Machine Learning

| Gini Impurity Index

Prof. Dr. Mostafa Elhosseini

https://youtube.com/ElhosseiniAcademy

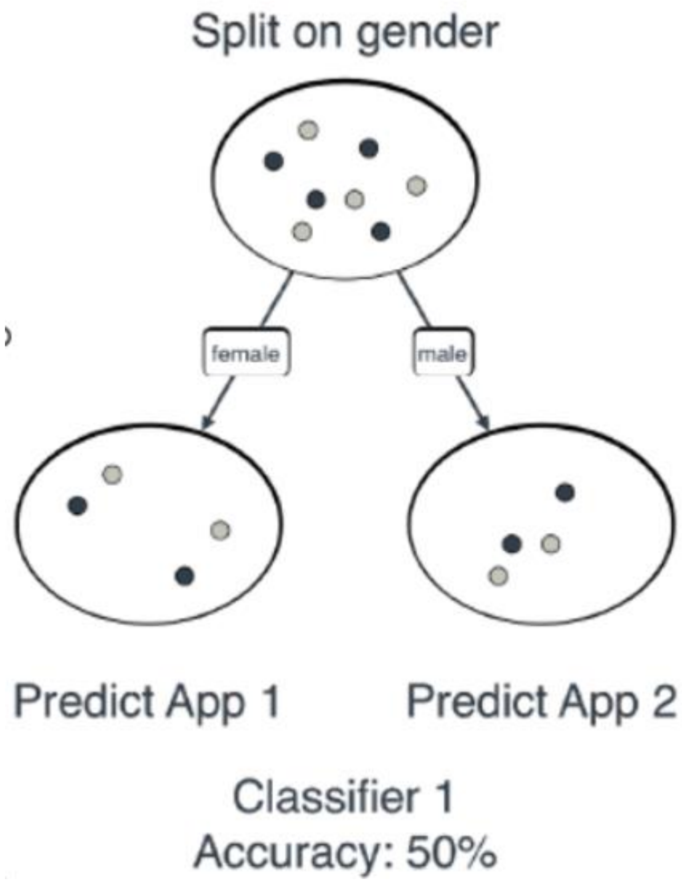
Agenda

- Accuracy
- Gini Impurity Index
- Recommendation System Application

Accuracy

Classifier 1: **What is your gender**?

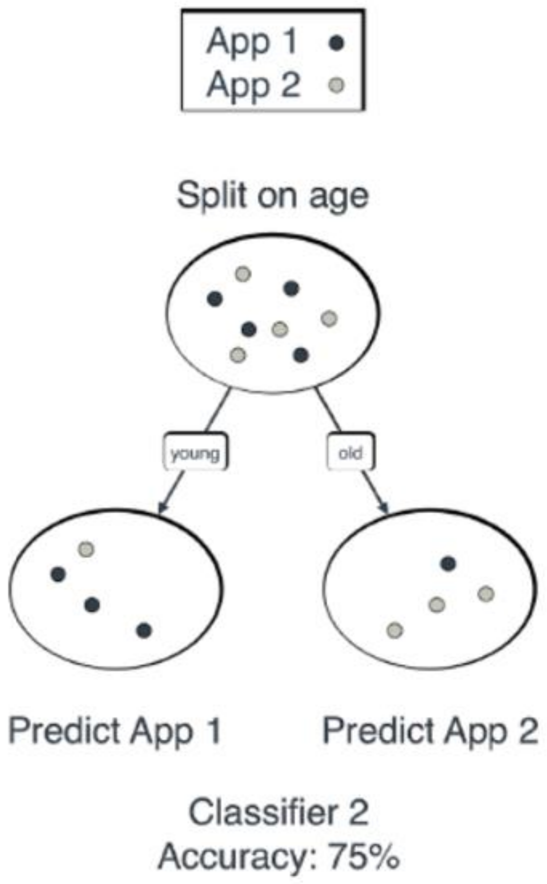
- The classifier was correct four times, and wrong four times. Therefore, the accuracy of this classifier is ½, or 50%.



Accuracy

<u>Classifier 2</u>: **What is your age**?

