Classification

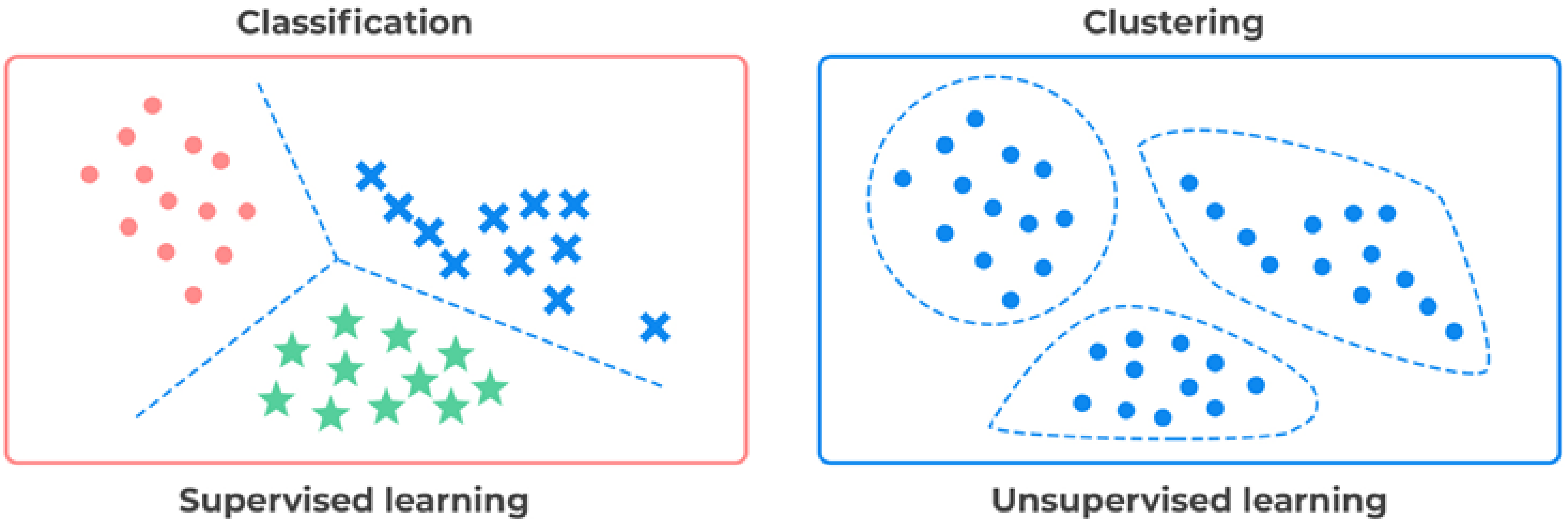
By 4V Analytics

Structured vs Unstructured Data

Structured data	Semi-structured data	Unstructured data
Databases	XML / JSON data	Audio
	Email	Video
	Web pages	Image data
		Natural language
		Documents

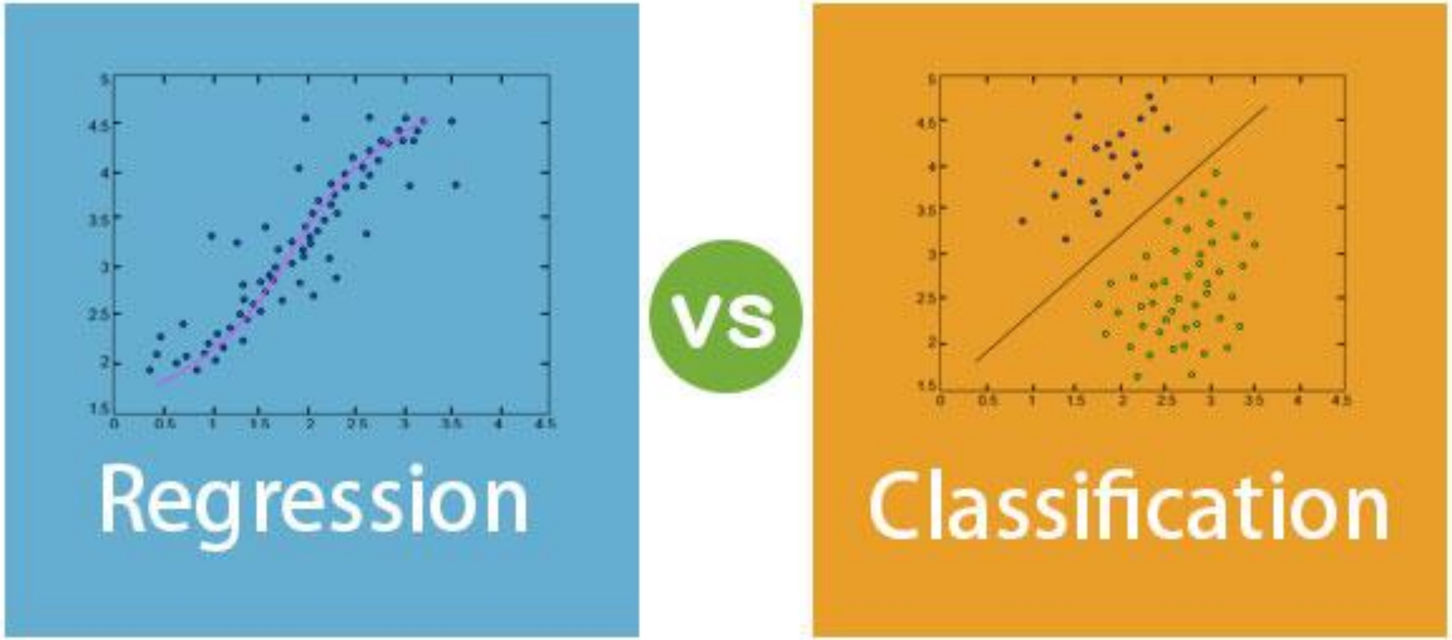
Supervised vs Unsupervised

- Predictive vs Descriptive
- Supervised learning uses labeled input and output data, while an **unsupervised learning algorithm does not.**

