Classification & Decision Trees

Business Analytics

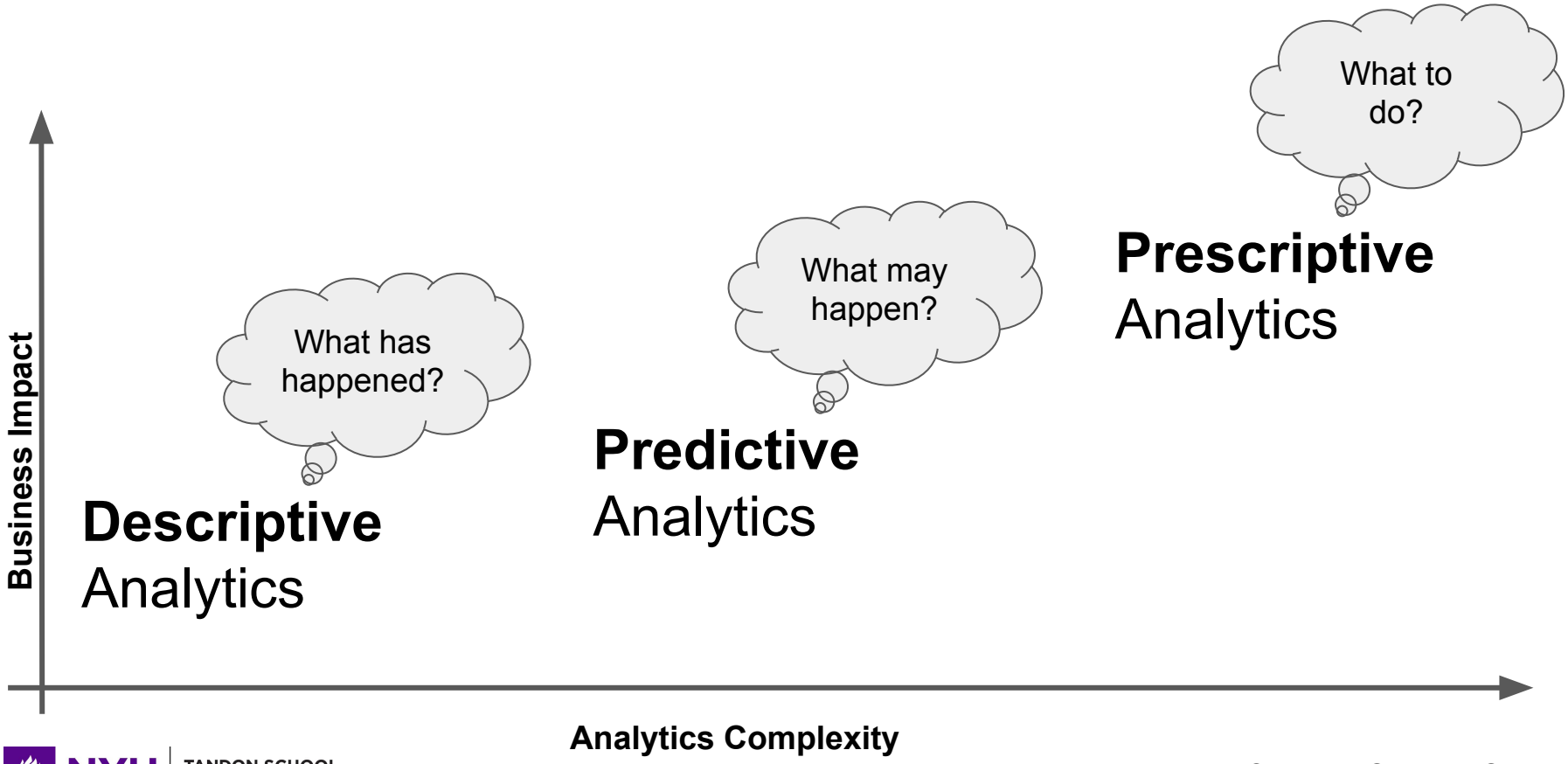
Business Analytics Definition

Business analytics refers to the application of data analysis and modeling techniques for understanding business situations and improving business decisions.

IMPLICATIONS:

- data → past business performance
- methods → statistics + mathematics + computational methods
- business decisions → actionable insight

Types of Analytics



Review

1. Regression - Theory
    a. Linear Models
    b. Ordinary Least Squares
    c. Simple Linear Regression
2. Regression - Application
    a. Model strength
    b. Model interpretation
    c. Dummy variables
    d. Non-linear transformations
3. Prediction

Lesson Objectives

1. Classification
    a. Statistical classification
    b. Decision Trees

Classification

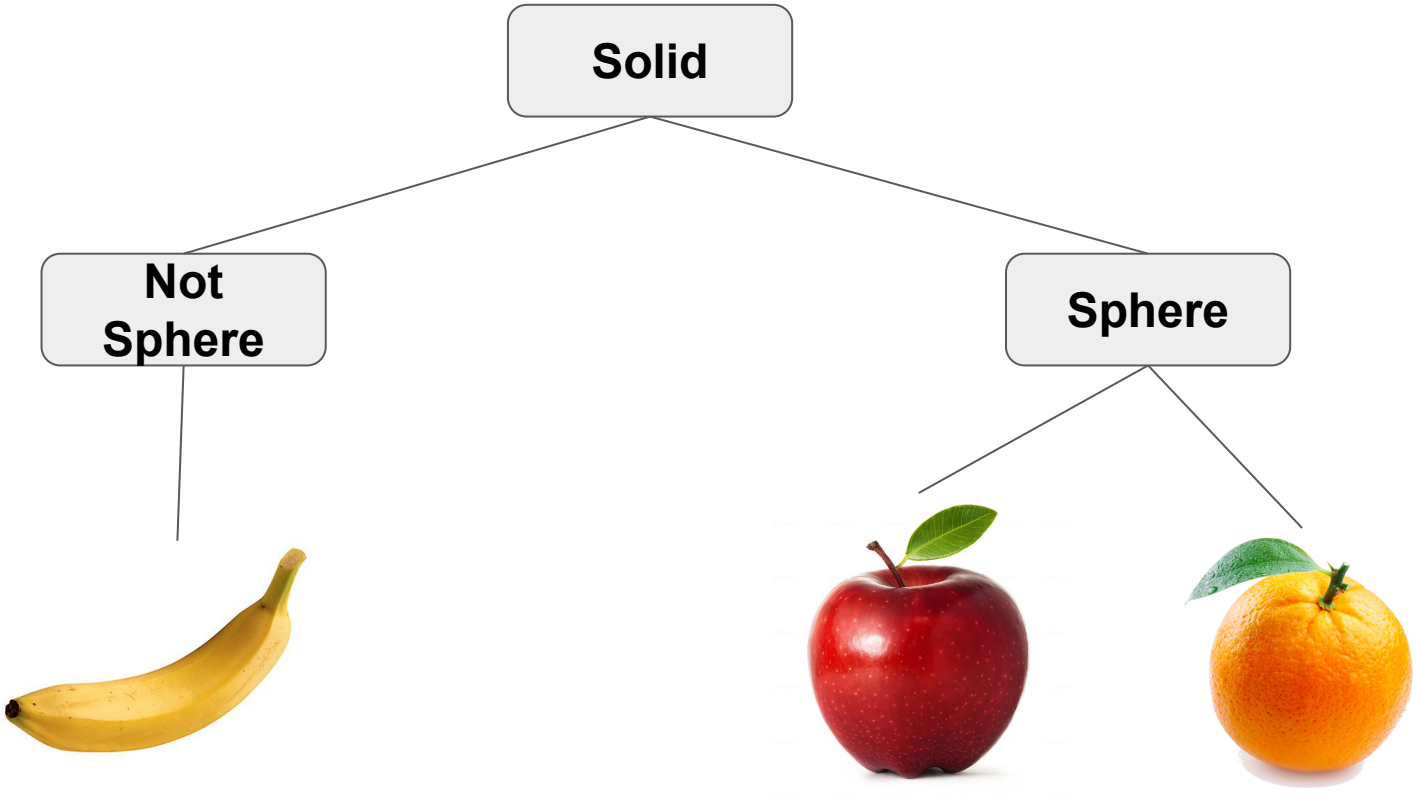
Classification is the task of assigning objects to one of several predefined categories.