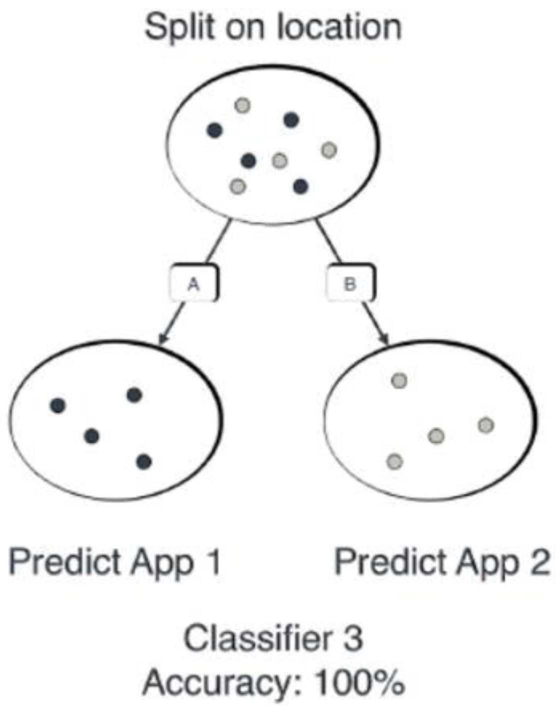
- Among our data, the classifier was correct six times, and wrong two times. Therefore, the accuracy of this classifier is ¾, or 75%.



