



Applied Malaria Modeling Network (AMMNet)



Presents

Training on Malaria Modeling in R & RStudio



Data Visualization & Malaria Modeling Techniques in R & RStudio



Dennis K. Muriithi, Ph.D

Wednesday, August 28,
2024

2



"Every 2 minutes, a child under 5 dies of malaria" - UNICEF

MALARIA

There were an estimated
241 million
cases worldwide in 2020



An estimated
627,000
of malaria deaths
in 2020

The African region was home
to 95% of malaria cases and
96% of malaria deaths

Source: World Malaria report 2021

PREMIUM
Times

Infographics by Nike Ade



Wednesday, August 28,
2024

“80% of the deaths among children under five years were found to be malaria related”-WHO 2022

MALARIA

There were an estimated
249 million
cases worldwide in 2022

An estimated
609,000
Of malaria deaths in 2022

Approximately
95%
of death re malaria Related

Source: WHO 2022 Report

Author: D K. Muriithi
Chuka University

2024

DATA VISUALIZATION: Turning Data into Stories

- ❖ This is the art and science of communicating information clearly and effectively through visual representations.
- ❖ It involves transforming raw data into meaningful and understandable graphics, charts, and maps.
- ❖ Why is it important?
 - ✓ **Clarity:** Complex data becomes easy to understand
 - ✓ **Insights:** Patterns, trends, and outliers become visible
 - ✓ **Storytelling:** Data-driven narratives can be created
 - ✓ **Decision Making:** Informed choices can be made based on visual evidence
- ❖ Focus will be on the 20% that is useful 80% of the time



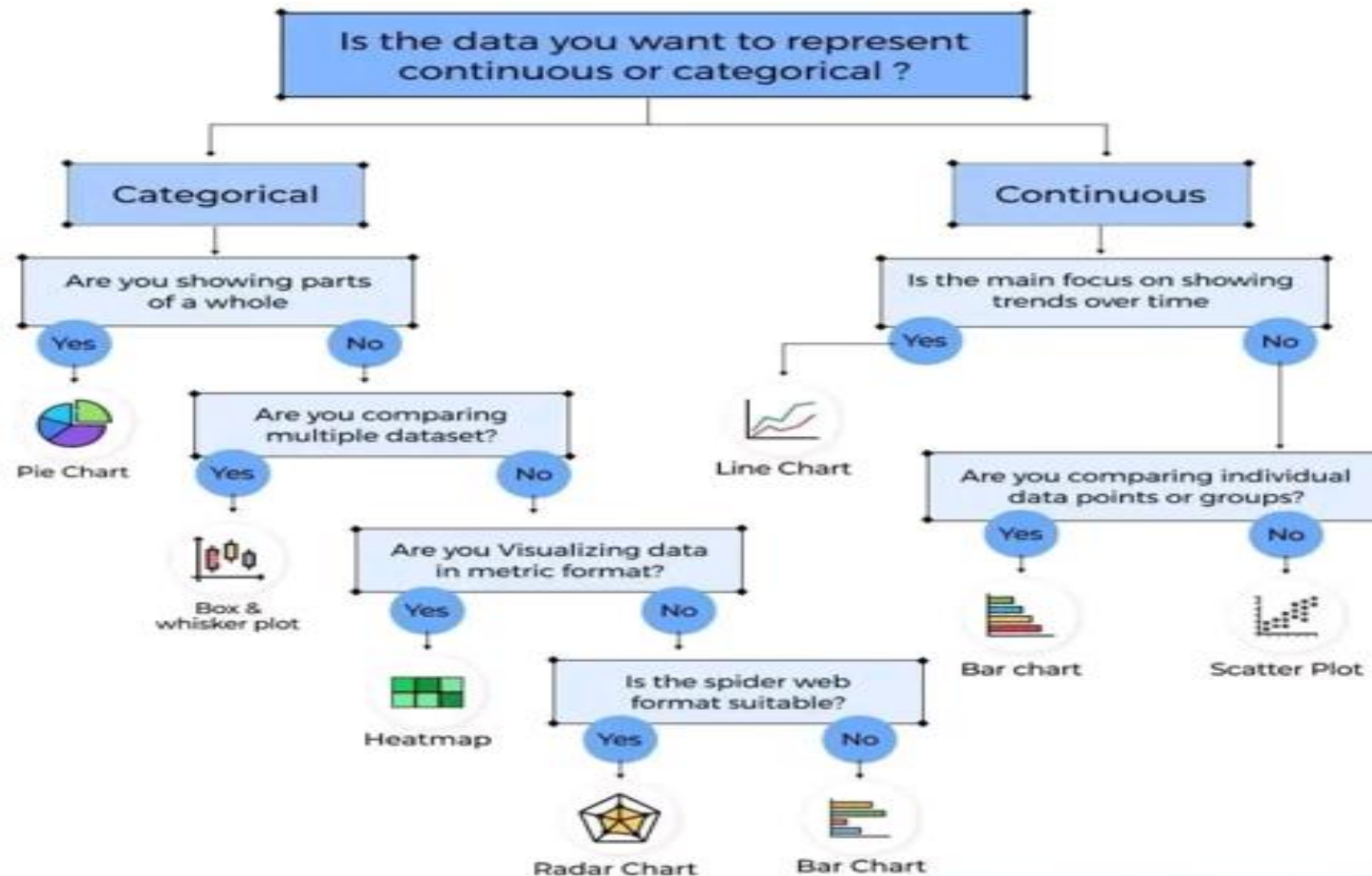
Types of Data Visualization

- **Charts:** Bar charts, line charts, pie charts, area charts, etc
- **Graphs:** Scatter plots, bubble charts, histograms, box plots, etc
- **Maps:** Geographic data visualization
- **Infographics:** Combining visuals with text for storytelling
- **Dashboards:** Interactive displays of multiple visualizations.





