

Tugas 05

Dari dataset berikut ini, silahkan buat perhitungan manual dari algoritma decision tree C45 secara lengkap dan gambarkan pohon keputusan yang diperoleh.

No	Outlook	Temperature	Humidity	Windy	Play
1	Sunny	Hot	High	FALSE	No
2	Sunny	Hot	High	TRUE	No
3	Cloudy	Hot	High	FALSE	Yes
4	Rainy	Mild	High	FALSE	Yes
5	Rainy	Cool	Normal	FALSE	Yes
6	Rainy	Cool	Normal	TRUE	Yes
7	Cloudy	Cool	Normal	TRUE	Yes
8	Sunny	Mild	High	FALSE	No
9	Sunny	Cool	Normal	FALSE	Yes
10	Rainy	Mild	Normal	FALSE	Yes
11	Sunny	Mild	Normal	TRUE	Yes
12	Cloudy	Mild	High	TRUE	Yes
13	Cloudy	Hot	Normal	FALSE	Yes
14	Rainy	Mild	High	TRUE	No

Kumpulkan dalam format MS Word. Sertakan penjelasan dan pendapat anda tentang algoritma ini, kelebihan dan kekurangannya berbanding metode lain yang sejenis (tentukan sendiri metode pembandingnya).

Tugas 05

Nama : Juminovario
NIM : 202420018
Kelas : MTI 23 Reguler A
MK : Advanced Database

Dari dataset berikutini, silahkan buat perhitungan manual dari algoritma decision tree C45 secara lengkap dan gambarkan pohon keputusan yang diperoleh.

No	Outlook	Temperature	Humidity	Windy	Play
1	Sunny	Hot	High	FALSE	No
2	Sunny	Hot	High	TRUE	No
3	Cloudy	Hot	High	FALSE	Yes
4	Rainy	Mild	High	FALSE	Yes
5	Rainy	Cool	Normal	FALSE	Yes
6	Rainy	Cool	Normal	TRUE	Yes
7	Cloudy	Cool	Normal	TRUE	Yes
8	Sunny	Mild	High	FALSE	No
9	Sunny	Cool	Normal	FALSE	Yes
10	Rainy	Mild	Normal	FALSE	Yes
11	Sunny	Mild	Normal	TRUE	Yes
12	Cloudy	Mild	High	TRUE	Yes
13	Cloudy	Hot	Normal	FALSE	Yes
14	Rainy	Mild	High	TRUE	No

Kumpulkan dalam format MS Word. Sertakan penjelasan dan pendapatan anda tentang algoritma ini, kelebihan dan kekurangannya berbanding metode lain yang sejenis (tentukan sendiri metode pembandingnya).

TABEL PERHITUNGAN NODE 1

Rumus :

Entropy = $((-\text{don't play} / \text{jumlah total kasus}) * \log((\text{don't play} / \text{jumlah total kasus}); 2)) + ((-\text{play} / \text{jumlah total kasus}) * \log((\text{play} / \text{jumlah total kasus}); 2))$

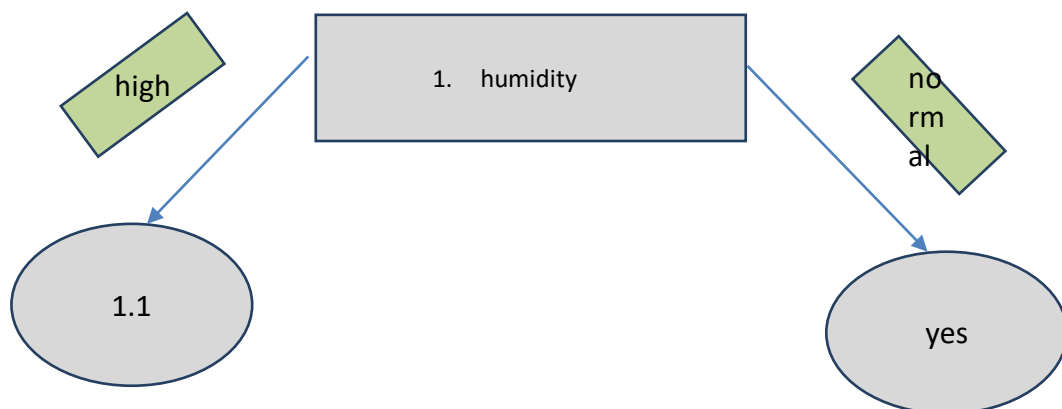
$$1. ((-4/14) * \log(4/14); 2) + ((-10/14) * \log(10/14); 2) = 0,863121$$

Gain = jumlah total entropy – ((jumlahkasus variable x/jumlah total kasus)*entropy variable x)+((jumlahkasus variable x/jumlah total kasus)*entropy variable x)

1. Outlook = $0,863121 - ((4/14) * 0) + ((5/14) * 0,721928) + ((5/14) * 0,970951) = 0,25852104$
2. Temp = $0,863121 - ((4/14) * 0) + ((4/14) * 1) + ((6/14) * 0,918296) = 0,18385093$
3. Humidity = $0,863121 - ((7/14) * 0,985228) + ((7/14) * 0) = 0,3705065$
4. Windy = $0,863121 - ((8/14) * 0,811278) + ((6/14) * 0,918296) = 0,00597771$

		Jumlahkasus	Don't play (s1)	Play (s2)	Entropy	Gain
TOTAL		14	4	10	0,863121	
OUTLOOK						0,25852104
	Cloudy	4	0	4	0	
	Raini	5	1	4	0,721928	
	Sunny	5	3	2	0,970951	
TEMP						0,18385093
	Cool	4	0	4	0	
	Hot	4	2	2	1	
	Mild	6	2	4	0,918296	
HUMIDITY						0,3705065
	Hight	7	4	3	0,985228	
	Normal	7	0	7	0	
WINDY						0,00597771
	No	8	2	6	0,811278	
	yes	6	2	4	0,918296	

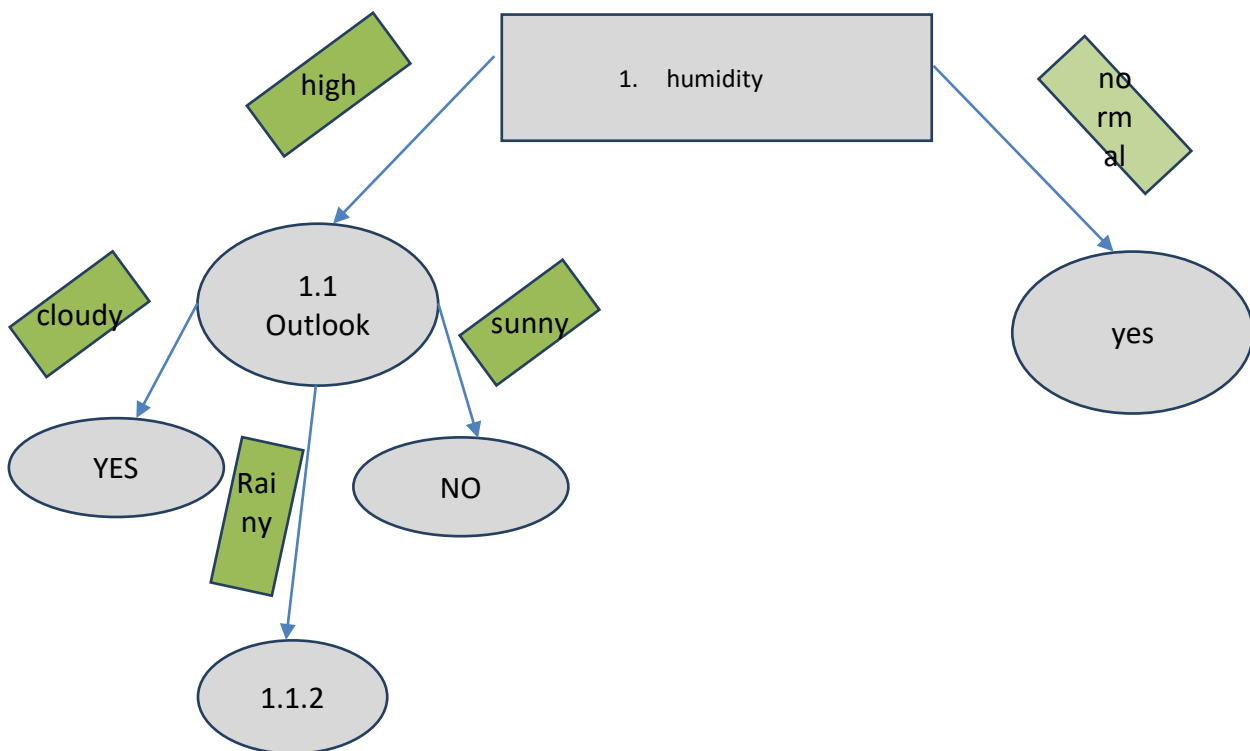
DECISION TREE



TABEL PERHITUNGAN 1.1

		Jumlahkasus	Don't play (s1)	Play (s2)	Entropy	Gain
HUMIDITY		7	4	3	0,985228136	
OUTLOOK						0,69951385
	Cloudy	2	0	2	0	
	Rainy	2	1	1	1	
	Sunny	3	3	0	0	
TEMP						0,020244207
	Cool	0	0	0	0	
	Hot	3	2	1	0,918295834	
	Mild	4	2	2	1	
WINDY						0,020244207
	No	4	2	2	1	
	yes	3	2	1	0,918295834	

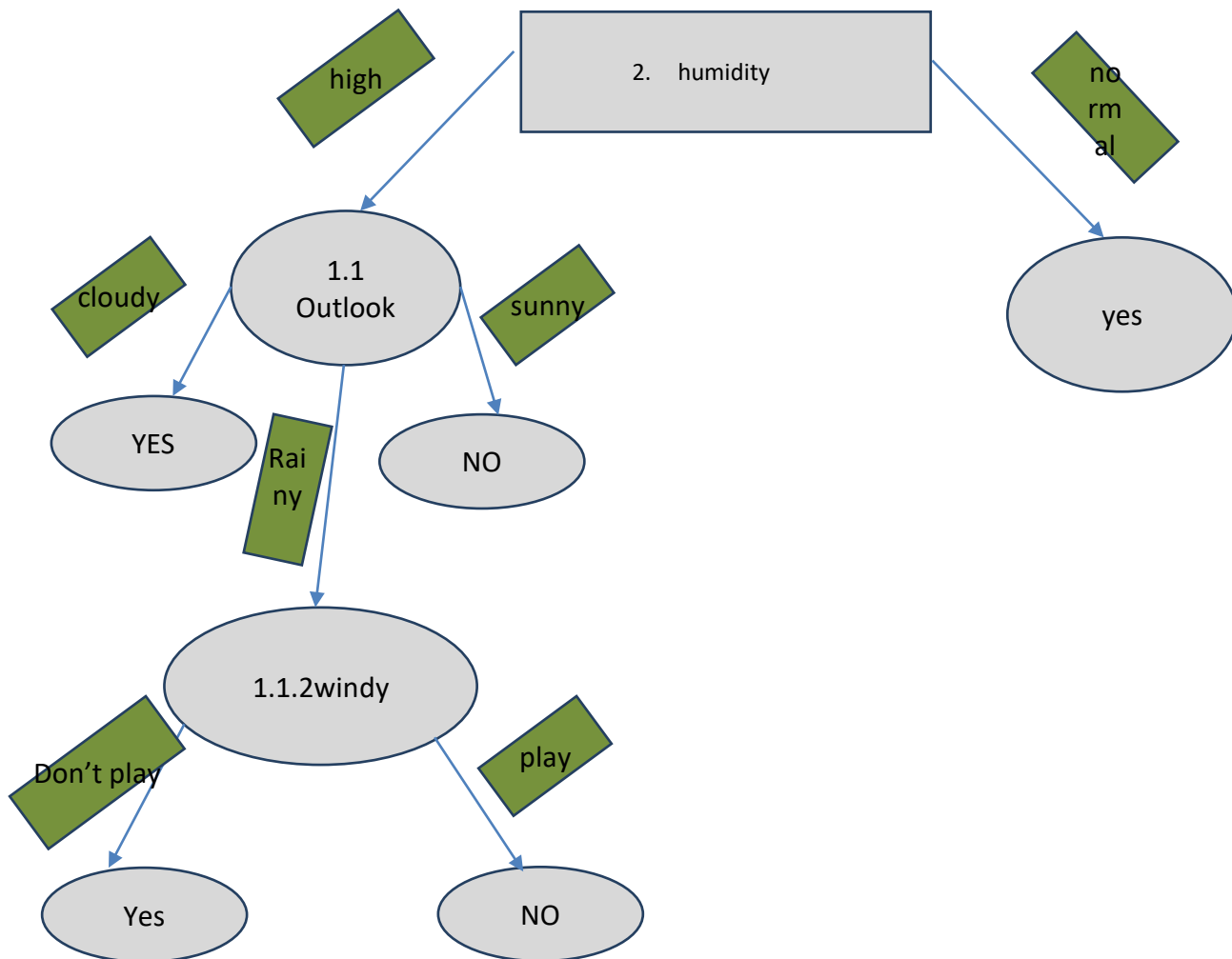
DECISION TREE



TABEL PERHITUNGAN 1.1.2

NODE 1.1.2		Jumlahkasus	Don't play (s1)	Play (s2)	Entropy	Gain
HUMIDITY HIGHT DAN OUTLOOK RAINY		2	1	1	1	
TEMP						0
	Cool	0	0	0	0	
	Hot	0	0	0	0	
	Mild	2	1	1	1	
WINDY						1
	No	1	0	1	0	
	yes	1	1	0	0	

DECISION TREE



Tugas 05

Dari dataset berikut ini, silahkan buat perhitungan manual dari algoritma decision tree C45 secara lengkap dan gambarkan pohon keputusan yang diperoleh.

No	Outlook	Temperature	Humidity	Windy	Play
1	Sunny	Hot	High	FALSE	No
2	Sunny	Hot	High	TRUE	No
3	Cloudy	Hot	High	FALSE	Yes
4	Rainy	Mild	High	FALSE	Yes
5	Rainy	Cool	Normal	FALSE	Yes
6	Rainy	Cool	Normal	TRUE	Yes
7	Cloudy	Cool	Normal	TRUE	Yes
8	Sunny	Mild	High	FALSE	No
9	Sunny	Cool	Normal	FALSE	Yes
10	Rainy	Mild	Normal	FALSE	Yes
11	Sunny	Mild	Normal	TRUE	Yes
12	Cloudy	Mild	High	TRUE	Yes
13	Cloudy	Hot	Normal	FALSE	Yes
14	Rainy	Mild	High	TRUE	No

Kumpulkan dalam format MS Word. Sertakan penjelasan dan pendapat anda tentang algoritma ini, kelebihan dan kekurangannya berbanding metode lain yang sejenis (tentukan sendiri metode pembandingnya).

Jawaban:

1. Perhitungan Gain

$$\text{Gain}(S,A) = \text{Entropy}(S) - \sum_{i=1}^n \frac{|S_i|}{|S|} * \text{Entropy}(S_i)$$

Keterangan:

- S : himpunan
- A : atribut
- n : jumlah partisi atribut A
- | S_i | : jumlah kasus pada partisi ke-i
- | S | : jumlah kasus dalam S

2. Menghitung Nilai Entropy

$$\text{Entropy}(S) = - \sum_{i=1}^n p_i * \log_2 p_i$$

Keterangan:

S : himpunan kasus

A : fitur

n : jumlah partisi S

pi : proporsi dari Si terhadap S

Perincian algoritma (langkah 1)

- Menghitung jumlah kasus seluruhnya, jumlah berkeputusan “Yes” maupun “No”.
- Menghitung Entropy dari semua kasus yg terbagi berdasarkan atribut “Outlook”, “Temperature”, “Humidity”, “Windy”.
- Lakukan penghitungan Gain utk setiap atributnya

Perhitungan

3). Perincian algoritma, perhitungan :

Node			total	yes	no	entropy	gain	
1	Outlook							
		Sunny	5	3	2	0.970951		
		Cloudy	4	0	4	0		
		Rainy	5	1	4	0.721928	0.71429	
	Temperature							
		Hot	4	2	2	1	0.57143	0.27094242
		Mild	6	2	4	0.918296	0.78711	
		Cool	4	0	4	0		-0.3733112
	Humidity							
		High	7	3	4	0.985228		
		Normal	7	0	7	0		
	Wind						0.02024	
		TRUE	3	1	2	0.918296		
		FALSE	4	2	2	1		
							0.28571	
							0.69951	

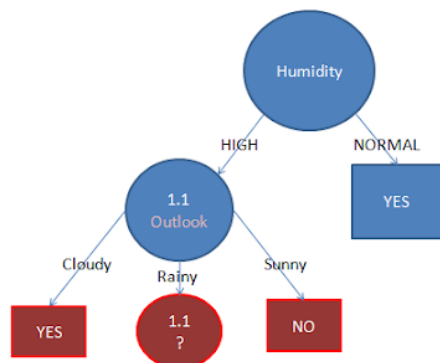
4). Perhitungan Total Entropy

Node			total	yes	no	entropy	gain
2	Outlook						
		Cloudy	2	0	2	0	
		Rainy	2	1	1	1	0.69951
		Sunny	3	3	0	0	
	Temperature						
		Hot	0	0	0	0	1
		Mild	2	1	1	1	0.30049

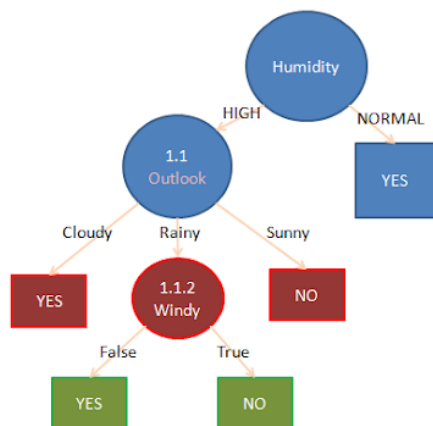
		Cool	0	0	0	0	
	Wind						
		TRUE	1	0	1	0	0.699
		FALSE	1	1	0	0	

5). Pohon keputusan node 2

Node			Jumlah	Yes	No	Entropy	Gain
2	Humidity "high" dan "Outlook" rainy		2	1	1	1	
	Temperature						0
		Hot	4	2	2	1	0.18385
		Mild	6	4	2	0.918296	
		Cool	4	4	0	0	
	Wind						
		TRUE	6	4	2	0.918296	0.00598
		FALSE	8	6	2	0.811278	



6). Pohon keputusan node



Kelebihan Algoritma C4.5 :

1. C4.5 mampu menangani atribut yang kosong (Missing Value).
2. C4.5 mampu menangani atribut dengan kontinu.
3. C4.5 memangkas pohon keputusan untuk mengatasi overfitting

Kelebihan dari metode pohon keputusan adalah sebagai berikut Faruz El Said, 2009 :

1. Daerah pengambilan keputusan yang sebelumnya kompleks dan sangat global, dapat diubah menjadi lebih simpel dan spesifik.
2. Eliminasi perhitungan-perhitungan yang tidak diperlukan, karena ketika menggunakan metode pohon keputusan maka sample diuji hanya berdasarkan kriteria atau kelas tertentu.
3. Fleksibel untuk memilih fitur dari internal node yang berbeda, fitur yang terpilih akan membedakan suatu kriteria dibandingkan kriteria yang lain dalam node yang sama.

Secara fleksibel metode pohon keputusan ini meningkatkan kualitas keputusan yang dihasilkan jika dibandingkan ketika menggunakan metode penghitungan satu tahap yang lebih konvensional.

Kekurangan Algoritma C4.5 :

1. Algoritma C4.5 hanya dapat digunakan untuk menangani sampel-sampel yang dapat disimpan secara keseluruhan dan pada waktu yang bersamaan di memori.

Kekurangan dari metode pohon keputusan adalah sebagai berikut :

1. Terjadi overlap terutama ketika kelas-kelas dan kriteria yang digunakan jumlahnya sangat banyak. Hal tersebut juga dapat menyebabkan meningkatnya waktu pengambilan keputusan dan jumlah memori yang diperlukan.
2. Terjadi akumulasi jumlah error dari setiap tingkat dalam sebuah pohon keputusan yang besar.
3. Kesulitan dalam mendesain pohon keputusan yang optimal.
4. Hasil kualitas keputusan yang didapatkan dari metode pohon keputusan sangat tergantung pada bagaimana pohon tersebut didesain

Perbandingan dengan metode K-Naive Bayes

Kekurangan Naïve Bayes :

1. Tidak berlaku jika probabilitas kondisionalnya adalah nol, apabila nol maka probabilitas prediksi akan bernilai nol juga
2. Mengasumsikan variabel bebas

Tugas 05

Dari dataset berikut ini, silahkan buat perhitungan manual dari algoritma decision tree C45 secara lengkap dan gambarkan pohon keputusan yang diperoleh.

No	Outlook	Temperature	Humidity	Windy	Play
1	Sunny	Hot	High	FALSE	No
2	Sunny	Hot	High	TRUE	No
3	Cloudy	Hot	High	FALSE	Yes
4	Rainy	Mild	High	FALSE	Yes
5	Rainy	Cool	Normal	FALSE	Yes
6	Rainy	Cool	Normal	TRUE	Yes
7	Cloudy	Cool	Normal	TRUE	Yes
8	Sunny	Mild	High	FALSE	No
9	Sunny	Cool	Normal	FALSE	Yes
10	Rainy	Mild	Normal	FALSE	Yes
11	Sunny	Mild	Normal	TRUE	Yes
12	Cloudy	Mild	High	TRUE	Yes
13	Cloudy	Hot	Normal	FALSE	Yes
14	Rainy	Mild	High	TRUE	No

Kumpulkan dalam format MS Word. Sertakan penjelasan dan pendapat anda tentang algoritma ini, kelebihan dan kekurangannya berbanding metode lain yang sejenis (tentukan sendiri metode pembandingnya).

Pembahasan :

No	Outlook	Temperature	Humidity	Windy	Play
1	Sunny	Hot	High	FALSE	No
2	Sunny	Hot	High	TRUE	No
3	Cloudy	Hot	High	FALSE	Yes
4	Rainy	Mild	High	FALSE	Yes
5	Rainy	Cool	Normal	FALSE	Yes
6	Rainy	Cool	Normal	TRUE	Yes
7	Cloudy	Cool	Normal	TRUE	Yes
8	Sunny	Mild	High	FALSE	No
9	Sunny	Cool	Normal	FALSE	Yes
10	Rainy	Mild	Normal	FALSE	Yes
11	Sunny	Mild	Normal	TRUE	Yes
12	Cloudy	Mild	High	TRUE	Yes
13	Cloudy	Hot	Normal	FALSE	Yes
14	Rainy	Mild	High	TRUE	No

Berdasarkan tabel diatas akan dibuat tabel keputusan dengan melihat keadaan Outlook (cuaca), Temperatur, Humidity (kelembaban), dan windy (keadaan angin).

Algoritma secara umum:

- Pilih atribut sebagai akar
 - Buat cabang untuk tiap2 nilai
 - Bagi kasus dalam cabang
 - Ulangi proses utk setiap cabang sampai semua kasus pada cabang memiliki kelas yang sama
- Memilih atribut berdasarkan nilai “gain” tertinggi dari atribut-atribut yang ada.

Perhitungan Gain

$$\text{Gain}(S,A) = \text{Entropy}(S) - \sum_{i=1}^n \frac{|S_i|}{|S|} * \text{Entropy}(S_i)$$

Keterangan:

- S : himpunan
- A : atribut
- n : jumlah partisi atribut A
- | S_i | : jumlah kasus pada partisi ke-i
- | S | : jumlah kasus dalam S

Menghitung Nilai Entropy

$$\text{Entropy (S)} = \sum_{i=1}^n - p_i * \log_2 p_i$$

Keterangan:

S : himpunan kasus

A : fitur

n : jumlah partisi S

pi : proporsi dari Si terhadap S

Perincian algoritma (langkah 1)

- Menghitung jumlah kasus seluruhnya, jumlah berkeputusan “Yes” maupun “No”.
- Menghitung Entropy dari semua kasus yg terbagi berdasarkan atribut “Outlook”, “Temperature”, “Humidity”, “Windy”.
- Lakukan penghitungan Gain utk setiap atributnya

Perhitungan

Node			Jumlah Kasus (S)	Tidak (S ₁)	Ya (S ₂)	Entropy	Gain
1	TOTAL		14	4	10	0,863120569	
	OUTLOOK						0,258521037
		CLOUDY	4	0	4	0	
		RAINY	5	1	4	0,721928095	
		SUNNY	5	3	2	0,970950594	
	TEMPERATURE						0,183850925
		COOL	4	0	4	0	
		HOT	4	2	2	1	
		MILD	6	2	4	0,918295834	
	HUMIDITY						0,370506501
		HIGH	7	4	3	0,985228136	
		NORMAL	7	0	7	0	
	WINDY						0,005977711
		FALSE	8	2	6	0,811278124	
		TRUE	6	2	4	0,918295834	

Perhitungan Total Entropy

Node			Jumlah Kasus (S)	Tidak (S ₁)	Ya (S ₂)	Entropy	Gain
1	TOTAL		14	4	10	0,863120569	
...	...	...	...	...	...	...	...

$$\text{Entropy (S)} = \sum_{i=1}^n - p_i * \log_2 p_i$$

$$\text{Entropy (Total)} = - 4/14 * \log_2 (4/14) + - 10/14 * \log_2 (10/14)$$

Menghitung gain pada baris Outlook

Gain (Total, Outlook)

$$= \text{Entropy(Total)} - \sum_{i=1}^n \frac{|\text{Outlook}_i|}{|\text{Total}|} * \text{Entropy(Outlook}_i)$$

$$= 0,863120569 - ((4/14 * 0) + (5/14 * 0,72) + (5/14 * 0,97))$$

$$= 0,258521037$$

Lakukan Hitung Gain untuk temperature, humidity dan windy

Seperti yg terlihat pada tabel, diperoleh bahwa atribut dengan Gain tertinggi adalah Humidity -> 0,37

Maka Humidity menjadi node akar

Humidity memiliki dua nilai yaitu "High" dan "Normal"

Humidity -> "Normal" sdh mengklasifikasikan kasus menjadi 1 yaitu keputusannya "yes"

Untuk humidity -> "High" msh perlu dilakukan perhitungn lagi (karena masih terdapat "yes" dan "no")

Pohon Keputusan Node 1

Perincian Algoritma (Langkah 2)

- Pilih node akar “High” dan hitung:
 - Jumlah kasus
 - Jumlah kasus keputusan “Yes”
 - Jumlah kasus keputusan “No”
 - Entropy



Atribut:

- Outlook
- Temperature
- Windy

Hasil perhitungan (Langkah 2)

Node			Jumlah Kasus (S)	Tidak (S _i)	Ya (S _j)	Entropy	Gain
1.1	HUMIDITY HIGH		7	4	3	0,985228136	
	OUTLOOK						0,69951385
		CLOUDY	2	0	2	0	
		RAINY	2	1	1	1	
		SUNNY	3	3	0	0	
	TEMPERATURE						0,020244207
		COOL	0	0	0	0	
		HOT	3	2	1	1	
		MILD	4	2	2	2	
	WINDY						0,020244207
		FALSE	4	2	2	1	
		TRUE	3	2	1	0,918295834	

Hasil perhitungan (Langkah 2)

Didapat Gain tertinggi -> outlook -> 0,69

Maka “Outlook” menjadi node cabang dari atribut humidity yg bernilai “High”

Berdasarkan atribut “Outlook” terdpt 3 nilai

Cloudy

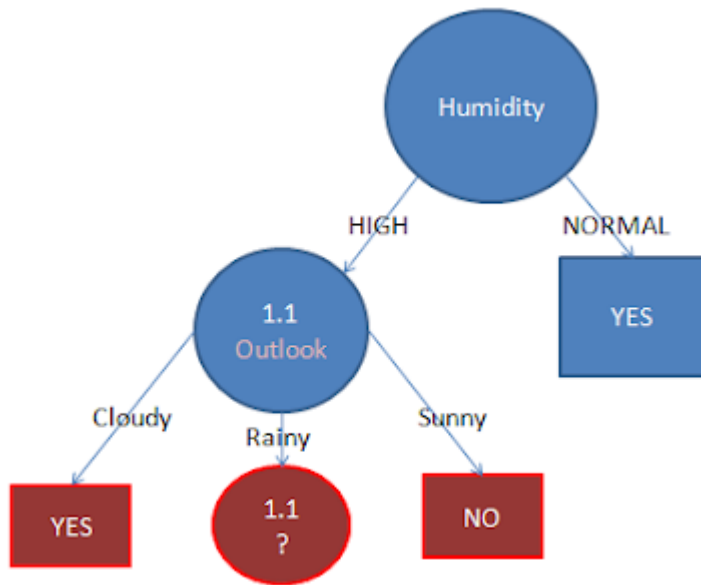
Rainy

Sunny

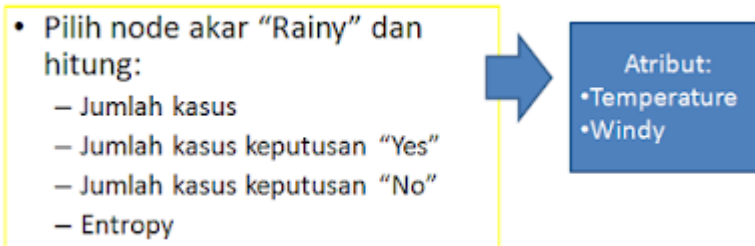
Krn “Cloudy” pasti bernilai “Yes” dan “Sunny” pasti bernilai “No”, maka tdk perlu dilakukan perhitungan lagi

Sedangkan “Rainy” bernilai “yes” dan “No”, maka masih perlu dilakukan perhitungan lagi

Pohon keputusan node 1.1



Perincian algoritma (Langkah 3)



Hasil perhitungan (Langkah 3)

Node			Jumlah Kasus (S)	Tidak (S ₁)	Ya (S ₂)	Entropy	Gain
1.1	HUMIDITY "HIGH" dan OUTLOOK "RAINY"		2	1	1	1	
	TEMPERATURE						0
		COOL	0	0	0	0	
		HOT	0	0	0	0	
		MILD	2	1	1	1	
	WINDY						1
		FALSE	1	0	1	0	
		TRUE	1	1	0	0	

Hasil perhitungan (Langkah 3)

Didapat Gain tertinggi -> Windy -> 1

Maka “Windy” menjadi node cabang dari atribut humidity yg bernilai “High” dan outlook yg bernilai “Rainy”

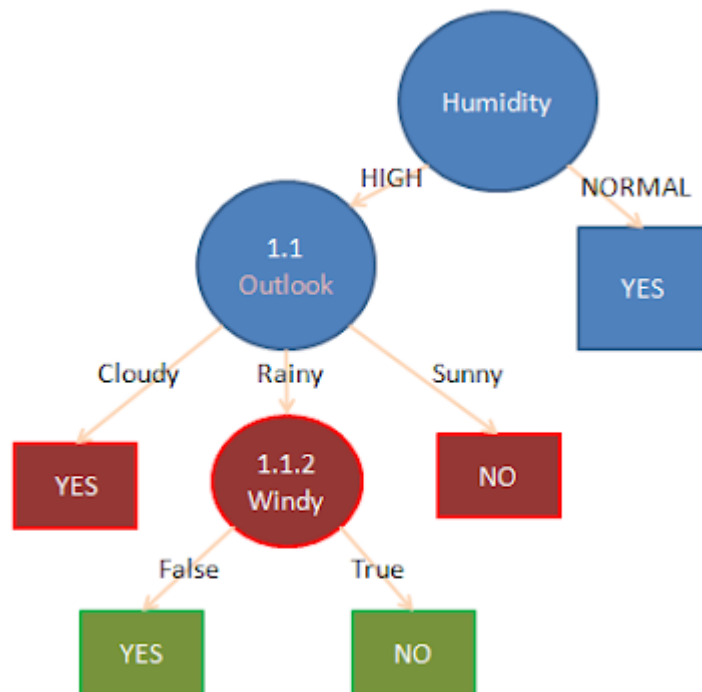
Berdasarkan atribut “Windy” terdpt 2 nilai

True

False

Karena “True” sdh terklasifikasi pasti bernilai “No” dan “False” pasti bernilai “Yes”, maka tidak perlu dilakukan perhitungan lagi

Pohon keputusan node 1.1.2



Hasil perhitungan (Langkah 3)

Berdasarkan node 1.1.2, maka:

“Semua kasus sudah masuk dapat kelas”

Sehingga pohon keputusan diatas merupakan pohon keputusan terakhir yang terbentuk

Pendapat saya tentang Algoritma C45

Adalah algoritma klasifikasi data dengan teknik pohon keputusan yang terkenal dan disukai karena memiliki kelebihan-kelebihan. Kelebihan ini misalnya dapat mengolah data numerik (kontinyu) dan diskret, dapat menangani nilai atribut yang hilang, menghasilkan aturan-aturan yang mudah diinterpretasikan dan tercepat di antara algoritma-algoritma yang menggunakan memori utama di komputer. Algoritma C4.5 mengkonstruksi pohon keputusan dari data

pelatihan, yang berupa kasus-kasus atau record (tupel) dalam basis data. Setiap kasus berisikan nilai dari atribut-atribut untuk sebuah kelas. Setiap atribut dapat berisi data diskret atau kontinyu (numerik). C4.5 juga menangani kasus yang tidak memiliki nilai untuk sebuah atau lebih atribut. Akan tetapi, atribut kelas hanya bertipe diskret dan tidak boleh kosong. Ada tiga prinsip kerja algoritma C4.5 pada tahap belajar dari data, yaitu sebagai berikut :

1. Pembuatan Pohon Keputusan

Obyektif dari algoritma pohon keputusan adalah mengkonstruksi struktur data pohon (dinamakan pohon keputusan) yang dapat digunakan untuk memprediksi kelas dari sebuah kasus atau record baru yang belum memiliki kelas. Algoritma ini memilih pemecahan kasus-kasus yang terbaik dengan menghitung dan membandingkan gain ratio, kemudian pada node-node yang terbentuk di level berikutnya. Demikian seterusnya sampai terbentuk daun - daun.

2. Pemangkasan Pohon Keputusan dan Evaluasi (Opsional)

Karena pohon yang dikonstruksi dapat berukuran besar dan tidak mudah dibaca, C4.5 dapat menyederhanakan pohon dengan melakukan pemangkasan berdasarkan nilai tingkat kepercayaan (confidence level). Selain untuk pengurangan ukuran pohon, pemangkasan juga bertujuan untuk mengurangi tingkat kesalahan prediksi pada kasus (rekord) baru.

3. Pembuatan Aturan-Aturan dari Pohon Keputusan (Opsional)

Aturan-aturan dalam bentuk if-then diturunkan dari pohon keputusan dengan melakukan penelusuran dari akar sampai ke daun. Setiap node dan syarat pencabangannya akan diberikan di if, sedangkan nilai pada daun akan menjadi ditulis di then. Setelah semua aturan dibuat, maka aturan akan disederhanakan (digabung atau diperumum).

3.1 Langkah-Langkah Konstruksi Pohon Keputusan dengan Algoritma

C4.5

Adapun langkah-langkah dalam konstruksi pohon keputusan adalah sebagai berikut :

Langkah 1: Pohon dimulai dengan sebuah simpul yang merepresentasikan sampel data pelatihan yaitu dengan membuat simpul akar.

Langkah 2 : Jika semua sampel berada dalam kelas yang sama, maka simpul ini menjadi daun dan dilabeli menjadi kelas. Jika tidak, gain ratio akan digunakan untuk memilih atribut split, yaitu atribut yang terbaik dalam memisahkan data sampel menjadi kelas-kelas individu.

Langkah 3 : Cabang akan dibuat untuk setiap nilai pada atribut dan data sampel akan dipartisi lagi.

Langkah 4 : Algoritma ini menggunakan proses rekursif untuk membentuk pohon keputusan pada setiap data partisi. Jika sebuah atribut sudah digunakan disebuah simpul, maka atribut ini tidak akan digunakan lagi di simpul anak-anaknya.

Langkah 5 : Proses ini berhenti jika dicapai kondisi seperti berikut :

- Semua sampel pada simpul berada di dalam satu kelas
- Tidak ada atribut lainnya yang dapat digunakan untuk mempartisi sampel lebih lanjut. Dalam hal ini akan diterapkan suara terbanyak. Ini berarti mengubah sebuah simpul menjadi daun dan melabelinya dengan kelas pada suara terbanyak.

Kelebihan dan kekurangan Algoritma C45

Merupakan salah satu algoritma yang sangat populer di kelompok pohon keputusan, algoritma ini sering digunakan untuk membentuk pohon keputusan (Decision Tree). Prinsip kerja algoritma C4.5 adalah dengan membaca seluruh sampel dari penyimpanan dan memuatnya ke memori kemudian melakukan komputasi dengan membaca sampel di memori untuk mengintruksikan Decision Tree.

Meski C4.5 menjadi salah satu algoritma yang sangat populer dan banyak digunakan, algoritma C4.5 memiliki beberapa kelebihan dan kekurangan, apa saja kelebihan dan kekurangannya? Silahkan simak penjelasan berikut dibawah ini ya.

Kelebihan Algoritma C45

1. C4.5 mampu menangani atribut yang kosong (Missing Value).
2. C4.5 mampu menangani atribut dengan kontinu.
3. C4.5 memangkas pohon keputusan untuk mengatasi overfitting.

Kekurangan Algoritma C45

Algoritma C4.5 hanya dapat digunakan untuk menangani sampel-sampel yang dapat disimpan secara keseluruhan dan pada waktu yang bersamaan di memori.

Perbandingan Algoritma C4.5 dengan Algoritma *Naive Bayes Classifier*

Naive Bayes Classifier

merupakan sebuah metoda klasifikasi yang berakar pada teorema Bayes. Sebelum menjelaskan Naive Bayes Classifier ini, akan dijelaskan terlebih dahulu Teorema Bayes yang menjadi dasar dari metoda tersebut. Pada teorema Bayes, bila terdapat dua kejadian yang terpisah (misalkan A dan B), maka teorema Bayes dirumuskan sebagai berikut:

$$P(B|A) = \frac{P(A|B)P(B)}{P(B)}$$

Keterangan:

$P(B|A)$: Probabilitas *posterior*, probabilitas muncul B jika diketahui A

$P(A|B)$: Probabilitas *posterior*, probabilitas muncul A jika diketahui B

$P(A)$: Probabilitas *prior*, probabilitas kejadian A

$P(B)$: Probabilitas *prior*, probabilitas kejadian B

Sedangkan Algoritma C4.5

C4.5 diperkenalkan Quinlan (1996) sebagai versi terbaru dari ID3. Dalam induksi *tree* hanya bisa dilakukan pada fitur bertipe kategorikal (nominal atau ordinal). Sedangkan tipe numerik (interval atau rasio) tidak dapat digunakan. Perbedaan yang membedakan algoritma C4.5 dan ID3 adalah dapat menangani fitur dengan tipe numerik, melakukan pemotongan *decision tree*. Algoritma C4.5 juga menggunakan kriteria dalam menentukan fitur yang menjadi pemecah pada pohon yang diinduksi [7].

$$Gain(S, A) = Entropy(S) - \sum_{i=1}^n \frac{|S_i|}{|S|} * Entropy(S_i)$$

Keterangan :

S : Himpunan Kasus

A : Atribut

N : Jumlah partisi atribut A

$|S_i|$: Jumlah kasus pada partisi ke i

$|S|$: Jumlah kasus dalam S

Sedangkan perhitungan nilai *Entropy* dapat dilihat pada persamaan 2-3:

$$entropy(A) = - \sum_{i=1}^n \frac{|S_i|}{|S|} \log_2 \frac{|S_i|}{|S|}$$

Keterangan :

S : Himpunan kasus

A : Atribut

N : Jumlah partisi atribut A

$|S_i|$: Jumlah kasus pada partisi ke i

$|S|$: Jumlah kasus dalam S

Kesimpulannya :

Kelebihan :

- Pada Metode Algoritma Naive Bayes Classifier lebih mudah untuk digunakan karena hanya memiliki alur perhitungan yang tidak panjang.
- Metode Algoritma C4.5 didapatkan tingkat akurasi lebih tinggi

Kekurangan :

- Pada Metode Algoritma Decision Tree (C4.5) jika data diubah atau ditambah maka perhitungan akan memerlukan waktu yang lebih lama lagi.
- Metode Naive Bayes Classifier didapatkan tingkat akurasi rendah

Nama : Oman Arrohman

NIM : 202420042

Kelas : MTI 23A

Mata Kuliah : Advanced Database

Dosen Pengasuh : Tri Basuki Kurniawan,S.Kom.,M.eng.,Ph.D.

Algoritma C 4.5

No	Outlook	Temperature	Humidity	Windy	Play
1	Sunny	Hot	High	FALSE	No
2	Sunny	Hot	High	TRUE	No
3	Cloudy	Hot	High	FALSE	Yes
4	Rainy	Mild	High	FALSE	Yes
5	Rainy	Cool	Normal	FALSE	Yes
6	Rainy	Cool	Normal	TRUE	Yes
7	Cloudy	Cool	Normal	TRUE	Yes
8	Sunny	Mild	High	FALSE	No
9	Sunny	Cool	Normal	FALSE	Yes
10	Rainy	Mild	Normal	FALSE	Yes
11	Sunny	Mild	Normal	TRUE	Yes
12	Cloudy	Mild	High	TRUE	Yes
13	Cloudy	Hot	Normal	FALSE	Yes
14	Rainy	Mild	High	TRUE	No

Perhitungan 1

Node			Jumlah Kasus	No (S1)	Yes (S2)	Entropy	Gain
1	Total		14	4	10	0,863120569	0,258521037
	Outlook						
		Cloudy	4	0	4	0	
		Rainy	5	1	4	0,721928095	
		Sunny	5	3	2	0,970950594	
	Temperature						0,183850925
		Cool	4	0	4	0	
		Hot	4	2	2	1	
		Mild	6	2	4	0,918295834	
	Humidity						0,370506501
		High	7	4	3	0,985228136	

		Normal	7	0	7	0	
	Windy						0,005977711
		FALSE	8	2	6	0,811278124	
		TRUE	6	2	4	0,918295834	

Perhitungan Entropy : rumus :

$$\text{Entropy (S)} = \sum_{i=1}^n - p_i * \log_2 p_i$$

Keterangan

S : himpunan kasus

A : fitur

n : jumlah partisi S

pi : proporsi dari Si terhadap S

$$\text{Entropy Total} = ((-10 \div 14 * \log_2 (10 \div 14) + -4 \div 14 * \log_2 (4 \div 14))) = 0,863120569$$

$$\text{Entropy Cloudy} = ((-4 \div 4 * \log_2 (4 \div 4) + -0 \div 4 * \log_2 (0 \div 4))) = 0$$

$$\text{Entropy Rainy} = ((-4 \div 5 * \log_2 (4 \div 5) + -1 \div 5 * \log_2 (1 \div 5))) = 0,721928095$$

$$\text{Entropy Sunny} = ((-2 \div 5 * \log_2 (2 \div 5) + -3 \div 5 * \log_2 (3 \div 5))) = 0,970950594$$

$$\text{Entropy Cool} = ((-4 \div 4 * \log_2 (4 \div 4) + -0 \div 4 * \log_2 (0 \div 4))) = 0$$

$$\text{Entropy Hot} = ((-2 \div 4 * \log_2 (2 \div 4) + -2 \div 4 * \log_2 (2 \div 4))) = 1$$

$$\text{Entropy Mild} = ((-4 \div 6 * \log_2 (4 \div 6) + -2 \div 6 * \log_2 (2 \div 6))) = 0,918295834$$

$$\text{Entropy High} = ((-3 \div 7 * \log_2 (3 \div 7) + -4 \div 7 * \log_2 (4 \div 7))) = 0,985228136$$

$$\text{Entropy Normal} = ((-7 \div 7 * \log_2 (7 \div 7) + -0 \div 7 * \log_2 (0 \div 7))) = 0$$

$$\text{Entropy False} = ((-6 \div 8 * \log_2 (6 \div 8) + -2 \div 8 * \log_2 (2 \div 8))) = 0,811278124$$

$$\text{Entropy True} = ((-4 \div 6 * \log_2 (4 \div 6) + -2 \div 6 * \log_2 (2 \div 6))) = 0,918295834$$

Perhitungan Gain : rumus :

$$\text{Gain (S,A)} = \text{Entropy (S)} - \sum_{i=1}^n \frac{|S_i|}{|S|} * \text{Entropy (Si)}$$

Keterangan:

- S : himpunan

- A : atribut

- n : jumlah partisi atribut A

- | Si | : jumlah kasus pada partisi ke-i

- | S | : jumlah kasus dalam S

$$\text{Gain Outlook} = (0,863120569) - ((4 \div 14) * 0) - ((5 \div 14) * 0,721928095) - ((5 \div 14) * 0,970950594)$$

$$= 0,258521037$$

$$\text{Gain Temperature} = (0,863120569) - (4 \div 14) * 0 - (4 \div 14) * 1 - (6 \div 14) * 0,918295834 \\ = 0,183850925$$

$$\text{Gain Humidity} = (0,863120569) - (7 \div 14) * 0,985228136 - (7 \div 14) * 0 \\ = 0,370506501$$

$$\text{Gain Windy} = (0,863120569) - (8 \div 14) * 0,811278124 - (6 \div 14) * 0,918295834 \\ = 0,005977711$$

Sepert yg terlihat pada perhitungan diatas, diperoleh bahwa atribut dgn Gain tertinggi adalah Humidity = **0,370506501**

Maka Humidity menjadi node akar

Humidity memiliki dua nilai yaitu "High" dan "Normal"

Humidity -> "Normal" sdh mengklasifikasikan kasus menjadi 1 yaitu keputusannya "yes"

Untuk humidity -> "High" msh perlu dilakukan perhitungan lagi (karena masih terdapat "yes" dan "no")

Humadity "High"

No	Outlook	Temperature	Humidity	Windy	Play
1	Sunny	Hot	High	FALSE	No
2	Sunny	Hot	High	TRUE	No
3	Cloudy	Hot	High	FALSE	Yes
4	Rainy	Mild	High	FALSE	Yes
5	Sunny	Mild	High	FALSE	No
6	Cloudy	Mild	High	TRUE	Yes
7	Rainy	Mild	High	TRUE	No

Perhitungan 1.1

Node			Jumlah Kasus	No (S1)	Yes (S2)	Entropy	Gain
1.1	Total		7	4	3	0,985228136	
	Outlook						0,69951385
		Cloudy	2	0	2	0	
		Rainy	2	1	1	1	
		Sunny	3	3	0	0	
	Temperature						0,020244207
		Cool	0	0	0	0	
		Hot	3	2	1	0,918295834	
		Mild	4	2	2	1	
	Windy						0,020244207

		FALSE	4	2	2	1	
		TRUE	3	2	1	0,918295834	

Perhitungan Entropy Humadity “High” 1.1

Entropy Total	$= ((-3 \div 7 * \log_2 (3 \div 7) + -4 \div 7 * \log_2 (4 \div 7))) = 0,985228136$					
Entropy Cloudy	$= ((-2 \div 2 * \log_2 (2 \div 2) + -0 \div 2 * \log_2 (0 \div 2))) = 0$					
Entropy Rainy	$= ((-1 \div 2 * \log_2 (1 \div 2) + -1 \div 2 * \log_2 (1 \div 2))) = 1$					
Entropy Sunny	$= ((-0 \div 3 * \log_2 (0 \div 3) + -3 \div 3 * \log_2 (3 \div 3))) = 0$					
Entropy Cool	$= ((-0 \div 0 * \log_2 (0 \div 0) + -0 \div 0 * \log_2 (0 \div 0))) = 0$					
Entropy Hot	$= ((-1 \div 3 * \log_2 (1 \div 3) + -2 \div 3 * \log_2 (2 \div 3))) = 0,918295834$					
Entropy Mild	$= ((-2 \div 4 * \log_2 (2 \div 4) + -2 \div 4 * \log_2 (2 \div 4))) = 1$					
Entropy False	$= ((-2 \div 4 * \log_2 (2 \div 4) + -2 \div 4 * \log_2 (2 \div 4))) = 1$					
Entropy True	$= ((-1 \div 3 * \log_2 (1 \div 3) + -2 \div 3 * \log_2 (2 \div 3))) = 0,918295834$					

Perhitungan Gain Humadity “High” 1.1

$$\begin{aligned} \text{Gain Outlook} &= (0,985228136) - (2 \div 7) * 0 - (2 \div 7) * 1 - (3 \div 7) * 0 \\ &= 0,69951385 \end{aligned}$$

$$\begin{aligned} \text{Gain Temperature} &= (0,985228136) - (0 \div 7) * 0 - (3 \div 7) * 0,918295834 - (4 \div 7) * 1 \\ &= 0,020244207 \end{aligned}$$

$$\begin{aligned} \text{Gain Windy} &= (0,985228136) - (4 \div 7) * 1 - (3 \div 7) * 0,918295834 \\ &= 0,020244207 \end{aligned}$$

Didapat Gain tertinggi -> outlook -> 0,69951385

Maka “Outlook” menjadi node cabang dari atribut humidity yg bernilai “High”

Berdasarkan atribut “Outlook” terdpt 3 nilai

Cloudy

Rainy

Sunny

Karena “Cloudy” pasti bernilai “Yes” dan “Sunny” pasti bernilai “No”, maka tidak perlu dilakukan perhitungan lagi.

