Bachelor of Engineering in Information Technology

CTE309 Machine Learning

Unit IV: Classification

Mr. Yeshi Jamtsho
Lecturer

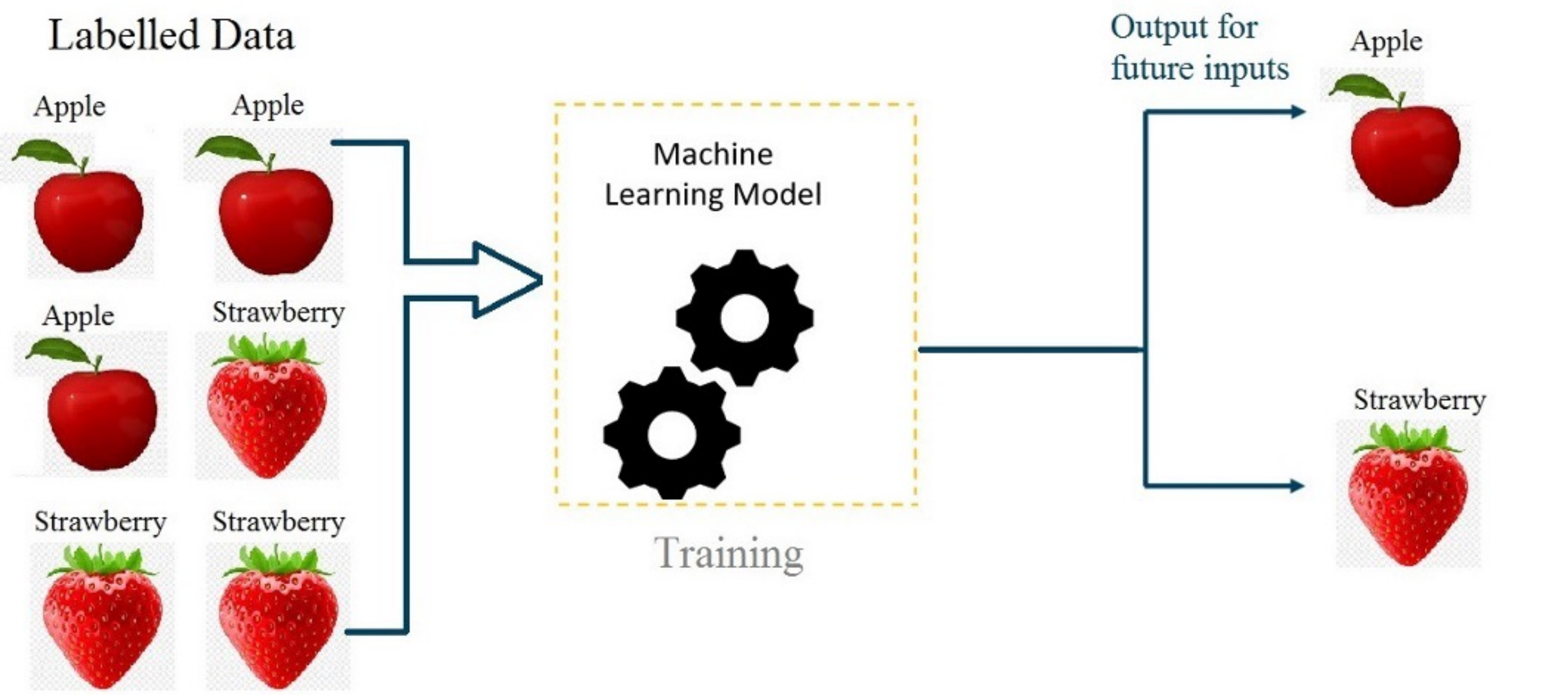
Overview

- Introduction
- Definition
- Applications
- Types of learners
- Classification tasks
- Classification algorithms

Introduction

- Supervised learning
- Choice between regression and classification
  - Predicting continuous value or category
- Considers the problem of identifying the categories of a data point on the basis of training data
- Example:
  - Fruit classification

Example

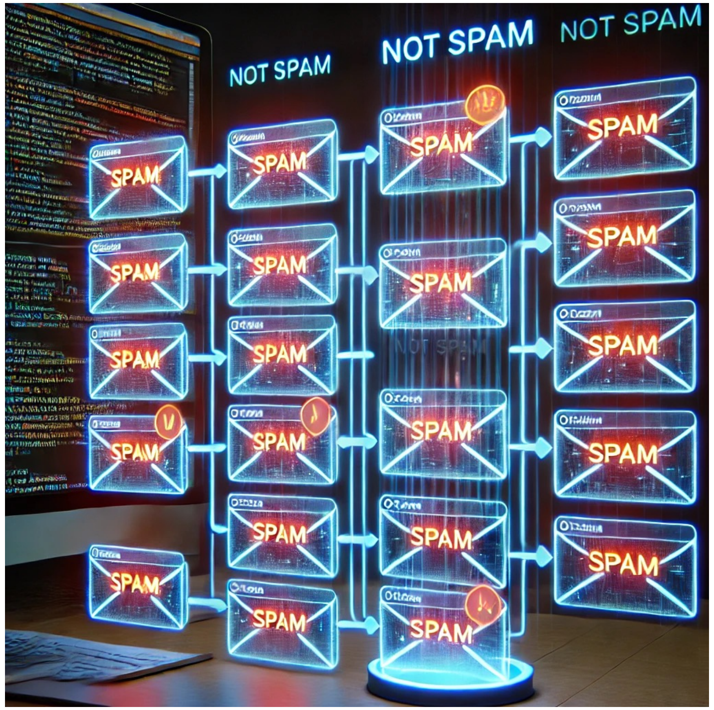

Labelled Data
Apple
Apple
Apple
Strawberry
Strawberry
Strawberry
Machine
Learning Model
Training
Output for
future inputs
Apple
Strawberry

Definition

Machine learning technique to identify the category of new observations based on the training dataset.

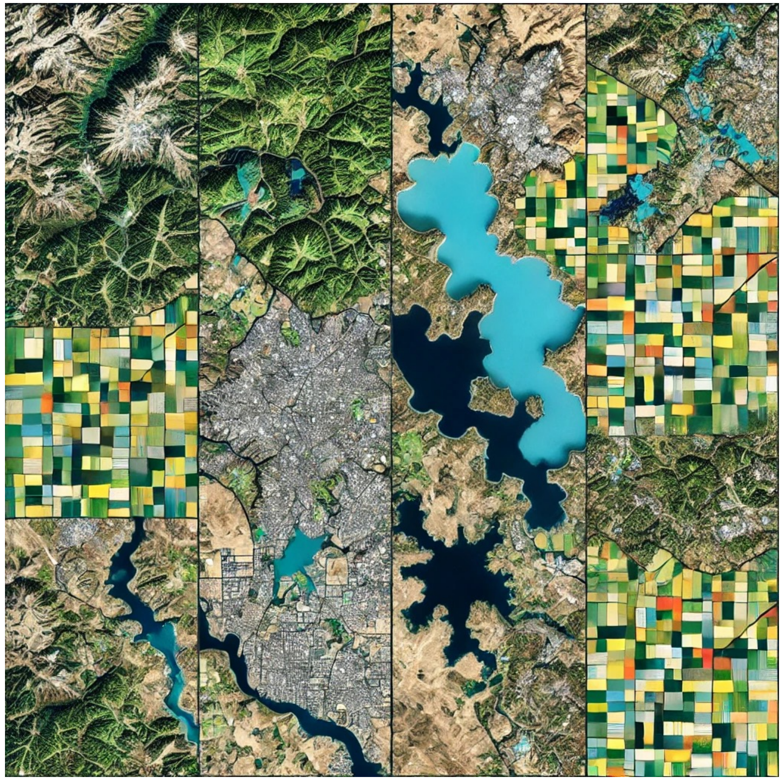
A supervised machine learning method where the model tries to predict the correct label of a given input data.

