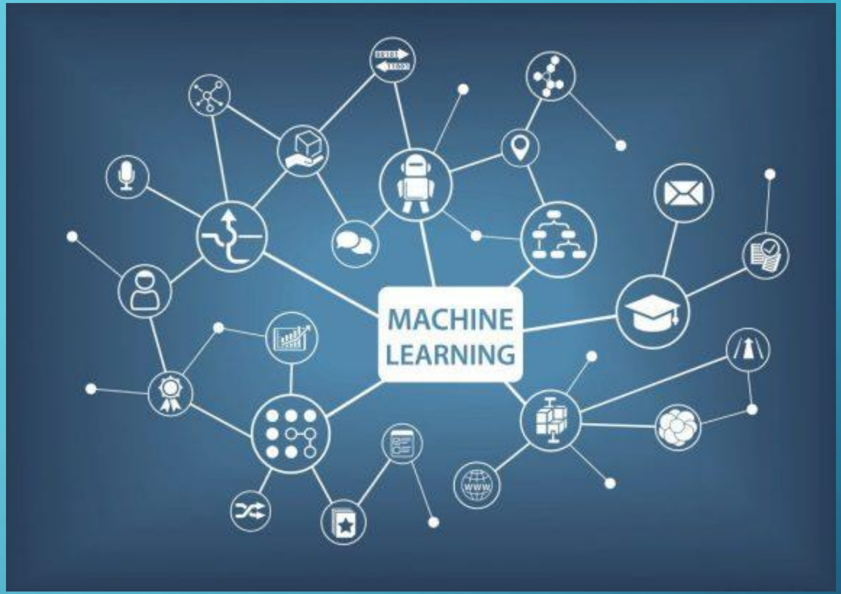
TC3020
SUPERVISED LEARNING
DECISION TREES



HÉCTOR G. CEBALLOS

CEBALLOS@TEC.MX

Image: https://www.mba-madrid.com/empresas/impacto-del-machine-learning-ambito-empresarial/

SUPERVISED CLASSIFICATION PROBLEM

- Supervised: Training data
  - Instance or feature vector
  - Attributes
  - Target attribute
- Classification: Target value is a category (not a number)

Attributes

Target attribute

Name	**Balance**	**Age**	**Employed**	**Write-off**
Mike	$200,000	42	no	yes
Mary	$35,000	33	yes	no
Claudio	$115,000	40	no	no
Robert	$29,000	23	yes	yes
Dora	$72,000	31	no	no

This is one row (example).
Feature vector is: **<Claudio,115000,40,no>**
Class label (value of Target attribute) is **no**

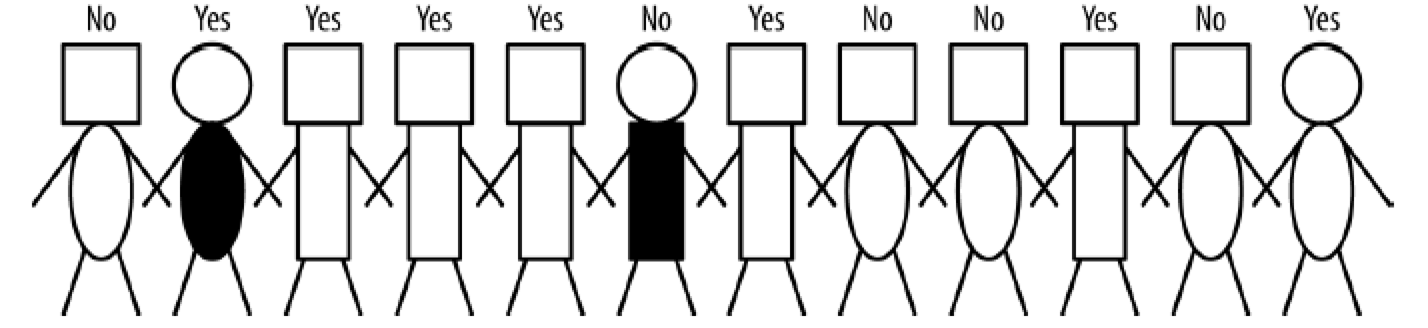
PREDICTIVE MODEL

- *Formula for estimating the unknown value of interest: the target*
  - Formula: mathematical, logical statement, hybrid
  - Prediction: estimation of an unknown (current, past or future) value.
- It is not a descriptive model (explain).
- It describes the relationship between some attributes and the target variable.

MODEL INDUCTION

- Creation of models from data.
- Generalization: specific cases -> general rules
- Induction algorithm or learner
    - Classification and Regression

SUPERVISED SEGMENTATION



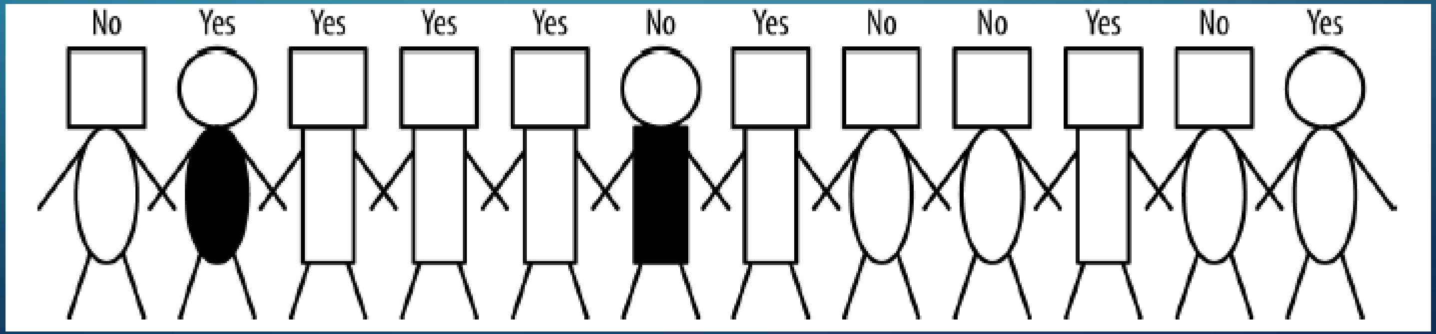
- *The label over each head represents the value of the target variable (write-off or not).*
- *Colors and shapes represent different predictor attributes.*