- detecting spam emails based on message header and content
- segmenting customers based on their response to an offer
- categorizing loan applications according to their risk level

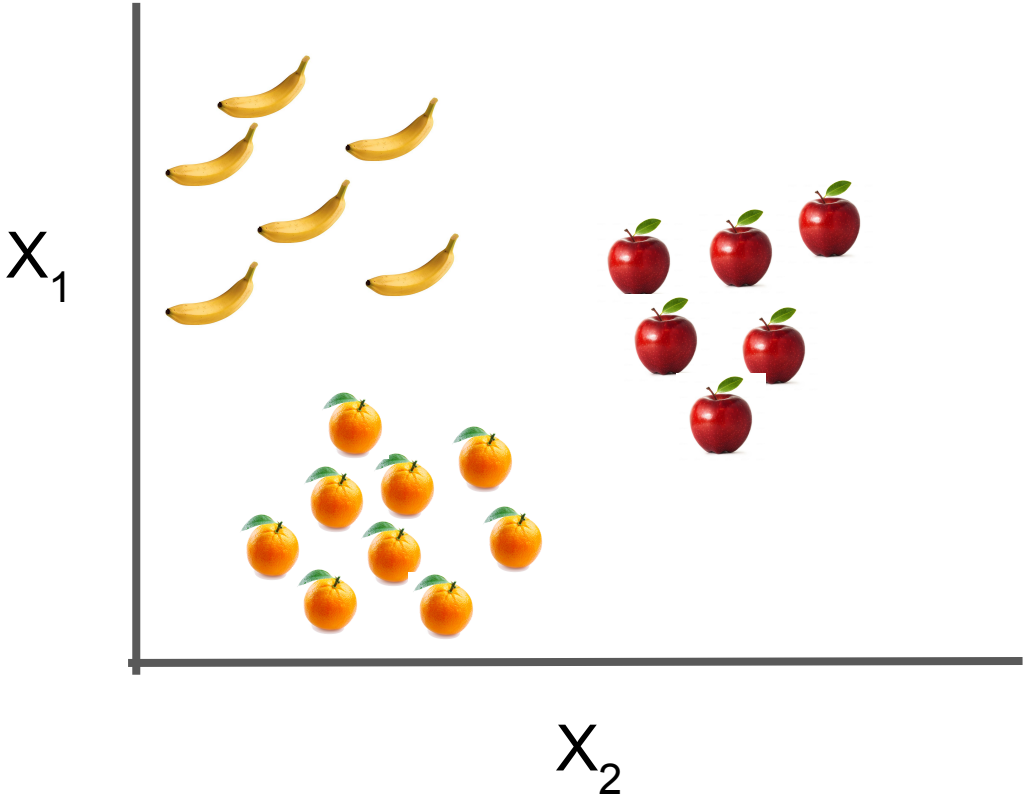
Classification

- A classification model can serve as an explanatory tool to distinguish between objects of different classes -- descriptive analytics
- It can also be used to predict the class label of unknown records -- predictive analytics
- Classification techniques are most suited for predicting or describing data sets with binary or nominal categories.
    - They are less effective for ordinal categories (e.g.: classify a person as a member of high-, medium-, or low-income group) because they do not consider the implicit order among the categories.
- Examples of classification techniques include decision tree classifiers, neural networks, support vector machines, naive Bayes classifiers...