Classification Applications



E-mail category

sentiment category



Land cover category

Learner types

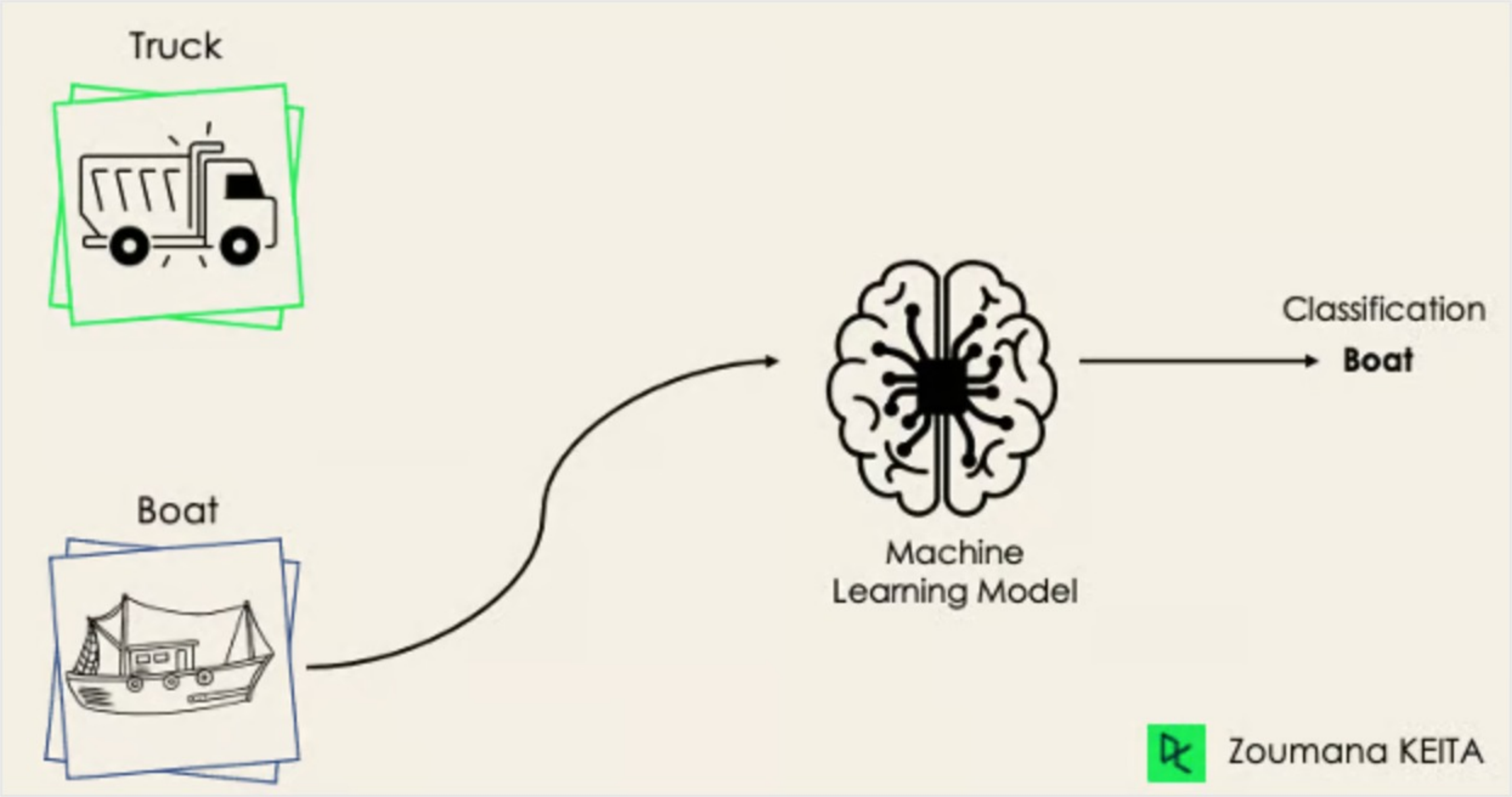
1. Eager learners

- first build a model from the training dataset before making any prediction on future datasets.
- spend more time during the training process
- but they require less time to make predictions.

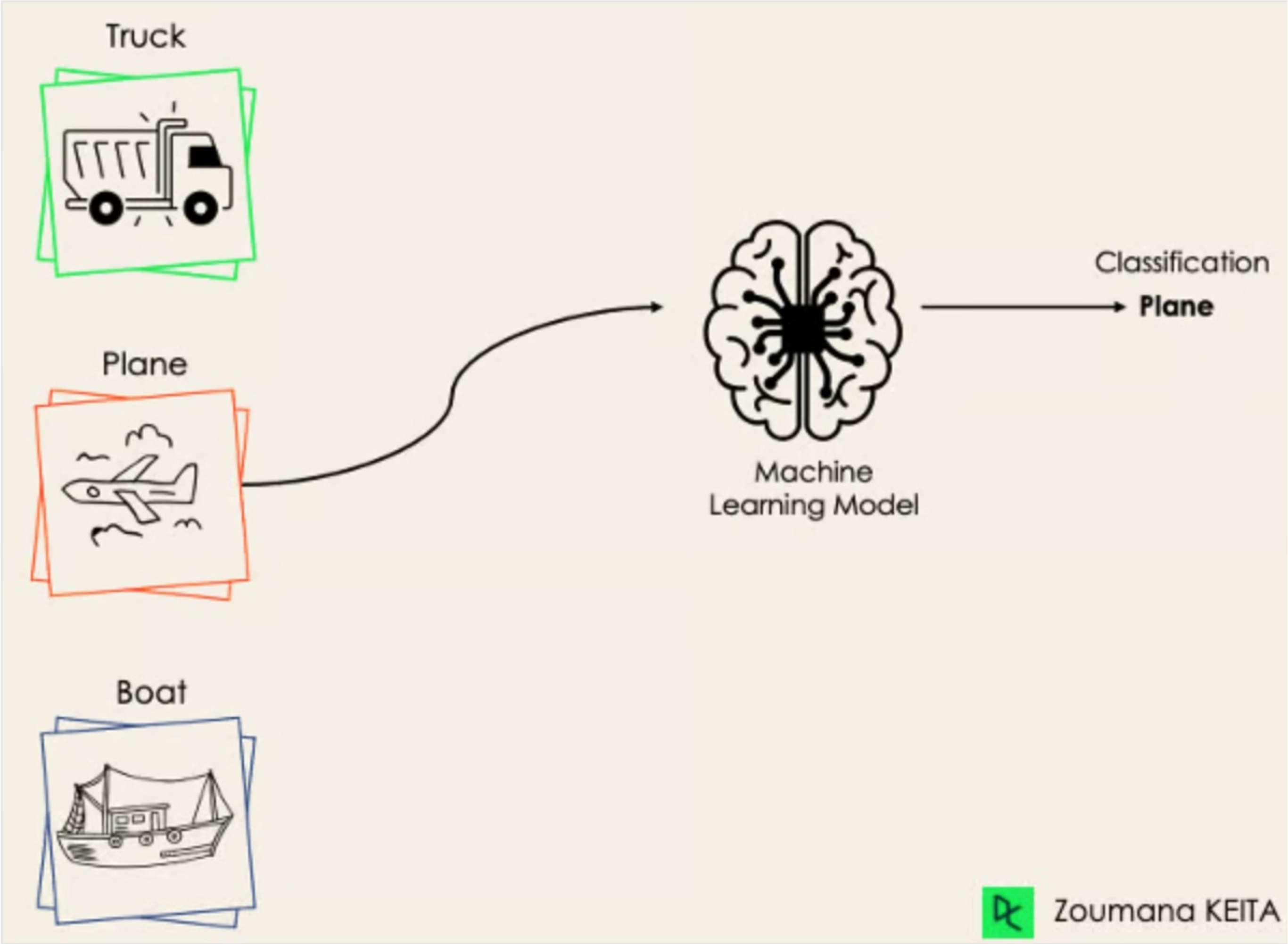
2. Lazy Learners or instance based learner