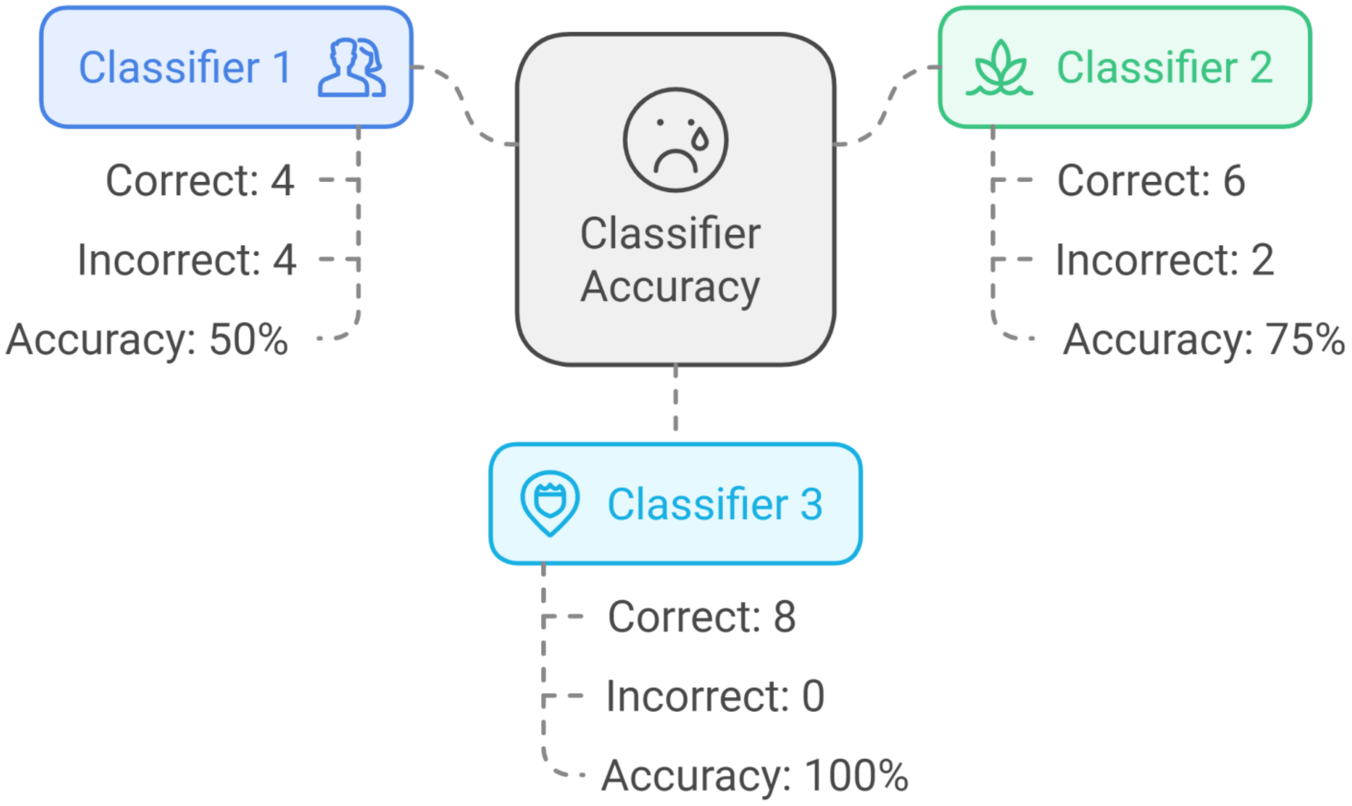
Accuracy

<u>Classifier 3</u>: What is your location?

- Well, among our data, it is correct eight times out of eight. Therefore, its accuracy is 100%.

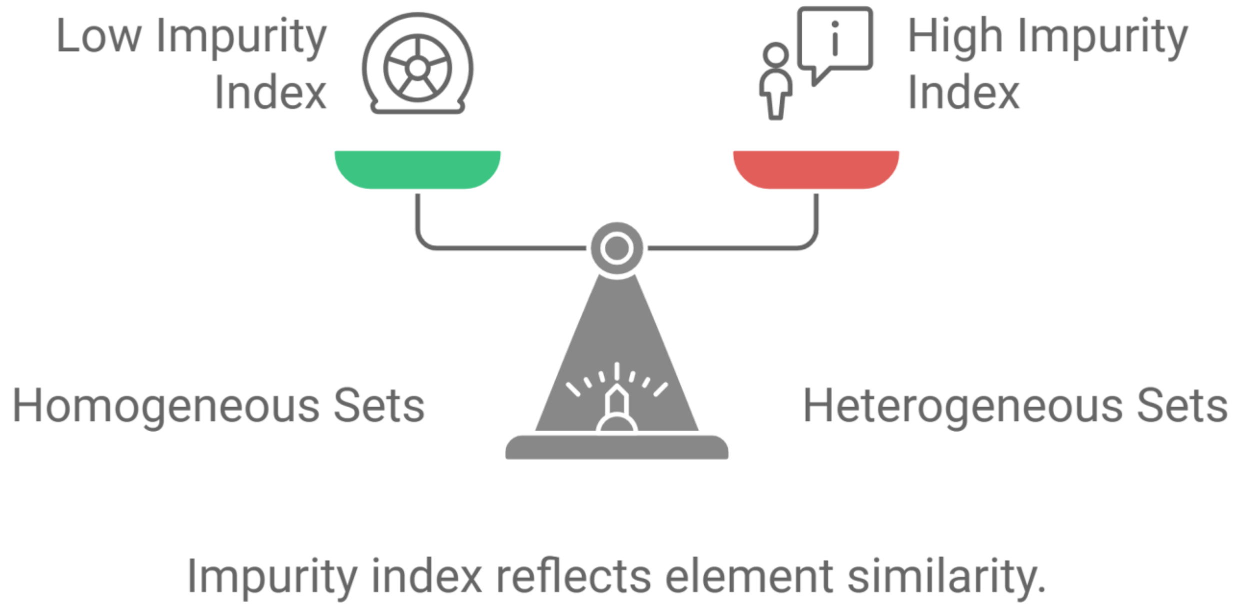


Accuracy


Classifier 1
Correct: 4
Incorrect: 4
Accuracy: 50%
Classifier Accuracy
Classifier 2
Correct: 6
Incorrect: 2
Accuracy: 75%
Classifier 3
Correct: 8
Incorrect: 0
Accuracy: 100%