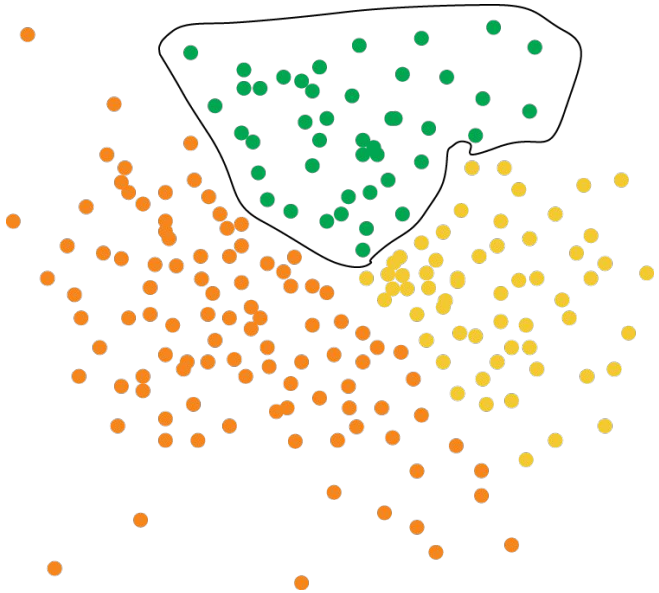
Classification Example

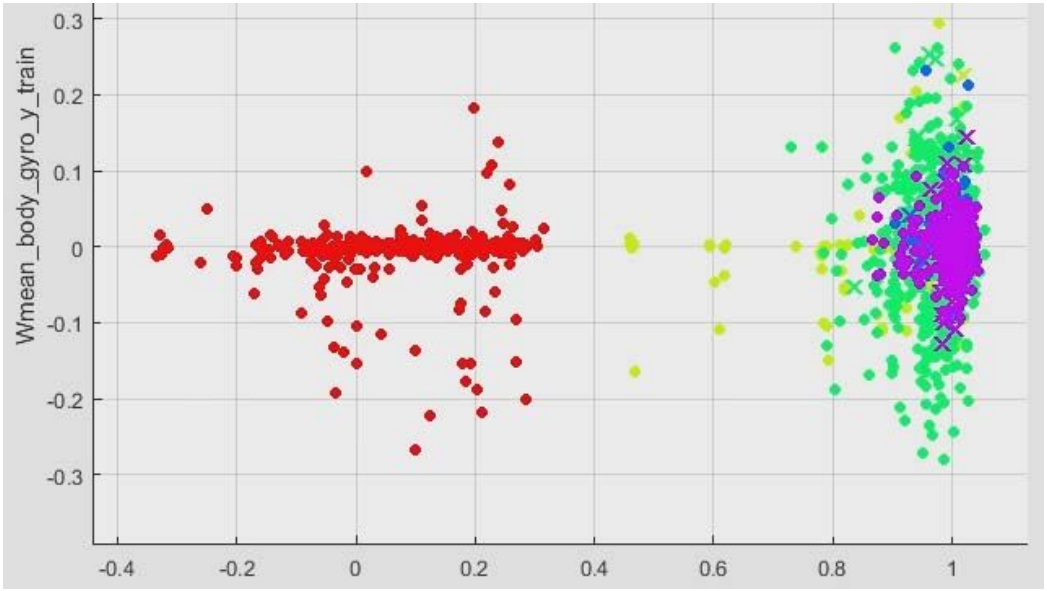


Statistical Classification

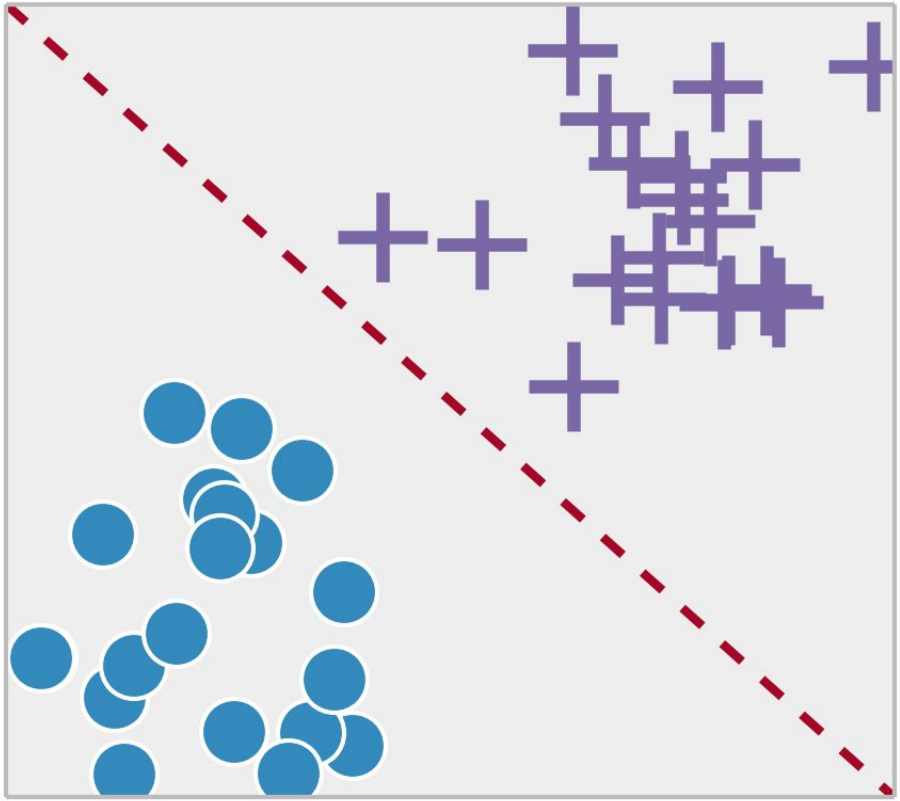


Statistical Classification Examples



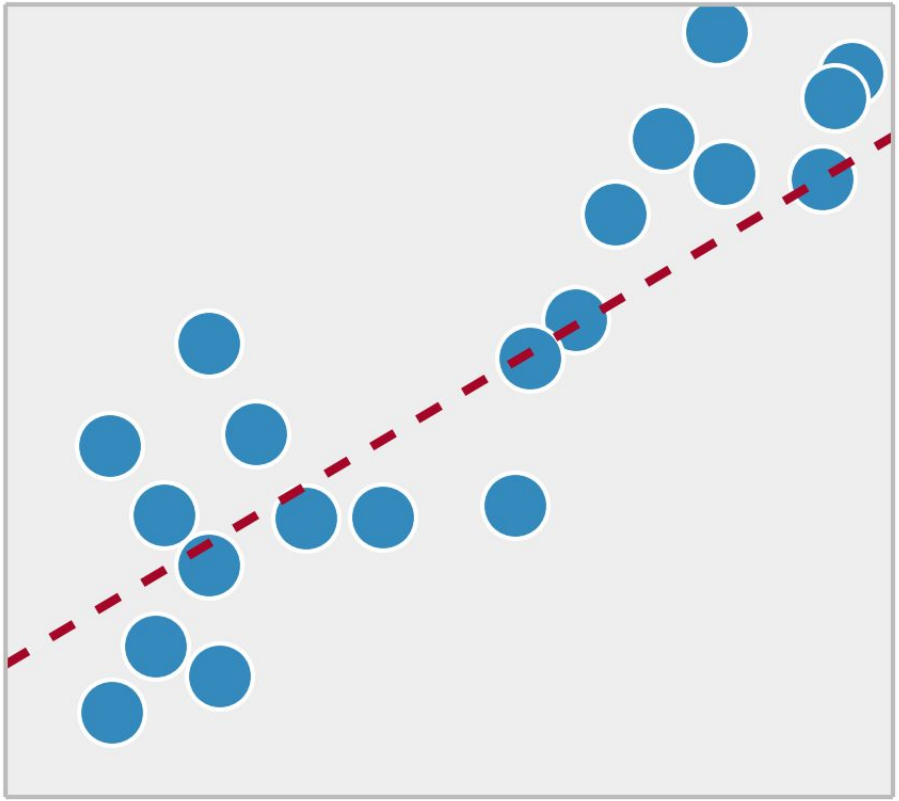


Classification



Output: Discrete (labels); Decision boundary
Evaluation: Accuracy;

