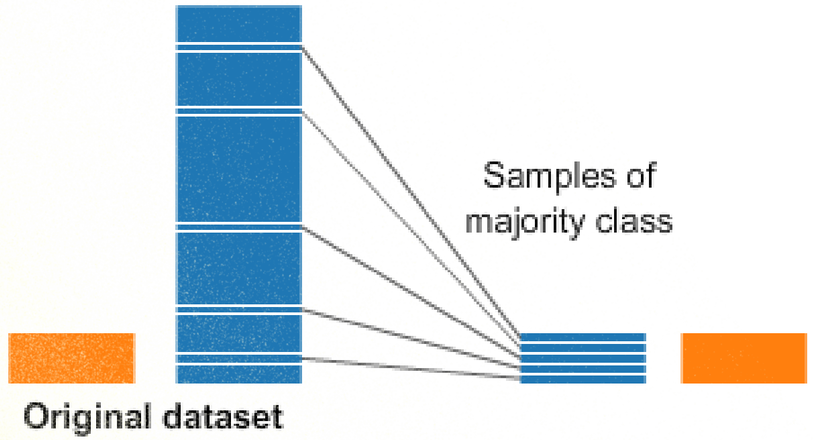


Techniques to Handle Imbalanced Data

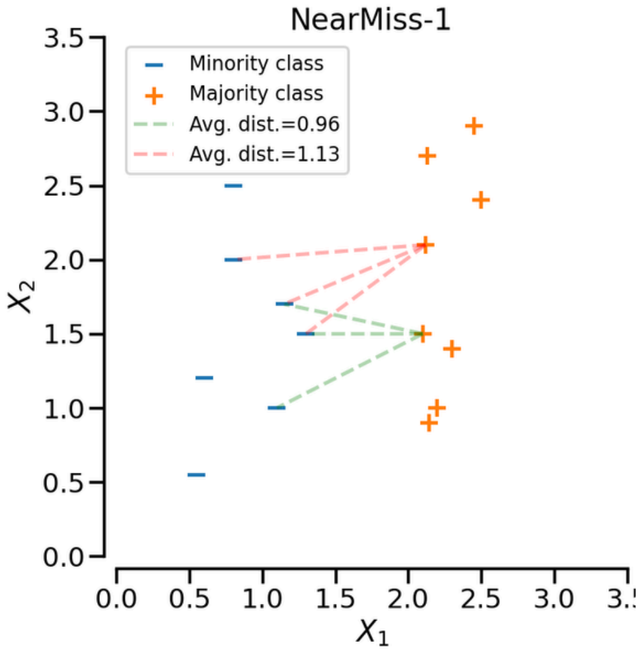
- Undersample
- Oversample
- Compination (smotetomek)
- Weighted class
- Ensemple
- Data Augmentation