Sedangkan “Rainy” bernilai “yes” dan “No”, maka masih perlu dilakukan perhitungan lagi

Humadity “High” dan Outlook “Rainy”

No	Outlook	Temperature	Humidity	Windy	Play
4	Rainy	Mild	High	FALSE	Yes
7	Rainy	Mild	High	TRUE	No

Perhitungan 1.1.2

Node			Jumlah Kasus	No (S1)	Yes (S2)	Entropy	Gain
1..1.2	Total		2	1	1	1	
	Temperature						0
		Cool	0	0	0	0	
		Hot	0	0	0	0	
		Mild	2	1	1	1	
	Windy						1
		FALSE	1	0	1	0	
		TRUE	1	1	0	0	

Perhitungan Entropy "Rainy" 1.1.2

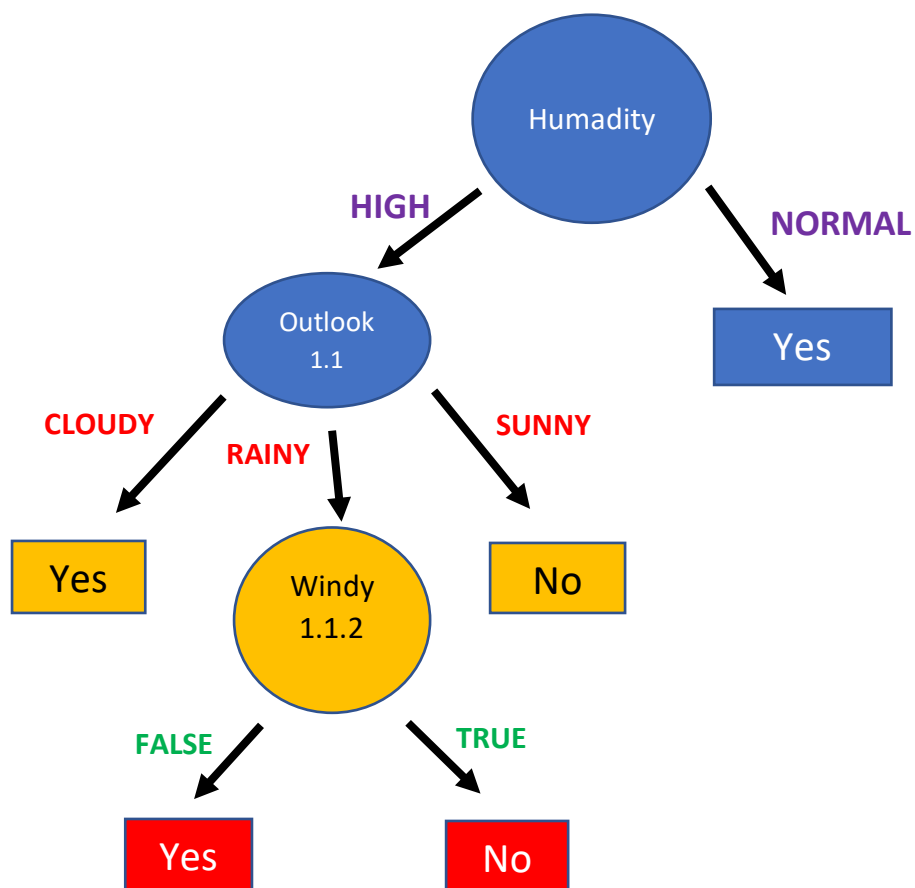
Entropy Total	$= ((-1 \div 2) * \log_2 (1 \div 2) + (-1 \div 2) * \log_2 (1 \div 2))$	= 1
Entropy Cool	$= ((-0 \div 0) * \log_2 (0 \div 0) + -0 \div 0 * \log_2 (0 \div 0))$	= 0
Entropy Hot	$= ((-0 \div 0) * \log_2 (0 \div 0) + -0 \div 0 * \log_2 (0 \div 0))$	= 0
Entropy Mild	$= ((-1 \div 2 * \log_2 (1 \div 2) + -1 \div 2 * \log_2 (1 \div 2))$	= 1
Entropy False	$= ((-1 \div 1 * \log_2 (1 \div 1) + -0 \div 1 * \log_2 (0 \div 1))$	= 0
Entropy True	$= ((-0 \div 1 * \log_2 (0 \div 1) + -1 \div 0 * \log_2 (1 \div 0))$	= 0

Perhitungan Gain "Rainy" 1.1.2

$$\text{Gain Temperature} = (1) - (0 \div 2) * 0 - (0 \div 2) * 0 - (2 \div 2) * 1 \\ = 0$$

$$\text{Gain Windy} = (1) - (1 \div 2) * 0 - (1 \div 2) * 0 \\ = 1$$

POHON KEPUTUSAN



Kelebihan Algoritma C4.5 :

1. C4.5 mampu menangani atribut yang kosong (Missing Value).
2. C4.5 mampu menangani atribut dengan kontinu.
3. C4.5 memangkas pohon keputusan untuk mengatasi overfitting.

Kekurangan Algoritma C4.5 :

1. Algoritma C4.5 hanya dapat digunakan untuk menangani sampel-sampel yang dapat disimpan secara keseluruhan dan pada waktu yang bersamaan di memori.

Nama : Puspita Dewi Setyadi
Nim : 202420011

Dari dataset berikut ini, silahkan buat perhitungan manual dari algoritma decision tree C45 secara lengkap dan gambarkan pohon keputusan yang diperoleh.

No	Outlook	Temperature	Humidity	Windy	Play
1	Sunny	Hot	High	FALSE	No
2	Sunny	Hot	High	TRUE	No
3	Cloudy	Hot	High	FALSE	Yes
4	Rainy	Mild	High	FALSE	Yes
5	Rainy	Cool	Normal	FALSE	Yes
6	Rainy	Cool	Normal	TRUE	Yes
7	Cloudy	Cool	Normal	TRUE	Yes
8	Sunny	Mild	High	FALSE	No
9	Sunny	Cool	Normal	FALSE	Yes
10	Rainy	Mild	Normal	FALSE	Yes
11	Sunny	Mild	Normal	TRUE	Yes
12	Cloudy	Mild	High	TRUE	Yes
13	Cloudy	Hot	Normal	FALSE	Yes
14	Rainy	Mild	High	TRUE	No

TABEL PERHITUNGAN NODE 1

Rumus :

Entropy = $((-\text{don't play} / \text{jumlah total kasus}) * \log((\text{don't play} / \text{jumlah total kasus});2)) + ((-\text{play} / \text{jumlah total kasus}) * \log((\text{play} / \text{jumlah total kasus});2))$

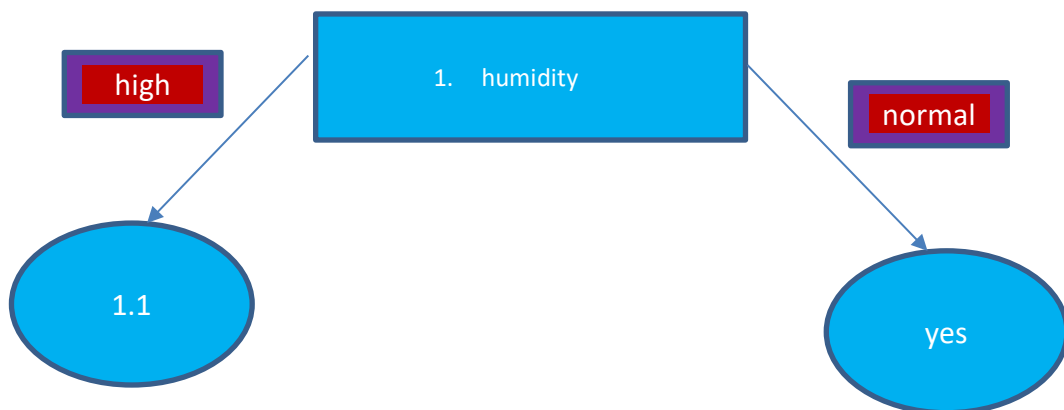
$$1. ((-4/14) * \log(4/14);2) + ((-10/14) * \log(10/14);2) = 0,863121$$

Gain = jumlah total entropy – ((jumlah kasus variable x/jumlah total kasus)*entropy variable x)+((jumlah kasus variable x/jumlah total kasus)*entropy variable x)

1. Outlook = $0,863121 - ((4/14)*0) + ((5/14)*0,721928) + ((5/14)*0,970951) = 0,25852104$
2. Temp = $0,863121 - ((4/14)*0) + ((4/14)*1) + ((6/14)*0,918296) = 0,18385093$
3. Humidity = $0,863121 - ((7/14)*0,985228) + ((7/14)*0) = 0,3705065$
4. Windy = $0,863121 - ((8/14)*0,811278) + ((6/14)*0,918296) = 0,00597771$

		Jumlah kasus	Don't play (s1)	Play (s2)	Entropy	Gain
TOTAL		14	4	10	0,863121	
OUTLOOK						0,25852104
	Cloudy	4	0	4	0	
	Raini	5	1	4	0,721928	
	Sunny	5	3	2	0,970951	
TEMP						0,18385093
	Cool	4	0	4	0	
	Hot	4	2	2	1	
	Mild	6	2	4	0,918296	
HUMIDITY						0,3705065
	Hight	7	4	3	0,985228	
	Normal	7	0	7	0	
WINDY						0,00597771
	No	8	2	6	0,811278	
	yes	6	2	4	0,918296	

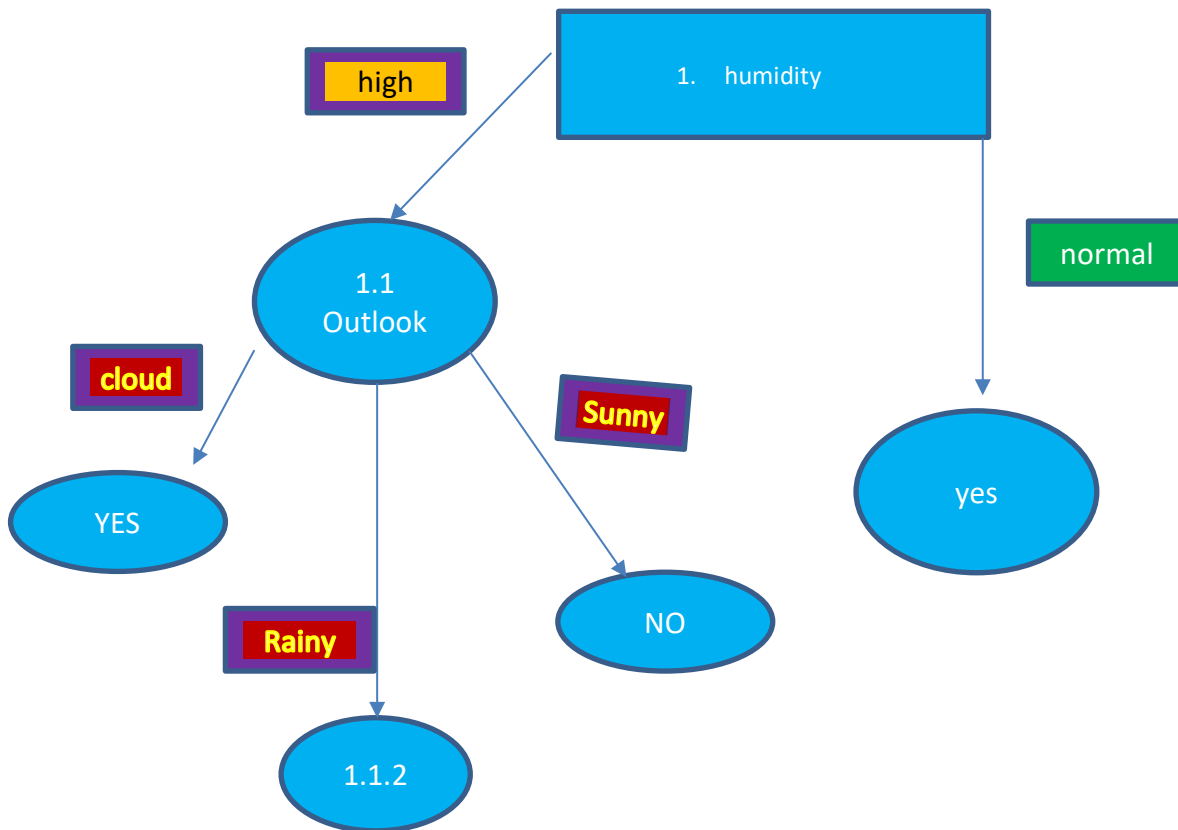
DECISION TREE



TABEL PERHITUNGAN 1.1

		Jumlah kasus	Don't play (s1)	Play (s2)	Entropy	Gain
HUMIDITY		7	4	3	0,985218136	
OUTLOOK						0,69951385
	Cloudy	2	0	2	0	
	Rainy	2	1	1	1	
	Sunny	3	3	0	0	
TEMP						0,020244207
	Cool	0	0	0	0	
	Hot	3	2	1	0,918295834	
	Mild	4	2	2	1	
WINDY						0,020244207
	No	4	2	2	1	
	yes	3	2	1	0,918295834	

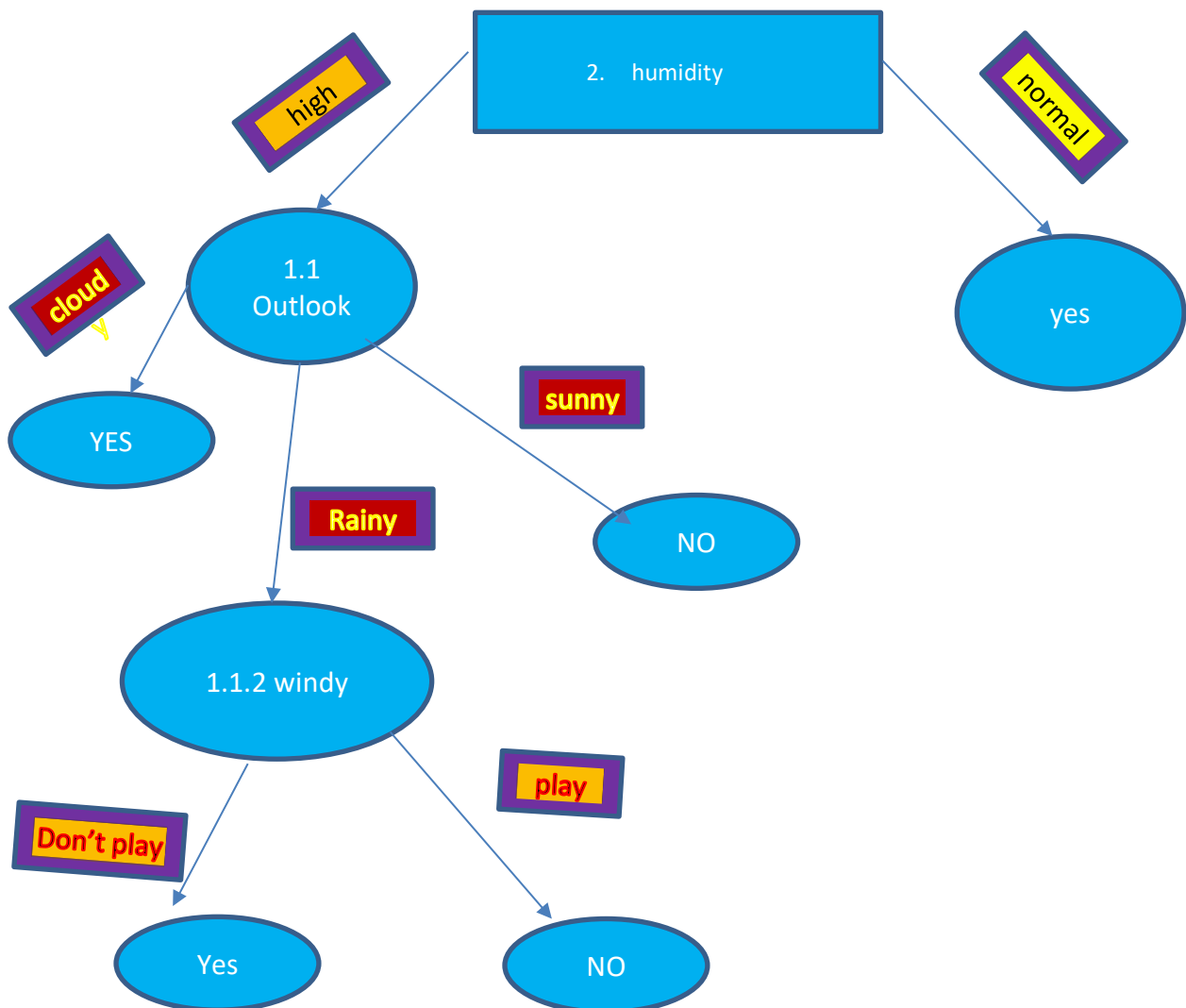
DECISION TREE



TABEL PERHITUNGAN 1.1.2

NODE 1.1.2		Jumlah kasus	Don't play (s1)	Play (s2)	Entropy	Gain
HUMIDITY HIGHT DAN OUTLOOK RAINY		2	1	1	1	
TEMP						0
	Cool	0	0	0	0	
	Hot	0	0	0	0	
	Mild	2	1	1	1	
WINDY						1
	No	1	0	1	0	
	yes	1	1	0	0	

DECISION TREE



Tugas 04

Dari dataset berikut ini, silahkan buat perhitungan manual dari algoritma decision tree C45 secara lengkap dan gambarkan pohon keputusan yang diperoleh

No	Outlook	Temperature	Humidity	Windy	Play
1	Sunny	Hot	High	FALSE	No
2	Sunny	Hot	High	TRUE	No
3	Cloudy	Hot	High	FALSE	Yes
4	Rainy	Mild	High	FALSE	Yes
5	Rainy	Cool	Normal	FALSE	Yes
6	Rainy	Cool	Normal	TRUE	Yes
7	Cloudy	Cool	Normal	TRUE	Yes
8	Sunny	Mild	High	FALSE	No
9	Sunny	Cool	Normal	FALSE	Yes
10	Rainy	Mild	Normal	FALSE	Yes
11	Sunny	Mild	Normal	TRUE	Yes
12	Cloudy	Mild	High	TRUE	Yes
13	Cloudy	Hot	Normal	FALSE	Yes
14	Rainy	Mild	High	TRUE	No

Kumpulkan dalam format MS Word. Sertakan penjelasan dan pendapat anda tentang algoritma ini, kelebihan dan kekurangannya berbanding metode lain yang sejenis (tentukan sendiri metode pembandingnya).

Jawabannya :

Pada kasus diatas untuk menentukan pengambil keputusan harus perlu dihitung bagaimana peluang – peluang yang muncul dengan menggunakan algoritma decision tree C45, adapun langkahnya sebagai berikut :

1. Menghitung jumlah kasus, jumlah kasus untuk keputusan Yes, jumlah kasus untuk keputusan No, dan Entropy dari semua kasus dan kasus yang dibagi berdasarkan atribut OUTLOOK, TEMPERATUR, HUMIDITY, dan WINDY.
2. Lakukan perhitungan Gain untuk setiap atribut.
 - a. Entropy

Entropy adalah suatu parameter untuk mengukur tingkat keberagaman (heterogenitas) dari kumpulan data. Jika nilai dari entropy semakin besar, maka tingkat keberagaman suatu kumpulan data semakin besar.

Rumus untuk menghitung entropy sebagai berikut:

$$Entropy(S) = - \sum_{i=1}^m p_i \log_2(p_i)$$

dimana:

m = jumlah kelas klasifikasi

p_i = jumlah proporsi sampel (peluang) untuk kelas i

Sedangkan rumus untuk entropy pada masing-masing variabel adalah:

$$Entropy_A(S) = \sum_v \frac{|S_v|}{|S|} Entropy(S_v)$$

dimana:

A = variabel

v = nilai yang mungkin untuk variabel A

$|S_v|$ = jumlah sampel untuk nilai v

$|S|$ = jumlah sampel untuk seluruh sampel data

$Entropy(S_v) = Entropy$ untuk sampel yang memiliki nilai v

b. Gain

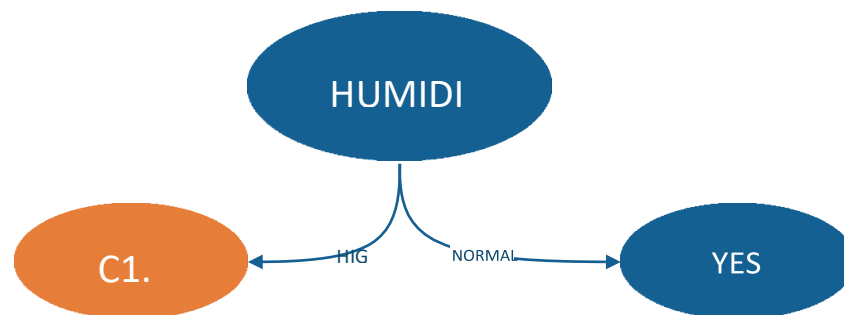
Gain adalah ukuran efektifitas suatu variabel dalam mengklasifikasikan data. Gain dari suatu variabel merupakan selisih antara nilai entropy total dengan entropy dari variabel tersebut. Gain dapat dirumuskan dengan :

$$Gain(A) = Entropy(S) - Entropy_A(S)$$

Dari data diatas didapatkan data sebagai berikut (detail perhitungan terlampir di excel) :

Node	Deskripsi	Jml Kasus (S)	Count Play "YES"	Count Play "NO"	Entropy	Gain
	Total	14	10	4	0,86312057	
1	OUTLOOK					0,25852104
	Cloudy	4	4			
	Rainy	5	4	1	0,72192809	
	Sunny	5	2	3	0,97095059	
	TEMPERATURE					0,18385093
	Cool	4	4		0	
	Hot	4	2	2	1	
	Mild	6	4	2	0,91829583	
	HUMIDITY					0,3705065
	High	7	3	4	0,98522814	
	Normal	7	7			
	WINDY					0,00597771
	FALSE	8	6	2	0,81127812	
	TRUE	6	4	2	0,91829583	

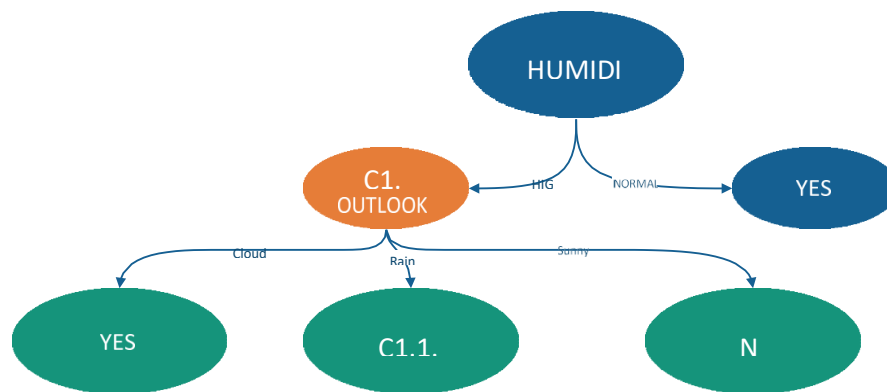
Dapat diketahui dari data diatas bahwa atribut dengan Gain tertinggi adalah HUMADITY , yaitu sebesar 0.37. Jadi HUMADITY yang menjadi node akar. Ada dua nilai atribut dari HUMADITY, yaitu HIGH dan NORMAL. Dari kedua atribut, nilai atribut NORMAL adalah 1, yaitu keputusan Yes, sehingga tidak perlu perhitungan lebih lanjut, tetapi untuk HIGH masih perlu dilakukan perhitungan lagi. Apabila dibuat tree nya sebagai berikut :



Menghitung jumlah kasus, jumlah kasus untuk keputusan Yes , jumlah kasus untuk keputusan No , dan Entropy dari semua kasus dan kasus yang dibagi berdasarkan atribut OUTLOOK, TEMPERATURE, dan WINDY yang dapat menjadi node akar dari nilai atribut HIGH . Kemudian lakukan perhitungan Gain untuk tiap tiap atribut. Hasil perhitungan sebagai berikut :

Node	Deskripsi	Jml Kasus (S)	Count Play "YES"	Count Play "NO"	Entropy	Gain
1	Total	7	3	4	0,98522814	
	OUTLOOK					0,69951385
	Cloudy	2	2	0	0	
	Rainy	2	1	1	1	
	Sunny	3	0	3	0	
	TEMPERATURE					0,02024421
	Cool	0	0	0	0	
	Hot	3	1	2	0,91829583	
	Mild	4	2	2	1	
	HUMIDITY					0
	High	7	3	4	0,98522814	
	Normal	7	7	0	0	
	WINDY					0,02024421
	FALSE	4	2	2	1	
	TRUE	3	1	2	0,91829583	

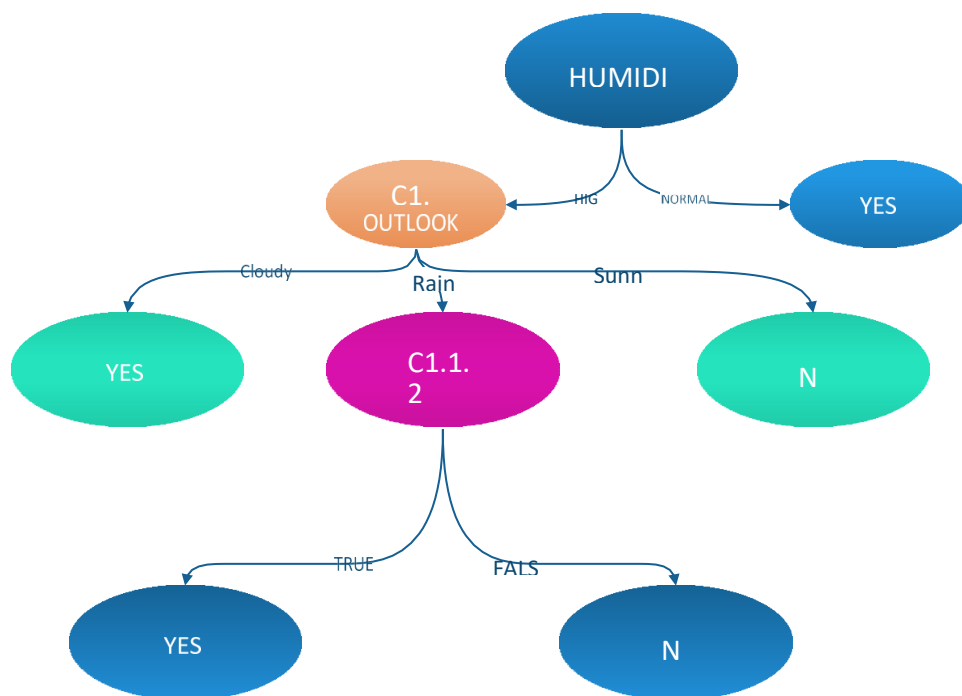
Dapat diketahui bahwa atribut dengan Gain tertinggi adalah OUTLOOK , yaitu sebesar 0.67. Jadi OUTLOOK dapat menjadi node cabang dari nilai atribut HIGH. Ada tiga nilai atribut dari OUTLOOK, yaitu CLOUDY, RAINY dan SUNNY. Dari ketiga nilai atribut tersebut, nilai atribut CLOUDY adalah 1, yaitu keputusannya Yes dan nilai atribut SUNNY menjadi keputusan No , sehingga tidak perlu dilakukan perhitungan lebih lanjut, tetapi nilai atribut RAINY masih perlu dilakukan perhitungan lagi.



Menghitung jumlah kasus, jumlah kasus untuk keputusan Yes, jumlah kasus untuk keputusan No, dan Entropy dari semua kasus dan kasus yang dibagi berdasarkan atribut TEMPERATURE dan WINDY yang dapat menjadi node cabang dari nilai atribut RAINY. Kemudian lakukan perhitungan Gain untuk tiap-tiap atribut. Hasil perhitungan seperti yang terdapat pada table dibawah ini :

Node	Deskripsi	Jml Kasus (S)	Count Play "YES"	Count Play "NO"	Entropy	Gain
1	Total	2	1	1	1	
	OUTLOOK					0
	Cloudy	2	2		0	
	Rainy	2	1	1	1	
	Sunny	3		3	0	
	TEMPERATURE					0
	Cool	0	0	0	0	
	Hot	0	0	0	0	
	Mild	2	1	1	1	
	HUMIDITY					-2,4482985
	High	7	3	4	0,98522814	
	Normal	7	7			
	WINDY					1
	FALSE	1	1	0	0	
	TRUE	1	0	1	0	

Dapat diketahui bahwa atribut dengan Gain tertinggi adalah WINDY, yaitu sebesar 1. Jadi WINDY dapat menjadi node cabang dari nilai atribut RAINY. Ada dua nilai atribut dari WINDY, yaitu FALSE dan TRUE. Dari kedua nilai atribut tersebut, nilai atribut FALSE adalah 1, yaitu keputusannya Yes dan nilai atribut TRUE menjadi keputusan No, sehingga tidak perlu dilakukan perhitungan lebih lanjut untuk nilai atribut ini. Dapat digambarkan untuk pohon keputusannya sebagai berikut:



Merupakan gambar akhir dari proses Data Mining menggunakan algoritma C4.5 pada contoh kasus keputusan bermain tenis. Berdasarkan gambar di atas dapat ditarik kesimpulan bahwa jika humidity normal maka keputusannya adalah yes, jika humidity adalah high, outlook adalah cloudy maka keputusannya adalah bermain (yes), selanjutnya jika humidity adalah high, outlook adalah rainy dan windy adalah true, maka keputusannya adalah bermain (yes). Dan jika humidity adalah high, outlook adalah rainy dan windy adalah false, maka keputusannya adalah tidak bermain (no). Terakhir jika humidity adalah high, outlook adalah no, maka keputusannya adalah tidak bermain (no).

Algoritma C4.5

Algoritma yang dapat digunakan untuk memprediksi atau mengklasifikasi suatu kejadian dengan pembentukan pohon keputusan antara lain algoritma C4.5, yang merupakan salah satu algoritma induksi pohon keputusan yang dikembangkan oleh J.Ross Quinlan (Yamin, et al, 2014).

Algoritma C4.5 merupakan kelompok algoritma decision tree. Algoritma ini mempunyai input berupa training samples dan samples. Training samples merupakan data contoh yang digunakan untuk membangun sebuah tree yang telah diuji kebenarannya. Sedangkan samples merupakan field-field data yang digunakan sebagai parameter dalam klasifikasi data (Sujana,2010).

Secara umum alur proses algoritma C4.5 untuk membangun pohon keputusan dalam data mining adalah sebagai berikut (Mahmud, et al, 2014):

- a. Pilih atribut sebagai akar
- b. Buat cabang untuk tiap-tiap nilai
- c. Bagi kasus dalam cabang.
- d. Ulangi proses untuk setiap cabang sampai semua kasus pada cabang memiliki kelas yang sama.

TUGAS
DECISION TREES ALGORITHM

Nama : Robby Prabowo

NIM : 202420001

Kelas MTIA1

Tugas 04

No	Outlook	Temperature	Humidity	Windy	Play
1	Sunny	Hot	High	False	No
2	Sunny	Hot	High	True	No
3	Cloudy	Hot	High	False	Yes
4	Rainy	Mild	High	False	Yes
5	Rainy	Cool	Normal	False	Yes
6	Rainy	Cool	Normal	True	Yes
7	Cloudy	Cool	Normal	True	Yes
8	Sunny	Mild	High	False	No
9	Sunny	Cool	Normal	False	Yes
10	Rainy	Mild	Normal	False	Yes
11	Sunny	Mild	Normal	True	Yes
12	Cloudy	Mild	High	True	Yes
13	Cloudy	Hot	Normal	False	Yes
14	Rainy	Mild	High	True	No

Rumus Penyelesaian

$$\text{Entropy}(S) = \sum - p(I) \cdot \log_2 p(I)$$

$$\text{Gain}(S, A) = \text{Entropy}(S) - \sum [p(S|A) \cdot \text{Entropy}(S|A)]$$

Entropy

Kita hitung dulu Nilai Entropy. Kolom Play terdiri dari 14 instance dan terdiri dari 2 label yakni Yes dan No. Play Yes = 10, _Play No= 4

$$\begin{aligned}\text{Entropy(Decision)} &= -p(\text{Yes}) \cdot \log_2 p(\text{Yes}) - p(\text{No}) \cdot \log_2 p(\text{No}) \\ \text{Entropy(Decision)} &= - (10/14) \cdot \log_2(10/14) - (4/14) \cdot \log_2(4/14) \\ &= -(0.714) \cdot \log_2(0.714) - (0.4) \cdot \log_2(0.4) \\ &= 0.347 + 0.528 \\ &= \mathbf{0.875}\end{aligned}$$

Menghitung faktor dominan untuk memutuskan bermain.

Faktor Windy dan Play

False Factor windy to Play

No	Outlook	Temperature	Humidity	Windy	Play
1	Sunny	Hot	High	False	No
3	Cloudy	Hot	High	False	Yes
4	Rainy	Mild	High	False	Yes
5	Rainy	Cool	Normal	False	Yes
8	Sunny	Mild	High	False	No
9	Sunny	Cool	Normal	False	Yes
10	Rainy	Mild	Normal	False	Yes
13	Cloudy	Hot	Normal	False	Yes

Tabel faktor false windy terdiri dari **8 instance**. False No = 2, False Yes = 6

$$\begin{aligned}\text{Entropy play|windy=False} &= -(2/8) \cdot \log_2(2/8) - (6/8) \cdot \log_2(6/8) \\ &= -(0.25) \cdot \log_2(0.25) - (0.75) \cdot \log_2(0.75) \\ &= -(0.25) \cdot (-2) - (0.75) \cdot (-0.415) \\ &= 0.50 + 0.311 \\ &= \mathbf{0.811}\end{aligned}$$

Menghitung True Factor windy to Play

No	Outlook	Temperature	Humidity	Windy	Play
2	Sunny	Hot	High	True	No
6	Rainy	Cool	Normal	True	Yes
7	Cloudy	Cool	Normal	True	Yes
11	Sunny	Mild	Normal	True	Yes
12	Cloudy	Mild	High	True	Yes
14	Rainy	Mild	High	True	No

Tabel faktor True windy terdiri dari **6 instance**. True No = 2, True Yes = 4

$$\begin{aligned}
 \text{Entropy play|windy=False} &= -(2/6) \cdot \log_2(2/6) - (4/6) \cdot \log_2(4/6) \\
 &= -(0.333) \cdot \log_2(0.333) - (0.667) \cdot \log_2(0.667) \\
 &= -(0.333) \cdot (-1.586) - (0.667) \cdot (-0.586) \\
 &= 0.528 + 0.390 \\
 &= \mathbf{0.918}
 \end{aligned}$$

Menghitung Gain-Play|Windy

$$\begin{aligned}
 \text{Gain(Play, Windy)} &= \text{Entropy(Play)} - [p(\text{Play|Windy=False}) \cdot \text{Entropy(Play|Windy=False)} \\
 &\quad] - [p(\text{Play|Windy=True}) \cdot \text{Entropy(Play|Windy=True)}]
 \end{aligned}$$

$$\begin{aligned}
 &= 0.875 - [(8/14) \cdot 0.811] - (6/14) \cdot 0.918 \\
 &= 0.804
 \end{aligned}$$

Maka Gain(Play, Windy) = **0.804**

Selanjutnya kolom 1

Menghitung Gain(Play, Outlook)

Faktor Outlook Sunny

No	Outlook	Temperature	Humidity	Windy	Play
1	Sunny	Hot	High	False	No
2	Sunny	Hot	High	True	No
8	Sunny	Mild	High	False	No
9	Sunny	Cool	Normal	False	Yes
11	Sunny	Mild	Normal	True	Yes

Tabel faktor Sunny Play terdiri dari **5 instance**. Play No = 3, Play Yes = 2
 Entropy $\text{play}|\text{Outlook}=\text{Sunny}$ $= -(3/5).\log_2(3/5) - (2/5).\log_2(2/5)$
 $= -(0.6).\log_2(0.6) - (0.4).\log_2(0.4)$
 $= -(0.6).(-0.737) - (0.4).(-1.322)$
 $= 0.442 + 0.5288$
 $= \mathbf{0.971}$

Faktor Cloudy Outlook

No	Outlook	Temperature	Humidity	Windy	Play
3	Cloudy	Hot	High	False	Yes
7	Cloudy	Cool	Normal	True	Yes
12	Cloudy	Mild	High	True	Yes
13	Cloudy	Hot	Normal	False	Yes

Tabel faktor Cloudy- Play terdiri dari **4 instance**. Play No = 0, Play Yes = 4
 Entropy $\text{play}|\text{Outlook}=\text{Cloudy}$ $= -(0/4).\log_2(0/4) - (4/4).\log_2(4/4)$
 $= \mathbf{0}$

Faktor Rainy Outlook

No	Outlook	Temperature	Humidity	Windy	Play
4	Rainy	Mild	High	False	Yes
5	Rainy	Cool	Normal	False	Yes
6	Rainy	Cool	Normal	True	Yes
10	Rainy	Mild	Normal	False	Yes
14	Rainy	Mild	High	True	No

Tabel faktor Rainy Play terdiri dari **5 instance**. Play No = 1, Play Yes = 4
 Entropy $\text{play}|\text{Outlook}=\text{Sunny}$ $= -(1/5).\log_2(1/5) - (4/5).\log_2(4/5)$
 $= -(0.2).\log_2(0.2) - (0.8).\log_2(0.8)$
 $= -(0.2).(-2.321) - (0.8).(-0.322)$
 $= 0.464 + 0.2576$
 $= \mathbf{0.722}$

Maka Gain (Play,Outlook) $= 0.875 - [(5/14.0.971) - (0) - (5/14.0.722)]$
 $= 0.787$

.....

Selanjutnya Kolom 2 (Temperature)
Menghitung Gain(play,temperature)

Faktor Hot Temperatur

No	Outlook	Temperature	Humidity	Windy	Play
1	Sunny	Hot	High	False	No
2	Sunny	Hot	High	True	No
3	Cloudy	Hot	High	False	Yes
13	Cloudy	Hot	Normal	False	Yes

Tabel faktor Hot Play terdiri dari **4 instance**. Play No = 2, Play Yes = 2

$$\begin{aligned}
 \text{Entropy play|Temperature=Hot} &= -(2/4) \cdot \log_2(2/4) - (2/4) \cdot \log_2(2/4) \\
 &= -(0.5) \cdot \log_2(0.5) - (0.5) \cdot \log_2(0.5) \\
 &= -(0.5) \cdot (-1) - (0.5) \cdot (-1) \\
 &= 0.5 + 0.5 \\
 &= \mathbf{1.00}
 \end{aligned}$$

Faktor Mild Temperature

No	Outlook	Temperature	Humidity	Windy	Play
4	Rainy	Mild	High	False	Yes
8	Sunny	Mild	High	False	No
10	Rainy	Mild	Normal	False	Yes
11	Sunny	Mild	Normal	True	Yes
12	Cloudy	Mild	High	True	Yes
14	Rainy	Mild	High	True	No

Tabel faktor Mild Play terdiri dari **6 instance**. Play No = 2, Play Yes = 4

$$\begin{aligned}
 \text{Entropy play|Temperature=Mild} &= -(2/6) \cdot \log_2(2/6) - (4/6) \cdot \log_2(4/6) \\
 &= -(0.333) \cdot \log_2(0.333) - (0.667) \cdot \log_2(0.667) \\
 &= -(0.333) \cdot (-1.586) - (0.667) \cdot (-0.584) \\
 &= 0.528 + 0.389 \\
 &= \mathbf{0.92}
 \end{aligned}$$

Faktor cold Temperature

No	Outlook	Temperature	Humidity	Windy	Play
5	Rainy	Cool	Normal	False	Yes
6	Rainy	Cool	Normal	True	Yes
7	Cloudy	Cool	Normal	True	Yes
9	Sunny	Cool	Normal	False	Yes

Tabel faktor Cold Play terdiri dari **4 instance**. Play No = 0, Play Yes = 4
 Entropy play|Temperature=Cool $= -(0/4) \cdot \log_2(0/4) - (4/4) \cdot \log_2(4/4)$
 $= 0$

Maka Gain (**Play, Temperature**) $= 0.875 - [(4/14) \cdot 1 - (6/14) \cdot 0.92] - (0)$
 $= 0.875 - [(0.286) - (0.395)] - (0)$
 $= 1.065$

Selanjutnya menghitung Kolom 3 (Humidity)

Menghitung Gain|Play, Humidity

Faktor High Humidity

No	Outlook	Temperature	Humidity	Windy	Play
1	Sunny	Hot	High	False	No
2	Sunny	Hot	High	True	No
3	Cloudy	Hot	High	False	Yes
4	Rainy	Mild	High	False	Yes
8	Sunny	Mild	High	False	No
12	Cloudy	Mild	High	True	Yes
14	Rainy	Mild	High	True	No

Tabel faktor High Play terdiri dari **7 instance**. Play No = 4, Play Yes = 3
 Entropy play|Outlook=Sunny $= -(4/7) \cdot \log_2(4/7) - (3/7) \cdot \log_2(3/7)$
 $= -(0.571) \cdot \log_2(0.571) - (0.428) \cdot \log_2(0.428)$
 $= -(0.571) \cdot (-0.808) - (0.428) \cdot (-1.224)$
 $= 0.461 + 0.524$
 $= 0.985$

Faktor Normal Humidity

No	Outlook	Temperature	Humidity	Windy	Play
5	Rainy	Cool	Normal	False	Yes
6	Rainy	Cool	Normal	True	Yes
7	Cloudy	Cool	Normal	True	Yes
9	Sunny	Cool	Normal	False	Yes
10	Rainy	Mild	Normal	False	Yes
13	Cloudy	Hot	Normal	False	Yes

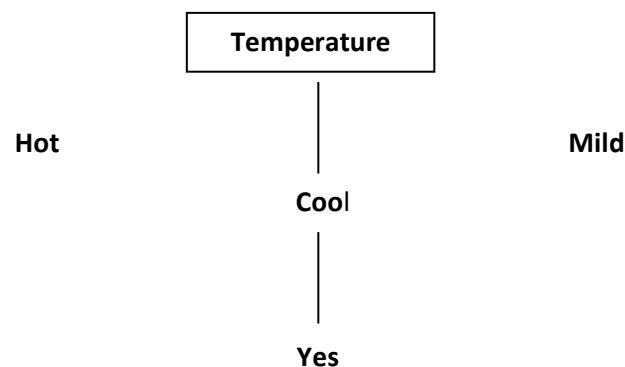
Tabel faktor Normal Play terdiri dari 6 instance. Play No = 0, Play Yes = 6
 Entropy play|Outlook=Sunny = $-(0/6) \cdot \log_2(0/6) - (6/6) \cdot \log_2(6/6)$
 = 0

Maka Gain (Play, Humadity) = $0.875 - [(7/14) \cdot 0.985] - (0)$
 = $0.875 - 0.492$
 = 0.383

Dari perhitungan diatas didapat

1. Gain (Play, Windy) = 0.804
2. Gain (Play, Outlook) = 0.787
3. Gain (Play, Temperature) = 1.605
4. Gain (Play, Humadity) = 0.383

Karena nilai temperatur paling besar, maka pohon keputusan paling atas adalah faktor Temperature.



Cool Temperatur On Play

Dari tabel dibawah, Keputusan untuk bermain diambil jika temperatur dingin/cool

No	Outlook	Temperature	Humidity	Windy	Play
5	Rainy	Cool	Normal	False	Yes
6	Rainy	Cool	Normal	True	Yes
7	Cloudy	Cool	Normal	True	Yes
9	Sunny	Cool	Normal	False	Yes

Hot Temperatur On Play.