- Attributes:
  - head-shape: square, circular
  - body-shape: rectangular, oval
  - body-color: gray, white
- Target variable:
  - write-off: Yes, No

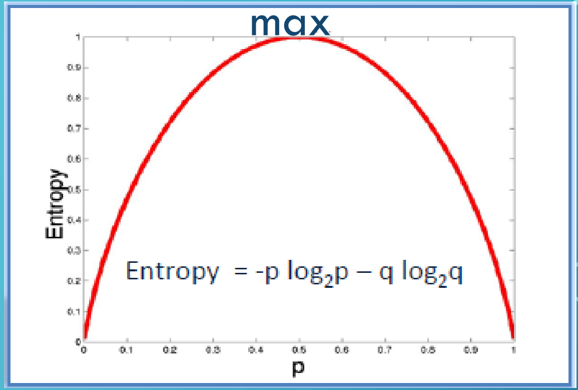
INFORMATIVE ATTRIBUTES

If every member of a group has the same value for the target, then the group is **pure.**

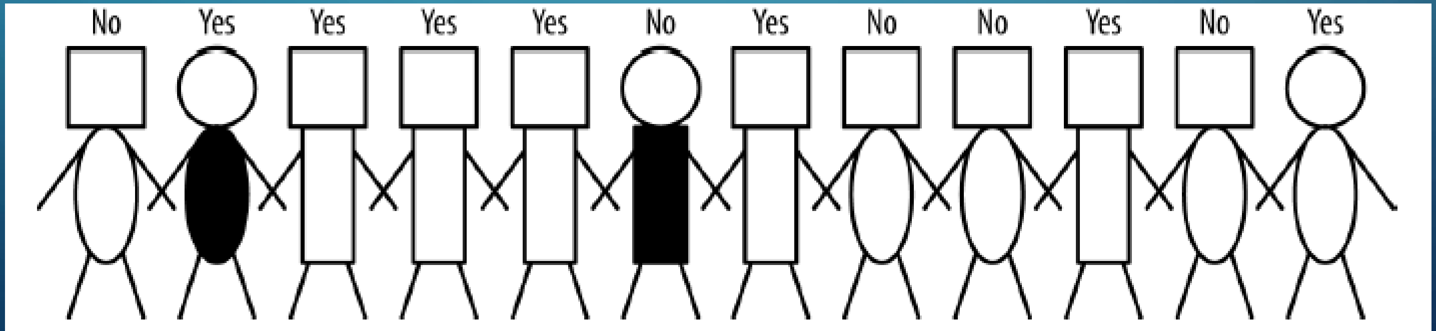
1. Attributes rarely split a group perfectly (*body-color=gray*)
2. Not all attributes are binary
3. Some attributes take on numeric values (continuous or integer).



ATTRIBUTE SELECTION FOR SPLIT



- Purity measure: formula that evaluates how well each attribute splits a set of examples into segments, w.r.t. target value.
- Splitting criteria
  - ***Entropy*** *is the measure of how homogeneous (0) or equally divided (1) is a sample.*
  - ***Information gain*** *is based on the decrease in entropy after a dataset is split on an attribute.*
- Which attribute splits the better our dataset?



- head-shape
- body-shape
- body-color

INFORMATION GAIN

Entropy favors smaller partitions with many distinct values.

$$E(S)=\sum_{i=1}^{c}-p_i\log_2 p_i$$

Play Golf	
Yes	No
9	5

Entropy(PlayGolf) = Entropy (5,9)
= Entropy (0.36, 0.64)
= - (0.36 $\log_2$ 0.36) - (0.64 $\log_2$ 0.64)
= 0.94

$$Gain(T,X)=Entropy(T)-Entropy(T,X)$$

G(PlayGolf, Outlook) = **E**(PlayGolf) – **E**(PlayGolf, Outlook)
= 0.940 – 0.693 = 0.247

$$E(T,X)=\sum_{c\in X}P(c)E(c)$$

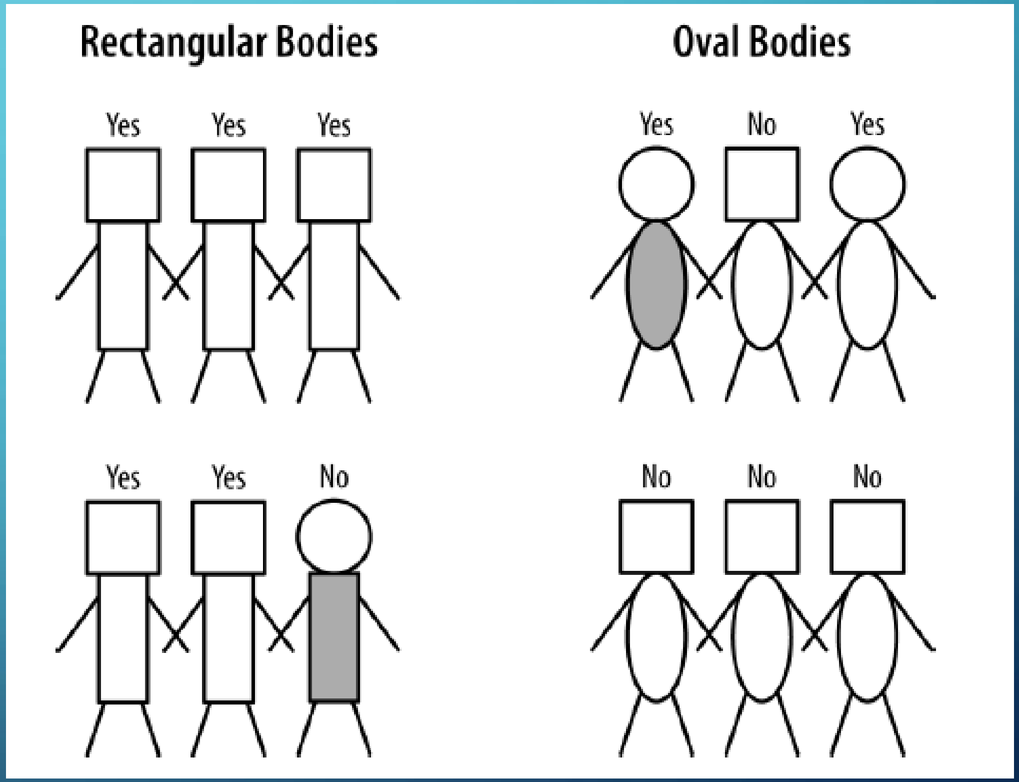
		Play Golf		
		Yes	No	
Outlook	Sunny	3	2	5
	Overcast	4	0	4
	Rainy	2	3	5
				14