How to choose a Right Graph for Data Visualization





DATA VISUALIZATION



In R, there are three main plotting systems:

- ❖ Base graphics
- ❖ ggplot2
- ❖ Leaflet & tmap

Prerequisites

- Install R & RStudio
- Install ggplot2, leaflet & tmap package on your R environment
- The repository on Github has files for the source, data and other important materials.

URL: <https://github.com/CUDataanalytics/CUAMMnet>



Get Started

- R is case sensitive
- Comment in your codes: start with #
- Get help:
 - ☐ ?
 - ☐ help()





	Class	Example
1	Integer	3L, as.integer(3)
2	Numeric	3, 3.0, π
3	Character	"a", "b", "UMD"
4	Logical	TRUE, FALSE
5	Complex	$1 + 4i$

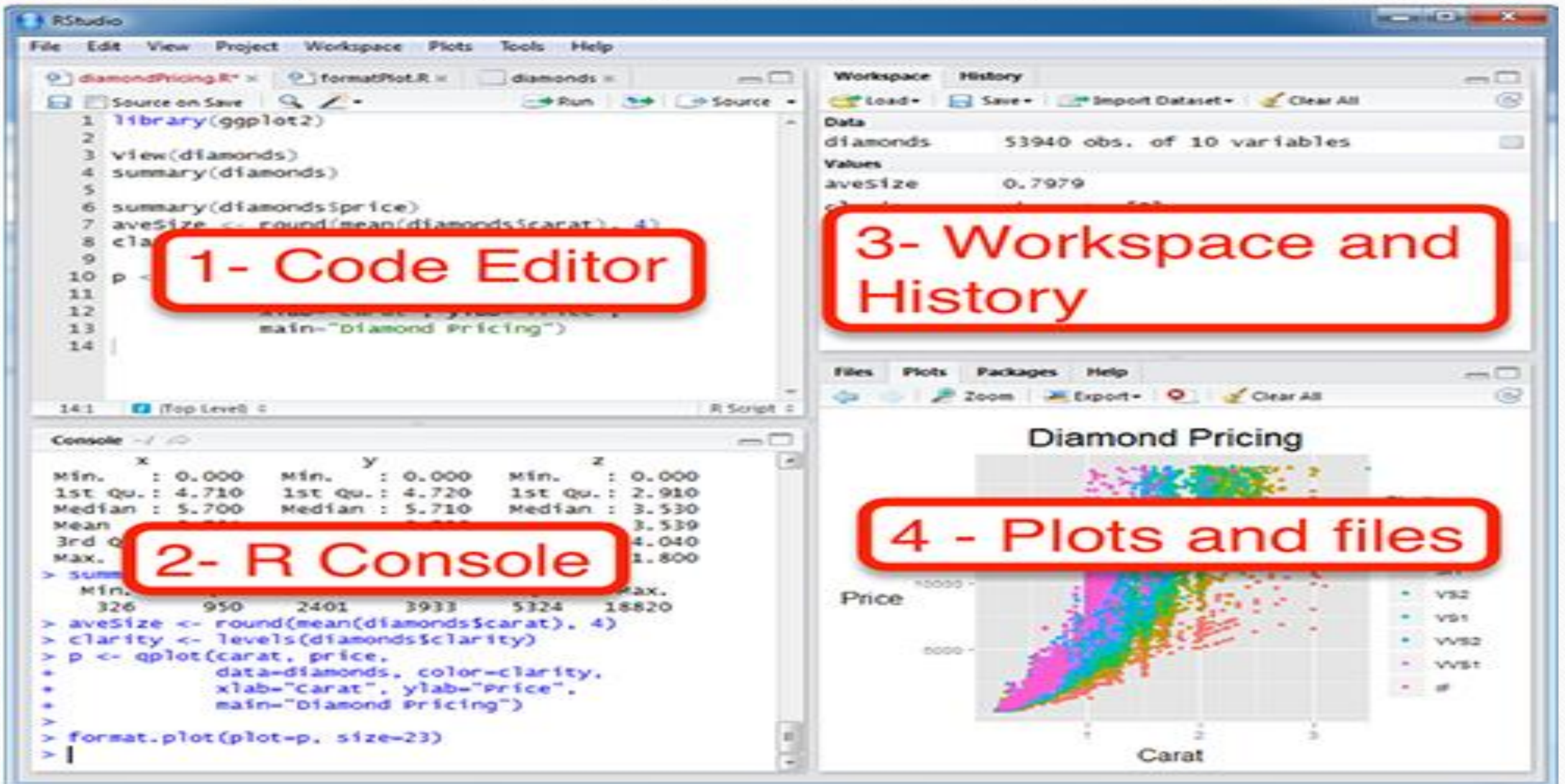
■ Get the class of a R object

- ❑ class()
- ❑ is.numeric()
- ❑ is.character()
- ❑ is.logical()

■ Change the class of a R object

- ❑ as.numeric()
- ❑ as.logical()
- ❑ as.character()







Malaria Modeling using Machine Learning Algorithms

Machine Learning (ML), sometimes referred to as **Statistical Learning**, is a subfield of artificial intelligence (AI) that focuses on the development of algorithms capable of learning and making predictions or decisions based on inputs and data.

