KOMDIGI
Kementerian Komunikasi dan Digital
Republik Indonesia

digitalent

VSGA Vocational School Graduate Academy

VOCATIONAL SCHOOL GRADUATE ACADEMY

Associate Data Scientist

Pertemuan: Membangun Model Classification

PROFIL PENGAJAR

Jabatan Akademik (tahun dan jabatan terakhir Pengajar)
Latar belakang Pendidikan Pengajar

- AAA
- BBB
- CCC

Riwayat Pekerjaan

- AAA
- BBB
- CCC

Contact Pengajar
Ponsel :
Email :

Course Definition

- Bagian Delapan dari Associate Data Scientist
- Berfokus pada Membangun Model
- Konteksnya yaitu :

Kursus ini akan menjelaskan Classification dan bagaimana membangun model Classification, yaitu:
a. menyiapkan parameter model,
b. menggunakan tools pemodelan,
selanjutnya menjelaskan algoritma dan menggunakan Classification dengan Rapid Miner

Learning Objective

Peserta mempelajari pengertian, cara menyiapkan, dan cara implementasi model Classification dengan algoritma:

A. Classification

B. K-Nearest-Neighbors

E. Decision Tree

Beserta pengukuran performansinya menggunakan Rapid Miner.

Classification

What is Classification?

Approach:

- Given a collection of records (*training set*)
- each record contains a set of *attributes*
- one of the attributes is the *class (label)* that should be predicted.
- Learn a *model* for the class attribute as a function of the values of other attributes.

Variants:

- single-class problems (class labels e.g. true/false or fraud/no fraud)
- multi-class problems (class labels e.g. low, medium, high)

Introduction to Classification

A Couple of Questions:

- What is this?
- Why do you know?
- How have you come to that knowledge?

Introduction to Classification

Goal: Learn a model for recognizing a concept, e.g. trees
 Training data:

"tree"

"tree"

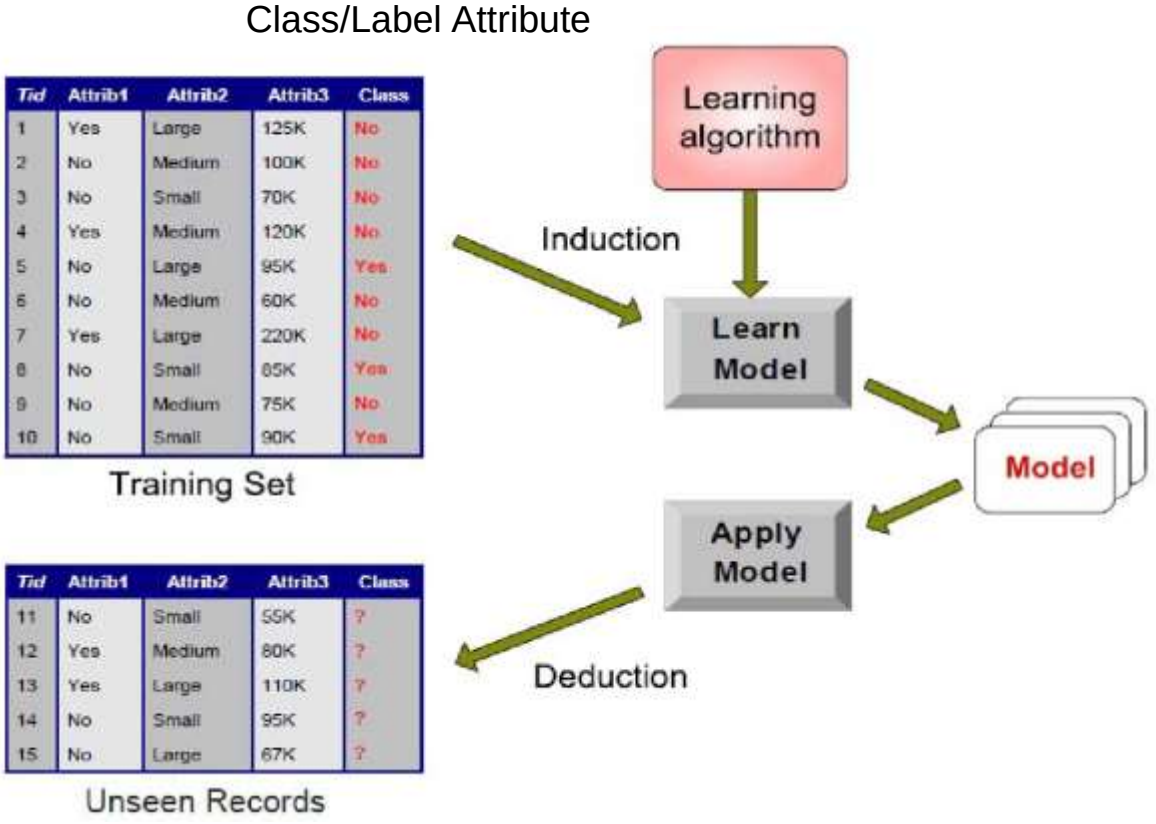
"tree"

"not a tree"

"not a tree"

"not a tree"

Model Learning and Model Application Process



Classification Examples

Credit Risk Assessment

- Attributes: your age, income, debts, …
- Class: Are you getting credit by your bank?

