

CLASSIFICATION: DECISION TREES

Gökhan Akçapınar

(gokhana@hacettepe.edu.tr)

Seminar in Methodology and Statistics

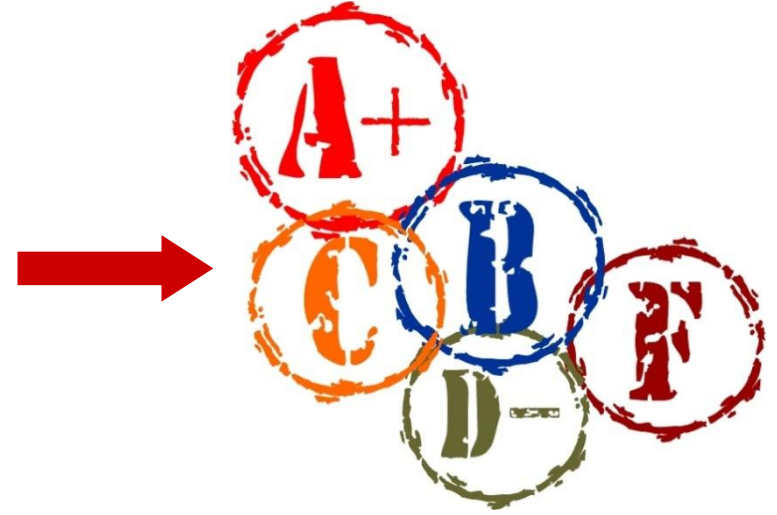
John Nerbonne, Çağrı Çöltekin

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?

nominal terim x using wiki in ed x Wiki-based Coll x www.wikisym.c x PB ickm2009.pbw x Wikis and colla x Integrating Tec x Uzaktan Eğitim x Bilgisayar Deste x W Revision history x 404 Not Found x

bote.hacettepe.edu.tr/wiki/index.php/Uzaktan_Eğitim

madde tartışma değiştir geçmiş

Uzaktan Eğitim

7 Mayıs 2012

Eğitim yaşayan ve sürekli olarak **gelişim** gösteren canlı bir **süreçtir**. **Eğitim bilimi** bireyleri hiçbir **bilimin** etkilemediği kadar derinden etkilemekte bireyin kişisel özelliklerinin oluşmasında, "birey" olunmasında en etkili olan **bilimdir**. Hayatımızda bu kadar önemli bir yere sahip olan **eğitim** doğal olarak insanı etkileyen her türlü alanla **etkileşim** içerisinde. Günümüze baktığımızda **teknoloji** hayatımızı derinden etkilemekte, bireyler hergün daha fazla bilgiyle karşılaşmakta bu bilgileri de çoğunlukla **teknoloji** yoluyla edinmektedir. Bu süreci izlediğimizde **eğitimin teknoloji** ile ne denli içiçe olabileceğini tahmin edebiliriz. **Eğitimi** temelden etkileyen **teknolojik** yenilikler ve **buluşlar**, her defasında bir önceki sisteme göre üstünlükler sağlamakta, yeni **kavramların** ortaya çıkmasına neden olmaktadır.

Günümüzde birey ve toplum olarak varlığını sürdürebilmek, her alanda başarılı ve öncü olmak ile eşdeğer bir yaklaşım ortaya koymayı gerektirmektedir. Yani artık bireyler öğrendikleri ile yetinmemeli sürekli olarak bilginin izini sürmelidir. Bilgi toplumunda bitmiş **eğitim** diye bir şey yoktur. **Yaşam boyu eğitim** sistemiyle beraber birey bütün yaşantısı boyunca **eğitim** sistemine dahil olur.

Bizler de genç **eğitimsiz** olarak burada sizlere yaşam boyu **eğitim** sistemi içerisinde en önemli yere sahip olan, belki de artık çağımızın **eğitim** sistemi haline dönüşmekte olan uzaktan **eğitimi** tanıtacağız. Uzaktan **eğitimi** kısaca **birbirinden zaman ve mekan bakımından ayrı olan öğrenci, öğretmen ve öğretim materyallerinin iletişim teknolojileri vasıtasıyla bir araya getirildiği eğitim sistemi** olarak tanımlayabiliriz.

Bir başka tanıma göre uzaktan **eğitim**, **öğrenci ile öğretmenin birbirinden uzakta olmalarına karşın eş zamanlı(senkron) ya da ayrı zamanlı(asenkron) olarak bir araçla iletişim kurdukları bir eğitim sistemidir**.

ara

Git Ara

araçlar

- Sayfaya bağlantılar
- İlgili değişiklikler
- Dosya yükle
- Özel sayfalar
- Basılmaya uygun görünüm
- Son haline bağlantı

Uzaktan Eğitim Kavramları
Uzaktan Eğitimle İlgili Sık Kullanılan Kavramlar

Uzaktan Eğitimin Tarihçesi
Geçmişten Günümüze Uzaktan Eğitim

Neden Uzaktan Eğitim
Uzaktan Eğitimin Yararları Günümüzde Gerekli

Uzaktan Eğitimde Teknolojiler
Uzaktan Eğitimde Kullanılan Teknolojiler

Dünyada Uzaktan Eğitim
Çeşitli Ülkelerdeki Uzaktan Eğitim Uygulamaları

Türkiye'de Uzaktan Eğitim
Ülkemizde Uzaktan Eğitim Uygulamaları Uzaktan Eğitim Merkezleri

Uzaktan Eğitimde Roller
Uzaktan Eğitim Katılımcılarının Eğitimdeki Roller

Uzaktan Eğitim Modelleri
Uzaktan Eğitimde Kullanılan Modeller

Uzaktan Eğitim Kuramları
Uzaktan Eğitimi Etkileyen Kuramlar

web tasarım antalya ve SEO

Bu sayfa son olarak 15:25, 28 Nisan 2012 tarihinde güncellenmiştir. Bu sayfaya 13.364 defa erişilmiştir. Gizlilik ikisi Bote Hakkında Feraoatname

Powered By

Sample wiki page

History Logs / Revisions

The screenshot shows the history page for the article "Bilgisayar Destekli Eğitim (BDE)". The page title is "Bilgisayar Destekli Eğitim (BDE)" and the URL is "bote.hacettepe.edu.tr/wiki/index.php?title=Bilgisayar_Destekli_Egitim_%28BDE%29&action=history". The page features a navigation menu with "madde", "tartisma", "degistir", and "gecmis" tabs. Below the title, there is a "Sürüm geçmişi" section with a link to "View logs for this page". The main content is a list of revisions, each with a date, time, and user name. The list is sorted by date and time, with the most recent revision at the top. The list includes revisions from 19:17 on 5 Mart 2008 to 22:27 on 6 Kasım 2007. Each revision is linked to a "Tartışma" (discussion) page and a "Katkılar" (contributions) page. The page also has a sidebar with navigation links and a search box.

Bilgisayar Destekli Eğitim (BDE)
Sürüm geçmişi
View logs for this page

(En yeni | En eski) (önceki 50) (sonraki 50) (20 | 50 | 100 | 250 | 500).