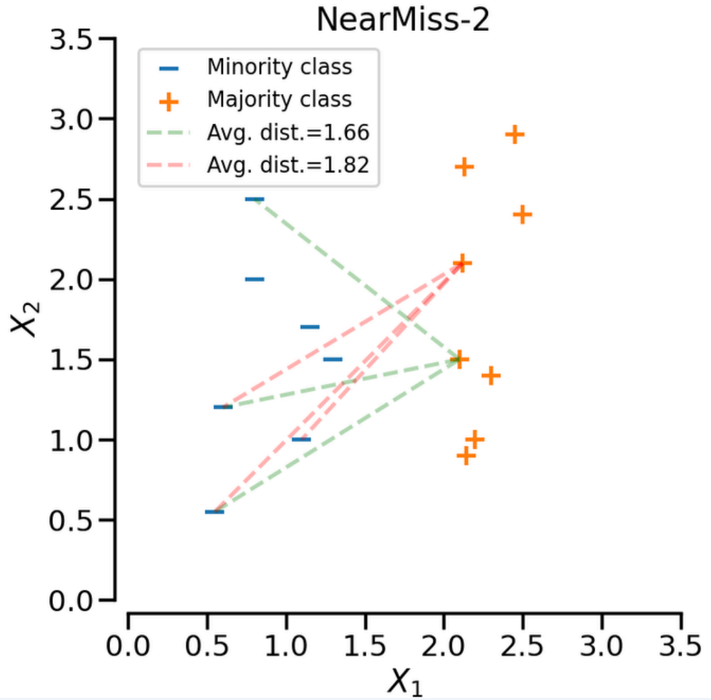
1-Random Undersampling

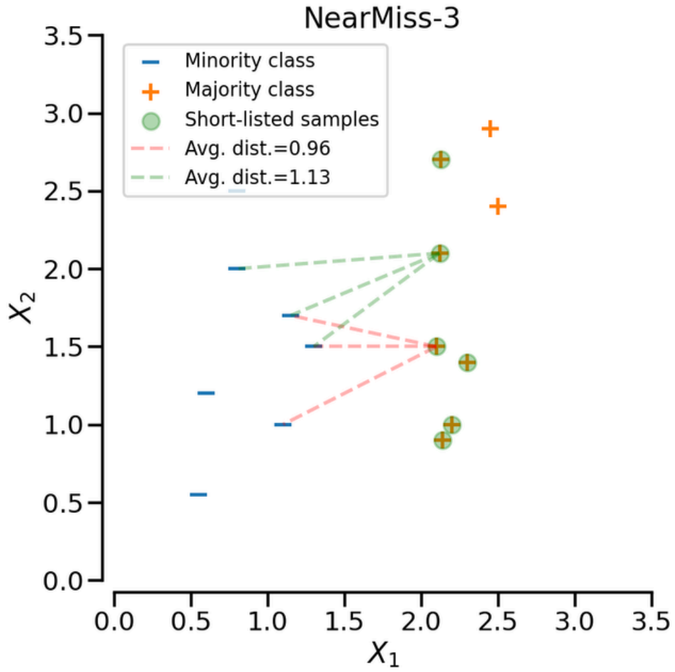
Undersampling cuts down the majority class samples to balance the classes, improving predictions for the minority class.



2-Near Miss





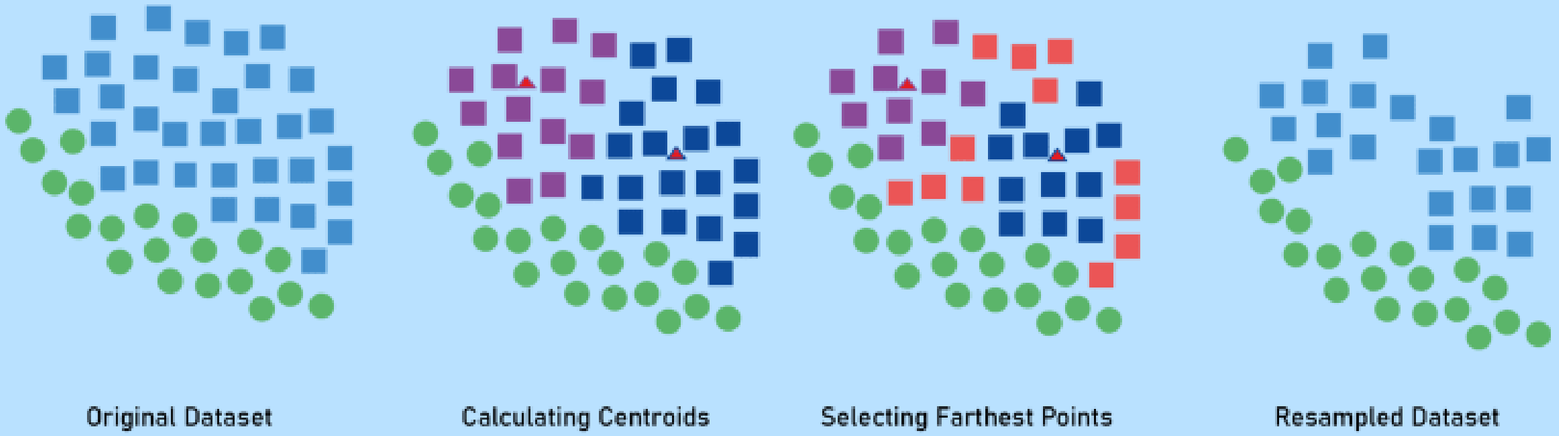


Majority class examples with minimum average distance to three closest minority class examples.

Majority class examples with minimum average distance to three furthest minority class examples.

Majority class examples with minimum distance to each minority class examples.

3-Cluster Centroid



This method involves finding centroids of the majority class and using them to balance the dataset.