Marketing

- Attributes: previously bought products, browsing behaviour
- Class: Are you a target customer for a new product?

Tax Fraud

- Attributes: the values in your tax declaration
- Class: Are you trying to cheat?

SPAM Detection

- Attributes: words and header fields of an e-mail
- Class: Is it a spam e-mail?

Classification Techniques

1. K-Nearest-Neighbors
2. Decision Trees
3. Rule Learning
4. Naïve Bayes
5. Support Vector Machines
6. Artificial Neural Networks
7. Deep Neural Networks
8. Many others …

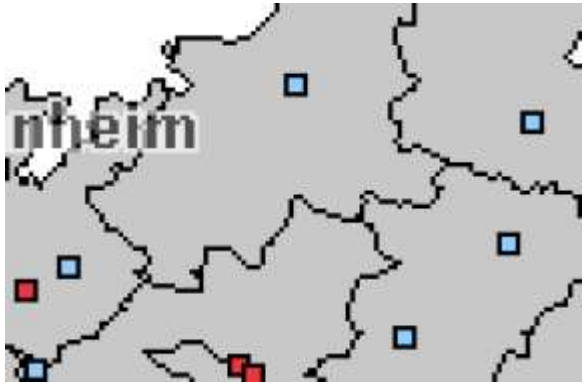
K-Nearest-Neighbors

Example Problem

– Predict what the current weather is in a certain place
– where there is no weather station.
– How could you do that?

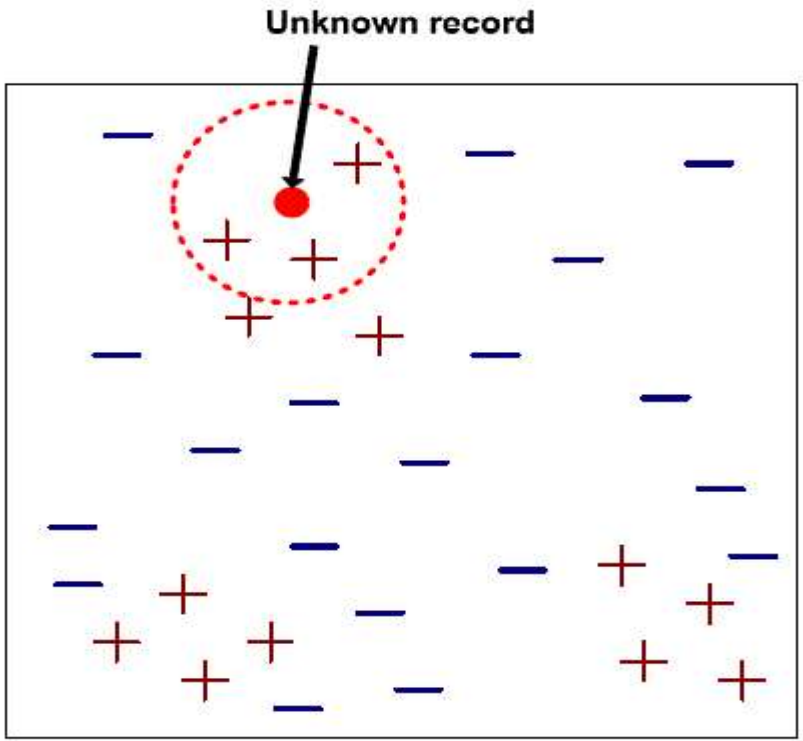
Basic Idea

- Use the average of the nearest stations
- Example:
  - 3x sunny
  - 2x cloudy
  - result = sunny



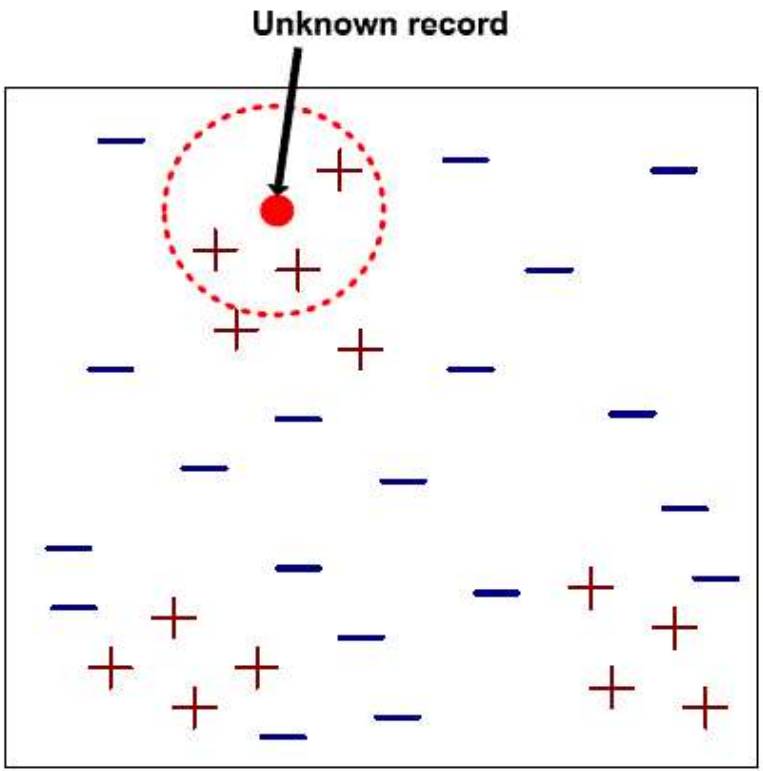
- This approach is called K-Nearest-Neighbors
  - where k is the number of neighbors to consider
  - in the example: k=5
  - in the example: “near” denotes geographical proximity

