CLASSIFICATION: DECISION TREES

Gökhan Akçapınar

(gokhana@hacettepe.edu.tr)

Seminar in Methodology and Statistics

John Nerbonne, Çağrı Çöltekin

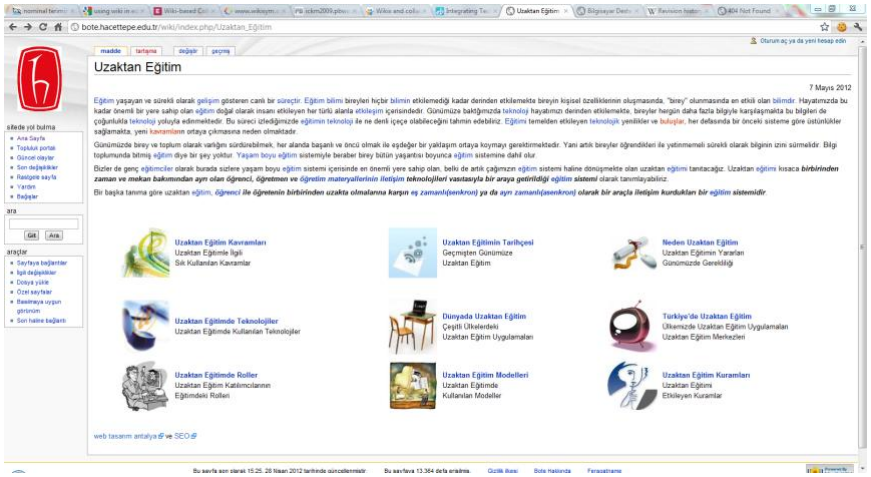
University of Groningen – May, 2012

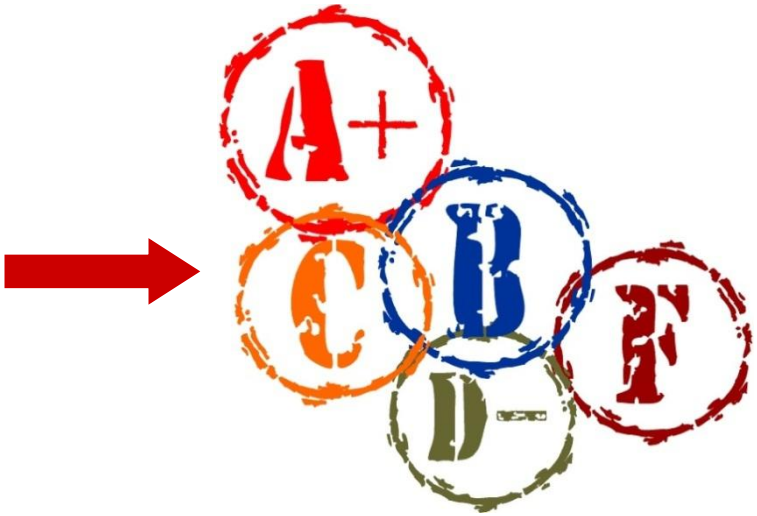
Outline

- Research question
- Background knowledge
- Data collection
- Classification with decision trees
- R example

Research Problem

- Predict student performance based on their activity data on wiki environment.





Wikis

- «A wiki is a website whose users can add, modify, or delete its content via a web browser.»

Wiki software

- Wikis are typically powered by wiki software and are often created collaboratively by multiple users.

Wiki in Education

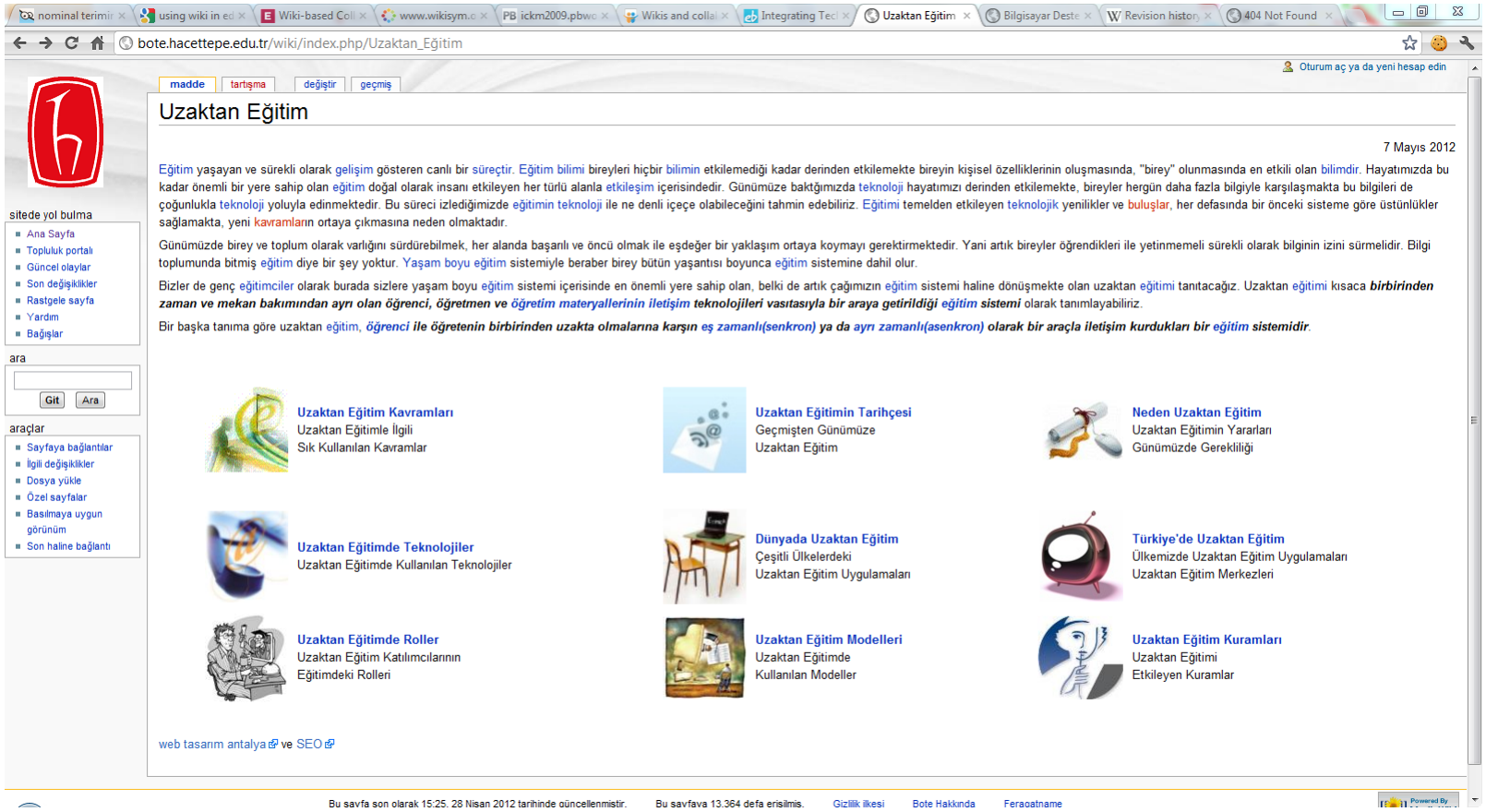
- Wikis are using mostly in group work and **collaboration**.
- Students create content, knowledge production

Assessment in Wiki?

- Assessment and to rate individual performance are the main problems in introducing wikis.

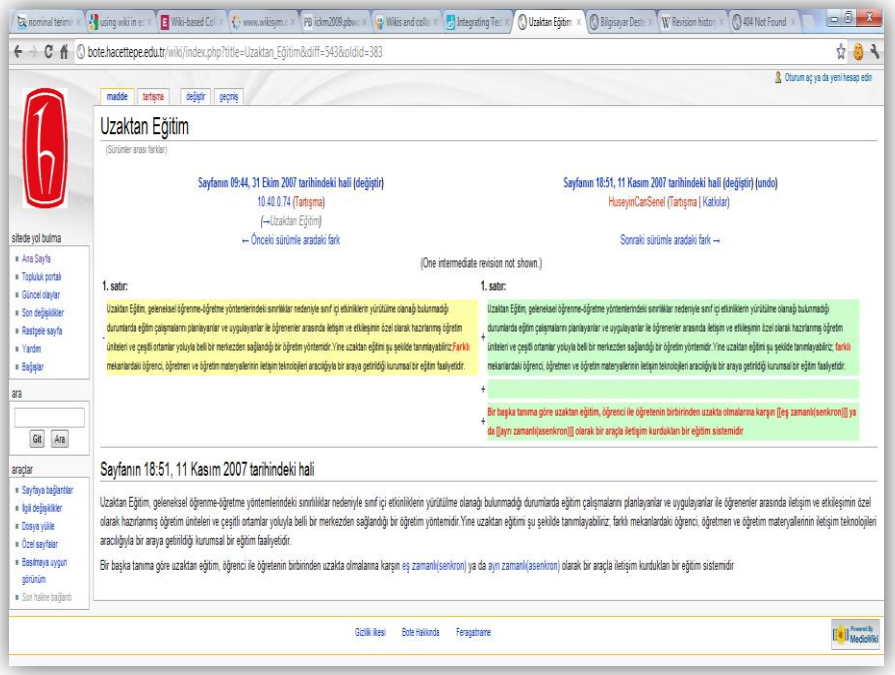
- If teachers cannot assess wiki work, we can not expect wiki to be adopted for education, despite the potential learning gains for students.

Why assessment is difficult?