- do not create any model immediately from the training data
- just memorize the training data, and
- each time there is a need to make a prediction, they search for the nearest neighbor from the whole training data,
- Very slow during prediction.

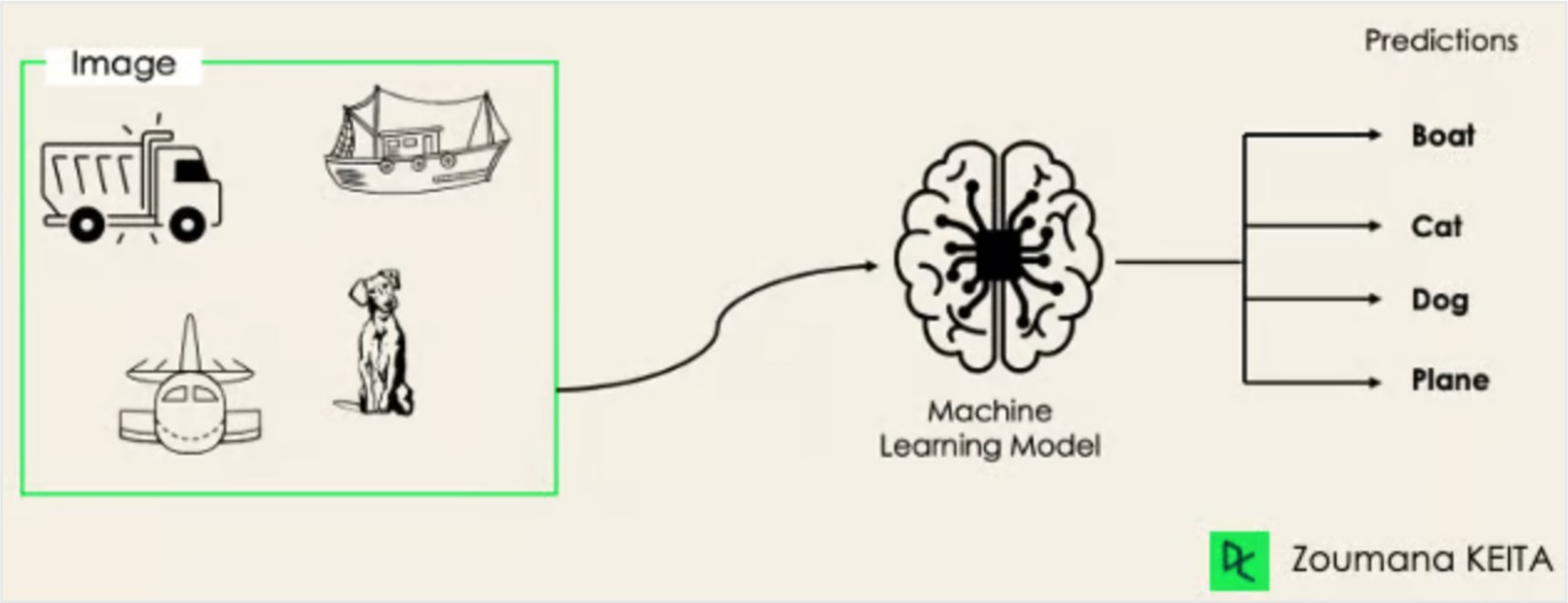
Classification task: Binary Classification



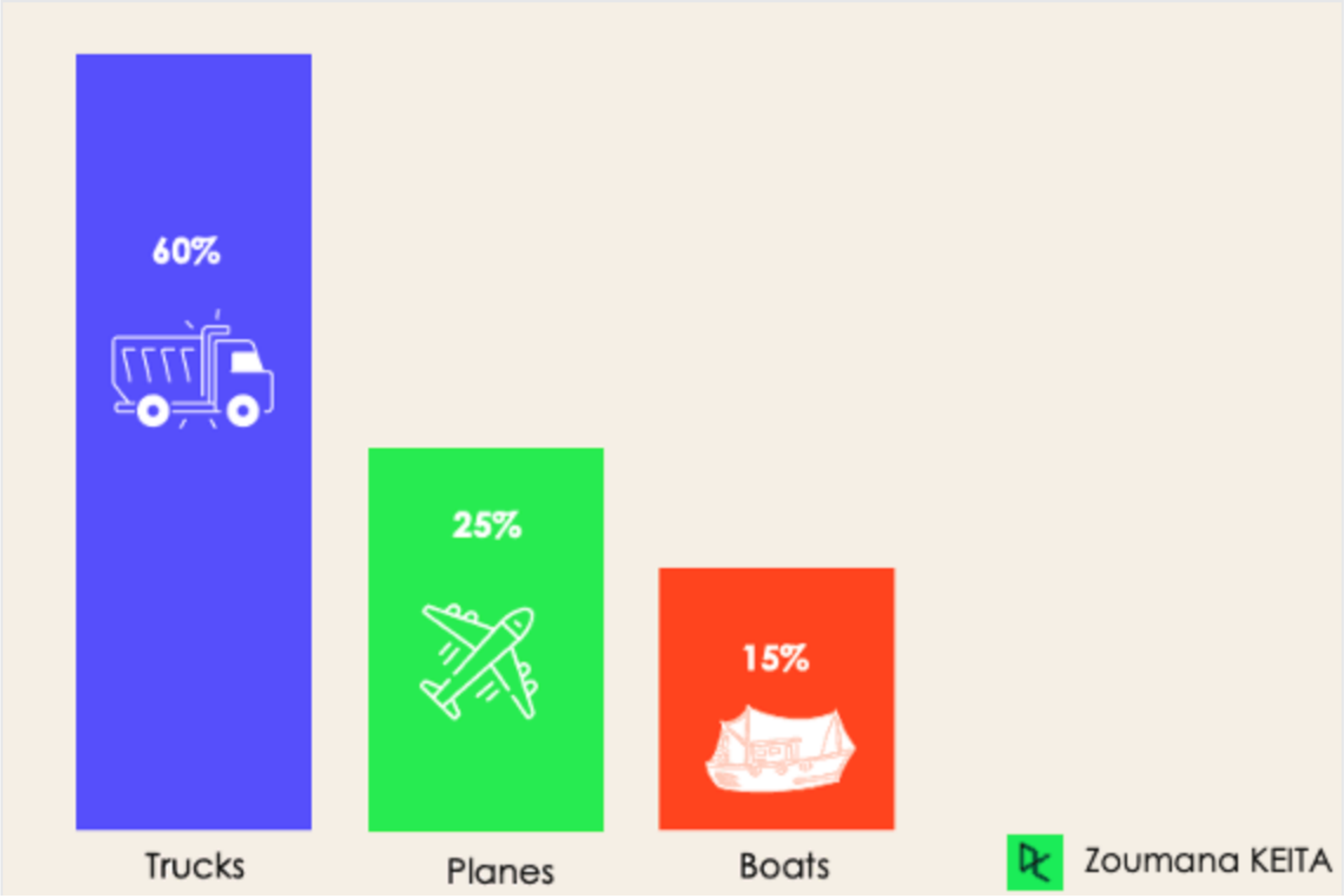
Classification task: Multi-class Classification



