# Data Mining Tutorial

Mark A. Austin

University of Maryland

*austin@umd.edu*
*ENCE 688P, Fall Semester 2021*

October 16, 2021

# Overview

1. Quick Review
2. Introduction to Data Mining
3. Entropy, Probability Distributions, and Information Gain
4. Information Gain in Decision Trees
5. Ensemble Learning
6. Metrics of Evaluation
7. Working with Weka
8. Data Mining Examples

Part 04

# Ensemble Learning

# Ensemble Methods (General Idea)

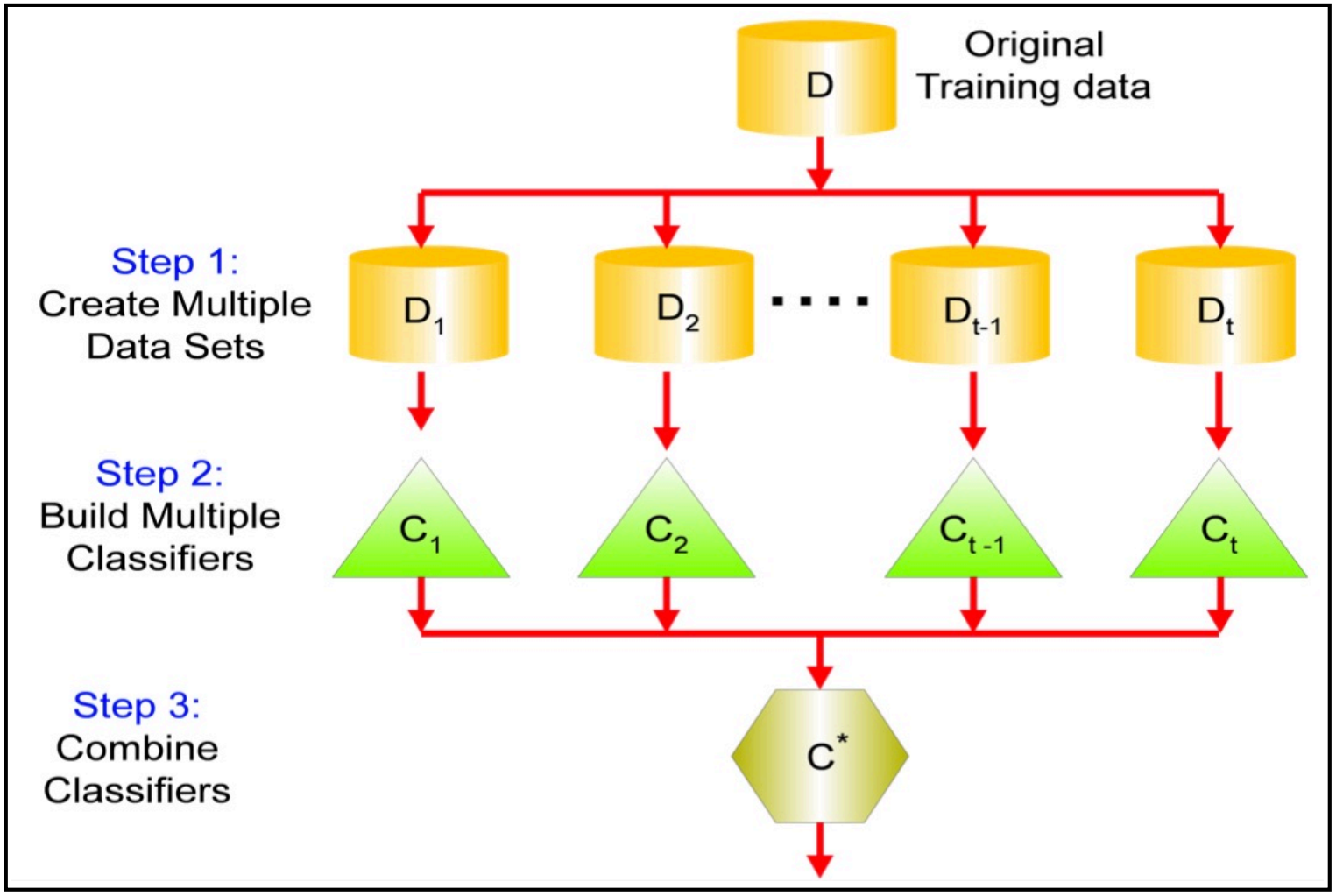
## Ensemble Methods

Ensemble methods use multiple learning algorithms to obtain better predictive performance than could be obtained from any one constituent learning algorithm.

**Motivation and Approach**

- Supervised learning algorithms search through a hypothesis space to find a hypothesis that will make good predictions.
- Even if the hypothesis space contains hypotheses that are well suited to a particular problem space, find a good hypothesis can still be very difficult.
- Ensembles combine hypotheses in the hope of finding a new one with superior predictive capabilities.