- ❖ The process of developing a machine learning model begins with the collection & preprocessing of data
- ❖ Data preprocessing includes data cleaning, transformation, organization, imputation & labeling





The relationship between AI, Machine Learning, and Deep Learning is summarized below

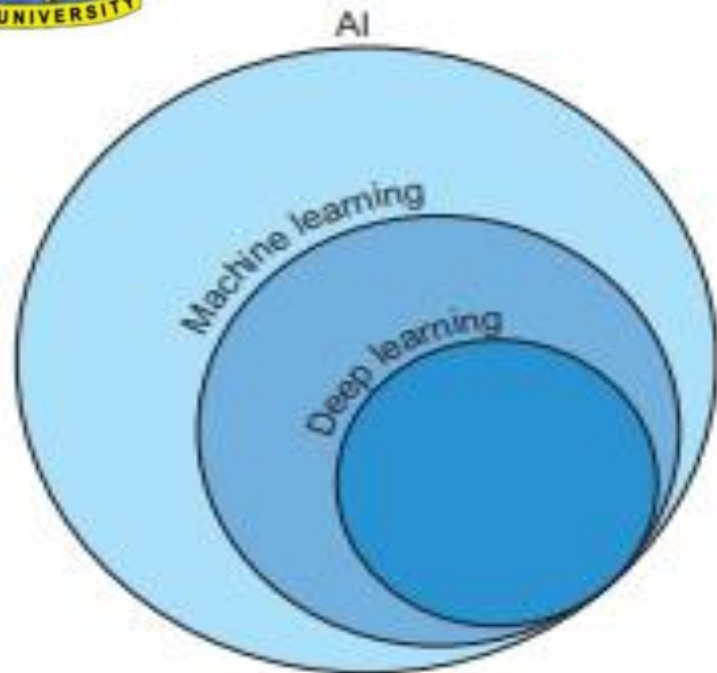


Figure 1.7 The relationship between artificial intelligence (AI), machine learning, and deep learning. Deep learning comprises a collection of techniques that form a subset of machine learning techniques, which themselves are a subfield of AI.

Credit: Hefin I. Rhys(2020) All rights reserved



Types of Machine Learning

All machine learning algorithms can be categorized by their learning type and the task they perform. There are three learning types:

❖ **Supervised learning** where models learn from labeled data.....

- ☐ Classification

- ☐ Regression

❖ **Unsupervised learning** discovers hidden patterns within unlabeled data, useful in...

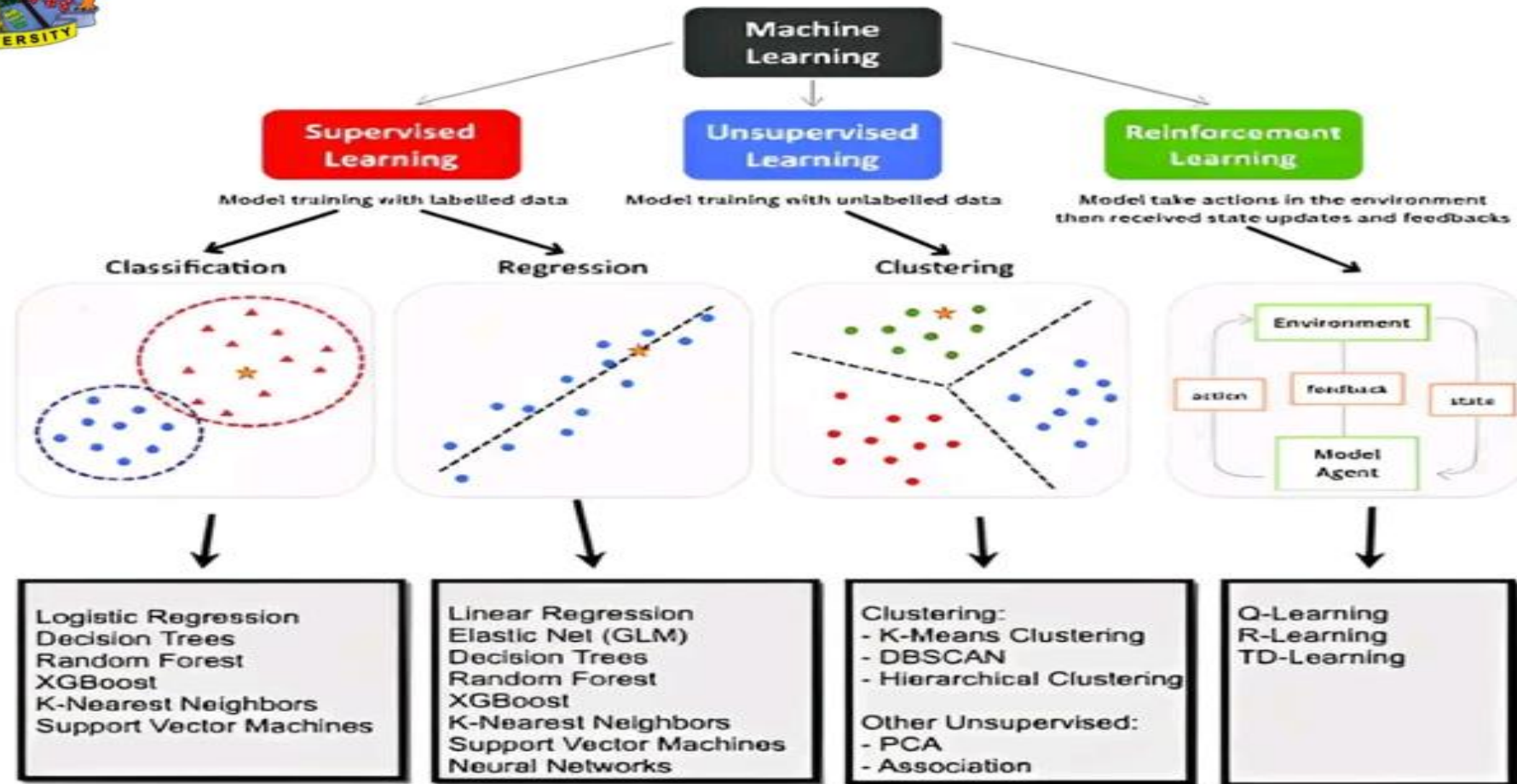
- ☐ Dimension reduction

- ☐ Clustering

❖ **Reinforcement learning** teaches agents to make decisions by trial and error, valuable in autonomous systems