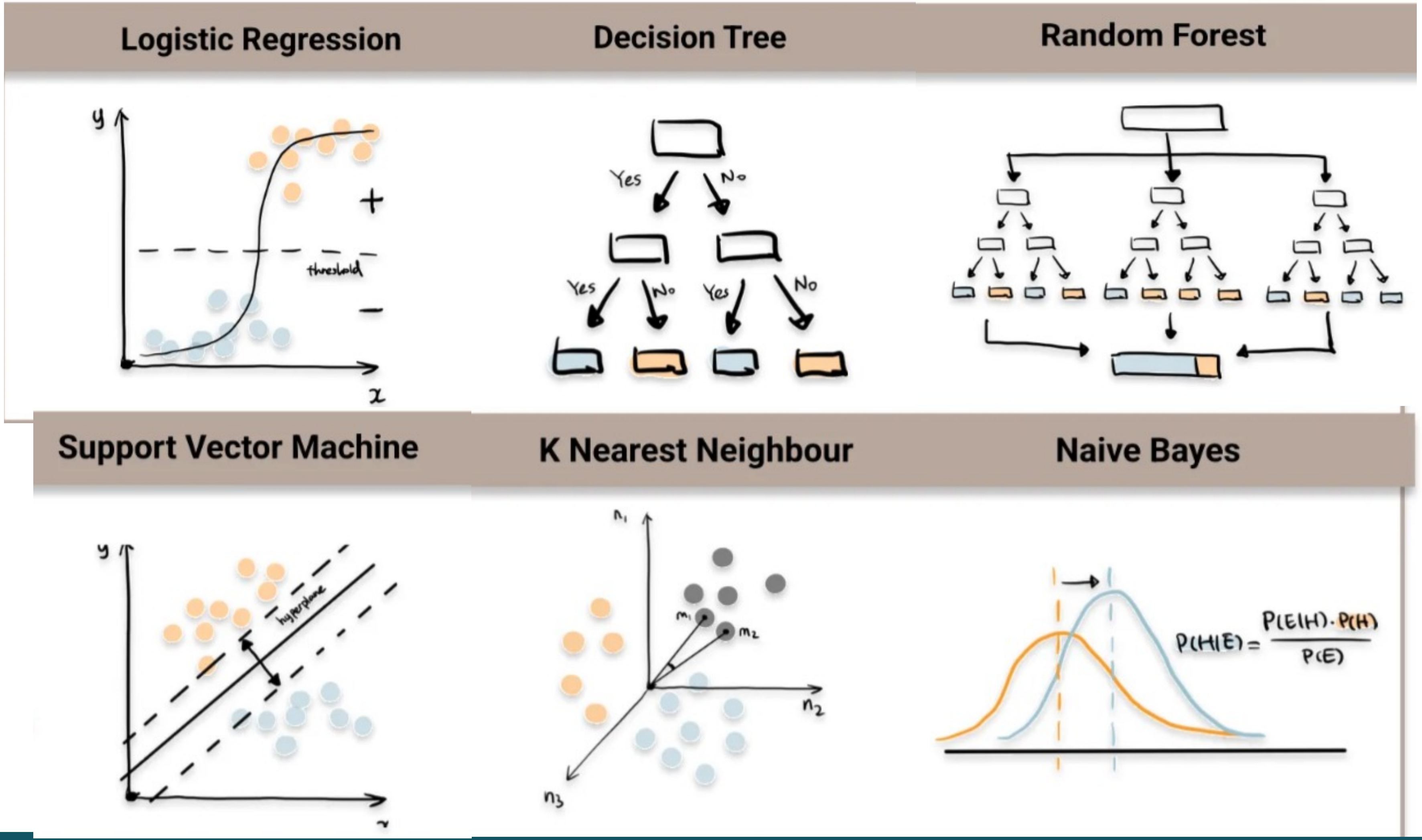
Classification task: Multi-label Classification



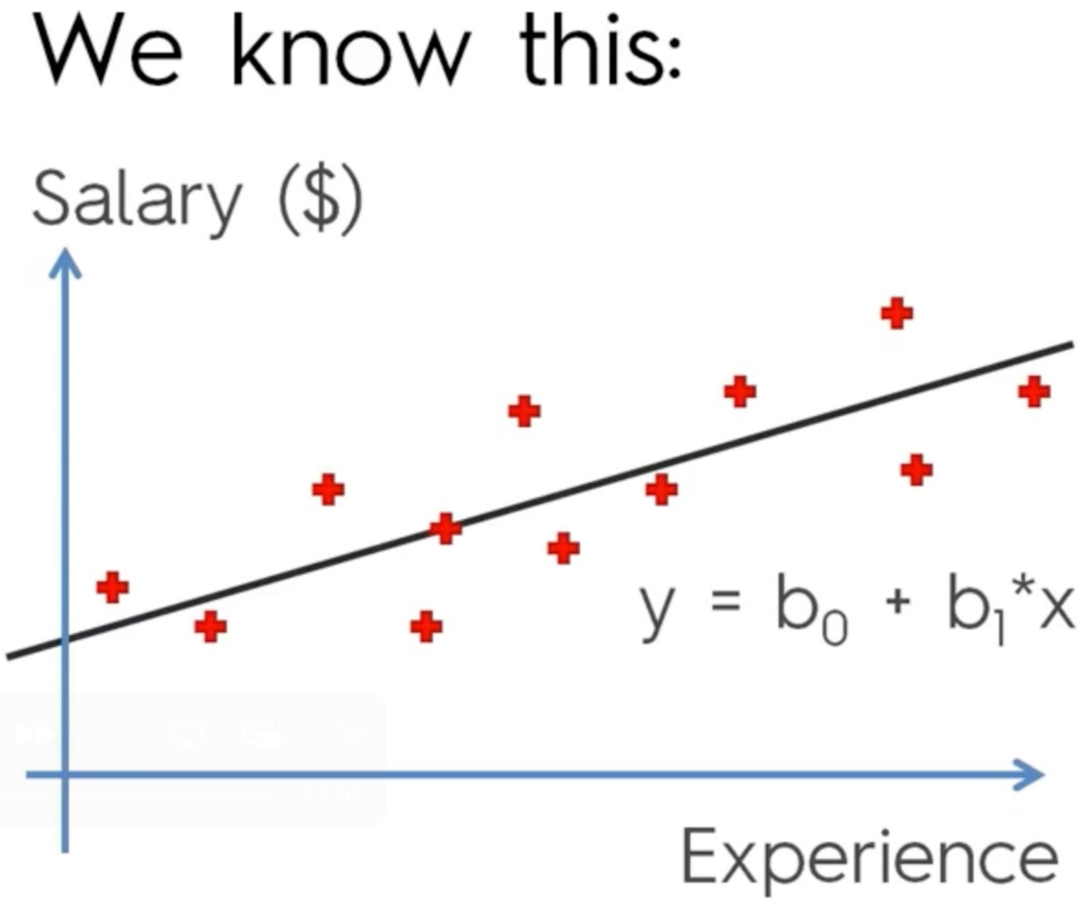
Classification task: Imbalance Classification