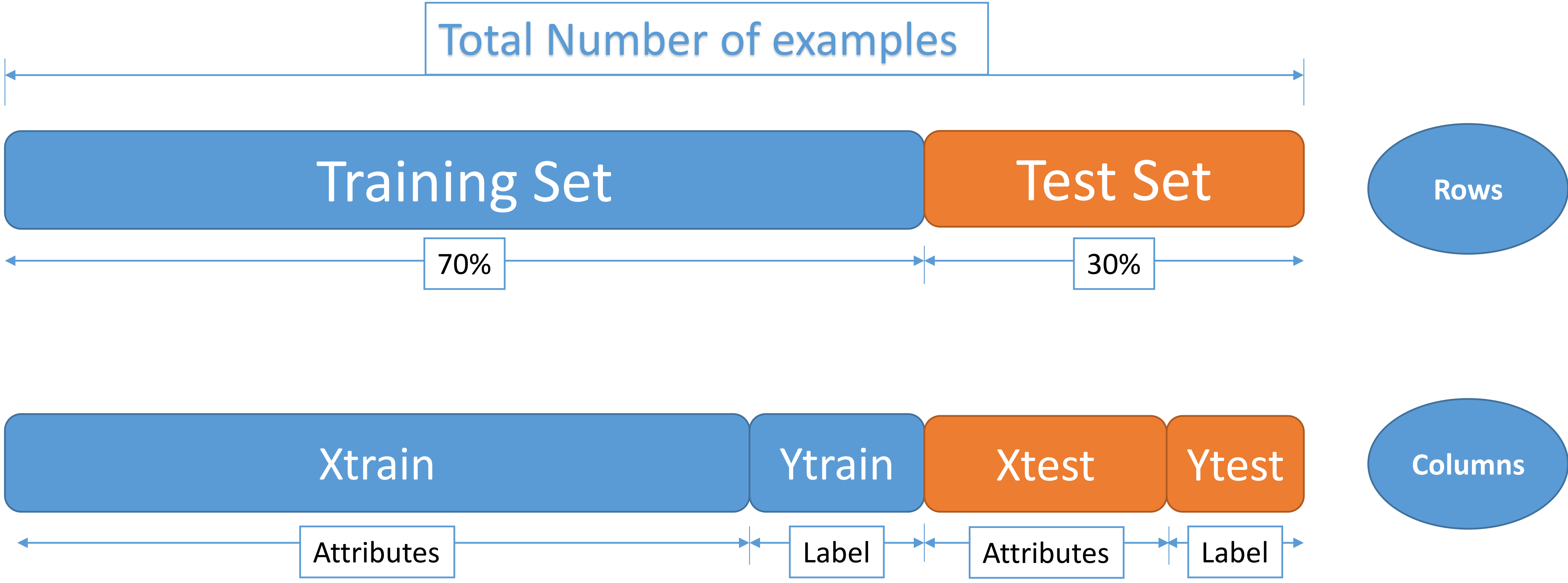


Classification vs Regression

- That classification is the problem of predicting a **discrete class label** output.
- Regression is the problem of predicting a **continuous quantity** output.



Train Test Split



Classification Predictive Modeling

In machine learning, classification refers to a predictive modeling problem where a class label is predicted for a given example of input data.

- Binary Classification
- Multi-Class Classification
- Multi-Label Classification
- Imbalanced Classification

Classification examples

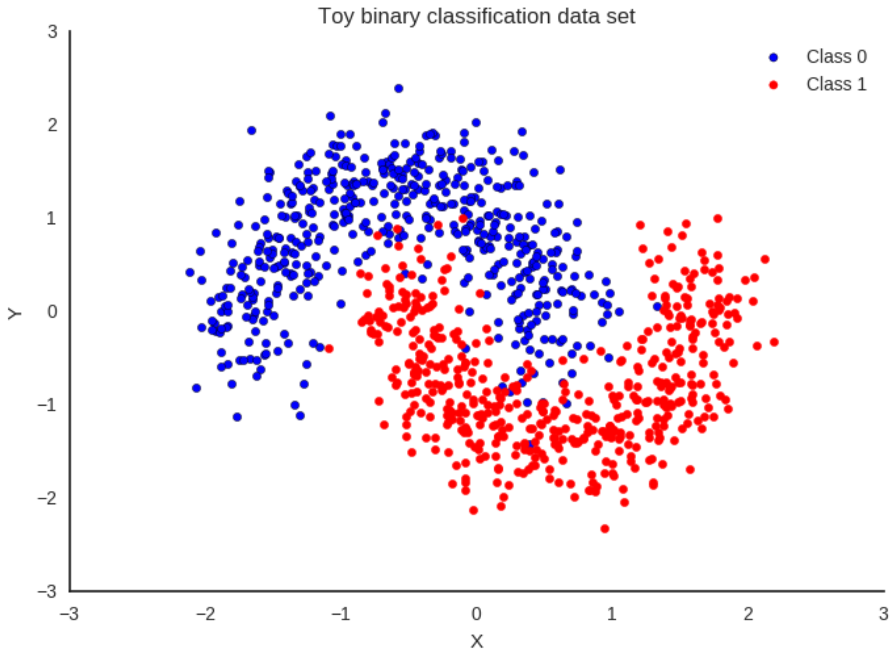
- Given an example, classify if it is **spam or not**.
- Given a handwritten character, classify it as one of the known **characters**.
- Given recent user behavior, classify as **churn or not**.

Binary Classification

It refers to those classification tasks that have **two class labels**.

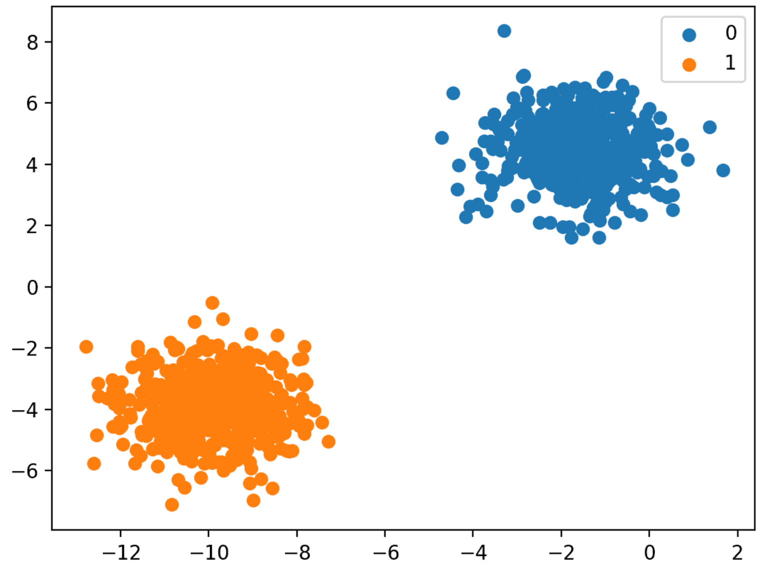
Examples include:

- Cancer detected or not
- Email spam detection (spam or not).
- Churn prediction (churn or not).
- Conversion prediction (buy or not).