3 Types of Machine Learning (Every Data Scientist Should Know)



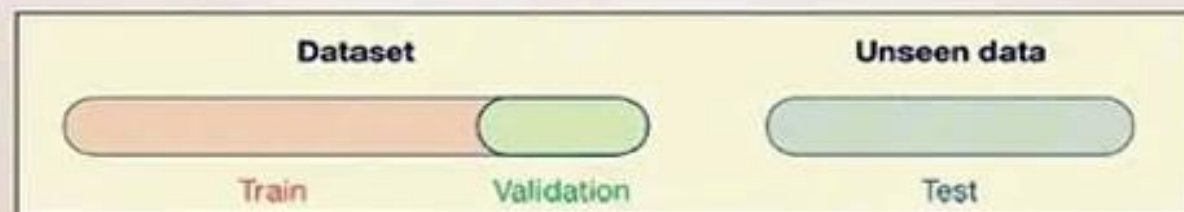


Model Training and Testing

When selecting a model, we distinguish 3 different parts of the data that we have as follows:

Training set	Validation set	Testing set
<ul style="list-style-type: none">- Model is trained- Usually 80% of the dataset	<ul style="list-style-type: none">- Model is assessed- Usually 20% of the dataset- Also called hold-out or development set	<ul style="list-style-type: none">- Model gives predictions- Unseen data

Once the model has been chosen, it is trained on the entire dataset and tested on the unseen test set. These are represented in the figure below:





Cross Validation (CV)

- ❑ CV is a powerful statistical technique used to assess the performance and generalization ability of machine learning models.
- ❑ It helps to ensure that the model generalizes well to unseen data by testing it on different subsets of the data.
- ❑ The primary goal of cross-validation is to prevent overfitting, ensuring that the model performs well on unseen data.

Common cross-validation approaches:

- ❖ Holdout cross-validation
- ❖ K-fold cross-validation
- ❖ Leave-one-out cross-validation



Holdout Cross-Validation

Holdout CV

Training set

Test set

1. The data is randomly split into a training and test set.
2. A model is trained using only the training set.
3. Predictions are made on the test set.
4. The predictions are compared to the true values.



K-fold Cross-Validation



K-fold CV

Fold 1		Training set		Test set
Fold 2			Test set	
Fold 3			Test set	
Fold 4		Test set		
Fold 5	Test set			

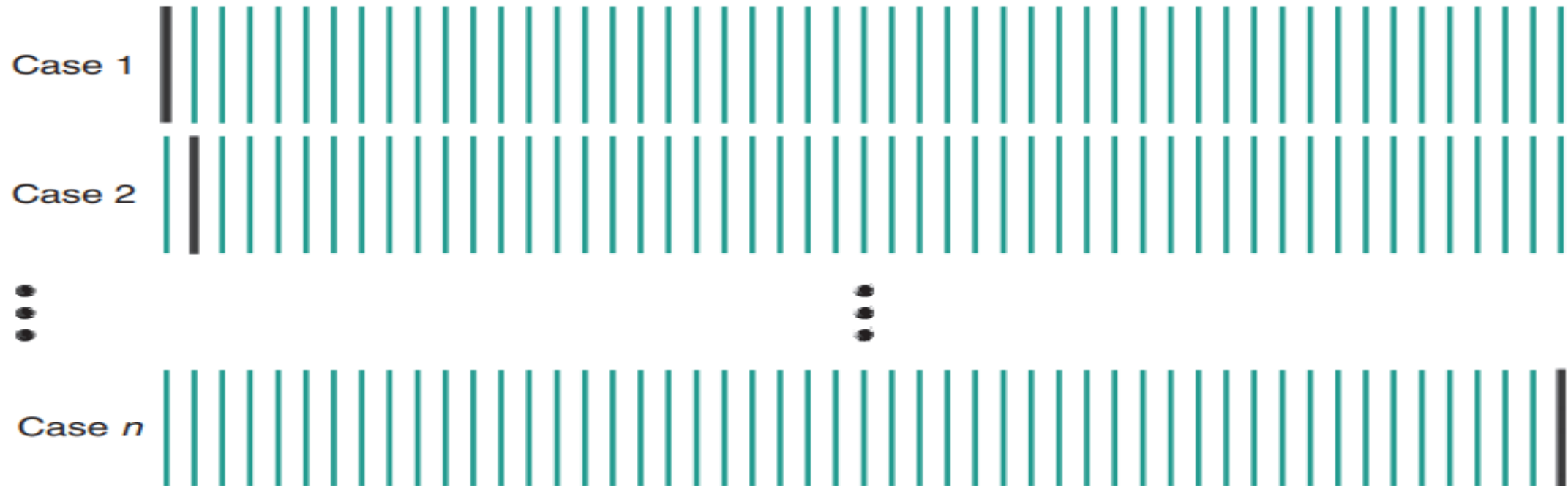
1. The data is randomly split into k equal-sized folds.
2. Each fold is used as the test set once, where the rest of the data makes the training set.
3. For each fold, predictions are made on the test set.
4. The predictions are compared to the true values.