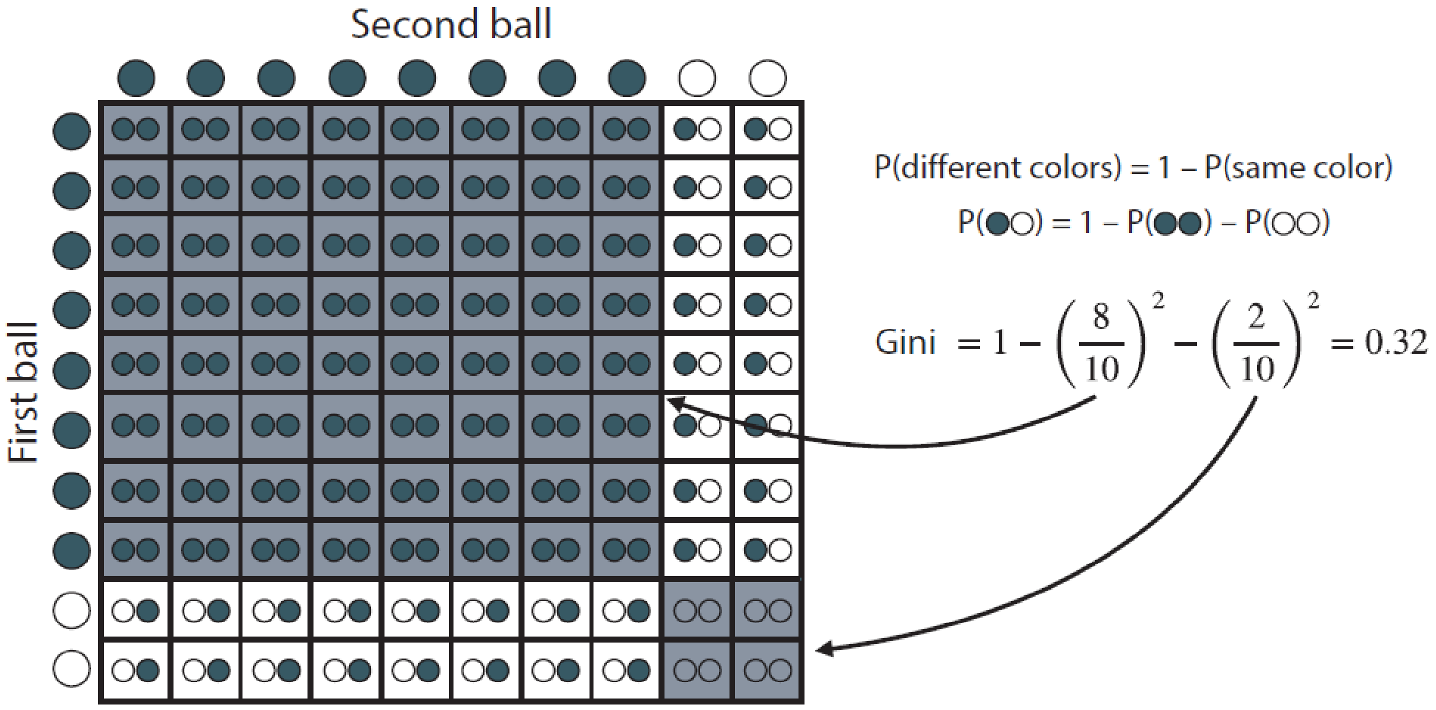
Gini Impurity

- The Gini index is a measure of diversity in a dataset.



Gini Impurity

- Note that if the total area of the square is 1, the probability of picking two black balls is 0.8^2, and the probability of picking two white balls is 0.2^2. the **probability of picking two balls of a different color** is the remaining area, which is $1 - 0.8^2 - 0.2^2 = 0.32$. That is the **Gini index**



Gini Impurity

- The **Gini Index**, also known as **Gini Impurity**, is a metric used to evaluate the quality of a split in decision trees.
- It measures the degree of impurity or heterogeneity in a dataset.
- In a set with m elements and n classes, with a_i elements belonging to the $i-th$ class, the Gini impurity index is

$$\text{Gini} = 1 - {p_1}^2 - {p_2}^2 - {p_3}^2 \ldots {p_n}^2$$

- $p_i = \frac{a_i}{m}$

Gini Impurity

- **Set 1**: {red, red, red, red, red, red, red, red, blue, blue} (eight red balls, two blue balls)