No	Outlook	Temperature	Humidity	Windy	Play
1	Sunny	Hot	High	False	No
2	Sunny	Hot	High	True	No
3	Cloudy	Hot	High	False	Yes
13	Cloudy	Hot	Normal	False	Yes

Mencari nilai hubungan Outlook-Temperatur / Gain (temp,Outlook)

Entropy Hot Temperatur = $2/4$ no dan $2/4$ yes

$$\text{Nilai Entrophy} = -(2/4) \cdot \log_2(2/4) - (2/4) \cdot \log_2(2/4) \\ = 1.00$$

Mencari entropy Sunny

karena, instance sunny = 2, nilai play No = 2 dan nilai Play yes = 0, maka
Entropy Sunny = 0

Mencari entropy Cloudy

Karena instance claudy = 2, nilai Play No = 0 dan nilai play yes = 2, maka
Entrophy Cloudy = 0

$$\text{Maka Gain(Outlook-Temperature)} = 1.00 - (2/4 \cdot 0) - (2/4 \cdot 0)$$

$$= 1.00$$

Mencari nilai hubungan Humidity-Temperatur / Gain (Temp,Hum)

Entropy High terdiri dari 3 instance dengan nilai play no = 2, yes = 1.

$$\begin{aligned} \text{Entropy High} &= -(2/3).\log_2(2/3) - (1/3).\log_2(1/3) \\ &= -(0.667).\log(0.667) - (0.333).\log(0.333) \\ &= -(0.667).(-0.584) - (0.333).(-1.586) \\ &= 0.389 + 0.528 \\ &= 0.917 \end{aligned}$$

$$\begin{aligned} \text{Entropy Normal} &= -(0/1).\log_2(0/1) - (1/1).\log_2(1/1) \\ &= 0 \end{aligned}$$

$$\begin{aligned} \text{Maka Gain(Humadity-Temperature)} &= 1.00 - (3/4)(0.971) - (1/4.0) \\ &= 1.00 - 0.728 \\ &= 0.272 \end{aligned}$$

Mencari nilai hubungan Windy-Temperatur / Gain (Temp,Windy)

Entropy False terdiri dari 3 instance, dengan nilai play no = 1 dan yes = 2.

$$\begin{aligned} \text{Entropy False} &= -(1/3).\log_2(1/3) - (2/3).\log_2(2/3) \\ &= -(0.333).\log(0.333) - (0.667).\log(0.667) \\ &= -(0.333).(-1.586) - (0.667).(-0.584) \\ &= 0.528 + 0.389 \\ &= 0.917 \end{aligned}$$

Entropy True, dengan nilai play No = 1, play yes = 0

Entropy True = 0

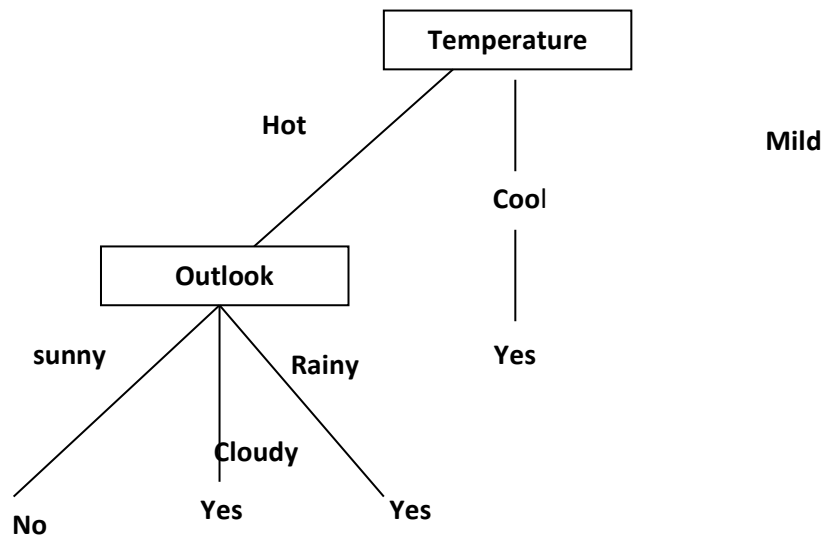
$$\begin{aligned} \text{Maka Gain(Windy-Temperature)} &= 1.00 - (3/4)(0.971) - (1/4.0) \\ &= 1.00 - 0.728 \\ &= 0.272 \end{aligned}$$

Maka nilai hubungan antara Temperature Hot dengan Outlook, Humidity dan windi adalah

$$\text{Maka Gain(Outlook-Temperature=hot)} = 1.00$$

$$\text{Maka Gain(Humadity-Temperature=hot)} = 0.272$$

$$\text{Maka Gain(Windy-Temperature =hot)} = 0.272$$



Mild Temperatur On Play.

No	Outlook	Temperature	Humidity	Windy	Play
4	Rainy	Mild	High	False	Yes
8	Sunny	Mild	High	False	No
10	Rainy	Mild	Normal	False	Yes
11	Sunny	Mild	Normal	True	Yes
12	Cloudy	Mild	High	True	Yes
14	Rainy	Mild	High	True	No

Mencari hubungan **Mild Temperatur On Play**

Mencari nilai hubungan Outlook-Temperatur / Gain (temp,Outlook)

Entropy **Mild**Temperatur = 2/6 No dan 4/6 yes

$$\begin{aligned}
 \text{Nilai Entropy} &= -(2/6) \cdot \log_2(2/6) - (4/6) \cdot \log_2(4/6) \\
 &= -(0.333) \cdot \log_2(0.333) - (0.667) \cdot \log_2(0.667) \\
 &= 0.917
 \end{aligned}$$

Mencari entropy Rainy

Instance = 3, Play No = 1 , Play Yes = 2

$$\begin{aligned}\text{Entropy Rainy} &= -(1/3).\log_2(1/3) - (2/3).\log_2(2/3) \\ &= -(0.333).(0.333) - (0.667).\log_2(0.667) \\ &= 0.917\end{aligned}$$

Mencari entropy Sunny

Instance = 2, Play No = 1 , Play Yes = 1

$$\begin{aligned}\text{Entropy Rainy} &= -(1/2).\log_2(1/2) - (1/2).\log_2(1/2) \\ &= -(0.5).\log_2(0.5) - (0.5).\log_2(0.5) \\ &= -(0.5).(-1.0) - (0.5).\log_2(-1.0) \\ &= 0.5 + 0.5 \\ &= 1.00\end{aligned}$$

Mencari entropy Cloudy

Instance = 1, Play No = 0 , Play Yes = 1

$$\begin{aligned}\text{Entropy Rainy} &= -(0/1).\log_2(0/1) - (1/1).\log_2(1/1) \\ &= 0\end{aligned}$$

$$\begin{aligned}\text{Maka Gain(Outlook|Temperature=mild)} &= 0.917 - [(3/6).(0.917)-(2/6).1]-(0) \\ &= 0.917 - [(0.458)-(0.667)-(0)] \\ &= 0.917 - (-0.209) \\ &= 1.126\end{aligned}$$

Mencari hubungan **Mild Temperatur On Play. (temp-humidity)**

Mencari nilai hubungan Humidity-Temperatur / Gain (temp,humidity)

Mencari entropy High

Instance = 4, Play No = 2 , Play Yes = 2

$$\begin{aligned}\text{Entropy Rainy} &= -(2/4).\log_2(2/4) - (2/4).\log_2(2/4) \\ &= -(0.5).(-1.00) - (0.5).(-1.0) \\ &= 1.00\end{aligned}$$

Mencari entropy Normal

Instance = 2, Play No = 0 , Play Yes = 2

$$\begin{aligned}\text{Entropy Rainy} &= -(0/2).\log_2(0/2) - (2/2).\log_2(2/2) \\ &= 0\end{aligned}$$

$$\begin{aligned}\text{Maka Gain(Humadity|Temperature=mild)} &= 0.917 - [(4/6).(1) - (0)] \\ &= 0.917 - [(0.667)-(0)] \\ &= 0.917 - 0.667 \\ &= 0.25\end{aligned}$$

--

Mencari hubungan **Mild Temperatur On Play. (temp-windy)**

Mencari nilai hubungan Windy-Temperatur / Gain (temp,windy)

Mencari entropy False pada Windy

Instance = 3, Play No = 1 , Play Yes = 2

$$\begin{aligned}\text{Entropy False} &= -(1/3) \cdot \log_2(1/3) - (2/3) \cdot \log_2(2/3) \\ &= -(0.333) \cdot (-1.586) - (0.667) \cdot (-0.584) \\ &= 0.528 + 0.389 \\ &= 0.917\end{aligned}$$

Mencari entropy True pada Windy

Instance = 3, Play No = 1 , Play Yes = 2

$$\begin{aligned}\text{Entropy True} &= -(1/3) \cdot \log_2(1/3) - (2/3) \cdot \log_2(2/3) \\ &= -(0.333) \cdot (-1.586) - (0.667) \cdot (-0.584) \\ &= 0.528 + 0.389 \\ &= 0.917\end{aligned}$$

$$\begin{aligned}\text{Maka Gain(Windy|Temperature=mild)} &= 0.917 - [(3/6) \cdot (0.917) - ((3/6) \cdot (0.917))] \\ &= 0.917 - (0) \\ &= 0.917\end{aligned}$$

Maka nilai hubungan antara Temperature Mild dengan Outlook, Humidity dan windy adalah

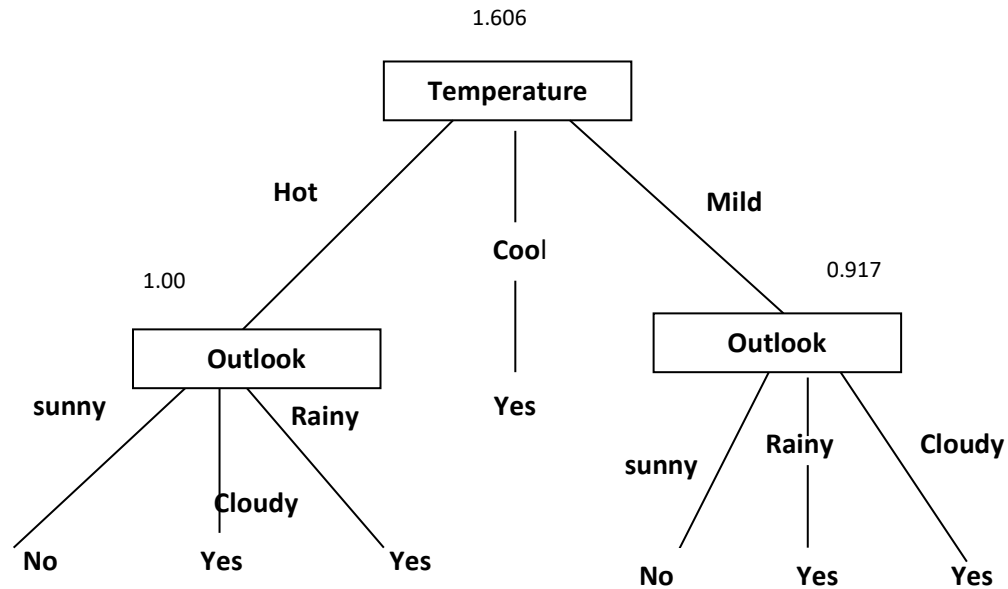
Maka Gain(Outlook-Temperature=mild) = 0.917

Maka Gain(Humadity-Temperature=mild) = 0.25

Maka Gain(Windy-Temperature =mild) = 0.917

Terdapat 2 nilai yang sama, yakni pada **outlook dan Windy**, namun kita bisa cek pada tabel berikut bahwa keputusan untuk bermain pada temperature mild lebih dipengaruhi oleh **Outlook**

No	Outlook	Temperature	Humidity	Windy	Play
4	Rainy	Mild	High	False	Yes
8	Sunny	Mild	High	False	No
10	Rainy	Mild	Normal	False	Yes
11	Sunny	Mild	Normal	True	Yes
12	Cloudy	Mild	High	True	Yes
14	Rainy	Mild	High	True	No



Kesimpulan :

Kelebihan

1. Algoritma ini mampu mentransformasikan dari data mentah menjadi data yang beraturan.
2. Dapat menggunakan atribut nominal dimana kebanyakan tidak dapat dilakukan oleh algoritma machine learning.

Kekurangan

1. Terlalu panjang/lama waktu untuk mencari hubungan satu faktor dengan faktor lainnya.
2. Cenderung kaku dan dapat menimbulkan kesalahan perhitungan.

Metode pembandingan selain menggunakan metode ini dapat dilakukan dengan metode chi-square / CHAID atau Random Forest.

Nama : Shabila Fitri Aulia
Nim : 202420024
Kelas : MTI A23
MK : Advanced Database

TUGAS 4

Dari dataset berikut ini, silahkan buat perhitungan manual dari algoritma decision tree C45 secara lengkap dan gambarkan pohon keputusan yang diperoleh.

No	Outlook	Temperature	Humidity	Windy	Play
1	Sunny	Hot	High	FALSE	No
2	Sunny	Hot	High	TRUE	No
3	Cloudy	Hot	High	FALSE	Yes
4	Rainy	Mild	High	FALSE	Yes
5	Rainy	Cool	Normal	FALSE	Yes
6	Rainy	Cool	Normal	TRUE	Yes
7	Cloudy	Cool	Normal	TRUE	Yes
8	Sunny	Mild	High	FALSE	No
9	Sunny	Cool	Normal	FALSE	Yes
10	Rainy	Mild	Normal	FALSE	Yes
11	Sunny	Mild	Normal	TRUE	Yes
12	Cloudy	Mild	High	TRUE	Yes
13	Cloudy	Hot	Normal	FALSE	Yes
14	Rainy	Mild	High	TRUE	No

Kumpulkan dalam format MS Word. Sertakan penjelasan dan pendapat anda tentang algoritma ini, kelebihan dan kekurangannya berbanding metode lain yang sejenis (tentukan sendiri metodepembandingnya).

Jawaban :

Pada kasus diatas untuk menentukan pengambil keputusan harus perlu dihitung bagaimana peluang – peluang yang muncul dengan menggunakan algoritma decision tree C45, adapun langkahnya sebagai berikut :

1. Menghitung jumlah kasus, jumlah kasus untuk keputusan Yes, jumlah kasus untuk keputusan No, dan Entropy dari semua kasus dan kasus yang dibagi berdasarkan atribut OUTLOOK, TEMPERATUR, HUMIDITY, dan WINDY.
2. Lakukan perhitungan Gain untuk setiap atribut.

- Entropy

Entropy adalah suatu parameter untuk mengukur tingkat keberagaman (heterogenitas) dari kumpulan data. Jika nilai dari entropy semakin besar, maka tingkat keberagaman suatu kumpulan data semakin besar. Rumus untuk menghitung entropy sebagai berikut:

$$Entropy(S) = - \sum_{i=1}^m p_i \log_2(p_i)$$

dimana:

m = jumlah kelas klasifikasi

p_i = jumlah proporsi sampel (peluang) untuk kelas i

Sedangkan rumus untuk entropy pada masing-masing variabel adalah:

$$Entropy_A(S) = \sum_v \frac{|S_v|}{|S|} Entropy(S_v)$$

dimana:

A = variabel

v = nilai yang mungkin untuk variabel A

$|S_v|$ = jumlah sampel untuk nilai v

$|S|$ = jumlah sampel untuk seluruh sampel data

$Entropy(S_v) = Entropy$ untuk sampel yang memiliki nilai v

- Gain

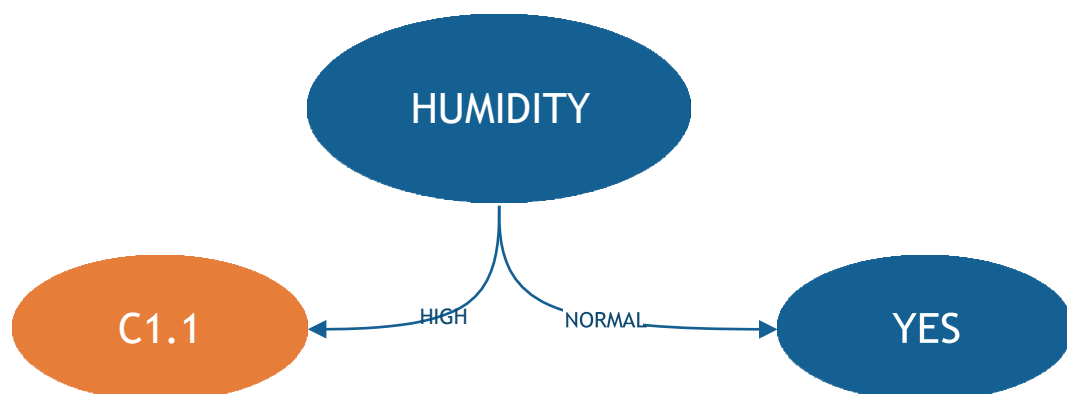
Gain adalah Ukuran efektifitas suatu variabel dalam mengklasifikasikan data. Gain dari suatu variabel merupakan selisih antara nilai entropy total dengan entropy dari variabel tersebut. Gain dapat dirumuskan dengan:

$$Gain(A) = Entropy(S) - Entropy_A(S)$$

Dari data diatas didapatkan data sebagai berikut (detail perhitungan terlampir di excel) :

Node	Deskripsi	Jml Kasus (S)	Count Play "YES"	Count Play "NO"	Entropy	Gain
1	Total	14	10	4	0.86312057	
	OUTLOOK					0.25852104
	Cloudy	4	4			
	Rainy	5	4	1	0.72192809	
	Sunny	5	2	3	0.97095059	
	TEMPERATURE					0.18385093
	Cool	4	4		0	
	Hot	4	2	2	1	
	Mild	6	4	2	0.91829583	
	HUMIDITY					0.3705065
	High	7	3	4	0.98522814	
	Normal	7	7			
	WINDY					0.00597771
	FALSE	8	6	2	0.81127812	
	TRUE	6	4	2	0.91829583	

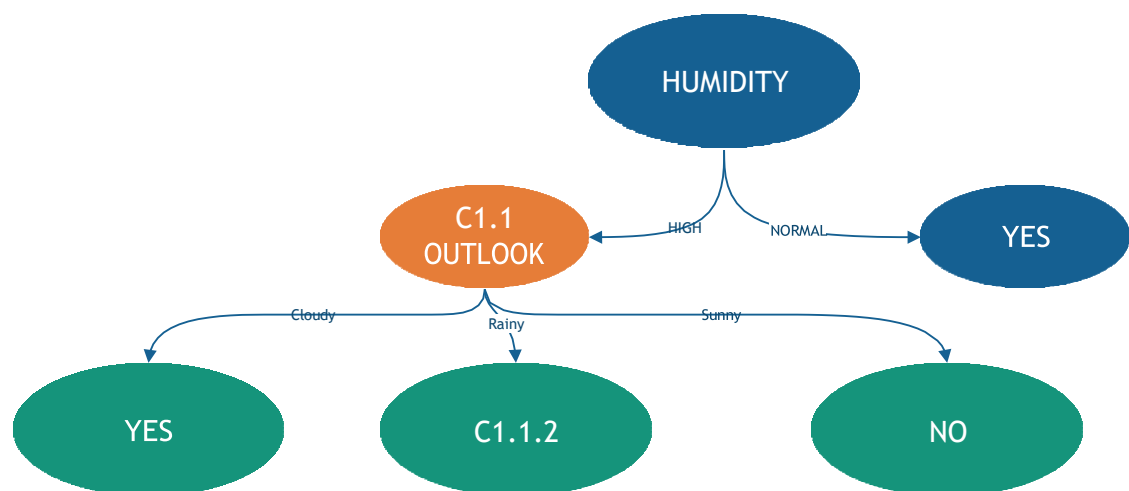
Dapat diketahui dari data diatas bahwa atribut dengan Gain tertinggi adalah HUMADITY ,yaitusebesar0.37.JadiHUMADITY yangmenjadi node akar. Adadua nilai atribut dari HUMADITY , yaitu HIGH dan NORMAL . Dari kedua atribut, nilai atribut NORMAL adalah 1, yaitu keputusan Yes , sehingga tidak perlu perhitungan lebih lanjut, tetapi untuk HIGH masih perlu dilakukan perhitungan lagi. Apabila dibuat tree nya sebagai berikut :



Menghitung jumlah kasus, jumlah kasus untuk keputusan Yes , jumlah kasus untuk keputusan No , dan Entropy dari semua kasus dan kasus yang dibagi berdasarkan atribut OUTLOOK, TEMPERATURE, dan WINDY yang dapat menjadi node akar dari nilai atribut HIGH . Kemudian lakukan perhitungan Gain untuk tiap tiap atribut. Hasil perhitungan sebagai berikut :

Nod e	Deskri psi		Jml Kasus (S)	Count Play "YES"	Count Play "NO"	Entropy	Gain
1	Total		7	3	4	0.9852281 4	
	OUTLOOK						0.699513 85
	Cloud y		2	2		0	
	Rainy		2	1	1	1	
	Sunny		3		3	0	
	TEMPERATURE						0.020244 21
	Cool		0	0	0	0	
	Hot		3	1	2	0.9182958 3	
	Mild		4	2	2	1	
	HUMIDITY						0
	High		7	3	4	0.9852281 4	
	Norm al		7	7			
	WINDY						0.020244 21
	FALSE		4	2	2	1	
	TRUE		3	1	2	0.9182958 3	

Dapat diketahui bahwa atribut dengan Gain tertinggi adalah OUTLOOK , yaitu sebesar 0.67. Jadi OUTLOOK dapat menjadi node cabang dari nilai atribut HIGH . Ada tiga nilai atribut dari OUTLOOK, yaitu CLOUDY, RAINY dan SUNNY. Dari ketiga nilai atribut tersebut, nilai atribut CLOUDY adalah 1, yaitu keputusannya Yes dan nilai atribut SUNNY menjadi keputusan No , sehingga tidak perlu dilakukan perhitungan lebih lanjut, tetapi nilai atribut RAINY masih perlu dilakukan perhitungan lagi.

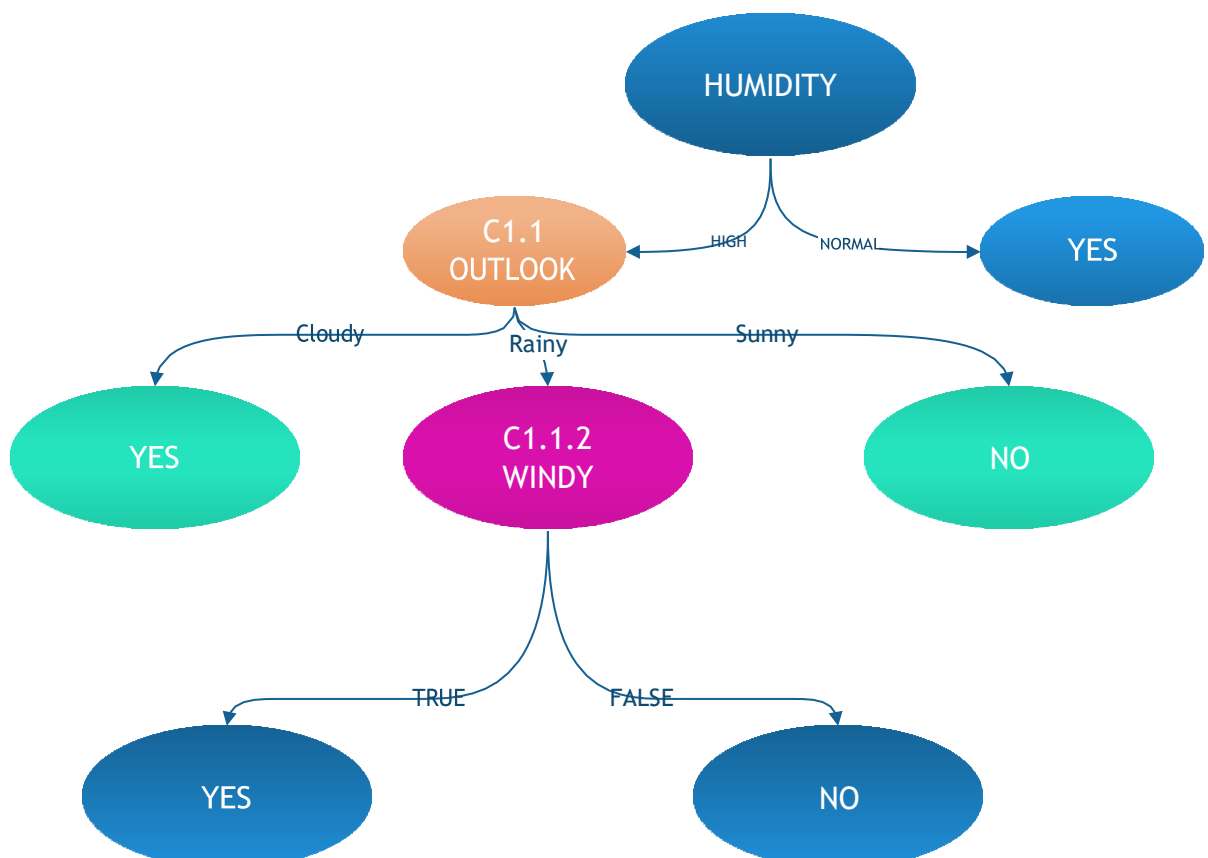


Menghitung jumlah kasus, jumlah kasus untuk keputusan Yes, jumlah kasus untuk keputusan No, dan Entropy dari semua kasus dan kasus yang dibagi berdasarkan atribut TEMPERATURE dan WINDY yang dapat menjadi node cabang dari nilai atribut RAINY.

Kemudian lakukan perhitungan *Gain* untuk tiap-tiap atribut. Hasil perhitungan seperti yang terdapat pada table dibawah ini :

Nod e	Deskri psi	Jml Kasus (S)	Count Play "YES"	Count Play "NO"	Entropy	Gain
1	Total	2	1	1	1	
	OUTLOOK					0
	Cloud y	2	2		0	
	Rainy	2	1	1	1	
	Sunny	3		3	0	
	TEMPERATURE					0
	Cool	0	0	0	0	
	Hot	0	0	0	0	
	Mild	2	1	1	1	
	HUMIDITY					2.44829 85
	High	7	3	4	0.9852281 4	
	Norm al	7	7			
	WINDY					1
	FALSE	1	1	0	0	
	TRUE	1	0	1	0	

Dapat diketahui bahwa atribut dengan *Gain* tertinggi adalah *WINDY*, yaitu sebesar 1. Jadi *WINDY* dapat menjadi *node* cabang dari nilai atribut *RAINY*. Ada dua nilai atribut dari *WINDY*, yaitu *FALSE* dan *TRUE*. Dari kedua nilai atribut tersebut, nilai atribut *FALSE* adalah 1, yaitu keputusannya *Yes* dan nilai atribut *TRUE* menjadi keputusan *No*, sehingga tidak perlu dilakukan perhitungan lebih lanjut untuk nilai atribut ini. Dapat digambarkan untuk pohon



keputusannya sebagai berikut:

Merupakan gambar akhir dari proses *Data Mining* menggunakan algoritma C4.5 pada contoh kasus keputusan bermain tenis. Berdasarkan gambar di atas dapat ditarik kesimpulan bahwa jika *humidity* normal maka keputusannya adalah *yes*, jika *humidity* adalah *high*, *outlook* adalah *cloudy* maka keputusannya adalah bermain (*yes*), selanjutnya jika *humidity* adalah *high*, *outlook* adalah *rainy* dan *windy* adalah *true*, maka keputusannya adalah bermain (*yes*). Dan jika *humidity* adalah *high*, *outlook* adalah *rainy* dan *windy* adalah *false*, maka keputusannya adalah tidak bermain (*no*). Terakhir jika *humidity* adalah *high*, *outlook* adalah *no*, maka keputusannya adalah tidak bermain (*no*).

Algoritma C4.5

Algoritma yang dapat digunakan untuk memprediksi atau mengklasifikasi suatu kejadian dengan pembentukan pohon keputusan antara lain algoritma C4.5, yang merupakan salah satu algoritma induksi pohon keputusan yang dikembangkan oleh J. Ross Quinlan (Yamin, *et al*, 2014).

Algoritma C4.5 merupakan kelompok algoritma *decision tree*. Algoritma ini mempunyai input berupa *training samples* dan *samples*. *Training samples* merupakan data contoh yang digunakan untuk membangun sebuah *tree* yang telah diuji kebenarannya. Sedangkan *samples* merupakan *field-field* data yang digunakan sebagai parameter dalam klasifikasi data (Sujana, 2010).

Secara umum alur proses algoritma C4.5 untuk membangun pohon keputusan dalam *data mining* adalah sebagai berikut (Mahmud, *et al*, 2014):

- a. Pilih atribut sebagai akar
- b. Buat cabang untuk tiap-tiap nilai
- c. Bagi kasus dalam cabang.
- d. Ulangi proses untuk setiap cabang sampai semua kasus pada cabang memiliki kelas yang sama.

Nama : Siti Ratu Delima

Nim : 202420025

Kelas : MTI24

TUGAS 04

Dari dataset berikut ini, silahkan buat perhitungan manual dari algoritma decision tree C45 secara lengkap dan gambarkan pohon keputusan yang diperoleh.

No	Outlook	Temperature	Humidity	Windy	Play
1	Sunny	Hot	High	FALSE	No
2	Sunny	Hot	High	TRUE	No
3	Cloudy	Hot	High	FALSE	Yes
4	Rainy	Mild	High	FALSE	Yes
5	Rainy	Cool	Normal	FALSE	Yes
6	Rainy	Cool	Normal	TRUE	Yes
7	Cloudy	Cool	Normal	TRUE	Yes
8	Sunny	Mild	High	FALSE	No
9	Sunny	Cool	Normal	FALSE	Yes
10	Rainy	Mild	Normal	FALSE	Yes
11	Sunny	Mild	Normal	TRUE	Yes
12	Cloudy	Mild	High	TRUE	Yes
13	Cloudy	Hot	Normal	FALSE	Yes
14	Rainy	Mild	High	TRUE	No

Kumpulkan dalam format MS Word. Sertakan penjelasan dan pendapat anda tentang algoritma ini, kelebihan dan kekurangannya berbanding metode lain yang sejenis (tentukan sendiri metode pembandingnya).

Jawabannya :

Pada kasus diatas untuk menentukan pengambil keputusan harus perlu dihitung bagaimana peluang - peluang yang muncul dengan menggunakan algoritmadecisiontree C45, adapun langkahnya sebagai berikut :

1. Menghitung jumlah kasus, jumlah kasus untuk keputusan Yes, jumlah kasus untuk keputusan No, dan Entropy dari semua kasus dan kasus yang dibagi berdasarkan atribut OUTLOOK, TEMPERATUR, HUMIDITY, dan WINDY.

2. Lakukan perhitungan Gain untuk setiap atribut.

a. Entropy

Entropy adalah suatu parameter untuk mengukur tingkat keberagaman (heterogenitas) dari kumpulan data. Jika nilai dari entropy semakin besar, maka tingkat keberagaman suatu kumpulan data semakin besar. Rumus untuk menghitung entropy sebagai berikut:

$$Entropy(S) = - \sum_{i=1}^m p_i \log_2(p_i)$$

dimana:

m = jumlah kelas klasifikasi

p_i = jumlah proporsi sampel (peluang) untuk kelas i

Sedangkan rumus untuk entropy pada masing-masing variabel adalah:

$$Entropy_A(S) = \sum_v \frac{|S_v|}{|S|} Entropy(S_v)$$

dimana:

A = variabel

v = nilai yang mungkin untuk variabel A

$|S_v|$ = jumlah sampel untuk nilai v

$|S|$ = jumlah sampel untuk seluruh sampel data

$Entropy(S_v) = Entropy$ untuk sampel yang memiliki nilai v

b. Gain

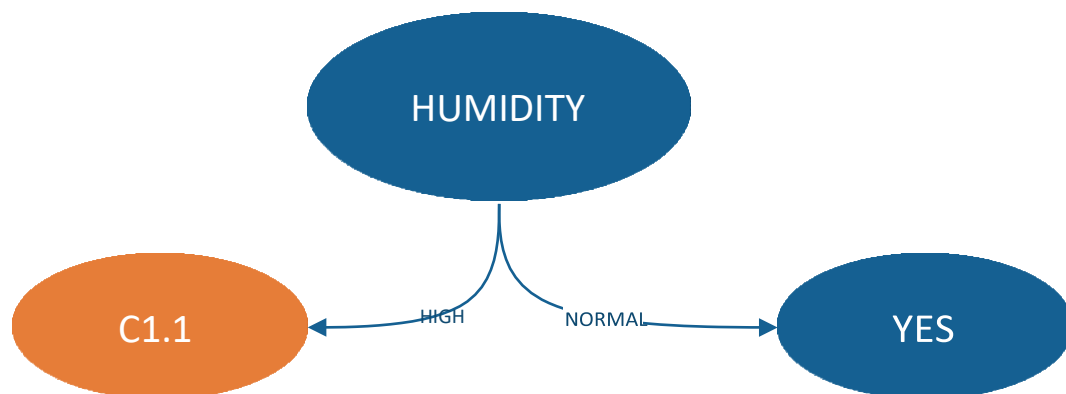
Gain adalah Ukuran efektifitas suatu variabel dalam mengklasifikasikan data. Gain dari suatu variabel merupakan selisih antara nilai entropy total dengan entropy dari variabel tersebut. Gain dapat dirumuskan dengan:

$$Gain(A) = Entropy(S) - Entropy_A(S)$$

Dari data diatas didapatkan data sebagai berikut (detail perhitungan terlampir di excel):

Node	Deskripsi	Jml Kasus (S)	Count Play "YES"	Count Play "NO"	Entropy	Gain
1	Total	14	10	4	0,86312057	
	OUTLOOK					0,25852104
	Cloudy	4	4			
	Rainy	5	4	1	0,72192809	
	Sunny	5	2	3	0,97095059	
	TEMPERATURE					0,18385093
	Cool	4	4		0	
	Hot	4	2	2	1	
	Mild	6	4	2	0,91829583	
	HUMIDITY					0,3705065
	High	7	3	4	0,98522814	
	Normal	7	7			
	WINDY					0,00597771
	FALSE	8	6	2	0,81127812	
	TRUE	6	4	2	0,91829583	

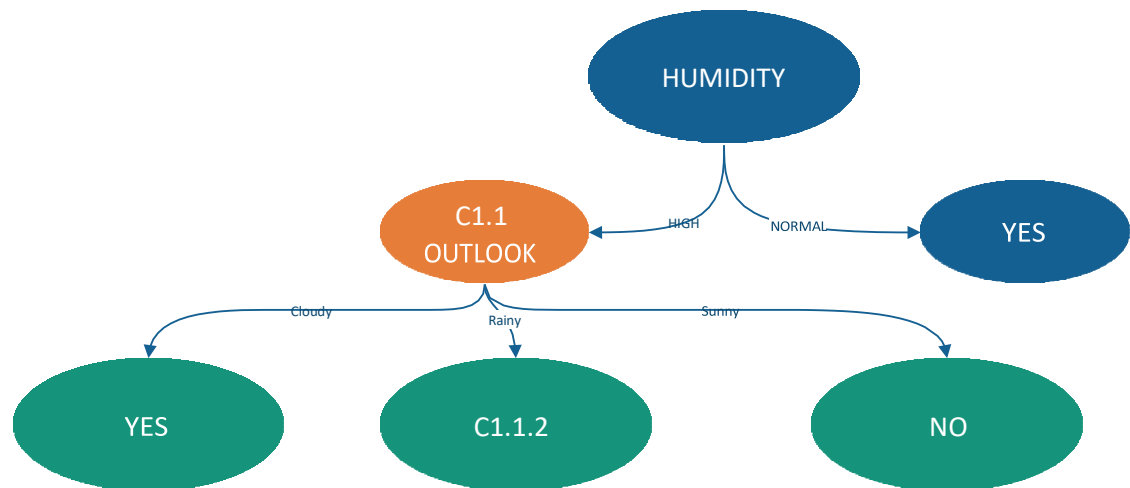
Dapat diketahui dari data diatas bahwa atribut dengan Gain tertinggi adalah HUMADITY , yaitu sebesar 0.37. Jadi HUMADITY yang menjadi node akar. Ada dua nilai atribut dari HUMADITY , yaitu HIGH dan NORMAL . Dari kedua atribut, nilai atribut NORMAL adalah 1, yaitu keputusan Yes , sehingga tidak perlu perhitungan lebih lanjut, tetapi untuk HIGH masih perlu dilakukan perhitungan lagi. Apabila dibuat tree nya sebagai berikut :



Menghitung jumlah kasus, jumlah kasus untuk keputusan Yes , jumlah kasus untuk keputusan No , dan Entropy dari semua kasus dan kasus yang dibagi berdasarkan atribut OUTLOOK, TEMPERATURE, dan WINDY yang dapat menjadi node akar dari nilai atribut HIGH . Kemudian lakukan perhitungan Gain untuk tiap atribut. Hasil perhitungan sebagai berikut :

Node	Deskripsi	Jml Kasus (S)	Count Play "YES"	Count Play "NO"	Entropy	Gain
1	Total	7	3	4	0,98522814	
	OUTLOOK					0,69951385
	Cloudy	2	2		0	
	Rainy	2	1	1	1	
	Sunny	3		3	0	
	TEMPERATURE					0,02024421
	Cool	0	0	0	0	
	Hot	3	1	2	0,91829583	
	Mild	4	2	2	1	
	HUMIDITY					0
	High	7	3	4	0,98522814	
	Normal	7	7			
	WINDY					0,02024421
	FALSE	4	2	2	1	
	TRUE	3	1	2	0,91829583	

Dapat diketahui bahwa atribut dengan Gain tertinggi adalah OUTLOOK , yaitu sebesar 0.67. Jadi OUTLOOK dapat menjadi node cabang dari nilai atribut HIGH . Ada tiga nilai atribut dari OUTLOOK, yaitu CLOUDY, RAINY dan SUNNY. Dari ketiga nilai atribut tersebut, nilai atribut CLOUDY adalah 1, yaitu keputusannya Yes dan nilai atribut SUNNY menjadi keputusan No , sehingga tidak perlu dilakukan perhitungan lebih lanjut, tetapi nilai atribut RAINY masih perlu dilakukan

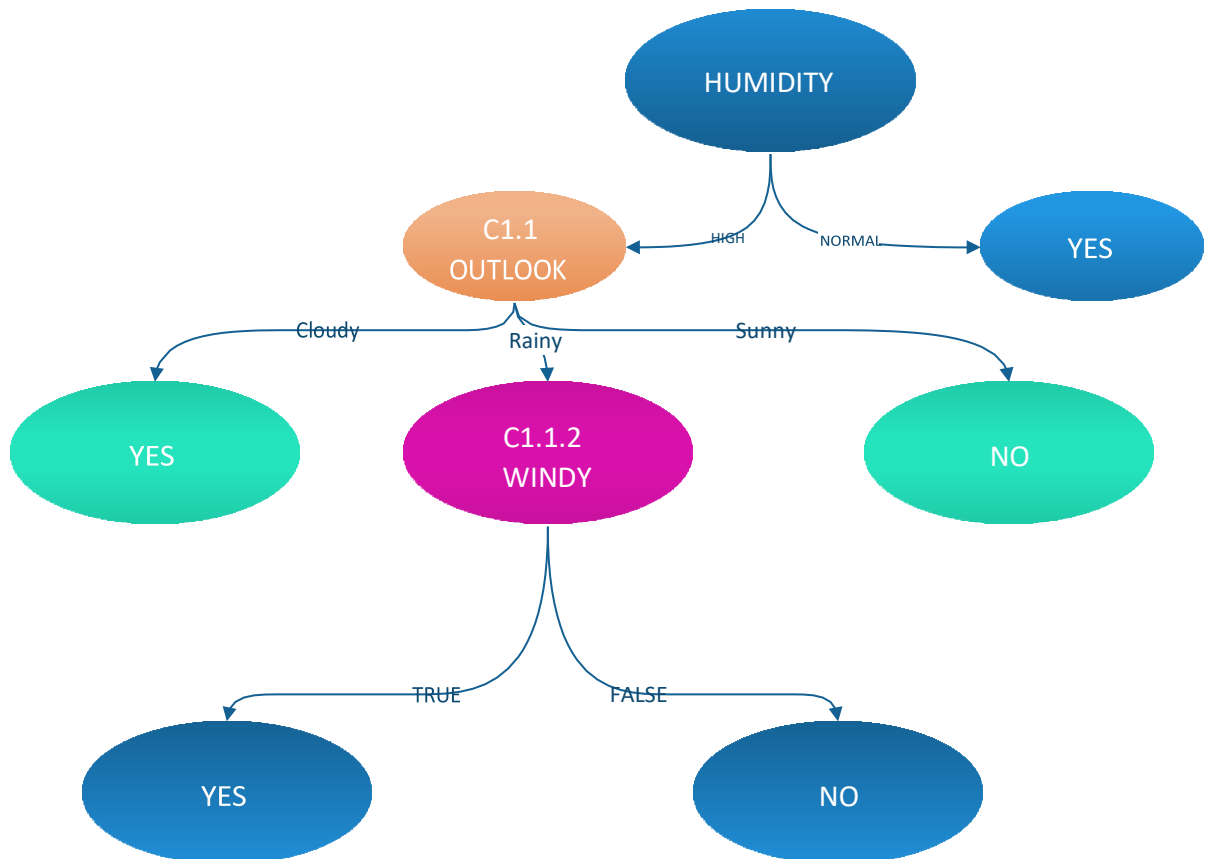


perhitungan lagi.

Menghitung jumlah kasus, jumlah kasus untuk keputusan Yes, jumlah kasus untuk keputusan No, dan Entropy dari semua kasus dan kasus yang dibagi berdasarkan atribut TEMPERATURE dan WINDY yang dapat menjadi node cabang dari nilai atribut RAINY. Kemudian lakukan perhitungan Gain untuk tiap-tiap atribut. Hasil perhitungan seperti yang terdapat pada table dibawah ini :

Node	Deskripsi	Jml Kasus (S)	Count Play "YES"	Count Play "NO"	Entropy	Gain
1	Total	2	1	1	1	
	OUTLOOK					0
	Cloudy	2	2	0	0	
	Rainy	2	1	1	1	
	Sunny	3	3	0	0	
	TEMPERATURE					0
	Cool	0	0	0	0	
	Hot	0	0	0	0	
	Mild	2	1	1	1	
	HUMIDITY					-2,4482985
	High	7	3	4	0,98522814	
	Normal	7	7	0	0	
	WINDY					1
	FALSE	1	1	0	0	
	TRUE	1	0	1	0	

Dapat diketahui bahwa atribut dengan *Gain* tertinggi adalah *WINDY*, yaitu sebesar 1. Jadi *WINDY* dapat menjadi *node* cabang dari nilai atribut *RAINY*. Ada dua nilai atribut dari *WINDY*, yaitu *FALSE* dan *TRUE*. Dari kedua nilai atribut tersebut, nilai atribut *FALSE* adalah 1, yaitu keputusannya *Yes* dan nilai atribut *TRUE* menjadi keputusan *No*, sehingga tidak perlu dilakukan perhitungan lebih lanjut untuk nilai atribut ini. Dapat digambarkan untuk pohon keputusannya sebagai berikut:



Merupakan gambar akhir dari proses *Data Mining* menggunakan algoritma C4.5 pada contoh kasus keputusan bermain tenis. Berdasarkan gambar di atas dapat ditarik kesimpulan bahwa jika *humidity* normal maka keputusannya adalah *yes*, jika *humidity* adalah *high*, *outlook* adalah *cloudy* maka keputusannya adalah bermain (*yes*), selanjutnya jika *humidity* adalah *high*, *outlook* adalah *rainy* dan *windy* adalah *true*, maka keputusannya adalah bermain (*yes*). Dan jika *humidity* adalah *high*, *outlook* adalah *rainy* dan *windy* adalah *false*, maka keputusannya adalah tidak bermain (*no*). Terakhir jika *humidity* adalah *high*, *outlook* adalah *no*, maka keputusannya adalah tidak bermain (*no*).

Algoritma C4.5

Algoritma yang dapat digunakan untuk memprediksi atau mengklasifikasi suatu kejadian dengan pembentukan pohon keputusan antara lain algoritma C4.5, yang merupakan salah satu algoritma induksi pohon keputusan yang dikembangkan oleh J. Ross Quinlan (Yamin, *et al*, 2014).

Algoritma C4.5 merupakan kelompok algoritma *decision tree*. Algoritma ini mempunyai input berupa *training samples* dan *samples*. *Training samples* merupakan data contoh yang digunakan untuk membangun sebuah *tree* yang telah diuji kebenarannya. Sedangkan *samples* merupakan *field-field* data yang digunakan sebagai parameter dalam klasifikasi data (Sujana, 2010).

Secara umum alur proses algoritma C4.5 untuk membangun pohon keputusan dalam *data mining* adalah sebagai berikut (Mahmud, *et al*, 2014):

- c. Pilih atribut sebagai akar
- d. Buat cabang untuk tiap-tiap nilai
- e. Bagi kasus dalam cabang.
- f. Ulangi proses untuk setiap cabang sampai semua kasus pada cabang memiliki kelas yang sama.

Tugas 05

Nama : Surta Wijaya

NIM : 202420014

Dari dataset berikut ini, silahkan buat perhitungan manual dari algoritma decision tree C45 secara lengkap dan gambarkan pohon keputusan yang diperoleh.

No	Outlook	Temperature	Humidity	Windy	Play
1	Sunny	Hot	High	FALSE	No
2	Sunny	Hot	High	TRUE	No
3	Cloudy	Hot	High	FALSE	Yes
4	Rainy	Mild	High	FALSE	Yes
5	Rainy	Cool	Normal	FALSE	Yes
6	Rainy	Cool	Normal	TRUE	Yes
7	Cloudy	Cool	Normal	TRUE	Yes
8	Sunny	Mild	High	FALSE	No
9	Sunny	Cool	Normal	FALSE	Yes
10	Rainy	Mild	Normal	FALSE	Yes
11	Sunny	Mild	Normal	TRUE	Yes
12	Cloudy	Mild	High	TRUE	Yes
13	Cloudy	Hot	Normal	FALSE	Yes
14	Rainy	Mild	High	TRUE	No

Kumpulkan dalam format MS Word. Sertakan penjelasan dan pendapat anda tentang algoritma ini, kelebihan dan kekurangannya berbanding metode lain yang sejenis (tentukan sendiri metode pembandingnya).

Jawaban :

Menghitung Nilai Entropy

$$\text{Entropy (S)} = \sum_{i=1}^n - p_i * \log_2 p_i$$

Keterangan:

S : himpunan kasus

A : fitur

n : jumlah partisi S

pi : proporsi dari Si terhadap S

Perincian algoritma (langkah 1)

- Menghitung jumlah kasus seluruhnya, jumlah berkeputusan “Yes” maupun “No”.
- Menghitung Entropy dari semua kasus yg terbagi berdasarkan atribut “Outlook”, “Temperature”, “Humidity”, “Windy”.