Leave-one-out Cross-Validation



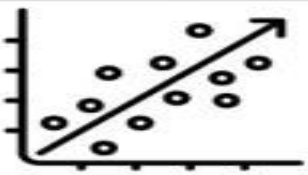
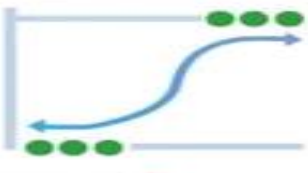



Leave-one-out CV



1. Use all of the data except a single case as the training set.
2. Predict the value of the single test case.
3. Repeat until every case has been the test case.
4. The predictions for each case are compared to the true values.



Hyperparameter Tuning in Machine Learning

Representation	Algorithm Name	Hyperparameter
	Linear Regression	Regularization parameter (alpha for Ridge/ Lasso Regression)
	Logistic Regression	C (Inverse of regularization strength), penalty (L1, L2)
	Decision Tree	Max_depth, min_samples_splits, min_samples_leaf, criterion
	K- Nearest Neighbors	n_neighbors, weights, metric
	Support Vector Machines	C, Kernel, gamma, degree (for polynomial kernel)



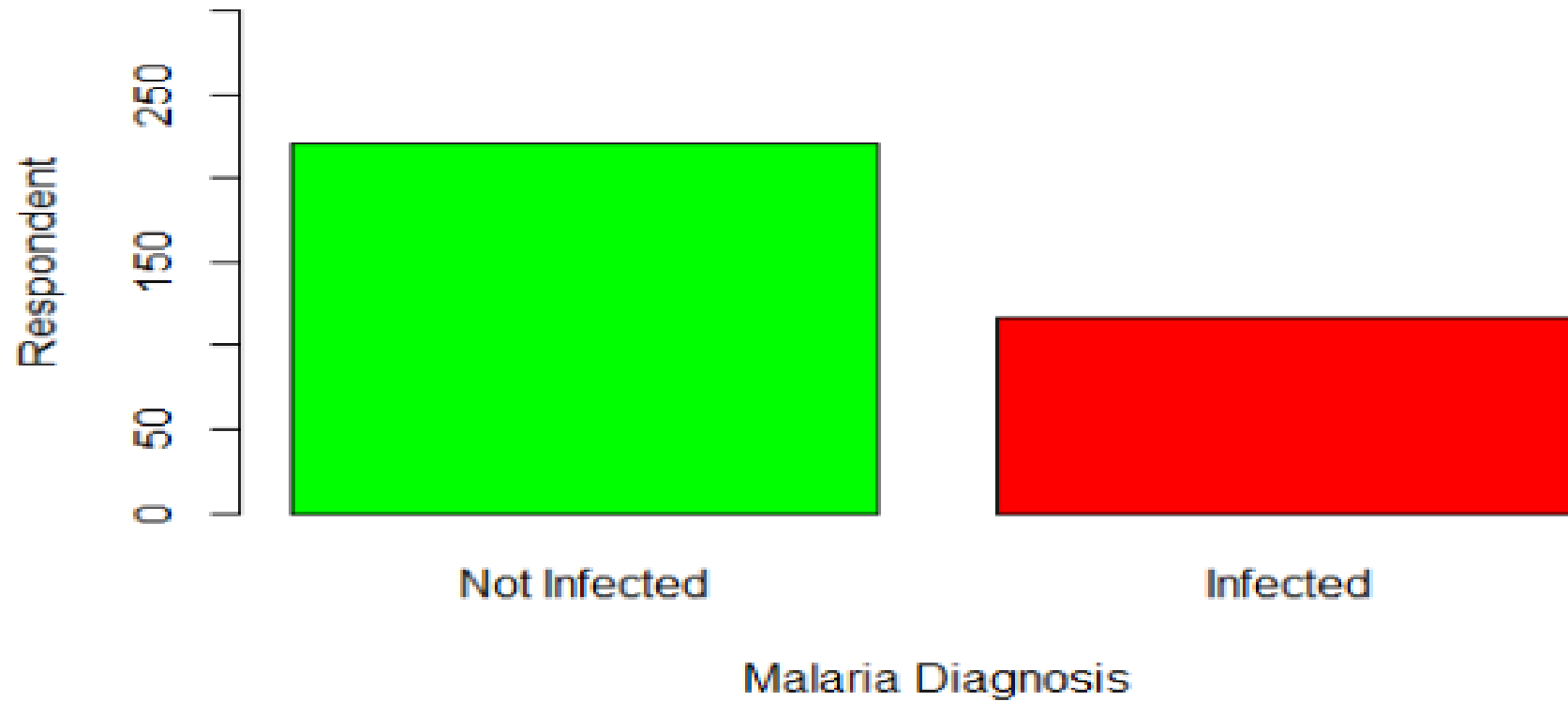


What is Imbalance Data?

- Refers to datasets where the classes are not represented equally.
- This issue can manifest in various real-world scenarios, such as fraud detection, medical diagnosis, text classification, and image recognition.
- Data imbalance is a common and critical challenge in the field of machine learning.
- This imbalance can lead to biased models that perform well on the majority class but poorly on the minority class.
- Machine learning algorithms are designed to optimize overall accuracy, which means they tend to favor the majority class.
- As a result, the minority class is underrepresented in the model's learning process, leading to skewed predictions and poor generalization to the minority class.