# Ensemble Learning (General Idea)



# Ensemble Learning (General Idea)

**Ensemble Learning**

- Combine predictions from multiple learning algorithms $\longrightarrow$ ensemble.
- Often leads to better predictive performance than a single learner.
- Works well then small differences in the training data produce very different classifiers (e.g., decision trees).

**Drawbacks**

- Increased computational effort.
- Reduced level of interpretability.

# Ensemble Learning (Why does it work?)

**Why does it work?**

- Assume classifiers $C_1, \cdots, C_k$ are independent, i.e.,

$$\text{correlation} \quad \sigma(C_1, C_2) = 0. \tag{21}$$

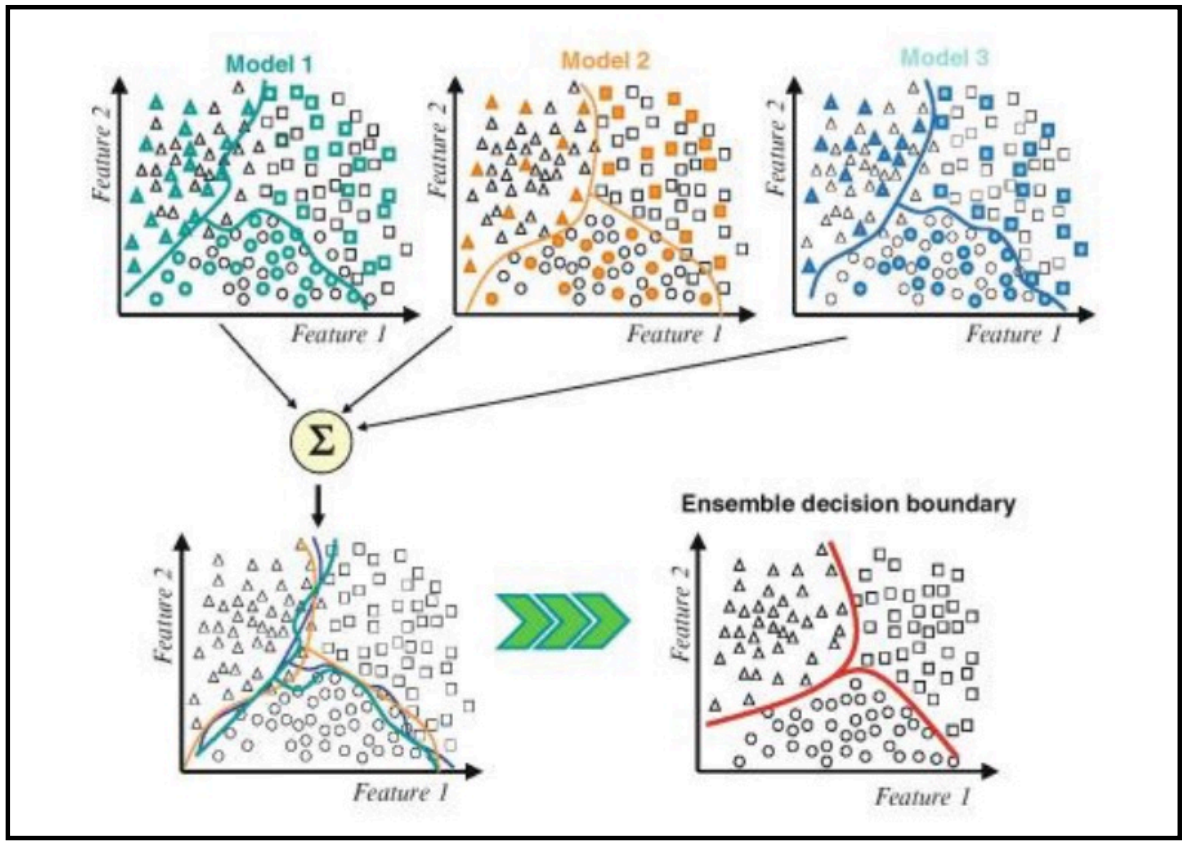
- Assume, for example, that there are 25 classifiers, each having an error rate $\eta = 0.35$.
- Probability that the ensemble classifier makes a wrong prediction:

$$\sum_{i=13}^{25} \binom{25}{i} \eta^i (1-\eta)^{25-i} = 0.06. \tag{22}$$

which is much lower than any individual classifier.

# Ensemble Learning (Diversity in Prediction)

Use of ensemble methods can lead to improvements in prediction accuracy through reduction of variability.




Source: Zhang, et al, Ensemble Machine Learning, Springer, 2012.