Node			JUMLAH KASUS	TIDAK (S1)	YA (S2)	ENTROPY
1	TOTAL		14	4	10	$= (-\frac{4}{14} * \log_2 \frac{4}{14}) + (-\frac{10}{14} * \log_2 \frac{10}{14})$ $= 0,8163120569$
	OUTLOOK					
		CLOUDY	4	0	4	$= (-\frac{0}{4} * \log_2 \frac{0}{4}) + (-\frac{4}{4} * \log_2 \frac{4}{4})$ $= 0$
		RAINY	5	1	4	$= (-\frac{1}{5} * \log_2 \frac{1}{5}) + (-\frac{4}{5} * \log_2 \frac{4}{5})$ $= 0,721928095$
		SUNNY	5	3	2	$= (-\frac{3}{5} * \log_2 \frac{3}{5}) + (-\frac{2}{5} * \log_2 \frac{2}{5})$ $= 0,970950594$
	TEMPERATURE					
		COOL	4	0	4	$= (-\frac{0}{4} * \log_2 \frac{0}{4}) + (-\frac{4}{4} * \log_2 \frac{4}{4})$ $= 0$
		HOT	4	2	2	$= (-\frac{2}{4} * \log_2 \frac{2}{4}) + (-\frac{2}{4} * \log_2 \frac{2}{4})$ $= 1$
		MILD	6	2	4	$= (-\frac{2}{6} * \log_2 \frac{2}{6}) + (-\frac{4}{6} * \log_2 \frac{4}{6})$ $= 0,918295834$
	HUMIDITY					
		HIGH	7	4	3	$= (-\frac{4}{7} * \log_2 \frac{4}{7}) + (-\frac{3}{7} * \log_2 \frac{3}{7})$ $= 0,985228136$
		NORMAL	7	0	7	$= (-\frac{0}{7} * \log_2 \frac{0}{7}) + (-\frac{7}{7} * \log_2 \frac{7}{7})$ $= 0$
	WINDY					
		FALSE	8	2	6	$= (-\frac{2}{8} * \log_2 \frac{2}{8}) + (-\frac{6}{8} * \log_2 \frac{6}{8})$ $= 0,811278124$
		TRUE	6	2	4	$= (-\frac{2}{6} * \log_2 \frac{2}{6}) + (-\frac{4}{6} * \log_2 \frac{4}{6})$ $= 0,918295834$

Menghitung gain

$$\text{Gain}(S,A) = \text{Entropy}(S) - \sum_{i=1}^n \frac{|S_i|}{|S|} * \text{Entropy}(S_i)$$

Gain (Total, Outlook)

$$= \text{Entropy}(\text{Total}) - \sum_{i=1}^n \frac{|\text{Outlook}_i|}{|\text{Total}|} * \text{Entropy}(\text{Outlook}_i)$$

$$= 0,863120569 - ((4/14*0) + (5/14*0,72) + (5/14*0,97))$$

$$= 0,258521037$$

Node			JUMLAH KASUS	TIDAK (S1)	YA (S2)	ENTROPY	GAIN
1	TOTAL		14	4	10	0,8163120569	
	OUTLOOK						= 0,8163120569 – ((4/14*0)+(5/14*0,72) +(5/14*0,97)) = 0,258521037
		CLOUDY	4	0	4	0	
		RAINY	5	1	4	0,721928095	
		SUNNY	5	3	2	0,970950594	
	TEMPERATURE						= 0,8163120569 – ((4/14*0)+(4/14*1) +(6/14*0,92)) = 0,183850925
		COOL	4	0	4	0	
		HOT	4	2	2	1	
		MILD	6	2	4	0,918295834	
	HUMIDITY						= 0,8163120569 – ((7/14*0,99)+(7/14*0)) = 0,370506501
		HIGH	7	4	3	0,985228136	
		NORMAL	7	0	7	0	
	WINDY						= 0,8163120569 – ((8/14*0,81)+(6/14*0,92)) = 0,005977711
		FALSE	8	2	6	0,811278124	
		TRUE	6	2	4	0,918295834	

Seperti yang terlihat pada tabel, diperoleh bahwa atribut dengan gain tertinggi adalah Humidity = 0,37, maka Humidity menjadi node akar. Humidity memiliki dua nilai yaitu High dan Normal . Humidity Normal sudah mengklasifikasi kasus menjadi 1 yaitu keputusan yes , untuk humidity High masih perlu dilakukan perhitungan lagi karena masih terdapat yes dan no .

Pohon Keputusan Node 1

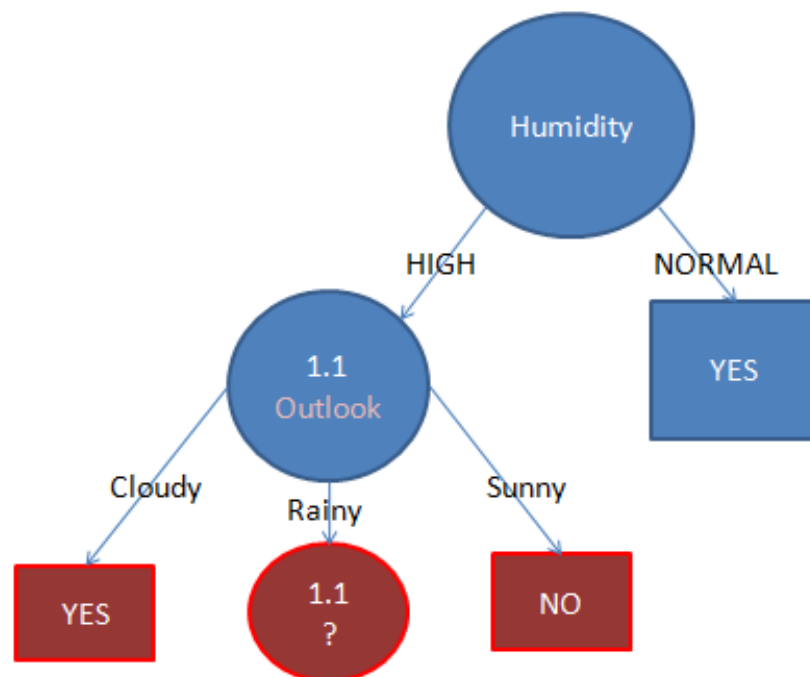
Perincian Algoritma (langkah 2):

Pilih node akar High dan menghitung jumlah kasus ,Entropy Gain

Node			Jumlah Kasus (S)	Tidak (S1)	Ya (S2)	Entropy	Gain
1.1	HUMIDITY HIGH		7	4	3	$= (-\frac{4}{7} * \log_2 \frac{4}{7}) + (-\frac{3}{7} * \log_2 \frac{3}{7})$ $= 0,985228136$	
	OUTLOOK						$= 0,985228136 - ((2/7*0)+(2/7*1) + (3/7*0))$ $= 0,69951385$
		CLOUDY	2	0	2	$= (-\frac{0}{2} * \log_2 \frac{0}{2}) + (-\frac{2}{2} * \log_2 \frac{2}{2})$ $= 0$	
		RAINY	2	1	1	$= (-\frac{1}{2} * \log_2 \frac{1}{2}) + (-\frac{1}{2} * \log_2 \frac{1}{2})$ $= 1$	
		SUNNY	3	3	0	$= (-\frac{3}{3} * \log_2 \frac{3}{3}) + (-\frac{0}{3} * \log_2 \frac{0}{3})$ $= 0$	
	TEMPERATURE						$= 0,985228136 - ((0)+(3/7*0,92) + (4/7*1))$ $= 0,020244207$
		COOL	0	0	0	0	
		HOT	3	2	1	$= (-\frac{2}{3} * \log_2 \frac{2}{3}) + (-\frac{1}{3} * \log_2 \frac{1}{3})$ $= 0,918296$	
		MILD	4	2	2	$= (-\frac{2}{4} * \log_2 \frac{2}{4}) + (-\frac{2}{4} * \log_2 \frac{2}{4})$ $= 1$	
	WINDY						$= 0,985228136 - ((4/7*1)+(3/7*0,92))$ $= 0,020244207$
		FALSE	4	2	2	$= (-\frac{2}{4} * \log_2 \frac{2}{4}) + (-\frac{2}{4} * \log_2 \frac{2}{4})$ $= 1$	
		TRUE	3	2	1	$= (-\frac{2}{3} * \log_2 \frac{2}{3}) + (-\frac{1}{3} * \log_2 \frac{1}{3})$ $= 0,918296$	

Dari hasil perhitungan , didapatkan gain tertinggi yaitu outlook = 0,69951, maka Outlook menjadi mode cabang dari atribut humidity yang bernilai high, berdasarkan atribut outlook terdapat 3 nilai yaitu : Cloudy , Rainy , Sunny . Karena Cloudy pasti bernilai Yes, dan sunny pasti bernilai No, maka tidak perlu perhitungan lagi. Sedangkan rainy bernilai yes dan no, maka perlu dilakukan perhitungan lagi.

Pohon keputusan Node 1.1



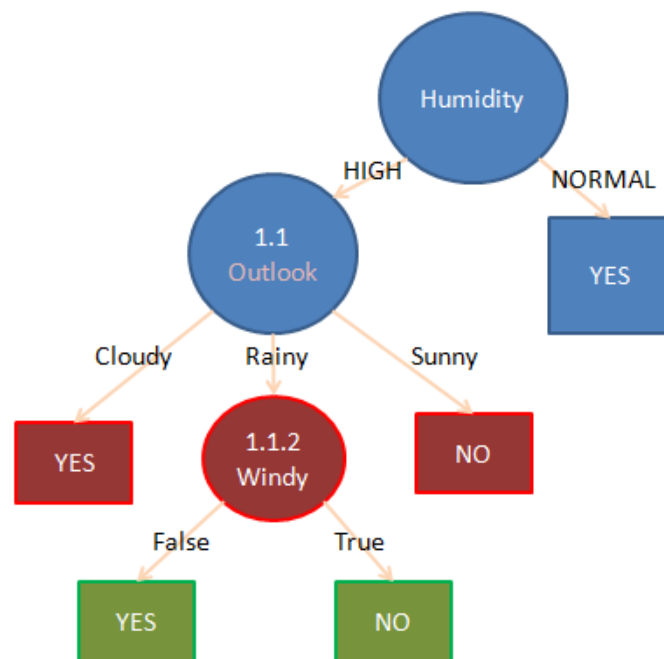
Perincian algoritma (langkah 3) :

Pilih node akar rainy dan hitung jumlah kasus, entropy dan gain. Atribut : Temperature dan Windy

Node			JUMLAH KASUS (S)	TIDAK (S1)	YA (S2)	ENTROPY	GAIN
1.1.2	HUMIDITY HIGH DAN OUTLOOK RAINY		2	1	1	1	
	TEMPERATURE						0
		COOL	0	0	0	0	
		HOT	0	0	0	0	
		MILD	2	1	1	1	
	WINDY						1
		FALSE	1	0	1	0	
		TRUE	1	1	0	0	

Dari hasil perhitungan, didapatkan gain tertinggi yaitu Windy = 1 , maka windy menjadi node cabang dari atribut humidity yang bernilai high dan outlook yang bernilai rainy. Berdasarkan atribut windy, terdapat dua nilai yaitu true dan false. Karena true sudah terklasifikasi pasti bernilai No, dan False pasti bernilai yes, maka tidak perlu dilakukan perhitungan lagi.

Pohon keputusan node 1.1.2



Berdasarkan node 1.1.2, maka semua kasus sudah masuk kelas sehingga pohon keputusan diatas merupakan pohon keputusan terakhir yang terbentuk.

Kelebihan Algoritma C4.5:

- Kecepatan membaca dan membentuk model
- Daerah pengambilan menjadi lebih simpel dan spesifik
- Eliminasi perhitungan – perhitungan yang tidak diperlukan
- Fleksibel sehingga akan meningkatkan kualitas keputusan yang dihasilkan
- Dapat mengolah data numerik (kontinyu) dan diskret
- Interpertabilitas yaitu model yang dihasilkan mudah dipahami

Kekurangan :

- Terjadi overlapping
- Pengakumulasian jumlah kesalahan dari setiap tingkat dalam sebuah pohon keputusan yang besar
- Kesulitan dalam mendesain pohon keputusan yang optimal
- Hasil kualitas keputusan sangat tergantung pada bagaimana pohon tersebut didesain
- Sampel yang harus disimpan secara keseluruhan dan pada waktu yang bersamaan di memori (kebutuhan memori tinggi)

Nama : Trada Ayang Pratiwi
NIM : 202420020
Tugas : 04 Advanced Database

Tugas 04

No	Outlook	Temperature	Humidity	Windy	Play
1	Sunny	Hot	High	False	No
2	Sunny	Hot	High	True	No
3	Cloudy	Hot	High	False	Yes
4	Rainy	Mild	High	False	Yes
5	Rainy	Cool	Normal	False	Yes
6	Rainy	Cool	Normal	True	Yes
7	Cloudy	Cool	Normal	True	Yes
8	Sunny	Mild	High	False	No
9	Sunny	Cool	Normal	False	Yes
10	Rainy	Mild	Normal	False	Yes
11	Sunny	Mild	Normal	True	Yes
12	Cloudy	Mild	High	True	Yes
13	Cloudy	Hot	Normal	False	Yes
14	Rainy	Mild	High	True	No

Rumus Penyelesaian

$$\text{Entropy}(S) = \sum -p(I) \cdot \log_2 p(I)$$

$$\text{Gain}(S, A) = \text{Entropy}(S) - \sum [p(S|A) \cdot \text{Entropy}(S|A)]$$

Entropy

Kita hitung dulu Nilai Entropy. Kolom Play terdiri dari 14 instance dan terdiri dari 2 label yakni Yes dan No. Play Yes = 10, _Play No= 4

$$\begin{aligned}\text{Entropy(Decision)} &= -p(\text{Yes}) \cdot \log_2 p(\text{Yes}) - p(\text{No}) \cdot \log_2 p(\text{No}) \\ \text{Entropy(Decision)} &= - (10/14) \cdot \log_2(10/14) - (4/14) \cdot \log_2(4/14) \\ &= -(0.714) \cdot \log_2(0.714) - (0.4) \cdot \log_2(0.4) \\ &= 0.347 + 0.528 \\ &= \mathbf{0.875}\end{aligned}$$

Menghitung faktor dominan untuk memutuskan bermain.

Faktor Windy dan Play

False Factor windy to Play

No	Outlook	Temperature	Humidity	Windy	Play
1	Sunny	Hot	High	False	No
3	Cloudy	Hot	High	False	Yes
4	Rainy	Mild	High	False	Yes
5	Rainy	Cool	Normal	False	Yes
8	Sunny	Mild	High	False	No
9	Sunny	Cool	Normal	False	Yes
10	Rainy	Mild	Normal	False	Yes
13	Cloudy	Hot	Normal	False	Yes

Tabel faktor false windy terdiri dari **8 instance**. False No = 2, False Yes = 6

$$\begin{aligned}
 \text{Entropy play|windy=False} &= -(2/8) \cdot \log_2(2/8) - (6/8) \cdot \log_2(6/8) \\
 &= -(0.25) \cdot \log_2(0.25) - (0.75) \cdot \log_2(0.75) \\
 &= -(0.25) \cdot (-2) - (0.75) \cdot (-0.415) \\
 &= 0.50 + 0.311 \\
 &= \mathbf{0.811}
 \end{aligned}$$

Menghitung True Factor windy to Play

No	Outlook	Temperature	Humidity	Windy	Play
2	Sunny	Hot	High	True	No
6	Rainy	Cool	Normal	True	Yes
7	Cloudy	Cool	Normal	True	Yes
11	Sunny	Mild	Normal	True	Yes
12	Cloudy	Mild	High	True	Yes
14	Rainy	Mild	High	True	No

Tabel faktor True windy terdiri dari **6 instance**. True No = 2, True Yes = 4

$$\begin{aligned}
 \text{Entropy play|windy=True} &= -(2/6) \cdot \log_2(2/6) - (4/6) \cdot \log_2(4/6) \\
 &= -(0.333) \cdot \log_2(0.333) - (0.667) \cdot \log_2(0.667) \\
 &= -(0.333) \cdot (-1.586) - (0.667) \cdot (-0.586) \\
 &= 0.528 + 0.390 \\
 &= \mathbf{0.918}
 \end{aligned}$$

Menghitung Gain-Play|Windy

$$\begin{aligned}
 \text{Gain(Play, Windy)} &= \text{Entropy(Play)} - [p(\text{Play|Windy=False}) \cdot \\
 &\text{Entropy(Play|Windy=False)}] - [p(\text{Play|Windy=True}) \cdot \text{Entropy(Play|Windy=True)}] \\
 &= 0.875 - [(8/14) \cdot 0.811] - [(6/14) \cdot 0.918] \\
 &= 0.804
 \end{aligned}$$

Maka Gain(Play, Windy) = **0.804**

Selanjutnya kolom 1

Menghitung Gain(Play, Outlook)

Faktor Outlook Sunny

No	Outlook	Temperature	Humidity	Windy	Play
1	Sunny	Hot	High	False	No
2	Sunny	Hot	High	True	No
8	Sunny	Mild	High	False	No
9	Sunny	Cool	Normal	False	Yes
11	Sunny	Mild	Normal	True	Yes

Tabel faktor Sunny Play terdiri dari **5 instance**. Play No = 3, Play Yes = 2

$$\begin{aligned}
 \text{Entropy play|Outlook=Sunny} &= -(3/5) \cdot \log_2(3/5) - (2/5) \cdot \log_2(2/5) \\
 &= -(0.6) \cdot \log_2(0.6) - (0.4) \cdot \log_2(0.4) \\
 &= -(0.6) \cdot (-0.737) - (0.4) \cdot (-1.322) \\
 &= 0.442 + 0.5288 \\
 &= \mathbf{0.971}
 \end{aligned}$$

Faktor Cloudy Outlook

No	Outlook	Temperature	Humidity	Windy	Play
3	Cloudy	Hot	High	False	Yes
7	Cloudy	Cool	Normal	True	Yes
12	Cloudy	Mild	High	True	Yes
13	Cloudy	Hot	Normal	False	Yes

Tabel faktor Cloudy- Play terdiri dari **4 instance**. Play No = 0, Play Yes = 4

$$\begin{aligned}
 \text{Entropy play|Outlook=Cloudy} &= -(0/4) \cdot \log_2(0/4) - (4/4) \cdot \log_2(4/4) \\
 &= \mathbf{0}
 \end{aligned}$$

Faktor Rainy Outlook

No	Outlook	Temperature	Humidity	Windy	Play
4	Rainy	Mild	High	False	Yes
5	Rainy	Cool	Normal	False	Yes
6	Rainy	Cool	Normal	True	Yes
10	Rainy	Mild	Normal	False	Yes
14	Rainy	Mild	High	True	No

Tabel faktor Rainy Play terdiri dari **5 instance**. Play No = 1, Play Yes = 4

$$\begin{aligned}
 \text{Entropy play|Outlook=Sunny} &= -(1/5) \cdot \log_2(1/5) - (4/5) \cdot \log_2(4/5) \\
 &= -(0.2) \cdot \log_2(0.2) - (0.8) \cdot \log_2(0.8) \\
 &= -(0.2) \cdot (-2.321) - (0.8) \cdot (-0.322) \\
 &= 0.464 + 0.2576 \\
 &= \mathbf{0.722}
 \end{aligned}$$

$$\begin{aligned}
 \text{Maka Gain (Play,Outlook)} &= 0.875 - [(5/14) \cdot 0.971 - (0) - (5/14) \cdot 0.722] \\
 &= \mathbf{0.787}
 \end{aligned}$$

.....

Selanjutnya Kolom 2 (Temperature)
Menghitung Gain(play,temperature)

Faktor Hot Temperatur

No	Outlook	Temperature	Humidity	Windy	Play
1	Sunny	Hot	High	False	No
2	Sunny	Hot	High	True	No
3	Cloudy	Hot	High	False	Yes
13	Cloudy	Hot	Normal	False	Yes

Tabel faktor Hot Play terdiri dari **4 instance**. Play No = 2, Play Yes = 2
Entropy play|Temperature=Hot

$$= -(2/4) \cdot \log_2(2/4) - (2/4) \cdot \log_2(2/4)$$

$$= -(0.5) \cdot \log_2(0.5) - (0.5) \cdot \log_2(0.5)$$

$$= -(0.5) \cdot (-1) - (0.5) \cdot (-1)$$

$$= 0.5 + 0.5$$

$$= \mathbf{1.00}$$

Faktor Mild Temperature

No	Outlook	Temperature	Humidity	Windy	Play
4	Rainy	Mild	High	False	Yes
8	Sunny	Mild	High	False	No
10	Rainy	Mild	Normal	False	Yes
11	Sunny	Mild	Normal	True	Yes
12	Cloudy	Mild	High	True	Yes
14	Rainy	Mild	High	True	No

Tabel faktor Mild Play terdiri dari **6 instance**. Play No = 2, Play Yes = 4
Entropy play|Temperature=Mild

$$= -(2/6) \cdot \log_2(2/6) - (4/6) \cdot \log_2(4/6)$$

$$= -(0.333) \cdot \log_2(0.333) - (0.667) \cdot \log_2(0.667)$$

$$= -(0.333) \cdot (-1.586) - (0.667) \cdot (-0.584)$$

$$= 0.528 + 0.389$$

$$= \mathbf{0.92}$$

Faktor cold Temperature

No	Outlook	Temperature	Humidity	Windy	Play
5	Rainy	Cool	Normal	False	Yes
6	Rainy	Cool	Normal	True	Yes
7	Cloudy	Cool	Normal	True	Yes
9	Sunny	Cool	Normal	False	Yes

Tabel faktor Cold Play terdiri dari **4 instance**. Play No = 0, Play Yes = 4
Entropy play|Temperature=Cool

$$= -(0/4) \cdot \log_2(0/4) - (4/4) \cdot \log_2(4/4)$$

$$= \mathbf{0}$$

Maka Gain (Play, Temperature)

$$= 0.875 - [(4/14) \cdot 1 - (6/14) \cdot 0.92 - (0)]$$

$$= 0.875 - [(0.286) - (0.395) - (0)]$$

$$= \mathbf{1.065}$$

Selanjutnya menghitung Kolom 3 (Humidity)

Menghitung Gain|Play,Humidity

Faktor High Humidity

No	Outlook	Temperature	Humidity	Windy	Play
1	Sunny	Hot	High	False	No
2	Sunny	Hot	High	True	No
3	Cloudy	Hot	High	False	Yes
4	Rainy	Mild	High	False	Yes
8	Sunny	Mild	High	False	No
12	Cloudy	Mild	High	True	Yes
14	Rainy	Mild	High	True	No

Tabel faktor High Play terdiri dari 7 instance. Play No = 4, Play Yes = 3

$$\begin{aligned}
 \text{Entropy play|Outlook=Sunny} &= -(4/7) \cdot \log_2(4/7) - (3/7) \cdot \log_2(3/7) \\
 &= -(0.571) \cdot \log_2(0.571) - (0.428) \cdot \log_2(0.428) \\
 &= -(0.571) \cdot (-0.808) - (0.428) \cdot (-1.224) \\
 &= 0.461 + 0.524 \\
 &= \mathbf{0.985}
 \end{aligned}$$

Faktor Normal Humidity

No	Outlook	Temperature	Humidity	Windy	Play
5	Rainy	Cool	Normal	False	Yes
6	Rainy	Cool	Normal	True	Yes
7	Cloudy	Cool	Normal	True	Yes
9	Sunny	Cool	Normal	False	Yes
10	Rainy	Mild	Normal	False	Yes
13	Cloudy	Hot	Normal	False	Yes

Tabel faktor Normal Play terdiri dari 6 instance. Play No = 0, Play Yes = 6

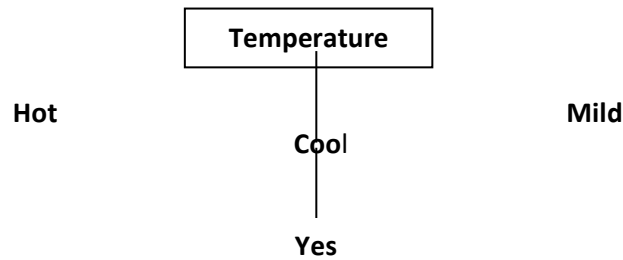
$$\begin{aligned}
 \text{Entropy play|Outlook=Sunny} &= -(0/6) \cdot \log_2(0/6) - (6/6) \cdot \log_2(6/6) \\
 &= \mathbf{0}
 \end{aligned}$$

$$\begin{aligned}
 \text{Maka Gain (Play,Humadity)} &= 0.875 - [(7/14) \cdot 0.985] - (0) \\
 &= 0.875 - 0.492 \\
 &= \mathbf{0.383}
 \end{aligned}$$

Dari perhitungan diatas didapat

1. Gain (Play,Windy) = 0.804
2. Gain (Play,Outlook) = 0.787
3. Gain (Play, Temperature) = 1.605
4. Gain (Play,Humadity) = 0.383

Karena nilai temperatur paling besar, maka pohon keputusan paling atas adalah faktor Temperature.



Cool Temperatur On Play

Dari tabel dibawah, Keputusan untuk bermain diambil jika temperatur dingin/cool

No	Outlook	Temperature	Humidity	Windy	Play
5	Rainy	Cool	Normal	False	Yes
6	Rainy	Cool	Normal	True	Yes
7	Cloudy	Cool	Normal	True	Yes
9	Sunny	Cool	Normal	False	Yes

Hot Temperatur On Play.

No	Outlook	Temperature	Humidity	Windy	Play
1	Sunny	Hot	High	False	No
2	Sunny	Hot	High	True	No
3	Cloudy	Hot	High	False	Yes
13	Cloudy	Hot	Normal	False	Yes

Mencari nilai hubungan Outlook-Temperatur / Gain (temp,Outlook)

Entropy Hot Temperatur = 2/4 no dan 2/4 yes

$$\text{Nilai Entropy} = -(2/4) \cdot \log_2(2/4) - (2/4) \cdot \log_2(2/4) \\ = 1.00$$

Mencari entropy Sunny

karena, instance sunny = 2, nilai play No = 2 dan nilai Play yes = 0, maka
Entropy Sunny = 0

Mencari entropy Cloudy

Karena instance cloudy = 2, nilai Play No = 0 dan nilai play yes = 2, maka
Entropy Cloudy = 0

$$\text{Maka Gain(Outlook-Temperature)} = 1.00 - (2/4 \cdot 0) - (2/4 \cdot 0) \\ = 1.00$$

Mencari nilai hubungan Humidity-Temperatur / Gain (Temp,Hum)

Entropy High terdiri dari 3 instance dengan nilai play no = 2, yes = 1.

$$\begin{aligned} \text{Entropy High} &= -(2/3) \cdot \log_2(2/3) - (1/3) \cdot \log_2(1/3) \\ &= -(0.667) \cdot \log_2(0.667) - (0.333) \cdot \log_2(0.333) \\ &= -(0.667) \cdot (-0.584) - (0.333) \cdot (-1.586) \\ &= 0.389 + 0.528 \\ &= 0.917 \end{aligned}$$

$$\begin{aligned}\text{Entropy Normal} &= -(0/1) \cdot \log_2(0/1) - (1/1) \cdot \log_2(1/1) \\ &= 0\end{aligned}$$

$$\begin{aligned}\text{Maka Gain(Humidity-Temperature)} &= 1.00 - (3/4)(0.971) - (1/4.0) \\ &= 1.00 - 0.728 \\ &= 0.272\end{aligned}$$

Mencari nilai hubungan Windy-Temperatur / Gain (Temp,Windy)

Entropy False terdiri dari 3 instance, dengan nilai play no = 1 dan yes = 2.

$$\begin{aligned}\text{Entropy False} &= -(1/3) \cdot \log_2(1/3) - (2/3) \cdot \log_2(2/3) \\ &= -(0.333) \cdot \log(0.333) - (0.667) \cdot \log(0.667) \\ &= -(0.333) \cdot (-1.586) - (0.667) \cdot (-0.584) \\ &= 0.528 + 0.389 \\ &= 0.917\end{aligned}$$

Entropy True, dengan nilai play No = 1, play yes = 0

Entropy True = 0

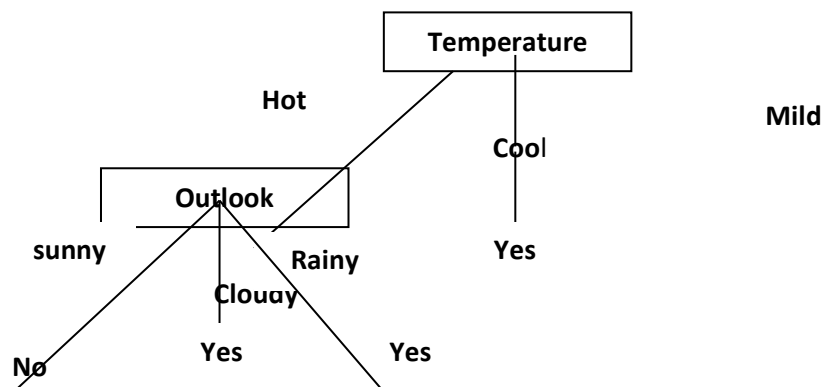
$$\begin{aligned}\text{Maka Gain(Windy-Temperature)} &= 1.00 - (3/4)(0.971) - (1/4.0) \\ &= 1.00 - 0.728 \\ &= 0.272\end{aligned}$$

Maka nilai hubungan antara Temperature Hot dengan Outlook, Humidity dan windy adalah

$$\text{Maka Gain(Outlook-Temperature=hot)} = 1.00$$

$$\text{Maka Gain(Humidity-Temperature=hot)} = 0.272$$

$$\text{Maka Gain(Windy-Temperature =hot)} = 0.272$$



Mild Temperatur On Play.

No	Outlook	Temperature	Humidity	Windy	Play
4	Rainy	Mild	High	False	Yes
8	Sunny	Mild	High	False	No
10	Rainy	Mild	Normal	False	Yes
11	Sunny	Mild	Normal	True	Yes
12	Cloudy	Mild	High	True	Yes
14	Rainy	Mild	High	True	No

Mencari hubungan **Mild Temperatur On Play**

Mencari nilai hubungan Outlook-Temperatur / Gain (temp,Outlook)

Entropy **Mild**Temperatur = 2/6 No dan 4/6 yes

$$\begin{aligned}\text{Nilai Entrophy} &= -(2/6).\log_2(2/6)-(4/6).\log_2(4/6) \\ &= -(0.333).\log_2(0.333)-(0.667).\log_2(0.667) \\ &= 0.917\end{aligned}$$

Mencari entropy Rainy

Instance = 3, Play No = 1 , Play Yes = 2

$$\begin{aligned}\text{Entropy Rainy} &= -(1/3).\log_2(1/3) - (2/3).\log_2(2/3) \\ &= -(0.333).(0.333) - (0.667).\log_2(0.667) \\ &= 0.917\end{aligned}$$

Mencari entropy Sunny

Instance = 2, Play No = 1 , Play Yes = 1

$$\begin{aligned}\text{Entropy Rainy} &= -(1/2).\log_2(1/2) - (1/2).\log_2(1/2) \\ &= -(0.5).\log_2(0.5) - (0.5).\log_2(0.5) \\ &= -(0.5).(-1.0) - (0.5).\log_2(-1.0) \\ &= 0.5 + 0.5 \\ &= 1.00\end{aligned}$$

Mencari entropy Cloudy

Instance = 1, Play No = 0 , Play Yes = 1

$$\begin{aligned}\text{Entropy Rainy} &= -(0/1).\log_2(0/1) - (1/1).\log_2(1/1) \\ &= 0\end{aligned}$$

$$\begin{aligned}\text{Maka Gain(Outlook|Temperature=mild)} &= 0.917 - [(3/6).(0.917)-(2/6).1]-(0) \\ &= 0.917 - [(0.458)-(0.667)-(0)] \\ &= 0.917 - (-0.209) \\ &= 1.126\end{aligned}$$

Mencari hubungan **Mild Temperatur On Play. (temp-humidity)**

Mencari nilai hubungan Humidity-Temperatur / Gain (temp,humidity)

Mencari entropy High

Instance = 4, Play No = 2 , Play Yes = 2

$$\begin{aligned}\text{Entropy Rainy} &= -(2/4) \cdot \log_2(2/4) - (2/4) \cdot \log_2(2/4) \\ &= -(0.5) \cdot (-1.00) - (0.5) \cdot (-1.0) \\ &= 1.00\end{aligned}$$

Mencari entropy Normal

Instance = 2, Play No = 0 , Play Yes = 2

$$\begin{aligned}\text{Entropy Rainy} &= -(0/2) \cdot \log_2(0/2) - (2/2) \cdot \log_2(2/2) \\ &= 0\end{aligned}$$

$$\begin{aligned}\text{Maka Gain(Humadity|Temperature=mild)} &= 0.917 - [(4/6) \cdot (1) - (0)] \\ &= 0.917 - [(0.667) - (0)] \\ &= 0.917 - 0.667 \\ &= 0.25\end{aligned}$$

--

Mencari hubungan **Mild Temperatur On Play. (temp-windy)**

Mencari nilai hubungan Windy-Temperatur / Gain (temp,windy)

Mencari entropy False pada Windy

Instance = 3, Play No = 1 , Play Yes = 2

$$\begin{aligned}\text{Entropy False} &= -(1/3) \cdot \log_2(1/3) - (2/3) \cdot \log_2(2/3) \\ &= -(0.333) \cdot (-1.586) - (0.667) \cdot (-0.584) \\ &= 0.528 + 0.389 \\ &= 0.917\end{aligned}$$

Mencari entropy True pada Windy

Instance = 3, Play No = 1 , Play Yes = 2

$$\begin{aligned}\text{Entropy True} &= -(1/3) \cdot \log_2(1/3) - (2/3) \cdot \log_2(2/3) \\ &= -(0.333) \cdot (-1.586) - (0.667) \cdot (-0.584) \\ &= 0.528 + 0.389 \\ &= 0.917\end{aligned}$$

$$\begin{aligned}\text{Maka Gain(Windy|Temperature=mild)} &= 0.917 - [(3/6) \cdot (0.917) - ((3/6) \cdot (0.917))] \\ &= 0.917 - (0) \\ &= 0.917\end{aligned}$$

Maka nilai hubungan antara Temperature Mild dengan Outlook, Humidity dan windy adalah

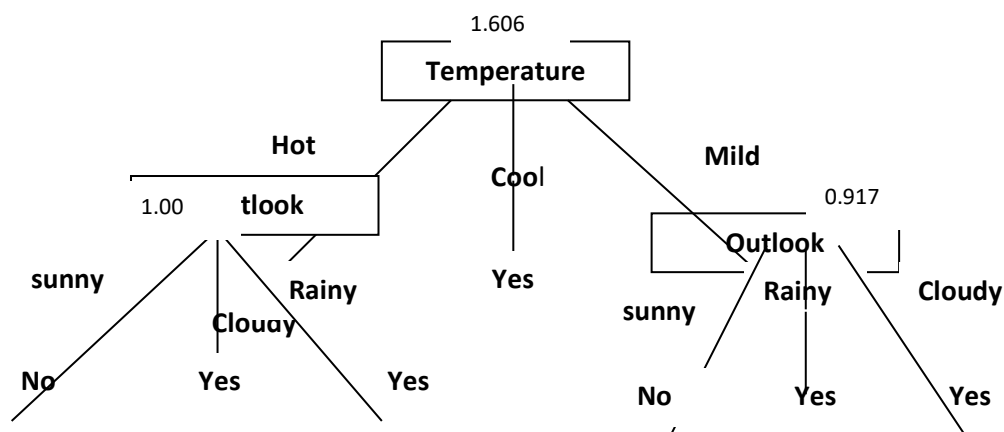
Maka Gain(Outlook-Temperature=mild) = 0.917

Maka Gain(Humadity-Temperature=mild) = 0.25

Maka Gain(Windy-Temperature =mild) = 0.917

Terdapat 2 nilai yang sama, yakni pada **outlook dan Windy**, namun kita bisa cek pada tabel berikut bahwa keputusan untuk bermain pada temperature mild lebih dipengaruhi oleh **Outlook**

No	Outlook	Temperature	Humidity	Windy	Play
4	Rainy	Mild	High	False	Yes
8	Sunny	Mild	High	False	No
10	Rainy	Mild	Normal	False	Yes
11	Sunny	Mild	Normal	True	Yes
12	Cloudy	Mild	High	True	Yes
14	Rainy	Mild	High	True	No



Kesimpulan :

Kelebihan

1. Algoritma ini mampu mentransformasikan dari data mentah menjadi data yang beraturan.
2. Dapat menggunakan atribut nominal dimana kebanyakan tidak dapat dilakukan oleh algoritma machine learning.

Kekurangan

1. Terlalu panjang/lama waktu untuk mencari hubungan satu faktor dengan faktor lainnya.
2. Cenderung kaku dan dapat menimbulkan kesalahan perhitungan.
Metode pembandingan selain menggunakan metode ini dapat dilakukan dengan metode chi-square / CHAID atau Random Forest.

-Tugas 05

Nama : Vero Faloris
Nim : 202420032
Kelas : MTI 23 Reguler A
Mk : Advanced Database

Dari dataset berikut ini, silahkan buat perhitungan manual dari algoritma decision tree C45 secara lengkap dan gambarkan pohon keputusan yang diperoleh.

No	Outlook	Temperature	Humidity	Windy	Play
1	Sunny	Hot	High	FALSE	No
2	Sunny	Hot	High	TRUE	No
3	Cloudy	Hot	High	FALSE	Yes
4	Rainy	Mild	High	FALSE	Yes
5	Rainy	Cool	Normal	FALSE	Yes
6	Rainy	Cool	Normal	TRUE	Yes
7	Cloudy	Cool	Normal	TRUE	Yes
8	Sunny	Mild	High	FALSE	No
9	Sunny	Cool	Normal	FALSE	Yes
10	Rainy	Mild	Normal	FALSE	Yes
11	Sunny	Mild	Normal	TRUE	Yes
12	Cloudy	Mild	High	TRUE	Yes
13	Cloudy	Hot	Normal	FALSE	Yes
14	Rainy	Mild	High	TRUE	No

Kumpulkan dalam format MS Word. Sertakan penjelasan dan pendapat anda tentang algoritma ini, kelebihan dan kekurangannya berbanding metode lain yang sejenis (tentukan sendiri metode pembandingnya).

Decision tree merupakan metode klasifikasi dan prediksi yang sangat kuat dan banyak di minati (Wu, 2009) . Dalam decision tree ini data yang berupa fakta dirubah menjadi sebuah pohon keputusan yang berisi aturan dan tentunya dapat lebih mudah dipahami dengan bahasa alami. Model pohon keputusan banyak digunakan pada kasus data dengan output yang bernilai diskrit . Walaupun tidak menutup kemungkinan dapat juga digunakan untuk kasus data dengan atribut numeric.

Sedangkan K-Nearest Neighbor Algoritma k-nearest neighbor (k-NN atau KNN) adalah sebuah metode untuk melakukan klasifikasi terhadap objek berdasarkan data pembelajaran yang jaraknya paling dekat dengan objek tersebut, Ketepatan algoritma k-NN ini sangat dipengaruhi oleh ada atau tidaknya fitur-fitur yang tidak relevan, atau jika bobot fitur tersebut tidak setara dengan relevansinya terhadap klasifikasi. Riset terhadap algoritma ini sebagian besar membahas bagaimana memilih dan memberi bobot terhadap fitur, agar performa klasifikasi menjadi lebih baik, menurut (Wu, 2009) KNN juga merupakan contoh teknik lazy learning, yaitu teknik yang menunggu sampai pertanyaan (query) datang agar sama dengan data training.

TABEL PERHITUNGAN NODE 1

Rumus :

Entropy = $((-\text{don't play} / \text{jumlah total kasus}) * \log((-\text{don't play} / \text{jumlah total kasus});2)) + ((-\text{play} / \text{jumlah total kasus}) * \log((\text{play} / \text{jumlah total kasus});2))$

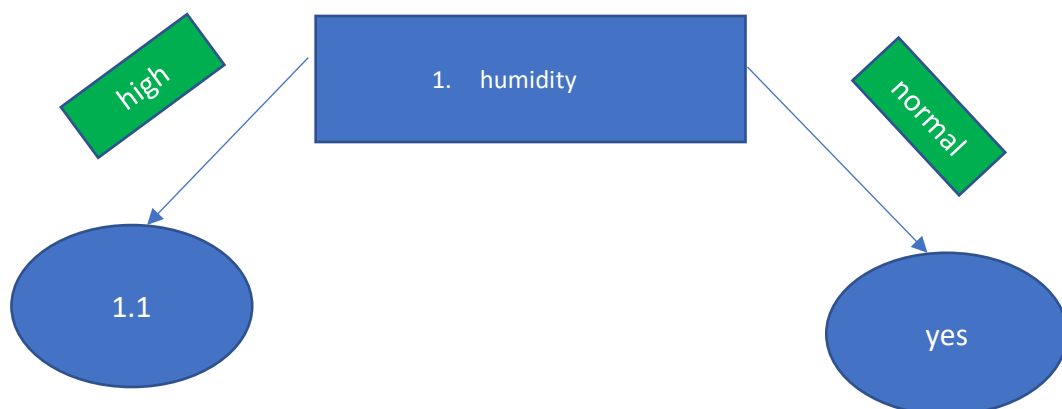
$$1. ((-4/14)*\log(4/14);2)+((-10/14)*\log(10/14);2)= \mathbf{0,863121}$$

Gain = jumlah total entropy – ((jumlah kasus variable x/jumlah total kasus)*entropy variable x)+((jumlah kasus variable x/jumlah total kasus)*entropy variable x)

1. Outlook = $0,863121 - ((4/14)*0) + ((5/14)*0,721928) + ((5/14)*0,970951) = \mathbf{0,25852104}$
2. Temp = $0,863121 - ((4/14)*0) + ((4/14)*1) + ((6/14)*0,918296) = \mathbf{0,18385093}$
3. Humidity = $0,863121 - ((7/14)*0,985228) + ((7/14)*0) = \mathbf{0,3705065}$
4. Windy = $0,863121 - ((8/14)*0,811278) + ((6/14)*0,918296) = \mathbf{0,00597771}$

		Jumlah kasus	Don't play (s1)	Play (s2)	Entropy	Gain
TOTAL		14	4	10	0,863121	
OUTLOOK						0,25852104
	Cloudy	4	0	4	0	
	Raini	5	1	4	0,721928	
	Sunny	5	3	2	0,970951	
TEMP						0,18385093
	Cool	4	0	4	0	
	Hot	4	2	2	1	
	Mild	6	2	4	0,918296	
HUMIDITY						0,3705065
	Hight	7	4	3	0,985228	
	Normal	7	0	7	0	
WINDY						0,00597771
	No	8	2	6	0,811278	
	yes	6	2	4	0,918296	

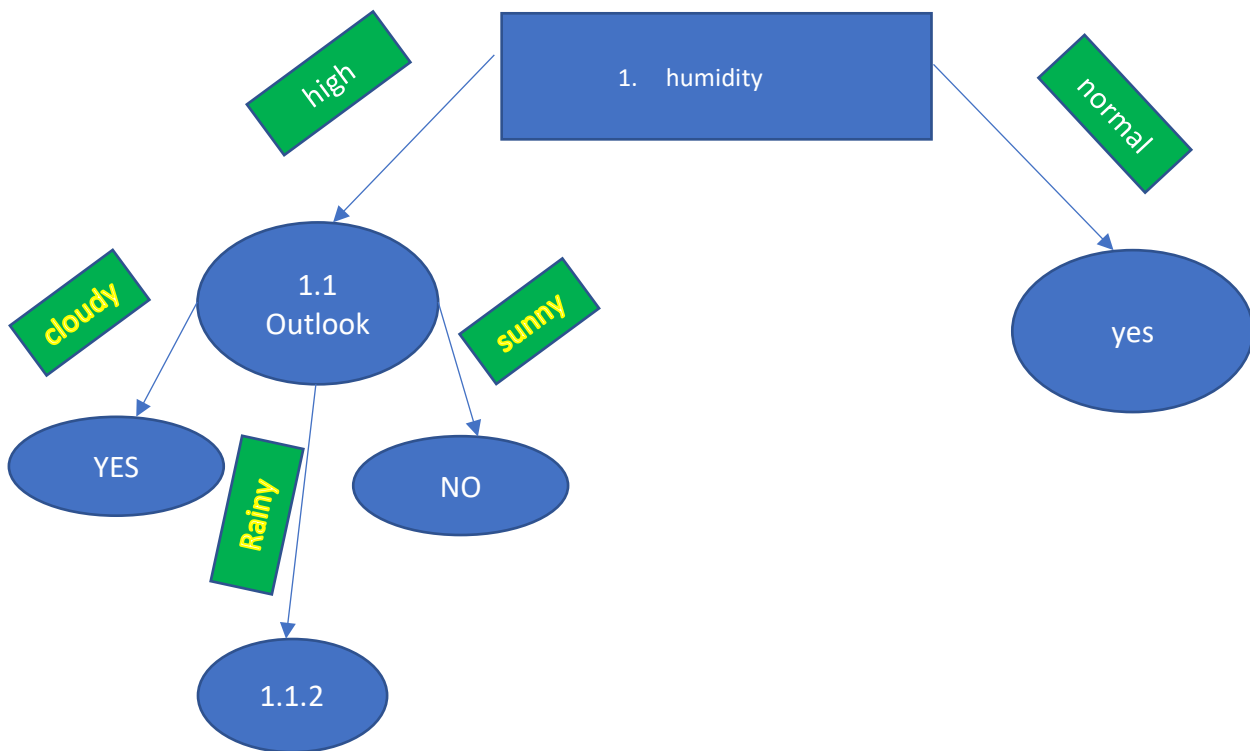
DECISION TREE



TABEL PERHITUNGAN 1.1

		Jumlah kasus	Don't play (s1)	Play (s2)	Entropy	Gain
HUMIDITY		7	4	3	0,985228136	
OUTLOOK						0,69951385
	Cloudy	2	0	2	0	
	Rainy	2	1	1	1	
	Sunny	3	3	0	0	
TEMP						0,020244207
	Cool	0	0	0	0	
	Hot	3	2	1	0,918295834	
	Mild	4	2	2	1	
WINDY						0,020244207
	No	4	2	2	1	
	yes	3	2	1	0,918295834	

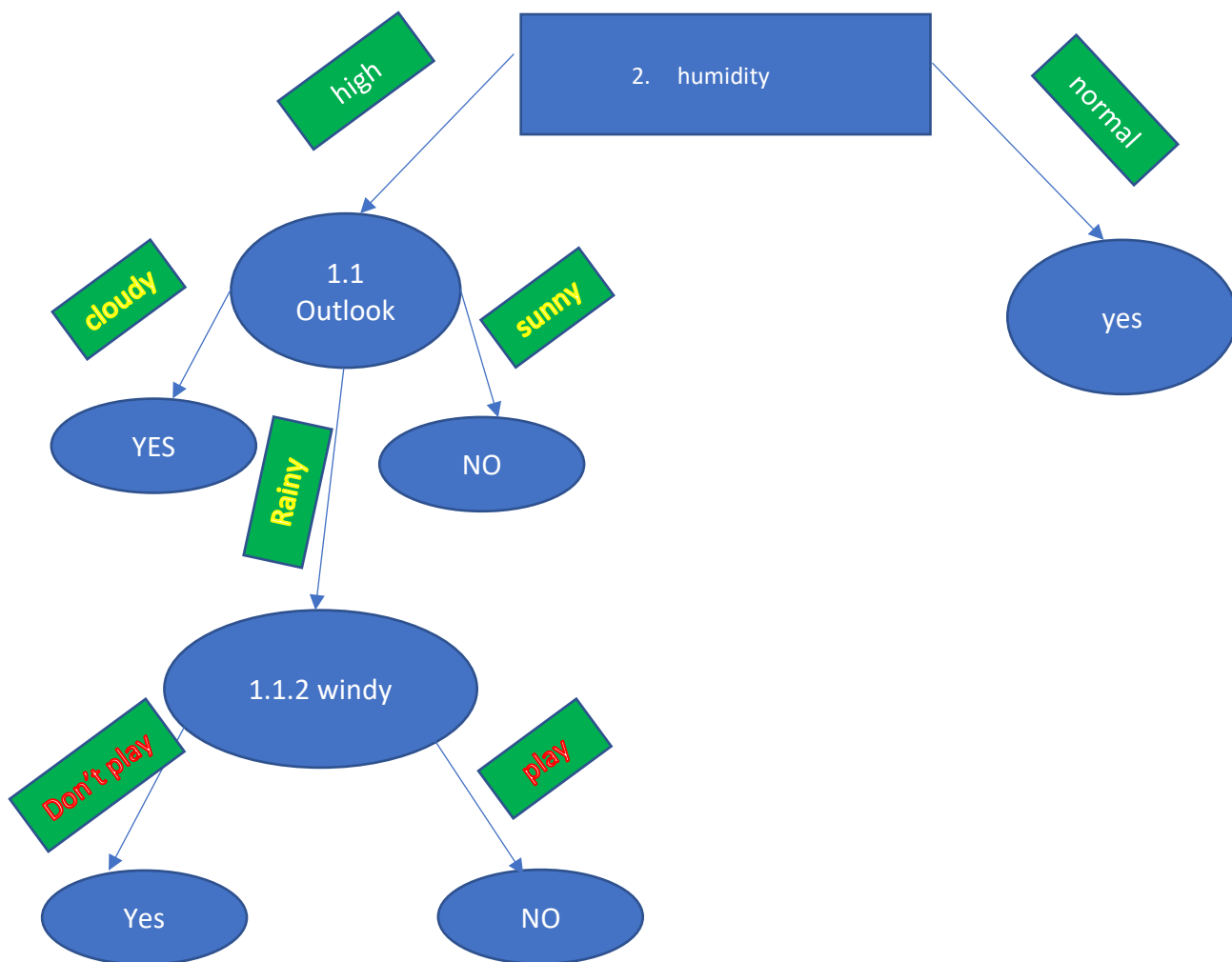
DECISION TREE



TABEL PERHITUNGAN 1.1.2

NODE 1.1.2		Jumlah kasus	Don't play (s1)	Play (s2)	Entropy	Gain
HUMIDITY HIGHT DAN OUTLOOK RAINY		2	1	1	1	
TEMP						0
	Cool	0	0	0	0	
	Hot	0	0	0	0	
	Mild	2	1	1	1	
WINDY						1
	No	1	0	1	0	
	yes	1	1	0	0	