E(PlayGolf, Outlook) = **P**(Sunny)***E**(3,2) + **P**(Overcast)***E**(4,0) + **P**(Rainy)***E**(2,3)
= (5/14)*0.971 + (4/14)*0.0 + (5/14)*0.971
= 0.693

FIRST PARTITIONING

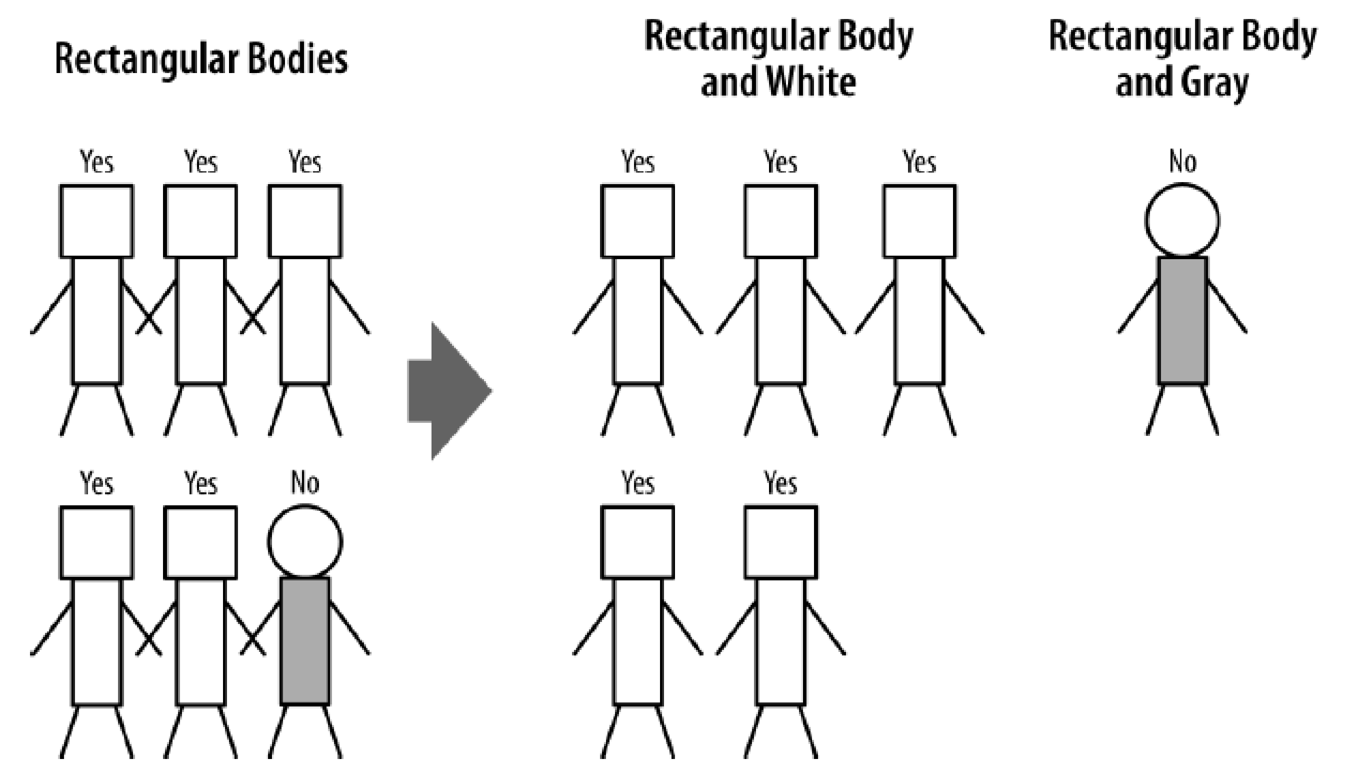
- *Splitting on body shape (rectangular versus oval).*