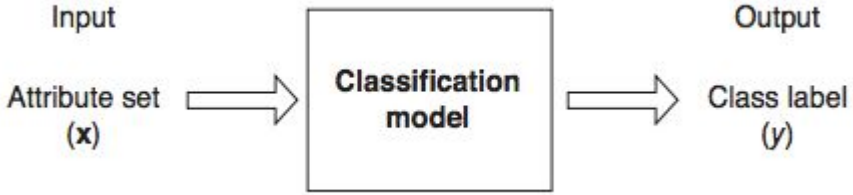
Regression



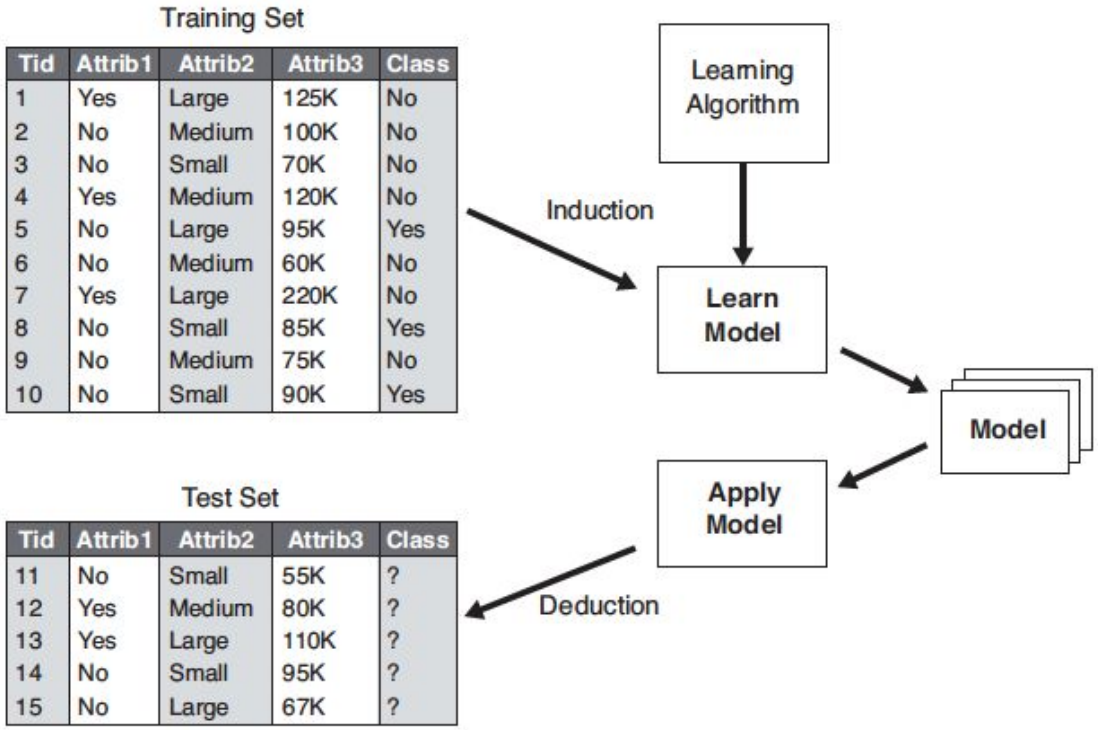
Output: Continuous (number); best fit line
Evaluation: Sum of Errors; R^2

Recap of Modeling Process

- Employ a learning algorithm to identify a model that best fits the relationship between the attribute set and class label of the input data.
- The model should both fit the input data well and correctly predict the class labels of records it has never seen before.
  - training set
  - test set
- Key objective is to build models with good generalization capability.



Recap of Modeling Process



Classification

- Evaluation of the performance of a classification model is based on the counts of test records correctly and incorrectly predicted by the model.
  - confusion matrix

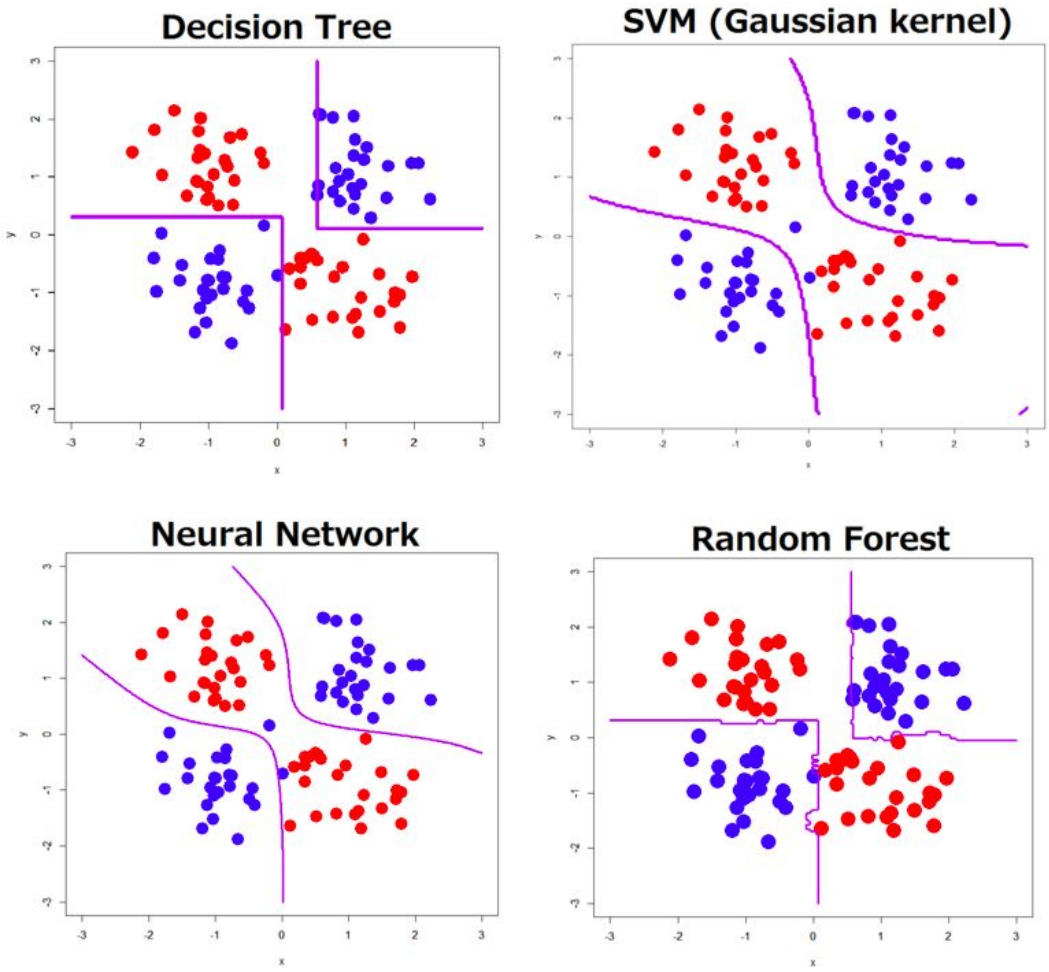
		Predicted Class	
		$Class = 1$	$Class = 0$
Actual Class	$Class = 1$	f_{11}	f_{10}
	$Class = 0$	f_{01}	f_{00}