Popular algorithms

- Logistic Regression (binary only)
- Support Vector Machine (binary only)
- k-Nearest Neighbors
- Decision Trees
- Naive Bayes



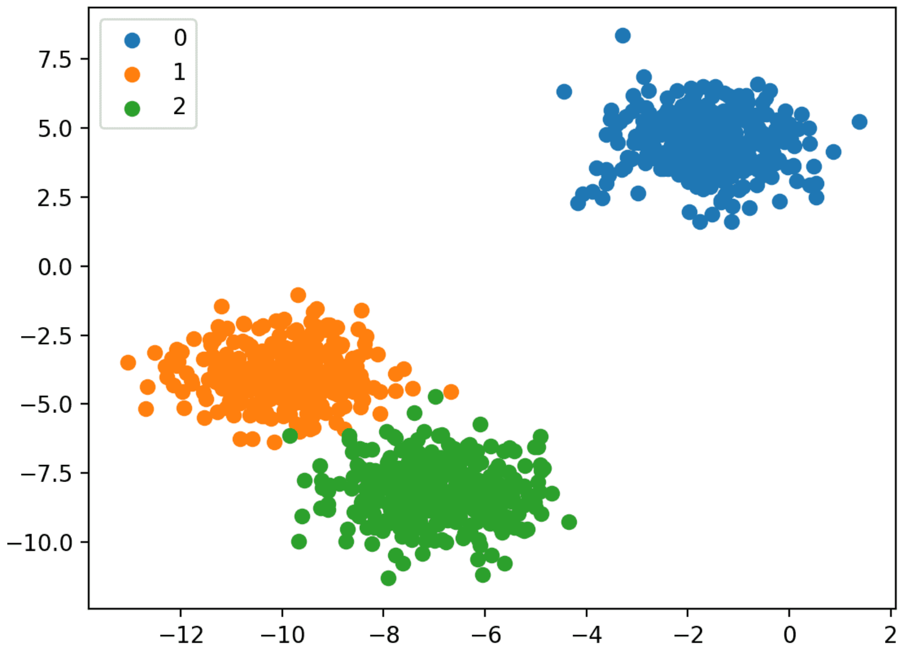
Multi-Class Classification

Unlike binary classification, multi-class classification does not have the notion of normal and abnormal outcomes. Instead, examples are classified as belonging to one among a **range of known classes**.

- Face classification.
- Plant species classification.
- Optical character recognition

Popular algorithms

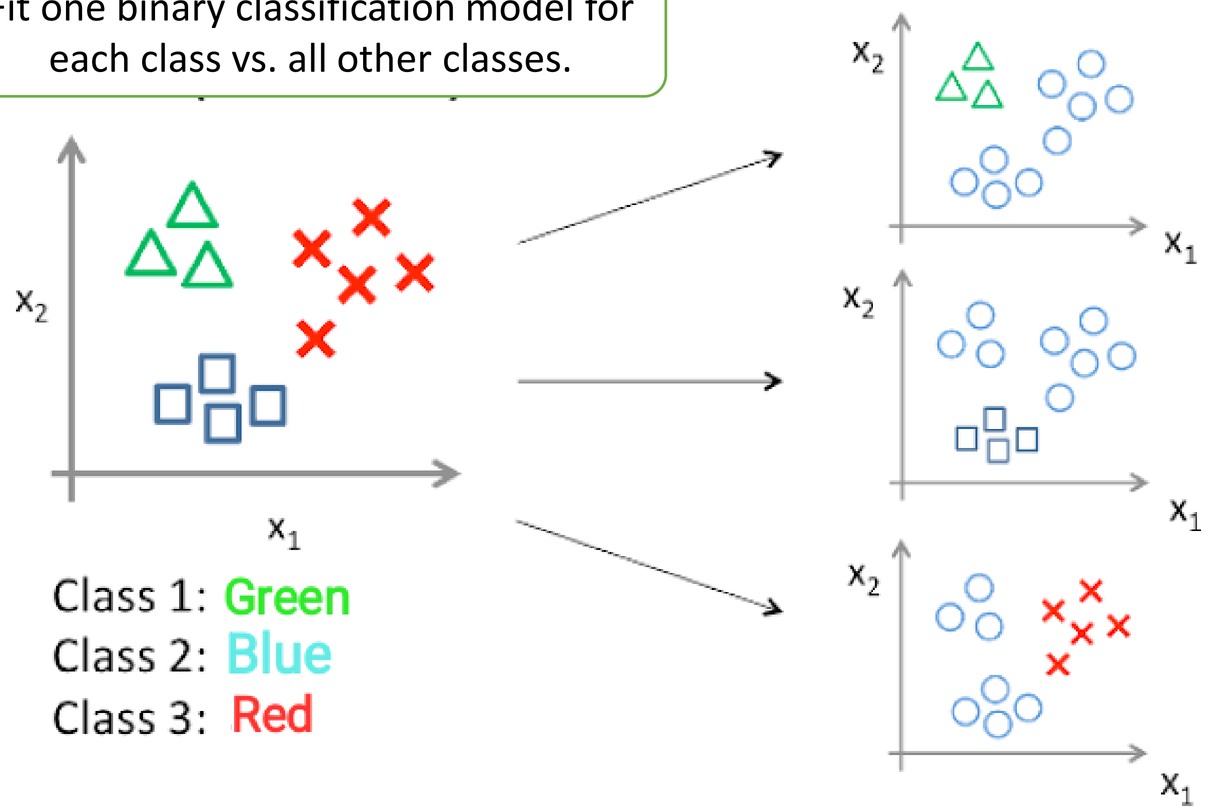
- k-Nearest Neighbors.
- Decision Trees.
- Naive Bayes.
- Random Forest.
- Gradient Boosting



Multi-Class Classification cont.