- *Which attribute splits the better to rectangular bodies?*



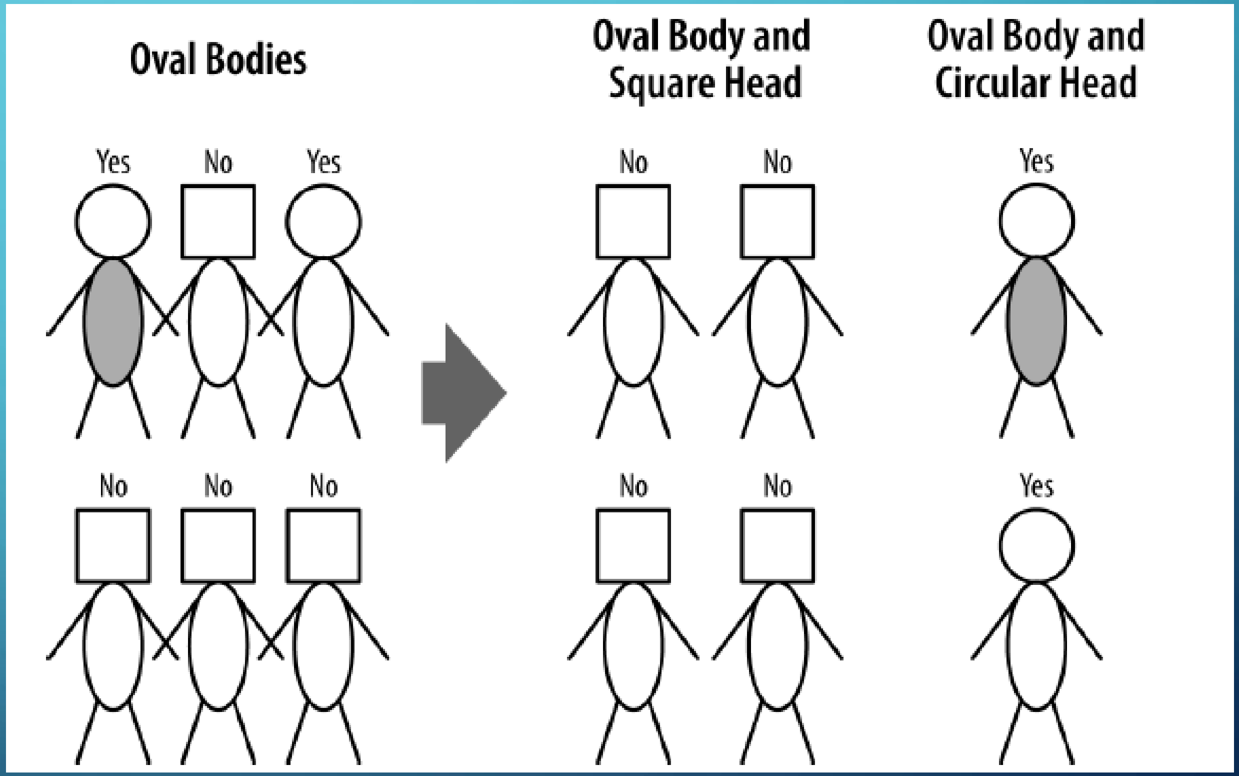
SECOND PARTITIONING

- *The rectangular body people subgrouped by body color*
- *Are subsets pure?*
- *Which attribute splits the better to circular bodies?*



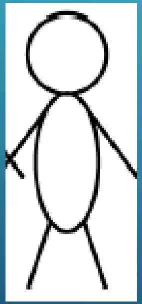
THIRD PARTITIONING

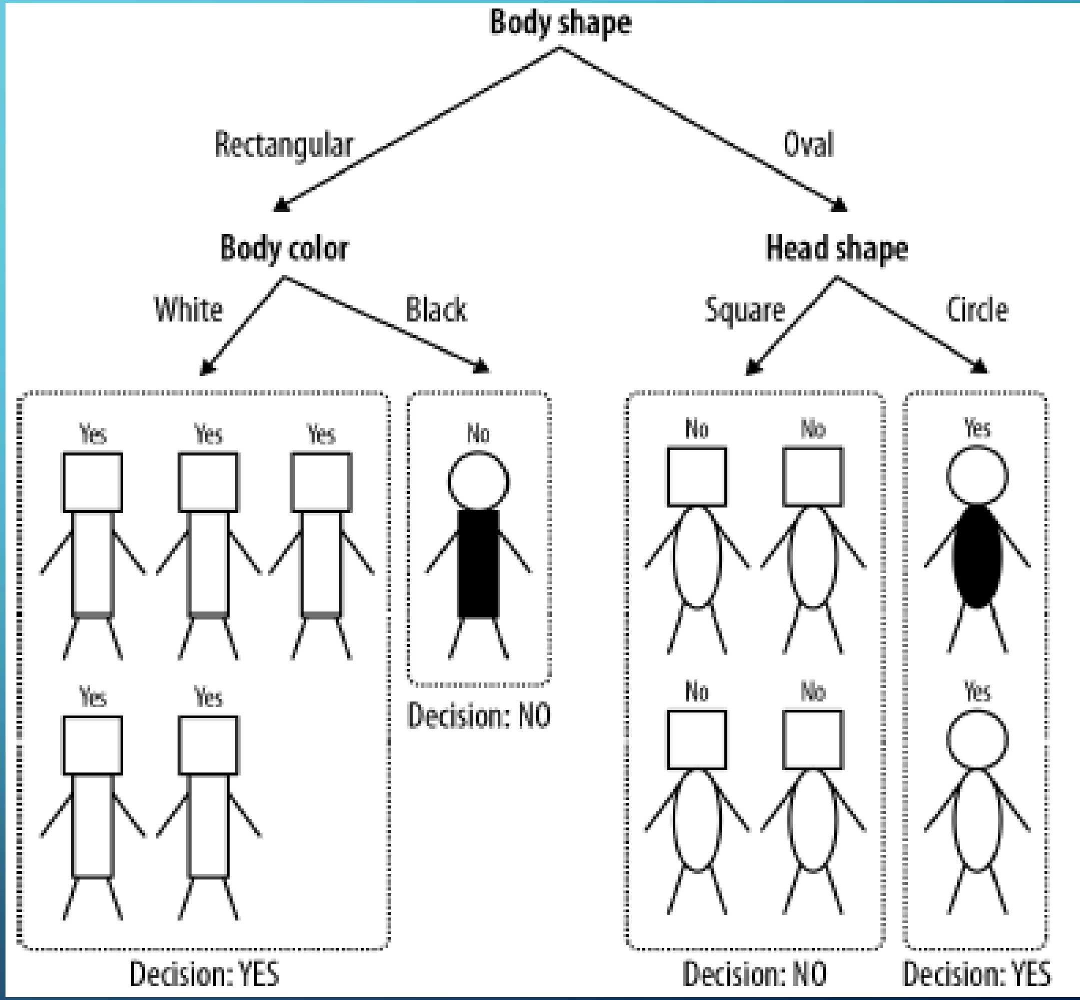
- *Second partitioning: the oval body people sub-grouped by head type*
- *Are subsets pure?*