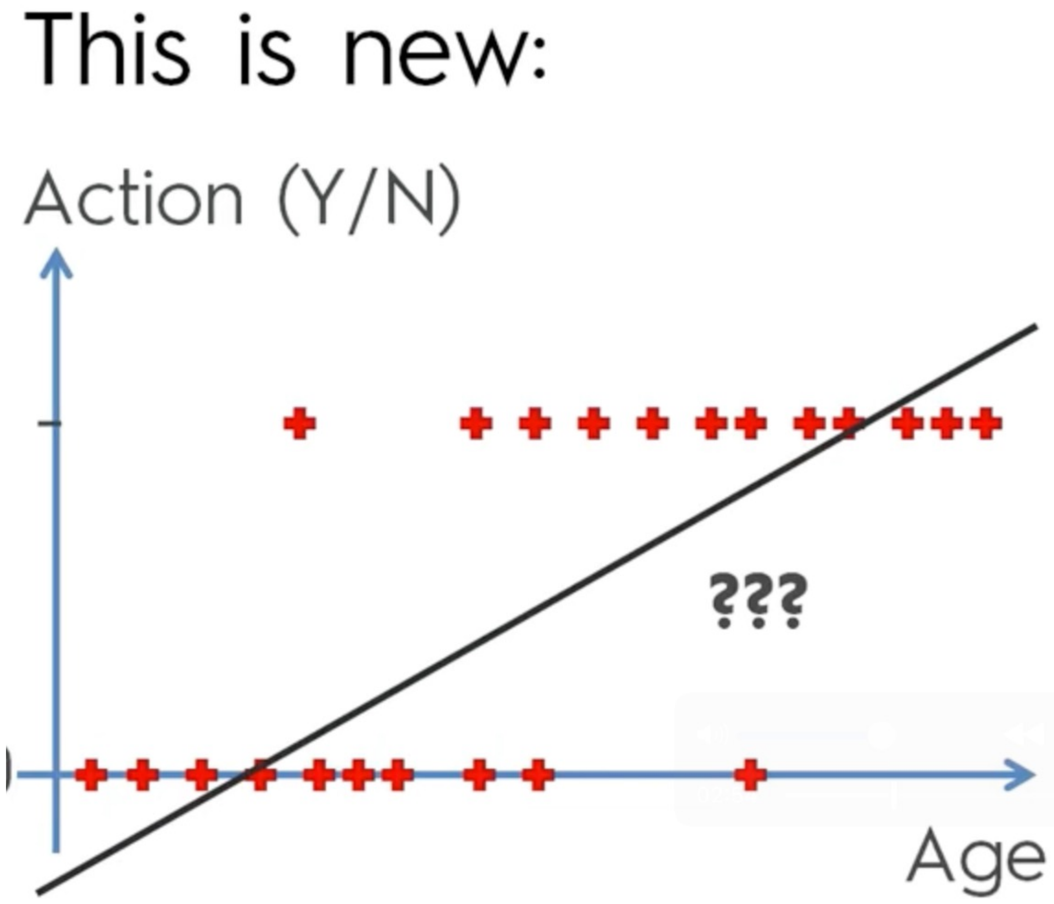


Classification Algorithms

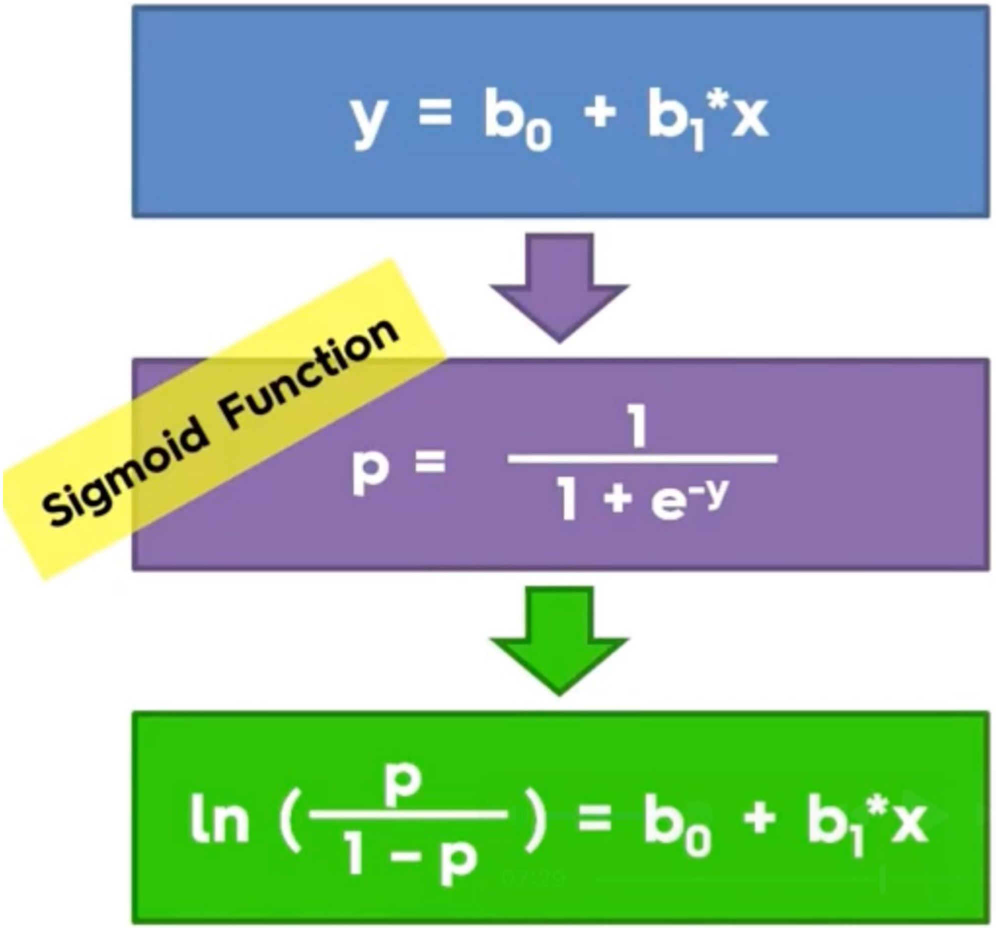


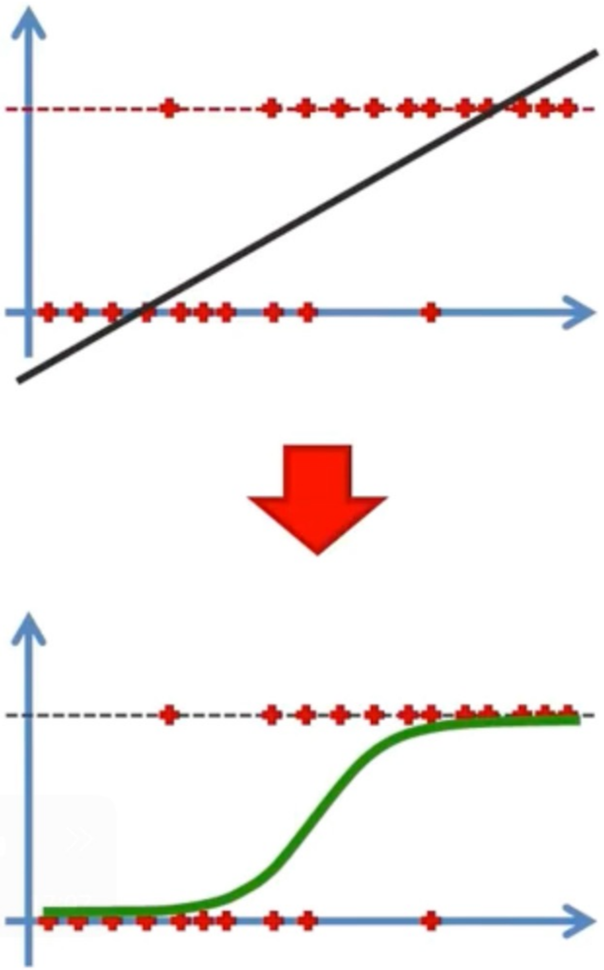
Logistic Regresion



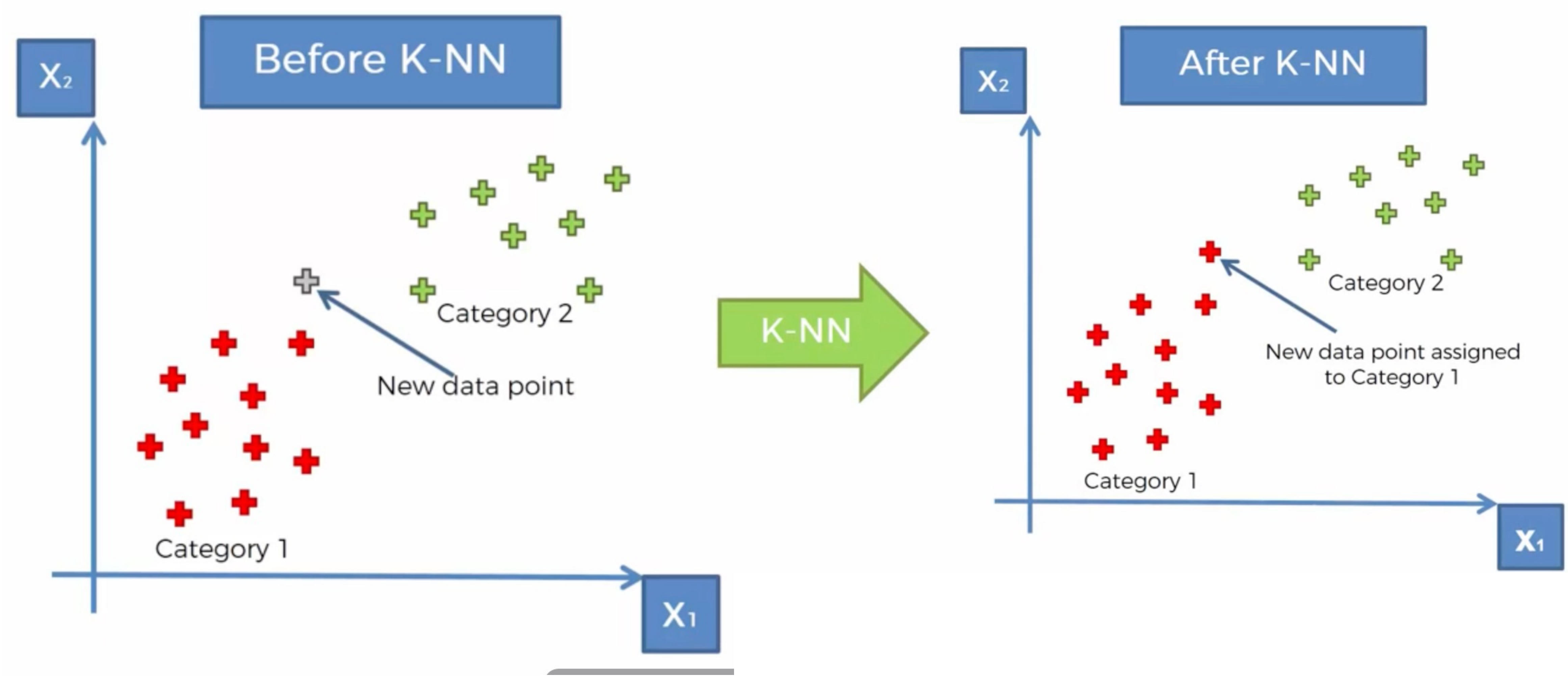