DECISION TREE



Algoritma C 4.5

No	Outlook	Temperature	Humidity	Windy	Play
1	Sunny	Hot	High	FALSE	No
2	Sunny	Hot	High	TRUE	No
3	Cloudy	Hot	High	FALSE	Yes
4	Rainy	Mild	High	FALSE	Yes
5	Rainy	Cool	Normal	FALSE	Yes
6	Rainy	Cool	Normal	TRUE	Yes
7	Cloudy	Cool	Normal	TRUE	Yes
8	Sunny	Mild	High	FALSE	No
9	Sunny	Cool	Normal	FALSE	Yes
10	Rainy	Mild	Normal	FALSE	Yes
11	Sunny	Mild	Normal	TRUE	Yes
12	Cloudy	Mild	High	TRUE	Yes
13	Cloudy	Hot	Normal	FALSE	Yes
14	Rainy	Mild	High	TRUE	No

Perhitungan 1

Node			Jumlah Kasus	No (S1)	Yes (S2)	Entropy	Gain
1	Total		14	4	10	0,863120569	0,258521037
	Outlook						
		Cloudy	4	0	4	0	
		Rainy	5	1	4	0,721928095	
		Sunny	5	3	2	0,970950594	
	Temperature						0,183850925
		Cool	4	0	4	0	
		Hot	4	2	2	1	
		Mild	6	2	4	0,918295834	
	HumidityWindy 1.1.2						0,370506501
		High	7	4	3	0,985228136	
		Normal	7	0	7	0	
	Windy						0,005977711
		FALSE	8	2	6	0,811278124	
		TRUE	6	2	4	0,918295834	

Perhitungan Entropy : rumus :

Keterangan

S : himpunan kasus

A : fitur

n : jumlah partisi S

pi : proporsi dari Si terhadap S

$$\text{Entropy (S)} = \sum_{i=1}^n -p_i * \log_2 p_i$$

$$\text{Entropy Total} = ((-\frac{10}{14} * \log_2 (\frac{10}{14}) + -\frac{4}{14} * \log_2 (\frac{4}{14})) = 0,863120569$$

$$\text{Entropy Cloudy} = ((-\frac{4}{4} * \log_2 (\frac{4}{4}) + -0 * \log_2 (0)) = 0$$

$$\text{Entropy Rainy} = ((-\frac{5}{5} * \log_2 (\frac{5}{5}) + -1 * \log_2 (1)) = 0,721928095$$

$$\text{Entropy Sunny} = ((-\frac{2}{5} * \log_2 (\frac{2}{5}) + -\frac{3}{5} * \log_2 (\frac{3}{5})) = 0,970950594$$

$$\text{Entropy Cool} = ((-\frac{4}{4} * \log_2 (\frac{4}{4}) + -0 * \log_2 (0)) = 0$$

$$\text{Entropy Hot} = ((-\frac{2}{4} * \log_2 (\frac{2}{4}) + -\frac{2}{4} * \log_2 (\frac{2}{4})) = 1$$

$$\text{Entropy Mild} = ((-\frac{4}{6} * \log_2 (\frac{4}{6}) + -\frac{2}{6} * \log_2 (\frac{2}{6})) = 0,918295834$$

$$\text{Entropy High} = ((-\frac{3}{7} * \log_2 (\frac{3}{7}) + -\frac{4}{7} * \log_2 (\frac{4}{7})) = 0,985228136$$

$$\text{Entropy Normal} = ((-\frac{7}{7} * \log_2 (\frac{7}{7}) + -0 * \log_2 (0)) = 0$$

$$\text{Entropy False} = ((-\frac{6}{8} * \log_2 (\frac{6}{8}) + -\frac{2}{8} * \log_2 (\frac{2}{8})) = 0,811278124$$

$$\text{Entropy True} = ((-\frac{4}{6} * \log_2 (\frac{4}{6}) + -\frac{2}{6} * \log_2 (\frac{2}{6})) = 0,918295834$$

Perhitungan Gain : rumus :

Keterangan:

- S : himpunan

- A : atribut

- n : jumlah partisi atribut A

- | Si | : jumlah kasus pada partisi ke-i

- | S | : jumlah kasus dalam S

$$\text{Gain (S,A)} = \text{Entropy (S)} - \sum_{i=1}^n \frac{|S_i|}{|S|} * \text{Entropy (Si)}$$

$$\text{Gain Outlook} = (0,863120569) - (\frac{4}{14} * 0) - (\frac{5}{14} * 0,721928095) - (\frac{5}{14} * 0,970950594) = 0,258521037$$

$$\text{Gain Temperature} = (0,863120569) - (\frac{4}{14} * 0) - (\frac{4}{14} * 1) - (\frac{6}{14} * 0,918295834) = 0,183850925$$

$$\text{Gain Humidity} = (0,863120569) - (\frac{7}{14} * 0,985228136) - (\frac{7}{14} * 0) = \mathbf{0,370506501}$$

$$\text{Gain Windy} = (0,863120569) - (\frac{8}{14} * 0,811278124) - (\frac{6}{14} * 0,918295834) = 0,005977711$$

Sepert yg terlihat pada perhitungan diatas, diperoleh bahwa atribut dgn Gain tertinggi adalah Humidity = **0,370506501**

Maka Humidity menjadi node akar

Humidity memiliki dua nilai yaitu "High" dan "Normal"

Humidity -> "Normal" sdh mengklasifikasikan kasus menjadi 1 yaitu keputusannya "yes"

Untuk humidity -> "High" msh perlu dilakukan perhitungan lagi (karena masih terdapat "yes" dan "no")

Humidity "High"

No	Outlook	Temperature	Humidity	Windy	Play
1	Sunny	Hot	High	FALSE	No
2	Sunny	Hot	High	TRUE	No
3	Cloudy	Hot	High	FALSE	Yes
4	Rainy	Mild	High	FALSE	Yes
5	Sunny	Mild	High	FALSE	No
6	Cloudy	Mild	High	TRUE	Yes
7	Rainy	Mild	High	TRUE	No

Perhitungan 1.1

Node			Jumlah Kasus	No (S1)	Yes (S2)	Entropy	Gain
1.1	Total		7	4	3	0,985228136	
	Outlook						0,69951385
		Cloudy	2	0	2	0	
		Rainy	2	1	1	1	
		Sunny	3	3	0	0	
	Temperature						0,020244207
		Cool	0	0	0	0	
		Hot	3	2	1	0,918295834	
		Mild	4	2	2	1	
	Windy						0,020244207
		FALSE	4	2	2	1	
		TRUE	3	2	1	0,918295834	

Perhitungan Entropy Humadity "High" 1.1

$$\text{Entropy Total} = \left(\left(-\frac{3}{7} \log_2 \left(\frac{3}{7} \right) \right) + \left(-\frac{4}{7} \log_2 \left(\frac{4}{7} \right) \right) \right) = 0,985228136$$

$$\text{Entropy Cloudy} = \left(\left(-\frac{2}{2} \log_2 \left(\frac{2}{2} \right) \right) + \left(-\frac{0}{2} \log_2 \left(\frac{0}{2} \right) \right) \right) = 0$$

$$\text{Entropy Rainy} = \left(\left(-\frac{1}{2} \log_2 \left(\frac{1}{2} \right) \right) + \left(-\frac{1}{2} \log_2 \left(\frac{1}{2} \right) \right) \right) = 1$$

$$\text{Entropy Sunny} = \left(\left(-\frac{0}{3} \log_2 \left(\frac{0}{3} \right) \right) + \left(-\frac{3}{3} \log_2 \left(\frac{3}{3} \right) \right) \right) = 0$$

$$\text{Entropy Cool} = \left(\left(-\frac{0}{0} \log_2 \left(\frac{0}{0} \right) \right) + \left(-\frac{0}{0} \log_2 \left(\frac{0}{0} \right) \right) \right) = 0$$

$$\text{Entropy Hot} = \left(\left(-\frac{1}{3} \log_2 \left(\frac{1}{3} \right) \right) + \left(-\frac{2}{3} \log_2 \left(\frac{2}{3} \right) \right) \right) = 0,918295834$$

$$\text{Entropy Mild} = \left(\left(-\frac{2}{4} \log_2 \left(\frac{2}{4} \right) \right) + \left(-\frac{2}{4} \log_2 \left(\frac{2}{4} \right) \right) \right) = 1$$

$$\text{Entropy False} = \left(\left(-\frac{2}{4} \log_2 \left(\frac{2}{4} \right) \right) + \left(-\frac{2}{4} \log_2 \left(\frac{2}{4} \right) \right) \right) = 1$$

$$\text{Entropy True} = \left(\left(-\frac{1}{3} \log_2 \left(\frac{1}{3} \right) \right) + \left(-\frac{2}{3} \log_2 \left(\frac{2}{3} \right) \right) \right) = 0,918295834$$

Perhitungan Gain Humadity "High" 1.1

$$\text{Gain Outlook} = (0,985228136) - (2 \frac{2}{7}) * 0 - (2 \frac{2}{7}) * 1 - (3 \frac{2}{7}) * 0 = 0,69951385$$

$$\text{Gain Temperature} = (0,985228136) - (0 \frac{2}{7}) * 0 - (3 \frac{2}{7}) * 0,918295834 - (4 \frac{2}{7}) * 1 = 0,020244207$$

$$\text{Gain Windy} = (0,985228136) - (4 \frac{2}{7}) * 1 - (3 \frac{2}{7}) * 0,918295834 = 0,020244207$$

Didapat Gain tertinggi -FALSE

> outlook -> 0,69951385

Maka "Outlook" menjadi node cabang dari atribut humidity yg bernilai "High"

Berdasarkan atribut "Outlook" terdpt 3 nilai

Cloudy

Rainy

Sunny

Karena "Cloudy" pasti bernilai "Yes" dan "Sunny" pasti bernilai "No", maka tidak perlu dilakukan perhitungan lagi.

Sedangkan "Rainy" bernilai "yes" dan "No", maka masih perlu dilakukan perhitungan lagi

Humadity "High" dan Outlook "Rainy"

No	Outlook	Temperature	Humidity	Windy	Play
4	Rainy	Mild	High	FALSE	Yes
7	Rainy	Mild	High	TRUE	No

Perhitungan 1.1.2

Node			Jumlah Kasus	No (S1)	Yes (S2)	Entropy	Gain
1..1.2	Total		2	1	1	1	
	Temperature						0
		Cool	0	0	0	0	
		Hot	0	0	0	0	
		Mild	2	1	1	1	
	Windy						1
		FALSE	1	0	1	0	
		TRUE	1	1	0	0	

Perhitungan Entropy "Rainy" 1.1.2

$$\text{Entropy Total} = ((-1 \frac{2}{2}) * \log_2 (1 \frac{2}{2}) + (-1 \frac{2}{2}) * \log_2 (1 \frac{2}{2})) = 1$$

$$\text{Entropy Cool} = ((-0 \frac{2}{0}) * \log_2 (0 \frac{2}{0}) + (-0 \frac{2}{0}) * \log_2 (0 \frac{2}{0})) = 0$$

$$\text{Entropy Hot} = ((-0 \frac{2}{0}) * \log_2 (0 \frac{2}{0}) + (-0 \frac{2}{0}) * \log_2 (0 \frac{2}{0})) = 0$$

$$\text{Entropy Mild} = ((-1 \frac{2}{2}) * \log_2 (1 \frac{2}{2}) + (-1 \frac{2}{2}) * \log_2 (1 \frac{2}{2})) = 1$$

$$\text{Entropy False} = ((-1 \frac{2}{1}) * \log_2 (1 \frac{2}{1}) + (-0 \frac{2}{1}) * \log_2 (0 \frac{2}{1})) = 0$$

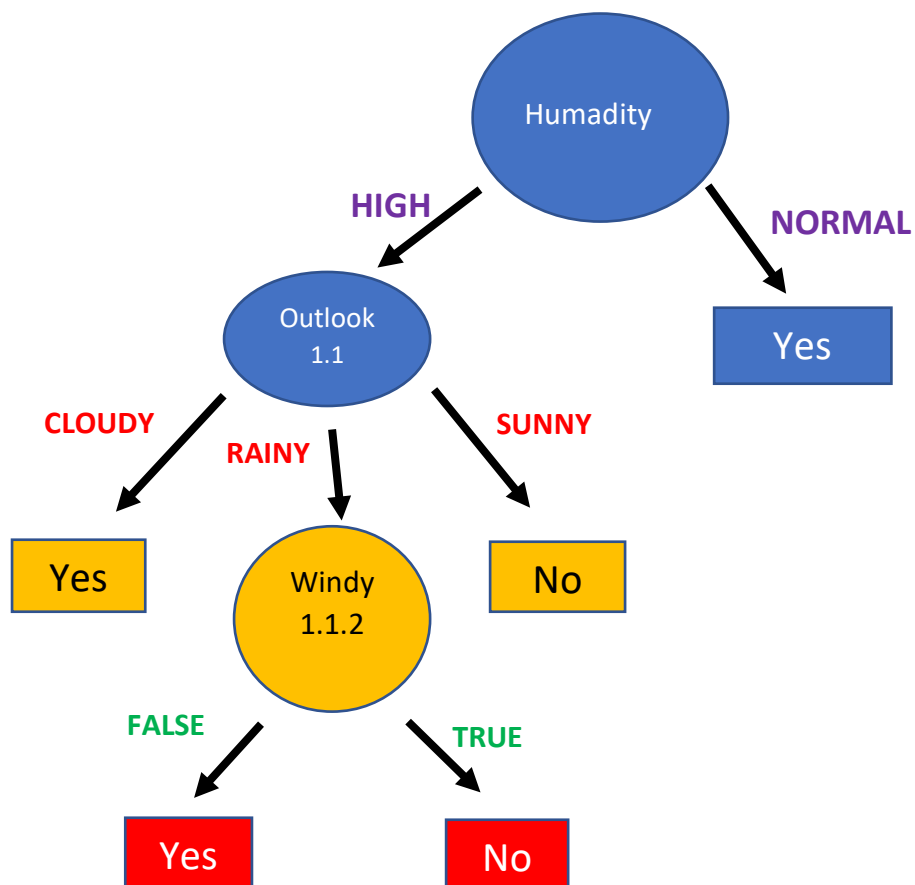
$$\text{Entropy True} = \left(-0 \frac{1}{2} \cdot 1 \cdot \log_2 \left(0 \frac{1}{2} \cdot 1 \right) + -1 \frac{1}{2} \cdot 0 \cdot \log_2 \left(1 \frac{1}{2} \cdot 0 \right) \right) = 0$$

Perhitungan Gain "Rainy" 1.1.2

$$\text{Gain Temperature} = (1) - \left(0 \frac{1}{2} \cdot 2 \cdot 0 \right) - \left(0 \frac{1}{2} \cdot 2 \cdot 0 \right) - \left(2 \frac{1}{2} \cdot 2 \cdot 1 \right) = 0$$

$$\text{Gain Windy} = (1) - \left(1 \frac{1}{2} \cdot 2 \cdot 0 \right) - \left(1 \frac{1}{2} \cdot 2 \cdot 0 \right) = 1$$

POHON KEPUTUSAN



Kelebihan Algoritma C4.5 :

1. C4.5 mampu menangani atribut yang kosong (Missing Value).
2. C4.5 mampu menangani atribut dengan kontinu.
3. C4.5 memangkas pohon keputusan untuk mengatasi overfitting.

Kekurangan Algoritma C4.5 :

1. Algoritma C4.5 hanya dapat digunakan untuk menangani sampel-sampel yang dapat disimpan secara keseluruhan dan pada waktu yang bersamaan di memori.

Nama : WIDIA ASTUTI

NIM : 202420021

MATA KULIAH : ADVANCED DATABASE

Tugas 05

Dari dataset berikut ini, silahkan buat perhitungan manual dari algoritma decision tree C45 secara lengkap dan gambarkan pohon keputusan yang diperoleh.

No	Outlook	Temperature	Humidity	Windy	Play
1	Sunny	Hot	High	FALSE	No
2	Sunny	Hot	High	TRUE	No
3	Cloudy	Hot	High	FALSE	Yes
4	Rainy	Mild	High	FALSE	Yes
5	Rainy	Cool	Normal	FALSE	Yes
6	Rainy	Cool	Normal	TRUE	Yes
7	Cloudy	Cool	Normal	TRUE	Yes
8	Sunny	Mild	High	FALSE	No
9	Sunny	Cool	Normal	FALSE	Yes
10	Rainy	Mild	Normal	FALSE	Yes
11	Sunny	Mild	Normal	TRUE	Yes
12	Cloudy	Mild	High	TRUE	Yes
13	Cloudy	Hot	Normal	FALSE	Yes
14	Rainy	Mild	High	TRUE	No

Kumpulkan dalam format MS Word. Sertakan penjelasan dan pendapat anda tentang algoritma ini, kelebihan dan kekurangannya berbanding metode lain yang sejenis (tentukan sendiri metode pembandingnya).

JAWAB :

ALGORITMA DECISION TREE C45

Rumus Perhitungan Entrophy

$$Entropy(s) = \sum_{i=1}^n -P_i \times \log_2 P_i$$

Rumus Perhitungan Gain

$$Gain(s, A) = Entropy(s) - \sum_{i=1}^n \times Entropy(s_i)$$

PERHITUNGAN NODE 1

A. Perhitungan total Entropy

$$\begin{aligned} &= ((-jumlahNo/totalkasus)*Log_2(jumlahNo/totalkasus) + (-jumlahYes/totalkasus)*Log_2(jumlahYes/totalkasus)) \\ &= ((-4/14)*Log_2(4/14) + ((-10/14)*Log_2(10/14))) \\ &= 0,863120569 \end{aligned}$$

1. Perhitungan Outlook

a. Perhitungan Entropy Cloudy

$$\begin{aligned} &= ((-jumlahNo/totalkasus)*Log_2(jumlahNo/totalkasus) + (-jumlahYes/totalkasus)*Log_2(jumlahYes/totalkasus)) \\ &= ((-0/4)*Log_2(0/4) + ((-4/4)*Log_2(4/4))) \\ &= 0 \end{aligned}$$

b. Perhitungan Entropy Rainy

$$\begin{aligned} &= ((-jumlahNo/totalkasus)*Log_2(jumlahNo/totalkasus) + (-jumlahYes/totalkasus)*Log_2(jumlahYes/totalkasus)) \\ &= ((-1/5)*Log_2(1/5) + ((-4/5)*Log_2(4/5))) \\ &= 0,721928095 \end{aligned}$$

c. Perhitungan Entropy Sunny

$$\begin{aligned} &= ((-jumlahNo/totalkasus)*Log_2(jumlahNo/totalkasus) + (-jumlahYes/totalkasus)*Log_2(jumlahYes/totalkasus)) \\ &= ((-3/5)*Log_2(3/5) + ((-2/5)*Log_2(2/5))) \\ &= 0,970950594 \end{aligned}$$

2. Perhitungan Temperature

a. Perhitungan Entropy Cool

$$\begin{aligned} &= ((-jumlahNo/totalkasus)*Log_2(jumlahNo/totalkasus) + (-jumlahYes/totalkasus)*Log_2(jumlahYes/totalkasus)) \\ &= ((-0/4)*Log_2(0/4) + ((-4/4)*Log_2(4/4))) \\ &= 0 \end{aligned}$$

b. Perhitungan Entropy Hot

$$\begin{aligned} &= ((-jumlahNo/totalkasus)*Log_2(jumlahNo/totalkasus) + (-jumlahYes/totalkasus)*Log_2(jumlahYes/totalkasus)) \\ &= ((-2/4)*Log_2(2/4) + ((-2/4)*Log_2(2/4))) \\ &= 1 \end{aligned}$$

c. Perhitungan Entropy Mild

$$\begin{aligned} &= ((-jumlahNo/totalkasus)*Log_2(jumlahNo/totalkasus) + (-jumlahYes/totalkasus)*Log_2(jumlahYes/totalkasus)) \\ &= ((-2/6)*Log_2(2/6) + ((-4/6)*Log_2(4/6))) \\ &= 0,918295834 \end{aligned}$$

3. Perhitungan Humidity

a. Perhitungan Entropy High

$$\begin{aligned}
&=((- \text{jumlahNo} / \text{total kasus}) * \text{Log}_2(\text{jumlahNo} / \text{total kasus}) + (- \text{jumlahYes} / \text{total kasus}) * \text{Log}_2(\text{jumlahYes} / \text{total kasus})) \\
&=((-4/7) * \text{Log}_2(4/7) + ((-3/7) * \text{Log}_2(3/7))) \\
&= 0,985228136
\end{aligned}$$

b. Perhitungan Entropy Normal

$$\begin{aligned}
&=((- \text{jumlahNo} / \text{total kasus}) * \text{Log}_2(\text{jumlahNo} / \text{total kasus}) + (- \text{jumlahYes} / \text{total kasus}) * \text{Log}_2(\text{jumlahYes} / \text{total kasus})) \\
&=((-0/7) * \text{Log}_2(0/7) + ((-7/7) * \text{Log}_2(7/7))) \\
&= 0
\end{aligned}$$

4. Perhitungan Windy

a. Perhitungan Entropy False

$$\begin{aligned}
&=((- \text{jumlahNo} / \text{total kasus}) * \text{Log}_2(\text{jumlahNo} / \text{total kasus}) + (- \text{jumlahYes} / \text{total kasus}) * \text{Log}_2(\text{jumlahYes} / \text{total kasus})) \\
&=((-2/8) * \text{Log}_2(2/8) + ((-6/8) * \text{Log}_2(6/8))) \\
&= 0,811278124
\end{aligned}$$

b. Perhitungan Entropy True

$$\begin{aligned}
&=((- \text{jumlahNo} / \text{total kasus}) * \text{Log}_2(\text{jumlahNo} / \text{total kasus}) + (- \text{jumlahYes} / \text{total kasus}) * \text{Log}_2(\text{jumlahYes} / \text{total kasus})) \\
&=((-4/6) * \text{Log}_2(4/6) + ((-2/6) * \text{Log}_2(2/6))) \\
&= 0,918295834
\end{aligned}$$

B. Perhitungan Gain

1. Gain Outlook

$$\begin{aligned}
&= (\text{Entropy total}) - ((\text{total cloudy} / \text{total kasus}) * \text{entropy cloudy}) - \\
&\quad ((\text{total rainy} / \text{total kasus}) * \text{entropy rainy}) - ((\text{total sunny} / \text{total kasus}) * \text{entropy sunny}) \\
&= (0,863120569) - ((4/14) * 0) - ((5/14) * 0,721928095) - ((5/14) * 0,970950594) \\
&= 0,258521037
\end{aligned}$$

2. Gain Temperature

$$\begin{aligned}
&= (\text{Entropy total}) - ((\text{total cool} / \text{total kasus}) * \text{entropy cool}) - \\
&\quad ((\text{total hot} / \text{total kasus}) * \text{entropy hot}) - ((\text{total mild} / \text{total kasus}) * \text{entropy mild}) \\
&= (0,863120569) - ((4/14) * 0) - ((4/14) * 1) - ((6/14) * 0,918295834) \\
&= 0,183850925
\end{aligned}$$

3. Gain Humidity

$$\begin{aligned}
&= (\text{Entropy total}) - ((\text{total high} / \text{total kasus}) * \text{entropy high}) - \\
&\quad ((\text{total normal} / \text{total kasus}) * \text{entropy normal}) \\
&= (0,863120569) - ((7/14) * 0,985228136) - ((7/14) * 0) \\
&= 0,370506501
\end{aligned}$$

4. Gain Windy

$$\begin{aligned}
&= (\text{Entropy total}) - ((\text{total high} / \text{total kasus}) * \text{entropy high}) - \\
&\quad ((\text{total normal} / \text{total kasus}) * \text{entropy normal}) \\
&= (0,863120569) - ((7/14) * 0,985228136) - ((7/14) * 0) \\
&= 0,005977711
\end{aligned}$$

Tabel hasil perhitungan Node 1

	jumlah kasus	NO	YES	ENTROPHY	GAIN
Total	14	4	10	0,863120569	
Outlook					0,258521037
cloudy	4	0	4	0	
rainy	5	1	4	0,721928095	
sunny	5	3	2	0,970950595	
Temperature					0,183850925

	cool	4	0	4	0	
	hot	4	2	2	1	
	mild	6	2	4	0,918295834	
Humidity						0,370506501
	high	7	4	3	0,985228136	
	normal	7	0	7	0	
Windy						0,005977711
	FALSE	8	2	6	0,811278125	
	TRUE	6	4	2	0,918295834	

Note : Nilai gain tertinggi pada humidity, sehingga perhitungan selanjutnya dimulai pada Humidi

PERHITUNGAN NODE 1.1

A. Perhitungan total Humidity

$$\begin{aligned}
 &= ((\text{jumlahNo}/\text{total kasus}) * \text{Log}_2(\text{jumlahNo}/\text{total kasus}) + (-\text{jumlahYes}/\text{total kasus}) * \text{Log}_2(\text{jumlahYes}/\text{total kasus})) \\
 &= ((-4/7) * \text{Log}_2(4/7) + ((-3/7) * \text{Log}_2(3/7))) \\
 &= 0,985228136
 \end{aligned}$$

1. Perhitungan Outlook

a. Perhitungan Entropy Cloudy

$$\begin{aligned}
 &= ((-\text{jumlahNo}/\text{total kasus}) * \text{Log}_2(\text{jumlahNo}/\text{total kasus}) + (-\text{jumlahYes}/\text{total kasus}) * \text{Log}_2(\text{jumlahYes}/\text{total kasus})) \\
 &= ((-0/2) * \text{Log}_2(0/2) + ((-2/2) * \text{Log}_2(2/2))) \\
 &= 0
 \end{aligned}$$

b. Perhitungan Entropy Rainy

$$\begin{aligned}
 &= ((-\text{jumlahNo}/\text{total kasus}) * \text{Log}_2(\text{jumlahNo}/\text{total kasus}) + (-\text{jumlahYes}/\text{total kasus}) * \text{Log}_2(\text{jumlahYes}/\text{total kasus})) \\
 &= ((-1/2) * \text{Log}_2(1/2) + ((-1/2) * \text{Log}_2(1/2))) \\
 &= 1
 \end{aligned}$$

c. Perhitungan Entropy Sunny

$$\begin{aligned}
 &= ((-\text{jumlahNo}/\text{total kasus}) * \text{Log}_2(\text{jumlahNo}/\text{total kasus}) + (-\text{jumlahYes}/\text{total kasus}) * \text{Log}_2(\text{jumlahYes}/\text{total kasus})) \\
 &= ((-3/3) * \text{Log}_2(3/3) + ((-0/3) * \text{Log}_2(0/3))) \\
 &= 0
 \end{aligned}$$

2. Perhitungan Temperature

a. Perhitungan Entropy Cool

$$\begin{aligned}
 &= ((-\text{jumlahNo}/\text{total kasus}) * \text{Log}_2(\text{jumlahNo}/\text{total kasus}) + (-\text{jumlahYes}/\text{total kasus}) * \text{Log}_2(\text{jumlahYes}/\text{total kasus})) \\
 &= ((-0/0) * \text{Log}_2(0/0) + ((-0/0) * \text{Log}_2(0/0))) \\
 &= 0
 \end{aligned}$$

b. Perhitungan Entropy Hot

$$\begin{aligned}
 &= ((-\text{jumlahNo}/\text{total kasus}) * \text{Log}_2(\text{jumlahNo}/\text{total kasus}) + (-\text{jumlahYes}/\text{total kasus}) * \text{Log}_2(\text{jumlahYes}/\text{total kasus})) \\
 &= ((-2/3) * \text{Log}_2(2/3) + ((-1/3) * \text{Log}_2(1/3))) \\
 &= 0,918295834
 \end{aligned}$$

c. Perhitungan Entropy Mild

$$\begin{aligned}
 &= ((-\text{jumlahNo}/\text{total kasus}) * \text{Log}_2(\text{jumlahNo}/\text{total kasus}) + (-\text{jumlahYes}/\text{total kasus}) * \text{Log}_2(\text{jumlahYes}/\text{total kasus}))
 \end{aligned}$$

$$= ((-2/4) * \log_2(2/4) + ((-2/4) * \log_2(2/4)))$$

$$= 1$$

3. Perhitungan Windy

a. Perhitungan Entrophy False

$$= ((- \text{jumlahNo} / \text{total kasus}) * \log_2(\text{jumlahNo} / \text{total kasus}) + (- \text{jumlahYes} / \text{total kasus}) * \log_2(\text{jumlahYes} / \text{total kasus}))$$

$$= ((-2/4) * \log_2(2/4) + ((-2/4) * \log_2(2/4)))$$

$$= 1$$

b. Perhitungan Entrophy True

$$= ((- \text{jumlahNo} / \text{total kasus}) * \log_2(\text{jumlahNo} / \text{total kasus}) + (- \text{jumlahYes} / \text{total kasus}) * \log_2(\text{jumlahYes} / \text{total kasus}))$$

$$= ((-2/3) * \log_2(2/3) + ((-1/3) * \log_2(1/3)))$$

$$= 0,918295834$$

C. Perhitungan Gain

1. Gain Outlook

$$= (\text{Entropytotal}) - ((\text{totalcloudy} / \text{total kasus}) * \text{entropycloudy}) - ((\text{totalrainy} / \text{total kasus}) * \text{entropyrainy}) - ((\text{totalsunny} / \text{total kasus}) * \text{entropysunny})$$

$$= (0,985228136) - ((2/7) * 0) - ((2/7) * 1) - ((3/7) * 0)$$

$$= 0,69951385$$

2. Gain Temperature

$$= (\text{Entropytotal}) - ((\text{totalcool} / \text{total kasus}) * \text{entropycool}) - ((\text{totalhot} / \text{total kasus}) * \text{entropyhot}) - ((\text{totalmild} / \text{total kasus}) * \text{entropymild})$$

$$= (0,985228136) - ((0/7) * 0) - ((4/7) * 0,918295834) - ((3/7) * 1)$$

$$= 0,020244207$$

3. Gain Windy

$$= (\text{Entropytotal}) - ((\text{totalhigh} / \text{total kasus}) * \text{entropyhigh}) - ((\text{totalnormal} / \text{total kasus}) * \text{entropynormal})$$

$$= (0,985228136) - ((4/7) * 1) - ((4/7) * 0,918295834)$$

$$= 0,020244207$$

Tabel hasil perhitungan Node 1.1

		jumlah kasus	NO	YES	ENTROPHY	GAIN
Humidity		7	4	3	0,985228134	
Outlook						0,69951385
	cloudy	2	0	2	0	
	rainy	2	1	1	1	
	sunny	3	3	0	0	
Temperature						0,020244207
	cool	0	0	0	0	
	hot	2	1	1	1	
	mild	4	2	2	1	
Windy						0,020244207
	FALSE	4	2	2	1	
	TRUE	3	2	1	0,918295834	

PERHITUNGAN NODE 1.1.2

A. Perhitungan Entropy total Humidity high and outlook rainny

$$\begin{aligned}
&= ((-\text{jumlahNo}/\text{total kasus}) * \text{Log}_2(\text{jumlahNo}/\text{total kasus}) + (-\text{jumlahYes}/\text{total kasus}) * \text{Log}_2(\text{jumlahYes}/\text{total kasus})) \\
&= ((-1/2) * \text{Log}_2(1/2) + ((-1/2) * \text{Log}_2(1/2))) \\
&= 1
\end{aligned}$$

1. Perhitungan Temperature

a. Perhitungan Entropy Cool

$$\begin{aligned}
&= ((-\text{jumlahNo}/\text{total kasus}) * \text{Log}_2(\text{jumlahNo}/\text{total kasus}) + (-\text{jumlahYes}/\text{total kasus}) * \text{Log}_2(\text{jumlahYes}/\text{total kasus})) \\
&= 0
\end{aligned}$$

b. Perhitungan Entropy Hot

$$\begin{aligned}
&= ((-\text{jumlahNo}/\text{total kasus}) * \text{Log}_2(\text{jumlahNo}/\text{total kasus}) + (-\text{jumlahYes}/\text{total kasus}) * \text{Log}_2(\text{jumlahYes}/\text{total kasus})) \\
&= 0
\end{aligned}$$

c. Perhitungan Entropy Mild

$$\begin{aligned}
&= ((-\text{jumlahNo}/\text{total kasus}) * \text{Log}_2(\text{jumlahNo}/\text{total kasus}) + (-\text{jumlahYes}/\text{total kasus}) * \text{Log}_2(\text{jumlahYes}/\text{total kasus})) \\
&= ((-1/2) * \text{Log}_2(1/2) + ((-1/2) * \text{Log}_2(1/2))) \\
&= 1
\end{aligned}$$

2. Perhitungan Windy

a. Perhitungan Entropy False

$$\begin{aligned}
&= ((-\text{jumlahNo}/\text{total kasus}) * \text{Log}_2(\text{jumlahNo}/\text{total kasus}) + (-\text{jumlahYes}/\text{total kasus}) * \text{Log}_2(\text{jumlahYes}/\text{total kasus})) \\
&= 0
\end{aligned}$$

b. Perhitungan Entropy True

$$\begin{aligned}
&= ((-\text{jumlahNo}/\text{total kasus}) * \text{Log}_2(\text{jumlahNo}/\text{total kasus}) + (-\text{jumlahYes}/\text{total kasus}) * \text{Log}_2(\text{jumlahYes}/\text{total kasus})) \\
&= 0
\end{aligned}$$

B. Perhitungan Gain

1. Gain Temperature

$$\begin{aligned}
&= (\text{Entropytotal}) - ((\text{totalcool}/\text{total kasus}) * \text{entropycool}) - \\
&\quad ((\text{totalhot}/\text{total kasus}) * \text{entropyhot}) - ((\text{totalmild}/\text{total kasus}) * \text{entropymild}) \\
&= 0
\end{aligned}$$

2. Gain Windy

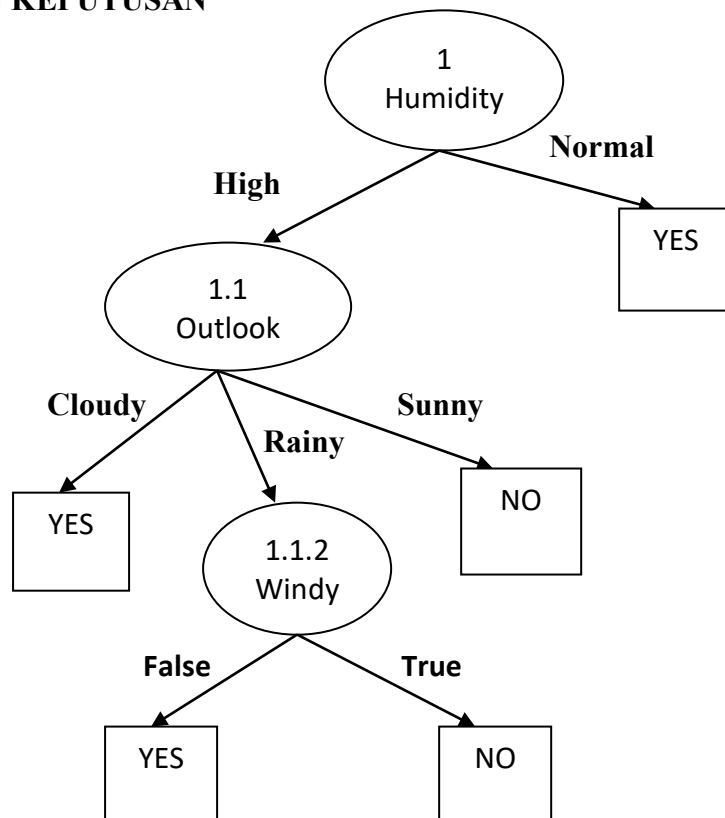
$$\begin{aligned}
&= (\text{Entropytotal}) - ((\text{totalhigh}/\text{total kasus}) * \text{entropyhigh}) - \\
&\quad ((\text{totalnormal}/\text{total kasus}) * \text{entropynormal}) \\
&= 1
\end{aligned}$$

Tabel hasil done 1.1.2

		Total	NO	YES	Entropy	Gain
humidity high and outlook rainy		2	1	1	1	
Temperatatur						0
	Cool	0	0	0	0	
	Hot	0	0	0	0	
	Mild	2	1	1	1	
Windy						1
	NO	1	0	1	0	

YES	1	1	0	0
-----	---	---	---	---

POHON KEPUTUSAN



Pada kasus tsb terlihat bahwa kita dapat menangani atribut dengan tipe data diskrit dan kontinyu, mampu menangani atribut yang kosong, dan dapat melakukan pemangkasan pohon. Adapun metode yang dapat membandingkan algoritma c4.5 yaitu Algoritma K-Nearest Neighbor (K-NN) yang merupakan sebuah metode untuk melakukan klasifikasi terhadap objek berdasarkan dari data pembelajaran yang jaraknya paling dekat dengan objek tersebut.

Algoritma metode k-nearest neighbor (KNN) bekerja berdasarkan jarak terdekat dari query instance ke training sample untuk menentukan KNN-nya. Training sample diproyeksikan, sehingga lebih akurat dibandingkan K-NN dan juga menunjukkan C4.5 memiliki akurasi tertinggi.

TUGAS
DECISION TREES ALGORITHM



Dibuat Oleh

Aan Novrianto

Dosen Pengampu

TRI BASUKI KURNIAWAN, S.Kom, M.Eng, Ph.D

Program Pasca Sarjana

Universitas Binadarma Palembang

Tugas 05

No	Outlook	Temperature	Humidity	Windy	Play
1	Sunny	Hot	High	False	No
2	Sunny	Hot	High	True	No
3	Cloudy	Hot	High	False	Yes
4	Rainy	Mild	High	False	Yes
5	Rainy	Cool	Normal	False	Yes
6	Rainy	Cool	Normal	True	Yes
7	Cloudy	Cool	Normal	True	Yes
8	Sunny	Mild	High	False	No
9	Sunny	Cool	Normal	False	Yes
10	Rainy	Mild	Normal	False	Yes
11	Sunny	Mild	Normal	True	Yes
12	Cloudy	Mild	High	True	Yes
13	Cloudy	Hot	Normal	False	Yes
14	Rainy	Mild	High	True	No

Rumus Penyelesaian

$$\text{Entropy}(S) = \sum - p(I) \cdot \log_2 p(I)$$

$$\text{Gain}(S, A) = \text{Entropy}(S) - \sum [p(S|A) \cdot \text{Entropy}(S|A)]$$

Entropy

Kita hitung dulu Nilai Entropy. Kolom Play terdiri dari 14 instance dan terdiri dari 2 label yakni Yes dan No. Play Yes = 10, _Play No= 4

$$\begin{aligned}\text{Entropy(Decision)} &= -p(\text{Yes}) \cdot \log_2 p(\text{Yes}) - p(\text{No}) \cdot \log_2 p(\text{No}) \\ \text{Entropy(Decision)} &= - (10/14) \cdot \log_2(10/14) - (4/14) \cdot \log_2(4/14) \\ &= -(0.714) \cdot \log_2(0.714) - (0.4) \cdot \log_2(0.4) \\ &= 0.347 + 0.528 \\ &= \mathbf{0.875}\end{aligned}$$

Menghitung faktor dominan untuk memutuskan bermain.

Faktor Windy dan Play

False Factor windy to Play

No	Outlook	Temperature	Humidity	Windy	Play
1	Sunny	Hot	High	False	No
3	Cloudy	Hot	High	False	Yes
4	Rainy	Mild	High	False	Yes
5	Rainy	Cool	Normal	False	Yes
8	Sunny	Mild	High	False	No
9	Sunny	Cool	Normal	False	Yes
10	Rainy	Mild	Normal	False	Yes
13	Cloudy	Hot	Normal	False	Yes

Tabel faktor false windy terdiri dari **8 instance**. False No = 2, False Yes = 6

$$\begin{aligned}\text{Entropy play|windy=False} &= -(2/8) \cdot \log_2(2/8) - (6/8) \cdot \log_2(6/8) \\ &= -(0.25) \cdot \log_2(0.25) - (0.75) \cdot \log_2(0.75) \\ &= -(0.25) \cdot (-2) - (0.75) \cdot (-0.415) \\ &= 0.50 + 0.311 \\ &= \mathbf{0.811}\end{aligned}$$

Menghitung True Factor windy to Play

No	Outlook	Temperature	Humidity	Windy	Play
2	Sunny	Hot	High	True	No
6	Rainy	Cool	Normal	True	Yes
7	Cloudy	Cool	Normal	True	Yes
11	Sunny	Mild	Normal	True	Yes

12	Cloudy	Mild	High	True	Yes
14	Rainy	Mild	High	True	No

Tabel faktor True windy terdiri dari **6 instance**. True No = 2, True Yes = 4

$$\begin{aligned}
 \text{Entropy play|windy=False} &= -(2/6) \cdot \log_2(2/6) - (4/6) \cdot \log_2(4/6) \\
 &= -(0.333) \cdot \log_2(0.333) - (0.667) \cdot \log_2(0.667) \\
 &= -(0.333) \cdot (-1.586) - (0.667) \cdot (-0.586) \\
 &= 0.528 + 0.390 \\
 &= \mathbf{0.918}
 \end{aligned}$$

Menghitung Gain-Play|Windy

$$\begin{aligned}
 \text{Gain(Play, Windy)} &= \text{Entropy(Play)} - [p(\text{Play|Windy=False}) \cdot \\
 &\text{Entropy(Play|Windy=False)}] - [p(\text{Play|Wind=True}) \cdot \\
 &\text{Entropy(Play|Windy=true)}]
 \end{aligned}$$

$$\begin{aligned}
 &= 0.875 - [(8/14) \cdot 0.811] - (6/14) \cdot 0.918 \\
 &= 0.804
 \end{aligned}$$

Maka Gain(Play, Windy) = **0.804**

Selanjutnya kolom 1

Menghitung Gain(Play, Outlook)

Faktor Outlook Sunny

No	Outlook	Temperature	Humidity	Windy	Play
1	Sunny	Hot	High	False	No
2	Sunny	Hot	High	True	No
8	Sunny	Mild	High	False	No
9	Sunny	Cool	Normal	False	Yes
11	Sunny	Mild	Normal	True	Yes

Tabel faktor Sunny Play terdiri dari **5 instance**. Play No = 3, Play Yes = 2

$$\begin{aligned}
 \text{Entropy play|Outlook=Sunny} &= -(3/5) \cdot \log_2(3/5) - (2/5) \cdot \log_2(2/5) \\
 &= -(0.6) \cdot \log_2(0.6) - (0.4) \cdot \log_2(0.4) \\
 &= -(0.6) \cdot (-0.737) - (0.4) \cdot (-1.322) \\
 &= 0.442 + 0.5288 \\
 &= \mathbf{0.971}
 \end{aligned}$$

Faktor Cloudy Outlook

No	Outlook	Temperature	Humidity	Windy	Play
3	Cloudy	Hot	High	False	Yes
7	Cloudy	Cool	Normal	True	Yes

12	Cloudy	Mild	High	True	Yes
13	Cloudy	Hot	Normal	False	Yes

Tabel faktor Cloudy- Play terdiri dari **4 instance**. Play No = 0, Play Yes = 4
 Entropy play|Outlook=Cloudy = $-(0/4) \cdot \log_2(0/4) - (4/4) \cdot \log_2(4/4)$
 = **0**

Faktor Rainy Outlook

No	Outlook	Temperature	Humidity	Windy	Play
4	Rainy	Mild	High	False	Yes
5	Rainy	Cool	Normal	False	Yes
6	Rainy	Cool	Normal	True	Yes
10	Rainy	Mild	Normal	False	Yes
14	Rainy	Mild	High	True	No

Tabel faktor Rainy Play terdiri dari **5 instance**. Play No = 1, Play Yes = 4
 Entropy play|Outlook=Sunny = $-(1/5) \cdot \log_2(1/5) - (4/5) \cdot \log_2(4/5)$
 = $-(0.2) \cdot \log_2(0.2) - (0.8) \cdot \log_2(0.8)$
 = $-(0.2) \cdot (-2.321) - (0.8) \cdot (-0.322)$
 = $0.464 + 0.2576$
 = **0.722**

Maka Gain (Play,Outlook) = $0.875 - [(5/14) \cdot 0.971] - (0) - (5/14) \cdot 0.722]$
 = 0.787

.....