K-Nearest-Neighbors Classifiers



- Require three things
  - The set of stored records
  - A distance measure to compute distance between records
  - The value of k, the number of nearest neighbors to consider
- To classify an unknown record:
  1. Compute distance to each training record
  2. Identify k-nearest neighbors
  3. Use class labels of nearest neighbors to determine the class label of unknown record
     - by taking majority vote or
     - by weighing the vote according to distance

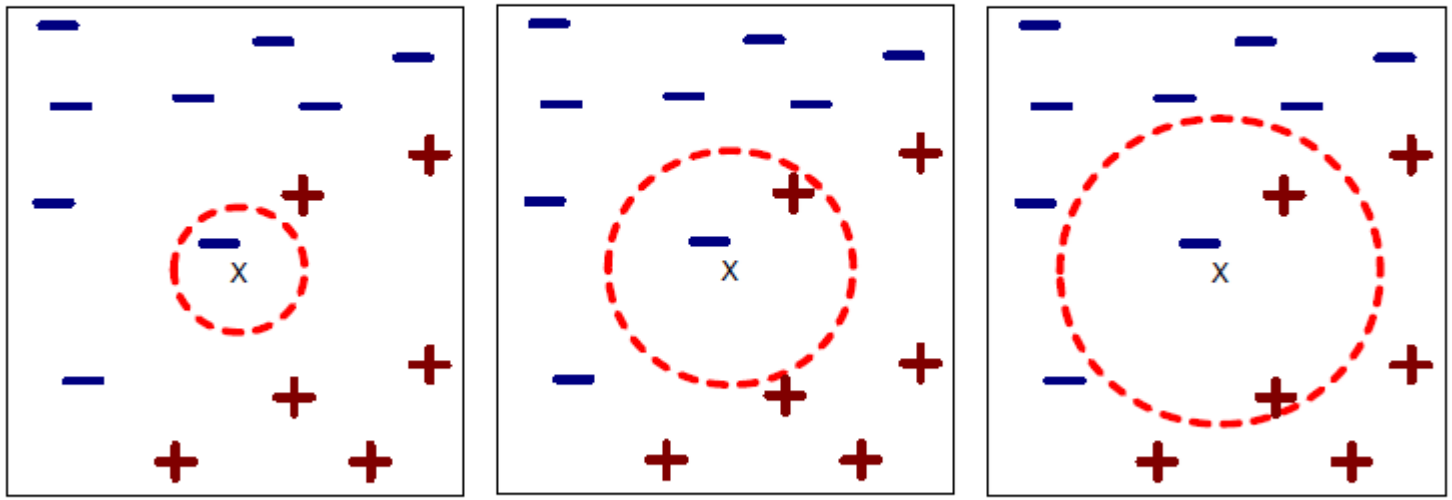
K-Nearest-Neighbors Classifiers



- Require three things
  - The set of stored records
  - A distance measure to compute distance between records
  - The value of k, the number of nearest neighbors to consider
- To classify an unknown record:
  1. Compute distance to each training record
  2. Identify k-nearest neighbors
  3. Use class labels of nearest neighbors to determine the class label of unknown record
     - by taking majority vote or
     - by weighing the vote according to distance

Examples of K-Nearest Neighbors

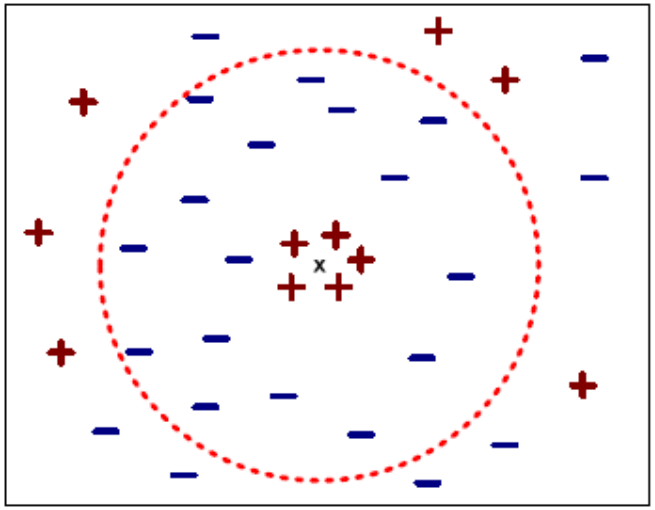
The k-nearest neighbors of a record x are data points that have the k smallest distances to x.