- Other performance metrics:

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}}$$

$$\text{Error rate} = \frac{\text{Number of wrong predictions}}{\text{Total number of predictions}}$$

- Base rate: how well would a classifier perform by simply choosing that class for every instance

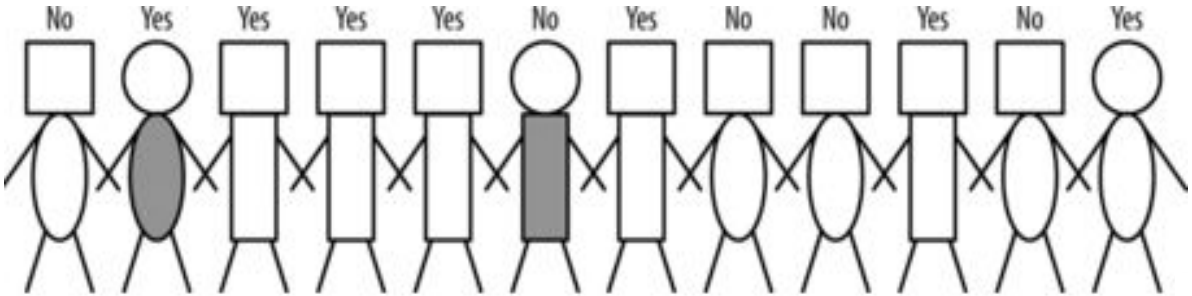

Decision Tree
SVM (Gaussian kernel)
Neural Network
Random Forest

Decision Trees

- Pose a series of questions about the characteristics of the target variable.
  - A follow-up question is asked until a conclusion is reached about the class label of the record
- The series of questions and their possible answers can be organized in the form of a decision tree.
  - nodes -- root node, internal nodes, leaf or terminal nodes
  - directed edges
- Each leaf node is assigned a class label.
- Non-terminal nodes contain attribute test conditions to separate records that have different characteristics.

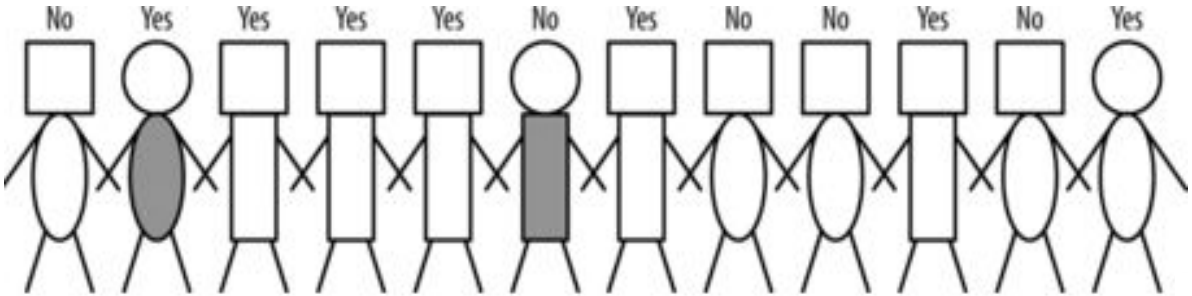
Classification Problem

- Determining whether a customer becomes a loan write-off
  - Binary classification problem with target variable “yes” or “no”



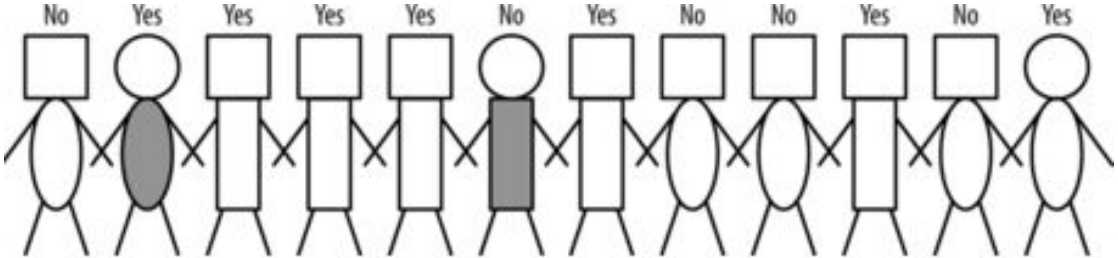
Classification Problem

- Determining whether a customer will default on a loan
  - Binary classification problem with target variable “yes” or “no”
  - Customers represented as stick figures with three attributes