(fark) = güncel sürümle aradaki fark, (son) = önceki sürümle aradaki fark, K= küçük değişiklik

Seçilen sürümleri karşılaştır

- (fark) (son) 19:17, 5 Mart 2008 Ahmetemre (Tartışma | Katkılar)
- (fark) (son) 16:19, 10 Aralık 2007 Kasım kale (Tartışma | Katkılar)
- (fark) (son) 19:09, 1 Aralık 2007 Gemrah korkmaz (Tartışma | Katkılar)
- (fark) (son) 21:33, 28 Kasım 2007 Burcin candan (Tartışma | Katkılar)
- (fark) (son) 21:32, 28 Kasım 2007 Burcin candan (Tartışma | Katkılar)
- (fark) (son) 09:42, 28 Kasım 2007 Kasım kale (Tartışma | Katkılar)
- (fark) (son) 08:54, 28 Kasım 2007 Kasım kale (Tartışma | Katkılar)
- (fark) (son) 18:17, 27 Kasım 2007 Kasım kale (Tartışma | Katkılar)
- (fark) (son) 18:05, 27 Kasım 2007 Kasım kale (Tartışma | Katkılar)
- (fark) (son) 18:02, 27 Kasım 2007 Kasım kale (Tartışma | Katkılar)
- (fark) (son) 17:58, 27 Kasım 2007 Kasım kale (Tartışma | Katkılar)
- (fark) (son) 16:21, 27 Kasım 2007 Kasım kale (Tartışma | Katkılar)
- (fark) (son) 16:20, 27 Kasım 2007 Kasım kale (Tartışma | Katkılar)
- (fark) (son) 11:34, 23 Kasım 2007 Burcin candan (Tartışma | Katkılar)
- (fark) (son) 18:34, 18 Kasım 2007 Burcin candan (Tartışma | Katkılar)
- (fark) (son) 22:52, 15 Kasım 2007 Burcin candan (Tartışma | Katkılar)
- (fark) (son) 22:29, 15 Kasım 2007 Burcin candan (Tartışma | Katkılar)
- (fark) (son) 22:29, 15 Kasım 2007 Burcin candan (Tartışma | Katkılar)
- (fark) (son) 21:11, 15 Kasım 2007 Burcin candan (Tartışma | Katkılar)
- (fark) (son) 21:03, 15 Kasım 2007 Burcin candan (Tartışma | Katkılar)
- (fark) (son) 21:02, 15 Kasım 2007 Burcin candan (Tartışma | Katkılar)
- (fark) (son) 09:10, 7 Kasım 2007 Burcin candan (Tartışma | Katkılar)
- (fark) (son) 08:41, 7 Kasım 2007 Kasım kale (Tartışma | Katkılar)
- (fark) (son) 22:58, 6 Kasım 2007 Burcin candan (Tartışma | Katkılar)
- (fark) (son) 22:42, 6 Kasım 2007 Burcin candan (Tartışma | Katkılar)
- (fark) (son) 22:27, 6 Kasım 2007 Burcin candan (Tartışma | Katkılar)

The screenshot shows the history page for the article "Uzaktan Eğitim". The page title is "Uzaktan Eğitim" and the URL is "bote.hacettepe.edu.tr/wiki/index.php?title=Uzaktan_Egitim&diff=343&oldid=383". The page features a navigation menu with "madde", "tartisma", "degistir", and "gecmis" tabs. Below the title, there is a "Sürüm geçmişi" section with a link to "View logs for this page". The main content is a list of revisions, each with a date, time, and user name. The list is sorted by date and time, with the most recent revision at the top. The list includes revisions from 09:44 on 31 Ekim 2007 to 18:51 on 11 Kasım 2007. Each revision is linked to a "Tartışma" (discussion) page and a "Katkılar" (contributions) page. The page also has a sidebar with navigation links and a search box.

Uzaktan Eğitim
Sürüm geçmişi
View logs for this page

(En yeni | En eski) (önceki 50) (sonraki 50) (20 | 50 | 100 | 250 | 500).

(fark) = güncel sürümle aradaki fark, (son) = önceki sürümle aradaki fark, K= küçük değişiklik

Seçilen sürümleri karşılaştır

Sayfann 09:44, 31 Ekim 2007 tarihindeki hali (değiştir)
10:40 0.74 (Tartışma)
(←Uzaktan Eğitim)
→ Önceki sürümle aradaki fark

Sayfann 18:51, 11 Kasım 2007 tarihindeki hali (değiştir) (sonda)
HuseyinCanSene (Tartışma | Katkılar)
Sonraları sürümle aradaki fark →

(One intermediate revision not shown.)

1. satır:

Uzaktan Eğitim, geleneksel öğrenme-öğretme yöntemlerindeki sınırlar nedeniyle sınıf içi etkinlikleri yürütüle olamıy bulmamış durumlarda eğitim çalışmalarını planlayanlar ve uygulayanlar ile öğrenciler arasında iletişim ve etkileşimi özel olarak hazırlanmış öğrenim üniteleri ve çeşitli ortamlar yoluyla belli bir merkezden sağlanıy bir öğrenim yöntemi.Yine uzaktan eğitim şu şekilde tanımlayabiliriz:Farklı mekanlardaki öğrenci, öğrenim ve öğrenim materyallerini iletişim teknolojileri aracılığıyla bir araya getirdiği kurumsal bir eğitim faaliyetidir.

Uzaktan Eğitim, geleneksel öğrenme-öğretme yöntemlerindeki sınırlar nedeniyle sınıf içi etkinlikleri yürütüle olamıy bulmamış durumlarda eğitim çalışmalarını planlayanlar ve uygulayanlar ile öğrenciler arasında iletişim ve etkileşimi özel olarak hazırlanmış öğrenim üniteleri ve çeşitli ortamlar yoluyla belli bir merkezden sağlanıy bir öğrenim yöntemi.Yine uzaktan eğitim şu şekilde tanımlayabiliriz: Farklı mekanlardaki öğrenci, öğrenim ve öğrenim materyallerini iletişim teknolojileri aracılığıyla bir araya getirdiği kurumsal bir eğitim faaliyetidir.

Bir başka tanıma göre uzaktan eğitim, öğrenci ile öğretmenin birbirinden uzaktaki olmalarına karşın eş zamanlı(senkron) ya da farklı zaman(senkron) olarak bir araya iletişim kurdukları bir eğitim sistemidir

Sayfann 18:51, 11 Kasım 2007 tarihindeki hali

Uzaktan Eğitim, geleneksel öğrenme-öğretme yöntemlerindeki sınırlar nedeniyle sınıf içi etkinlikleri yürütüle olamıy bulmamış durumlarda eğitim çalışmalarını planlayanlar ve uygulayanlar ile öğrenciler arasında iletişim ve etkileşimi özel olarak hazırlanmış öğrenim üniteleri ve çeşitli ortamlar yoluyla belli bir merkezden sağlanıy bir öğrenim yöntemi.Yine uzaktan eğitim şu şekilde tanımlayabiliriz: Farklı mekanlardaki öğrenci, öğrenim ve öğrenim materyallerini iletişim teknolojileri aracılığıyla bir araya getirdiği kurumsal bir eğitim faaliyetidir.