# Ensemble Learning

**Constructing Ensembles:** Methods for obtaining sets of classifiers

- **Bagging.**
- **Random Forest.**
- **Cross-Validation.** Two key ideas: (1) instead of different classifiers, train same classifier on different data, (2) since training data is expensive, reuse data bu subsampling.

**Combining Classifiers:** Methods for combining different classifiers

- Stacking
- Bayesian Model Averaging
- Boosting
- AdaBoost

# Ensemble Techniques (Bagging)

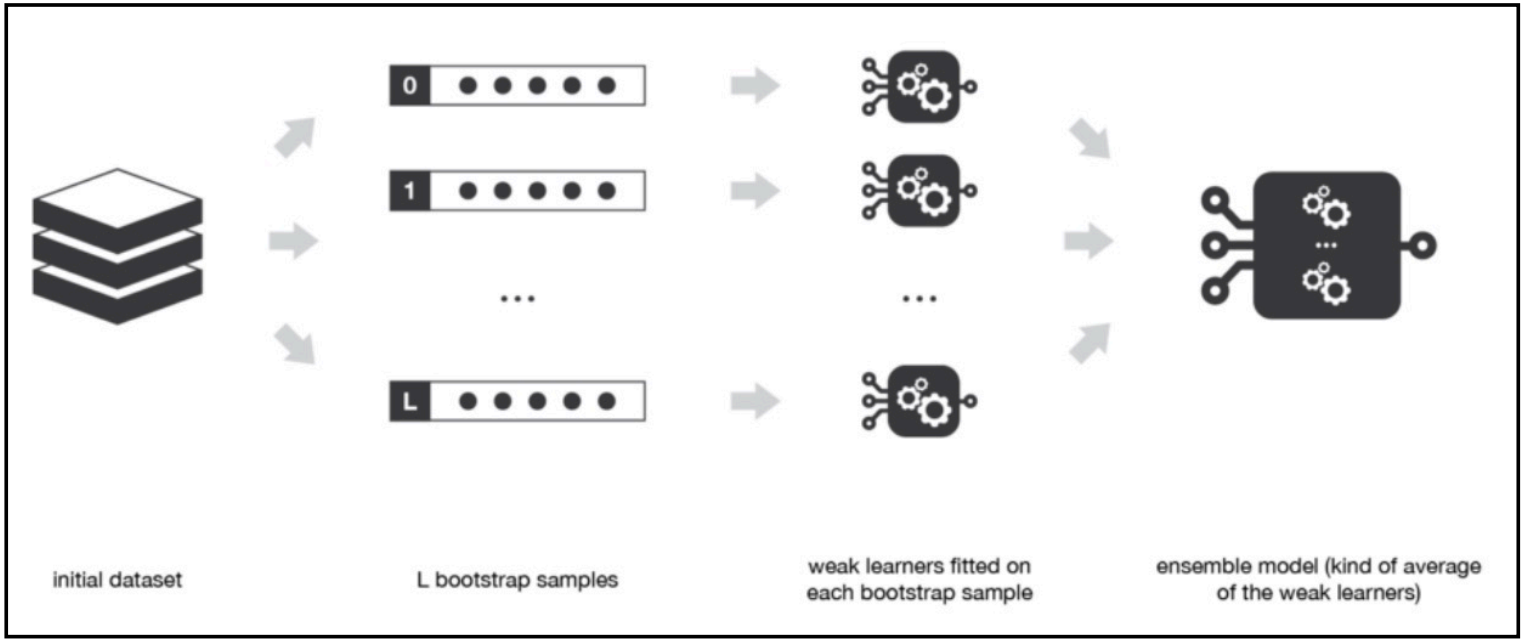
**Bagging** (Breiman, 1996). Bootstrapping on data.

- Create a data set by sampling data points with replacement.

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Original Data : | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| Bagging (Round 1): | 7 | 2 | 9 | 7 | 3 | 2 | 1 | 1 | 4 | 5 |
| Bagging (Round 2): | 6 | 10 | 4 | 2 | 10 | 3 | 8 | 9 | 7 | 4 |
| Bagging (Round 3): | 4 | 6 | 8 | 2 | 5 | 1 | 6 | 3 | 1 | 9 |
| Bagging (Round 4): | ..... | | | | | | | | | |
| Bagging (Round 5): | ..... | | | | | | | | | |

- Create models based on the data sets.
- Generate more data sets and models.
- Make predictions by combining votes – Classification $\rightarrow$ majority vote; prediction $\rightarrow$ average.