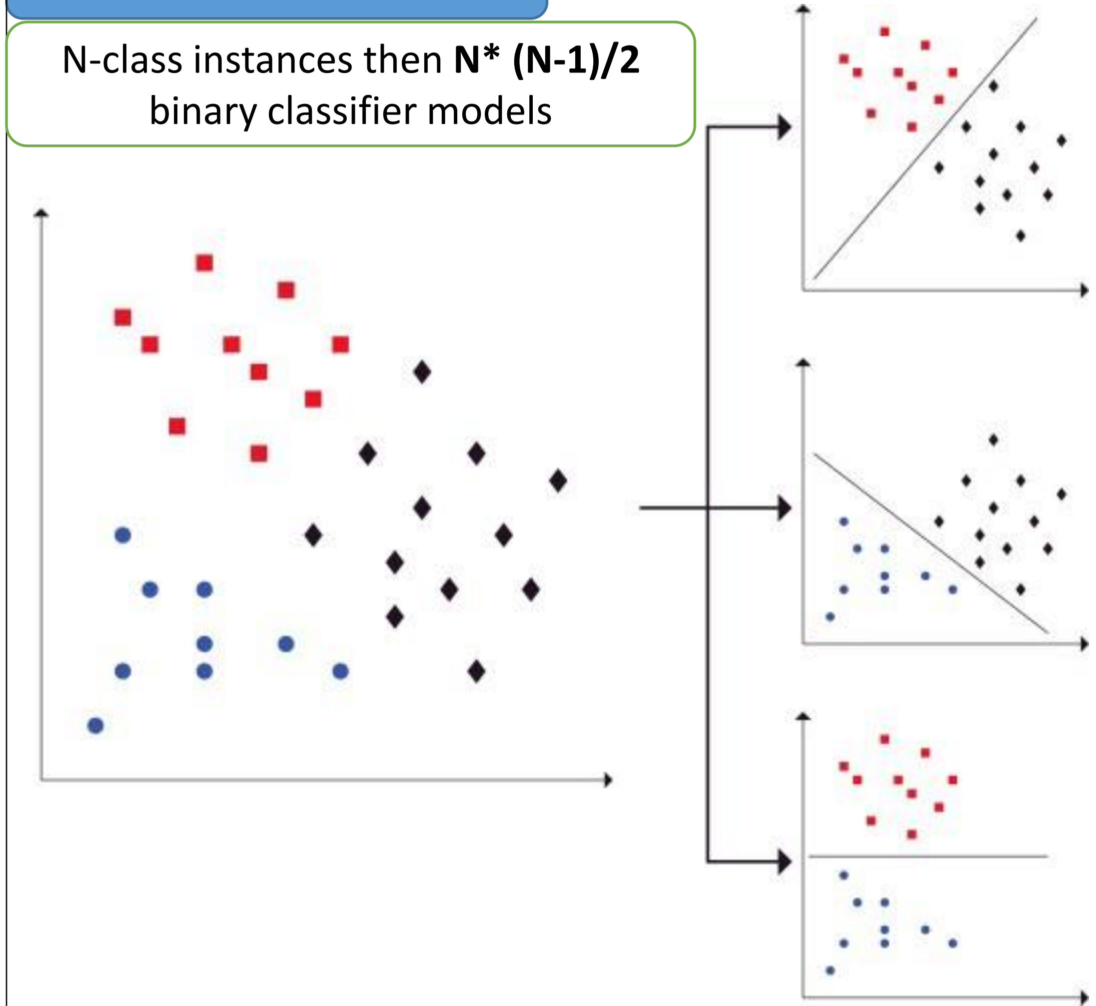
One-vs-All (One-vs-Rest)

Fit one binary classification model for each class vs. all other classes.



One-vs-One

N-class instances then **N* (N-1)/2** binary classifier models



Multi-Label Classification

Multi-Label Classification refers to those classification tasks that **have two or more class labels**, where one or more class labels may be predicted for each example.

Consider the example of **photo classification**, where a given photo may have **multiple objects in the scene** and a model may predict the presence of multiple known objects in the photo, such as “bicycle,” “apple,” “person,” etc.

Popular algorithms

Classification algorithms used for binary or multi-class classification ***cannot*** be used directly for multi-label classification. Specialized versions of standard classification algorithms can be used, so-called multi-label versions of the algorithms, including:

- Multi-label Decision Trees
- Multi-label Random Forests
- Multi-label Gradient Boosting

Note: Another approach is to use a separate classification algorithm to predict the labels for each class.

Imbalanced Classification

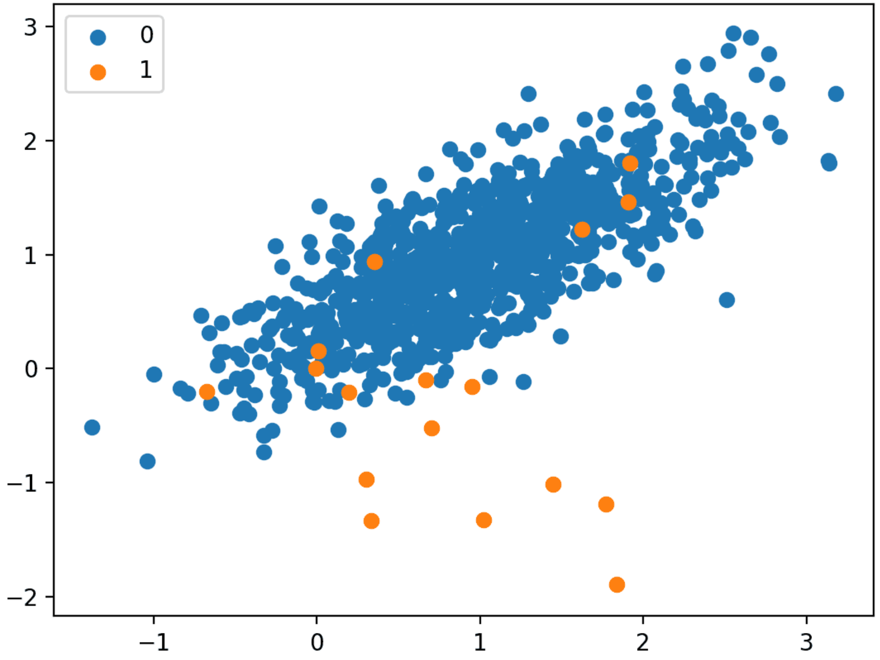
Imbalanced Classification refers to classification tasks where the number of examples in each class is **unequally distributed.**
Typically, imbalanced classification tasks are binary classification tasks where the majority of examples in the training dataset belong to the normal class and a minority of examples belong to the abnormal class.