(a) 1-nearest neighbor (b) 2-nearest neighbor (c) 3-nearest neighbor

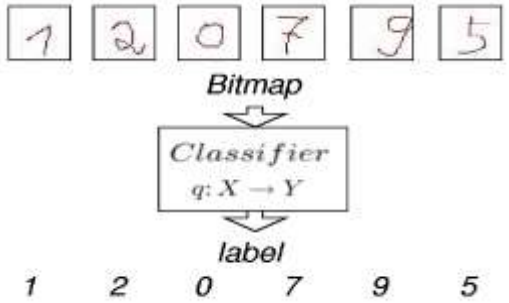
Choosing a Good Value for K

- If k is too small, the result is sensitive to noise points
- If k is too large, the neighborhood may include points from other classes



- Rule of thumb: Test k values between 1 and 10.

Discussion of K-Nearest-Neighbor Classification



 Often very accurate

• for instance for optical character recognition (OCR)

 … but slow

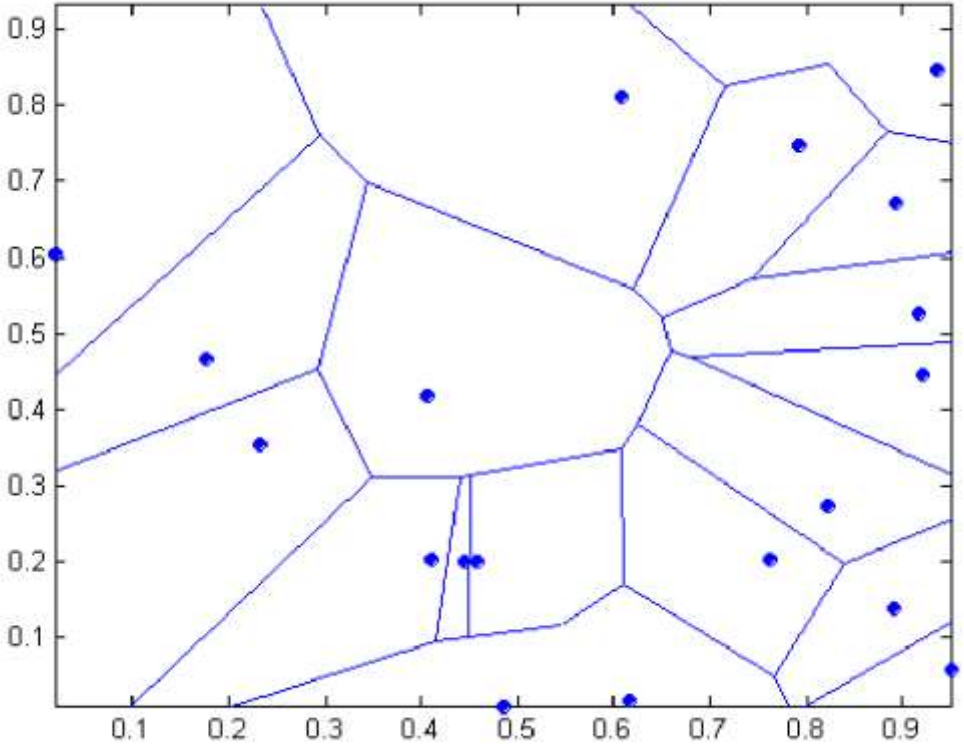
• as training data needs to be searched

 Assumes that all attributes are equally important

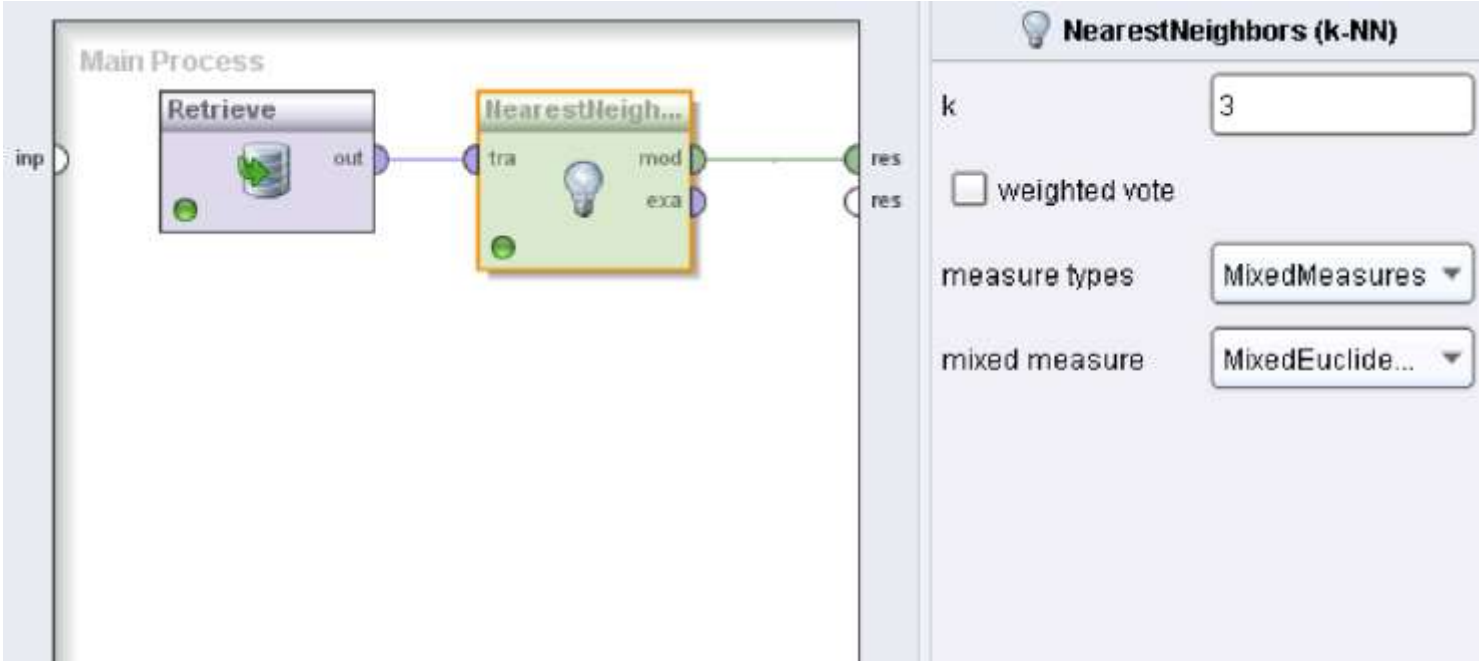
• remedy: attribute selection or attribute weights