# Ensemble Techniques (Bagging)



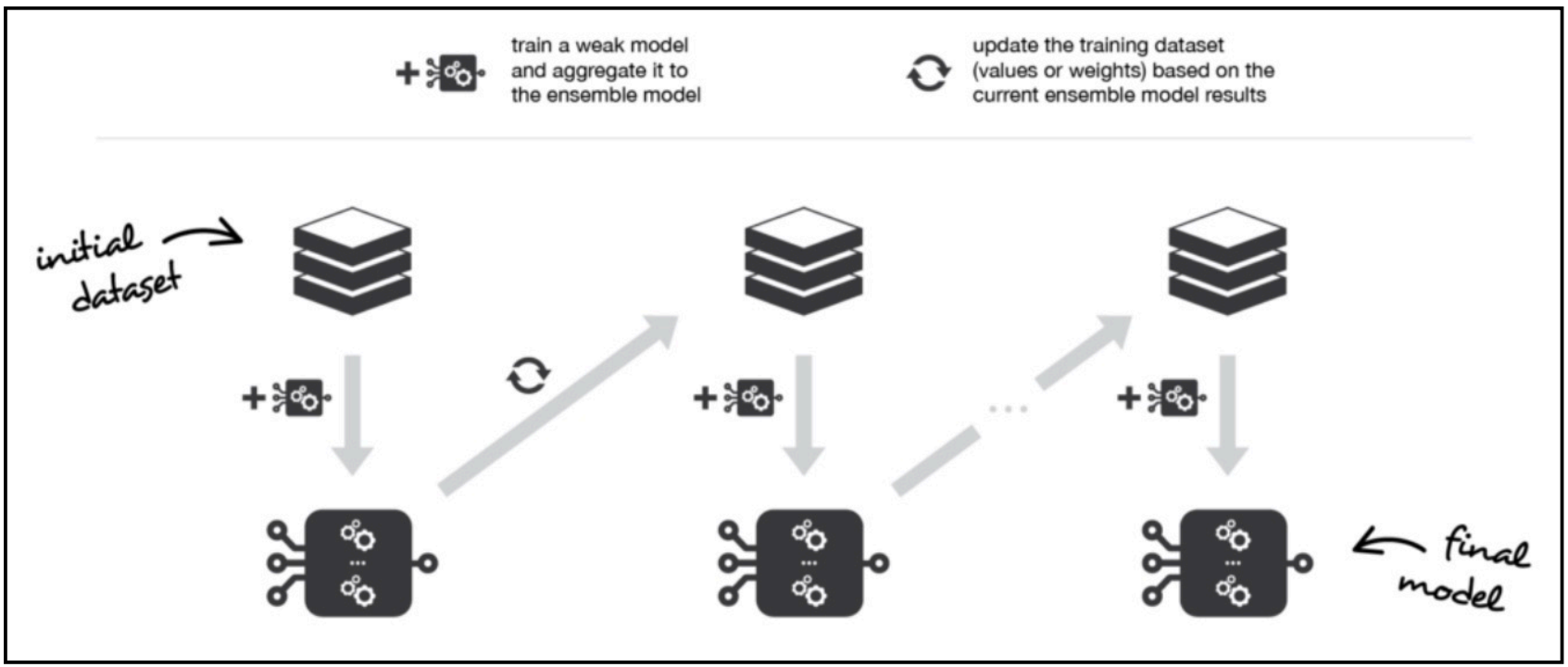

**Advantages/Disadvantages:**

- Helps when classifier is unstable (has high variance).
- Not helpful when classifier is stable and has large bias.