- Fraud detection.
- Outlier detection.
- Medical diagnostic tests

Popular algorithms

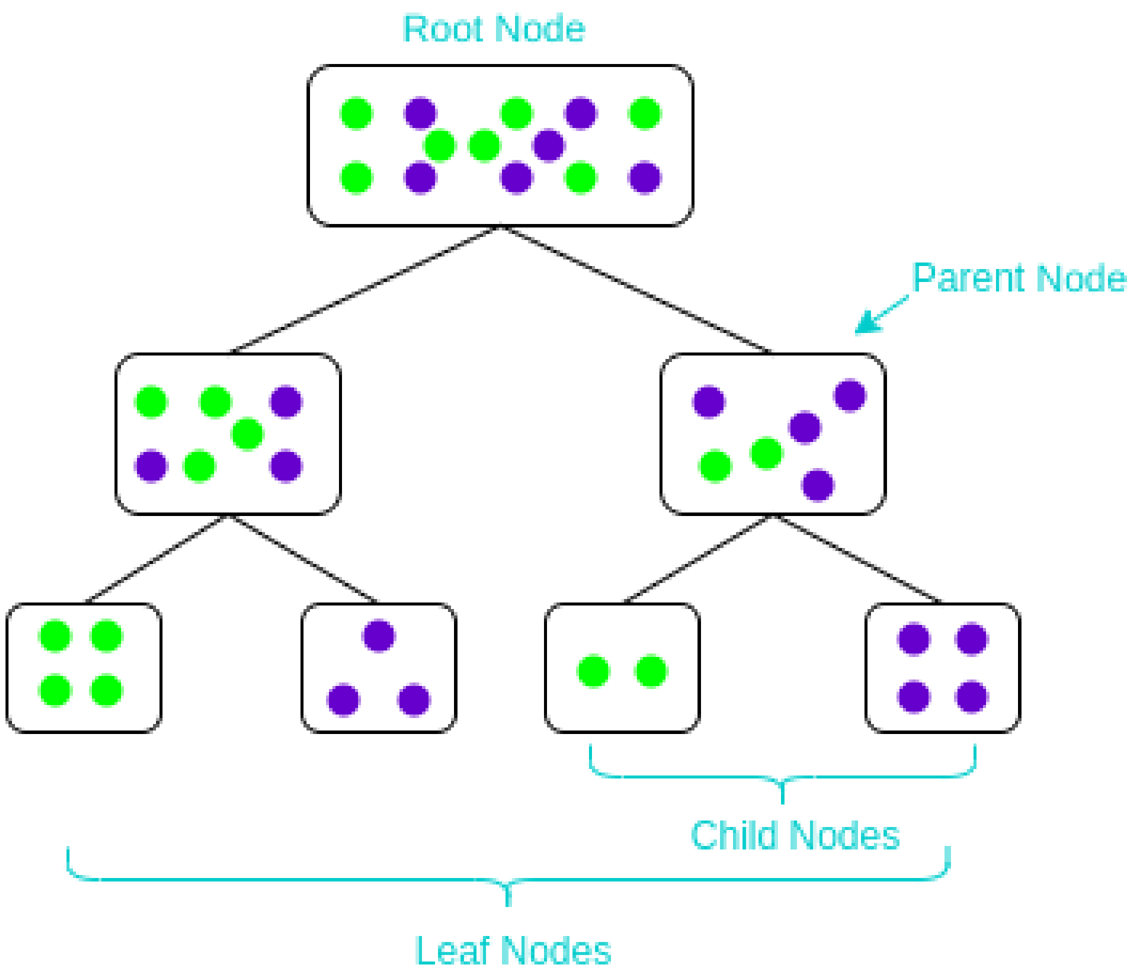
Specialized modeling algorithms may be used that pay more attention to the minority class when fitting the model on the training dataset, such as cost-sensitive machine learning algorithms.

- Cost-sensitive Decision Tree
- Cost-sensitive Logistic Regression



Decision Tree Algorithm

- Decision Tree is a powerful machine learning algorithm that also serves as the building block for other widely used and complicated machine learning algorithms like **Random Forest, XGBoost** and **LightGBM**.

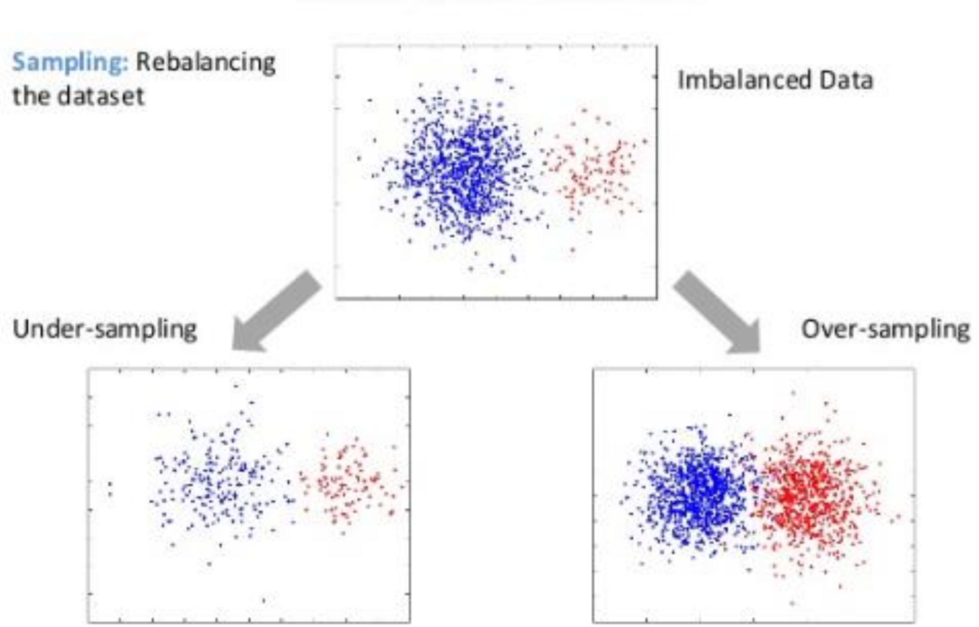


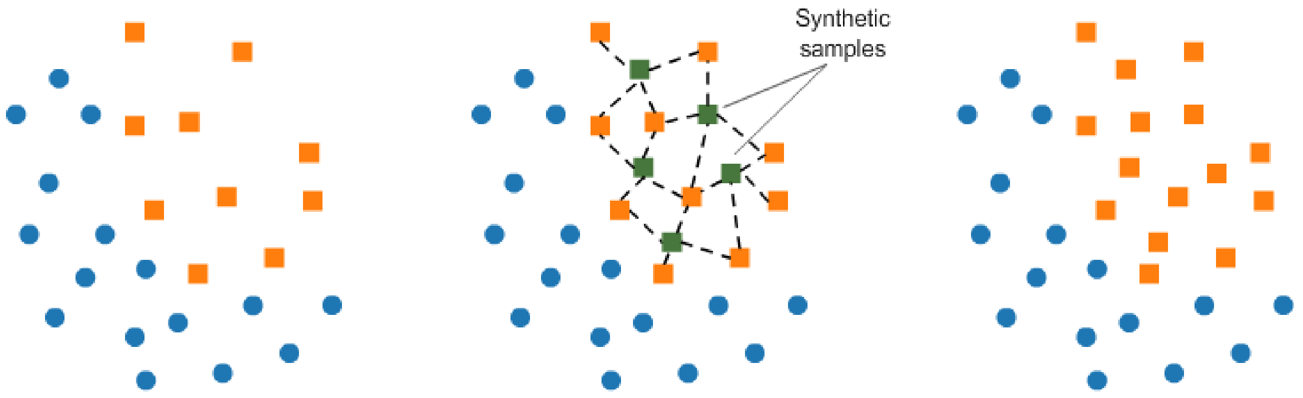
Imbalanced Classification

Specialized techniques may be used to change the composition of samples in the training dataset by under sampling the majority class or oversampling the minority class.