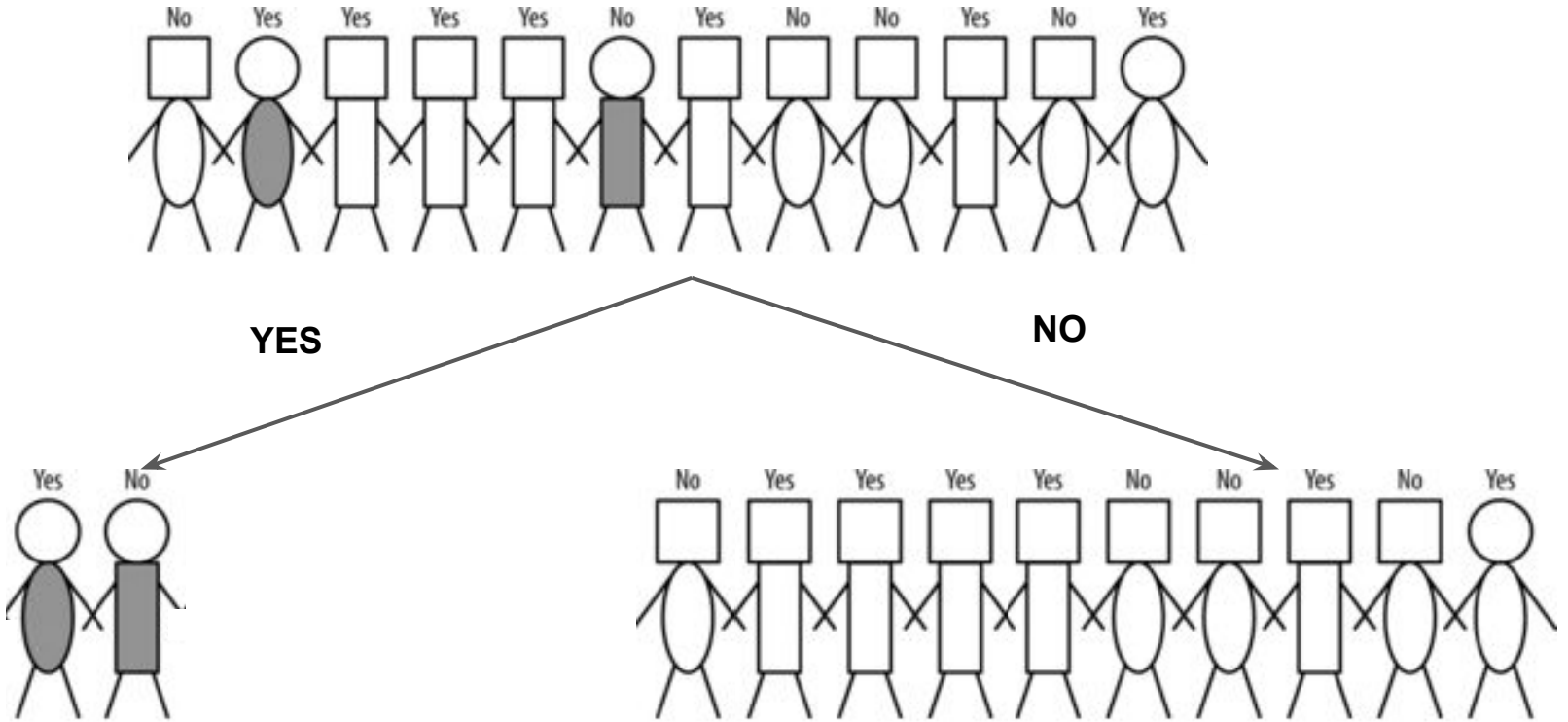


Classification Problem

- Determining whether a customer will default on a loan
  - Binary classification problem with target variable “yes” or “no”
  - Customers represented as stick figures with three attributes
    - head shape
    - body shape
    - body color
  - Which of the attributes would be best to segment these people into groups to distinguish defaults from non defaults?
  - We would like the resultant groups to be as pure as possible with respect to the target variable.