CLASSIFICATION TREE

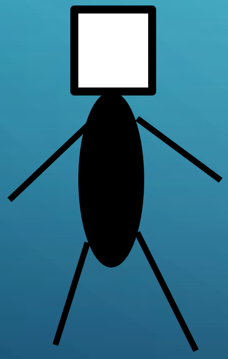
- *It establishes the attributes to evaluate and the order to do it.*
- *It can be use for classifying a new instance.*
- *Classiffy:*

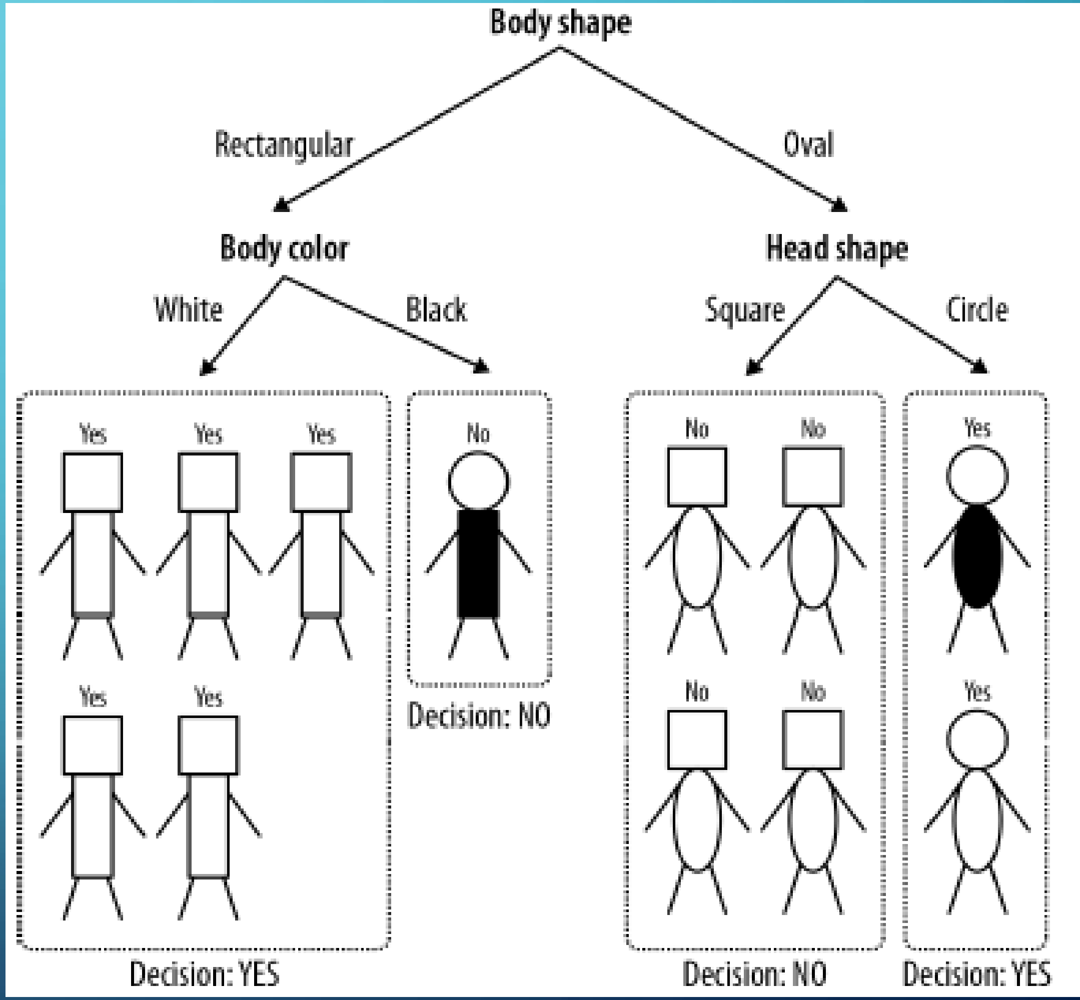




CLASSIFICATION TREE

- *It establishes the attributes to evaluate and the order to do it.*
- *It can be use for classifying a new instance.*
- *Classiffy:*





GINI INDEX

GINI index favors larger partitions.

$$Gini = 1 - \sum_{i=1}^{C} (p_i)^2$$

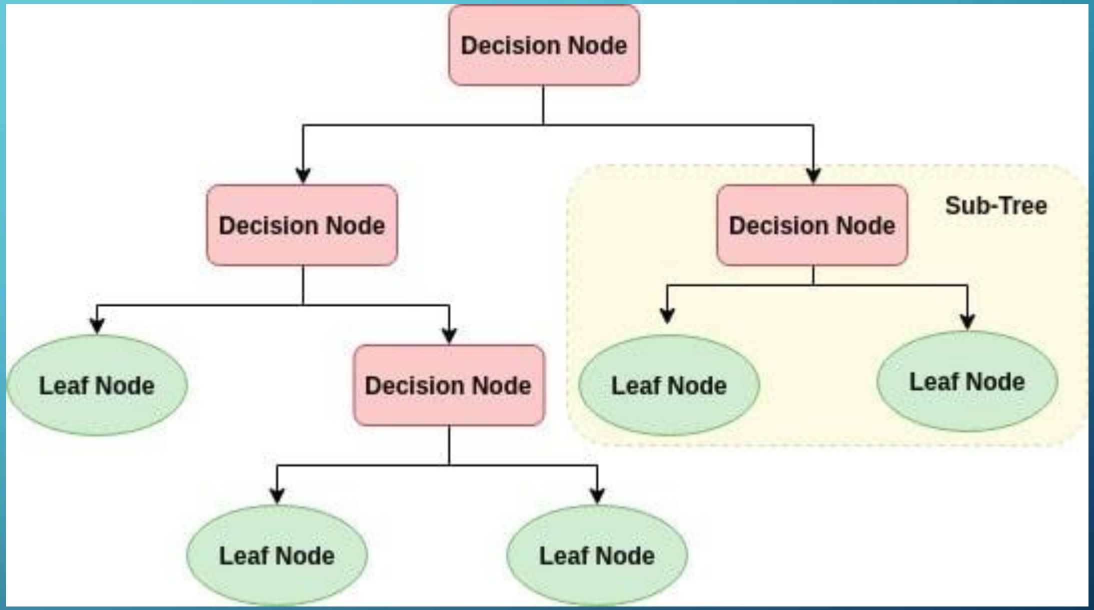
- Favors larger partitions.
- Uses squared proportion of classes.
- Perfectly classified, Gini Index would be zero.
- Evenly distributed would be 1 – (1/# Classes).
- You want a variable split that has a low Gini Index.
- The algorithm works as 1 – ($P(class_1)^2 + P(class_2)^2 + \ldots + P(class_N)^2$)