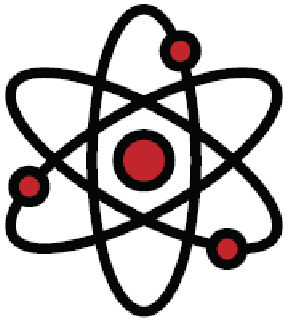
$$\text{Gini impurity index} = 1 - \left(\frac{8}{10}\right)^2 - \left(\frac{2}{10}\right)^2 = \frac{32}{100} = 0.32$$

- **Set 2**: {red, red, red, red, blue, blue, blue, yellow, yellow, green}

$$\text{Gini impurity index} = 1 - \left(\frac{4}{10}\right)^2 - \left(\frac{3}{10}\right)^2 - \left(\frac{2}{10}\right)^2 - \left(\frac{1}{10}\right)^2 = \frac{70}{100} = 0.7$$

Gini Impurity

- Build a system that recommends to users **which app to download** among the following options
  - **Atom Count**: an app that counts the number of atoms in your body
  - **Beehive Finder**: an app that maps your location and finds the closest beehives
  - **Check Mate**: an app for finding Australian chess players



Atom Count

Beehive Finder

Check Mate Mate

Gini Impurity

- A dataset with users of an app store

Platform	Age	App
iPhone	15	Atom Count
iPhone	25	Check Mate Mate
Android	32	Beehive Finder
iPhone	35	Check Mate Mate
Android	12	Atom Count
Android	14	Atom Count

Gini Impurity

- A simplified version of the dataset

Platform	Age	App
iPhone	Young	Atom Count
iPhone	Adult	Check Mate Mate
Android	Adult	Beehive Finder
iPhone	Adult	Check Mate Mate
Android	Young	Atom Count
Android	Young	Atom Count

Platform	Age	App
iPhone	15	Atom Count
iPhone	25	Check Mate Mate
Android	32	Beehive Finder
iPhone	35	Check Mate Mate
Android	12	Atom Count
Android	14	Atom Count

Gini Impurity

- How do we use the Gini index to decide which of the two ways to split the data (**age or platform**) is better? Clearly, if we can split the data into two **purer datasets**, we have performed a better split.
- Classifier 1 (by **platform**):
  - Left leaf (iPhone): {A, C, C}
  - Right leaf (Android): {A, A, B}
- Classifier 2 (by **age**):
  - Left leaf (young): {A, A, A}
  - Right leaf (adult): {B, C, C}