Bir başka tanıma göre uzaktan eğitim, öğrenci ile öğretmenin birbirinden uzaktaki olmalarına karşın eş zamanlı(senkron) ya da farklı zaman(senkron) olarak bir araya iletişim kurdukları bir eğitim sistemidir

Gözetici Bote Hakkında Fergatname

WikLog

WikLog v1.2 - Mediawiki Veritabanı Analiz Programı

Bağlantı Ayarları Genel bilgiler Grup oluşturma Analiz Değerlendirme Sonuçlar: Bireysel Sonuçlar: Grup Dışa aktar

Bağlantı

Bağlantı Ayarları

Sunucu adresi

Kullanıcı adı

Şifre

Veritabanı adı

Durum

Genel Ayarlar

Mediawiki versiyonu 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

00:00:05.0140

Analizi Başlat

Bilgisayarınızın hızına ve veri sayısına göre analizin tamamlanması zaman alabilir.

WikLog

WikLog v1.2 - Mediawiki Veritabanı Analiz Programı

Bağlantı Ayarları Genel bilgiler Grup oluşturma Analiz Değerlendirme Sonuçlar: Bireysel Sonuçlar: Grup Dışa aktar

Kullanıcı İstatistikleri

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

B1: Yeni sayfa sayısı. B2: Düzenleme sayısı. B3: İç bağlantı sayısı. B4: Kelime sayısı.

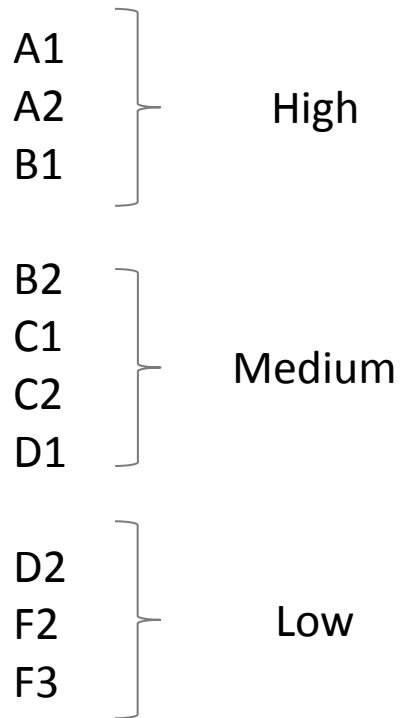
Metrics (Attributes)

- **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



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	?

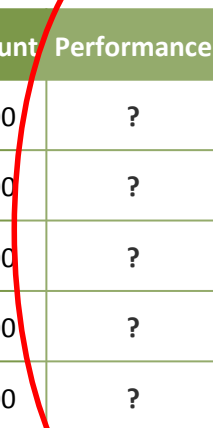
Research Problem

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	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

Prediction

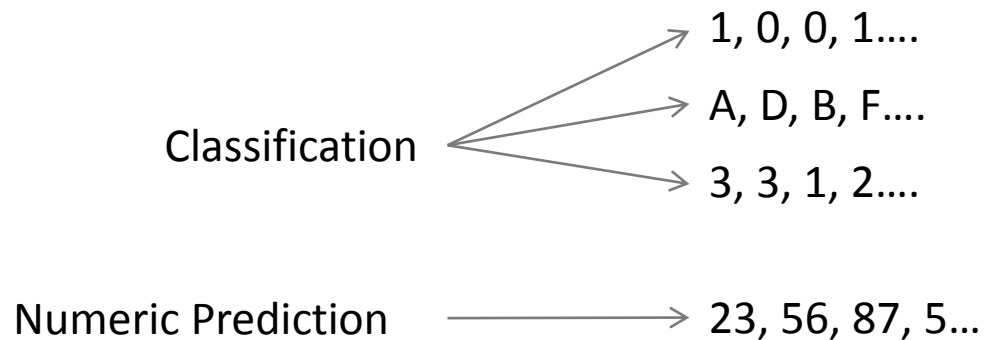


ID	PageCount	EditCount	LinkCount	WordCount	Performance
11	80,00	547,00	193,00	1269,00	?
12	65,00	271,00	273,00	2132,00	?
13	47,00	252,00	231,00	1213,00	?
14	106,00	278,00	399,00	2675,00	?
15	55,00	266,00	49,00	5713,00	?



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.