 Can handle decision boundaries which are not parallel to the axes (unlike decision trees)

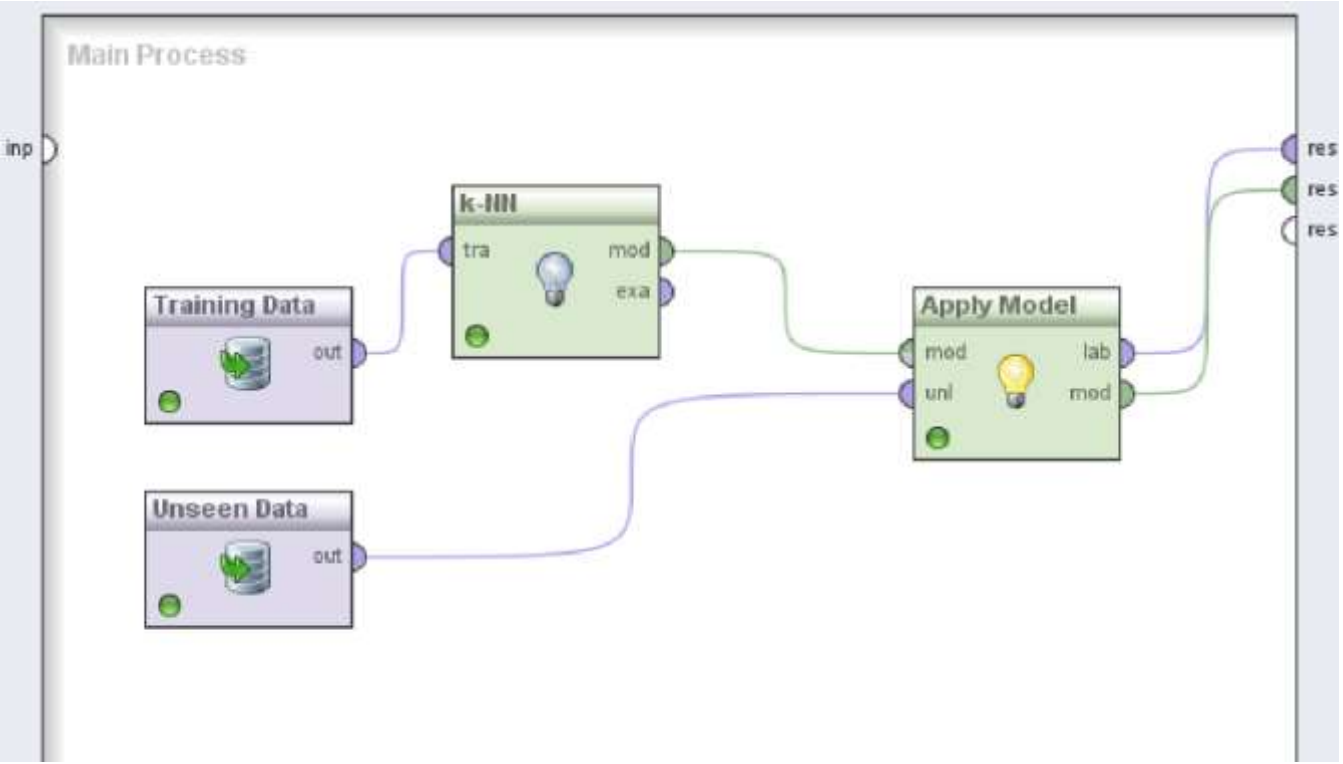
Decision Boundaries of a 1-NN Classifier