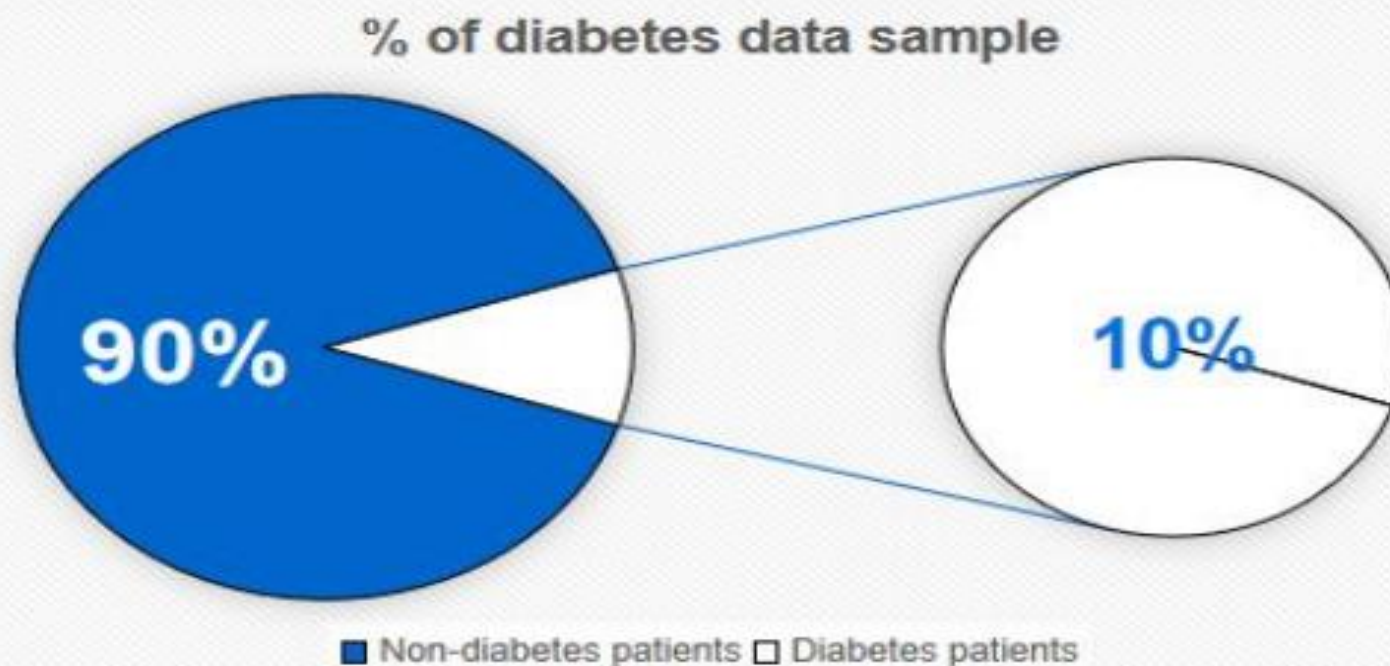


Example of Imbalance Malaria Data





Example of Imbalance Diabetes Data





Key Points About Imbalanced Data



- ❖ **Class Distribution:** One class (majority) has many more instances than the other (minority).
- ❖ **Impact on Models:** Standard machine learning algorithms may become biased toward the majority class, resulting in poor performance on the minority class.
- ❖ **Evaluation Metrics:** Accuracy is not a reliable metric for imbalanced datasets. Instead, metrics such as precision, recall, F1-score, and area under the ROC curve (AUC-ROC) are more informative.
- ❖ **Techniques to Handle Imbalance:**
 - ❑ **Resampling:** Techniques like oversampling the minority class (e.g., SMOTE) or undersampling the majority class.
 - ❑ **Algorithmic Adjustments:** Modifying algorithms to account for imbalance, such as adjusting class weights.
 - ❑ **Synthetic Data Generation:** Creating synthetic samples for the minority class.
 - ❑ **Anomaly Detection Methods:** Treating the minority class as anomalies or outliers