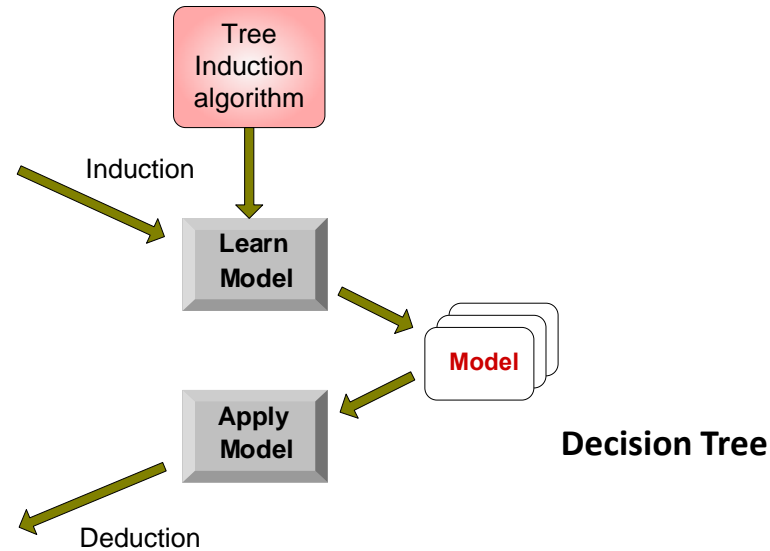
Classification—A Two-Step Process

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



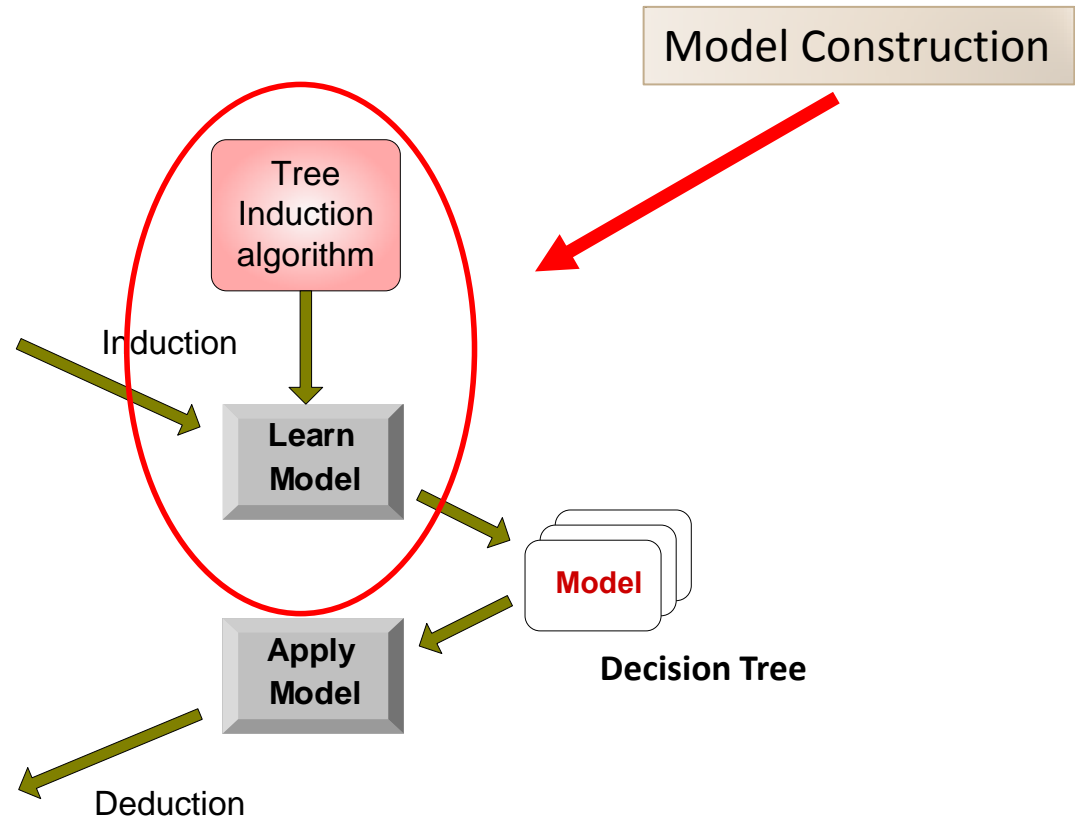
Classification—A Two-Step Process

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



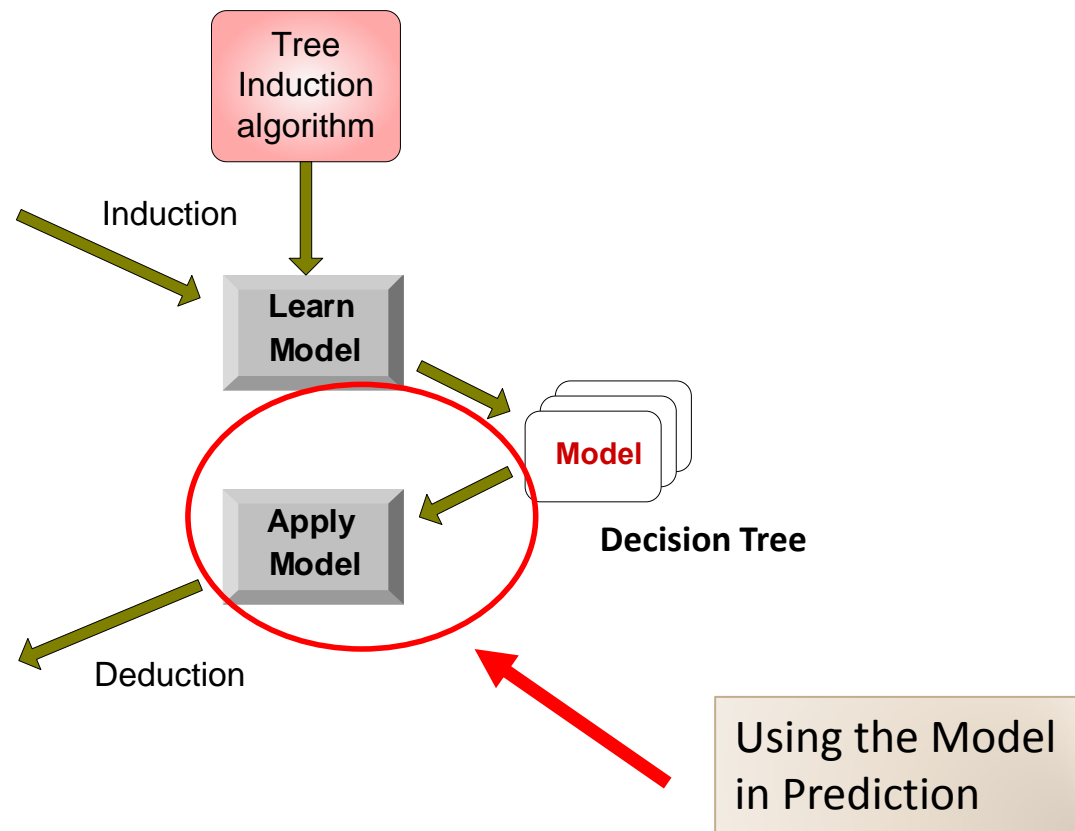
Classification—A Two-Step Process

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

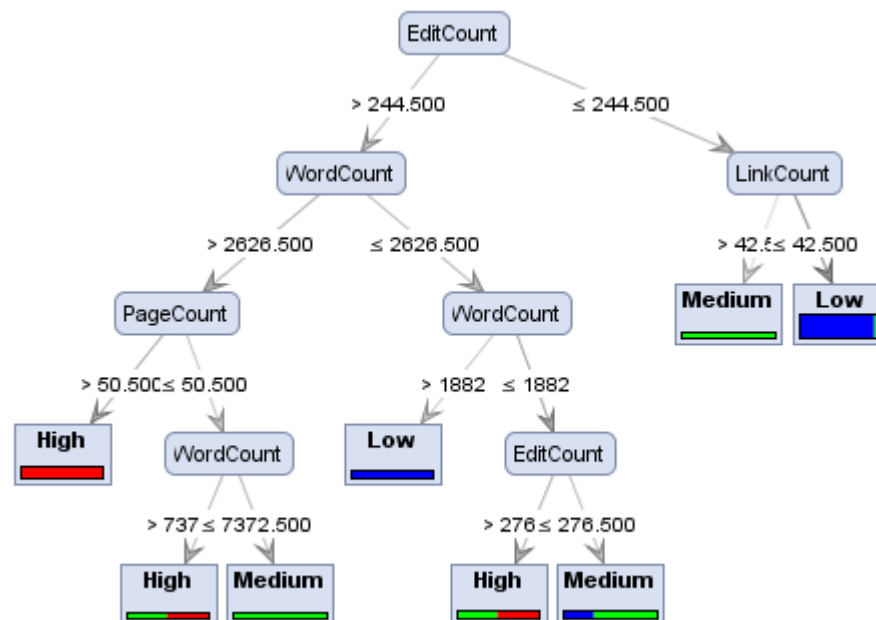


Classification Techniques

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

Classification Techniques

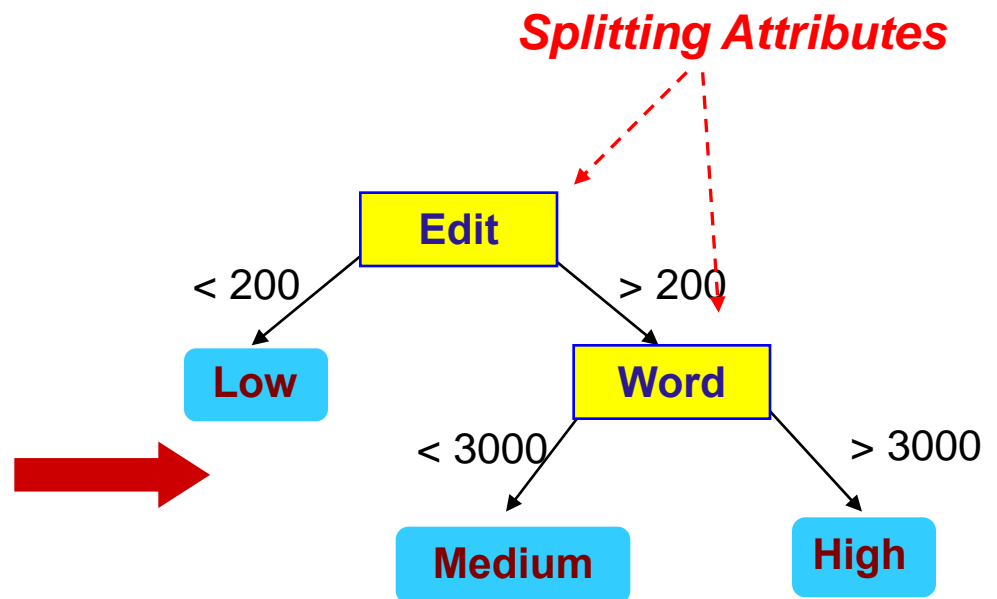
- Decision Tree based Methods