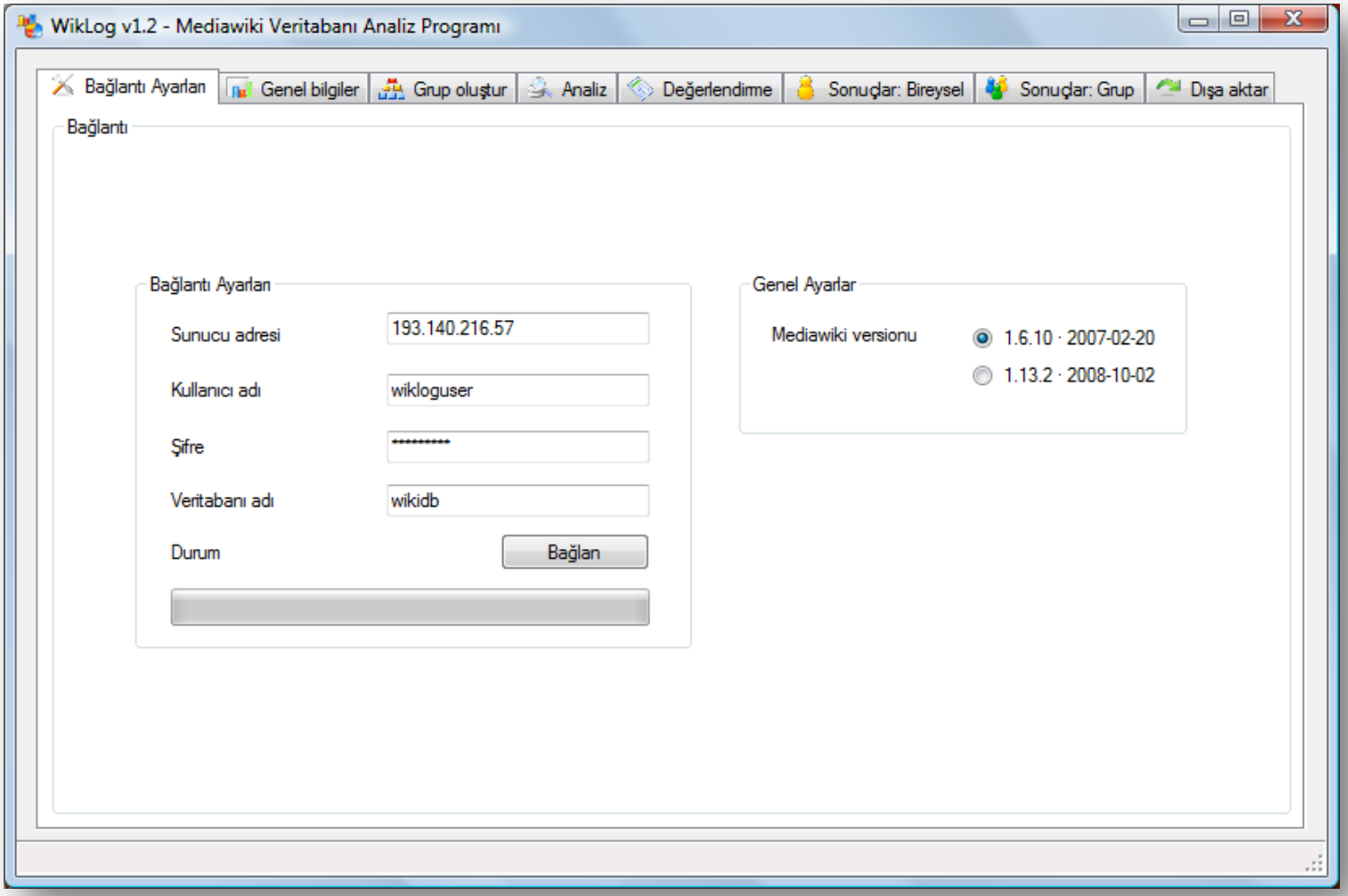


Sample wiki page

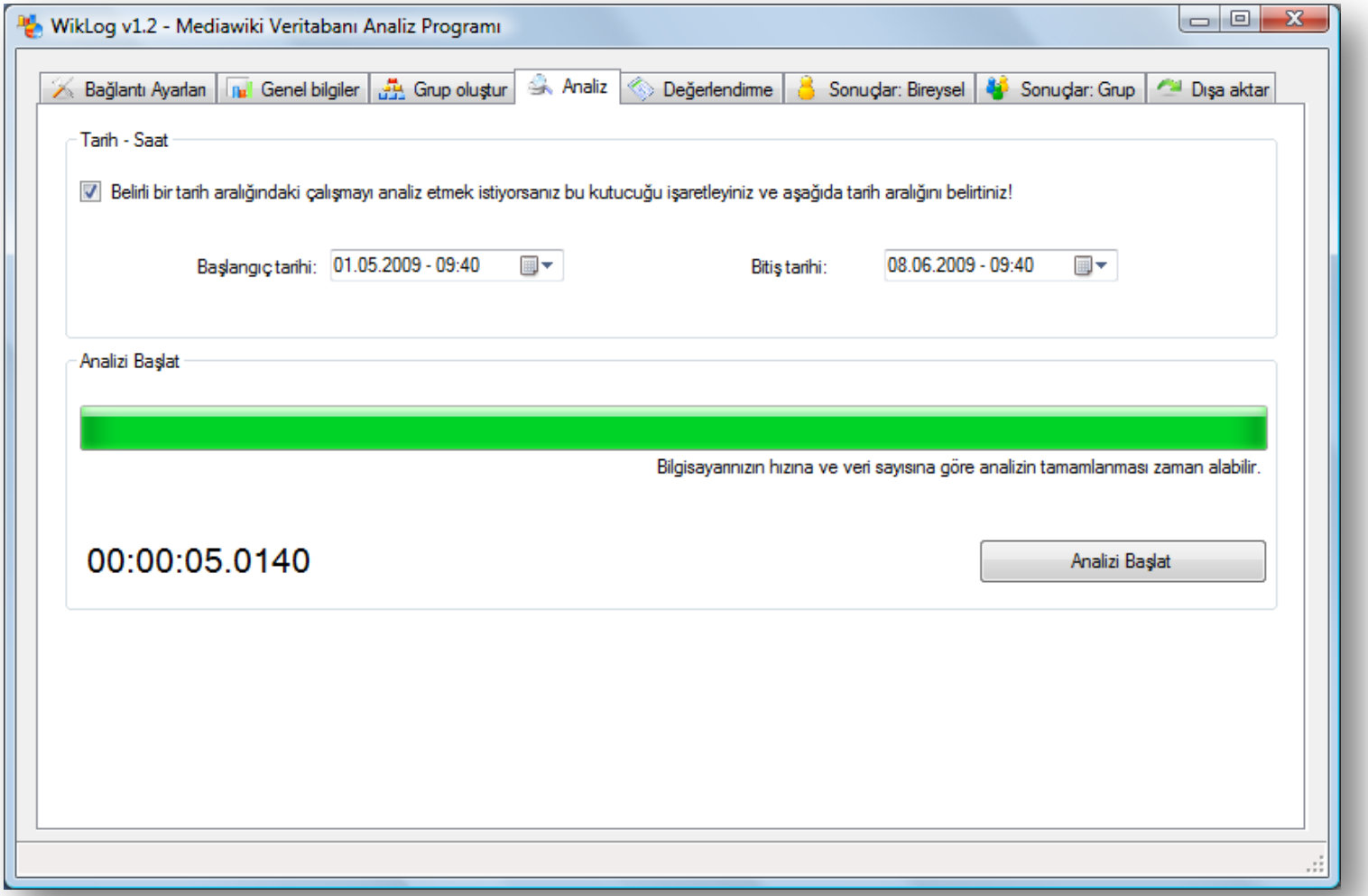
History Logs / Revisions



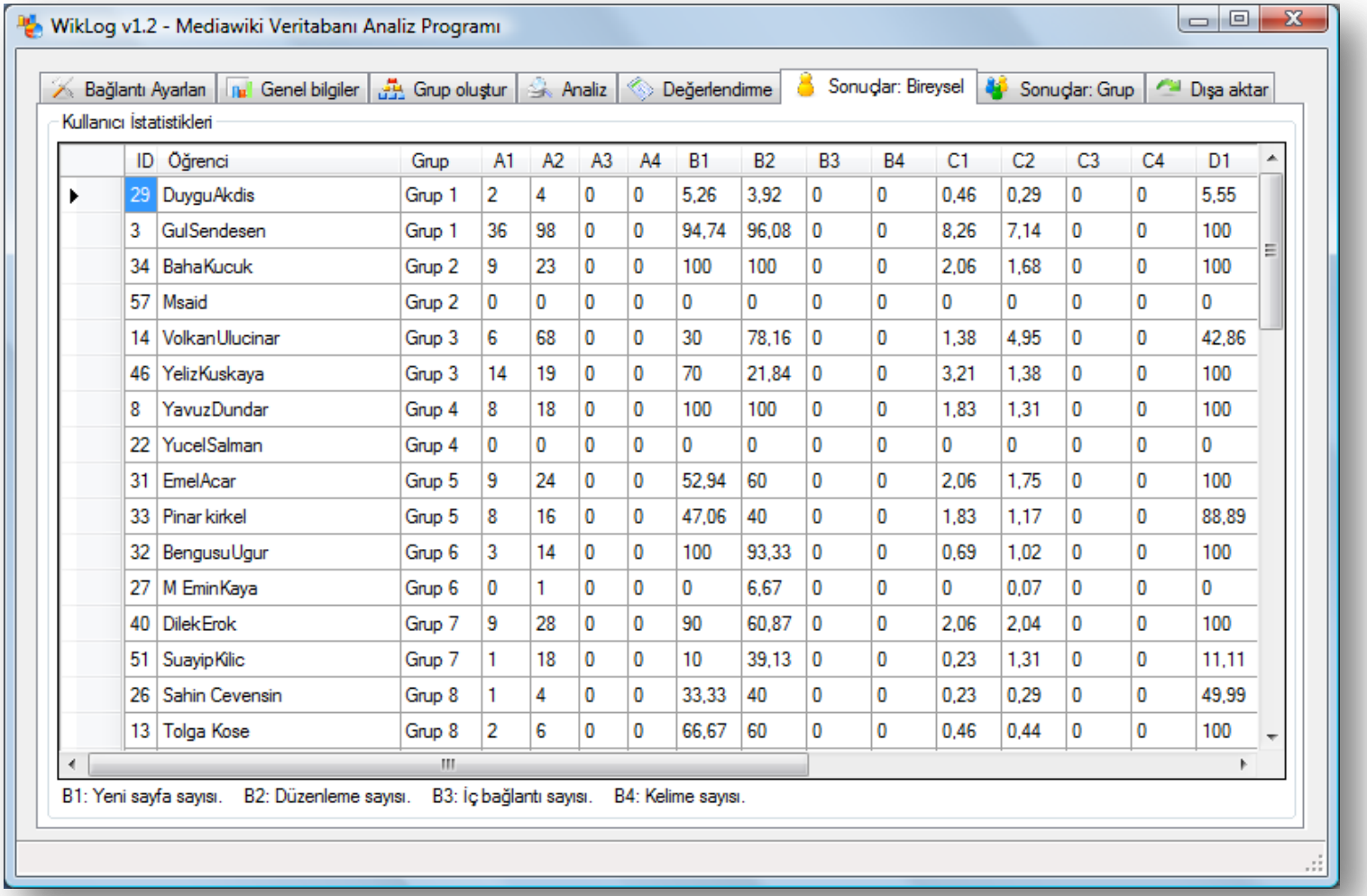
WikLog


WikLog v1.2 - Mediawiki Veritabanı Analiz Programı
Bağlantı Ayarları
Genel bilgiler
Grup oluştur
Analiz
Değerlendirme
Sonuçlar: Bireysel
Sonuçlar: Grup
Dışa aktar
Bağlantı
Bağlantı Ayarları
Sunucu adresi
193.140.216.57
Kullanıcı adı
wikloguser
Şifre
Veritabanı adı
wikidb
Durum
Bağlan
Genel Ayarlar
Mediawiki versionu
1.6.10 · 2007-02-20
1.13.2 · 2008-10-02

WikLog


WikLog v1.2 - Mediawiki Veritabanı Analiz Programı
Bağlantı Ayarları
Genel bilgiler
Grup oluştur
Analiz
Değerlendirme
Sonuçlar: Bireysel
Sonuçlar: Grup
Dışa aktar
Tarih - Saat
Belirli bir tarih aralığındaki çalışmayı analiz etmek istiyorsanız bu kutucuğu işaretleyiniz ve aşağıda tarih aralığını belirtiniz!
Başlangıç tarihi:
01.05.2009 - 09:40
Bitiş tarihi:
08.06.2009 - 09:40
Analizi Başlat
Bilgisayarınızın hızına ve veri sayısına göre analizin tamamlanması zaman alabilir.
00:00:05.0140
Analizi Başlat

WikLog



ID	Öğrenci	Grup	A1	A2	A3	A4	B1	B2	B3	B4	C1	C2	C3	C4	D1
29	DuyguAkdis	Grup 1	2	4	0	0	5,26	3,92	0	0	0,46	0,29	0	0	5,55
3	GulSendesen	Grup 1	36	98	0	0	94,74	96,08	0	0	8,26	7,14	0	0	100
34	BahaKucuk	Grup 2	9	23	0	0	100	100	0	0	2,06	1,68	0	0	100
57	Msaid	Grup 2	0	0	0	0	0	0	0	0	0	0	0	0	0
14	VolkanUlucinar	Grup 3	6	68	0	0	30	78,16	0	0	1,38	4,95	0	0	42,86
46	YelizKuskaya	Grup 3	14	19	0	0	70	21,84	0	0	3,21	1,38	0	0	100
8	YavuzDundar	Grup 4	8	18	0	0	100	100	0	0	1,83	1,31	0	0	100
22	YucelSalman	Grup 4	0	0	0	0	0	0	0	0	0	0	0	0	0
31	EmelAcar	Grup 5	9	24	0	0	52,94	60	0	0	2,06	1,75	0	0	100
33	Pinar kirkel	Grup 5	8	16	0	0	47,06	40	0	0	1,83	1,17	0	0	88,89
32	BengusuUgur	Grup 6	3	14	0	0	100	93,33	0	0	0,69	1,02	0	0	100
27	M EminKaya	Grup 6	0	1	0	0	0	6,67	0	0	0	0,07	0	0	0
40	DilekErok	Grup 7	9	28	0	0	90	60,87	0	0	2,06	2,04	0	0	100
51	SuayipKilic	Grup 7	1	18	0	0	10	39,13	0	0	0,23	1,31	0	0	11,11
26	Sahin Cevensin	Grup 8	1	4	0	0	33,33	40	0	0	0,23	0,29	0	0	49,99
13	Tolga Kose	Grup 8	2	6	0	0	66,67	60	0	0	0,46	0,44	0	0	100

Metrics (Attributes)

- PageCount: The number of pages created by the user.
- EditCount: The number of edits conducted by the user.
- LinkCount: The number of links created by the user.
- WordCount: The number of words created by the user.

Sample Data

ID	PageCount	EditCount	LinkCount	WordCount	Final Grade
1	55,00	334,00	30,00	5251,00	**B1**
2	5,00	194,00	0,00	430,00	**F**