Selanjutnya Kolom 2 (Temperature)
 Menghitung Gain(play,temperature)

Faktor Hot Temperatur

No	Outlook	Temperature	Humidity	Windy	Play
1	Sunny	Hot	High	False	No
2	Sunny	Hot	High	True	No
3	Cloudy	Hot	High	False	Yes
13	Cloudy	Hot	Normal	False	Yes

Tabel faktor Hot Play terdiri dari **4 instance**. Play No = 2, Play Yes = 2
 Entropy play|Temperature=Hot = $-(2/4) \cdot \log_2(2/4) - (2/4) \cdot \log_2(2/4)$
 = $-(0.5) \cdot \log_2(0.5) - (0.5) \cdot \log_2(0.5)$
 = $-(0.5) \cdot (-1) - (0.5) \cdot (-1)$
 = $0.5 + 0.5$
 = **1.00**

Faktor Mild Temperature

No	Outlook	Temperature	Humidity	Windy	Play
4	Rainy	Mild	High	False	Yes
8	Sunny	Mild	High	False	No
10	Rainy	Mild	Normal	False	Yes
11	Sunny	Mild	Normal	True	Yes
12	Cloudy	Mild	High	True	Yes
14	Rainy	Mild	High	True	No

Tabel faktor Mild Play terdiri dari **6 instance**. Play No = 2, Play Yes = 4
 Entropy play|Temperature=Mild = $-(2/6) \cdot \log_2(2/6) - (4/6) \cdot \log_2(4/6)$
 = $-(0.333) \cdot \log_2(0.333) - (0.667) \cdot \log_2(0.667)$
 = $-(0.333) \cdot (-1.586) - (0.667) \cdot (-0.584)$
 = $0.528 + 0.389$
 = **0.92**

Faktor cold Temperature

No	Outlook	Temperature	Humidity	Windy	Play
5	Rainy	Cool	Normal	False	Yes
6	Rainy	Cool	Normal	True	Yes
7	Cloudy	Cool	Normal	True	Yes
9	Sunny	Cool	Normal	False	Yes

Tabel faktor Cold Play terdiri dari **4 instance**. Play No = 0, Play Yes = 4
 Entropy play|Temperature=Cool = $-(0/4) \cdot \log_2(0/4) - (4/4) \cdot \log_2(4/4)$
 = **0**

$$\begin{aligned}
 \text{Maka Gain (Play, Temperature)} &= 0.875 - [(4/14.1) - (6/14.0.92) - (0)] \\
 &= 0.875 - [(0.286) - (0.395) - (0)] \\
 &= \mathbf{1.065}
 \end{aligned}$$

Selanjutnya menghitung Kolom 3 (Humidity)

Menghitung Gain|Play, Humidity

Faktor High Humidity

No	Outlook	Temperature	Humidity	Windy	Play
1	Sunny	Hot	High	False	No
2	Sunny	Hot	High	True	No
3	Cloudy	Hot	High	False	Yes
4	Rainy	Mild	High	False	Yes
8	Sunny	Mild	High	False	No
12	Cloudy	Mild	High	True	Yes
14	Rainy	Mild	High	True	No

Tabel faktor High Play terdiri dari **7 instance**. Play No = 4, Play Yes = 3
 Entropy play|Outlook=Sunny = $-(4/7) \cdot \log_2(4/7) - (3/7) \cdot \log_2(3/7)$
 $= -(0.571) \cdot \log_2(0.571) - (0.428) \cdot \log_2(0.428)$
 $= -(0.571) \cdot (-0.808) - (0.428) \cdot (-1.224)$
 $= 0.461 + 0.524$
 $= \mathbf{0.985}$

Faktor Normal Humidity

No	Outlook	Temperature	Humidity	Windy	Play
5	Rainy	Cool	Normal	False	Yes
6	Rainy	Cool	Normal	True	Yes

7	Cloudy	Cool	Normal	True	Yes
9	Sunny	Cool	Normal	False	Yes
10	Rainy	Mild	Normal	False	Yes
13	Cloudy	Hot	Normal	False	Yes

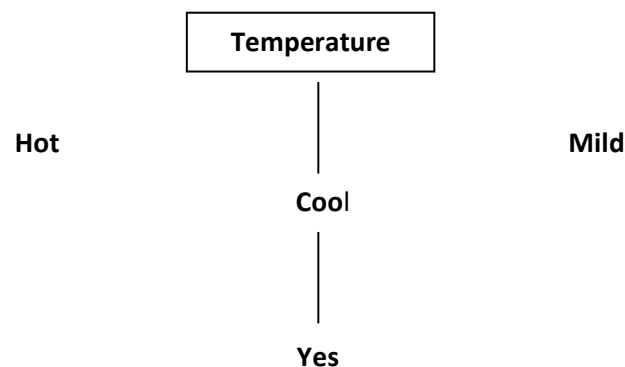
Tabel faktor Normal Play terdiri dari 6 **instance**. Play No = 0, Play Yes = 6
Entropy play|Outlook=Sunny = $-(0/6) \cdot \log_2(0/6) - (6/6) \cdot \log_2(6/6)$
= 0

Maka Gain (**Play, Humadity**) = $0.875 - [(7/14) \cdot 0.985] - (0)$
= $0.875 - 0.492$
= **0.383**

Dari perhitungan diatas didapat

1. Gain (**Play, Windy**) = 0.804
2. Gain (**Play, Outlook**) = 0.787
3. Gain (**Play, Temperature**) = 1.605
4. Gain (**Play, Humadity**) = 0.383

Karena nilai temperatur paling besar, maka pohon keputusan paling atas adalah faktor Temperature.



Cool Temperatur On Play

Dari tabel dibawah, Keputusan untuk bermain diambil jika temperatur dingin/cool

No	Outlook	Temperature	Humidity	Windy	Play
5	Rainy	Cool	Normal	False	Yes
6	Rainy	Cool	Normal	True	Yes
7	Cloudy	Cool	Normal	True	Yes
9	Sunny	Cool	Normal	False	Yes

Hot Temperatur On Play.

No	Outlook	Temperature	Humidity	Windy	Play
1	Sunny	Hot	High	False	No
2	Sunny	Hot	High	True	No
3	Cloudy	Hot	High	False	Yes
13	Cloudy	Hot	Normal	False	Yes

Mencari nilai hubungan Outlook-Temperatur / Gain (temp,Outlook)

Entropy Hot Temperatur = 2/4 no dan 2/4 yes

$$\text{Nilai Entrophy} = -(2/4) \cdot \log_2(2/4) - (2/4) \cdot \log_2(2/4) \\ = 1.00$$

Mencari entropy Sunny

karena, instance sunny = 2, nilai play No = 2 dan nilai Play yes = 0, maka
Entropy Sunny = 0

Mencari entropy Cloudy

Karena instance cloudy = 2, nilai Play No = 0 dan nilai play yes = 2, maka
Entropy Cloudy = 0

$$\text{Maka Gain(Outlook-Temperature)} = 1.00 - (2/4 \cdot 0) - (2/4 \cdot 0) \\ = 1.00$$

Mencari nilai hubungan Humidity-Temperatur / Gain (Temp,Hum)

Entropy High terdiri dari 3 instance dengan nilai play no = 2, yes = 1.

$$\text{Entropy High} = -(2/3) \cdot \log_2(2/3) - (1/3) \cdot \log_2(1/3)$$

$$\begin{aligned}
&= -(0.667) \cdot \log(0.667) - (0.333) \cdot \log(0.333) \\
&= -(0.667) \cdot (-0.584) - (0.333) \cdot (-1.586) \\
&= 0.389 + 0.528 \\
&= 0.917
\end{aligned}$$

$$\begin{aligned}
\text{Entropy Normal} &= -(0/1) \cdot \log_2(0/1) - (1/1) \cdot \log_2(1/1) \\
&= 0
\end{aligned}$$

$$\begin{aligned}
\text{Maka Gain(Humadity-Temperature)} &= 1.00 - (3/4)(0.971) - (1/4 \cdot 0) \\
&= 1.00 - 0.728 \\
&= 0.272
\end{aligned}$$

Mencari nilai hubungan Windy-Temperatur / Gain (Temp,Windy)

Entropy False terdiri dari 3 instance, dengan nilai play no = 1 dan yes = 2.

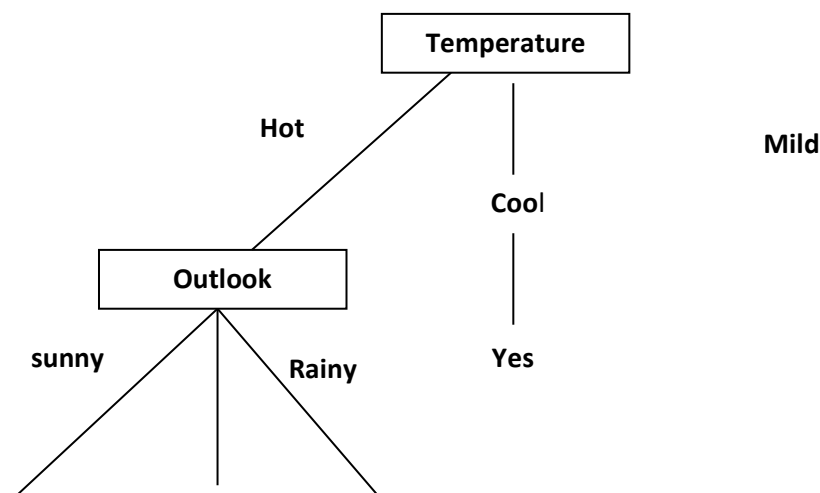
$$\begin{aligned}
\text{Entropy False} &= -(1/3) \cdot \log_2(1/3) - (2/3) \cdot \log_2(2/3) \\
&= -(0.333) \cdot \log(0.333) - (0.667) \cdot \log(0.667) \\
&= -(0.333) \cdot (-1.586) - (0.667) \cdot (-0.584) \\
&= 0.528 + 0.389 \\
&= 0.917
\end{aligned}$$

$$\begin{aligned}
\text{Entropy True, dengan nilai play No} &= 1, \text{ play yes} = 0 \\
\text{Entropy True} &= 0
\end{aligned}$$

$$\begin{aligned}
\text{Maka Gain(Windy-Temperature)} &= 1.00 - (3/4)(0.971) - (1/4 \cdot 0) \\
&= 1.00 - 0.728 \\
&= 0.272
\end{aligned}$$

Maka nilai hubungan antara Temperature Hot dengan Outlook, Humidity dan windi adalah

$$\begin{aligned}
\text{Maka Gain(Outlook-Temperature=hot)} &= 1.00 \\
\text{Maka Gain(Humadity-Temperature=hot)} &= 0.272 \\
\text{Maka Gain(Windy-Temperature =hot)} &= 0.272
\end{aligned}$$



Cloudy

No **Yes** **Yes**

Mild Temperatur On Play.

No	Outlook	Temperature	Humidity	Windy	Play
4	Rainy	Mild	High	False	Yes
8	Sunny	Mild	High	False	No
10	Rainy	Mild	Normal	False	Yes
11	Sunny	Mild	Normal	True	Yes
12	Cloudy	Mild	High	True	Yes
14	Rainy	Mild	High	True	No

Mencari hubungan **Mild Temperatur On Play**

Mencari nilai hubungan Outlook-Temperatur / Gain (temp,Outlook)

Entropy **Mild**Temperatur = 2/6 No dan 4/6 yes

$$\begin{aligned}
 \text{Nilai Entrophy} &= -(2/6) \cdot \log_2(2/6) - (4/6) \cdot \log_2(4/6) \\
 &= -(0.333) \cdot \log_2(0.333) - (0.667) \cdot \log_2(0.667) \\
 &= 0.917
 \end{aligned}$$

Mencari entropy Rainy

Instance = 3, Play No = 1, Play Yes = 2

$$\begin{aligned}
 \text{Entropy Rainy} &= -(1/3) \cdot \log_2(1/3) - (2/3) \cdot \log_2(2/3) \\
 &= -(0.333) \cdot (0.333) - (0.667) \cdot \log_2(0.667) \\
 &= 0.917
 \end{aligned}$$

Mencari entropy Sunny

Instance = 2, Play No = 1, Play Yes = 1

$$\begin{aligned}
 \text{Entropy Rainy} &= -(1/2) \cdot \log_2(1/2) - (1/2) \cdot \log_2(1/2) \\
 &= -(0.5) \cdot \log_2(0.5) - (0.5) \cdot \log_2(0.5)
 \end{aligned}$$

$$\begin{aligned}
&= -(0.5).(-1.0) - (0.5).\log(-1.0) \\
&= 0.5 + 0.5 \\
&= 1.00
\end{aligned}$$

Mencari entropy Cloudy

Instance = 1, Play No = 0, Play Yes = 1

$$\begin{aligned}
\text{Entropy Rainy} &= -(0/1).\log_2(0/1) - (1/1).\log_2(1/1) \\
&= 0
\end{aligned}$$

$$\begin{aligned}
\text{Maka Gain(Outlook|Temperature=mild)} &= 0.917 - [(3/6).(0.917) - (2/6).1 - (0)] \\
&= 0.917 - [(0.458) - (0.667) - (0)] \\
&= 0.917 - (-0.209) \\
&= 1.126
\end{aligned}$$

Mencari hubungan **Mild Temperatur On Play. (temp-humadity)**

Mencari nilai hubungan Humidity-Temperatur / Gain (temp,humidity)

Mencari entropy High

Instance = 4, Play No = 2, Play Yes = 2

$$\begin{aligned}
\text{Entropy Rainy} &= -(2/4).\log_2(2/4) - (2/4).\log_2(2/4) \\
&= -(0.5).(-1.00) - (0.5).(-1.0) \\
&= 1.00
\end{aligned}$$

Mencari entropy Normal

Instance = 2, Play No = 0, Play Yes = 2

$$\begin{aligned}
\text{Entropy Rainy} &= -(0/2).\log_2(0/2) - (2/2).\log_2(2/2) \\
&= 0
\end{aligned}$$

$$\begin{aligned}
\text{Maka Gain(Humadity|Temperature=mild)} &= 0.917 - [(4/6).(1) - (0)] \\
&= 0.917 - [(0.667) - (0)] \\
&= 0.917 - 0.667 \\
&= 0.25
\end{aligned}$$

--

Mencari hubungan **Mild Temperatur On Play. (temp-windy)**

Mencari nilai hubungan Windy-Temperatur / Gain (temp,windy)

Mencari entropy False pada Windy

Instance = 3, Play No = 1, Play Yes = 2

$$\begin{aligned}
\text{Entropy False} &= -(1/3).\log_2(1/3) - (2/3).\log_2(2/3) \\
&= -(0.333).(-1.586) - (0.667).(-0.584) \\
&= 0.528 + 0.389 \\
&= 0.917
\end{aligned}$$

Mencari entropy True pada Windy

Instance = 3, Play No = 1 , Play Yes = 2

$$\begin{aligned}\text{Entropy True} &= -(1/3) \cdot \log_2(1/3) - (2/3) \cdot \log_2(2/3) \\ &= -(0.333) \cdot (-1.586) - (0.667) \cdot (-0.584) \\ &= 0.528 + 0.389 \\ &= 0.917\end{aligned}$$

$$\begin{aligned}\text{Maka Gain(Windy|Temperature=mild)} &= 0.917 - [(3/6) \cdot (0.917) - ((3/6) \cdot (0.917))] \\ &= 0.917 - (0) \\ &= 0.917\end{aligned}$$

Maka nilai hubungan antara Temperature Mild dengan Outlook, Humidity dan windy adalah

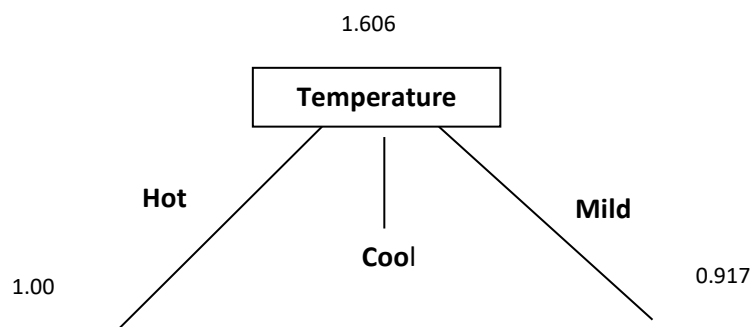
$$\text{Maka Gain(Outlook-Temperature=mild)} = 0.917$$

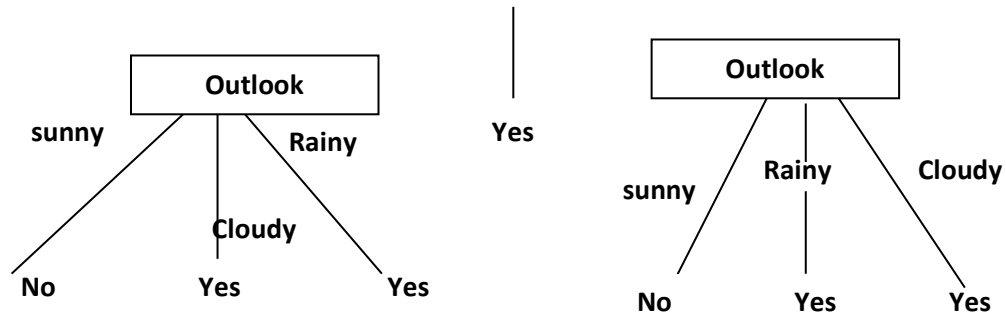
$$\text{Maka Gain(Humadity-Temperature=mild)} = 0.25$$

$$\text{Maka Gain(Windy-Temperature =mild)} = 0.917$$

Terdapat 2 nilai yang sama, yakni pada **outlook dan Windy**, namun kita bisa cek pada tabel berikut bahwa keputusan untuk bermain pada temperature mild lebih dipengaruhi oleh **Outlook**

No	Outlook	Temperature	Humidity	Windy	Play
4	Rainy	Mild	High	False	Yes
8	Sunny	Mild	High	False	No
10	Rainy	Mild	Normal	False	Yes
11	Sunny	Mild	Normal	True	Yes
12	Cloudy	Mild	High	True	Yes
14	Rainy	Mild	High	True	No





Kesimpulan :

Kelebihan

1. Algoritma ini mampu mentransformasikan dari data mentah menjadi data yang beraturan.
2. Dapat menggunakan atribut nominal dimana kebanyakan tidak dapat dilakukan oleh algoritma machine learning.

Kekurangan

1. Terlalu panjang/lama waktu untuk mencari hubungan satu faktor dengan faktor lainnya.
2. Cenderung kaku dan dapat menimbulkan kesalahan perhitungan.

Metode pembandingan selain menggunakan metode ini dapat dilakukan dengan metode chi-square / CHAID atau Random Forest.

NAMA : AHMAD ALI MA'MUN

NIM : 202420037

No	Outlook	Temperature	Humidity	Windy	Play
1	Sunny	Hot	High	FALSE	No
2	Sunny	Hot	High	TRUE	No
3	Cloudy	Hot	High	FALSE	Yes
4	Rainy	Mild	High	FALSE	Yes
5	Rainy	Cool	Normal	FALSE	Yes
6	Rainy	Cool	Normal	TRUE	Yes
7	Cloudy	Cool	Normal	TRUE	Yes
8	Sunny	Mild	High	FALSE	No
9	Sunny	Cool	Normal	FALSE	Yes
10	Rainy	Mild	Normal	FALSE	Yes
11	Sunny	Mild	Normal	TRUE	Yes
12	Cloudy	Mild	High	TRUE	Yes
13	Cloudy	Hot	Normal	FALSE	Yes
14	Rainy	Mild	High	TRUE	No

KETERANGAN :

Outlook, temperature, humidity dan windy adalah Atrribute

Dan Play adalah keputusan

Dari data diatas terdapat 14 kasus

Maka :

Setelah kita ketahui data atribut, keputusan dan jumlah kasusnya maka kita cari entropy-nya dengan rumus sebagai berikut :

$$\text{Entropy (S)} = \sum_{i=1}^n - p_i * \log_2 p_i$$

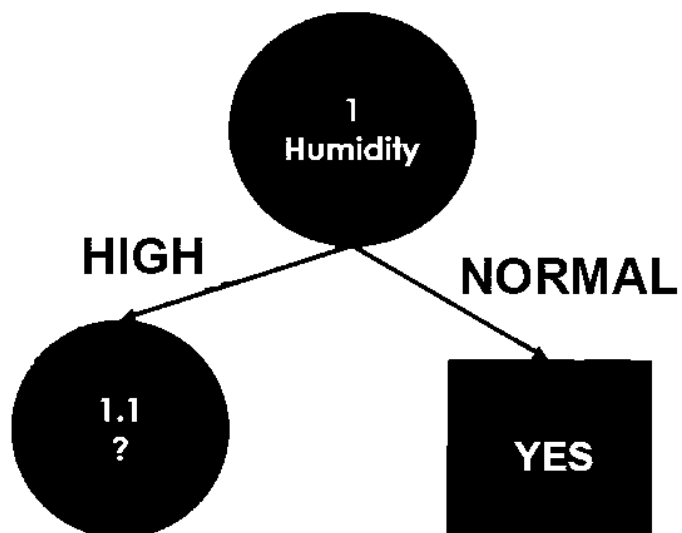
Lalu kita hitung gain-nya

$$\text{Gain (S,A)} = \text{Entropy (S)} - \sum_{i=1}^n \text{ * Entropy (Si)}$$

Hasil perhitungan 1

		Jumlah kasus(S)	No (S1)	Yes (S2)	Entropy	Gain
Total		14	4	10	0,86312	
outlook						0,258512
	Cloudy	4	0	4	0	
	Rainy	5	1	4	0,721928	
	Sunny	5	3	2	0,970950	
Temperatur						0,183850
	Cool	4	0	4	0	
	Hot	4	2	2	1	
	Mild	6	2	4	0,918295	
Humidity						0,370506
	High	7	4	3	0,985228	
	Normal	7	0	7	0	
Windy						0,005977
	False	8	2	6	0,811278	
	True	6	4	2	0,918295	

Maka dapat digambarkan dalam pohon data :



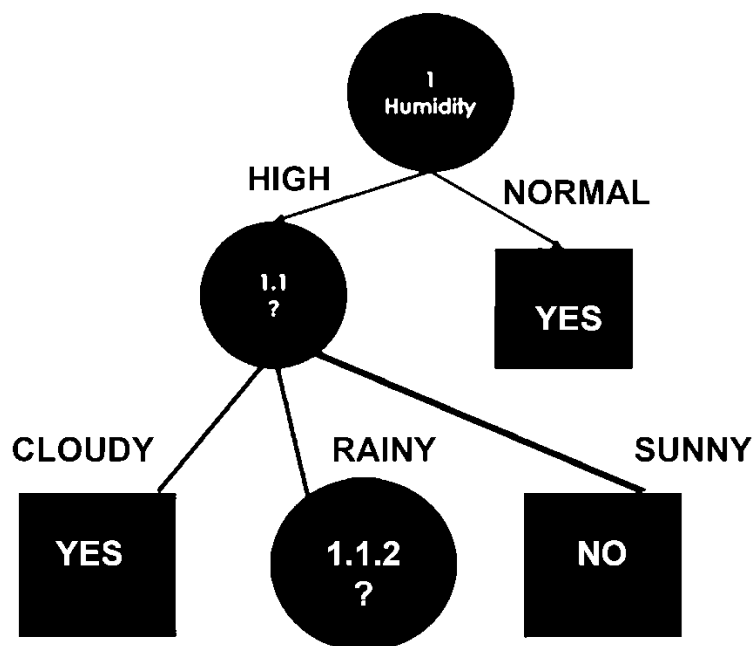
Karena nilai Humadity High masih mempunyai nilai semua maka masih perlu diperhitungkan lebih lanjut ke perhitungan 1.1

Hasil perhitungan 1.1

		Jumlah kasus(S)	No (S1)	Yes (S2)	Entropy	Gain
Humadity Hlgh		7	4	3	0,985228	
outlook						0,699513
	Cloudy	2	0	2	0	
	Rainy	2	1	1	1	
	Sunny	3	3	0	0	
Temperatur						0,020244
	Cool	0	0	0	0	
	Hot	3	2	1	0,918295	
	Mild	4	2	2	1	
Windy						0,020244
	False	4	2	2	1	
	True	3	2	1	0,918295	

Karena Outlook mempunyai 3 nilai Cloudy, Rainy, dan Sunny maka ia mempunyai 3 cabang pohon data yang dapat digambarkan sebagai berikut :

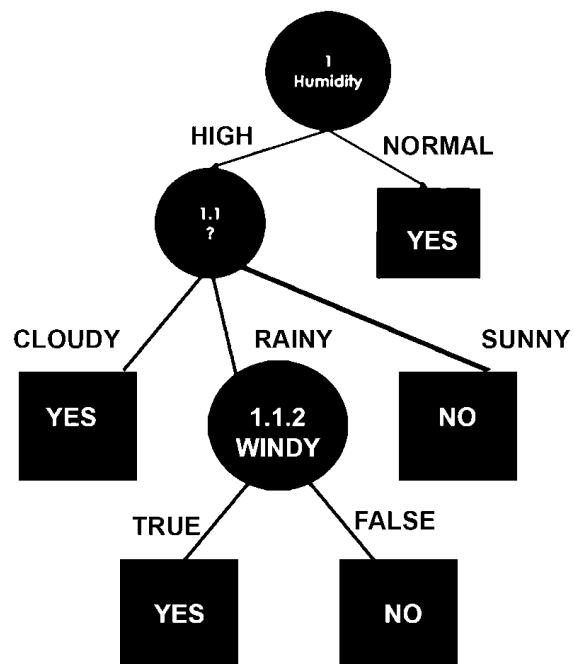
Karena rainy masih mepunyai nilai yang belum memiliki klasifikasi maka diperlukan perhitungan lebh lanjut



Hasil perhitungan Humidity High and Outlook Rainy

		Jumlah kasus(S)	No (S1)	Yes (S2)	Entropy	Gain
Humadity High and outlook Rainy		2	1	1	1	
Temperatur						0
	Cool	0	0	0	0	
	Hot	0	0	0	0	
	Mild	2	1	1	1	
Windy						1
	False	1	0	1	0	
	True	1	1	0	0	

Dari data tersebut dapat digambarkan dalam pohon data sebagai berikut:



Tugas 05

Dari dataset berikut ini, silahkan buat perhitungan manual dari algoritma decision tree C45 secara lengkap dan gambarkan pohon keputusan yang diperoleh.

No	Outlook	Temperature	Humidity	Windy	Play
1	Sunny	Hot	High	FALSE	No
2	Sunny	Hot	High	TRUE	No
3	Cloudy	Hot	High	FALSE	Yes
4	Rainy	Mild	High	FALSE	Yes
5	Rainy	Cool	Normal	FALSE	Yes
6	Rainy	Cool	Normal	TRUE	Yes
7	Cloudy	Cool	Normal	TRUE	Yes
8	Sunny	Mild	High	FALSE	No
9	Sunny	Cool	Normal	FALSE	Yes
10	Rainy	Mild	Normal	FALSE	Yes
11	Sunny	Mild	Normal	TRUE	Yes
12	Cloudy	Mild	High	TRUE	Yes
13	Cloudy	Hot	Normal	FALSE	Yes
14	Rainy	Mild	High	TRUE	No

Penyelesaian:

LANGKAH 1

- Menghitung jumlah kasus, jumlah keputusan **Yes** ataupun **No**
- Menghitung Entropy dari semua kasus yg terbagi berdasarkan atribut “Outlook”, “Temperature”, “Humidity”, “Windy”.
- Penghitungan Gain untuk setiap atributnya

Node			Jumlah Kasus	Tidak	Ya	Entropy	Gain
1	TOTAL		14	4	10	0,863120569	
	OUTLOOK						0,258521037
		CLOUDY	4	0	4	0	
		RAINY	5	1	4	0,721928095	
		SUNNY	5	3	2	0,970950594	
	TEMPERATURE						0,183850925
		COOL	4	0	4	0	
		HOT	4	2	2	1	
		MILD	6	2	4	0,918295834	
	HUMIDITY						0,370506501
		HIGH	7	4	3	0,985228136	

		NORMAL	7	0	7	0	
	WINDY						0,005977711
		FALSE	8	2	6	0,811278124	
		TRUE	6	2	4	0,918295834	

Perhitungan Total Entropy

$$\text{Entropy (S)} = - \sum_{i=1}^n p_i * \log_2 p_i$$

$$\text{Entropy(Total)} = -4/14 * \log_2 (4/14) + -10/14 * \log_2 (10/14)$$

Menghitung gain pada baris outlook

Gain (total, outlook)

$$\begin{aligned} \text{Gain (S,A)} &= \text{Entropy (S)} - \sum_{i=1}^n p_i * \text{Entropy (Si)} \\ &= 0,863120569 - ((4/14*0) + (5/14*0,72) + (5/14*0,97)) \\ &= 0,258521037 \end{aligned}$$

Menghitung temperature, humidity, dan windy

- Atribut dengan gain tertinggi adalah Humidity = 0,37, maka humidity menjadi node akar
- Humidity memiliki 2 nilai yaitu
 1. HIGH
Humidity HIGH masih perlu dilakukan perhitungan karena masih terdapat “yes” dan “no”
 2. NORMAL
Humidity NORMAL sudah mengklarifikasikan kasus menjadi 1 yaitu “yes”

LANGKAH 2

Node			Jumlah Kasus	Tidak	Ya	Entropy	Gain
1.1	HUMIDITY HIGH		7	4	3	0,985228136	
	OUTLOOK						0,69951385
		CLOUDY	2	0	2	0	
		RAINY	2	1	1	1	
		SUNNY	3	3	0	0	
	TEMPERATURE						0,020244207
		COOL	0	0	0	0	
		HOT	3	2	1	1	
		MILD	4	2	2	2	
	WINDY						
		FALSE	4	2	2	1	
		TRUE	3	2	1	0,918295834	

Gain Tertinggi = Outlook = 0,69

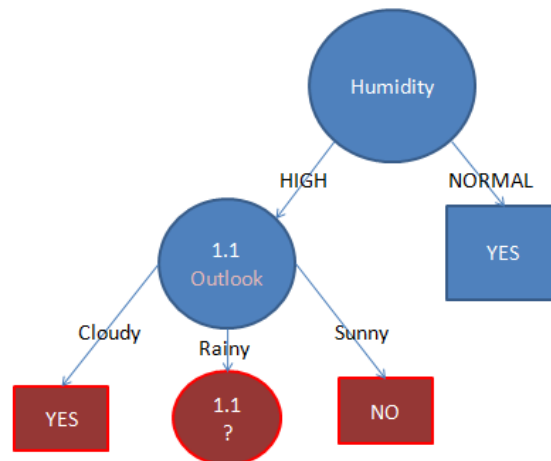
Maka **Outlook** menjadi node cabang dari humidity yang bernilai **High**.

Berdasarkan atribut **Outlook** terdapat 3 nilai yaitu:

1. Cloudy
2. Rainy
3. Sunny

Karena **Cloudy** pasti bernilai **Yes**, **Sunny** pasti bernilai **No**, maka tidak perlu dilakukan perhitungan lagi.

Pohon keputusan node 1.1



LANGKAH 3

Node			Jumlah Kasus	Tidak	Ya	Entropy	Gain
1.1	HUMIDITY HIGH dan OUTLOOK RAINY		2	1	1	1	
	TEMPERATURE						0
		COOL	0	0	0	0	
		HOT	0	0	0	0	
		MILD	2	1	1	1	
	WINDY						1
		FALSE	1	0	1	0	
		TRUE	1	1	0	0	

Gain tertinggi = Windy = 1

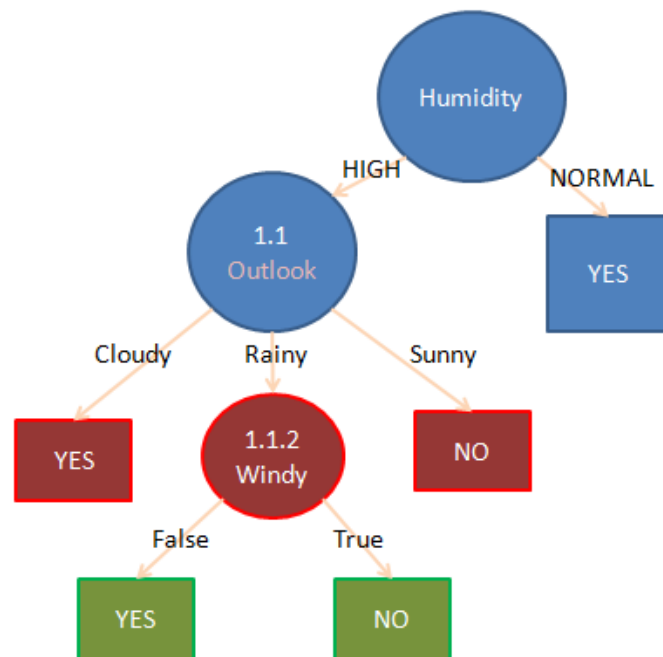
Maka **Windy** menjadi node cabang atribut humidity yang bernilai **High** dan outlook yang bernilai **Rainy**.

Berdasarkan atribut **Windy** terdapat 2 nilai, yaitu:

1. **True**, karena sudah terklarifikasi pasti bernilai **No**
2. **False**, pasti bernilai **Yes**.

Maka tidak perlu dilakukan perhitungan lagi.

Pohon keputusan node 1.1.2



Berdasarkan node 1.1.2, maka:

Semua kasus sudah masuk dapat kelasnya sehingga pohon keputusan diatas merupakan pohon keputusan terakhir yang terbentuk.

Penjelasan kekurangan dan kelebihan Algoritma C4.5

Kelebihan Algoritma C4.5	Kekurangan Algoritma C4.5
C4.5 mampu menangani atribut yang kosong (Missing Value).	Algoritma C4.5 hanya dapat digunakan untuk menangani sampel-sampel yang dapat disimpan secara keseluruhan dan pada waktu yang bersamaan di memori.
C4.5 mampu menangani atribut dengan kontinu.	
C4.5 memangkas pohon keputusan untuk mengatasi overfitting.	

Perbandingan Algoritma Decision Tree dengan SVM (Support Vector Machine Multiclass)

- Decision Tree data dibangun berdasarkan objek yang dimiliki.
- Sedangkan SVM harus menyimpan sebagian kecil data untuk digunakan kembali pada saat prediksi.

Nama : Andry Meylani

NIM : 202420009

Tugas 05

Dari dataset berikut ini, silahkan buat perhitungan manual dari algoritma decision tree C45 secara lengkap dan gambarkan pohon keputusan yang diperoleh.

No	Outlook	Temperature	Humidity	Windy	Play
1	Sunny	Hot	High	FALSE	No
2	Sunny	Hot	High	TRUE	No
3	Cloudy	Hot	High	FALSE	Yes
4	Rainy	Mild	High	FALSE	Yes
5	Rainy	Cool	Normal	FALSE	Yes
6	Rainy	Cool	Normal	TRUE	Yes
7	Cloudy	Cool	Normal	TRUE	Yes
8	Sunny	Mild	High	FALSE	No
9	Sunny	Cool	Normal	FALSE	Yes
10	Rainy	Mild	Normal	FALSE	Yes
11	Sunny	Mild	Normal	TRUE	Yes
12	Cloudy	Mild	High	TRUE	Yes
13	Cloudy	Hot	Normal	FALSE	Yes
14	Rainy	Mild	High	TRUE	No

Kumpulkan dalam format MS Word. Sertakan penjelasan dan pendapat anda tentang algoritma ini, kelebihan dan kekurangannya berbanding metode lain yang sejenis (tentukan sendiri metode pembandingnya).

1. Perhitungan manual dari algoritma decision tree C45

Secara umum algoritma C4.5 untuk membangun pohon keputusan adalah sebagai berikut :

- Pilih atribut sebagai akar.
- Buat cabang untuk tiap-tiap nilai.
- Bagi kasus dalam cabang.
- Ulangi proses untuk setiap cabang sampai semua kasus pada cabang memiliki kelas yang sama.

Perhitungan Gain

$$\text{Gain}(S,A) = \text{Entropy}(S) - \sum_{i=1}^n \frac{|S_i|}{|S|} * \text{Entropy}(S_i)$$

Keterangan:

- S : himpunan
- A : atribut
- n : jumlah partisi atribut A
- | S_i | : jumlah kasus pada partisi ke-i
- | S | : jumlah kasus dalam S

Menghitung Nilai Entropy

$$\text{Entropy}(S) = - \sum_{i=1}^n p_i * \log_2 p_i$$

Keterangan:

- S : himpunan kasus
- A : fitur
- n : jumlah partisi S
- p_i : proporsi dari S_i terhadap S

Langkah 1. Perincian Algoritma

- Menghitung jumlah kasus seluruhnya, jumlah berkeputusan “Yes” maupun “No”.
- Menghitung Entropy dari semua kasus yg terbagi berdasarkan atribut “Outlook”, “Temperature”, “Humidity”, “Windy”.
- Lakukan penghitungan Gain utk setiap atributnya

Perhitungan :

$$\begin{aligned} \text{Entropy (Total)} &= \left(-\frac{4}{14} \times \log_2 \left(\frac{4}{14} \right) \right) + \left(-\frac{10}{14} \times \log_2 \left(\frac{10}{14} \right) \right) \\ &= \mathbf{0,86312} \end{aligned}$$

Outlook

$$\begin{aligned}\text{Gain} &= 0,86312 - \left(\frac{4}{14} \times 0 \right) + \left(\frac{5}{14} \times (0,72193) + \frac{5}{14} \times (0,97095) \right) \\ &= 0,86312 - (0 + 0,25783 + 0,34676) \\ &= 0,86312 - 0,60459 \\ &= 0,25853\end{aligned}$$

Temperature

$$\begin{aligned}\text{Gain} &= 0,86312 - \left(\frac{4}{14} \times 0 \right) + \left(\frac{4}{14} \times 1 \right) + \left(\frac{6}{14} \times 0,9183 \right) \\ &= 0,86312 - (0 + 0,28571 + 0,39355) \\ &= 0,86312 - 0,67926 \\ &= 0,18386\end{aligned}$$

Humidity

$$\begin{aligned}\text{Gain} &= 0,86312 - \left(\frac{7}{14} \times 0,98523 \right) + \left(\frac{7}{14} \times 0 \right) \\ &= 0,86312 - (0,49261 + 0) \\ &= 0,86312 - 0,49261 \\ &= 0,37051\end{aligned}$$

Windy

$$\begin{aligned}\text{Gain} &= 0,86312 - \left(\frac{8}{14} \times 0,81128 \right) + \left(\frac{6}{14} \times 0,9183 \right) \\ &= 0,86312 - (0,46358 + 0,39355) \\ &= 0,86312 - 0,85713 \\ &= 0,00599\end{aligned}$$

		Jumlah Kasus (S)	No (S ₁)	Yes (S ₂)	Entropy	Gain
Total		14	4	10	0,86312	
Outlook						0,25853
	Cloudy	4	0	4	0	
	Rainy	5	1	4	0,72193	
	Sunny	5	3	2	0,97095	
Temp						0,18386
	Cool	4	0	4	0	
	Hot	4	2	2	1	
	Mild	6	2	4	0,9183	
Humidity						0,37051
	High	7	4	3	0,98523	
	Normal	7	0	7	0	
Windy						0,00599
	False	8	2	6	0,81128	
	True	6	4	2	0,9183	

- Dari hasil diketahui bahwa atribut dengan gain tertinggi adalah HUMIDITY yaitu sebesar 0.37051. Sehingga HUMIDITY dapat menjadi node akar.
- Ada dua nilai atribut dari HUMIDITY, yaitu HIGH dan NORMAL.
- Nilai atribut NORMAL sudah mengklasifikasikan kasus menjadi 1, yaitu keputusannya Yes, sehingga tidak perlu dilakukan perhitungan lebih lanjut.
- Tetapi untuk nilai HIGH masih perlu dilakukan perhitungan lagi.

Langkah 2.

- Menghitung jumlah kasus, jumlah kasus untuk keputusan Yes, jumlah kasus untuk keputusan No.
- Entropy dari semua kasus dan kasus yang dibagi berdasarkan atribut OUTLOOK, TEMPERATURE dan WINDY, yang dapat menjadi node akar dari nilai atribut HIGH.
- Setelah itu lakukan perhitungan Gain, untuk tiap-tiap atribut.

Perhitungan :

$$\begin{aligned}\text{Entropy (Total)} &= \left(-\frac{4}{7} \times \log_2 \left(\frac{4}{7}\right)\right) + \left(-\frac{3}{7} \times \log_2 \left(\frac{3}{7}\right)\right) \\ &= \mathbf{0,98522}\end{aligned}$$

Outlook

$$\begin{aligned}\text{Gain} &= 0,98522 - \left(\frac{2}{7} \times 0\right) + \left(\frac{2}{7} \times (1) + \frac{3}{7} \times (0)\right) \\ &= 0,98522 - (0 + 0,28571 + 0) \\ &= 0,98522 - 0,28571 \\ &= \mathbf{0,69951}\end{aligned}$$

Temperature

$$\begin{aligned}\text{Gain} &= 0,98522 - \left(\frac{0}{7} \times 0\right) + \left(\frac{3}{7} \times (0,9183) + \frac{4}{7} \times (1)\right) \\ &= 0,98522 - (0 + 0,39355 + 0,57142) \\ &= 0,98522 - 0,96497 \\ &= \mathbf{0,02025}\end{aligned}$$

Windy

$$\begin{aligned}\text{Gain} &= 0,98522 - \left(\frac{4}{7} \times 1\right) + \left(\frac{3}{7} \times (0,9183)\right) \\ &= 0,98522 - (0,57142 + 0,39355) \\ &= 0,98522 - 0,96497 \\ &= \mathbf{0,02025}\end{aligned}$$

		Jumlah Kasus (S)	No (S ₁)	Yes (S ₂)	Entropy	Gain
Humidity		7	4	3	0,98522	
High						
Outlook						0,69951
	Cloudy	2	0	2	0	
	Rainy	2	1	1	1	
	Sunny	3	3	0	0	
Temp						0,02025
	Cool	0	0	0	0	
	Hot	3	2	1	0,9183	
	Mild	4	2	2	1	
Windy						0,02025
	False	4	2	2	1	
	True	3	2	1	0,9183	

- Atribut dengan Gain tertinggi adalah OUTLOOK, yaitu sebesar 0.6995.
- Sehingga OUTLOOK dapat menjadi node cabang dari nilai atribut HIGH.
- Ada tiga nilai dari atribut OUTLOOK yaitu CLOUDY, RAINY dan SUNNY.

- CLOUDY : klasifikasi kasus 1 (Yes)
- SUNNY : klasifikasi kasus 1 (No)
- RAINY : masih perlu perhitungan lagi.

Langkah 3.

- Menghitung jumlah kasus, jumlah kasus untuk keputusan Yes, jumlah kasus untuk keputusan No.
- Entropy dari semua kasus dan kasus yang dibagi berdasarkan atribut TEMPERATURE dan WINDY, yang dapat menjadi node cabang dari nilai atribut RAINY.
- Setelah itu lakukan perhitungan Gain, untuk tiap-tiap atribut.

Perhitungan :

$$\text{Entropy (Total)} = \left(-\frac{1}{2} \times \log_2 \left(\frac{1}{2} \right) \right) + \left(-\frac{1}{2} \times \log_2 \left(\frac{1}{2} \right) \right) = 1$$

Temperature

$$\begin{aligned} \text{Gain} &= 1 - \left(\frac{0}{2} \times 0 \right) + \left(\frac{0}{2} \times (0) + \frac{2}{2} \times (1) \right) \\ &= 1 - 1 \\ &= 0 \end{aligned}$$

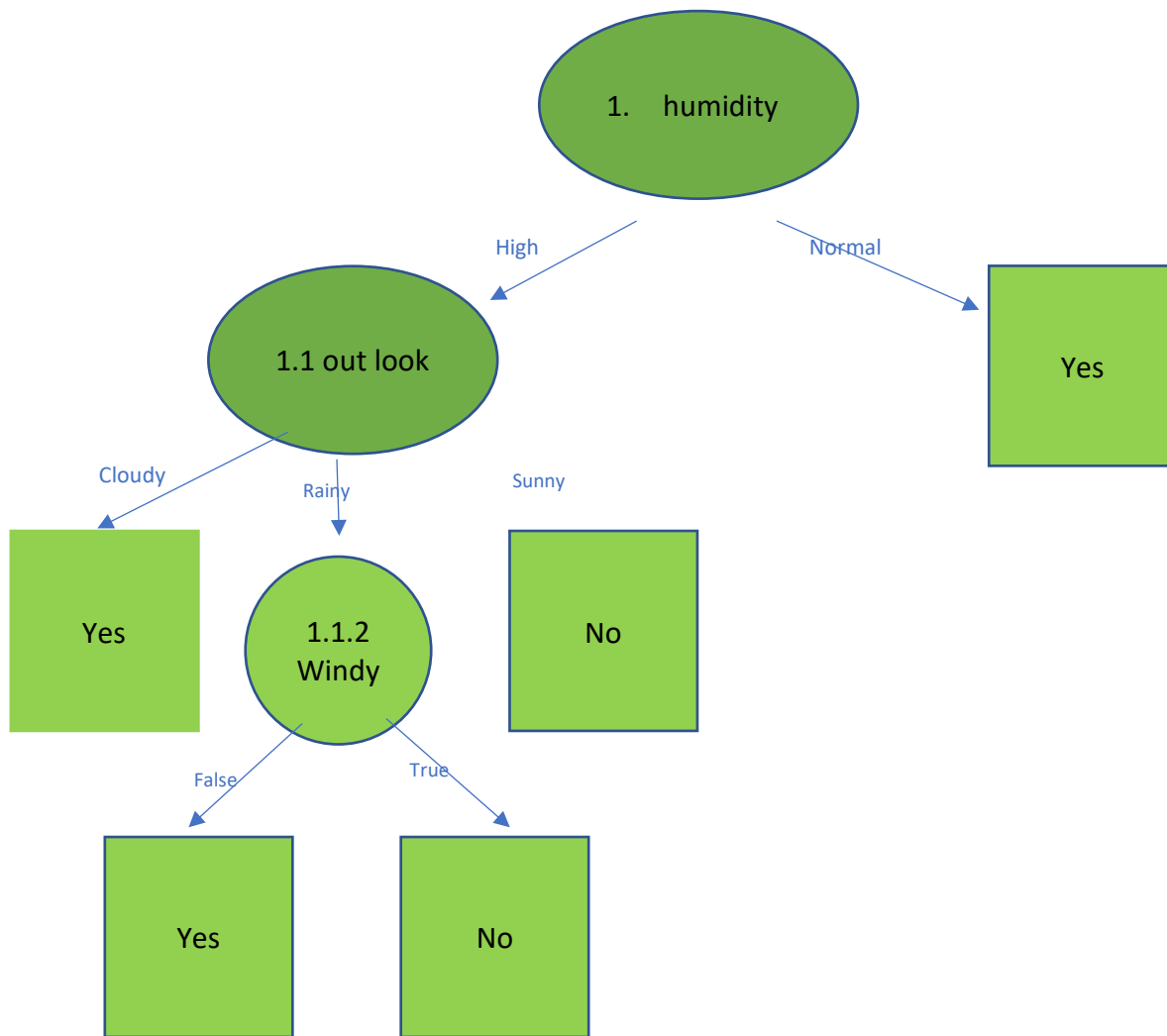
Windy

$$\begin{aligned}\text{Gain} &= 1 - \left(\frac{1}{2} \times 0\right) + \left(\frac{1}{2} \times 0\right) \\ &= 1 - 0 \\ &= \mathbf{1}\end{aligned}$$

		Jumlah Kasus (S)	No (S ₁)	Yes (S ₂)	Entropy	Gain
Humidity High and Outlook Rainy		2	1	1	1	
Temp						0
	Cool	0	0	0	0	
	Hot	0	0	0	0	
	Mild	2	1	1	1	
Windy						1
	False	1	0	1	0	
	True	1	1	0	0	

- Atribut dengan Gain tertinggi adalah WINDY, yaitu sebesar 1.
- Sehingga WINDY dapat menjadi node cabang dari nilai atribut RAINY.
- Ada dua nilai dari atribut WINDY, yaitu FALSE dan TRUE.

- Nilai atribut FALSE sudah mengklasifikasikan kasus menjadi 1 (Yes).
- Nilai atribut TRUE sudah mengklasifikasikan kasus menjadi 1 (No).
- Sehingga tidak perlu dilakukan perhitungan lagi.



Pendapat saya mengenai algoritma ini yaitu Algoritma data mining C4.5 merupakan salah satu algoritma yang digunakan untuk melakukan klasifikasi atau segmentasi atau pengelompokan dan bersifat prediktif. Klasifikasi merupakan salah satu proses pada data mining yang bertujuan untuk menemukan pola yang berharga dari data yang berukuran relatif besar hingga sangat besar.