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

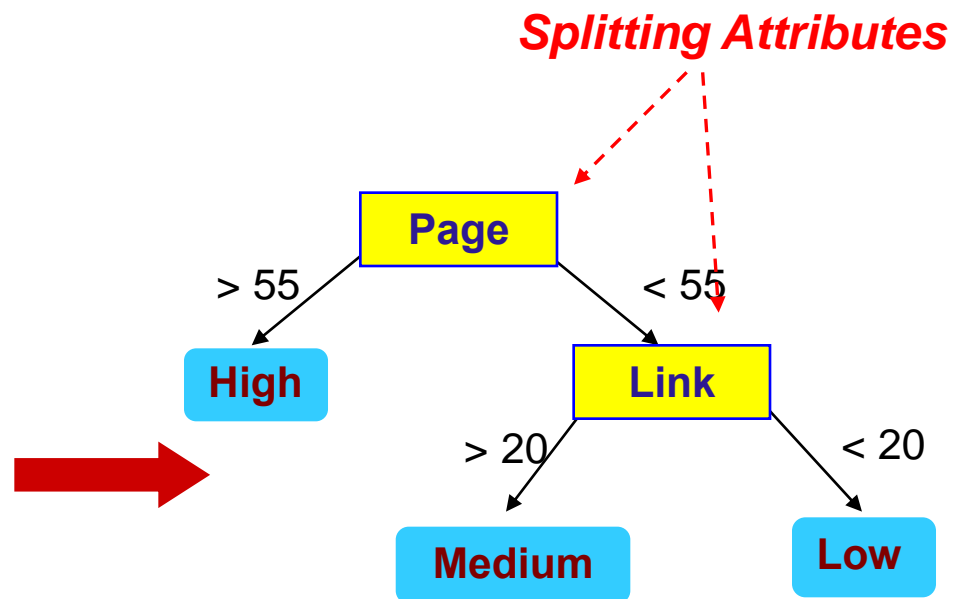


Model: Decision Tree

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

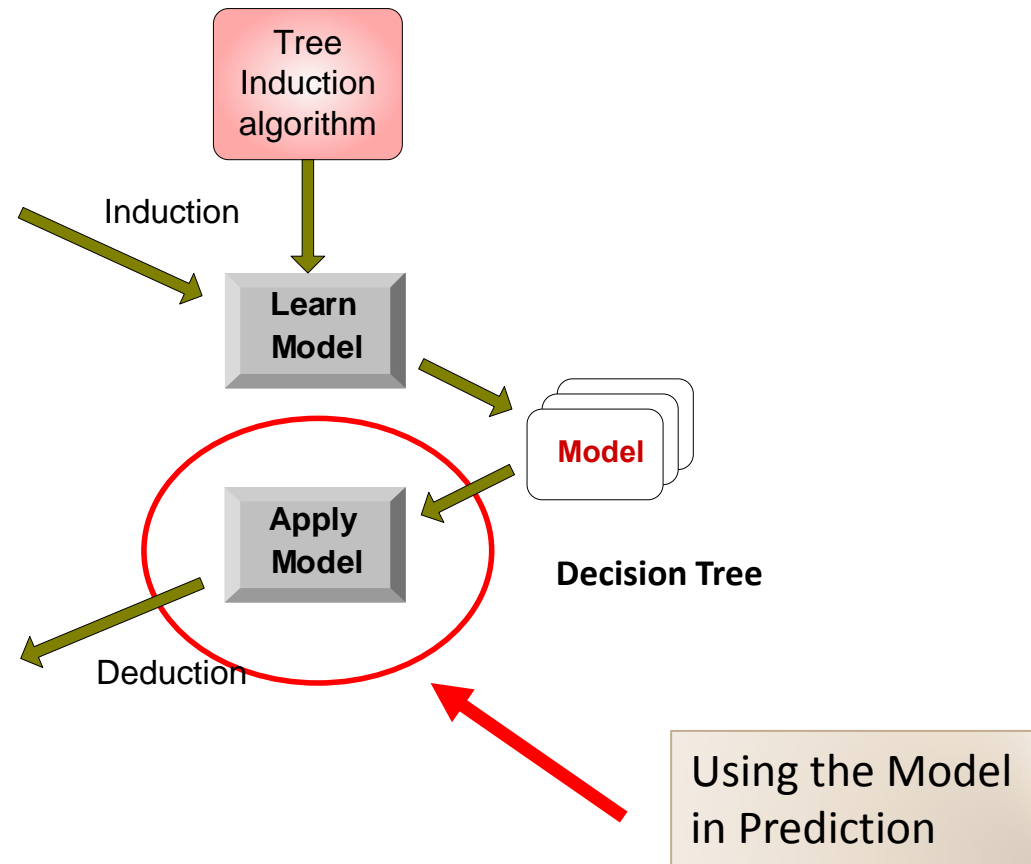
Example of Apply Model to Test Data

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

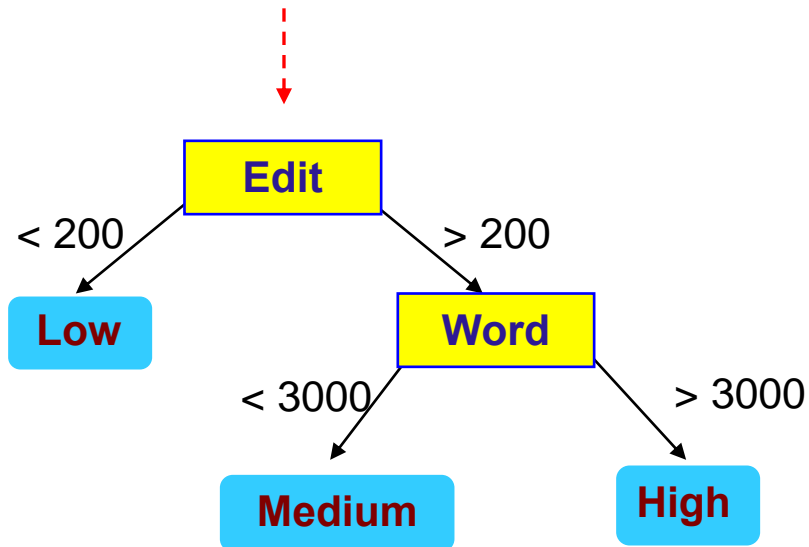
ID	PageCount	EditCount	LinkCount	WordCount	Performance
11	80,00	547,00	193,00	1269,00	?
12	65,00	271,00	273,00	2132,00	?
13	47,00	252,00	231,00	1213,00	?
14	106,00	278,00	399,00	2675,00	?
15	55,00	266,00	49,00	5713,00	?

Test Set



Apply Model to Test Data

Start from the root of tree.