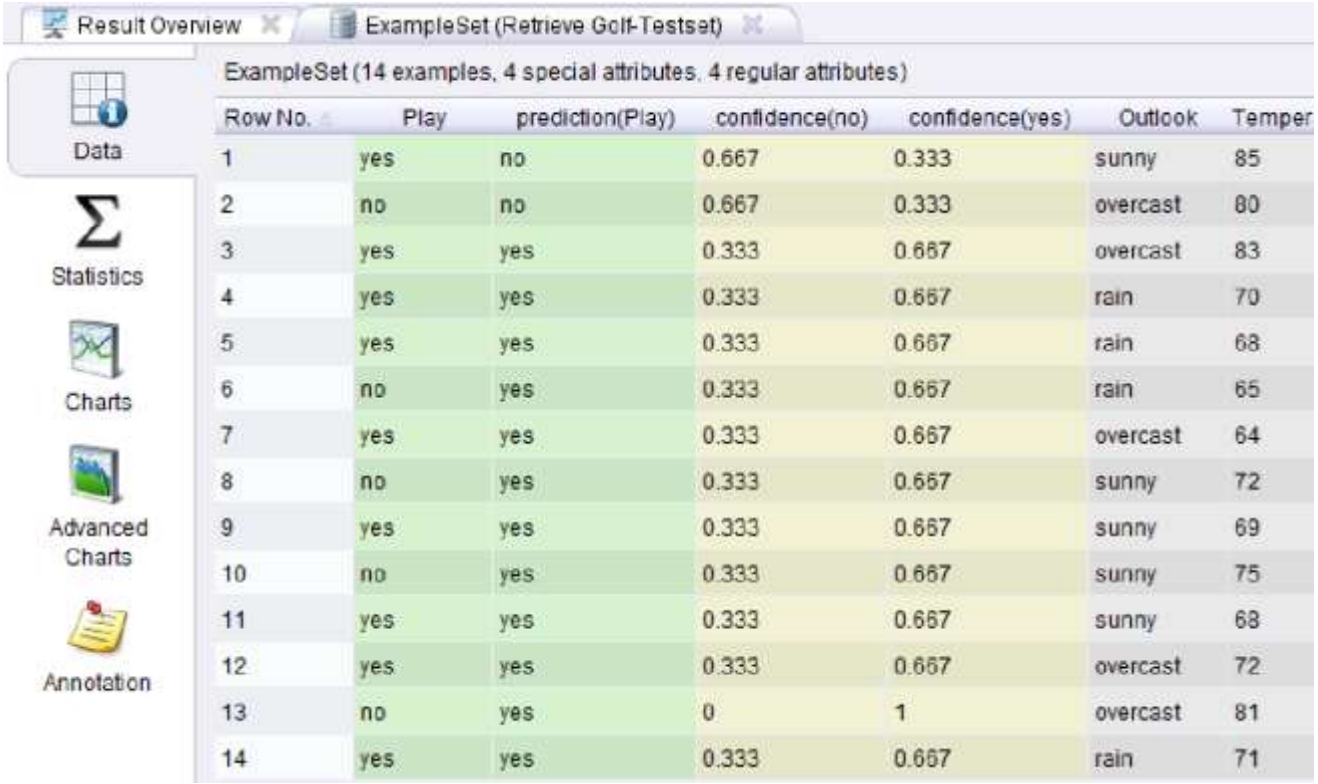
KNN in RapidMiner



Applying the Model



Resulting Dataset



ExampleSet (14 examples, 4 special attributes, 4 regular attributes)

Row No.	Play	prediction(Play)	confidence(no)	confidence(yes)	Outlook	Temper
1	yes	no	0.667	0.333	sunny	85
2	no	no	0.667	0.333	overcast	80
3	yes	yes	0.333	0.667	overcast	83
4	yes	yes	0.333	0.667	rain	70
5	yes	yes	0.333	0.667	rain	68
6	no	yes	0.333	0.667	rain	65
7	yes	yes	0.333	0.667	overcast	64
8	no	yes	0.333	0.667	sunny	72
9	yes	yes	0.333	0.667	sunny	69
10	no	yes	0.333	0.667	sunny	75
11	yes	yes	0.333	0.667	sunny	68
12	yes	yes	0.333	0.667	overcast	72
13	no	yes	0	1	overcast	81
14	yes	yes	0.333	0.667	rain	71

Lazy versus Eager Learning

Lazy Learning

- Instance-based learning approaches, like KNN, are also called lazy learning as no explicit knowledge (model) is learned
- Single goal: Classify unseen records as accurately as possible