Logistic Regression





K-Nearest Neigbors (K-NN)



KNN

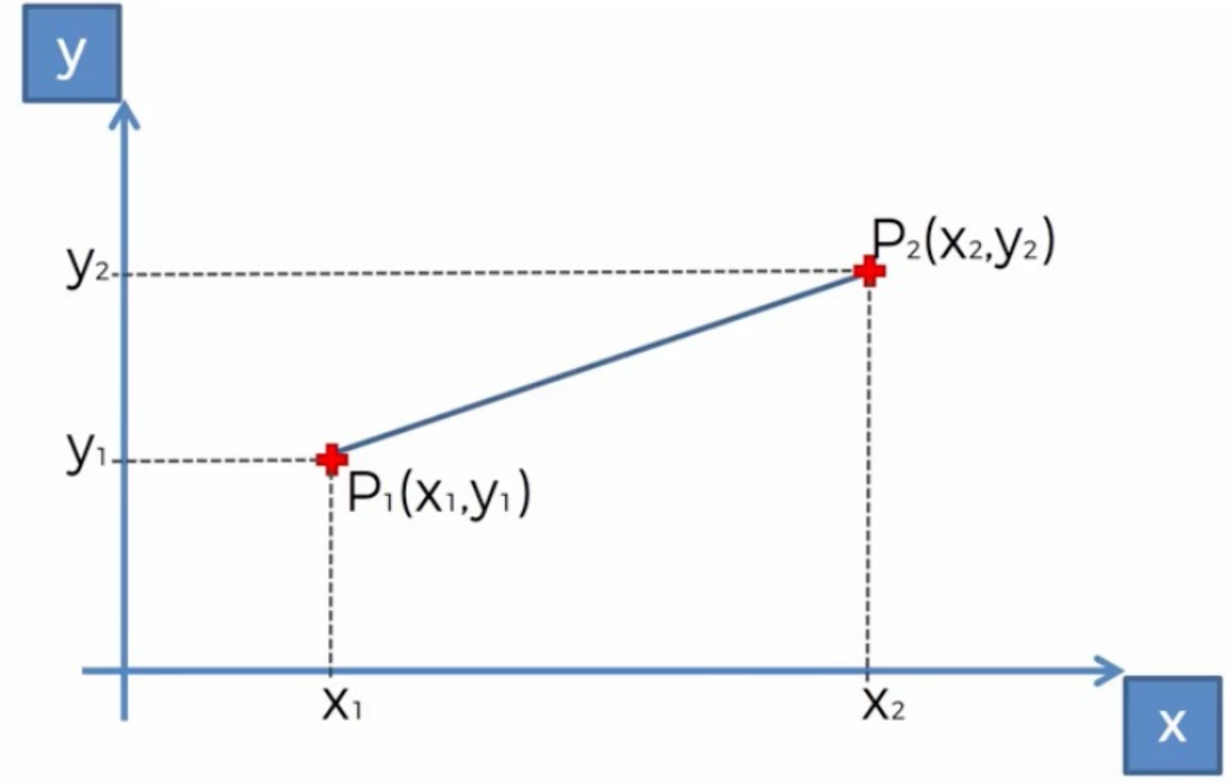
STEP 1: Choose the number K of neigbors

STEP 2: Take the K-Nearest neigbors of the new data point, according to to the Euclidean Distance