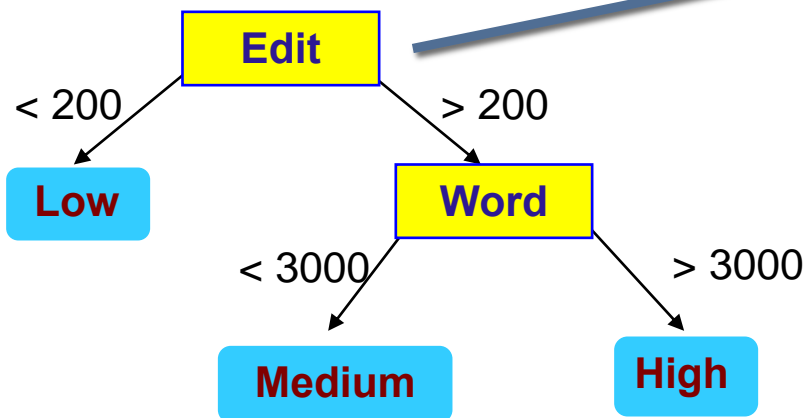
Test Data

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

Apply Model to Test Data

Test Data

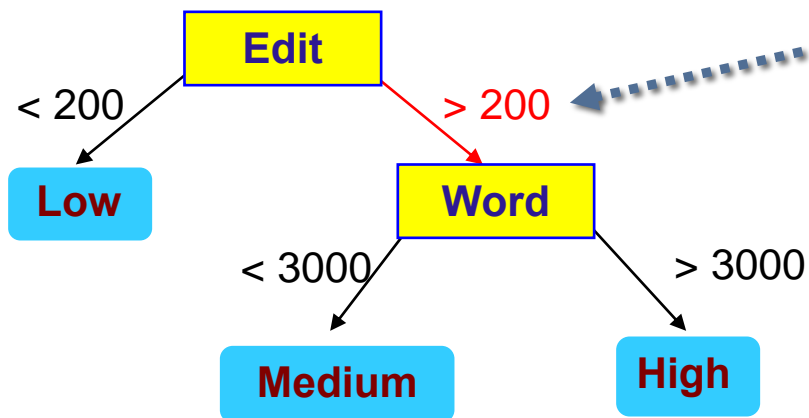
ID	PageCount	EditCount	LinkCount	WordCount	Performance
15	55,00	266,00	49,00	5713,00	?



Apply Model to Test Data

Test Data

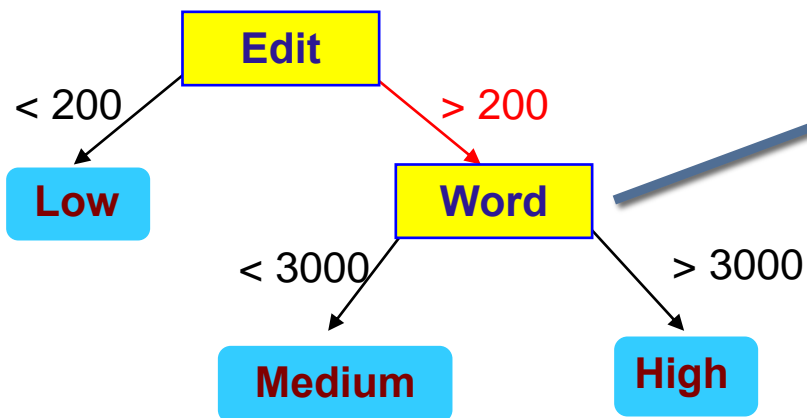
ID	PageCount	EditCount	LinkCount	WordCount	Performance
15	55,00	266,00	49,00	5713,00	?



Apply Model to Test Data

Test Data

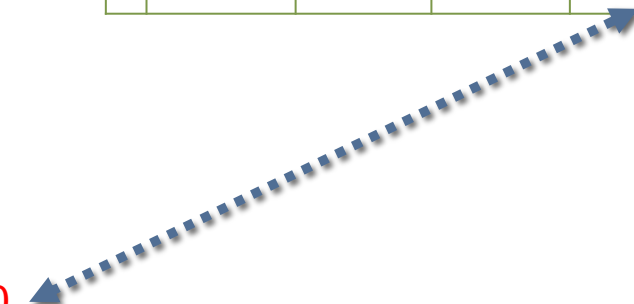
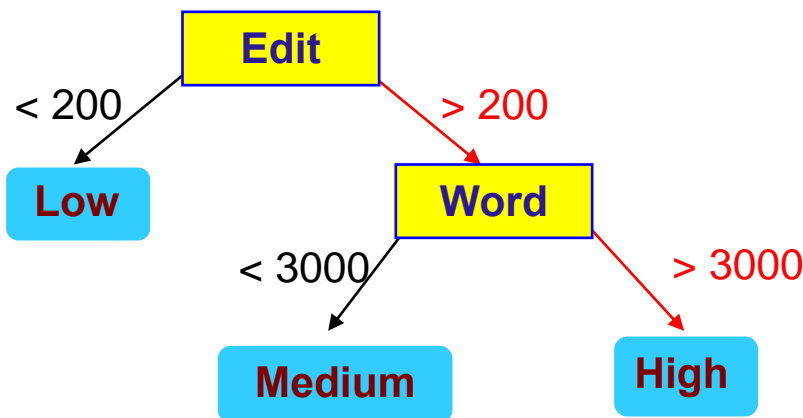
ID	PageCount	EditCount	LinkCount	WordCount	Performance
15	55,00	266,00	49,00	5713,00	?