3	37,00	267,00	243,00	9494,00	**A1**
4	75,00	402,00	138,00	1635,00	**A2**
5	24,00	183,00	1,00	2,00	**F**
6	40,00	232,00	83,00	1872,00	**C1**
7	8,00	128,00	13,00	1622,00	**F**
8	28,00	283,00	29,00	1361,00	**B2**
9	27,00	99,00	10,00	432,00	**D2**
10	32,00	113,00	9,00	1001,00	**F**

Class / Output Variable

A1
A2
B1 } High

B2
C1
C2
D1 } Medium

D2
F2
F3 } Low

ID	Final Grade
1	B1
2	F
3	A1
4	A2
5	F
6	C1
7	F
8	B2
9	D2
10	F

ID	Performance
1	High
2	Low
3	High
4	High
5	Low
6	Medium
7	Low
8	Medium
9	Low
10	Low

Research Problem

ID	PageCount	EditCount	LinkCount	WordCount	Performance
1	55,00	334,00	30,00	5251,00	**High**
2	5,00	194,00	0,00	430,00	**Low**
3	37,00	267,00	243,00	9494,00	**High**
4	75,00	402,00	138,00	1635,00	**High**
5	24,00	183,00	1,00	2,00	**Low**
6	40,00	232,00	83,00	1872,00	**Medium**
7	8,00	128,00	13,00	1622,00	**Low**
8	28,00	283,00	29,00	1361,00	**Medium**
9	27,00	99,00	10,00	432,00	**Low**
10	32,00	113,00	9,00	1001,00	**Low**

Research Problem

ID	PageCount	EditCount	LinkCount	WordCount	Performance
1	55,00	334,00	30,00	5251,00	**High**
2	5,00	194,00	0,00	430,00	**Low**
3	37,00	267,00	243,00	9494,00	**High**
4	75,00	402,00	138,00	1635,00	**High**
5	24,00	183,00	1,00	2,00	**Low**
6	40,00	232,00	83,00	1872,00	**Medium**
7	8,00	128,00	13,00	1622,00	**Low**
8	28,00	283,00	29,00	1361,00	**Medium**
9	27,00	99,00	10,00	432,00	**Low**
10	32,00	113,00	9,00	1001,00	**Low**

ID	PageCount	EditCount	LinkCount	WordCount	Performance
11	80,00	547,00	193,00	1269,00	?
12	65,00	271,00	273,00	2132,00	?
13	47,00	252,00	231,00	1213,00	?
14	106,00	278,00	399,00	2675,00	?
15	55,00	266,00	49,00	5713,00	?

Research Problem

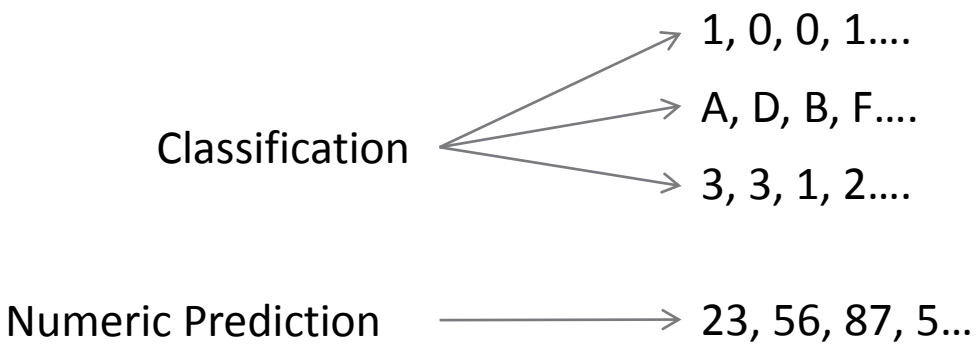
ID	PageCount	EditCount	LinkCount	WordCount	Performance
1	55,00	334,00	30,00	5251,00	**High**
2	5,00	194,00	0,00	430,00	**Low**
3	37,00	267,00	243,00	9494,00	**High**
4	75,00	402,00	138,00	1635,00	**High**
5	24,00	183,00	1,00	2,00	**Low**
6	40,00	232,00	83,00	1872,00	**M[illegible]**
7	8,00	128,00	13,00	1622,[illegible]	**Low**
8	28,00	283,00	29,00	1361,00	**Medium**
9	27,00	99,00	10,00	432,00	**Low**
10	32,00	113,00	9,00	1001,00	**Low**

Prediction

ID	PageCount	EditCount	LinkCount	WordCount	Performance
[illegible]	80,00	547,00	193,00	1269,00	**?**
12	65,00	271,00	273,00	2132,00	**?**
13	47,00	252,00	231,00	1213,00	**?**
14	106,00	278,00	399,00	2675,00	**?**
15	55,00	266,00	49,00	5713,00	**?**

Prediction: Classification or Numeric Prediction?