- Random Oversampling
- Random Under sampling
- **SMOTE (Synthetic Minority Oversampling Technique)**

SMOTE first selects a minority class instance at random and finds its **k nearest minority class neighbors**. The synthetic instance is then created by choosing one of the k nearest neighbors b at random and connecting a and b to form a line segment in the feature space.





Confusion Matrix

10, 000 PATIENTS

DIAGNOSIS

PATIENTS

	Diagnosed Sick	Diagnosed Healthy
Sick	1000 True Positives	200 False Negatives
Healthy	800 False Positives	8000 True Negatives

Confusion Matrix

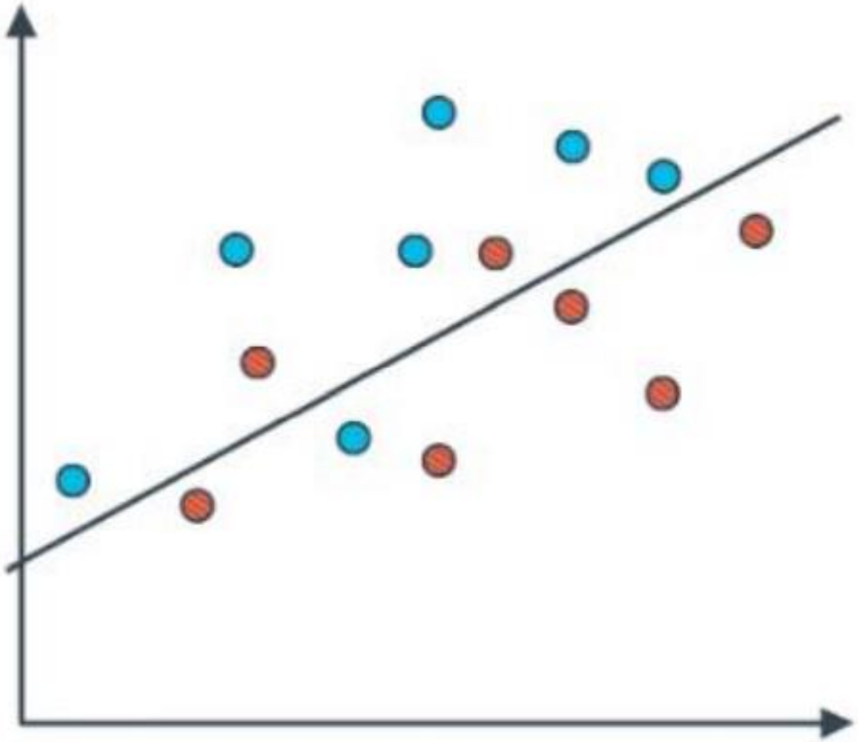
1000 EMAILS

SPAM

EMAIL

	Spam Folder	Inbox
Spam	100 True Positives	170 False Negatives
Not Spam	30 False Positives	700 True Negatives

Confusion Matrix



	Guessed Positive	Guessed Negative
Positive	6 True Positives	1 False Negatives
Negative	2 False Positives	5 True Negatives

TYPE-1 Error

TYPE-2 Error

In this image, the blue points are labelled positive, and the red points are labelled negative. Furthermore, the points on top of the line are predicted (guessed) to be positive, and the points below the line are predicted to be negative.

Accuracy

Out of <u>total patients</u>, how many <u>identified correctly</u>

	Diagnosed Sick	Diagnosed Healthy
Sick	1000 True Positives	200 False Negatives
Healthy	800 False Positives	8000 True Negatives

$$\frac{1000+8000}{1000+8000+200+800} = 90\%$$

Out of <u>total emails</u>, how many <u>identified correctly</u>

	Spam Folder	Inbox
Spam	100 True Positives	170 False Negatives
Not Spam	30 False Positives	700 True Negatives

$$\frac{100+700}{100+700+30+170} = 80\%$$

Precision

Out of all patients diagnosed as sick, how many diagnosed sick correctly

	Diagnosed Sick	Diagnosed Healthy
Sick	1000 True Positives	200 False Negatives
Healthy	800 False Positives	8000 True Negatives