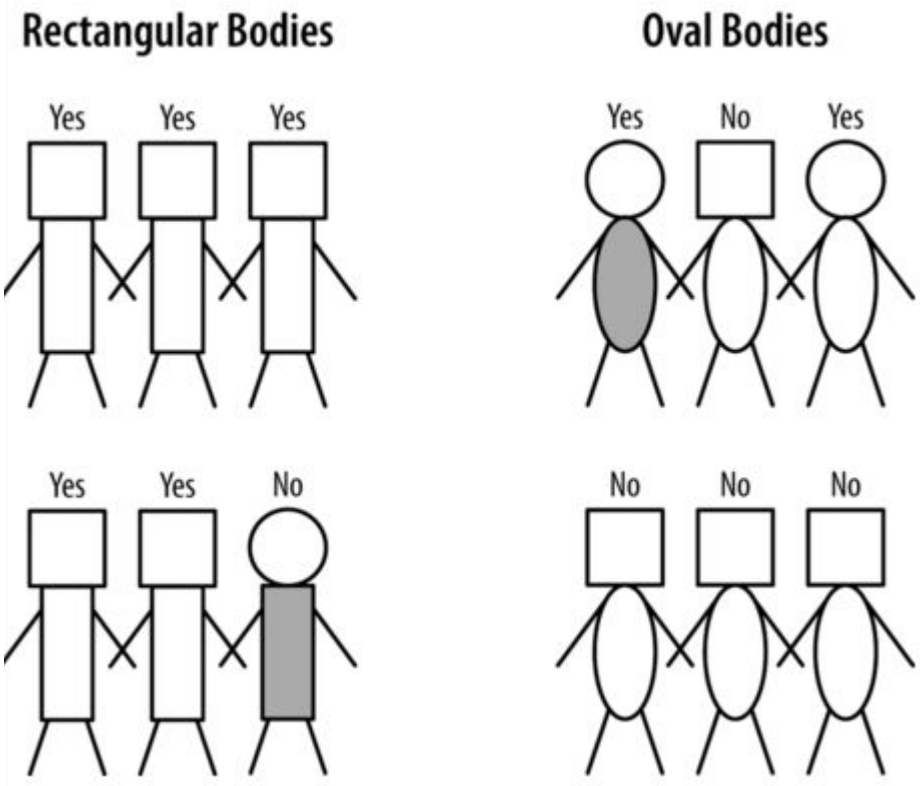


body-color = gray

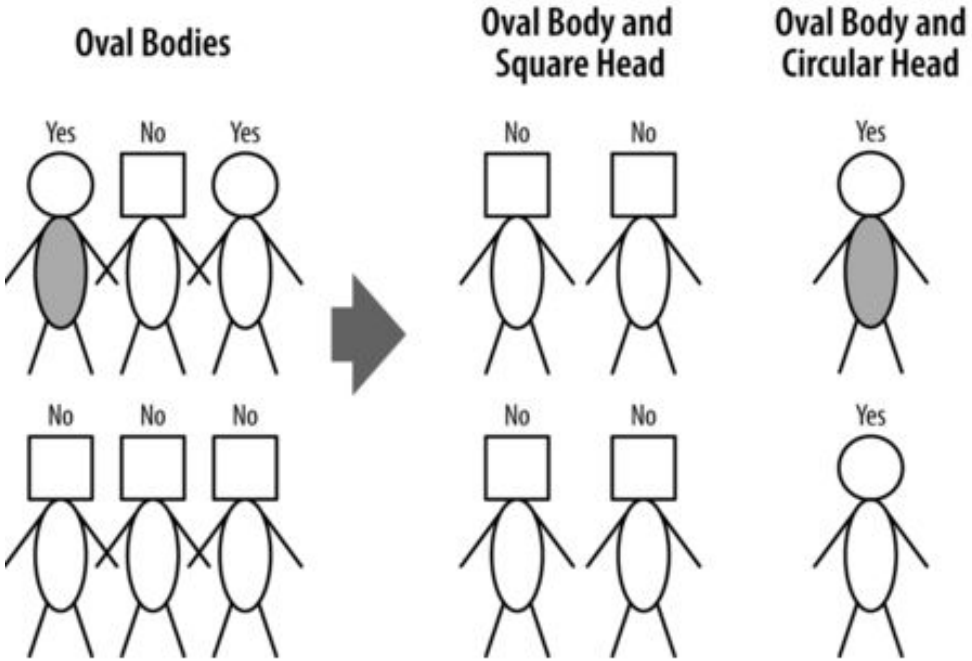


Are these groups pure?

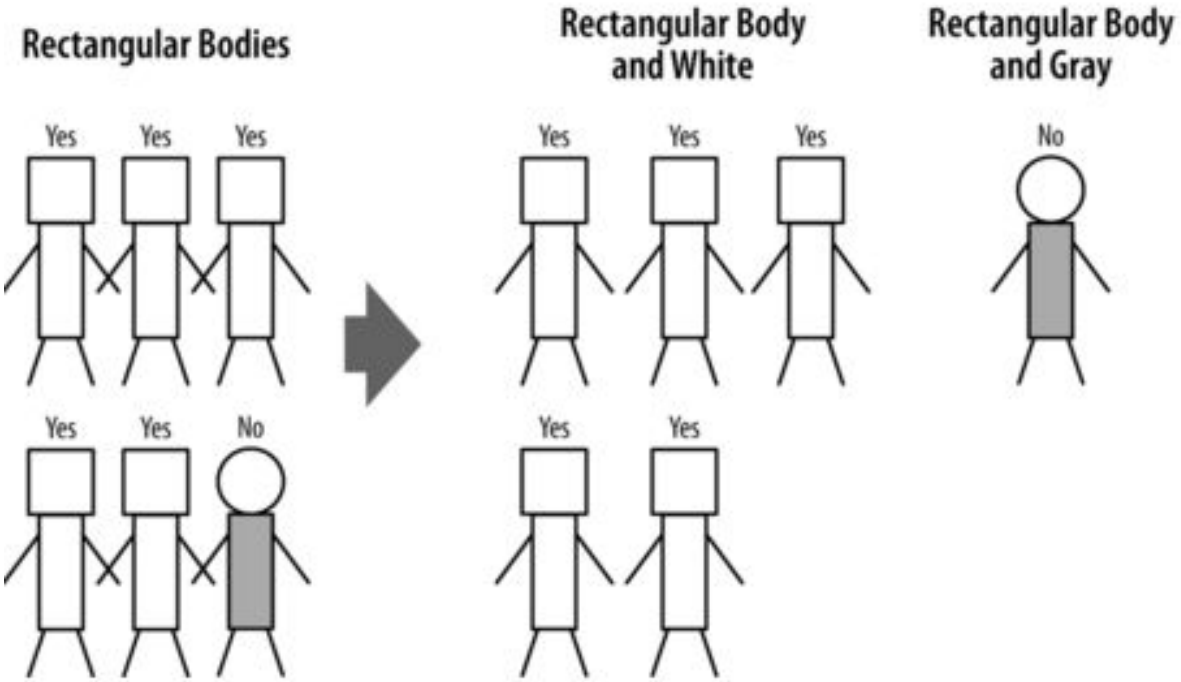
First partitioning: Body shape



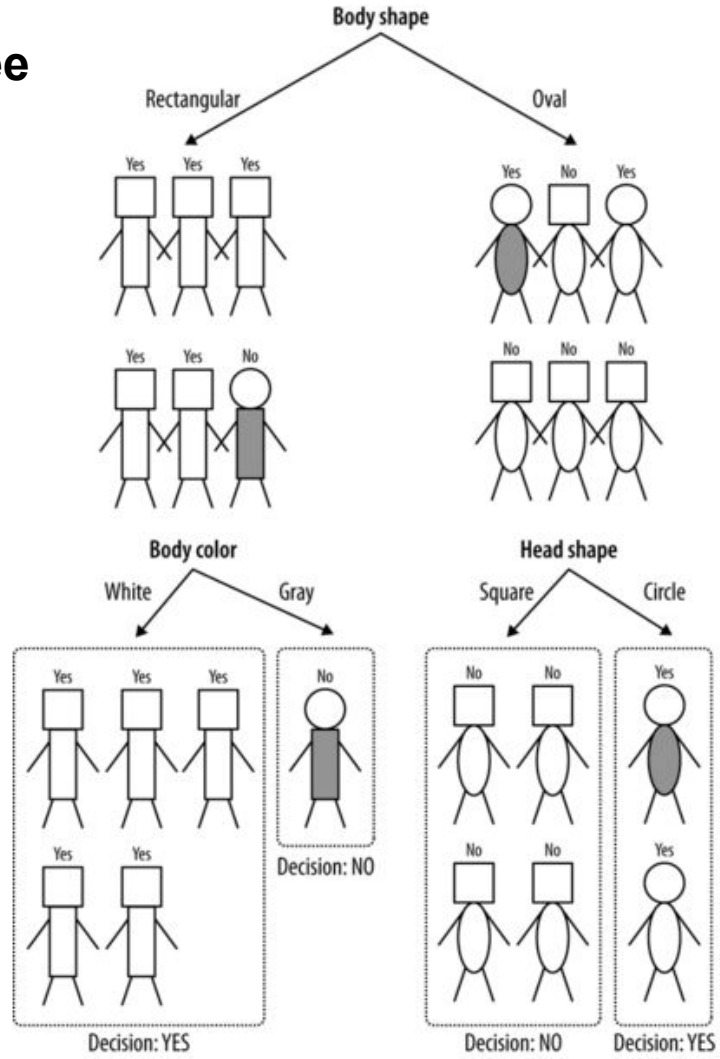
Second partitioning: Oval body people subgrouped by head type