- The objective of prediction is to estimate the unknown value of a variable.
- In education, the values can be knowledge, score, or mark.
- This value can be numerical/continuous value (regression task) or categorical/discrete value (classification task).



Classification

- Classification is a procedure in which individual items are placed into groups based on quantitative information regarding one or more characteristics inherent in the items and based on a training set of previously labeled items.

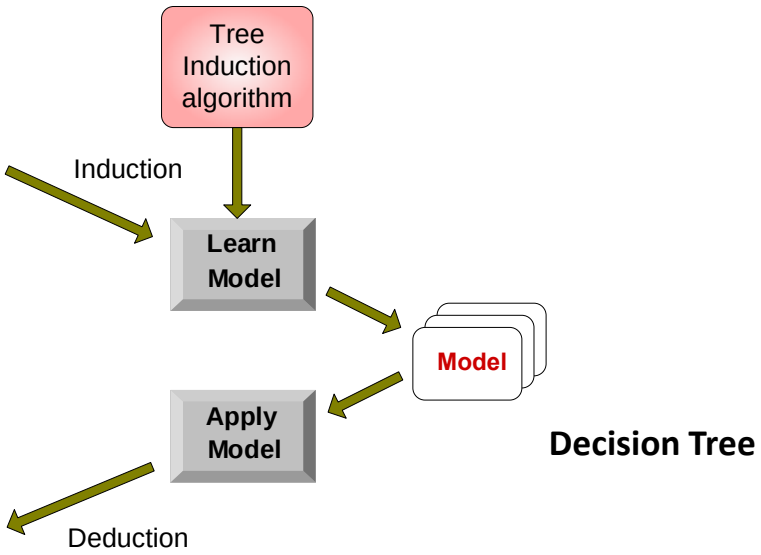
Classification—A Two-Step Process

ID	PageCount	EditCount	LinkCount	WordCount	Performance
1	55,00	334,00	30,00	5251,00	High
2	5,00	194,00	0,00	430,00	Low
3	37,00	267,00	243,00	9494,00	High
4	75,00	402,00	138,00	1635,00	High
5	24,00	183,00	1,00	2,00	Low
6	40,00	232,00	83,00	1872,00	Medium
7	8,00	128,00	13,00	1622,00	Low
8	28,00	283,00	29,00	1361,00	Medium
9	27,00	99,00	10,00	432,00	Low
10	32,00	113,00	9,00	1001,00	Low

Training Set

ID	PageCount	EditCount	LinkCount	WordCount	Performance
11	80,00	547,00	193,00	1269,00	?
12	65,00	271,00	273,00	2132,00	?
13	47,00	252,00	231,00	1213,00	?
14	106,00	278,00	399,00	2675,00	?
15	55,00	266,00	49,00	5713,00	?

Test Set



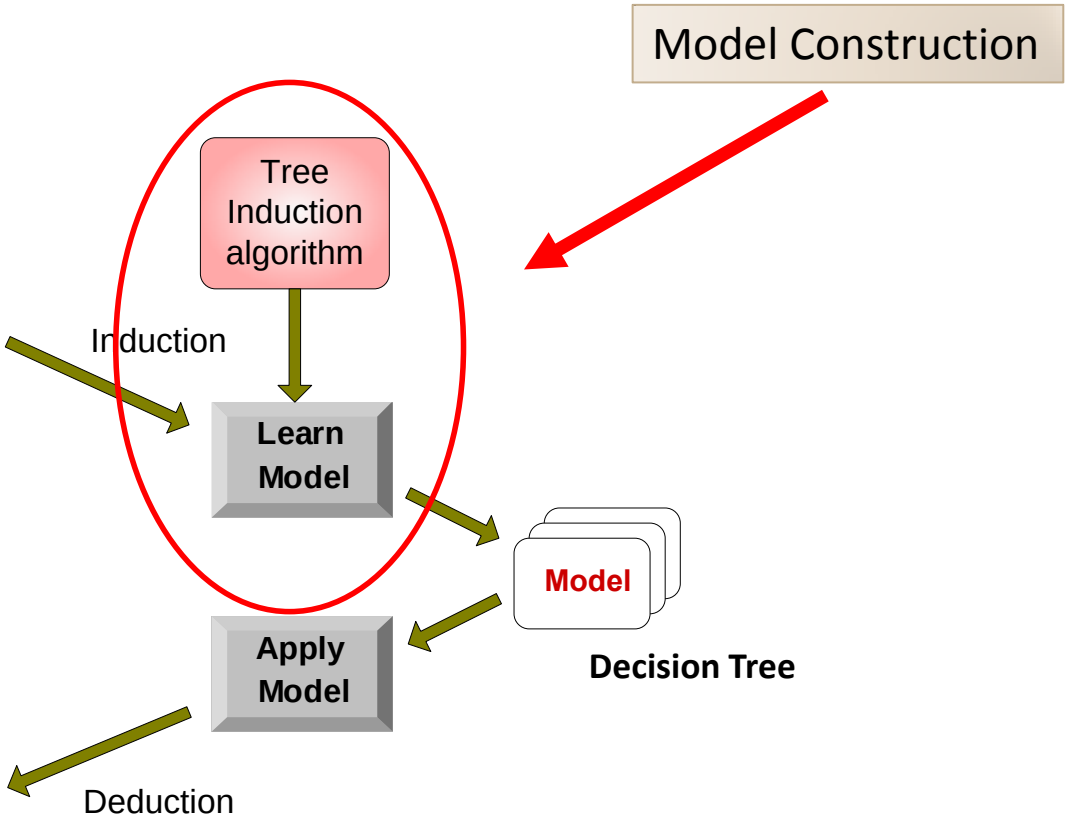
Classification—A Two-Step Process

ID	PageCount	EditCount	LinkCount	WordCount	Performance
1	55,00	334,00	30,00	5251,00	High
2	5,00	194,00	0,00	430,00	Low
3	37,00	267,00	243,00	9494,00	High
4	75,00	402,00	138,00	1635,00	High
5	24,00	183,00	1,00	2,00	Low
6	40,00	232,00	83,00	1872,00	Medium
7	8,00	128,00	13,00	1622,00	Low
8	28,00	283,00	29,00	1361,00	Medium
9	27,00	99,00	10,00	432,00	Low
10	32,00	113,00	9,00	1001,00	Low

Training Set

ID	PageCount	EditCount	LinkCount	WordCount	Performance
11	80,00	547,00	193,00	1269,00	?
12	65,00	271,00	273,00	2132,00	?
13	47,00	252,00	231,00	1213,00	?
14	106,00	278,00	399,00	2675,00	?
15	55,00	266,00	49,00	5713,00	?

Test Set



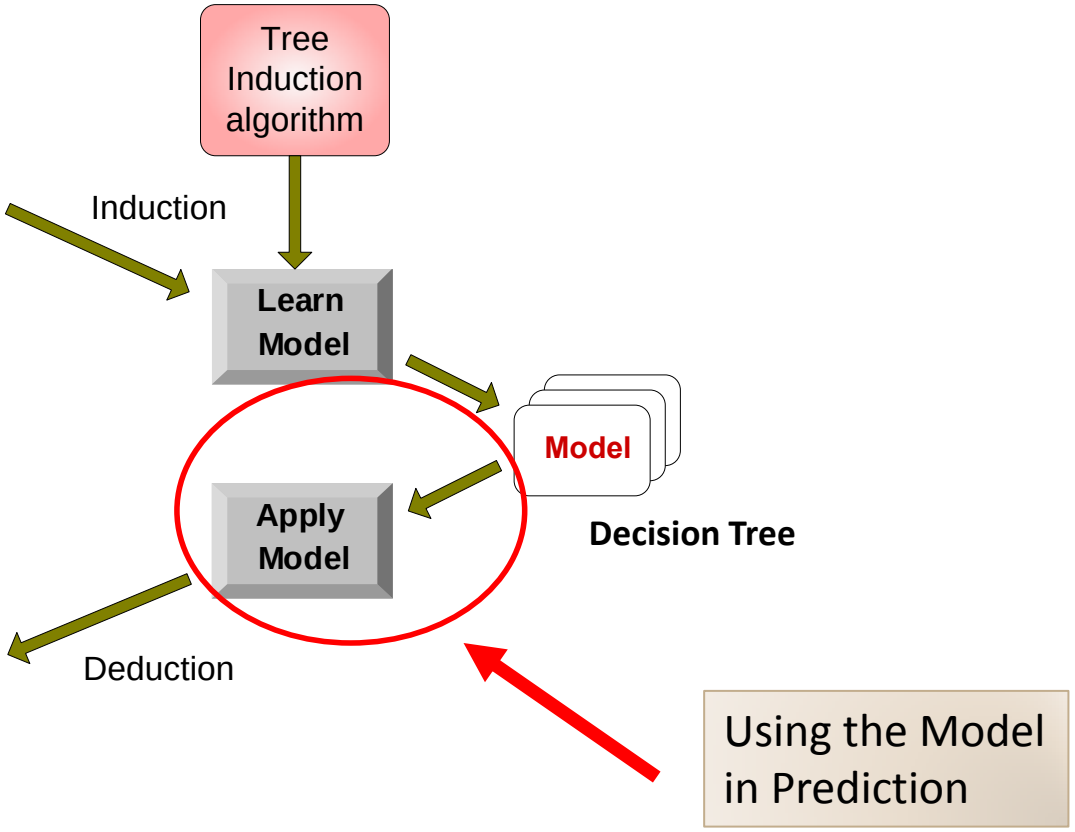
Classification—A Two-Step Process

ID	PageCount	EditCount	LinkCount	WordCount	Performance
1	55,00	334,00	30,00	5251,00	High
2	5,00	194,00	0,00	430,00	Low
3	37,00	267,00	243,00	9494,00	High
4	75,00	402,00	138,00	1635,00	High
5	24,00	183,00	1,00	2,00	Low
6	40,00	232,00	83,00	1872,00	Medium
7	8,00	128,00	13,00	1622,00	Low
8	28,00	283,00	29,00	1361,00	Medium
9	27,00	99,00	10,00	432,00	Low
10	32,00	113,00	9,00	1001,00	Low

Training Set

ID	PageCount	EditCount	LinkCount	WordCount	Performance
11	80,00	547,00	193,00	1269,00	?
12	65,00	271,00	273,00	2132,00	?
13	47,00	252,00	231,00	1213,00	?
14	106,00	278,00	399,00	2675,00	?
15	55,00	266,00	49,00	5713,00	?