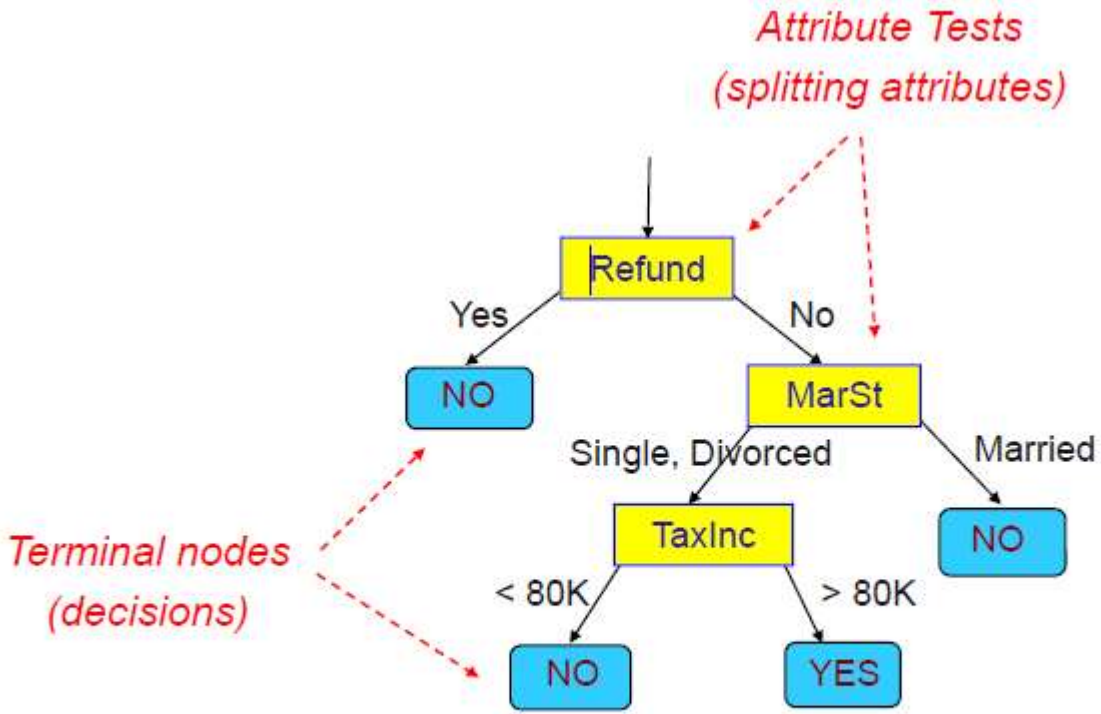
Eager Learning

- but actually, we might have two goals

1. classify unseen records
2. understand the application domain as a human

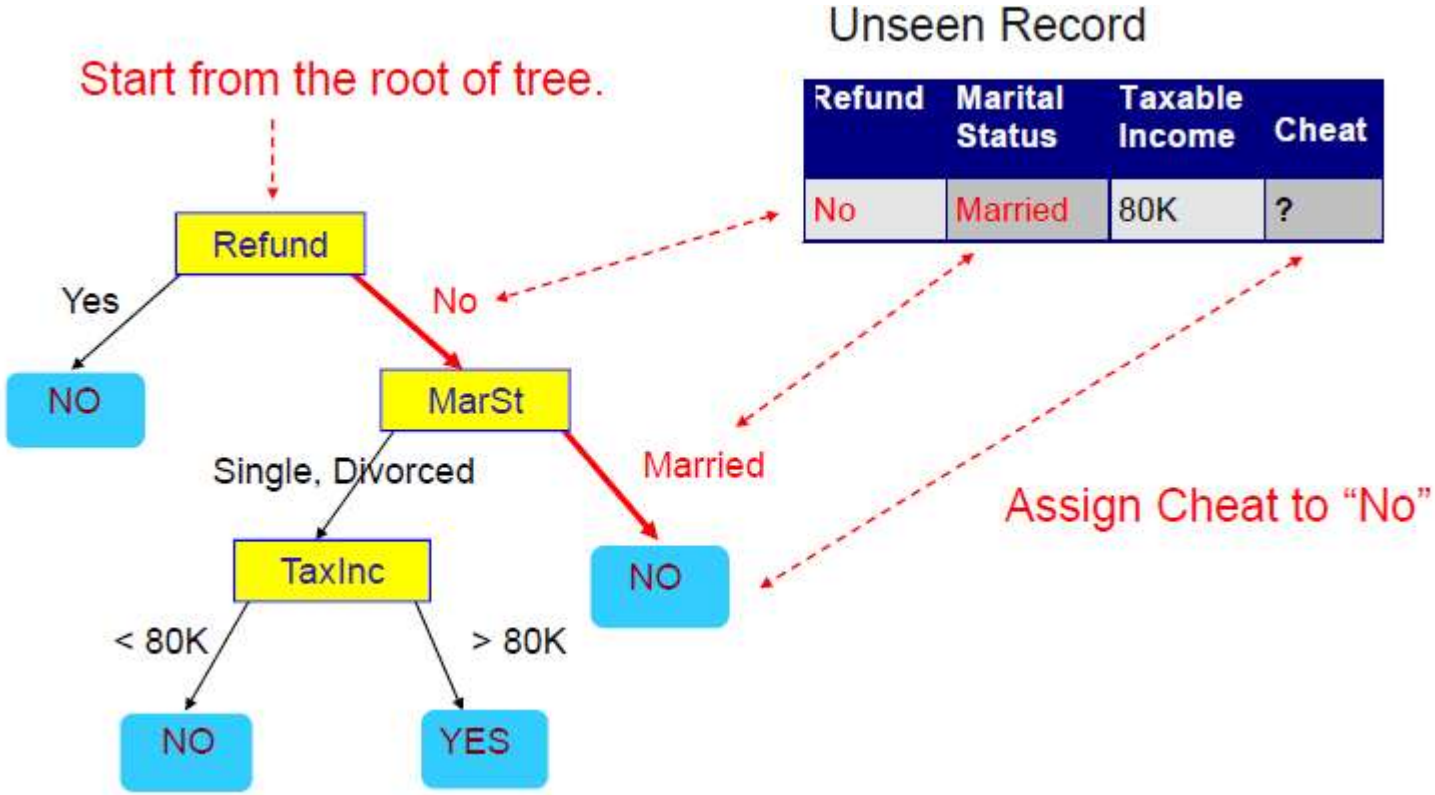
- Eager learning approaches generate models that are (might be) interpretable by humans
- Examples of eager techniques: Decision Tree Learning, Rule Learning