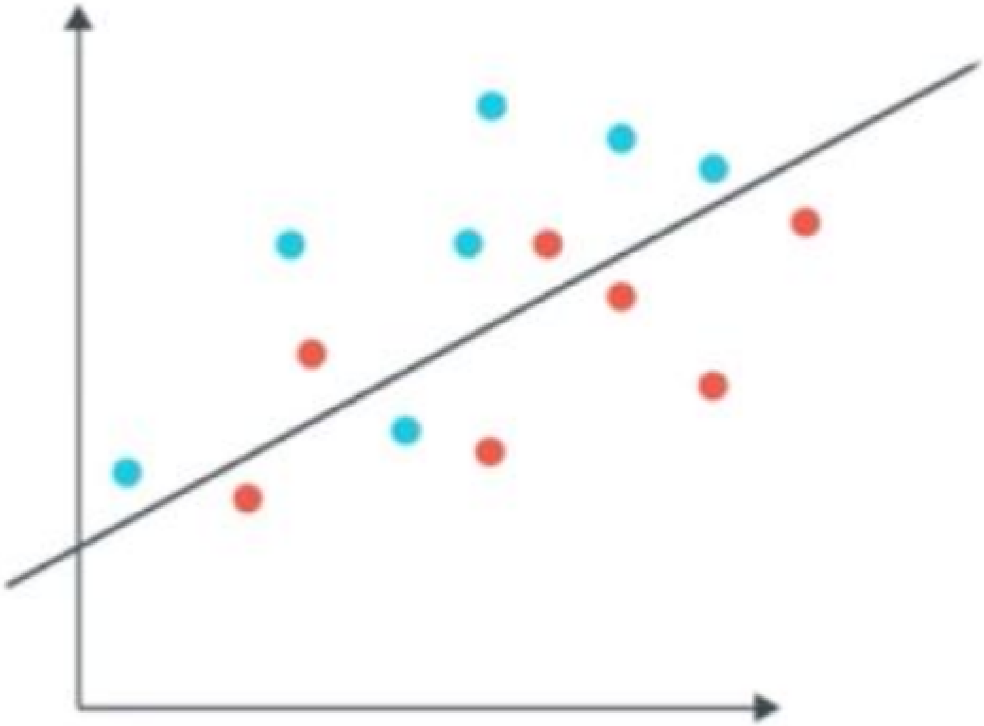
$$\frac{1000}{1000+800} = 55.6\%$$

Out of all emails sent to Spam folder, how many emails sent correctly

	Spam Folder	Inbox
Spam	100 True Positives	170 False Negatives
Not Spam	30 False Positives	700 True Negatives

$$\frac{100}{100+30} = 76.9\%$$

QUIZ



OUT OF THE POINTS WE HAVE PREDICTED TO BE POSITIVE, HOW MANY ARE CORRECT?

Recall

Out of all sick patients, how many were correctly diagnosed as sick

	Diagnosed Sick	Diagnosed Healthy
Sick	1000 True Positives	200 False Negatives
Healthy	800 False Positives	8000 True Negatives

$$\frac{1000}{1000+200} = 83.3\%$$

Out of all spam emails, how many were correctly sent to spam folder

	Spam Folder	Inbox
Spam	100 True Positives	170 False Negatives
Not Spam	30 False Positives	700 True Negatives

$$\frac{100}{100+170} = 37\%$$

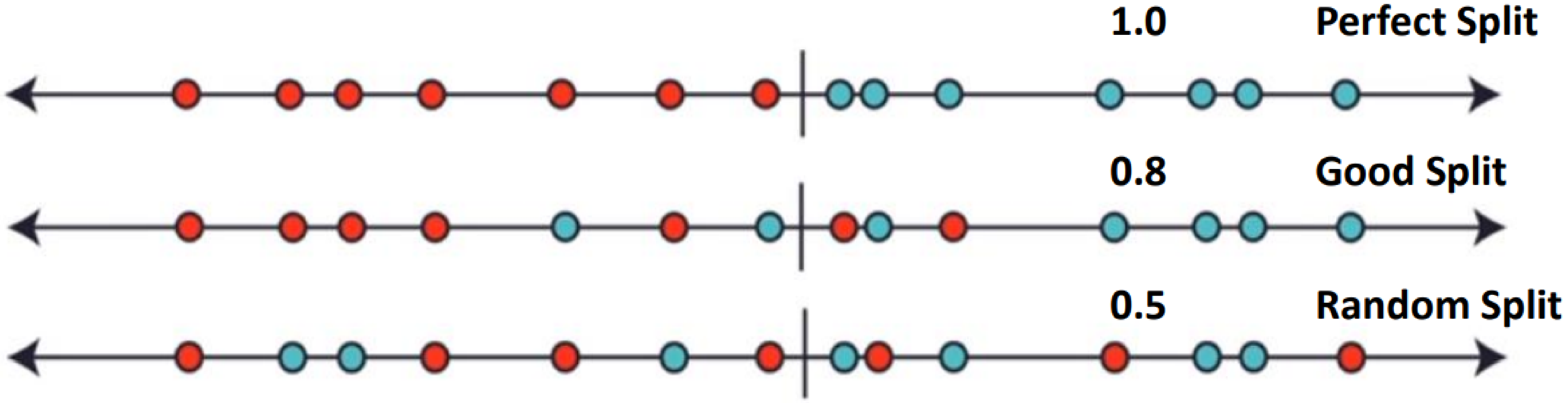
Confusion Matrix

n=165	Predicted: NO	Predicted: YES
Actual: NO	50	10
Actual: YES	5	100

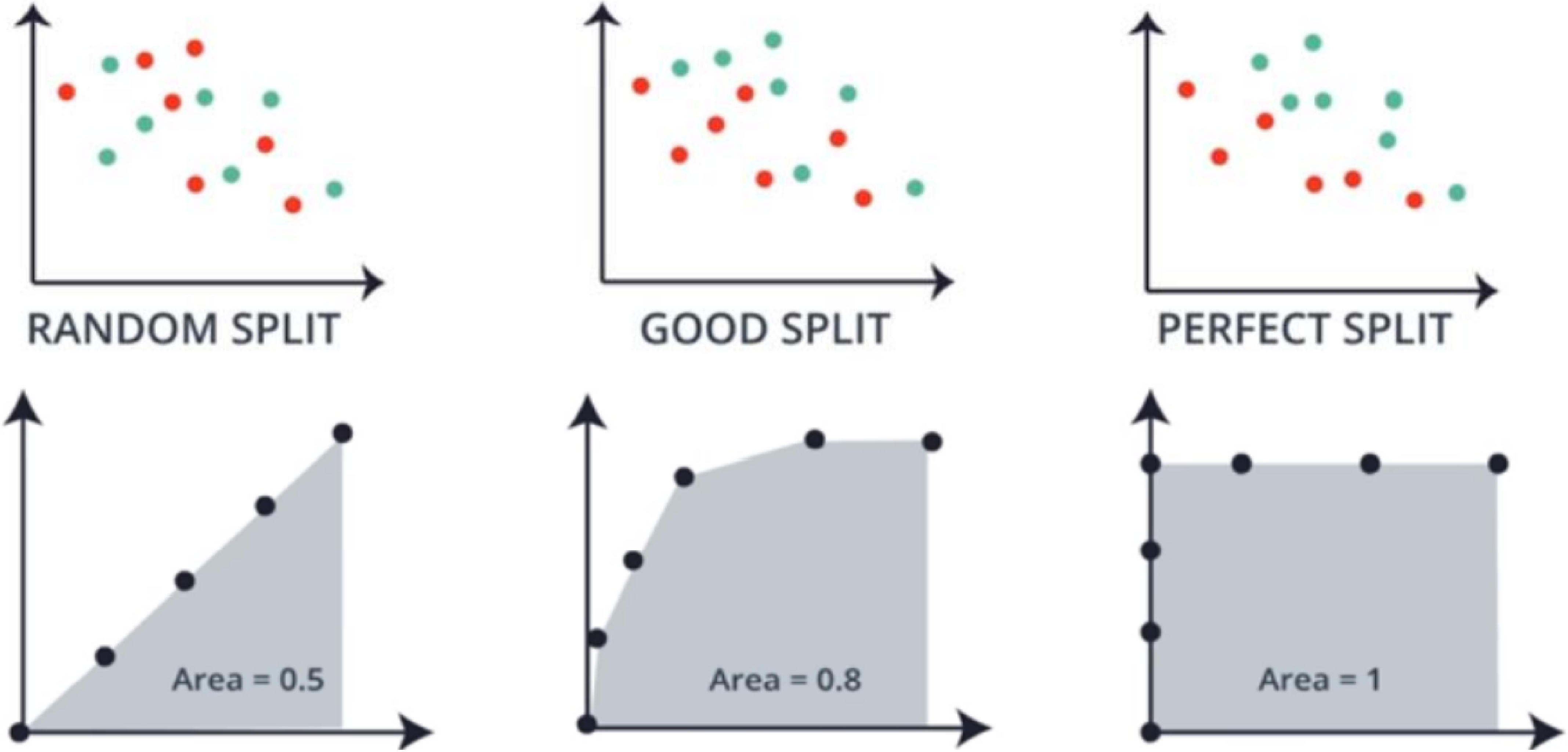
	Predicted: NO	Predicted: YES
Actual: NO	TN	FP
Actual: YES	FN	TP