Berikut perbandingan Algoritma C4.5 dengan Algoritma ID3 beserta kelebihan dan kekurangannya,

a. **Mampu Menangani Atribut dengan Tipe Diskrit atau Kontinu**

Pemilihan atribut pada algoritma induksi decision tree menggunakan ukuran berdasarkan entropy yang dikenal dengan information gain sebagai sebuah heuristic untuk memilih atribut yang merupakan bagian terbaik dari contoh ke dalam kelas. Semua atribut adalah bersifat kategori yang bernilai diskrit. Atribut dengan nilai *continuous* harus didiskritkan.

Outlook	Temperature	Humidity	Windy	Play
Sunny	85	85	False	No
Sunny	80	90	True	No
Overcast	83	86	False	Yes
Rainy	75	80	False	Yes
...	...	...	...	...

Continuous attributes

Diskritisasi atribut bertujuan untuk mempermudah pengelompokan nilai berdasarkan kriteria yang telah ditetapkan. Hal ini juga bertujuan untuk menyederhanakan permasalahan dan meningkatkan akurasi dalam proses pembelajaran.

Terdapat beberapa teknik yang bisa dilakukan, diantaranya ‘binning’, yang mendefinisikan kumpulan kelas nominal untuk setiap atribut, kemudian menetapkan setiap nilai atribut ke dalam salah satu kelas. Misalnya, jika domain atribut numerik mempunyai nilai dari 0 sampai dengan 100, domain tersebut dapat dibagi menjadi empat bin (0..24, 25..49, 50..74, 75..100). Setiap nilai atribut akan dikonversi menjadi atribut nominal yang berkorespondensi dengan salah satu bin. Oleh karena itu metode binning disebut *unsupervised discretization method*.

Teknik berikutnya dapat dilakukan dengan pengukuran entropi, dengan mempertimbangkan semua kemungkinan pembagian dan cari pemotongan terbaik.

64	65	68	69	70	71	72	72	75	75	80	81	83	85
Yes	No	Yes	Yes	Yes	No	No	Yes	Yes	Yes	No	Yes	Yes	No

Langkah pertama adalah urutkan nilai atribut kontinu beserta nilai kelas utama nya. Kemudian lakukan pemotongan dengan pertimbangan pembagian jumlah kelas utama 50:50 atau mendekati, untuk memerkirakan mendapat nilai entropi tertinggi atau mendekati 1. Contoh:

$temperature < 71.5: yes/4, no/2$

$temperature \geq 71.5: yes/5, no/3$

$Info([4,2],[5,3]) = 6/14 info([4,2]) + 8/14 info([5,3]) = 0.939$

Teknik lain, normalisasi, juga dapat dilakukan dengan cara mencari nilai rata dan standart deviasi pada rentang atribut kontinu.

b. Mampu Menangani Atribut yang Kosong (Missing Value)

Nilai kosong pada sebuah dataset wajib kita isi terlebih dahulu sebelum diproses untuk tahap machine learning atau bentuk kedalam sebuah model decision tree.

BI-RAD	Age	shape	Margin	Density	Class
4	48	4	5	?	1
5	67	3	5	3	1
5	57	4	4	3	1
5	60	?	5	1	1
4	53	?	4	3	1
4	28	1	1	3	0
4	70	?	2	3	0
2	66	1	1	?	0
5	63	3	?	3	0
4	78	1	1	1	0

Cara paling mudah dalam pengisian atribut kosong adalah dengan memberikan nilai berdasar nilai yang paling banyak atau dominan dalam atribut tersebut. Contoh, atribut Shape baris ketiga dan keempat kita beri nilai 4 dikarenakan pada kelompok class 1 mayoritas atribut Shape memiliki angka 4. Sedangkan pada class 0, baris 7 kita berikan nilai 1.

BI-RAD	Age	shape	Margin	Density	Class
4	48	4	5	3	1
5	67	3	5	3	1
5	57	4	4	3	1
5	60	4	5	1	1
4	53	4	4	3	1
4	28	1	1	3	0
4	70	1	2	3	0
2	66	1	1	3	0
5	63	3	?	3	0
4	78	1	1	1	0

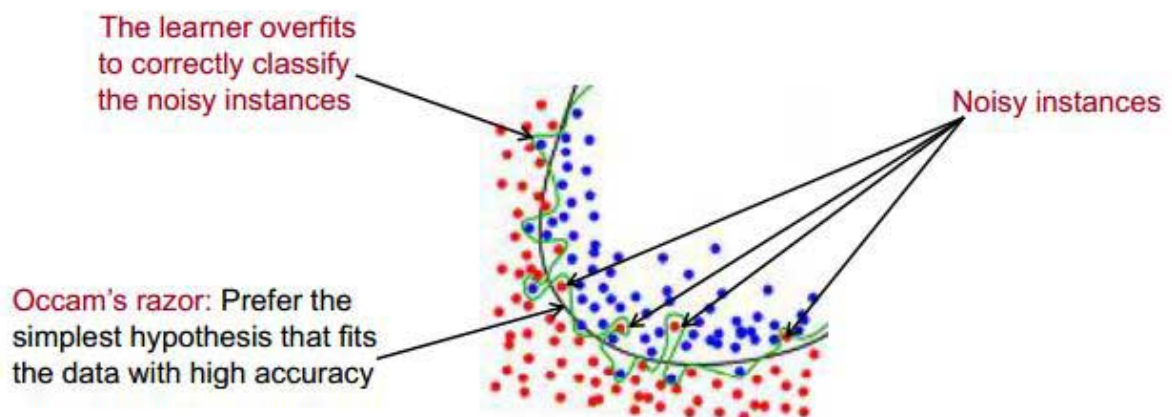
c. Pemangkasan Pohon Keputusan

Pada saat pembangunan pohon keputusan, banyaknya cabang mungkin mencerminkan adanya noise atau outlier pada training data. Pemangkasan pohon dapat dilakukan untuk mengenali dan menghapus cabang-cabang tersebut. Pohon yang dipangkas akan menjadi lebih kecil dan lebih mudah dipahami. Pohon semacam itu biasanya juga menjadi lebih cepat dan lebih baik dalam melakukan klasifikasi.



Jika suatu fungsi atau dataset memiliki terlalu banyak atribut (fitur), model pelatihan yang terbentuk akan menghasilkan akurasi yang baik, namun akan sangat mungkin mengalami kegagalan untuk memprediksi data masukan baru seperti pada contoh diatas.

Pemangkasan pohon juga dapat digunakan untuk mengatasi overfitting. Overfitting terjadi karena ada noise data training, yaitu data yang tidak relevan sehingga mengakibatkan pohon memiliki subtree yang panjang dan tidak seimbang. Misal internal node memiliki kelas YA = 5 dan TIDAK = 1. Data yang berada pada kelas TIDAK merupakan noise, sehingga apabila data tersebut diolah akan menghasilkan pohon dengan subtree yang panjang. Overfitting juga dapat terjadi karena data training yang sedikit.



TUGAS 04
ADVANCED DATABASE
KELAS MTI 23 A

Dosen Pengasuh :
Tri Basuki Kurniawan , S.Kom., M.Eng. Ph.D



Di Susun Oleh :
Ari Hardiyantoro Susanto
NIM : 202420015

Program Pasca Sarjana
Universitas Bina Darma Palembang
2020/2021

Dari data set berikut ini, silahkan buat perhitungan manual dari algoritma decision tree C45 secara lengkap dan gambarkan pohon keputusan yang diperoleh

Jawaban :

No	Outlook	Temperature	Humidity	Windy	Play
1	Sunny	Hot	High	False	No
2	Sunny	Hot	High	True	No
3	Cloudy	Hot	High	False	Yes
4	Rainy	Mild	High	False	Yes
5	Rainy	Cool	Normal	False	Yes
6	Rainy	Cool	Normal	True	Yes
7	Cloudy	Cool	Normal	True	Yes
8	Sunny	Mild	High	False	No
9	Sunny	Cool	Normal	False	Yes
10	Rainy	Mild	Normal	False	Yes
11	Sunny	Mild	Normal	True	Yes
12	Cloudy	Mild	High	True	Yes
13	Cloudy	Hot	Normal	False	Yes
14	Rainy	Mild	High	True	No

Rumus Penyelesaian

$$\text{Entropy}(S) = \sum - p(I) \cdot \log_2 p(I)$$

$$\text{Gain}(S, A) = \text{Entropy}(S) - \sum [p(S|A) \cdot \text{Entropy}(S|A)]$$

Entropy

Kita hitung dulu Nilai Entropy. Kolom Play terdiri dari 14 instance dan terdiri dari 2 label yakni Yes dan No. Play Yes = 10, Play No = 4

$$\begin{aligned}
 \text{Entropy(Decision)} &= - p(\text{Yes}) \cdot \log_2 p(\text{Yes}) - p(\text{No}) \cdot \log_2 p(\text{No}) \\
 \text{Entropy(Decision)} &= - (10/14) \cdot \log_2 (10/14) - (4/14) \cdot \log_2 (4/14) \\
 &= -(0.714) \cdot \log_2 (0.714) - (0.4) \cdot \log_2 (0.4) \\
 &= 0.347 + 0.528 \\
 &= \mathbf{0.875}
 \end{aligned}$$

Menghitung faktor dominan untuk memutuskan bermain.

Faktor Windy dan Play

False Factor windy to Play

No	Outlook	Temperature	Humidity	Windy	Play
1	Sunny	Hot	High	False	No
3	Cloudy	Hot	High	False	Yes
4	Rainy	Mild	High	False	Yes
5	Rainy	Cool	Normal	False	Yes
8	Sunny	Mild	High	False	No
9	Sunny	Cool	Normal	False	Yes
10	Rainy	Mild	Normal	False	Yes
13	Cloudy	Hot	Normal	False	Yes

Tabel faktor false windy terdiri dari **8 instance**. False No = 2, False Yes = 6

$$\begin{aligned}
 \text{Entropy play|windy=False} &= -(2/8) \cdot \log_2(2/8) - (6/8) \cdot \log_2(6/8) \\
 &= -(0.25) \cdot \log_2(0.25) - (0.75) \cdot \log_2(0.75) \\
 &= -(0.25) \cdot (-2) - (0.75) \cdot (-0.415) \\
 &= 0.50 + 0.311 \\
 &= \mathbf{0.811}
 \end{aligned}$$

Menghitung True Factor windy to Play

No	Outlook	Temperature	Humidity	Windy	Play
2	Sunny	Hot	High	True	No
6	Rainy	Cool	Normal	True	Yes
7	Cloudy	Cool	Normal	True	Yes
11	Sunny	Mild	Normal	True	Yes
12	Cloudy	Mild	High	True	Yes
14	Rainy	Mild	High	True	No

Tabel faktor True windy terdiri dari **6 instance**. True No = 2, True Yes = 4

$$\begin{aligned}
 \text{Entropy play|windy=True} &= -(2/6) \cdot \log_2(2/6) - (4/6) \cdot \log_2(4/6) \\
 &= -(0.333) \cdot \log_2(0.333) - (0.667) \cdot \log_2(0.667) \\
 &= -(0.333) \cdot (-1.586) - (0.667) \cdot (-0.586) \\
 &= 0.528 + 0.390 \\
 &= \mathbf{0.918}
 \end{aligned}$$

Menghitung Gain-Play|Windy

$$\text{Gain}(\text{Play}, \text{Windy}) = \text{Entropy}(\text{Play}) - [p(\text{Play}|\text{Windy}=\text{False}) \cdot \text{Entropy}(\text{Play}|\text{Windy}=\text{False})] - [p(\text{Play}|\text{Windy}=\text{True}) \cdot \text{Entropy}(\text{Play}|\text{Windy}=\text{True})]$$

$$= 0.875 - [(8/14) \cdot 0.8111] - [(6/14) \cdot 0.918]$$

$$= 0.804$$

Maka $\text{Gain}(\text{Play}, \text{Windy}) = \mathbf{0.804}$

Selanjutnya kolom 1

Menghitung $\text{Gain}(\text{Play}, \text{Outlook})$

Faktor Outlook Sunny

No	Outlook	Temperature	Humidity	Windy	Play
1	Sunny	Hot	High	False	No
2	Sunny	Hot	High	True	No
8	Sunny	Mild	High	False	No
9	Sunny	Cool	Normal	False	Yes
11	Sunny	Mild	Normal	True	Yes

Tabel faktor Sunny Play terdiri dari **5 instance**. Play No = 3, Play Yes = 2

$$\begin{aligned}\text{Entropy play}|\text{Outlook}=\text{Sunny} &= -(3/5) \cdot \log_2(3/5) - (2/5) \cdot \log_2(2/5) \\ &= -(0.6) \cdot \log_2(0.6) - (0.4) \cdot \log_2(0.4) \\ &= -(0.6) \cdot (-0.737) - (0.4) \cdot (-1.322) \\ &= 0.442 + 0.5288 \\ &= \mathbf{0.971}\end{aligned}$$

Faktor Cloudy Outlook

No	Outlook	Temperature	Humidity	Windy	Play
3	Cloudy	Hot	High	False	Yes
7	Cloudy	Cool	Normal	True	Yes
12	Cloudy	Mild	High	True	Yes
13	Cloudy	Hot	Normal	False	Yes

Tabel faktor Cloudy- Play terdiri dari **4 instance**. Play No = 0, Play Yes = 4

$$\begin{aligned}\text{Entropy play}|\text{Outlook}=\text{Cloudy} &= -(0/4) \cdot \log_2(0/4) - (4/4) \cdot \log_2(4/4) \\ &= \mathbf{0}\end{aligned}$$

Faktor Rainy Outlook

No	Outlook	Temperature	Humidity	Windy	Play
4	Rainy	Mild	High	False	Yes
5	Rainy	Cool	Normal	False	Yes
6	Rainy	Cool	Normal	True	Yes
10	Rainy	Mild	Normal	False	Yes
14	Rainy	Mild	High	True	No

Tabel faktor Rainy Play terdiri dari **5 instance**. Play No = 1, Play Yes = 4

$$\begin{aligned}
 \text{Entropy play|Outlook=Sunny} &= -(1/5) \cdot \log_2(1/5) - (4/5) \cdot \log_2(4/5) \\
 &= -(0.2) \cdot \log_2(0.2) - (0.8) \cdot \log_2(0.8) \\
 &= -(0.2) \cdot (-2.321) - (0.8) \cdot (-0.322) \\
 &= 0.464 + 0.2576 \\
 &= \mathbf{0.722}
 \end{aligned}$$

$$\begin{aligned}
 \text{Maka Gain (Play,Outlook)} &= 0.875 - [(5/14) \cdot 0.971] - (0) - [(5/14) \cdot 0.722] \\
 &= 0.787
 \end{aligned}$$

.....

Selanjutnya Kolom 2 (Temperature)
Menghitung Gain(play,temperature)

Faktor Hot Temperatur

No	Outlook	Temperature	Humidity	Windy	Play
1	Sunny	Hot	High	False	No
2	Sunny	Hot	High	True	No
3	Cloudy	Hot	High	False	Yes
13	Cloudy	Hot	Normal	False	Yes

Tabel faktor Hot Play terdiri dari **4 instance**. Play No = 2, Play Yes = 2

$$\begin{aligned}
 \text{Entropy play|Temperature=Hot} &= -(2/4) \cdot \log_2(2/4) - (2/4) \cdot \log_2(2/4) \\
 &= -(0.5) \cdot \log_2(0.5) - (0.5) \cdot \log_2(0.5) \\
 &= -(0.5) \cdot (-1) - (0.5) \cdot (-1) \\
 &= 0.5 + 0.5 \\
 &= \mathbf{1.00}
 \end{aligned}$$

Faktor Mild Temperature

No	Outlook	Temperature	Humidity	Windy	Play
4	Rainy	Mild	High	False	Yes
8	Sunny	Mild	High	False	No
10	Rainy	Mild	Normal	False	Yes
11	Sunny	Mild	Normal	True	Yes
12	Cloudy	Mild	High	True	Yes
14	Rainy	Mild	High	True	No

Tabel faktor Mild Play terdiri dari **6 instance**. Play No = 2, Play Yes = 4
 Entropy play|Temperature=Mild = $-(2/6) \cdot \log_2(2/6) - (4/6) \cdot \log_2(4/6)$
 $= -(0.333) \cdot \log_2(0.333) - (0.667) \cdot \log_2(0.667)$
 $= -(0.333) \cdot (-1.586) - (0.667) \cdot (-0.584)$
 $= 0.528 + 0.389$
 $= \mathbf{0.92}$

Faktor cold Temperature

No	Outlook	Temperature	Humidity	Windy	Play
5	Rainy	Cool	Normal	False	Yes
6	Rainy	Cool	Normal	True	Yes
7	Cloudy	Cool	Normal	True	Yes
9	Sunny	Cool	Normal	False	Yes

Tabel faktor Cold Play terdiri dari **4 instance**. Play No = 0, Play Yes = 4
 Entropy play|Temperature=Cool = $-(0/4) \cdot \log_2(0/4) - (4/4) \cdot \log_2(4/4)$
 $= \mathbf{0}$

Maka Gain (**Play, Temperature**) = $0.875 - [(4/14) \cdot 1 - (6/14) \cdot 0.92] - (0)$
 $= 0.875 - [(0.286) - (0.395)] - (0)$
 $= \mathbf{1.065}$

Selanjutnya menghitung Kolom 3 (Humidity)

Menghitung Gain|Play, Humidity

Faktor High Humidity

No	Outlook	Temperature	Humidity	Windy	Play
1	Sunny	Hot	High	False	No
2	Sunny	Hot	High	True	No
3	Cloudy	Hot	High	False	Yes
4	Rainy	Mild	High	False	Yes
8	Sunny	Mild	High	False	No
12	Cloudy	Mild	High	True	Yes
14	Rainy	Mild	High	True	No

Tabel faktor High Play terdiri dari 7 instance. Play No = 4, Play Yes = 3
 Entropy play|Outlook=Sunny = $-(4/7) \cdot \log_2(4/7) - (3/7) \cdot \log_2(3/7)$
 $= -(0.571) \cdot \log_2(0.571) - (0.428) \cdot \log_2(0.428)$
 $= -(0.571) \cdot (-0.808) - (0.428) \cdot (-1.224)$
 $= 0.461 + 0.524$
 $= 0.985$

Faktor Normal Humidity

No	Outlook	Temperature	Humidity	Windy	Play
5	Rainy	Cool	Normal	False	Yes
6	Rainy	Cool	Normal	True	Yes
7	Cloudy	Cool	Normal	True	Yes
9	Sunny	Cool	Normal	False	Yes
10	Rainy	Mild	Normal	False	Yes
13	Cloudy	Hot	Normal	False	Yes

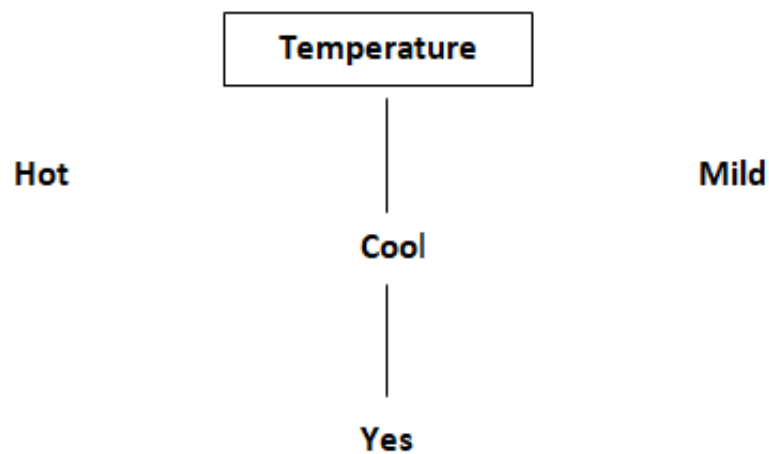
Tabel faktor Normal Play terdiri dari 6 instance. Play No = 0, Play Yes = 6
 Entropy play|Outlook=Sunny = $-(0/6) \cdot \log_2(0/6) - (6/6) \cdot \log_2(6/6)$
 $= 0$

Maka Gain (Play, Humadity) = $0.875 - [(7/14) \cdot 0.985] - (0)$
 $= 0.875 - 0.492$
 $= 0.383$

Dari perhitungan diatas didapat

1. Gain (Play,Windy) = 0.804
2. Gain (Play,Outlook) = 0.787
3. Gain (Play, Temperature) = 1.605
4. Gain (Play,Humadity) = 0.383

Karena nilai temperatur paling besar, maka pohon keputusan paling atas adalah faktor Temperature.



Cool Temperatur On Play

Dari tabel dibawah, Keputusan untuk bermain diambil jika temperatur dingin/cool

No	Outlook	Temperature	Humidity	Windy	Play
5	Rainy	Cool	Normal	False	Yes
6	Rainy	Cool	Normal	True	Yes
7	Cloudy	Cool	Normal	True	Yes
9	Sunny	Cool	Normal	False	Yes

Hot Temperatur On Play.

No	Outlook	Temperature	Humidity	Windy	Play
1	Sunny	Hot	High	False	No
2	Sunny	Hot	High	True	No
3	Cloudy	Hot	High	False	Yes
13	Cloudy	Hot	Normal	False	Yes

Mencari nilai hubungan Outlook-Temperatur / Gain (temp,Outlook)

Entropy Hot Temperatur = $2/4$ no dan $2/4$ yes

$$\begin{aligned}\text{Nilai Entrophy} &= -(2/4).\log_2(2/4) - (2/4).\log_2(2/4) \\ &= \mathbf{1.00}\end{aligned}$$

Mencari entropy Sunny

karena, instance sunny = 2, nilai play No = 2 dan nilai Play yes = 0, maka
Entropy Sunny = 0

Mencari entropy Cloudy

Karena instance cloudy = 2, nilai Play No = 0 dan nilai play yes = 2, maka
Entropy Cloudy = 0

$$\begin{aligned}\text{Maka Gain(Outlook-Temperature)} &= \mathbf{1.00 - (2/4.0) - (2/4.0)} \\ &= \mathbf{1.00}\end{aligned}$$

Mencari nilai hubungan Humidity-Temperatur / Gain (Temp,Hum)

Entropy High terdiri dari 3 instance dengan nilai play no = 2, yes = 1.

$$\begin{aligned}\text{Entropy High} &= -(2/3).\log_2(2/3) - (1/3).\log_2(1/3) \\ &= -(0.667).\log(0.667) - (0.333).\log(0.333) \\ &= -(0.667).(-0.584) - (0.333).(-1.586) \\ &= 0.389 + 0.528 \\ &= 0.917\end{aligned}$$

$$\begin{aligned}\text{Entropy Normal} &= -(0/1).\log_2(0/1) - (1/1).\log_2(1/1) \\ &= 0\end{aligned}$$

$$\begin{aligned}\text{Maka Gain(Humadity-Temperature)} &= \mathbf{1.00 - (3/4)(0.971) - (1/4.0)} \\ &= \mathbf{1.00 - 0.728} \\ &= \mathbf{0.272}\end{aligned}$$

Mencari nilai hubungan Windy-Temperatur / Gain (Temp,Windy)

Entropy False terdiri dari 3 instance, dengan nilai play no = 1 dan yes = 2.

$$\begin{aligned}\text{Entropy False} &= -(1/3) \cdot \log_2(1/3) - (2/3) \cdot \log_2(2/3) \\ &= -(0.333) \cdot \log(0.333) - (0.667) \cdot \log(0.667) \\ &= -(0.333) \cdot (-1.586) - (0.667) \cdot (-0.584) \\ &= 0.528 + 0.389 \\ &= 0.917\end{aligned}$$

Entropy True, dengan nilai play No = 1, play yes = 0

Entropy True = 0

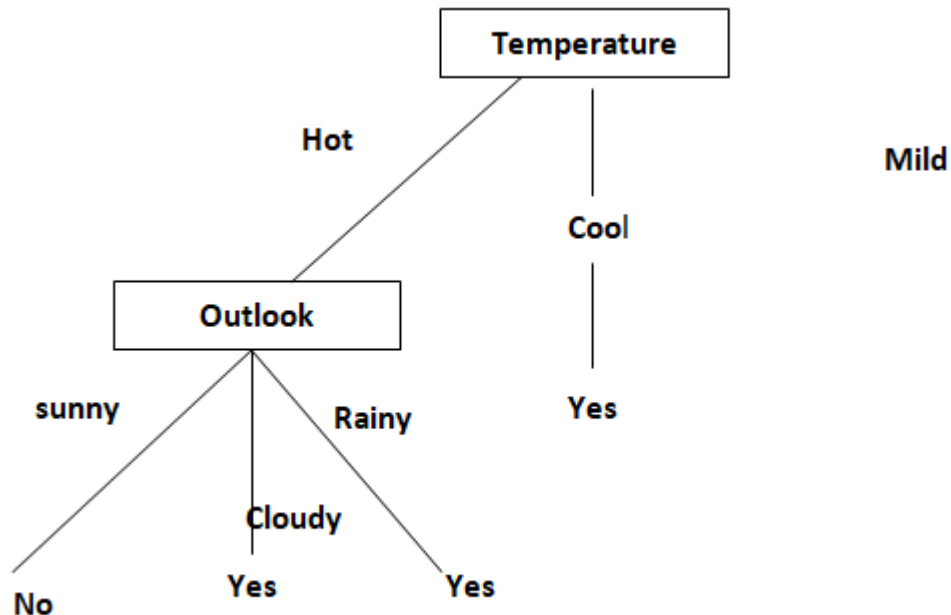
$$\begin{aligned}\text{Maka Gain(Windy-Temperature)} &= 1.00 - (3/4)(0.971) - (1/4.0) \\ &= 1.00 - 0.728 \\ &= 0.272\end{aligned}$$

Maka nilai hubungan antara Temperature Hot dengan Outlook, Humidity dan windi adalah

$$\text{Maka Gain(Outlook-Temperature=hot)} = 1.00$$

$$\text{Maka Gain(Humadity-Temperature=hot)} = 0.272$$

$$\text{Maka Gain(Windy-Temperature =hot)} = 0.272$$



Mild Temperatur On Play.

No	Outlook	Temperature	Humidity	Windy	Play
4	Rainy	Mild	High	False	Yes
8	Sunny	Mild	High	False	No
10	Rainy	Mild	Normal	False	Yes
11	Sunny	Mild	Normal	True	Yes
12	Cloudy	Mild	High	True	Yes
14	Rainy	Mild	High	True	No

Mencari hubungan **Mild Temperatur On Play**

Mencari nilai hubungan Outlook-Temperatur / Gain (temp,Outlook)

Entropy **Mild**Temperatur = 2/6 No dan 4/6 yes

$$\begin{aligned}\text{Nilai Entrophy} &= -(2/6).\log_2(2/6) - (4/6).\log_2(4/6) \\ &= -(0.333).\log_2(0.333) - (0.667).\log_2(0.667) \\ &= 0.917\end{aligned}$$

Mencari entropy Rainy

Instance = 3, Play No = 1 , Play Yes = 2

$$\begin{aligned}\text{Entropy Rainy} &= -(1/3).\log_2(1/3) - (2/3).\log_2(2/3) \\ &= -(0.333).(0.333) - (0.667).\log_2(0.667) \\ &= 0.917\end{aligned}$$

Mencari entropy Sunny

Instance = 2, Play No = 1 , Play Yes = 1

$$\begin{aligned}\text{Entropy Rainy} &= -(1/2).\log_2(1/2) - (1/2).\log_2(1/2) \\ &= -(0.5).\log_2(0.5) - (0.5).\log_2(0.5) \\ &= -(0.5).(-1.0) - (0.5).\log_2(-1.0) \\ &= 0.5 + 0.5 \\ &= 1.00\end{aligned}$$

Mencari entropy Cloudy

Instance = 1, Play No = 0 , Play Yes = 1

$$\begin{aligned}\text{Entropy Rainy} &= -(0/1).\log_2(0/1) - (1/1).\log_2(1/1) \\ &= 0\end{aligned}$$

$$\begin{aligned}\text{Maka Gain(Outlook|Temperature=mild)} &= 0.917 - [(3/6).(0.917) - (2/6).1] - (0) \\ &= 0.917 - [(0.458) - (0.667) - (0)] \\ &= 0.917 - (-0.209) \\ &= 1.126\end{aligned}$$

Mencari hubungan **Mild Temperatur On Play. (temp-humadity)**

Mencari nilai hubungan Humidity-Temperatur / Gain (temp,humidity)

Mencari entropy High

Instance = 4, Play No = 2 , Play Yes = 2

$$\begin{aligned}\text{Entrophy Rainy} &= -(2/4).\log_2(2/4) - (2/4).\log_2(2/4) \\ &= -(0.5).(-1.00) - (0.5).(-1.0) \\ &= 1.00\end{aligned}$$

Mencari entropy Normal

Instance = 2, Play No = 0 , Play Yes = 2

$$\begin{aligned}\text{Entrophy Rainy} &= -(0/2).\log_2(0/2) - (2/2).\log_2(2/2) \\ &= 0\end{aligned}$$

$$\begin{aligned}\text{Maka Gain(Humadity|Temperature=mild)} &= 0.917 - [(4/6).(1) - (0)] \\ &= 0.917 - [(0.667)-(0)] \\ &= 0.917 - 0.667 \\ &= 0.25\end{aligned}$$

--

Mencari hubungan **Mild Temperatur On Play. (temp-windy)**

Mencari nilai hubungan Windy-Temperatur / Gain (temp,windy)

Mencari entropy False pada Windy

Instance = 3, Play No = 1 , Play Yes = 2

$$\begin{aligned}\text{Entrophy False} &= -(1/3).\log_2(1/3) - (2/3).\log_2(2/3) \\ &= -(0.333).(-1.586) - (0.667).(-0.584) \\ &= 0.528+0.389 \\ &= 0.917\end{aligned}$$

Mencari entropy True pada Windy

Instance = 3, Play No = 1 , Play Yes = 2

$$\begin{aligned}\text{Entrophy True} &= -(1/3).\log_2(1/3) - (2/3).\log_2(2/3) \\ &= -(0.333).(-1.586) - (0.667).(-0.584) \\ &= 0.528+0.389 \\ &= 0.917\end{aligned}$$

$$\begin{aligned}\text{Maka Gain(Windy|Temperature=mild)} &= 0.917 - [(3/6).(0.917) - ((3/6).(0.917))] \\ &= 0.917 - (0) \\ &= 0.917\end{aligned}$$

Maka nilai hubungan antara Temperature Mild dengan Outlook, Humidity dan windy adalah

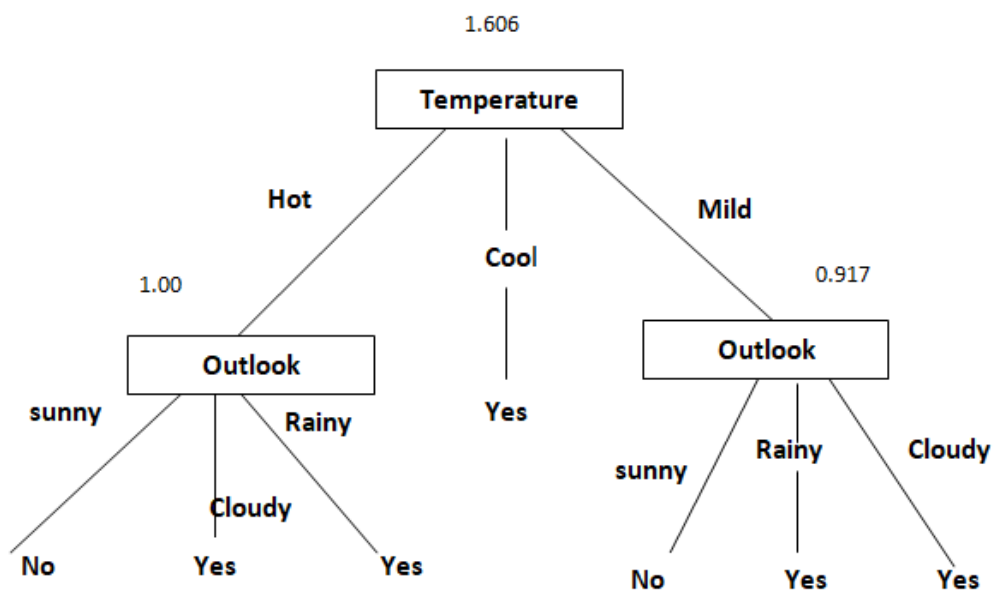
Maka Gain(Outlook-Temperature=mild) = 0.917

Maka Gain(Humadity-Temperature=mild) = 0.25

Maka Gain(Windy-Temperature =mild) = 0.917

Terdapat 2 nilai yang sama, yakni pada **outlook dan Windy**, namun kita bisa cek pada tabel berikut bahwa keputusan untuk bermain pada temperature mild lebih dipengaruhi oleh **Outlook**

No	Outlook	Temperature	Humidity	Windy	Play
4	Rainy	Mild	High	False	Yes
8	Sunny	Mild	High	False	No
10	Rainy	Mild	Normal	False	Yes
11	Sunny	Mild	Normal	True	Yes
12	Cloudy	Mild	High	True	Yes
14	Rainy	Mild	High	True	No



Kesimpulan :***Kelebihan***

1. Dapat menggunakan atribut nominal dimana kebanyakan tidak dapat dilakukan oleh algoritma machine learning.
2. Algoritma ini mampu mentransformasikan dari data mentah menjadi data yang beraturan.

Kekurangan

1. Cenderung kaku dan dapat menimbulkan kesalahan perhitungan.
Metode pembandingan selain menggunakan metode ini dapat dilakukan dengan metode chi-square / CHAID atau Random Forest.
2. Terlalu panjang/lama waktu untuk mencari hubungan satu faktor dengan faktor lainnya.

Diketahui Dataset adalah sebagai berikut :

No	Outlook	Temperature	Humidity	Windy	Play
1	Sunny	Hot	High	FALSE	No
2	Sunny	Hot	High	TRUE	No
3	Cloudy	Hot	High	FALSE	Yes
4	Rainy	Mild	High	FALSE	Yes
5	Rainy	Cool	Normal	FALSE	Yes
6	Rainy	Cool	Normal	TRUE	Yes
7	Cloudy	Cool	Normal	TRUE	Yes
8	Sunny	Mild	High	FALSE	No
9	Sunny	Cool	Normal	FALSE	Yes
10	Rainy	Mild	Normal	FALSE	Yes
11	Sunny	Mild	Normal	TRUE	Yes
12	Cloudy	Mild	High	TRUE	Yes
13	Cloudy	Hot	Normal	FALSE	Yes
14	Rainy	Mild	High	TRUE	No

Atribut : Outlook, Temperature, Humidity, Windy

Kelas : Play = Yes atau No

Jumlah data ada 12, terdiri dari :

Ya = 10

Tidak = 4

1. Menghitung Entropy Total Dataset

$$Entropi (S) = \sum_{j=1}^k -p_j \log_2 p_j$$

Entropy [Total]

$$S = \left\{ -\frac{10}{14} * (\log_2 \left(\frac{10}{14} \right)) \right\} + \left\{ -\frac{4}{14} * (\log_2 \left(\frac{4}{14} \right)) \right\} = 0.863120569$$

2. Menghitung Entropy dan Gain tiap Atribut

$$\text{Gain (S,A)} = \text{Entropy (s)} - \sum_{i=1}^n \frac{|s_i|}{|S|} * \text{Entropy (s}_i\text{)}$$

- Menghitung Gain pada baris Outlook
 $= 0.863120569 - ((4/14 * 0) + (5/14 * 0.72) + (5/14 * 0.97))$
 $= 0.258521037$
- Menghitung Gain pada baris Temperature
 $= 0.863120569 - ((4/14 * 0) + (4/14 * 1) + (6/14 * 0.91))$
 $= 0.183850925$
- Menghitung Gain pada baris Humidity
 $= 0.863120569 - ((7/14 * 0.98) + (7/14 * 0))$
 $= 0.370506501$
- Menghitung Gain pada baris Windy
 $= 0.863120569 - ((8/14 * 0.81) + (6/14 * 0.91))$
 $= 0.005977711$

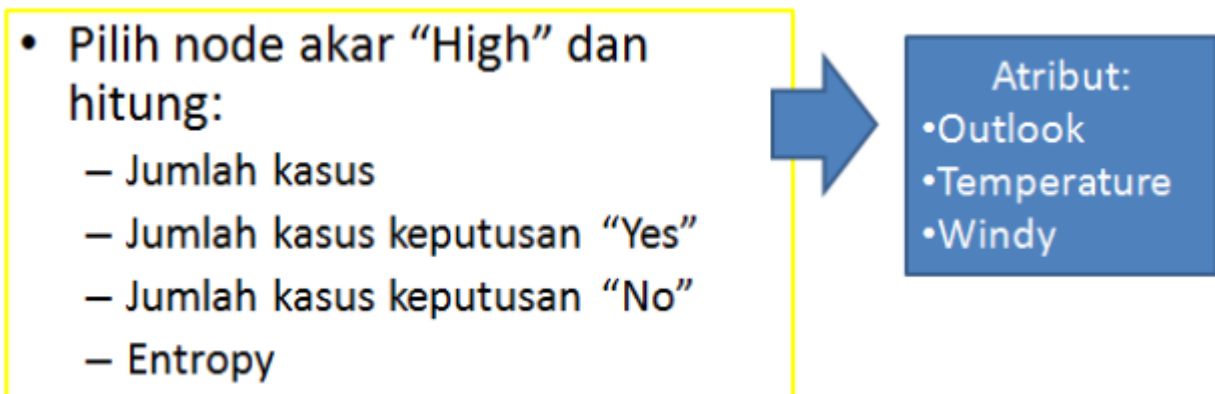
3. Perhitungan Langkah 1

Node			Jumlah Kasus (S)	Tidak (S1)	Ya (S2)	Entropy	Gain
1	TOTAL		14	10	4	0.863120569	

Node			Jumlah Kasus (S)	Tidak (S1)	Ya (S2)	Entropy	Gain
1	TOTAL		14	10	4	0.863120569	
	OUTLOOK						0.258521037
		CLOUDY	4	0	4	0	
		RAINY	5	1	4	0.721928095	
		SUNNY	5	3	2	0.970950594	
	TEMPERATURE						0.183850925
		COOL	4	0	4	0	
		HOOT	4	2	2	1	
		MILD	6	2	4	0.918295834	
	HUMIDITY						0.370506501
		HIGH	7	4	3	0.985228136	
		NORMAL	7	0	7	0	
	WINDY						0.005977711
		FALSE	8	2	6	0.811278124	
		TRUE	6	2	4	0.918295834	

- Seperti yg terlihat pd tabel, diperoleh bhw atribut dgn Gain tertinggi adalah Humidity -> 0,37
- Maka Humidity menjadi node akar
- Humidity memiliki dua nilai yaitu "High" dan "Normal"
- Humidity -> "Normal" sdh mengklasifikasikan kasus menjadi 1 yaitu keputusannya "yes"
- Untuk humidity -> "High" msh perlu dilakukan perhitungan lagi (karena masih terdapat "yes" dan "no")

4. Perincian Algoritma Langkah 2



5. Hasil Perhitungan langkah 2

Node			Jumlah Kasus (S)	Tidak (S1)	Ya (S2)	Entropy	Gain
1.1	HUMIDITY HIGH		7	4	3	0.985228136	
	OUTLOOK						0.69951385
		CLOUDY	2	0	2	0	
		RAINY	2	1	1	1	
		SUNNY	3	3	0	0	
	TEMPERATURE						0.020244207
		COOL	0	0	0	0	
		HOOT	3	2	1	0.918295834	
		MILD	4	2	2	1	
	WINDY						0.020244207
		FALSE	4	2	2	1	
		TRUE	3	2	1	0.918295834	

- Didapat Gain tertinggi -> outlook -> 0,69
- Maka "Outlook" menjadi node cabang dari atribut humidity yg bernilai "High"
- Berdasarkan atribut "Outlook" terdpt 3 nilai

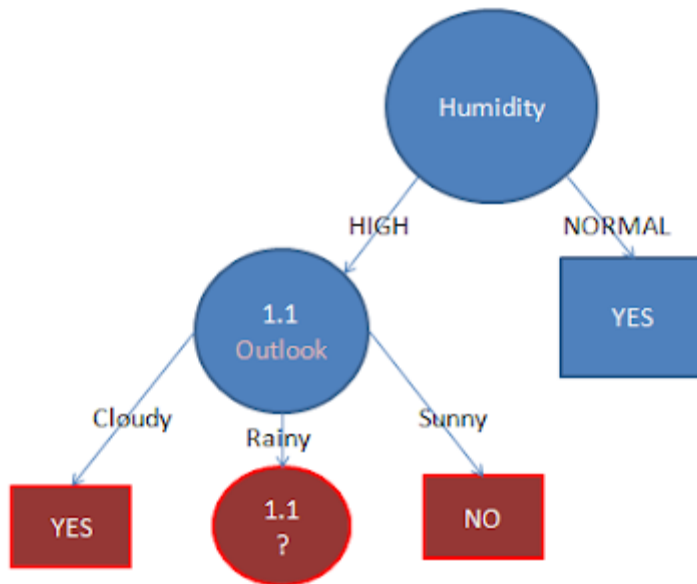
Cloudy

Rainy

Sunny

- Krn "Cloudy" pasti bernilai "Yes" dan "Sunny" pasti bernilai "No", maka tdk perlu dilakukan perhitungan lagi
- Sedangkan "Rainy" bernilai "yes" dan "No", maka masih perlu dilakukan perhitungan lagi

6. Pohon Keputusan Node 1.1



7. Perincian Algoritma langkah 3

- Pilih node akar "Rainy" dan hitung:
 - Jumlah kasus
 - Jumlah kasus keputusan "Yes"
 - Jumlah kasus keputusan "No"
 - Entropy



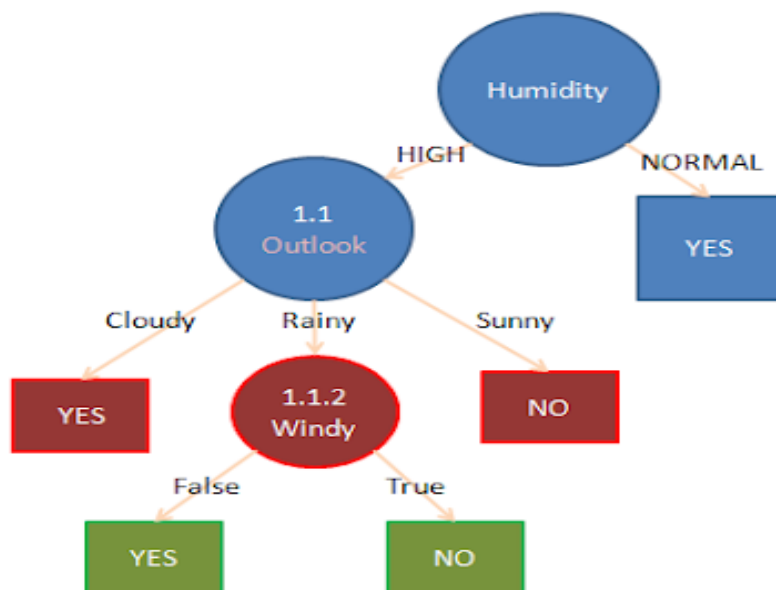
Atribut:
•Temperature
•Windy

8. Hasil Perhitungan langkah 3

Node			Jumlah Kasus (S)	Tidak (S1)	Ya (S2)	Entropy	Gain
1.1.2	HUMIDITY "HIGH" & OUTLOOK "RAINY"		2	1	1	1	
	TEMPERATURE						0
		COOL	0	0	0	0	
		HOOT	0	0	0	0	
		MILD	2	1	1	1	
	WINDY						1
		FALSE	1	0	1	0	
		TRUE	1	1	0	0	

- Didapat Gain tertinggi -> Windy -> 1
- Maka "Windy" menjadi node cabang dari atribut humidity yg bernilai "High" dan outlook yg bernilai "Rainy"
- Berdasarkan atribut "Windy" terdpt 2 nilai
True
False
- Karena "True" sdh terklasifikasi pasti bernilai "No" dan "False" pasti bernilai "Yes", maka tidak perlu dilakukan perhitungan lagi

9. Pohon Keputusan Node 1.1.2



Berdasarkan Node 1.1.2, Maka :

Semua kasus sudah mendapatkan kelas

Sehingga pohon keputusan diatas merupakan pohon keputusan terakhir yang dibentuk.

Pendapat saya tentang Algoritma ini :

Algoritma C4.5 merupakan kelompok algoritma decision tree. Algoritma ini mempunyai input berupa training samples dan samples. Training samples merupakan data contoh yang digunakan untuk membangun sebuah tree yang telah diuji kebenarannya. Sedangkan samples merupakan field-field data yang digunakan sebagai parameter dalam klasifikasi data.

Kelebihan Algoritma C4.5 :

- C4.5 mampu menangani atribut yang kosong (Missing Value).
- C4.5 mampu menangani atribut dengan kontinu.
- C4.5 memangkas pohon keputusan untuk mengatasi overfitting.

Kekurangan Algoritma C4.5 :

- Algoritma C4.5 hanya dapat digunakan untuk menangani sampel-sampel yang dapat disimpan secara keseluruhan dan pada waktu yang bersamaan di memori.

Di dilihat perbandingan dari Algoritma C4.5 dengan Algoritma ID3, dimana Algoritma ID3 merupakan algoritma yang dipergunakan untuk membangun sebuah decision tree atau pohon keputusan. Algoritma C4.5 merupakan pengembangan dari Algoritma ID3, beberapa perbedaannya antara lain :

- Mampu menangani atribut dengan tipe diskrit atau kontinu.
- Mampu menangani atribut yang kosong (missing value).
- Bisa memangkas cabang.