http://www.learnbymarketing.com/481/decision-tree-flavors-gini-info-gain/

DECISION TREES

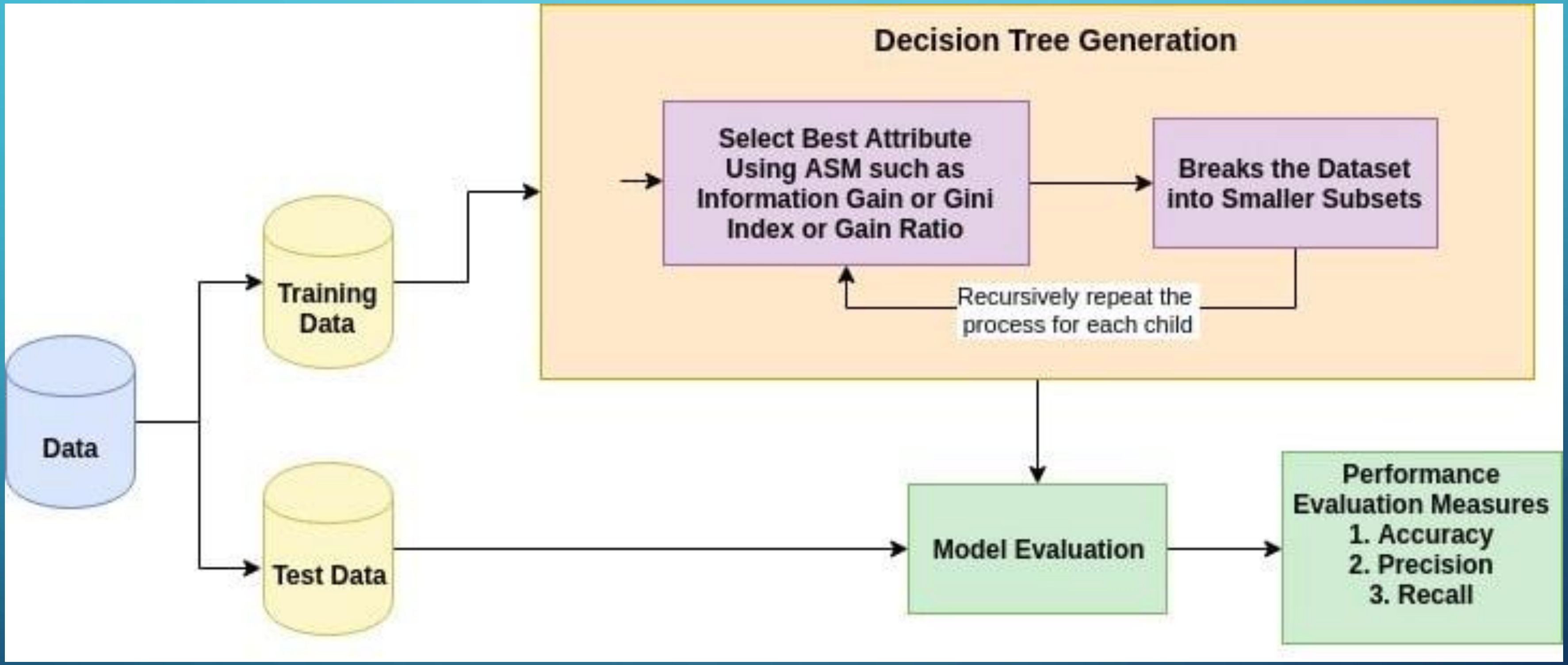
- A decision tree is a flowchart-like tree structure where
  - an internal node represents feature
  - the branch represents a decision rule
  - each leaf node represents the outcome.
- It learns to partition on the basis of the attribute value.
- It partitions the tree in recursively manner call recursive partitioning.