# Ensemble Techniques (Overview)

**Boosting** (Schapire, 1998). Recursively reweight data.

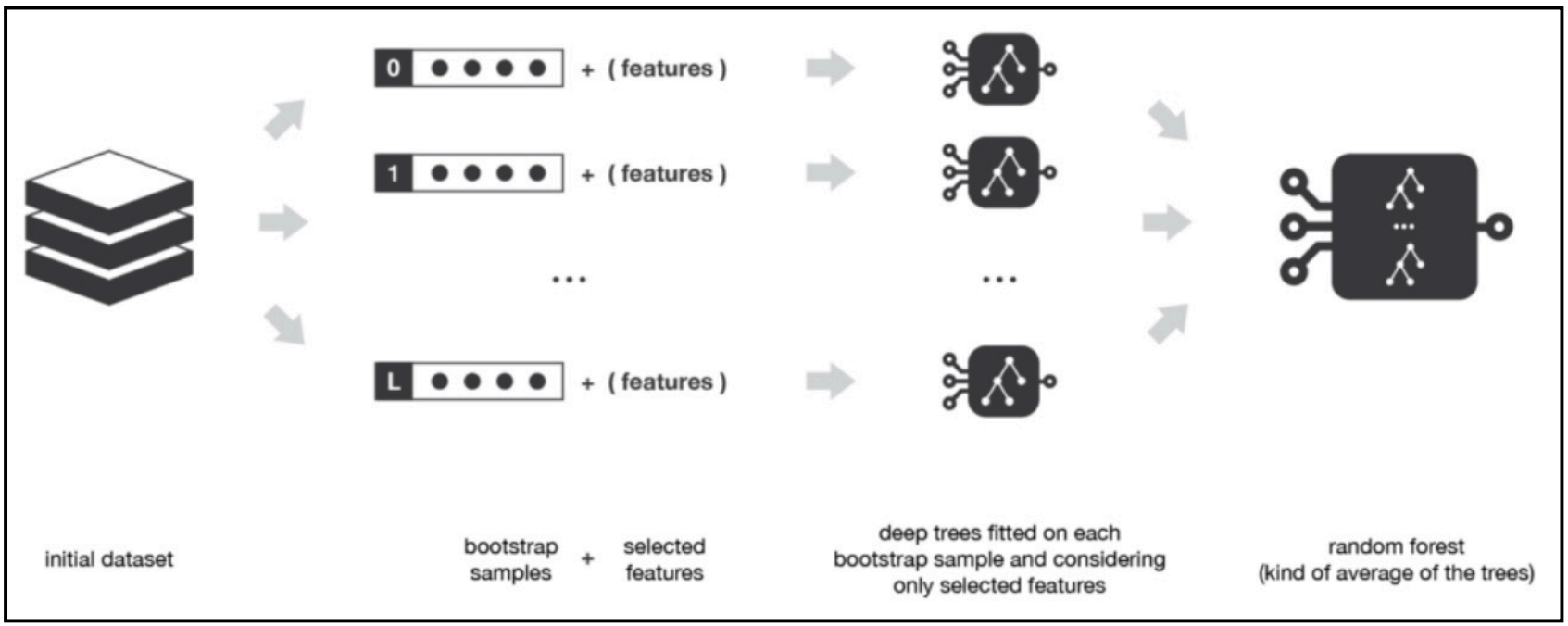
- Records wrongly classified will have their weights increased.
- Records correctly classified will have their weights decreased.



# Ensemble Techniques (Random Forest)

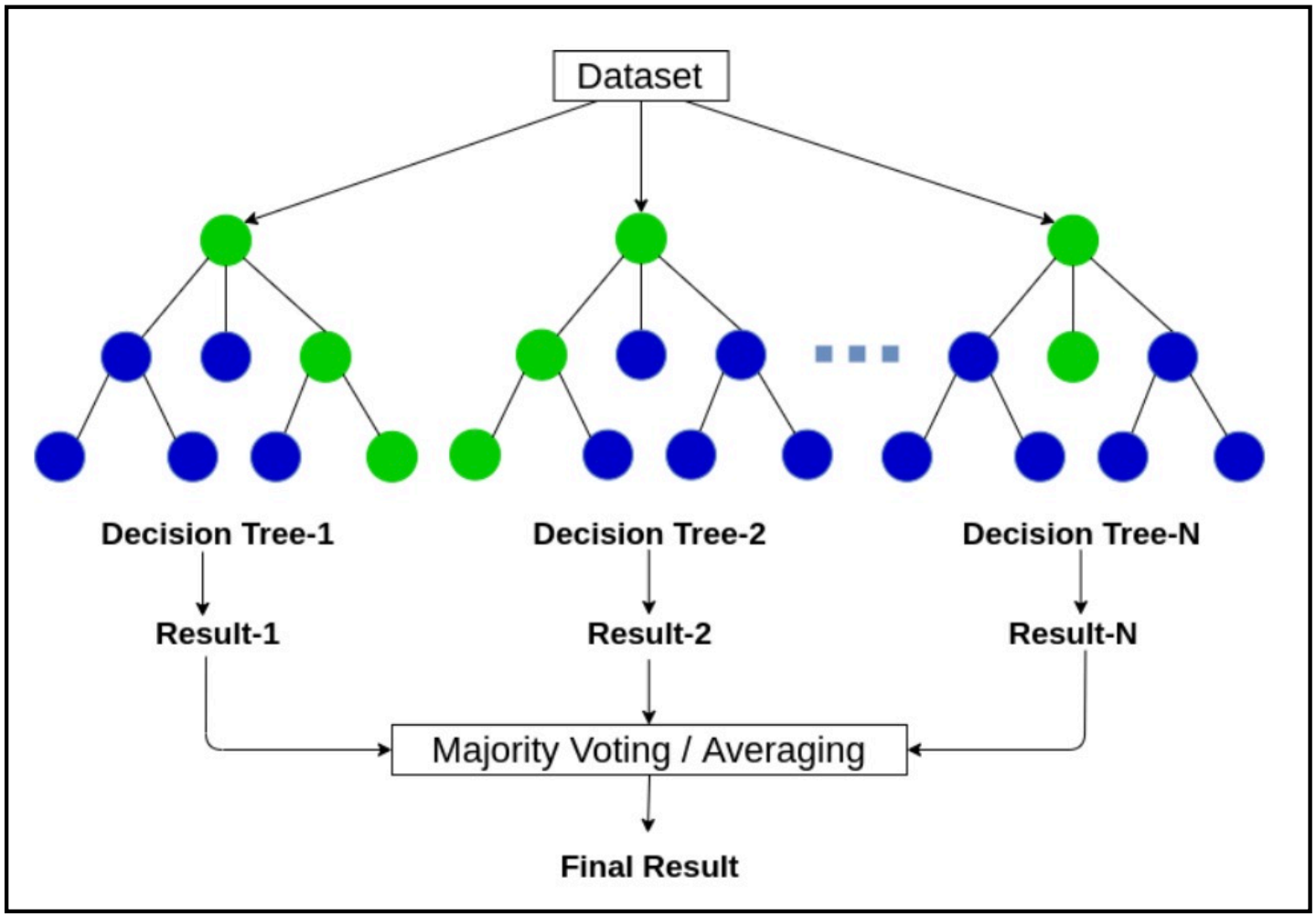
**Random Forest** (Breiman, 2001).