Apply Model to Test Data

Test Data

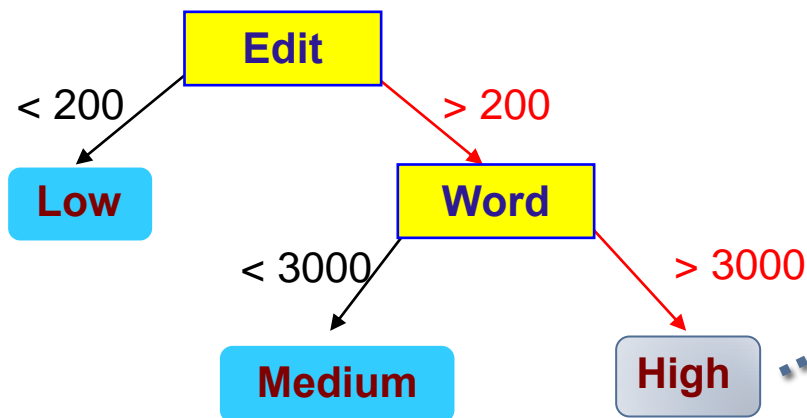
ID	PageCount	EditCount	LinkCount	WordCount	Performance
15	55,00	266,00	49,00	5713,00	?



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:

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

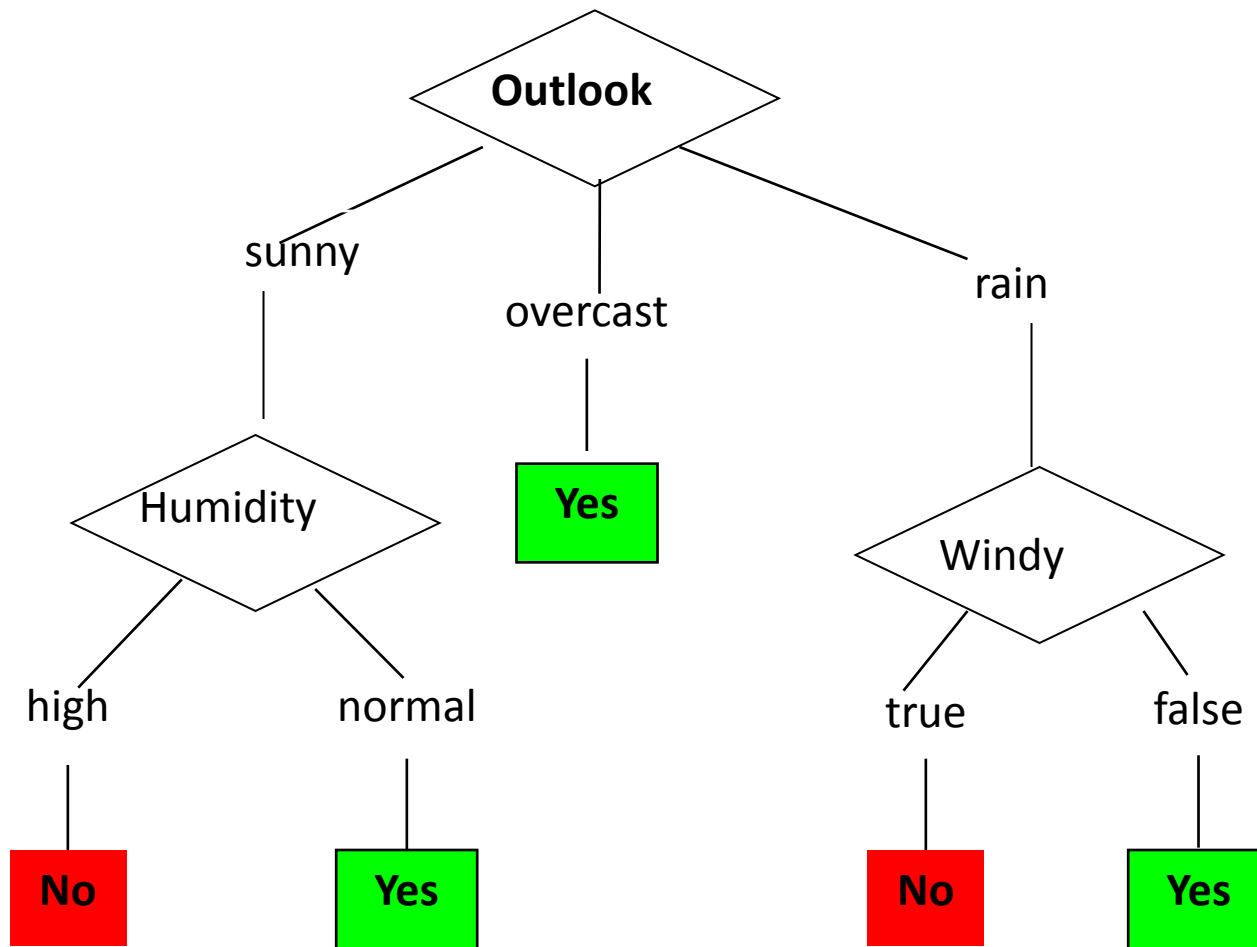
- **Information gained** by branching on attribute A

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

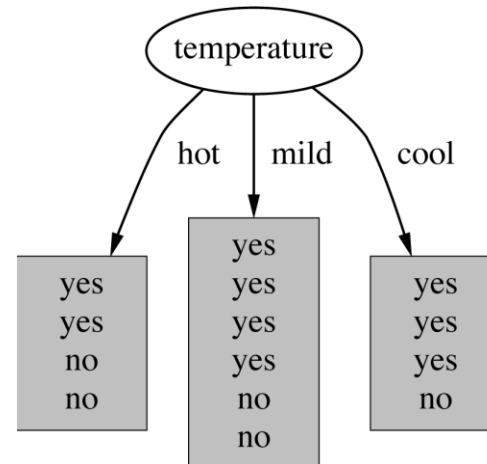
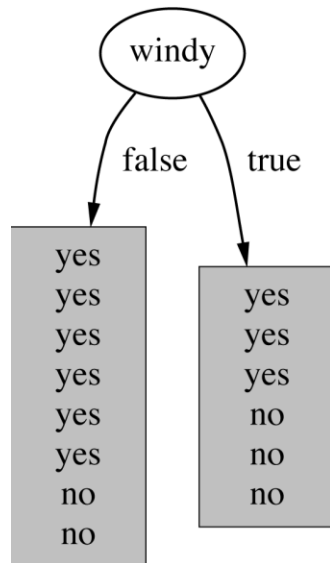
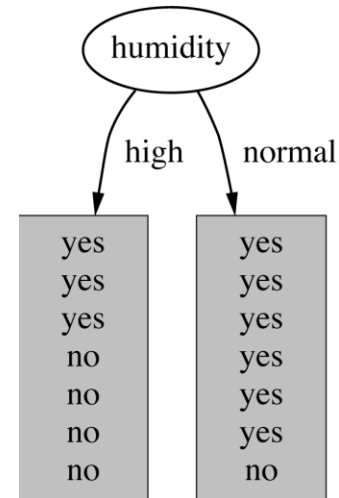
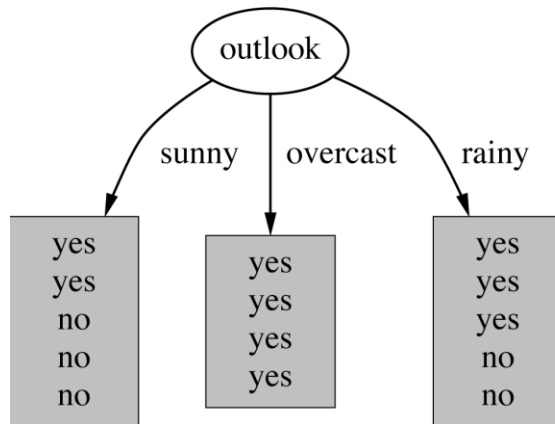
When do I play tennis?