Pros VS Cons

Advantages

- Reduced Complexity
- Prevents Overfitting
- Simpler Models

Disadvantages

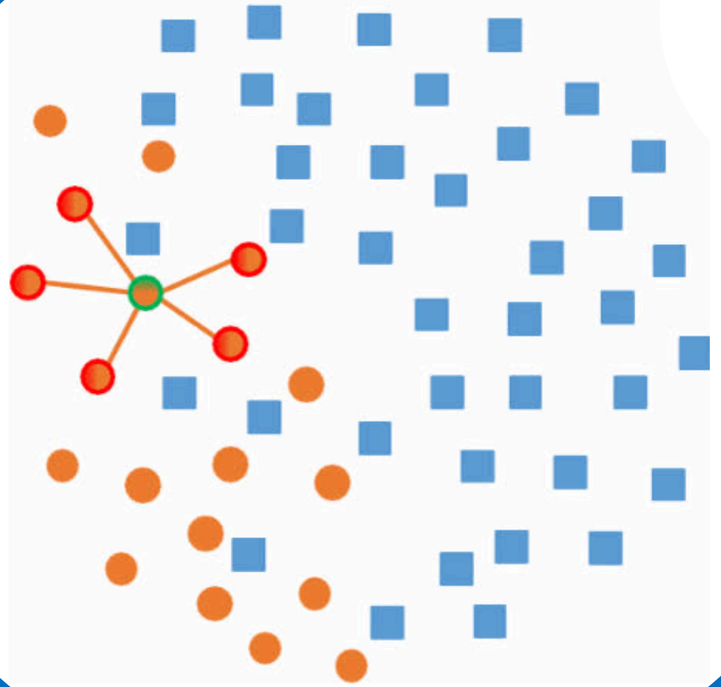
- Loss of Information
- Risk of Bias
- Potential for Instability

4-Oversamling

Random Oversampling:
This involves randomly duplicating samples from the minority class to balance the dataset.

5-SMOTE

- Identify a minority class point.
- Find its K nearest neighbors.
- Generate synthetic data between the point and its neighbors.
- Repeat steps 1 & 3



Pros VS Cons

Advantages

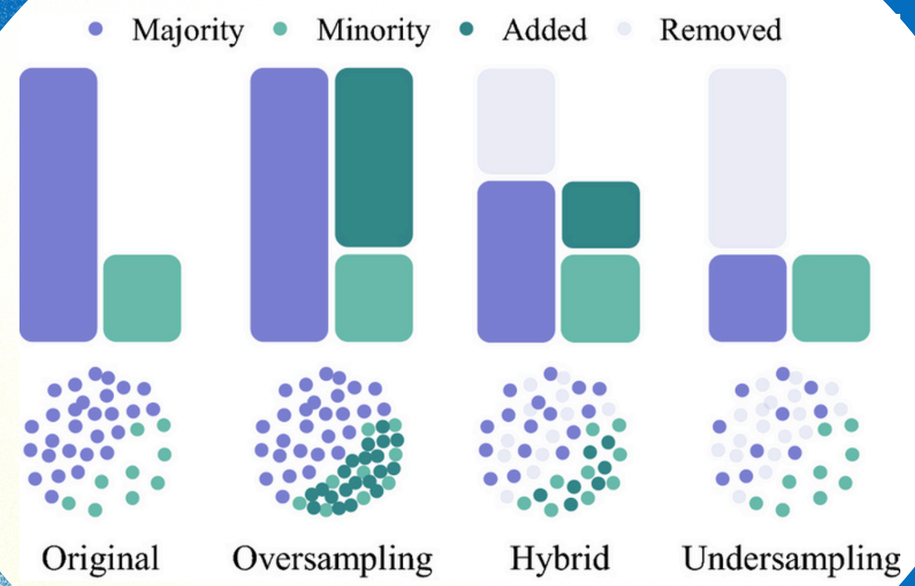
- Enhanced Model Performance
- More Robust Models
- Reduced Risk of Information Loss

Disadvantages

- Increased Complexity
- Potential Overfitting
- Algorithm Sensitivity

6-compination(smotetomek)

combination of SMOTE and Tomek Links, This combination helps by oversampling the minority class using SMOTE and then cleaning the resulting dataset using Tomek Links,



Advantages of smotetomek

SMOTETomek ensures that the model is trained on a balanced dataset, which typically leads to better generalization to new data.