root

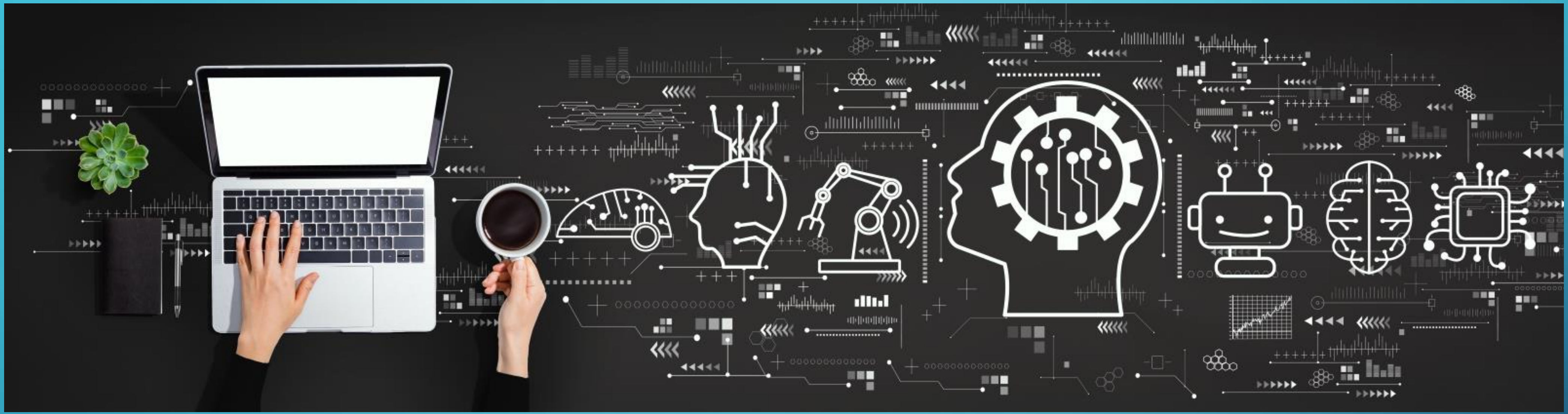


https://www.datacamp.com/community/tutorials/decision-tree-classification-python

HOW DOES IT WORK?



https://www.datacamp.com/community/tutorials/decision-tree-classification-python



HANDS ON …

DECISION TREES

Credits: https://expertsystem.com/machine-learning-definition/

PRACTICE

- Build a Decision Tree (DT) model to predict the onset of diabetes based on diagnostic measures.
  - **Pima Indians Diabetes Database**
  - https://www.kaggle.com/uciml/pima-indians-diabetes-database
- Train, evaluate and visualize a DT model

APPLICATION

DECISION TREES (PART 2)

https://ecaldima.com/machine-learning/

DT PROS

- Decision trees are easy to interpret and visualize.
- It can easily capture Non-linear patterns.
- It requires fewer data preprocessing from the user, for example, there is no need to normalize columns.
- It can be used for feature engineering such as predicting missing values, suitable for variable selection.
- The decision tree has no assumptions about distribution because of the non-parametric nature of the algorithm.

https://scikit-learn.org/stable/modules/tree.html

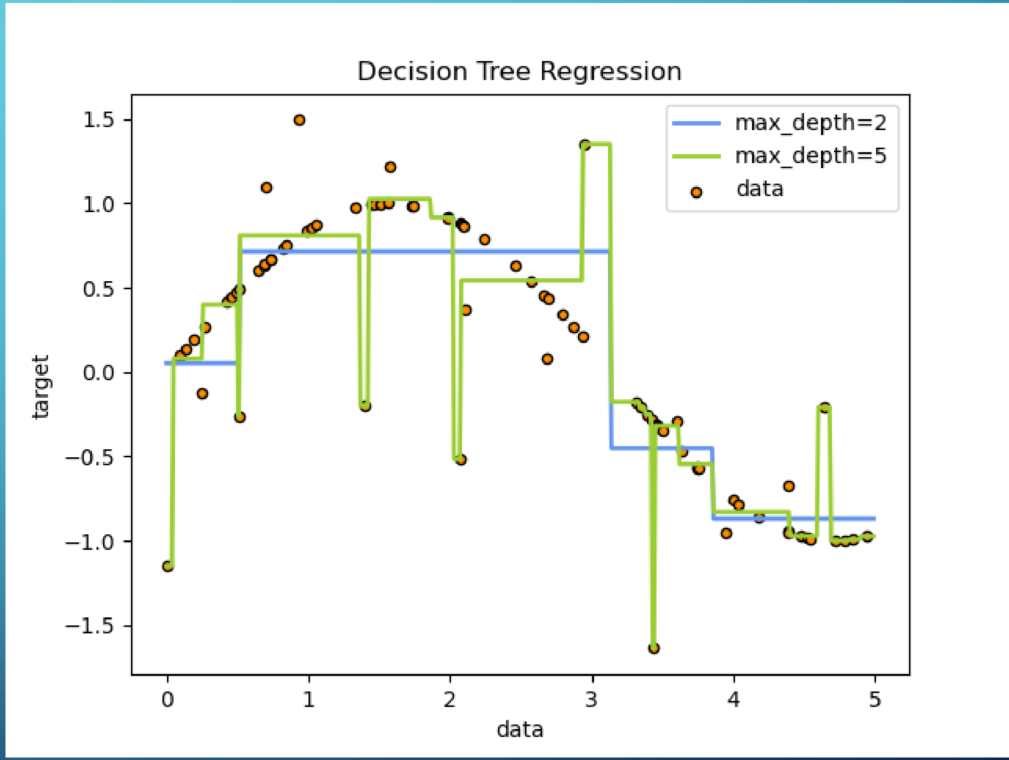
DT CONS

- Sensitive to noisy data. It can overfit noisy data.
- The small variation(or variance) in data can result in the different decision tree. This can be reduced by bagging and boosting algorithms.
- Decision trees are biased with imbalance dataset, so it is recommended that balance out the dataset before creating the decision tree.

https://scikit-learn.org/stable/modules/tree.html

WHAT KIND OF MODEL IT BUILDS?

- The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features.
- If the maximum depth of the tree is set too high, the decision trees learn too fine details of the training data and learn from the noise, i.e. they **overfit**.