- **true positives (TP):** These are cases in which we predicted yes (they have the disease), and they do have the disease.
- **true negatives (TN):** We predicted no, and they don't have the disease.
- **false positives (FP):** We predicted yes, but they don't actually have the disease. (Also known as a "Type I error.")
- **false negatives (FN):** We predicted no, but they actually do have the disease. (Also known as a "Type II error.")

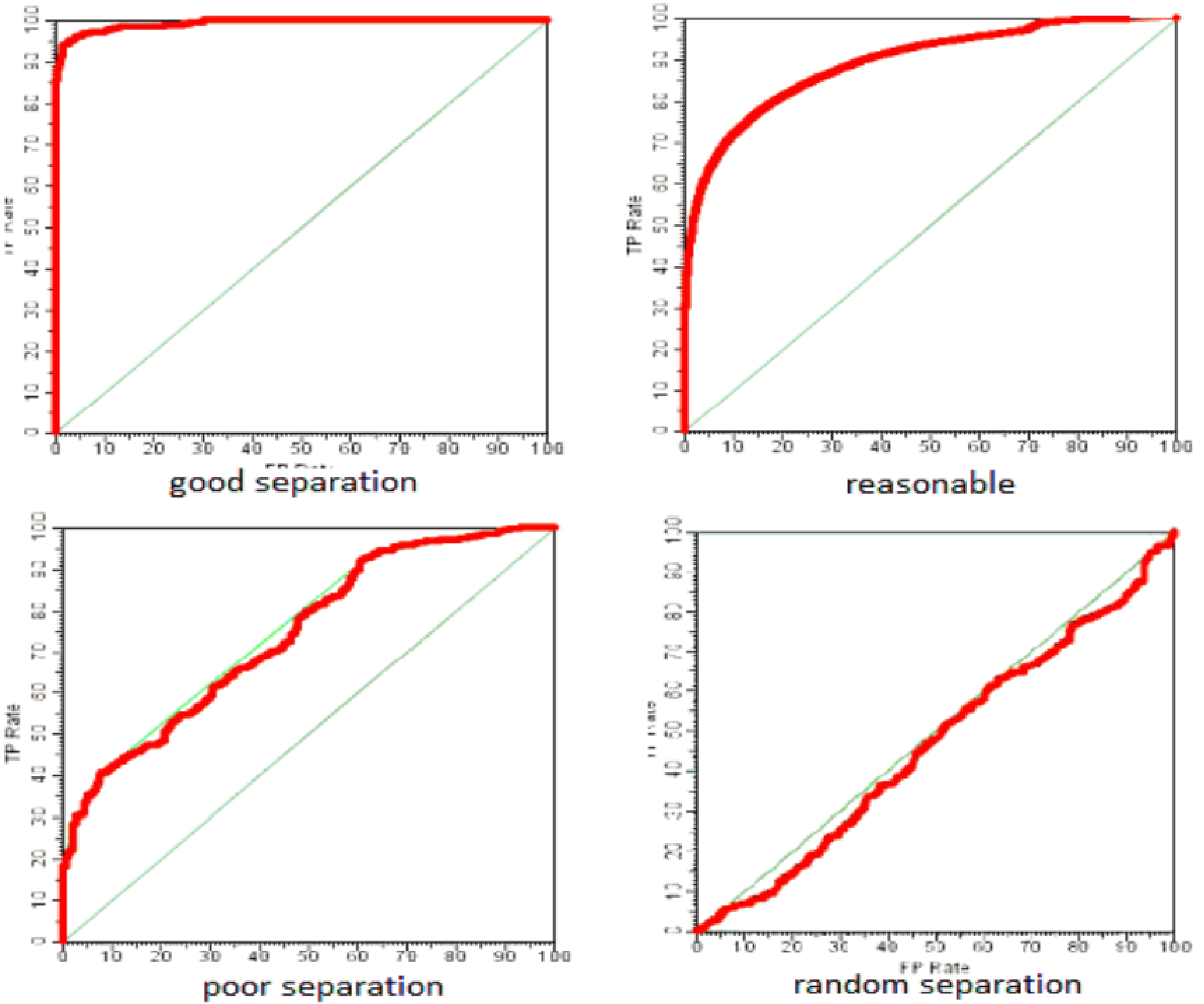
Receiver Operating Characteristic



AREA UNDER ROC Curve



ROC AUC Curve



F1 Score

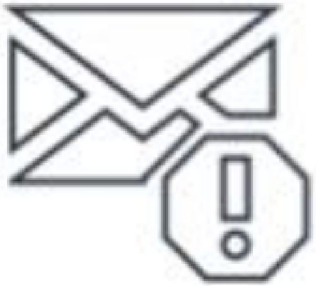
MEDICAL MODEL

PRECISION: 55.7%

RECALL: 83.3%

AVERAGE: 69.5%

ONE SCORE?



SPAM DETECTOR

PRECISION: 76.9%

RECALL: 37%

AVERAGE: 56.95%

$$(10.1)\quad Accuracy = \frac{T_p + T_n}{T_p + T_n + F_p + F_n}$$

$$(10.2)\quad Precision = \frac{T_p}{T_p + F_p}$$

$$(10.3)\quad Recall = \frac{T_p}{T_p + T_n}$$

$$(10.4)\quad F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$