Test Set

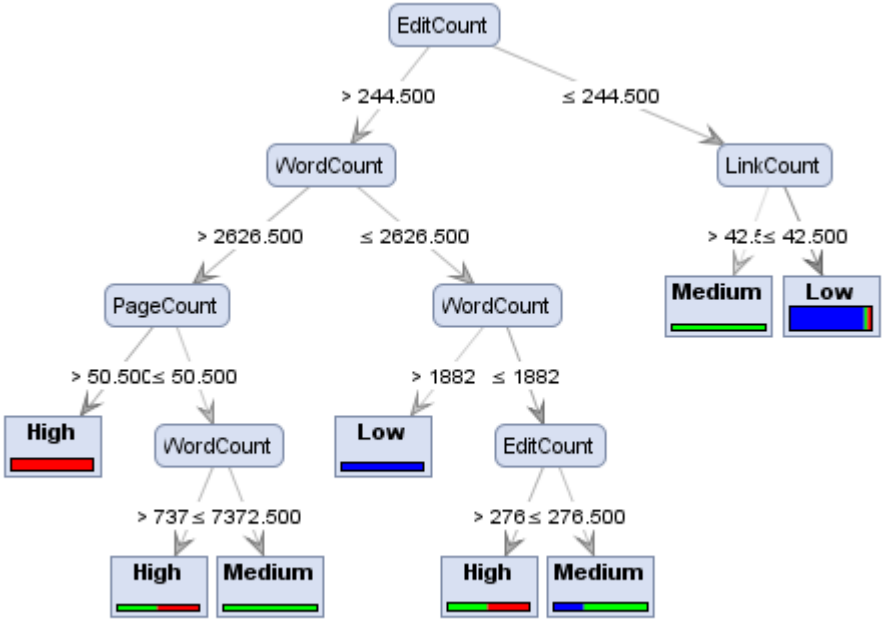


Classification Techniques

- Decision Tree based Methods
- Rule-based Methods
- Memory based reasoning
- Neural Networks
- Naïve Bayes and Bayesian Belief Networks
- Support Vector Machines

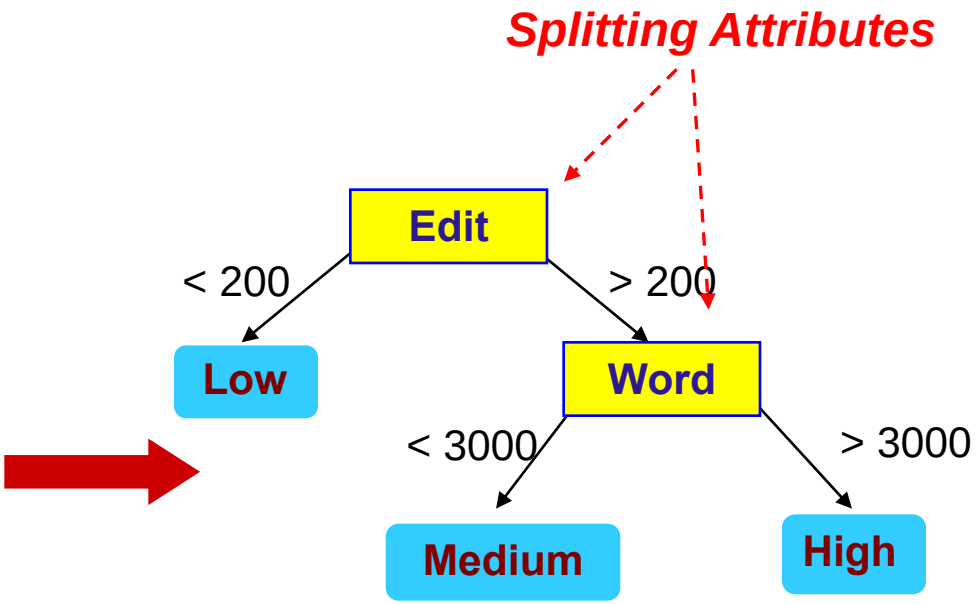
Classification Techniques

- Decision Tree based Methods



Example of a Decision Tree

ID	PageCount	EditCount	LinkCount	WordCount	Performance
1	55,00	334,00	30,00	5251,00	**High**
2	5,00	194,00	0,00	430,00	**Low**
3	37,00	267,00	243,00	9494,00	**High**
4	75,00	402,00	138,00	1635,00	**High**
5	24,00	183,00	1,00	2,00	**Low**
6	40,00	232,00	83,00	1872,00	**Medium**
7	8,00	128,00	13,00	1622,00	**Low**
8	28,00	283,00	29,00	1361,00	**Medium**
9	27,00	99,00	10,00	432,00	**Low**
10	32,00	113,00	9,00	1001,00	**Low**

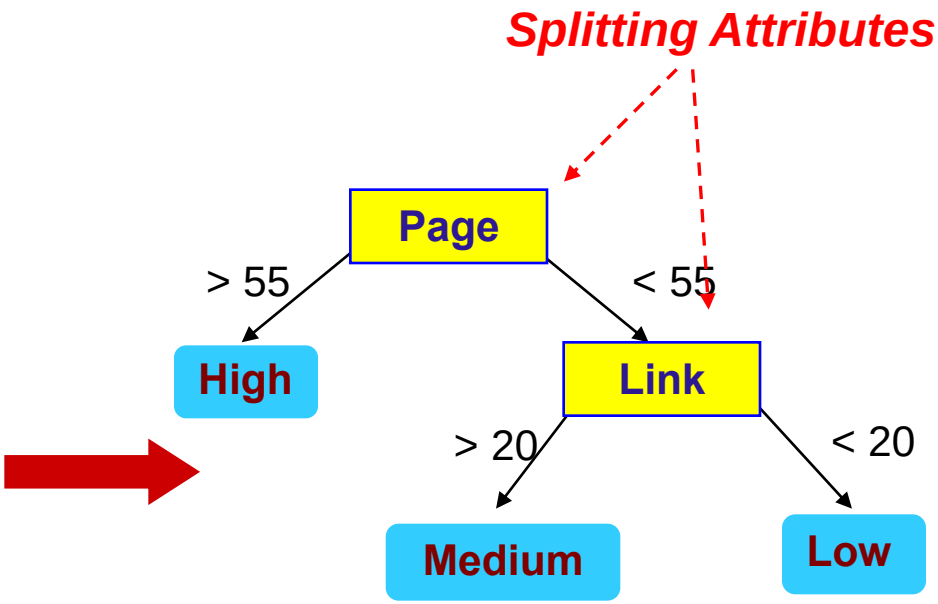


Training Data

Model: Decision Tree

Example of a Decision Tree

ID	PageCount	EditCount	LinkCount	WordCount	Performance
1	55,00	334,00	30,00	5251,00	**High**
2	5,00	194,00	0,00	430,00	**Low**
3	37,00	267,00	243,00	9494,00	**High**
4	75,00	402,00	138,00	1635,00	**High**
5	24,00	183,00	1,00	2,00	**Low**
6	40,00	232,00	83,00	1872,00	**Medium**
7	8,00	128,00	13,00	1622,00	**Low**
8	28,00	283,00	29,00	1361,00	**Medium**
9	27,00	99,00	10,00	432,00	**Low**
10	32,00	113,00	9,00	1001,00	**Low**