Third partitioning: Rectangular body people subgrouped by body color

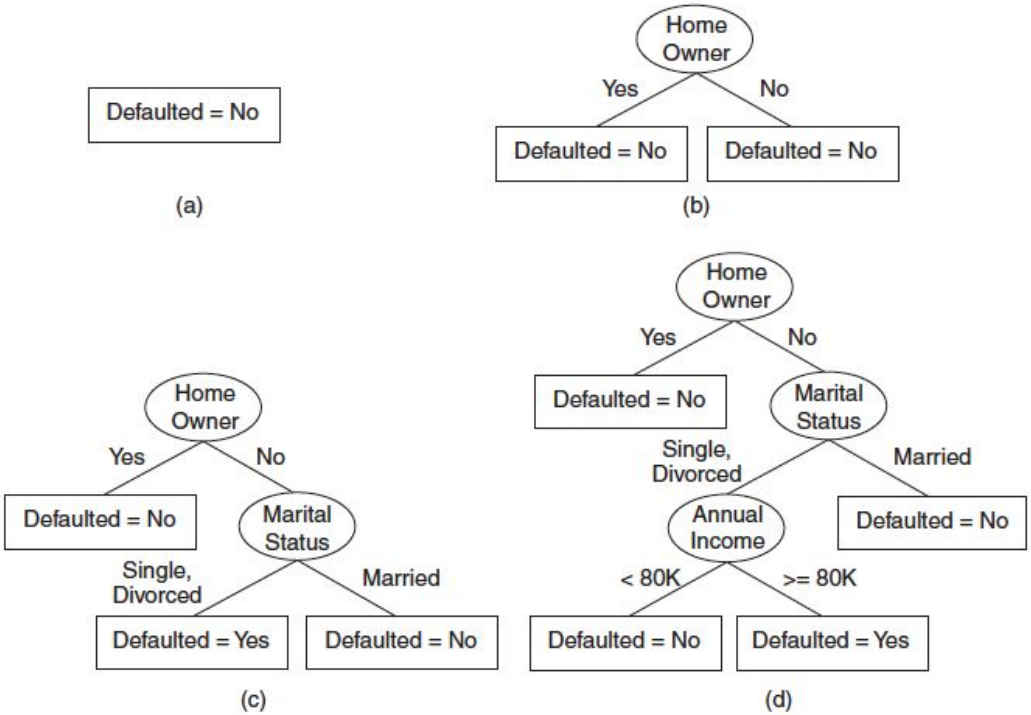


The classification tree resulting from the splits



Classification Problem

Tid	Home Owner	Marital Status	Annual Income	Defaulted Borrower
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes



Classification Example

- Data from direct marketing campaigns of a Portuguese banking institution
- The marketing campaigns were based on phone calls.
- Often, more than one contact to the same client was required
- Outcome: customer signed up for a bank term deposit or not
  - subscribe = yes/no
- The classification goal is to predict if a given client will subscribe (yes/no) for a term deposit.
- https://archive.ics.uci.edu/ml/datasets/Bank+Marketing

Data Attributes

 # bank client data:

1 - age (numeric)

2 - job : type of job (categorical: "admin.","unknown","unemployed","management","housemaid","entrepreneur","student", "blue-collar","self-employed","retired","technician","services")

3 - marital : marital status (categorical: "married","divorced","single"; note: "divorced" means divorced or widowed)

4 - education (categorical: "unknown","secondary","primary","tertiary")

5 - default: has credit in default? (binary: "yes","no")

6 - balance: average yearly balance, in euros (numeric)

7 - housing: has housing loan? (binary: "yes","no")

8 - loan: has personal loan? (binary: "yes","no")

 # related with the last contact of the current campaign:

9 - contact: contact communication type (categorical: "unknown","telephone","cellular")

10 - day: last contact day of the month (numeric)

11 - month: last contact month of year (categorical: "jan", "feb", "mar", ..., "nov", "dec")

12 - duration: last contact duration, in seconds (numeric)

 # other attributes:

13 - campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)

14 - pdays: number of days that passed by after the client was last contacted from a previous campaign (numeric, -1 means client was not previously contacted)

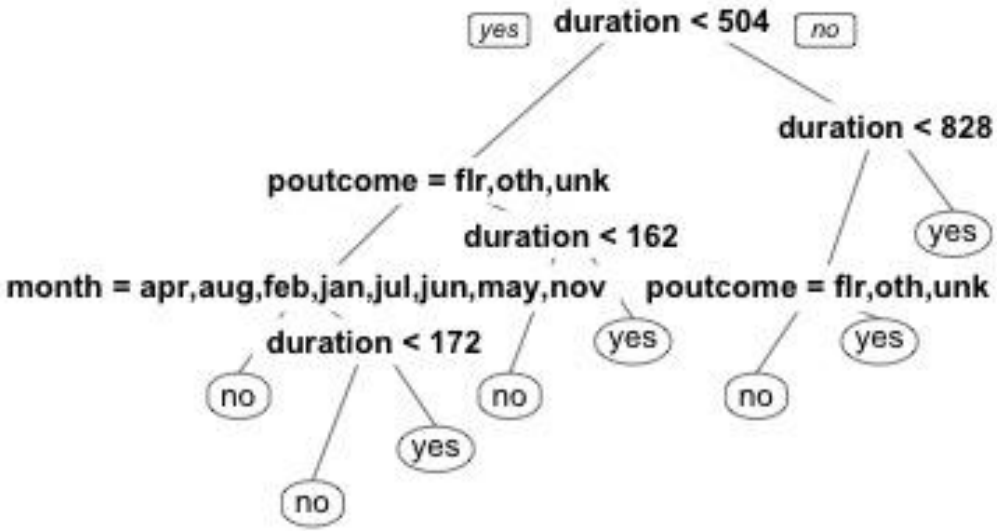
15 - previous: number of contacts performed before this campaign and for this client (numeric)

16 - poutcome: outcome of the previous marketing campaign (categorical: "unknown","other","failure","success")

Output variable (desired target):

17 - subscribe - has the client subscribed a term deposit? (binary: "yes","no")

Subscribe for Deposit?



Homework - Part 1

The goal of this assignment is to develop a classification model for the census income dataset you worked on last time.

- Census Income Data Set in the UCI Machine Learning Repository
  - https://archive.ics.uci.edu/ml/datasets/Census+Income
- Create a classification model to determine whether the income of an individual is greater than $50K.
- Identify the top 2/3 criteria that distinguish those whose income is greater than $50K and those whose income is less than $50K.
- Are the decision tree results in line with your findings from your exploratory analysis?

Homework - Part 2

The goal of to discover a something with Math data we used in the exam

- Load the portugal math data
- Create a classification model to determine student success
- How would you incorporate this new findings into a regression model
- With your new findings, what are the business implications here