7-Weighted class

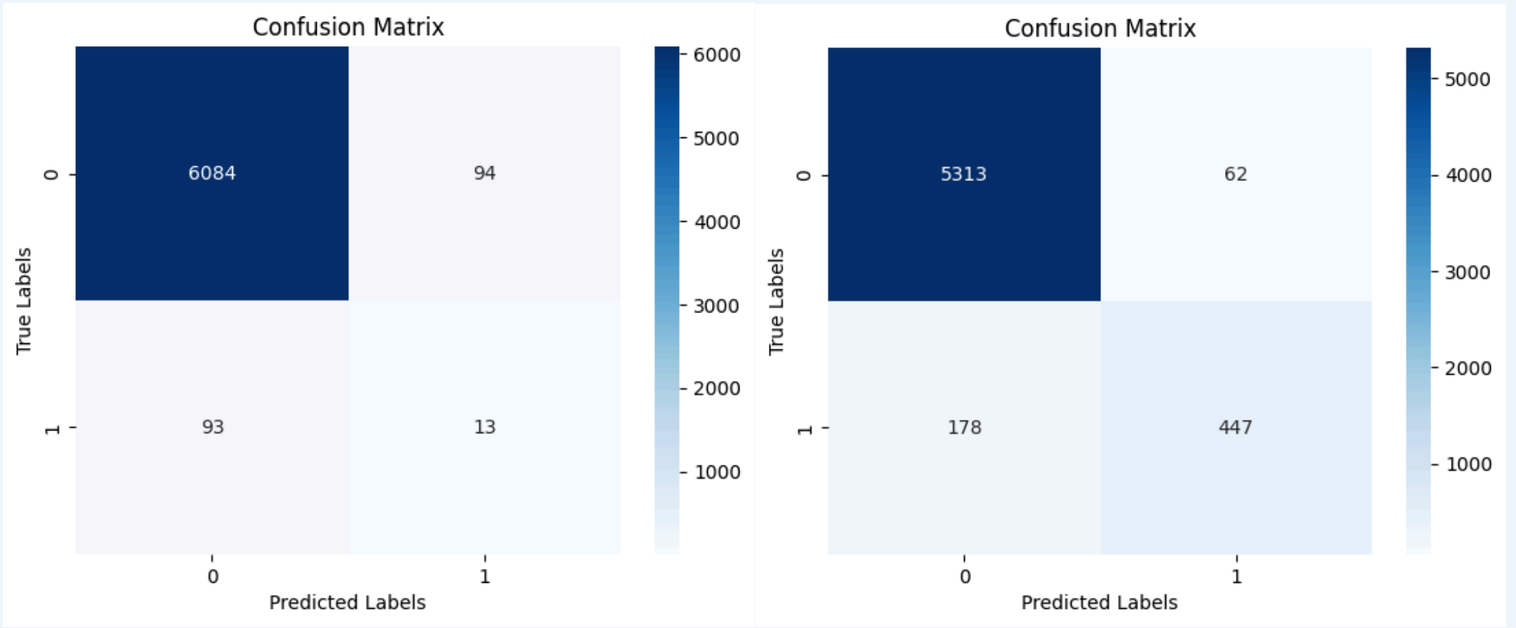
The idea is to make the model more sensitive to the minority class by increasing the cost of misclassification of that class.

$$\text{weight\_for\_class\_i} = \frac{\text{total\_samples}}{(\text{num\_samples\_in\_class\_i} * \text{num\_classes})}$$

$$\ell(x, y) = L = \{l_1, \dots, l_N\}^\top, \quad l_n = -w_n \left[y_n \cdot \log \sigma(x_n) + (1 - y_n) \cdot \log(1 - \sigma(x_n)) \right]$$