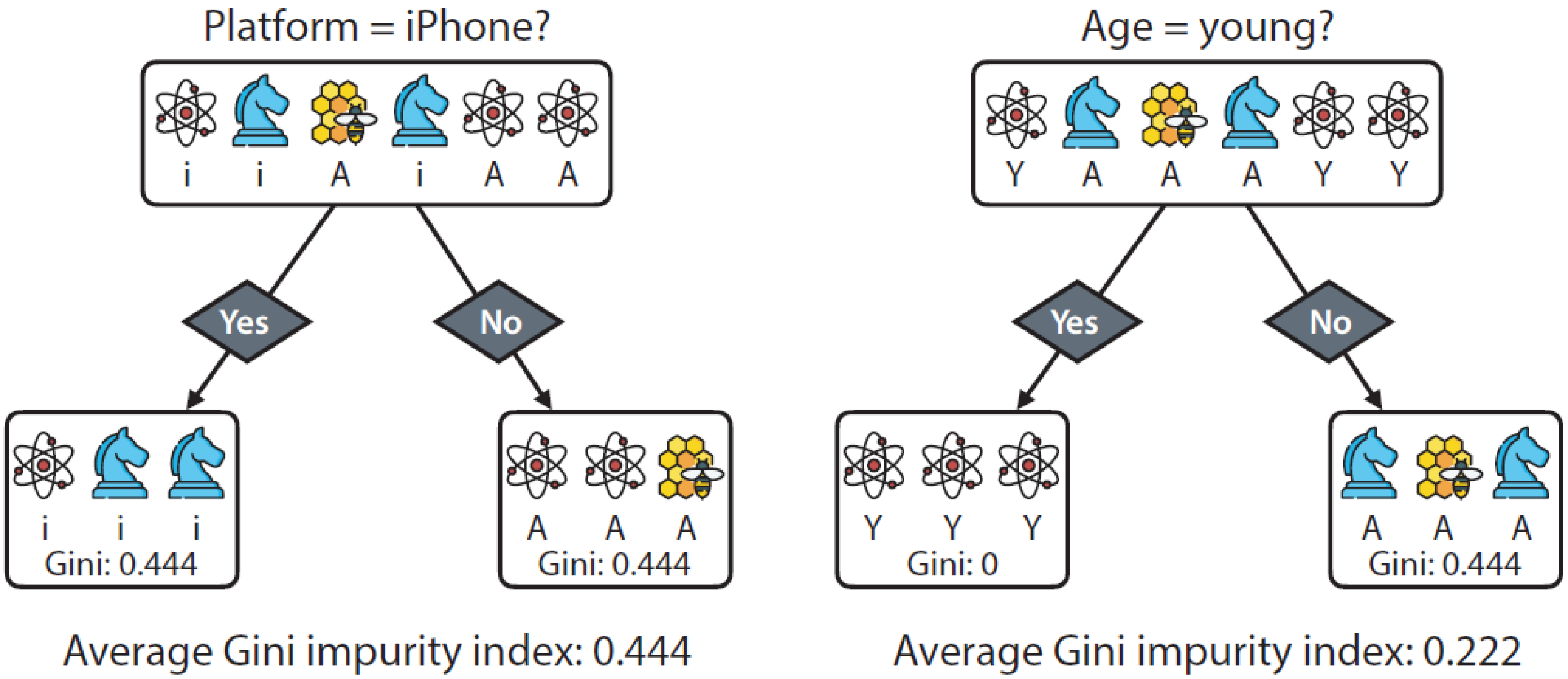
Platform	Age	App
iPhone	Young	Atom Count
iPhone	Adult	Check Mate Mate
Android	Adult	Beehive Finder
iPhone	Adult	Check Mate Mate
Android	Young	Atom Count
Android	Young	Atom Count

Gini Impurity

- Classifier 1 (by **platform**):
  - Left leaf (iPhone): {A, C, C} → Gini = $1 - \left(\frac{1}{3}\right)^2 - \left(\frac{2}{3}\right)^2 = \frac{4}{9} = 0.444$
  - Right leaf (Android): {A, A, B} → Gini = $1 - \left(\frac{1}{3}\right)^2 - \left(\frac{2}{3}\right)^2 = \frac{4}{9} = 0.444$
  - Average Gini index = 0.444
- Classifier 2 (by **age**):
  - Left leaf (young): {A, A, A} → Gini = 0
  - Right leaf (adult): {B, C, C} → Gini = $1 - \left(\frac{1}{3}\right)^2 - \left(\frac{2}{3}\right)^2 = \frac{4}{9} = 0.444$
  - Average Gini index = $\frac{1}{2}(0 + 0.444) = \mathbf{0.222}$
  - We conclude that the **second split is better**, because it has a lower average Gini index

Gini Impurity

- We conclude that the second split is better, because it has a **lower average Gini index**



Practical Example

- Building a decision tree to classify whether a person buys a product based on age.

- Dataset Split:
  - Before Split:
    - 40 Buy, 60 Not Buy → Gini = 1 - (0.4^2 + 0.6^2) = 0.48
  - After Split on Age ≤ 30:
    - **Left Node**: 30 Buy, 10 Not Buy → Gini = 1 - (0.75^2 + 0.25^2) = 0.375
    - **Right Node**: 10 Buy, 50 Not Buy → Gini = 1 - (0.166^2 + 0.833^2) ≈ 0.278
  - Weighted Gini:

$$\frac{40}{100} \times 0.375 + \frac{60}{100} \times 0.278 = 0.15 + 0.1668 = 0.3168$$

- **Interpretation**: The split on "Age ≤ 30" reduces the Gini Impurity from **0.48** to **0.3168**, indicating a better, purer division.