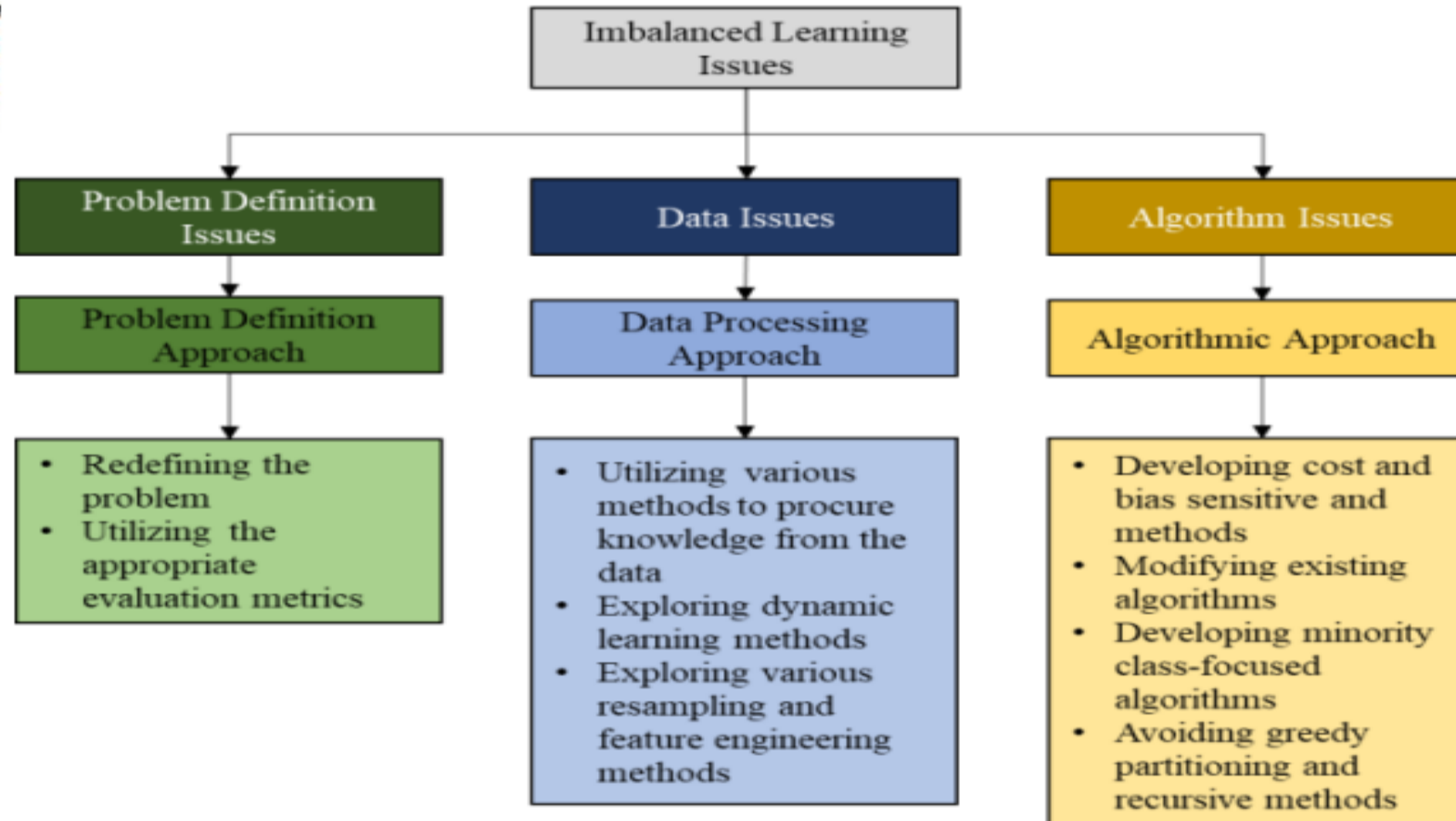


Figure 2: General Approaches in Imbalanced Learning



Consequences of Imbalance Data



- ❑ **Model Bias:** Imbalanced data can lead to model bias, where the model is influenced by the majority class. It may struggle to make accurate predictions for the minority class.
- ❑ **High Accuracy, Low Performance:** A model trained on imbalanced malaria data may appear to have high accuracy but may perform poorly on minority classes.
- ❑ **Missed Insights:** Data imbalance can result in the loss of important insights and patterns present in the minority class, leading to missed opportunities or critical errors



- ❖ In practical terms, models trained on imbalanced datasets may show a high accuracy rate, but they are ineffective at identifying and correctly classifying instances of the minority class.
- ❖ To address the challenges posed by data imbalance, various techniques and strategies have been developed.
- ❖ These methods aim to rebalance the dataset, adjust the model's learning process, or use specialized evaluation metrics that better reflect the performance on imbalanced data.
- ❖ The selection of the most appropriate approach depends on the specific problem, the dataset, and the desired outcome.
- ❖ During this training, we will explore different techniques for handling data imbalance in machine learning and discuss when and how to use them effectively for malaria modeling



Resampling Techniques for Handling Malaria Imbalance Data

- ❑ **Resampling techniques** are a common set of strategies used to address data imbalance in machine learning.
- ❑ These techniques involve modifying the dataset by either increasing the number of minority class samples (over-sampling) or reducing the number of majority class samples (under-sampling).
- ❑ **Key resampling techniques:**
 - ❖ Over-sampling
 - ❖ Under-sampling
 - ❖ Combined(Hybrid) Resampling



Oversampling:



❑ Random Over-sampling:

In this method, random instances from the minority class are duplicated until a more balanced distribution is achieved. While this can balance the class distribution, it may lead to overfitting.

❑ SMOTE (Synthetic Minority Over-sampling Technique):

SMOTE generates synthetic instances for the minority class by interpolating between neighboring instances. This approach creates new, realistic data points and helps prevent overfitting compared to random oversampling.

❑ ADASYN (Adaptive Synthetic Sampling)

- ✓ **Description:** An extension of SMOTE that focuses on generating more synthetic data for minority class
- ✓ **Advantages:** Improves the focus on difficult minority class examples, potentially enhancing model performance.
- ✓ **Disadvantages:** Similar to SMOTE, it can introduce noise if not applied carefully.

❑ SMOTEN

❑ SVM-SMOTE

❑ Random oversampler

❑ K-means-SMOTE



Oversampling





Undersampling



❑ Random Under-sampling

Description: Involves reducing the number of instances in the majority class to balance the dataset. This is done by randomly selecting and removing examples from the majority class until the class distribution is balanced

Advantages: Reduces the size of the dataset, making the training process faster.

Disadvantages: Can lead to loss of valuable information and under-fitting.

❑ Tomek Links

Description: Under-sampling technique used to identify and remove overlapping instances in a dataset

Advantages: Helps clean the boundary between classes, improving model performance.

Disadvantages: Only removes a small number of majority class examples, may not fully balance the dataset.