Training Data

Model: Decision Tree

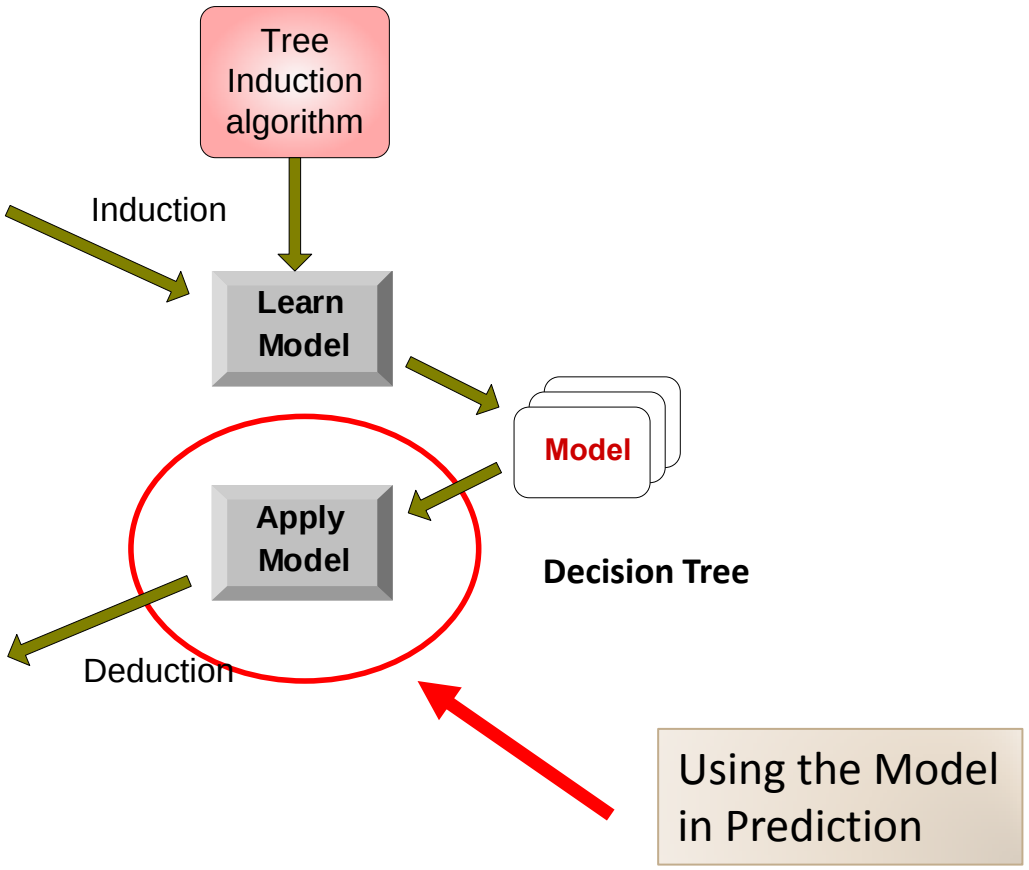
Example of Apply Model to Test Data

ID	PageCount	EditCount	LinkCount	WordCount	Performance
1	55,00	334,00	30,00	5251,00	High
2	5,00	194,00	0,00	430,00	Low
3	37,00	267,00	243,00	9494,00	High
4	75,00	402,00	138,00	1635,00	High
5	24,00	183,00	1,00	2,00	Low
6	40,00	232,00	83,00	1872,00	Medium
7	8,00	128,00	13,00	1622,00	Low
8	28,00	283,00	29,00	1361,00	Medium
9	27,00	99,00	10,00	432,00	Low
10	32,00	113,00	9,00	1001,00	Low

Training Set

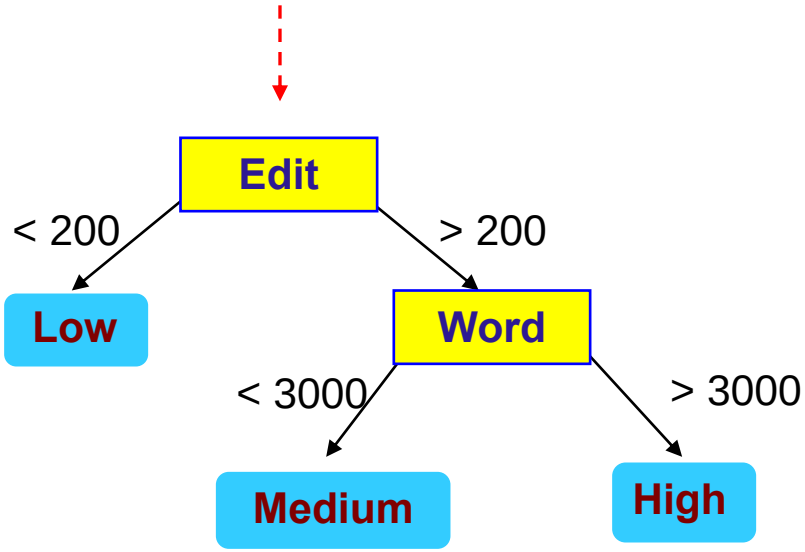
ID	PageCount	EditCount	LinkCount	WordCount	Performance
11	80,00	547,00	193,00	1269,00	?
12	65,00	271,00	273,00	2132,00	?
13	47,00	252,00	231,00	1213,00	?
14	106,00	278,00	399,00	2675,00	?
15	55,00	266,00	49,00	5713,00	?

Test Set



Apply Model to Test Data

Start from the root of tree.



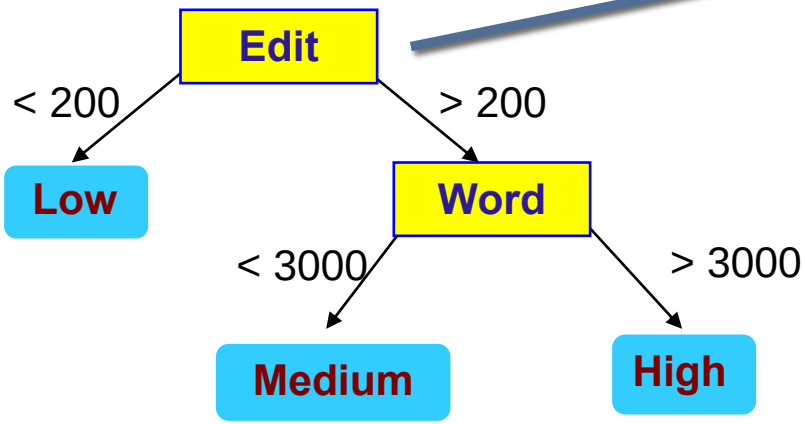
Test Data

ID	PageCount	EditCount	LinkCount	WordCount	Performance
15	55,00	266,00	49,00	5713,00	?

Apply Model to Test Data

Test Data

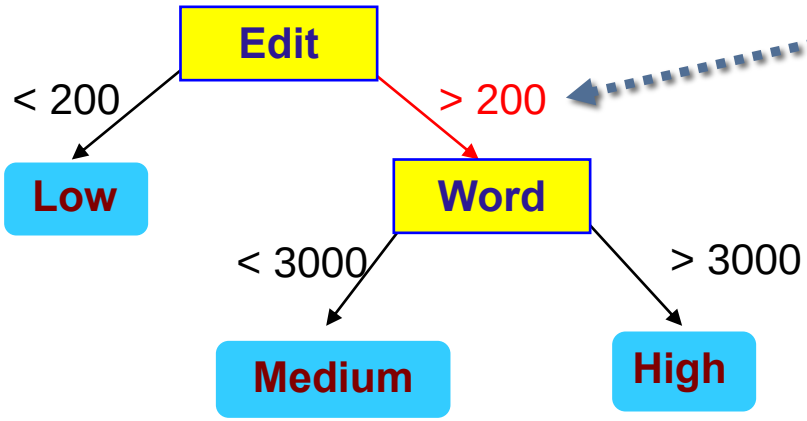
ID	PageCount	EditCount	LinkCount	WordCount	Performance
15	55,00	266,00	49,00	5713,00	?



Apply Model to Test Data

Test Data

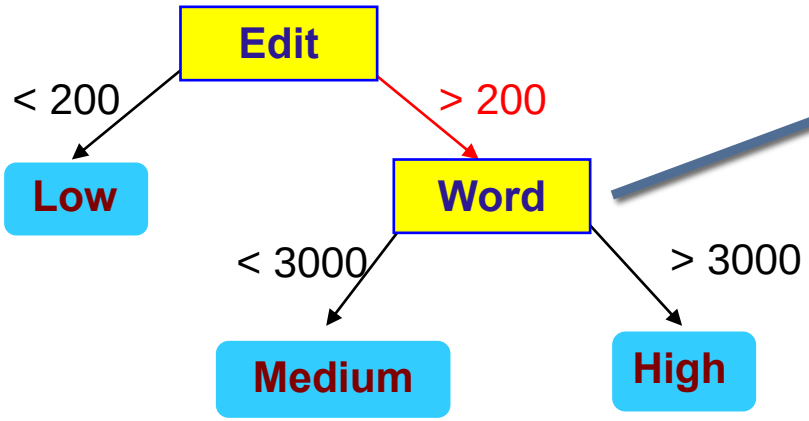
ID	PageCount	EditCount	LinkCount	WordCount	Performance
15	55,00	266,00	49,00	5713,00	?



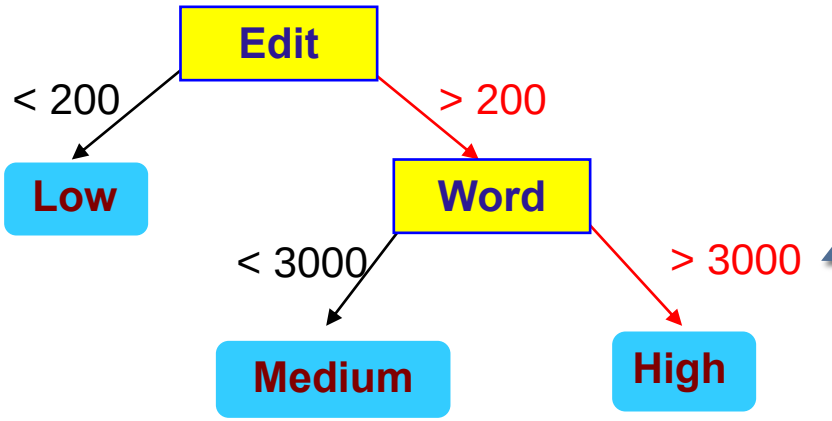
Apply Model to Test Data

Test Data

ID	PageCount	EditCount	LinkCount	WordCount	Performance
15	55,00	266,00	49,00	5713,00	?