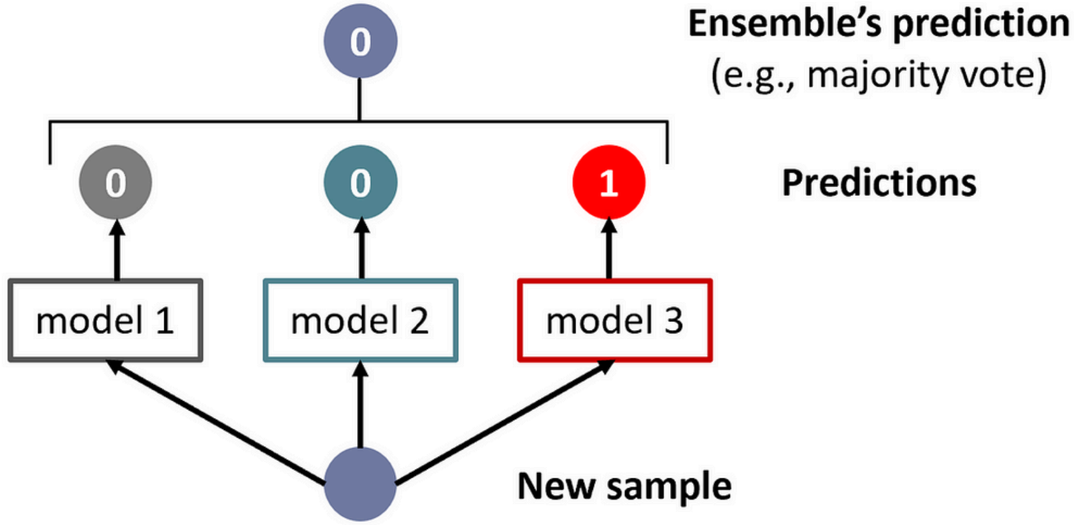
The result on our data

- accuracy = 0.96
- precision = 0.88
- recall = 0.72



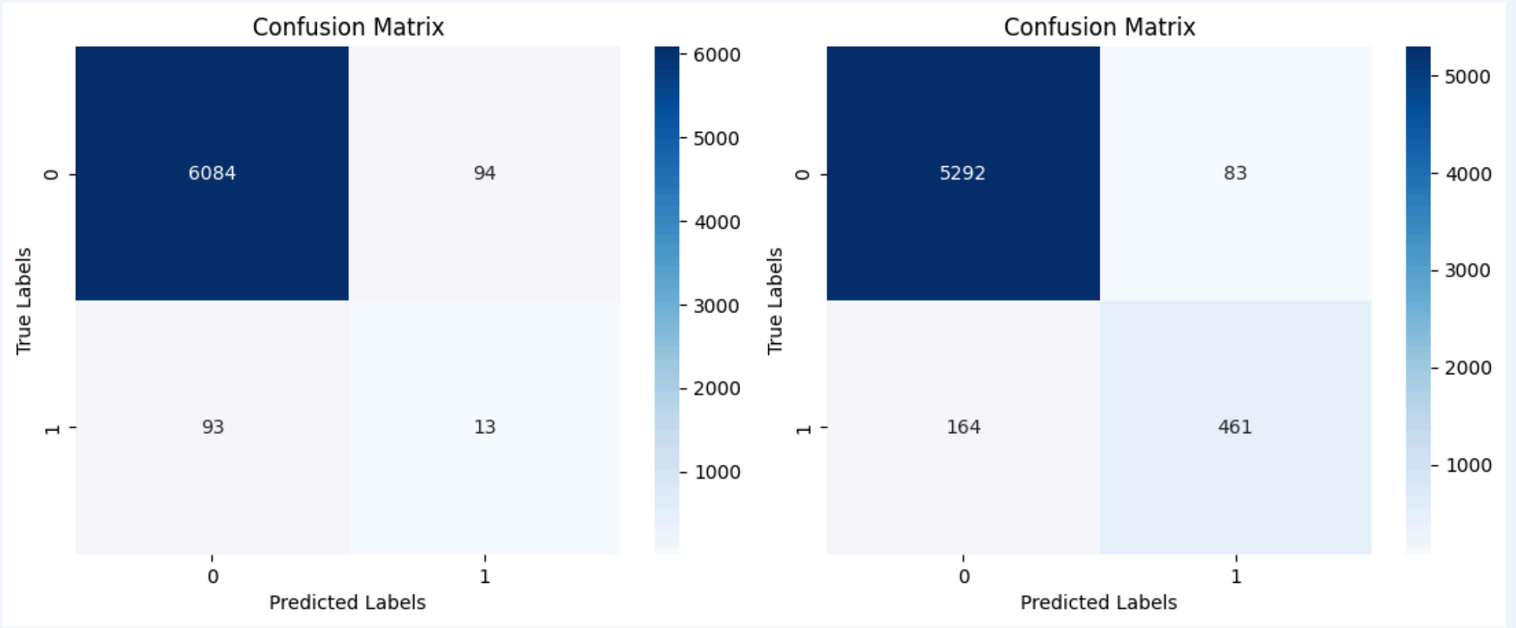
8-Ensemble

The idea behind ensemble learning is that by combining multiple models, each with its strengths and weaknesses, the ensemble can achieve better results than any single model alone