3. Decision Tree Classifiers

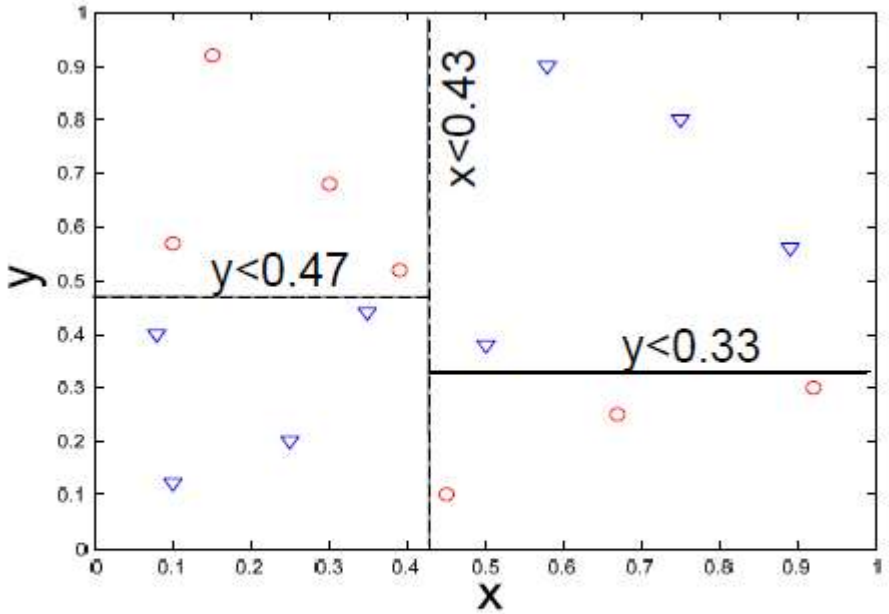


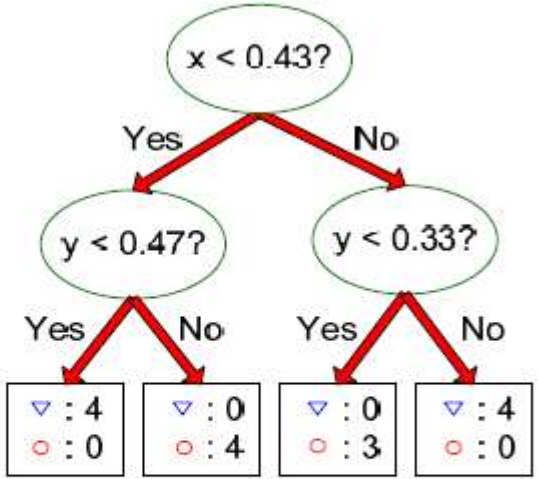
Decision trees encode a procedure for taking a classification decision.

Applying a Decision Tree to Unseen Data



Decision Boundary





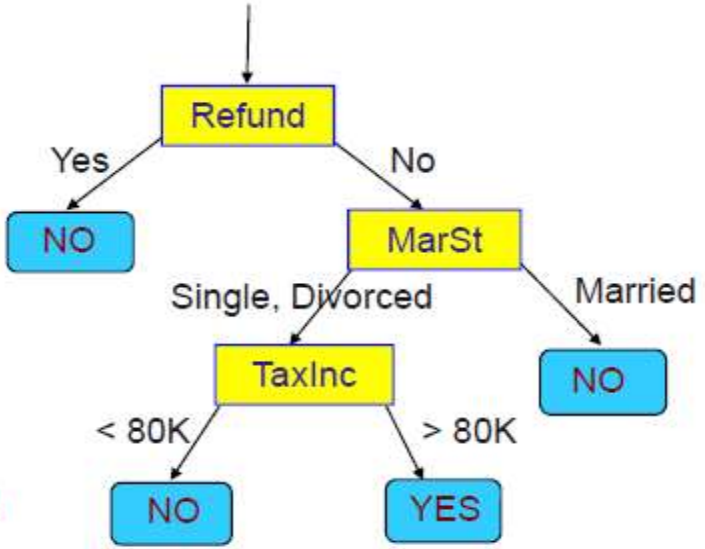
The decision boundaries are parallel to the axes because the test condition involves a single attribute at-a-time.

Decision Tree Induction

Tid	Refund	Marital Status	Taxable Income	Cheat
	categorical	categorical	continuous	class
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

Training Data

Learning Algorithm



Model: Decision Tree

Team Teaching

KOMDIGI
Kementerian Komunikasi dan Digital
Republik Indonesia

digitalent

Terima Kasih