Apply Model to Test Data



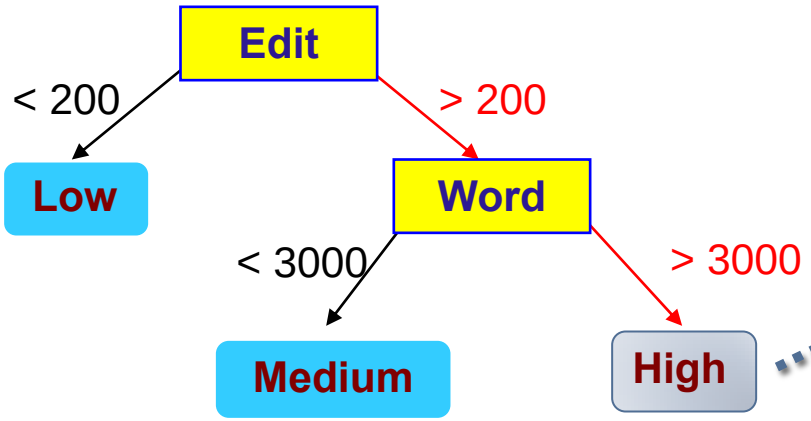
Test Data

ID	PageCount	EditCount	LinkCount	WordCount	Performance
15	55,00	266,00	49,00	5713,00	?

Apply Model to Test Data

Test Data

ID	PageCount	EditCount	LinkCount	WordCount	Performance
15	55,00	266,00	49,00	5713,00	?



Assign Performance to "High"

Choosing the Splitting Attribute

- Typical goodness functions:
  - information gain (ID3/C4.5)
  - information gain ratio
  - gini index

- Which is the best attribute?
  - The one which will result in the smallest tree
  - Choose the attribute that produces the “purest” nodes

- Strategy: choose attribute that results in greatest information gain

Information Gain (ID3/C4.5)

- Select the attribute with the highest information gain
- Expected information (entropy) needed to classify a tuple in D:

$$Info(D) = -\sum_{i=1}^{m} p_i \log_2(p_i)$$

- Information needed (after using A to split D into v partitions) to classify D:

$$Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times Info(D_j)$$

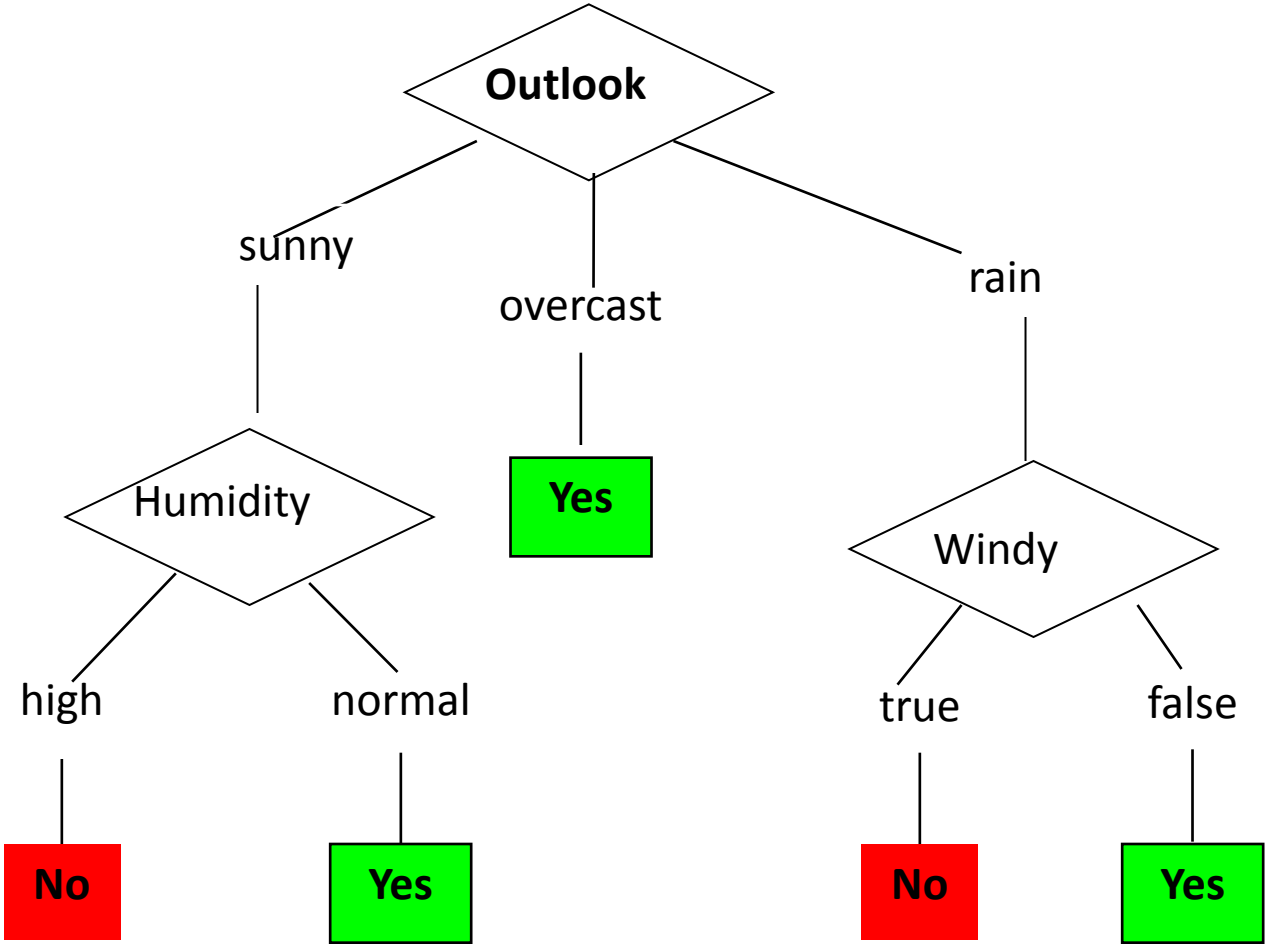
- Information gained by branching on attribute A

$$Gain(A) = Info(D) - Info_A(D)$$

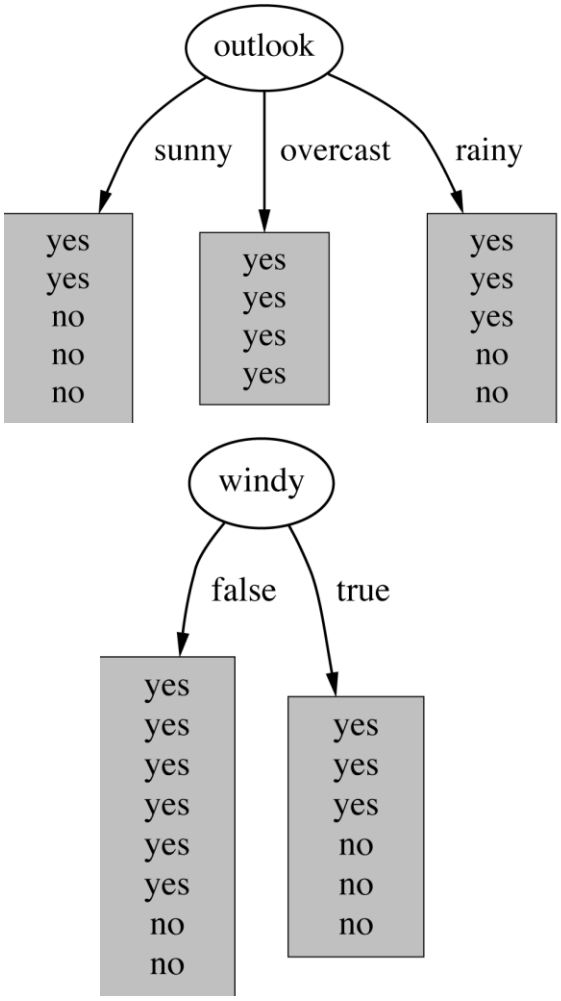
When do I play tennis?

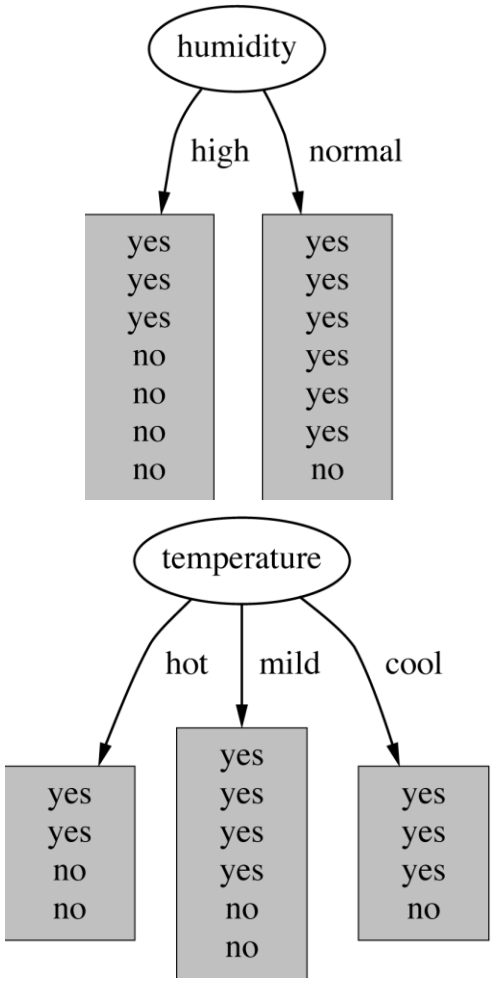
Outlook	Temperature	Humidity	Windy	Play?
sunny	hot	high	false	No
sunny	hot	high	true	No
overcast	hot	high	false	Yes
rain	mild	high	false	Yes
rain	cool	normal	false	Yes
rain	cool	normal	true	No
overcast	cool	normal	true	Yes
sunny	mild	high	false	No
sunny	cool	normal	false	Yes
rain	mild	normal	false	Yes
sunny	mild	normal	true	Yes
overcast	mild	high	true	Yes
overcast	hot	normal	false	Yes
rain	mild	high	true	No

Example Tree for “Play?”