- Randomly pick features and data to generate diversity of classifiers (decision trees).



# Ensemble Techniques (Random Forest)

**Random Forest** (Breiman, 2001).



# Metrics of Evaluation

# Metrics of Evaluation

## Cross Validation Model

Cross validation is a method for assessing how the results of a data mining (statistical) analysis will generalize to an independent dataset. It is mainly used in predictive model applications.

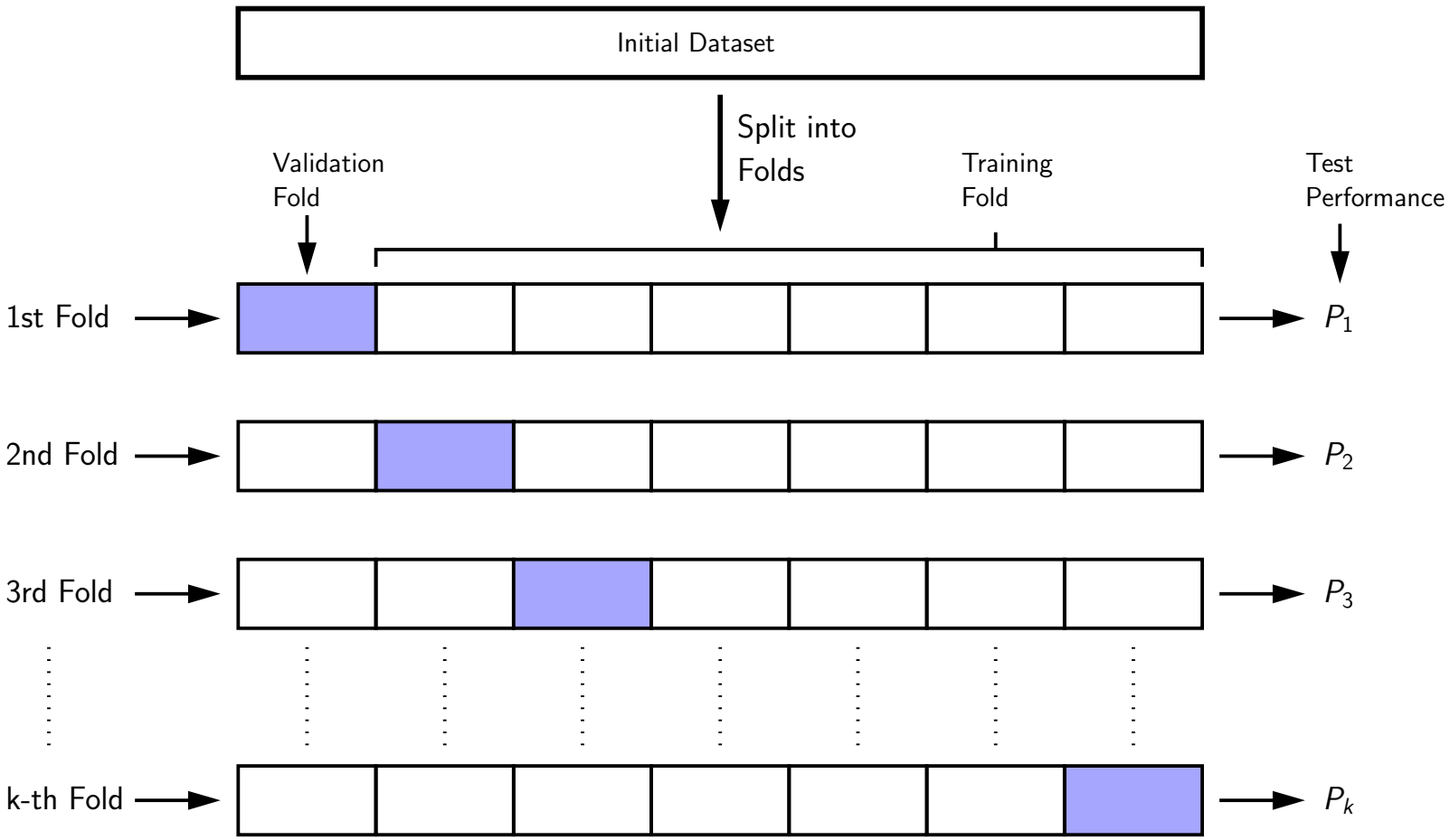
### K-Fold Cross Validation Method

- Divide the sample data into $k$ equal parts.
- Use $k-1$ parts for training and one for testing.
- Repeat the procedure $k$ times, rotating the test dataset.
- Compute metrics of performance across the iterations, i.e.,