STEP 3: Among these K neigbors, count the number of data points in each category

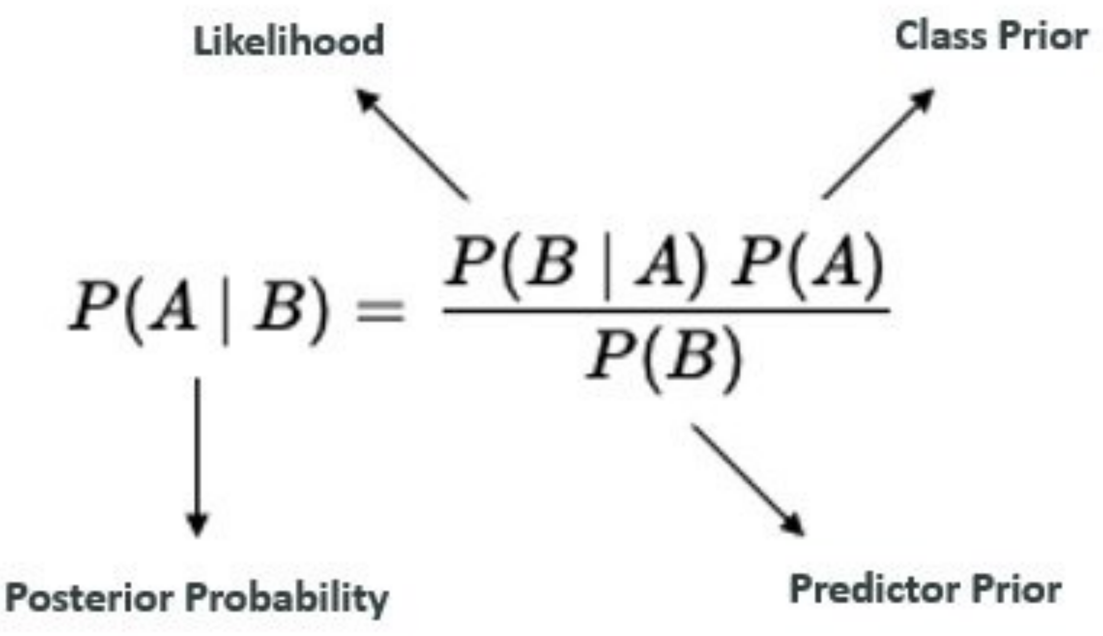
STEP 4: Assign the new data point to the category where you counted the most neighbors.

KNN – Euclidean DIstance



$$\text{Euclidean Distance between } P_1 \text{ and } P_2 = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

Naïve Bayes Algorithm



Bayes Theorem

Example

Weather	Play
Sunny	No
Overcast	Yes
Rainy	Yes
Sunny	Yes
Sunny	Yes
Overcast	Yes
Rainy	No
Rainy	No
Sunny	Yes
Rainy	Yes
Sunny	No
Overcast	Yes
Overcast	Yes
Rainy	No

Frequency Table		
Weather	No	Yes
Overcast		4
Rainy	3	2
Sunny	2	3
Grand Total	5	9

Likelihood table				
Weather	No	Yes		
Overcast		4	=4/14	0.29
Rainy	3	2	=5/14	0.36
Sunny	2	3	=5/14	0.36
All	5	9		
	=5/14	=9/14		
	0.36	0.64		

Problem: Players will play if the weather is sunny. Is this statement correct?

Decision Tree

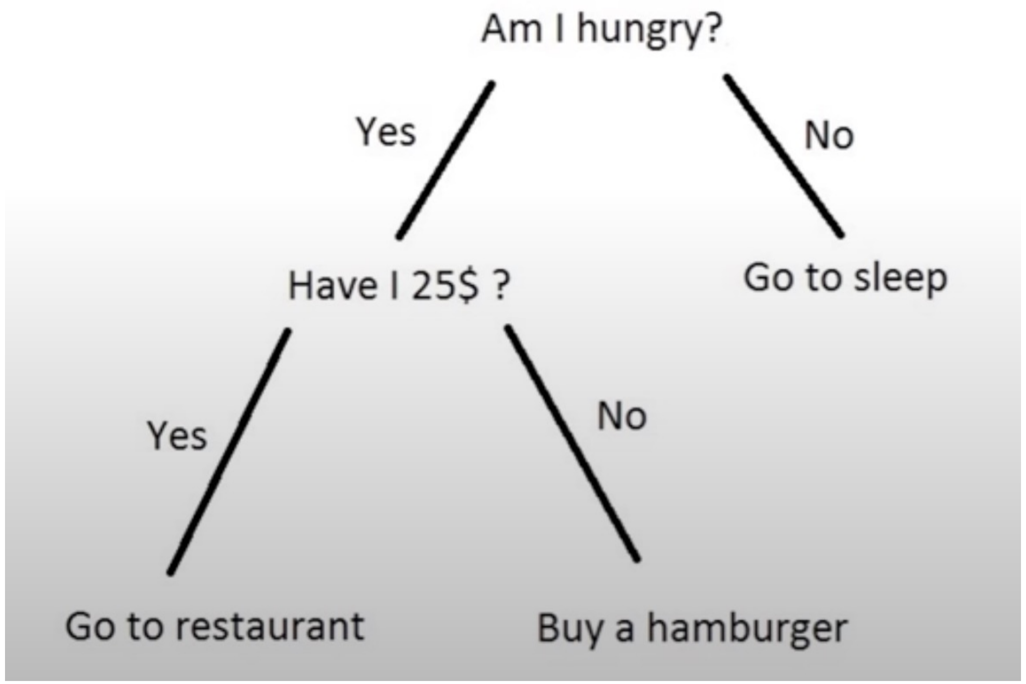
- Classification- to decide class for the record
- Also concerned with generating a description or a model for each class
- Supervised classification
  - Training set – generate description of the classes
  - Test set – determine the effectiveness of the classification
- **Decision trees** or classification tree – represent rules

Decision Tree

- Specific requirements to be considered while designing any decision tree construction algorithm
  - Efficient method to handle very large sized database
  - Method should be able to handle categorical attributes