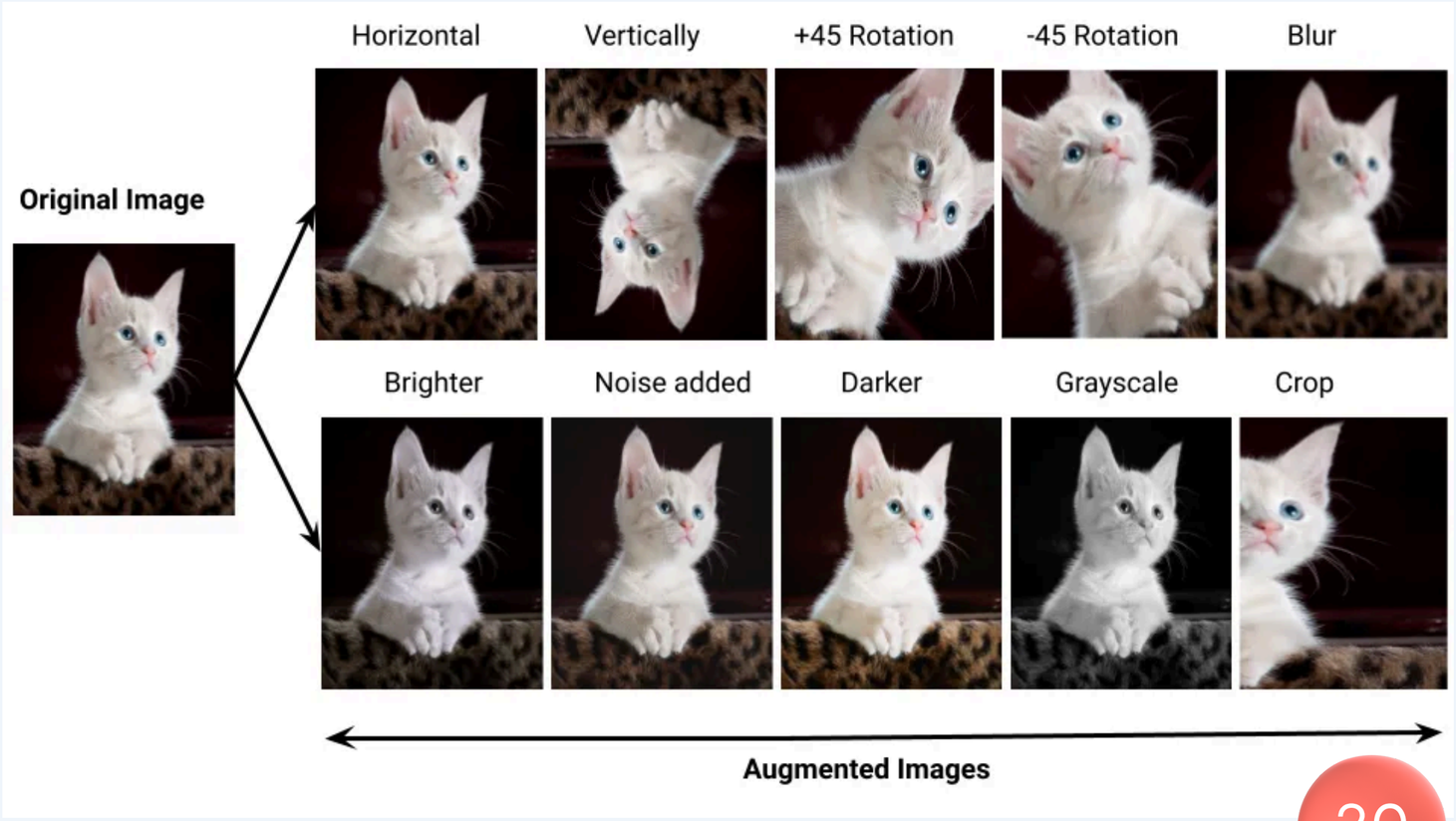
The result on our data

- accuracy = 0.96
- precision = 0.85
- recall = 0.74



9-Image Data Augmentation

- Flipping
- Rotation
- Scaling
- Translation
- Color Jittering
- Adding Noise
- Random Cropping



Conclusions

- **No Single Solution Fits All: The effectiveness of each technique depends on the specific characteristics of your dataset and the problem at hand.**
- **Combination of Techniques: Often, the best results are achieved by combining multiple techniques, such as using SMOTETomek with weighted classes or ensemble methods.**

References

kaggle

Thank you!