Which attribute to select?





Example: attribute “Outlook”, 2

- “Outlook” = “Sunny”:

$$\text{info}([2,3]) = \text{entropy}(2/5,3/5) = -2/5\log(2/5) - 3/5\log(3/5) = 0.971\,\text{bits}$$

- “Outlook” = “Overcast”:

$$\text{info}([4,0]) = \text{entropy}(1,0) = -1\log(1) - 0\log(0) = 0\,\text{bits}$$

- “Outlook” = “Rainy”:

$$\text{info}([3,2]) = \text{entropy}(3/5,2/5) = -3/5\log(3/5) - 2/5\log(2/5) = 0.971\,\text{bits}$$

- Expected information for attribute:

$$\text{info}([3,2],[4,0],[3,2]) = (5/14)\times 0.971 + (4/14)\times 0 + (5/14)\times 0.971$$
$$= 0.693\,\text{bits}$$

Computing the information gain

- Information gain:

(information before split) – (information after split)

$$\text{gain}(\text{" Outlook"}) = \text{info}([9,5]) - \text{info}([2,3],[4,0],[3,2]) = 0.940 - 0.693$$
$$= 0.247 \text{ bits}$$

- Compute for attribute “Humidity”

Example: attribute “Humidity”

- “Humidity” = “High”:

$$\text{info}([3,4]) = \text{entropy}(3/7,4/7) = -3/7\log(3/7) - 4/7\log(4/7) = 0.985\,\text{bits}$$

- “Humidity” = “Normal”:

$$\text{info}([6,1]) = \text{entropy}(6/7,1/7) = -6/7\log(6/7) - 1/7\log(1/7) = 0.592\,\text{bits}$$

- Expected information for attribute:

$$\text{info}([3,4],[6,1]) = (7/14)\times 0.985 + (7/14)\times 0.592 \quad = 0.79\,\text{bits}$$

- Information Gain:

$$\text{info}([9,5]) - \text{info}([3,4],[6,1]) = 0.940 - 0.788 = 0.152$$

Computing the information gain

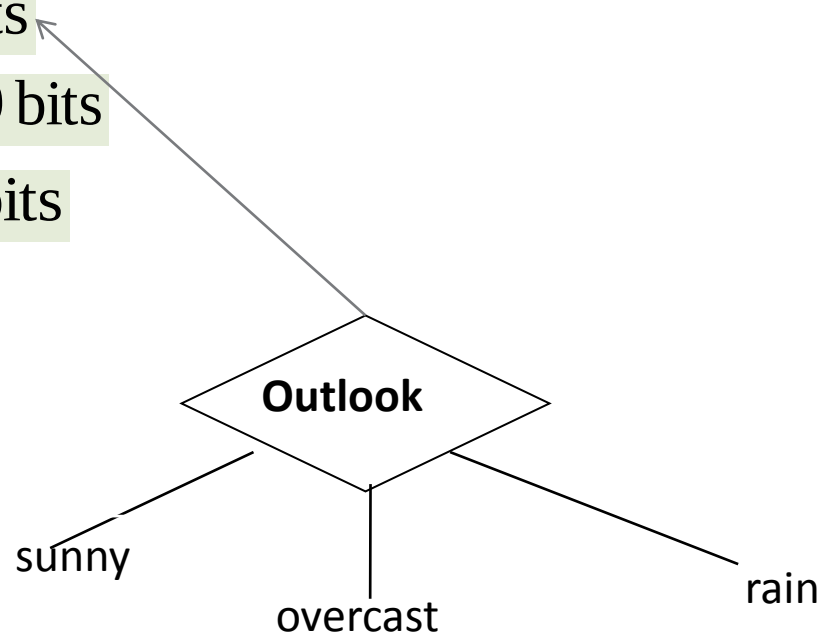
- Information gain for attributes from weather data:

$$\text{gain}(\text{"Outlook"}) = 0.247 \text{ bits}$$

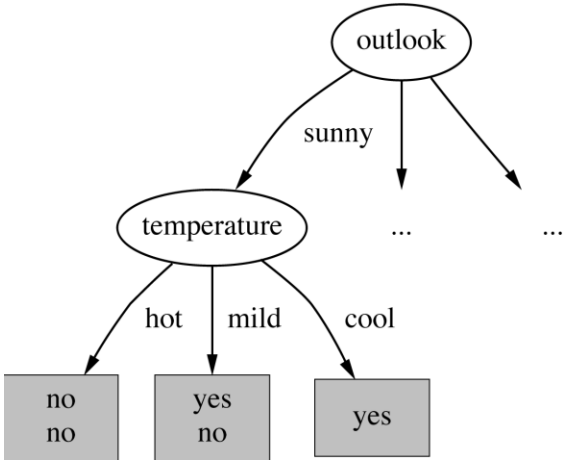
$$\text{gain}(\text{"Temperature"}) = 0.029 \text{ bits}$$

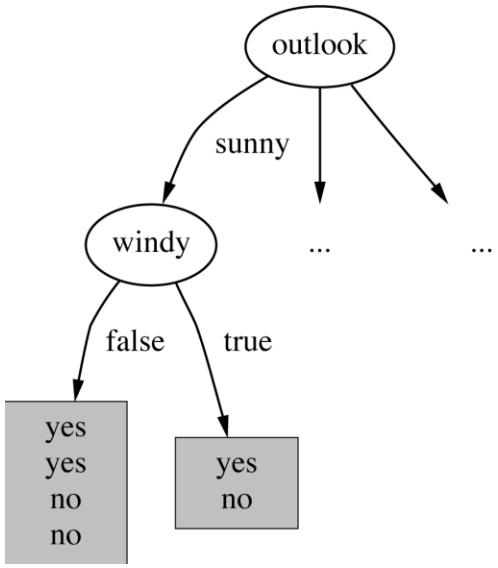
$$\text{gain}(\text{"Humidity"}) = 0.152 \text{ bits}$$

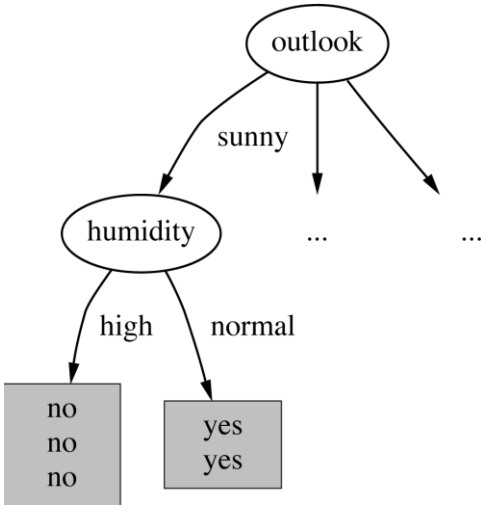
$$\text{gain}(\text{"Windy"}) = 0.048 \text{ bits}$$



Continuing to split





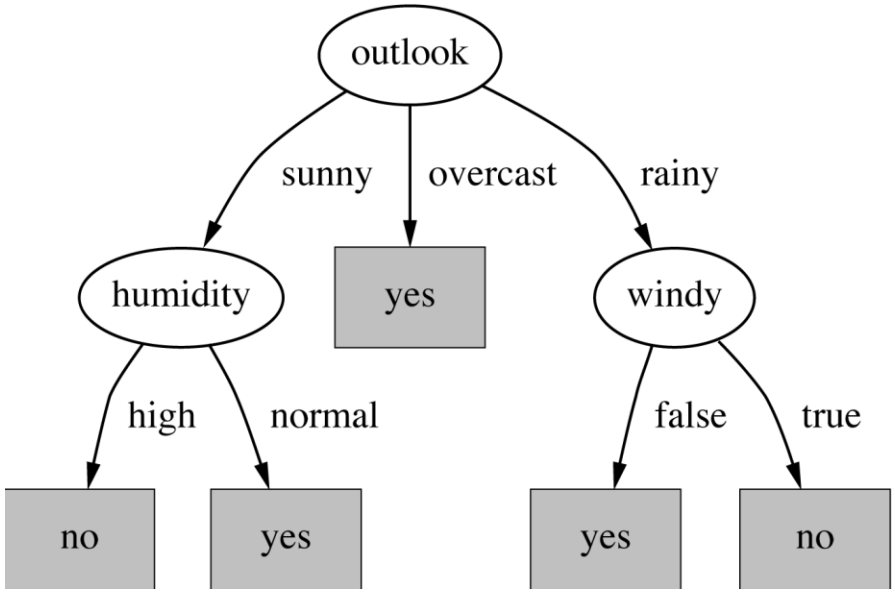


$$\text{gain}(\text{"Humidity"}) = 0.971\,\text{bits}$$

$$\text{gain}(\text{"Temperature"}) = 0.571\,\text{bits}$$

$$\text{gain}(\text{"Windy"}) = 0.020\,\text{bits}$$

The final decision tree



$\Rightarrow$ Splitting stops when data can't be split any further

Rpart()

- install.packages('rpart')
- library(rpart)
- data = read.xls("C://tree_data.xls",colNames = TRUE)
- results = rpart(Performance~PageCount+EditCount+LinkCount+WordCount, data=data, method="class", parms=list(split='information'))
- printcp(results)
- plot(results)
- text(results)

CLASSIFICATION: DECISION TREES

Gökhan Akçapınar

(gokhana@hacettepe.edu.tr)

Seminar in Methodology and Statistics

John Nerbonne, Çağrı Çöltekin

University of Groningen – May, 2012