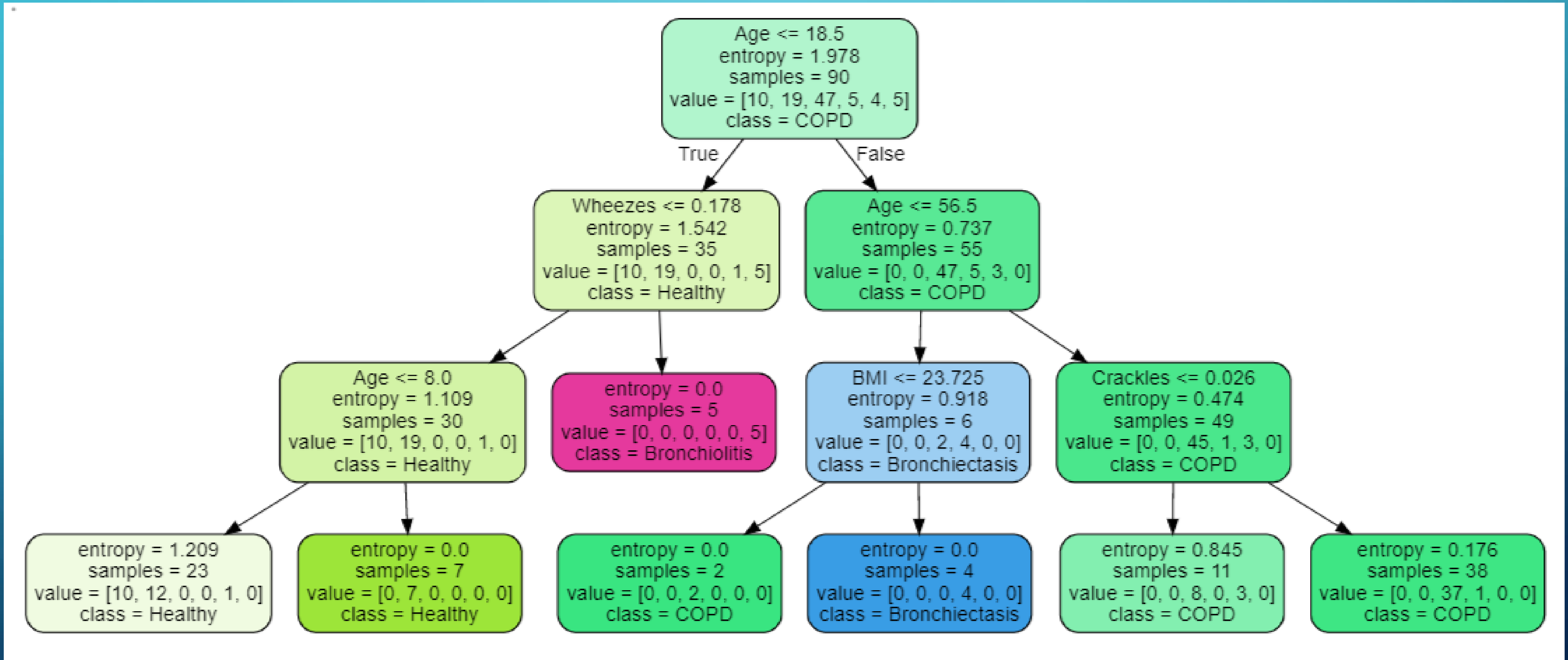


https://scikit-learn.org/stable/auto_examples/tree/plot_tree_regression.html

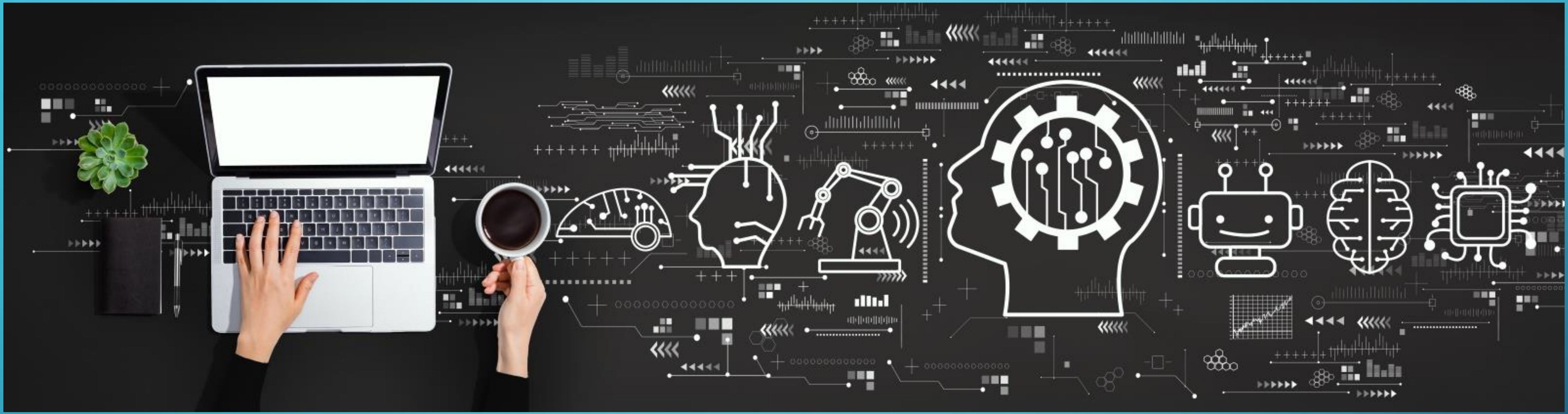
OPTIMIZING DECISION TREE PERFORMANCE

- **criterion : optional (default=”gini”) or Choose attribute selection measure:** This parameter allows us to use the different-different attribute selection measure. Supported criteria are “gini” for the Gini index and “entropy” for the information gain.
- **splitter : string, optional (default=”best”) or Split Strategy:** This parameter allows us to choose the split strategy. Supported strategies are “best” to choose the best split and “random” to choose the best random split.
- **max_depth : int or None, optional (default=None) or Maximum Depth of a Tree:** The maximum depth of the tree. If None, then nodes are expanded until all the leaves contain less than min_samples_split samples. The higher value of maximum depth causes overfitting, and a lower value causes underfitting (Source).

MULTI-CLASS CLASSIFICATION BY DECISION TREE



https://www.kaggle.com/gizemtanriver/multi-class-classification-by-decision-tree



HANDS ON …

DECISION TREE OPTIMIZATION

Credits: https://expertsystem.com/machine-learning-definition/

DECISION TREE OPTIMIZATION

- Learn and visualize a DT with max_depth = 3 and entropy criterion.
- Repeat using the gini criterion.

EXERCISE

- Answer in teams the following questions
  - Which is the best DT model with depth in [2-5] and given both information gain criterions (Entropy and GINI)?