❑ Random under-sampler

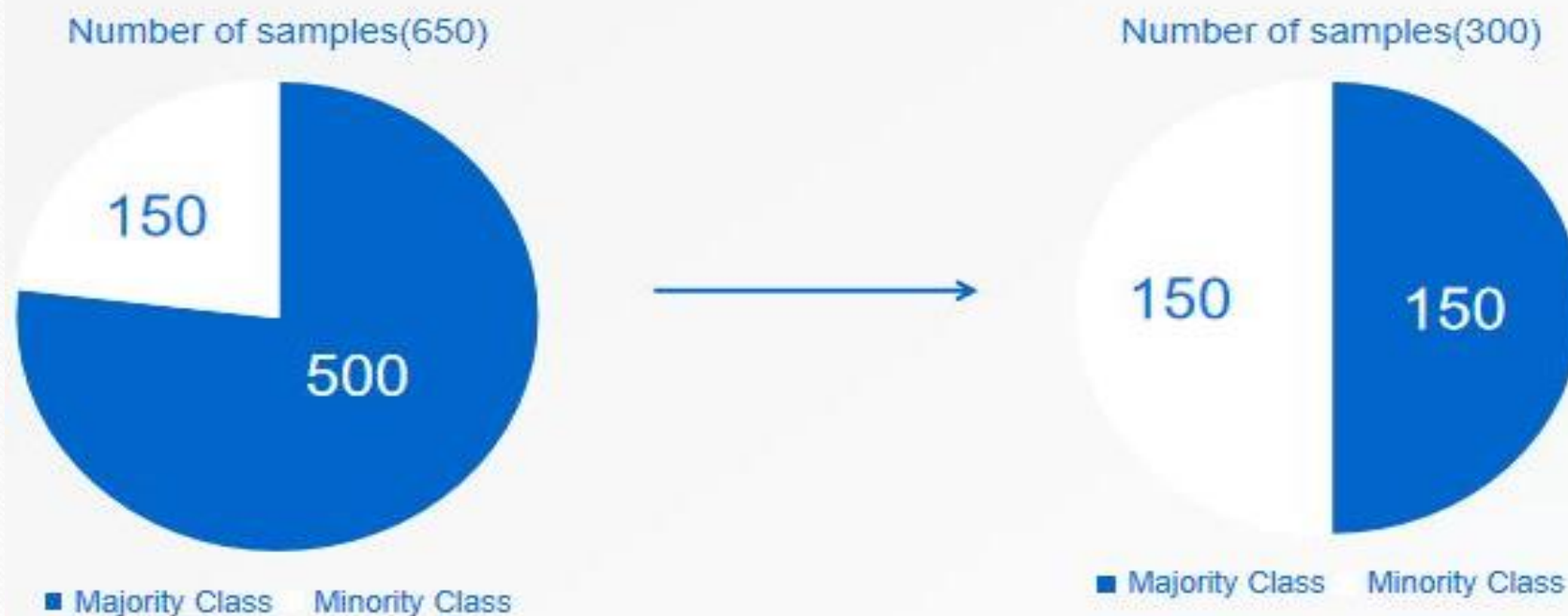
❑ NearMiss

❑ condensed Nearest Neighbour

❑ Edited Nearest Neighbour



Undersampling





Combined Resampling



Hybrid Methods

- ❑ Combining resampling techniques is a powerful strategy for addressing data imbalance in machine learning
- ❑ Involves using both oversampling and undersampling techniques to handle imbalanced datasets.
- ❑ This method leverages the strengths of both approaches to create a more balanced and representative dataset for training machine learning models

Advantages: Can provide a more balanced and effective approach.

Disadvantages: More complex to implement and require careful tuning.



Benefits of Combining Resampling Techniques



Balanced Dataset: More balanced representation of classes can lead to better model

Class Boundaries: Cleaning the class boundaries can enhance the classifier's ability to distinguish between classes.

Robust Models: Models trained on a combined resampled dataset can generalize better to unseen data.

Enhanced Model Performance: Models trained on balanced, informative datasets often demonstrate better performance, particularly when working with imbalanced data



Considerations



- ❑ The choice of combining resampling techniques should be guided by the characteristics of the dataset and the specific problem. It may not be the best approach for all situations.
- ❑ Depending on the problem, you can also experiment with different combinations of oversampling and undersampling techniques to find the most effective balance.
- ❑ Care must be taken when selecting and fine-tuning the specific resampling methods and their parameters to achieve the desired balance and model performance.



Evaluation Metrics for Handling Data Imbalance



Confusion Matrix

		Actual	
		Positives(1)	Negatives(0)
Predicted	Positives(1)	TP	FP
	Negatives(0)	FN	TN

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

Evaluation Metrics for Handling Data Imbalance



1. Confusion Matrix:

- A table that summarizes the performance of a classification algorithm by displaying the true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN).
- Helps in calculating various other metrics.

2. Precision:

- The ratio of correctly predicted positive observations to the total predicted positives.
- $$\text{Precision} = \frac{TP}{TP+FP}$$
- High precision indicates a low false positive rate.

3. Recall (Sensitivity or True Positive Rate):

- The ratio of correctly predicted positive observations to all the actual positives.
- $$\text{Recall} = \frac{TP}{TP+FN}$$
- High recall indicates a low false negative rate.





4. F1-Score:

- The harmonic mean of precision and recall.
- $$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$
- Balances precision and recall in a single metric.

5. Specificity (True Negative Rate):

- The ratio of correctly predicted negative observations to all the actual negatives.
- $$\text{Specificity} = \frac{TN}{TN + FP}$$

6. Balanced Accuracy:

- The average of recall (sensitivity) and specificity.
- $$\text{Balanced Accuracy} = \frac{\text{Sensitivity} + \text{Specificity}}{2}$$





Application of Machine Learning in Malaria Modeling



- ❖ **Disease Diagnosis:** ML algorithms can analyze medical images (like blood smears) and predict malaria presence with high accuracy.
- ❖ **Epidemiological Predictions:** ML models can forecast malaria outbreaks by analyzing patterns in climate, population movements, and historical data.
- ❖ **Drug Discovery:** ML can accelerate the discovery of new antimalarial drugs by predicting which compounds are most likely to be effective.
- ❖ **Treatment Optimization:** Personalizing treatment plans based on patient data to improve outcomes and reduce drug resistance.
- ❖ **Vector Control:** ML models can predict mosquito population dynamics and breeding sites, aiding in targeted vector control measures.



Applications to Malaria Modeling/Diagnosis



MALÁRIA



Wednesday, August 28,
2024





Important Resources

1. Introduction to Statistical Learning, 2nd edition

<https://www.statlearning.com/>