$$\text{Performance} = \sum_{i=1}^{k} P_i. \tag{23}$$

# Metrics of Evaluation

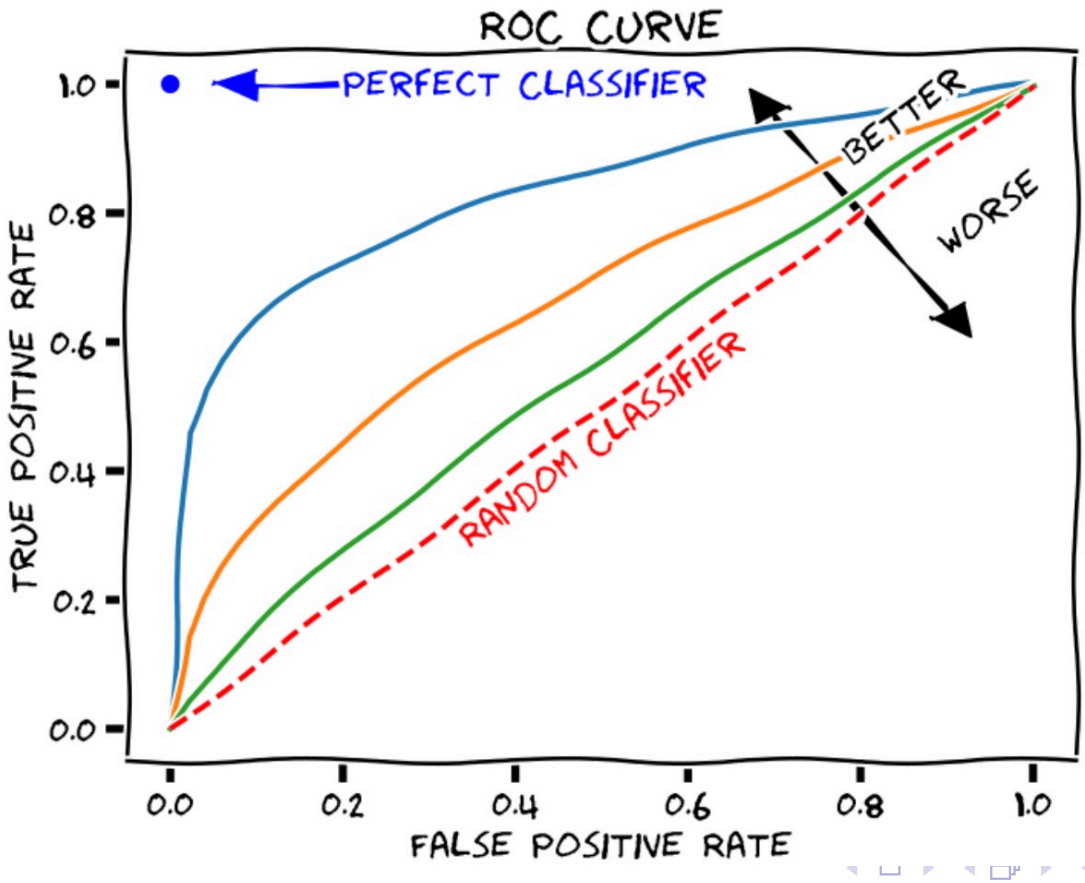
## K-Fold Cross Validation



# Metrics of Evaluation

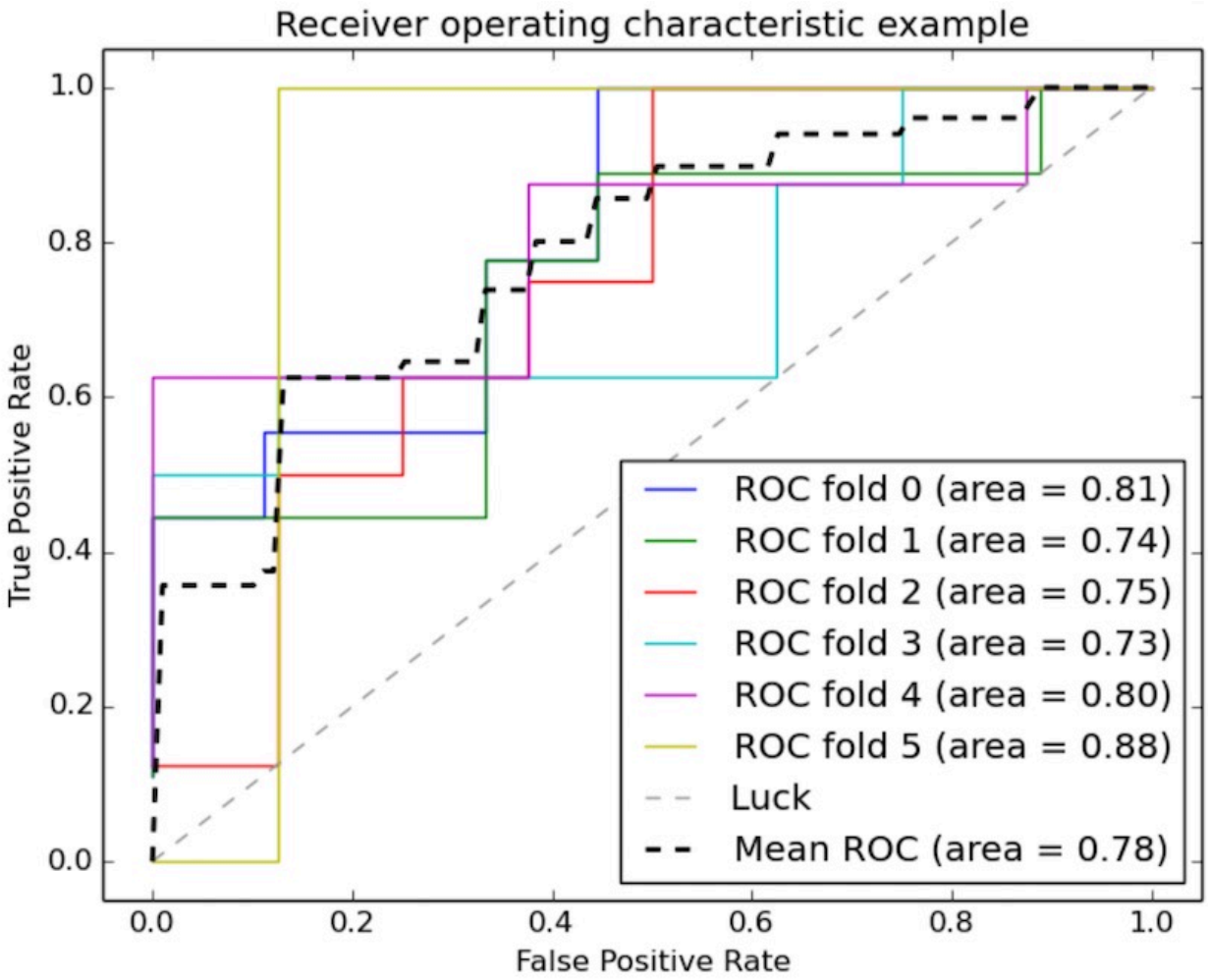
## Receiver Operating Curve

A receiver operating curve (ROC) illustrates diagnostic ability of a binary classifier as its discrimination threshold is varied.



# Metrics of Evaluation

## Typical ROC Curves



# References


- Jaynes E.T., Information Theory and Statistical Mechanics. II, Phys. Rev. 108, 171, October 1957.
- Kapur J.N., Maximum-Entropy Models in Science and Engineering, John Wiley and Sons, 1989.
- Mitchell T.M., Machine Learning and Data Mining, Communications of the ACM, Vol. 42., No. 11, November 1999.
- Russell S., and Norvig P., Artificial Intelligence: A Modern Approach (Third Edition), Prentice-Hall, 2010.
- Shanon C.E., and Weaver W., The Mathematical Theory of Communication, University of Illinois, Urbana, Chicago, 1949.
- Witten I.H., Frank E., Hall M.A., and Pal C.J., Data Mining: Practical Machine Learning Tools and Techniques, Morgan Kaufmann, 2017.