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:

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

Computing the information gain

- Information gain:

(information before split) – (information after split)

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

- 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:

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

- Information Gain:

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

Computing the information gain

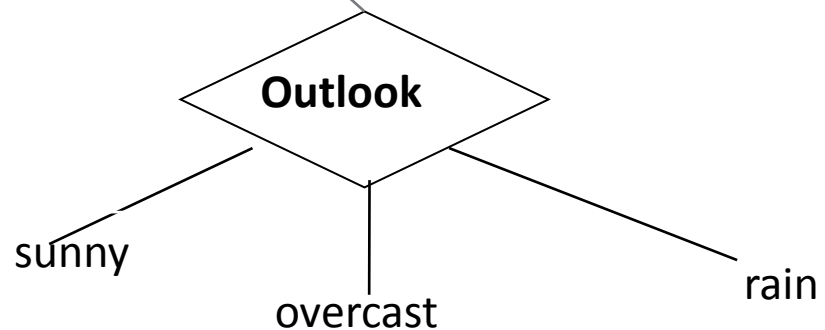
- Information gain for attributes from weather data:

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

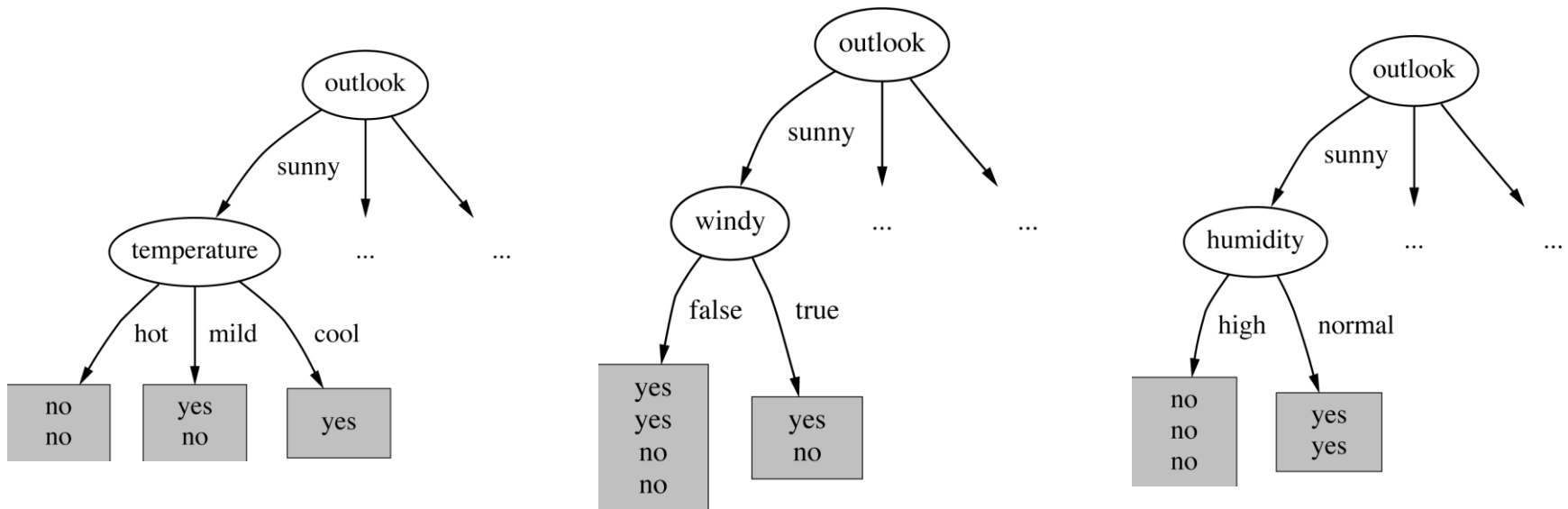
$\text{gain}(\text{" Temperature"}) = 0.029 \text{ bits}$

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

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



Continuing to split

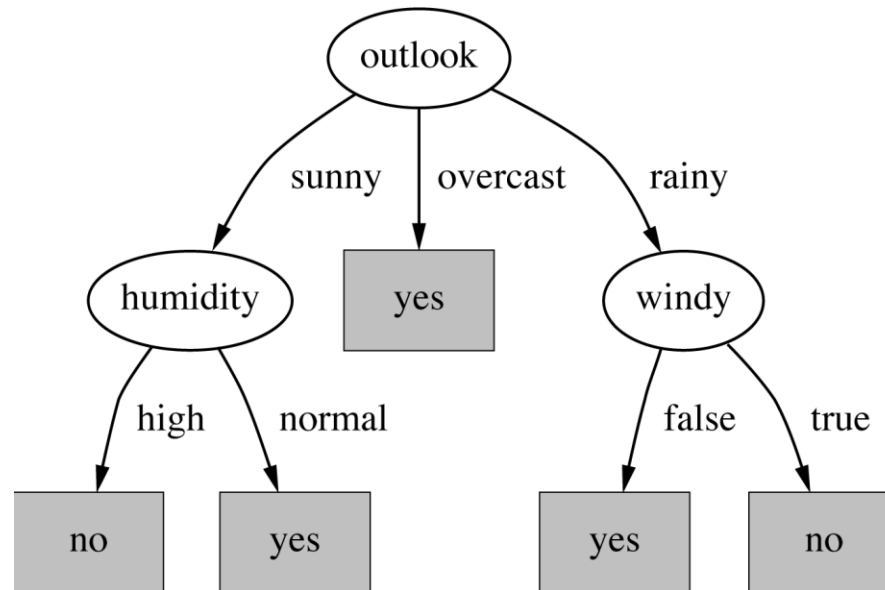


gain("Temperature") = 0.571 bits

gain("Humidity") = 0.971 bits

gain("Windy") = 0.020 bits

The final decision tree



⇒ 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+WordCo`
`unt, 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