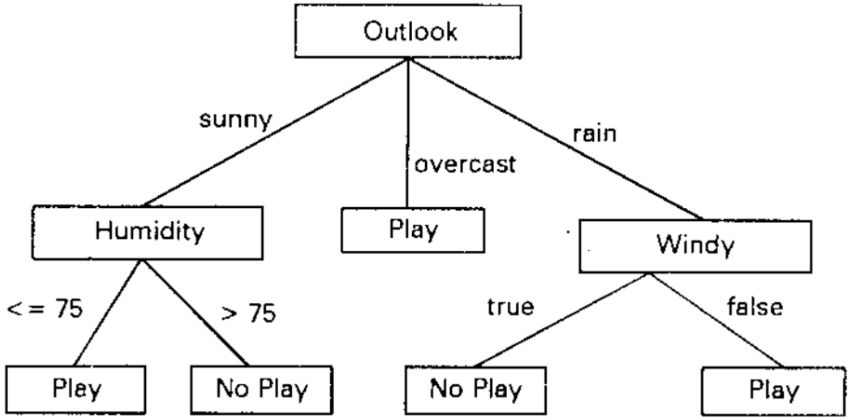
What is decision tree

- Classification scheme which generates a tree and a set of rules, representing the model of different classes, from a given set



Example

OUTLOOK	TEMP(F)	HUMIDITY(%)	WINDY	CLASS
sunny	79	90	true	no play
sunny	56	70	false	play
sunny	79	75	true	play
sunny	60	90	true	no play
overcast	88	88	false	no play
overcast	63	75	true	play
overcast	88	95	false	play
rain	78	60	false	play
rain	66	70	false	no play
rain	68	60	true	no play



RULE 1 *If it is sunny and the humidity is not above 75%, then play.*
RULE 2 *If it is sunny and the humidity is above 75%, then do not play.*
RULE 3 *If it is overcast, then play.*
RULE 4 *If it is rainy and not windy, then play.*
RULE 5 *If it is rainy and windy, then don't play.*

Decision Tree

- Note:
  - Every node is a splitting attribute
  - Every path from root to leaf node represents a rule
    - Different leaf → same class but each leaf → different rule
- Accuracy – percentage of the test data set that is correctly classified

Decision Tree

- A decision tree construction concernd with
  - Identifying the splitting attributes
  - Splitting criteria at every level of the tree
- Advantages:
  - Generate understandable rules
  - Handle both numerical and categorical attribute
  - Provide a clear indication of which fields are most important for prediction or classification

Decision Tree

- Weaknesses
  - Some decision trees can only deal with binary-valued target classes. Others are able to assign records to an arbitrary number of classes but are error prone when the number of training examples per classes gets small. This can happen rather quickly in a tree with many levels and/or many branches per node.
  - The process of growing a decision tree is computationally expensive. At each node, each candidate splitting field is examined before its best split can be found.

Splitting Indices

- Two methods of determining the goodness of split
  - Information gain based on entropy
  - Gini index – derived from economics as measure of diversity
- Entropy

If we are given a probability distribution $P = (p_1, p_2, \ldots, p_n)$, then the information conveyed by this distribution, also called the entropy of P, is

$$\textbf{Entropy}(\boldsymbol{P}) = -[p_1 \log(p_1) + p_2 \log(p_2) + \ldots + p_n \log(p_n)].$$

- If T is partitioned into a set of disjoint exhaustive classes $C_1, C_2, C_3, .., C_n$ on basis of the class attribute, then information needed to identify the class of an element of T is

  Info(T) = Entropy(T)

Entropy

T	C1	C2	C3
100	40	30	30

The value of the entropy of the whole data set is

$$Info(T) = -\frac{40}{100}\log\frac{40}{100} - \frac{30}{100}\log\frac{30}{100} - \frac{30}{100}\log\frac{30}{100} = 1.09$$

Information For a partition on X

If T is partitioned based on the value of the non-class attribute X, into sets T_1, T_2, …, T_n, then the information needed to identify the class of an element of T becomes the weighted average of the information to identify the class of the element of T_i, i.e., the weighted average of $Info(T_i)$

$$Info(X,T) = \sum_{i=1}^{n} \frac{|T_i|}{|T|} Info(T_i) \cdot$$

Information For a partition on X

Let us consider splitting the data set into two subsets, S_1 and S_2, with n_1 and n_2 data, respectively, where $n_1 + n_2 = n$. If we assume $n_1 = 60$ and $n_2 = 40$, the splitting is as follows (Table 6.5):

Table 6.5a

S2	C1	C2	C3
40	0	20	20

Table 6.5b

S1	C1	C2	C3
60	40	10	10

The entropy index value of the data set after the segmentation is

$$\frac{40}{100}\left(-\frac{20}{40}\log\frac{20}{40}-\frac{20}{40}\log\frac{20}{40}\right)+\frac{60}{100}\left(-\frac{40}{60}\log\frac{40}{60}-\frac{10}{60}\log\frac{10}{60}-\frac{10}{60}\log\frac{10}{60}\right)=0.80.$$

Gain

- We define the information gain due to split on X as

$$Gain(X, T) = Info(T) - Info(X, T).$$

- The information gain represents the difference between information need to identify an element of T and the information need to identify an element of T after the value of attribute is obtained ie. Information gain due to X

- Gain is 0.29

Example

Day	Outlook	Temp	Humidity	Wind	Play Tennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

Gain Ratio

- The notion of gain → favour attributes with large number of distinct value
- Quinlan suggest using the gain ratio

$$Gain\_ratio(X,T) = \frac{Gain(X,T)}{Info(X,T)}.$$

- Suppose Gain (outlook,T)=0.246 and Info(outlook,T) =0.694, then

$$Gain\_ratio(outlook,T) = \frac{Gain(outlook,T)}{Info(outlook,T)} = \frac{0.246}{0.694} = 0.3544.$$

Gini Index

If a data set T contains n classes, then $gini(T)$ is defined as

$$gini(T) = 1 - \sum p_i^2 \cdot$$

where p_j is the relative frequency of class j in T. If the split divides T into T_1 and T_2, then the index of the divided data is given as

$$gini_{spli}(T) = \frac{n_1}{n} gini(T_1) + \frac{n_2}{n} gini(T_2) \cdot$$

Gini Index

- Considering T with 14 records with classes c1 = 9records and c2 = 5 records and attribute outlook with sunny =5 (c1=3, c2=2), overcast = 4 (c1=4) and rain = 5(c1=3, c2=2).

$$gini(T) = 1 - \left[\frac{9}{14}\right]^2 - \left[\frac{5}{14}\right]^2 = 0.46.$$

Thus, the gini index due to splitting on *outlook* is

$$gini_{outlook}(T) = \frac{5}{14}\left[1 - \left[\frac{3}{5}\right]^2 - \left[\frac{2}{5}\right]^2\right] + \frac{4}{14}[1-1] + \frac{5}{14}\left[1 - \left[\frac{3}{5}\right]^2 - \left[\frac{2}{5}\right]^2\right] = 0.343.$$

The best splitter is determined as the attribute which has the smallest gini value.

Problem

Calculate overall Entropy, information gain and gain ratio against each attributes and Gini-index for each attributes. Create a decision tree using these two techniques.

Resp srl no	Target variable	Predictor variable	Predictor variable	Predictor variable
	Exam Result	Other online courses	Student background	Working Status
1	Pass	Y	Maths	NW
2	Fail	N	Maths	W
3	Fail	y	Maths	W
4	Pass	Y	CS	NW
5	Fail	N	Other	W
6	Fail	Y	Other	W
7	Pass	Y	Maths	NW
8	Pass	Y	CS	NW
9	Pass	n	Maths	W
10	Pass	n	CS	W
11	Pass	y	CS	W
12	Pass	n	Maths	NW
13	Fail	y	Other	W
14	Fail	n	Other	NW
15	Fail	n	Maths	W

Thank you