Row Labels	Sum of NO
Cool	27
Hot	19
Mild	59
Grand Total	105

NO	OUTLOOK	TEMPERATURE	HUMIDITY	WINDY	PLAY
1	Sunny	Hot	High	FALSE	No
2	Sunny	Hot	High	TRUE	No
3	Cloudy	Hot	High	FALSE	Yes
4	Rainy	Mild	High	FALSE	Yes
5	Rainy	Cool	Normal	FALSE	Yes
6	Rainy	Cool	Normal	TRUE	Yes
7	Cloudy	Cool	Normal	TRUE	Yes
8	Sunny	Mild	High	FALSE	No
9	Sunny	Cool	Normal	FALSE	Yes
10	Rainy	Mild	Normal	FALSE	Yes
11	Sunny	Mild	Normal	TRUE	Yes
12	Cloudy	Mild	High	TRUE	Yes
13	Cloudy	Hot	Normal	FALSE	Yes
14	Rainy	Mild	High	TRUE	No

Node
1

Row Labels	Count of PLAY
No	4
Yes	10
Grand Total	14

Count of PLAY	Column Labels		
Row Labels	No	Yes	Grand Total
Cloudy			4
Rainy	1	4	5
Sunny	3	2	5
Grand Total	4	10	14

Count of PLAY	Column Labels		
Row Labels	No	Yes	Grand Total
Cool			4
Hot	2	2	4
Mild	2	4	6
Grand Total	4	10	14

Count of PLAY	Column Labels		
Row Labels	No	Yes	Grand Total
High	4	3	7
Normal		7	7
Grand Total	4	10	14

Count of PLAY	Column Labels		
Row Labels	No	Yes	Grand Total
FALSE	2	6	8
TRUE	2	4	6
Grand Total	4	10	14

No	Outlook
1	Sunny
2	Sunny
3	Cloudy
4	Rainy
5	Rainy
6	Rainy
7	Cloudy
8	Sunny
9	Sunny
10	Rainy
11	Sunny
12	Cloudy
13	Cloudy
14	Rainy

Deskripsi		Jml Kasus (S)	Count Play "YES"	Count Play "NO"	Entropy
Total		14	10	4	0,86312057
OUTLOOK					
Cloudy		4	4		
Rainy		5	4	1	0,72192809
Sunny		5	2	3	0,97095059
TEMPERATURE					
Cool		4	4		0
Hot		4	2	2	1
Mild		6	4	2	0,91829583
HUMIDITY					
High		7	3	4	0,98522814
Normal		7	7		
WINDY					
FALSE		8	6	2	0,81127812
TRUE		6	4	2	0,91829583

Deskripsi	Temperature	Humidity	Windy	Play
Hot	Hot	High	FALSE	No
Hot	Hot	High	TRUE	No
Hot	Hot	High	FALSE	Yes
Mild	Mild	High	FALSE	Yes
Cool	Cool	Normal	FALSE	Yes
Cool	Cool	Normal	TRUE	Yes
Cool	Cool	Normal	TRUE	Yes
Mild	Mild	High	FALSE	No
Cool	Cool	Normal	FALSE	Yes
Mild	Mild	Normal	FALSE	Yes
Mild	Mild	Normal	TRUE	Yes
Mild	Mild	High	TRUE	Yes
Hot	Hot	Normal	FALSE	Yes
Mild	Mild	High	TRUE	No

Gain
0,25852104
0,18385093
0,3705065
0,00597771

NO	OUTLOOK	TEMPERATURE	HUMIDITY	WINDY	PLAY
1	Sunny	Hot	High	FALSE	No
2	Sunny	Hot	High	TRUE	No
3	Cloudy	Hot	High	FALSE	Yes
4	Rainy	Mild	High	FALSE	Yes
5	Rainy	Cool	Normal	FALSE	Yes
6	Rainy	Cool	Normal	TRUE	Yes
7	Cloudy	Cool	Normal	TRUE	Yes
8	Sunny	Mild	High	FALSE	No
9	Sunny	Cool	Normal	FALSE	Yes
10	Rainy	Mild	Normal	FALSE	Yes
11	Sunny	Mild	Normal	TRUE	Yes
12	Cloudy	Mild	High	TRUE	Yes
13	Cloudy	Hot	Normal	FALSE	Yes
14	Rainy	Mild	High	TRUE	No

HUMIDITY High

Row Labels	Count of PLAY
No	4
Yes	3
Grand Total	7

HUMIDITY High

Count of PLAY	Column Labels		
Row Labels	No	Yes	Grand Total
Cloudy			2
Rainy	1	1	2
Sunny	3		3
Grand Total	4	3	7

HUMIDITY High

Count of PLAY	Column Labels		
Row Labels	No	Yes	Grand Total
Hot	2	1	3
Mild	2	2	4
Grand Total	4	3	7

HUMIDITY High

Count of PLAY	Column Labels		
Row Labels	No	Yes	Grand Total
FALSE	2	2	4
TRUE	2	1	3
Grand Total	4	3	7

Node
1

Deskripsi		Jml Kasus (S)	Count Play "YES"	Count Play "NO"	Entropy
Total		7	3	4	0,98522814
OUTLOOK					
Cloudy		2	2		0
Rainy		2	1	1	1
Sunny		3		3	0
TEMPERATURE					
Cool		0	0	0	0
Hot		3	1	2	0,91829583
Mild		4	2	2	1
HUMIDITY					
High		7	3	4	0,98522814
Normal		7	7		
WINDY					
FALSE		4	2	2	1
TRUE		3	1	2	0,91829583

Gain
0,69951385
0,02024421
0
0,02024421

No	Outlook	Temperature	Humidity	Win
1	Sunny	Hot	High	FAL
2	Sunny	Hot	High	TR
3	Cloudy	Hot	High	FAL
4	Rainy	Mild	High	FAL
5	Rainy	Cool	Normal	FAL
6	Rainy	Cool	Normal	TR
7	Cloudy	Cool	Normal	TR
8	Sunny	Mild	High	FAL
9	Sunny	Cool	Normal	FAL
10	Rainy	Mild	Normal	FAL
11	Sunny	Mild	Normal	TR
12	Cloudy	Mild	High	TR
13	Cloudy	Hot	Normal	FAL
14	Rainy	Mild	High	TR

ndy	Play
LSE	No
UE	No
LSE	Yes
LSE	Yes
LSE	Yes
UE	Yes
UE	Yes
LSE	No
LSE	Yes
LSE	Yes
UE	Yes
UE	Yes
LSE	Yes
UE	No

NO	OUTLOOK	TEMPERATURE	HUMIDITY	WINDY	PLAY
1	Sunny	Hot	High	FALSE	No
2	Sunny	Hot	High	TRUE	No
3	Cloudy	Hot	High	FALSE	Yes
4	Rainy	Mild	High	FALSE	Yes
5	Rainy	Cool	Normal	FALSE	Yes
6	Rainy	Cool	Normal	TRUE	Yes
7	Cloudy	Cool	Normal	TRUE	Yes
8	Sunny	Mild	High	FALSE	No
9	Sunny	Cool	Normal	FALSE	Yes
10	Rainy	Mild	Normal	FALSE	Yes
11	Sunny	Mild	Normal	TRUE	Yes
12	Cloudy	Mild	High	TRUE	Yes
13	Cloudy	Hot	Normal	FALSE	Yes
14	Rainy	Mild	High	TRUE	No

Node
1

OUTLOOK	Rainy
HUMIDITY	High

Row Labels	Count of PLAY
No	1
Yes	1
Grand Total	2

OUTLOOK	Rainy
HUMIDITY	High

Count of PLAY	Column Labels		
Row Labels	No	Yes	Grand Total
Mild	1	1	2
Grand Total	1	1	2

OUTLOOK	Rainy
HUMIDITY	High

Count of PLAY	Column Labels		
Row Labels	No	Yes	Grand Total
FALSE		1	1
TRUE	1		1
Grand Total	1	1	2

Deskripsi		Jml Kasus (S)	Count Play "YES"	Count Play "NO"	Entropy
Total		2	1	1	1
OUTLOOK					
Cloudy		2	2		0
Rainy		2	1	1	1
Sunny		3		3	0
TEMPERATURE					
Cool		0	0	0	0
Hot		0	0	0	0
Mild		2	1	1	1
HUMIDITY					
High		7	3	4	0,98522814
Normal		7	7		
WINDY					
FALSE		1	1	0	0
TRUE		1	0	1	0

Gain
0
0
-2,4482985
1

No	Outlook	Temperature	Humidity	Win
1	Sunny	Hot	High	FAL
2	Sunny	Hot	High	TR
3	Cloudy	Hot	High	FAL
4	Rainy	Mild	High	FAL
5	Rainy	Cool	Normal	FAL
6	Rainy	Cool	Normal	TR
7	Cloudy	Cool	Normal	TR
8	Sunny	Mild	High	FAL
9	Sunny	Cool	Normal	FAL
10	Rainy	Mild	Normal	FAL
11	Sunny	Mild	Normal	TR
12	Cloudy	Mild	High	TR
13	Cloudy	Hot	Normal	FAL
14	Rainy	Mild	High	TR

ndy	Play
LSE	No
UE	No
LSE	Yes
LSE	Yes
LSE	Yes
UE	Yes
UE	Yes
LSE	No
LSE	Yes
LSE	Yes
UE	Yes
UE	Yes
LSE	Yes
UE	No

Nama : Bhijanta Wyasa WM
NIM : 202420019
Kelas : MTI 23
Tugas 04

Dari dataset berikut ini, silahkan buat perhitungan manual dari algoritma decision tree C45 secara lengkap dan gambarkan pohon keputusan yang diperoleh.

No	Outlook	Temperature	Humidity	Windy	Play
1	Sunny	Hot	High	FALSE	No
2	Sunny	Hot	High	TRUE	No
3	Cloudy	Hot	High	FALSE	Yes
4	Rainy	Mild	High	FALSE	Yes
5	Rainy	Cool	Normal	FALSE	Yes
6	Rainy	Cool	Normal	TRUE	Yes
7	Cloudy	Cool	Normal	TRUE	Yes
8	Sunny	Mild	High	FALSE	No
9	Sunny	Cool	Normal	FALSE	Yes
10	Rainy	Mild	Normal	FALSE	Yes
11	Sunny	Mild	Normal	TRUE	Yes
12	Cloudy	Mild	High	TRUE	Yes
13	Cloudy	Hot	Normal	FALSE	Yes
14	Rainy	Mild	High	TRUE	No

Kumpulkan dalam format MS Word. Sertakan penjelasan dan pendapat anda tentang algoritma ini, kelebihan dan kekurangannya berbanding metode lain yang sejenis (tentukan sendiri metode pembandingnya).

Jawaban :

Pada kasus diatas untuk menentukan pengambil keputusan harus perlu dihitung bagaimana peluang – peluang yang muncul dengan menggunakan algoritma decision tree C45, adapun langkahnya sebagai berikut :

1. Menghitung jumlah kasus, jumlah kasus untuk keputusan Yes, jumlah kasus untuk keputusan No, dan Entropy dari semua kasus dan kasus yang dibagi berdasarkan atribut OUTLOOK, TEMPERATUR, HUMIDITY, dan WINDY.
2. Lakukan perhitungan Gain untuk setiap atribut.

Nama : Bhijanta Wyasa WM
NIM : 202420019
Kelas : MTI 23

- **Entropy**

Entropy adalah suatu parameter untuk mengukur tingkat keberagaman (heterogenitas) dari kumpulan data. Jika nilai dari entropy semakin besar, maka tingkat keberagaman suatu kumpulan data semakin besar. Rumus untuk menghitung entropy sebagai berikut:

$$Entropy(S) = - \sum_{i=1}^m p_i \log_2(p_i)$$

dimana:

m = jumlah kelas klasifikasi

p_i = jumlah proporsi sampel (peluang) untuk kelas i

Sedangkan rumus untuk entropy pada masing-masing variabel adalah:

$$Entropy_A(S) = \sum_v \frac{|S_v|}{|S|} Entropy(S_v)$$

dimana:

A = variabel

v = nilai yang mungkin untuk variabel A

$|S_v|$ = jumlah sampel untuk nilai v

$|S|$ = jumlah sampel untuk seluruh sampel data

$Entropy(S_v) = Entropy$ untuk sampel yang memiliki nilai v

- **Gain**

Gain adalah Ukuran efektifitas suatu variabel dalam mengklasifikasikan data. Gain dari suatu variabel merupakan selisih antara nilai entropy total dengan entropy dari variabel tersebut. Gain dapat dirumuskan dengan:

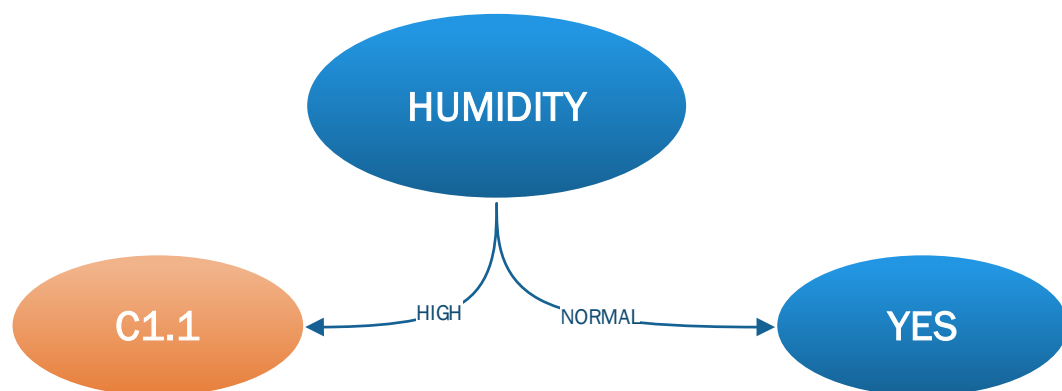
$$Gain(A) = Entropy(S) - Entropy_A(S)$$

Dari data diatas didapatkan data sebagai berikut (*detail perhitungan terlampir di excel*):

Nama : Bhijanta Wyasa WM
 NIM : 202420019
 Kelas : MTI 23

Node	Deskripsi	Jml Kasus (S)	Count Play "YES"	Count Play "NO"	Entropy	Gain
1	Total	14	10	4	0.86312057	
	OUTLOOK					0.25852104
	Cloudy	4	4			
	Rainy	5	4	1	0.72192809	
	Sunny	5	2	3	0.97095059	
	TEMPERATURE					0.18385093
	Cool	4	4		0	
	Hot	4	2	2	1	
	Mild	6	4	2	0.91829583	
	HUMIDITY					0.3705065
	High	7	3	4	0.98522814	
	Normal	7	7			
	WINDY					0.00597771
	FALSE	8	6	2	0.81127812	
	TRUE	6	4	2	0.91829583	

Dapat diketahui dari data diatas bahwa atribut dengan Gain tertinggi adalah **HUMADITY** , yaitu sebesar 0.37. Jadi HUMADITY yang menjadi node akar. Ada dua nilai atribut dari HUMADITY , yaitu HIGH dan NORMAL . Dari kedua atribut, nilai atribut NORMAL adalah 1, yaitu keputusan Yes , sehingga tidak perlu perhitungan lebih lanjut, tetapi untuk HIGH masih perlu dilakukan perhitungan lagi. Apabila dibuat tree nya sebagai berikut :

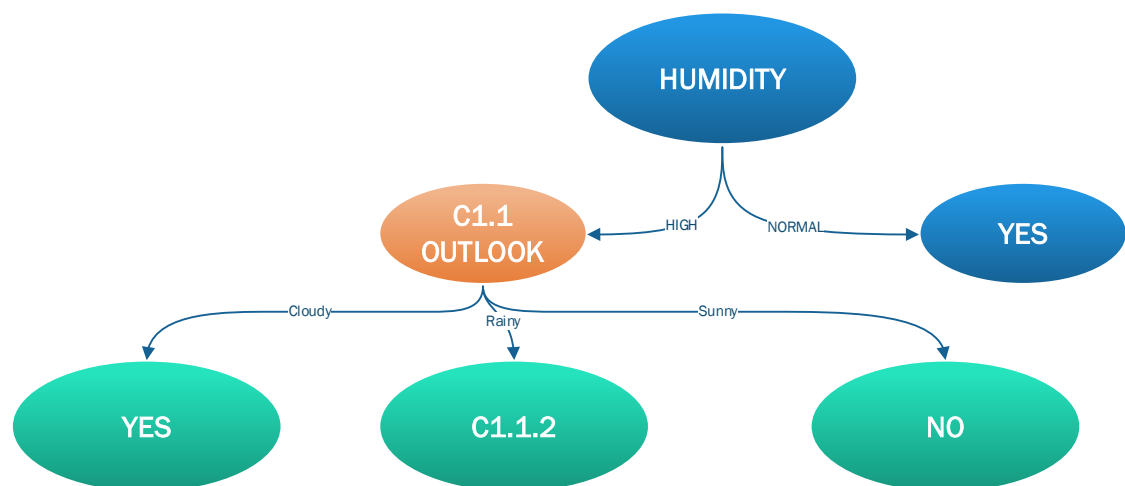


Menghitung jumlah kasus, jumlah kasus untuk keputusan Yes , jumlah kasus untuk keputusan No , dan Entropy dari semua kasus dan kasus yang dibagi berdasarkan atribut **OUTLOOK**, **TEMPERATURE**, dan **WINDY** yang dapat menjadi node akar dari nilai atribut HIGH . Kemudian lakukan perhitungan Gain untuk tiap atribut. Hasil perhitungan sebagai berikut :

Nama : Bhijanta Wyasa WM
 NIM : 202420019
 Kelas : MTI 23

Node	Deskripsi	Jml Kasus (S)	Count Play "YES"	Count Play "NO"	Entropy	Gain
1	Total	7	3	4	0.98522814	
	OUTLOOK					0.69951385
	Cloudy	2	2		0	
	Rainy	2	1	1	1	
	Sunny	3		3	0	
	TEMPERATURE					0.02024421
	Cool	0	0	0	0	
	Hot	3	1	2	0.91829583	
	Mild	4	2	2	1	
	HUMIDITY					0
	High	7	3	4	0.98522814	
	Normal	7	7			
	WINDY					0.02024421
	FALSE	4	2	2	1	
	TRUE	3	1	2	0.91829583	

Dapat diketahui bahwa atribut dengan Gain tertinggi adalah OUTLOOK , yaitu sebesar **0.67**. Jadi OUTLOOK dapat menjadi node cabang dari nilai atribut HIGH . Ada tiga nilai atribut dari OUTLOOK, yaitu CLOUDY, RAINY dan SUNNY. Dari ketiga nilai atribut tersebut, nilai atribut CLOUDY adalah 1, yaitu keputusannya Yes dan nilai atribut SUNNY menjadi keputusan No , sehingga tidak perlu dilakukan perhitungan lebih lanjut, tetapi nilai atribut RAINY masih perlu dilakukan perhitungan lagi.

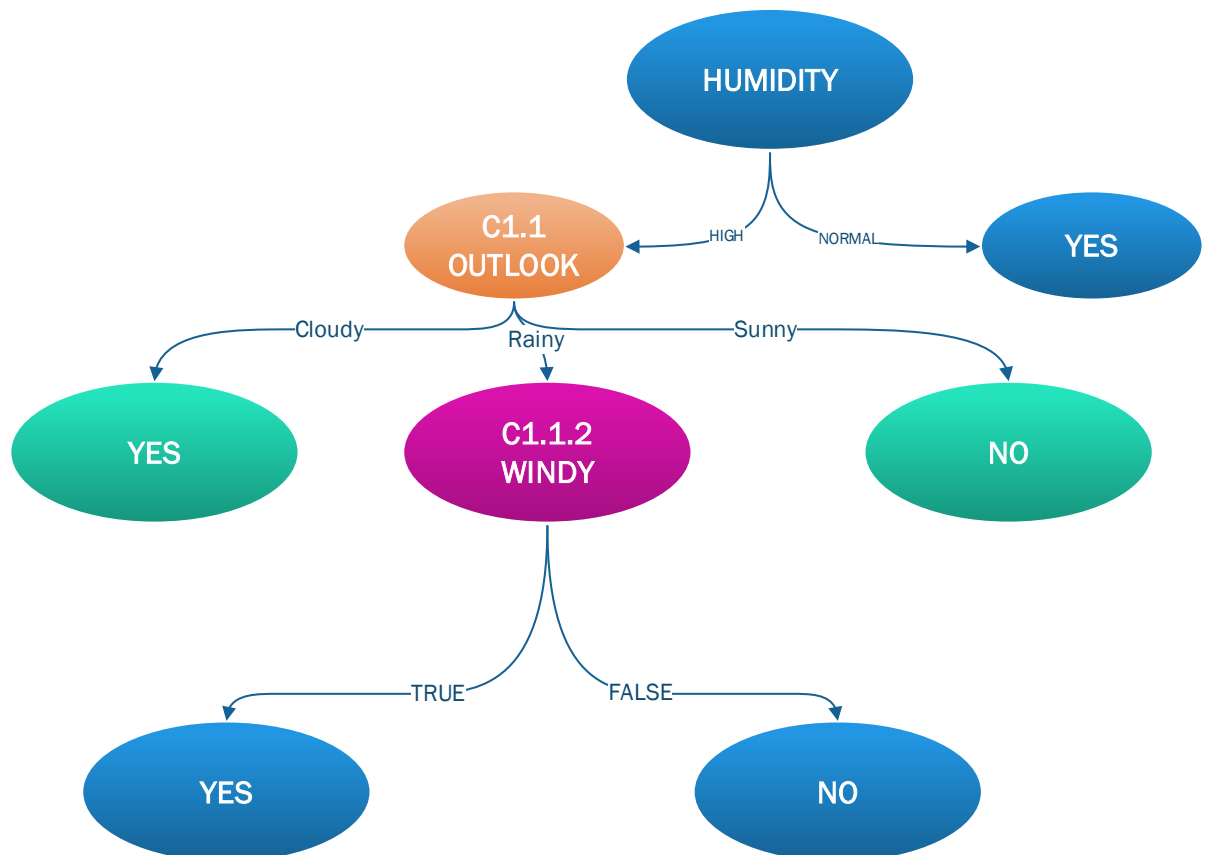


Menghitung jumlah kasus, jumlah kasus untuk keputusan Yes, jumlah kasus untuk keputusan No, dan **Entropy** dari semua kasus dan kasus yang dibagi berdasarkan atribut **TEMPERATURE** dan **WINDY** yang dapat menjadi node cabang dari nilai atribut **RAINY**. Kemudian lakukan perhitungan **Gain** untuk tiap-tiap atribut. Hasil perhitungan seperti yang terdapat pada table dibawah ini :

Nama : Bhijanta Wyasa WM
 NIM : 202420019
 Kelas : MTI 23

Node	Deskripsi	Jml Kasus (S)	Count Play "YES"	Count Play "NO"	Entropy	Gain
1	Total	2	1	1	1	
	OUTLOOK					0
	Cloudy	2	2		0	
	Rainy	2	1	1	1	
	Sunny	3		3	0	
	TEMPERATURE					0
	Cool	0	0	0	0	
	Hot	0	0	0	0	
	Mild	2	1	1	1	
	HUMIDITY					-2.4482985
	High	7	3	4	0.98522814	
	Normal	7	7			
	WINDY					1
	FALSE	1	1	0	0	
	TRUE	1	0	1	0	

Dapat diketahui bahwa atribut dengan *Gain* tertinggi adalah *WINDY*, yaitu sebesar 1. Jadi *WINDY* dapat menjadi *node* cabang dari nilai atribut *RAINY*. Ada dua nilai atribut dari *WINDY*, yaitu *FALSE* dan *TRUE*. Dari kedua nilai atribut tersebut, nilai atribut *FALSE* adalah 1, yaitu keputusannya *Yes* dan nilai atribut *TRUE* menjadi keputusan *No*, sehingga tidak perlu dilakukan perhitungan lebih lanjut untuk nilai atribut ini. Dapat digambarkan untuk pohon keputusannya sebagai berikut :



Nama : Bhijanta Wyasa WM
NIM : 202420019
Kelas : MTI 23

Merupakan gambar akhir dari proses *Data Mining* menggunakan algoritma C4.5 pada contoh kasus keputusan bermain tenis. Berdasarkan gambar di atas dapat ditarik kesimpulan bahwa jika *humidity* normal maka keputusannya adalah *yes*, jika *humidity* adalah *high*, *outlook* adalah *cloudy* maka keputusannya adalah bermain (*yes*), selanjutnya jika *humidity* adalah *high*, *outlook* adalah *rainy* dan *windy* adalah *true*, maka keputusannya adalah bermain (*yes*). Dan jika *humidity* adalah *high*, *outlook* adalah *rainy* dan *windy* adalah *false*, maka keputusannya adalah tidak bermain (*no*). Terakhir jika *humidity* adalah *high*, *outlook* adalah *no*, maka keputusannya adalah tidak bermain (*no*).

Algoritma C4.5

Algoritma yang dapat digunakan untuk memprediksi atau mengklasifikasi suatu kejadian dengan pembentukan pohon keputusan antara lain algoritma C4.5, yang merupakan salah satu algoritma induksi pohon keputusan yang dikembangkan oleh J. Ross Quinlan (Yamin, et al, 2014).

Algoritma C4.5 merupakan kelompok algoritma *decision tree*. Algoritma ini mempunyai input berupa *training samples* dan *samples*. *Training samples* merupakan data contoh yang digunakan untuk membangun sebuah *tree* yang telah diuji kebenarannya. Sedangkan *samples* merupakan *field-field* data yang digunakan sebagai parameter dalam klasifikasi data (Sujana, 2010).

Secara umum alur proses algoritma C4.5 untuk membangun pohon keputusan dalam *data mining* adalah sebagai berikut (Mahmud, et al, 2014):

- a. Pilih atribut sebagai akar
- b. Buat cabang untuk tiap-tiap nilai
- c. Bagi kasus dalam cabang.
- d. Ulangi proses untuk setiap cabang sampai semua kasus pada cabang memiliki kelas yang sama.

Nama : Cornelia Tri Wahyuni

NIM : 202420044

Kelas : MTI 23A

Mata Kuliah : Advanced Database

Dosen Pengasuh : Tri Basuki Kurniawan,S.Kom.,M.eng.,Ph.D.

Algoritma C 4.5

No	Outlook	Temperature	Humidity	Windy	Play
1	Sunny	Hot	High	FALSE	No
2	Sunny	Hot	High	TRUE	No
3	Cloudy	Hot	High	FALSE	Yes
4	Rainy	Mild	High	FALSE	Yes
5	Rainy	Cool	Normal	FALSE	Yes
6	Rainy	Cool	Normal	TRUE	Yes
7	Cloudy	Cool	Normal	TRUE	Yes
8	Sunny	Mild	High	FALSE	No
9	Sunny	Cool	Normal	FALSE	Yes
10	Rainy	Mild	Normal	FALSE	Yes
11	Sunny	Mild	Normal	TRUE	Yes
12	Cloudy	Mild	High	TRUE	Yes
13	Cloudy	Hot	Normal	FALSE	Yes
14	Rainy	Mild	High	TRUE	No

Perhitungan 1

Node			Jumlah Kasus	No (S1)	Yes (S2)	Entropy	Gain
1	Total		14	4	10	0,863120569	0,258521037
	Outlook						
		Cloudy	4	0	4	0	
		Rainy	5	1	4	0,721928095	
		Sunny	5	3	2	0,970950594	
	Temperature						0,183850925
		Cool	4	0	4	0	
		Hot	4	2	2	1	
		Mild	6	2	4	0,918295834	
	Humidity						0,370506501
		High	7	4	3	0,985228136	

		Normal	7	0	7	0	
	Windy						0,005977711
		FALSE	8	2	6	0,811278124	
		TRUE	6	2	4	0,918295834	

Perhitungan Entropy : rumus :

$$\text{Entropy (S)} = \sum_{i=1}^n - p_i * \log_2 p_i$$

Keterangan

S : himpunan kasus

A : fitur

n : jumlah partisi S

pi : proporsi dari Si terhadap S

$$\text{Entropy Total} = ((-10 \div 14 * \log_2 (10 \div 14) + -4 \div 14 * \log_2 (4 \div 14))) = 0,863120569$$

$$\text{Entropy Cloudy} = ((-4 \div 4 * \log_2 (4 \div 4) + -0 \div 4 * \log_2 (0 \div 4))) = 0$$

$$\text{Entropy Rainy} = ((-4 \div 5 * \log_2 (4 \div 5) + -1 \div 5 * \log_2 (1 \div 5))) = 0,721928095$$

$$\text{Entropy Sunny} = ((-2 \div 5 * \log_2 (2 \div 5) + -3 \div 5 * \log_2 (3 \div 5))) = 0,970950594$$

$$\text{Entropy Cool} = ((-4 \div 4 * \log_2 (4 \div 4) + -0 \div 4 * \log_2 (0 \div 4))) = 0$$

$$\text{Entropy Hot} = ((-2 \div 4 * \log_2 (2 \div 4) + -2 \div 4 * \log_2 (2 \div 4))) = 1$$

$$\text{Entropy Mild} = ((-4 \div 6 * \log_2 (4 \div 6) + -2 \div 6 * \log_2 (2 \div 6))) = 0,918295834$$

$$\text{Entropy High} = ((-3 \div 7 * \log_2 (3 \div 7) + -4 \div 7 * \log_2 (4 \div 7))) = 0,985228136$$

$$\text{Entropy Normal} = ((-7 \div 7 * \log_2 (7 \div 7) + -0 \div 7 * \log_2 (0 \div 7))) = 0$$

$$\text{Entropy False} = ((-6 \div 8 * \log_2 (6 \div 8) + -2 \div 8 * \log_2 (2 \div 8))) = 0,811278124$$

$$\text{Entropy True} = ((-4 \div 6 * \log_2 (4 \div 6) + -2 \div 6 * \log_2 (2 \div 6))) = 0,918295834$$

Perhitungan Gain : rumus :

$$\text{Gain (S,A)} = \text{Entropy (S)} - \sum_{i=1}^n \frac{|S_i|}{|S|} * \text{Entropy (Si)}$$

Keterangan:

- S : himpunan

- A : atribut

- n : jumlah partisi atribut A

- | Si | : jumlah kasus pada partisi ke-i

- | S | : jumlah kasus dalam S

$$\text{Gain Outlook} = (0,863120569) - ((4 \div 14) * 0) - ((5 \div 14) * 0,721928095) - ((5 \div 14) * 0,970950594)$$

$$= 0,258521037$$

$$\text{Gain Temperature} = (0,863120569) - (4 \div 14) * 0 - (4 \div 14) * 1 - (6 \div 14) * 0,918295834 \\ = 0,183850925$$

$$\text{Gain Humidity} = (0,863120569) - (7 \div 14) * 0,985228136 - (7 \div 14) * 0 \\ = 0,370506501$$

$$\text{Gain Windy} = (0,863120569) - (8 \div 14) * 0,811278124 - (6 \div 14) * 0,918295834 \\ = 0,005977711$$

Seperti yg terlihat pada perhitungan diatas, diperoleh bahwa atribut dgn Gain tertinggi adalah Humidity = **0,370506501**

Maka Humidity menjadi node akar

Humidity memiliki dua nilai yaitu "High" dan "Normal"

Humidity -> "Normal" sdh mengklasifikasikan kasus menjadi 1 yaitu keputusannya "yes"

Untuk humidity -> "High" msh perlu dilakukan perhitungan lagi (karena masih terdapat "yes" dan "no")

Humidity "High"

No	Outlook	Temperature	Humidity	Windy	Play
1	Sunny	Hot	High	FALSE	No
2	Sunny	Hot	High	TRUE	No
3	Cloudy	Hot	High	FALSE	Yes
4	Rainy	Mild	High	FALSE	Yes
5	Sunny	Mild	High	FALSE	No
6	Cloudy	Mild	High	TRUE	Yes
7	Rainy	Mild	High	TRUE	No

Perhitungan 1.1

Node			Jumlah Kasus	No (S1)	Yes (S2)	Entropy	Gain
1.1	Total		7	4	3	0,985228136	
	Outlook						0,69951385
		Cloudy	2	0	2	0	
		Rainy	2	1	1	1	
		Sunny	3	3	0	0	
	Temperature						0,020244207
		Cool	0	0	0	0	
		Hot	3	2	1	0,918295834	
		Mild	4	2	2	1	
	Windy						0,020244207

		FALSE	4	2	2	1	
		TRUE	3	2	1	0,918295834	

Perhitungan Entropy Humadity “High” 1.1

Entropy Total	$= ((-3 \div 7 * \log_2 (3 \div 7) + -4 \div 7 * \log_2 (4 \div 7))) = 0,985228136$					
Entropy Cloudy	$= ((-2 \div 2 * \log_2 (2 \div 2) + -0 \div 2 * \log_2 (0 \div 2))) = 0$					
Entropy Rainy	$= ((-1 \div 2 * \log_2 (1 \div 2) + -1 \div 2 * \log_2 (1 \div 2))) = 1$					
Entropy Sunny	$= ((-0 \div 3 * \log_2 (0 \div 3) + -3 \div 3 * \log_2 (3 \div 3))) = 0$					
Entropy Cool	$= ((-0 \div 0 * \log_2 (0 \div 0) + -0 \div 0 * \log_2 (0 \div 0))) = 0$					
Entropy Hot	$= ((-1 \div 3 * \log_2 (1 \div 3) + -2 \div 3 * \log_2 (2 \div 3))) = 0,918295834$					
Entropy Mild	$= ((-2 \div 4 * \log_2 (2 \div 4) + -2 \div 4 * \log_2 (2 \div 4))) = 1$					
Entropy False	$= ((-2 \div 4 * \log_2 (2 \div 4) + -2 \div 4 * \log_2 (2 \div 4))) = 1$					
Entropy True	$= ((-1 \div 3 * \log_2 (1 \div 3) + -2 \div 3 * \log_2 (2 \div 3))) = 0,918295834$					

Perhitungan Gain Humadity “High” 1.1

$$\begin{aligned} \text{Gain Outlook} &= (0,985228136) - (2 \div 7) * 0 - (2 \div 7) * 1 - (3 \div 7) * 0 \\ &= 0,69951385 \end{aligned}$$

$$\begin{aligned} \text{Gain Temperature} &= (0,985228136) - (0 \div 7) * 0 - (3 \div 7) * 0,918295834 - (4 \div 7) * 1 \\ &= 0,020244207 \end{aligned}$$

$$\begin{aligned} \text{Gain Windy} &= (0,985228136) - (4 \div 7) * 1 - (3 \div 7) * 0,918295834 \\ &= 0,020244207 \end{aligned}$$

Didapat Gain tertinggi -> outlook -> **0,69951385**

Maka “Outlook” menjadi node cabang dari atribut humidity yg bernilai “High”

Berdasarkan atribut “Outlook” terdpt 3 nilai

Cloudy

Rainy

Sunny

Karena “Cloudy” pasti bernilai “Yes” dan “Sunny” pasti bernilai “No”, maka tidak perlu dilakukan perhitungan lagi.

Sedangkan “Rainy” bernilai “yes” dan “No”, maka masih perlu dilakukan perhitungan lagi

Humadity “High” dan Outlook “Rainy”

No	Outlook	Temperature	Humidity	Windy	Play
4	Rainy	Mild	High	FALSE	Yes
7	Rainy	Mild	High	TRUE	No

Perhitungan 1.1.2

Node			Jumlah Kasus	No (S1)	Yes (S2)	Entropy	Gain
1..1.2	Total		2	1	1	1	
	Temperature						0
		Cool	0	0	0	0	
		Hot	0	0	0	0	
		Mild	2	1	1	1	
	Windy						1
		FALSE	1	0	1	0	
		TRUE	1	1	0	0	

Perhitungan Entropy "Rainy" 1.1.2

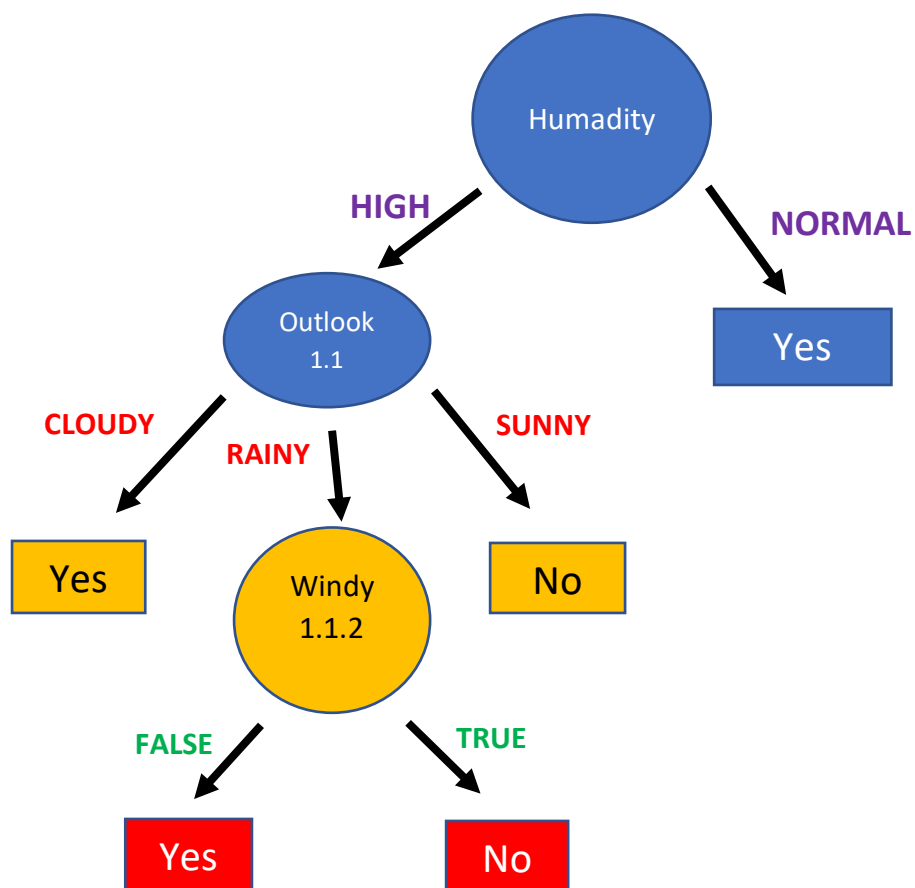
Entropy Total	$= ((-1 \div 2) * \log_2 (1 \div 2) + (-1 \div 2) * \log_2 (1 \div 2))$	= 1
Entropy Cool	$= ((-0 \div 0) * \log_2 (0 \div 0) + -0 \div 0 * \log_2 (0 \div 0))$	= 0
Entropy Hot	$= ((-0 \div 0) * \log_2 (0 \div 0) + -0 \div 0 * \log_2 (0 \div 0))$	= 0
Entropy Mild	$= ((-1 \div 2 * \log_2 (1 \div 2) + -1 \div 2 * \log_2 (1 \div 2))$	= 1
Entropy False	$= ((-1 \div 1 * \log_2 (1 \div 1) + -0 \div 1 * \log_2 (0 \div 1))$	= 0
Entropy True	$= ((-0 \div 1 * \log_2 (0 \div 1) + -1 \div 0 * \log_2 (1 \div 0))$	= 0

Perhitungan Gain "Rainy" 1.1.2

$$\text{Gain Temperature} = (1) - (0 \div 2) * 0 - (0 \div 2) * 0 - (2 \div 2) * 1 \\ = 0$$

$$\text{Gain Windy} = (1) - (1 \div 2) * 0 - (1 \div 2) * 0 \\ = 1$$

POHON KEPUTUSAN



Kelebihan Algoritma C4.5 :

1. C4.5 mampu menangani atribut yang kosong (Missing Value).
2. C4.5 mampu menangani atribut dengan kontinu.
3. C4.5 memangkas pohon keputusan untuk mengatasi overfitting.

Kekurangan Algoritma C4.5 :

1. Algoritma C4.5 hanya dapat digunakan untuk menangani sampel-sampel yang dapat disimpan secara keseluruhan dan pada waktu yang bersamaan di memori.

Nama : Cynthia Anisa Agatha

NIM 202420022

No	Outlook	Temperature	Humidity	Windy	Play
1.	Sunny	Hot	High	FALSE	No
2.	Sunny	Hot	High	TRUE	No
3.	Cloudy	Hot	High	FALSE	Yes
4.	Rainy	Mild	High	FALSE	Yes
5.	Rainy	Cool	Normal	FALSE	Yes
6.	Rainy	Cool	Normal	TRUE	Yes
7.	Cloudy	Cool	Normal	TRUE	Yes
8.	Sunny	Mild	High	FALSE	No
9.	Sunny	Cool	Normal	FALSE	Yes
10.	Rainy	Mild	Normal	FALSE	Yes
11.	Sunny	Mild	Normal	TRUE	Yes
12.	Cloudy	Mild	High	TRUE	Yes
13.	Cloudy	Hot	Normal	FALSE	Yes
14.	Rainy	Mild	High	TRUE	No

Untuk menentukan bermain atau tidak, maka kriteria yang diperlukan untuk dilakukan perhitungan yaitu:

- Cuaca
- Angin
- Kelembapan
- Temperatur Udara

Dari tabel diatas akan dibuat tabel dan pohon keputusan untuk menentukan bermain atau tidak dengan melihat *Outlook*, *Temperature*, *Humidity* dan *Windy*.

Berdasarkan algoritma langkah-langkah pada umumnya meliputi:

1. Pilih atribut sebagai akar
2. Buat cabang untuk setiap nilai
3. Bagikan kasus berupa cabang
4. Lakukan secara berulang untuk setiap cabang sampai semua kasus pada cabang memiliki kelas yang sama
5. Memilih atribut berdasarkan nilai *gain* tertinggi dari atribut-atribut yang ada.

Formula untuk menemukan nilai *Gain*

$$Gain(S, A) = Entropy(S) - \sum_{i=1}^n * Entropy(S_i)$$

Keterangan :

S : himpunan

A : atribut

n : jumlah partisi atribut A

| Si | : jumlah kasus pada partisi ke-i

| S | : jumlah kasus dalam S

Formula untuk menghitung nilai *Entropy*

$$Entropy(S) = \sum_{i=1}^n \rho_i * \log_2 \rho_i$$

Keterangan :

S : himpunan kasus

A : fitur

n : jumlah partisi S

ρ_i : proporsi dari Si terhadap S

Lakukan Langkah 1

- Menghitung jumlah seluruh kasus yang ada, per jumlah kasus berkeputusan “Yes” maupun “No”. Seperti yang akan ditunjukkan dalam tabel dibawah ini.

Node			Jumlah Kasus (S)	Yes (S_2)	No (S_1)	<i>Entropy</i>	<i>Gain</i>
1	Total		14	10	4		
	Outlook						
		Cloudy	4	0	4		
		Rainy	5	1	4		
		Sunny	5	3	2		
	Temperature						
		Cool	4	4	0		
		Mild	6	4	2		
		Hot	4	2	2		
	Humidity						
		High	7	3	4		
		Normal	7	7	0		
	Windy						
		TRUE	6	4	2		
		FALSE	8	6	2		

Setelah didapatkan data diatas kemudian dilanjutkan dengan perhitungan total nilai *Entropy* dengan menggunakan formula sebelumnya

$$Entropy(S) = \sum_{i=1}^n p_i * \log_2 p_i$$

$$Entropy(Total) = -4/14 * \log_2(4/14) + -10/14 * \log_2(10/14) \\ = 0,863120569$$

Hasil yang didapatkan apabila dimasukkan ke dalam tabel sebagai berikut

Node			Jumlah Kasus (S)	Yes (S_2)	No (S_1)	<i>Entropy</i>	<i>Gain</i>
1	Total		14	10	4	0,863120569	
...							

- Menghitung *Entropy* dari semua kasus yang terbagi berdasarkan atribut *Outlook*, *Temperature*, *Humidity* dan *Windy*.

Menghitung *Gain* pada baris *Outlook*

$$Gain(Total, Outlook) \\ = Entropy(S) - \sum_{i=1}^n Outlook_i / Total * Entropy(Outlook_i) \\ = 0,863120569 - ((4/14 * 0) + (5/14 * 0,72) + (5/14 * 0,97)) \\ = 0,258521037$$

Kemudian dari hasil *gain* pada *outlook* yang didapatkan ditunjukkan pada tabel dibawah ini.

Node			Jumlah Kasus (S)	Yes (S_2)	No (S_1)	<i>Entropy</i>	<i>Gain</i>
1	Total		14	10	4	0,863120569	
	Outlook						0,258521037

Setelah didapatkan nilai *gain* pada baris *outlook* dilanjutkan dengan melakukan perhitungan *gain* untuk baris *temperature*, *humidity* dan *windy*

- Melakukan perhitungan untuk nilai *Gain* di setiap atributnya.

Menghitung nilai *Gain* pada baris *temperature*

$$Gain(Total, Temperature) \\ = Entropy(Total) - \sum_{i=1}^n Temperature / Total * Entropy(Temperature_i) \\ = 0,863120569 - ((4/14 * 1) + (6/14 * 0,918295834) + (4/14 * 0)) \\ = 0,183850925$$

Kemudian dari hasil *gain* pada *temperature* yang didapatkan ditunjukkan pada tabel dibawah ini.

Node			Jumlah Kasus (S)	Yes (S ₂)	No (S ₁)	Entropy	Gain
1	Total		14	10	4	0,863120569	
	Outlook						0,258521037
		Sunny	5	2	3	0,970950594	
		Cloudy	4	4	0	0	
		Rainy	5	4	1	0,721928095	
	Temperature						0,183850925
		Hot	4	2	2	1	
		Mild	6	4	2	0,918295834	
		Cool	4	4	0	0	

Menghitung nilai *gain* pada baris *humidity*

$$\begin{aligned}
 & \text{Gain (Total, Humidity)} \\
 &= \text{Entropy(Total)} - \sum_{i=1}^n \text{Humidity/Total} * \text{Entropy(Humidity)} \\
 &= 0,863120569 - ((7/14 * 0,985228136) + (7/14 * 0)) \\
 &= 0,370506501
 \end{aligned}$$

Kemudian dari hasil *gain* pada *humidity* yang didapatkan ditunjukkan pada tabel dibawah ini.

Node			Jumlah Kasus (S)	Yes (S ₂)	No (S ₁)	Entropy	Gain
1	Total		14	10	4	0,863120569	
	Outlook						0,258521037
		Sunny	5	2	3	0,970950594	
		Cloudy	4	4	0	0	
		Rainy	5	4	1	0,721928095	
	Temperature						0,183850925
		Hot	4	2	2	1	
		Mild	6	4	2	0,918295834	
		Cool	4	4	0	0	
	Humidity						0,370506501
		High	7	3	4	0,985228136	
		Normal	7	7	0	0	

Menghitung nilai *gain* pada baris *windy*

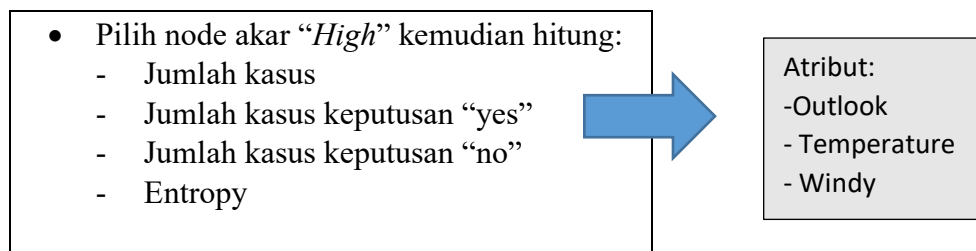
$$\begin{aligned}
 & \text{Gain (Total, Windy)} \\
 &= \text{Entropy(Total)} - \sum_{i=1}^n \text{Windy/Total} * \text{Entropy(Windy)} \\
 &= 0,863120569 - ((8/14 * 0,811278124) + (6/14 * 0,918295834)) = 0,005977711
 \end{aligned}$$

Node			Jumlah Kasus (S)	Yes (S_2)	No (S_1)	Entropy	Gain
1	Total		14	10	4	0,863120569	
	Outlook						0,258521037
		Sunny	5	2	3	0,970950594	
		Cloudy	4	4	0	0	
		Rainy	5	4	1	0,721928095	
	Temperature						0,183850925
		Hot	4	2	2	1	
		Mild	6	4	2	0,918295834	
		Cool	4	4	0	0	
	Humidity						0,370506501
		High	7	3	4	0,985228136	
		Normal	7	7	0	0	
	Windy						0,005977711
		TRUE	6	4	2	0,918295834	
		FALSE	8	6	2	0,811278124	

Dari tabel diatas diperoleh bahwa atribut dengan *Gain* tertinggi yaitu *Humidity* 0,37. Maka dari itu, *humidity* dijadikan node akar. *Humidity* memiliki dua *instance* yaitu *High* dan *Normal*. *Instance* dari *Humidity* yaitu *Normal* sudah mengklasifikasikan kasus menjadi 1 yaitu keputusannya “yes”. Untuk *instance high* pada atribut *humidity* masih perlu dilakukan perhitungan lagi karena masih terdapat “yes” dan “no”).

Pohon Keputusan Node 1

Perincian Algoritma (Langkah 2)

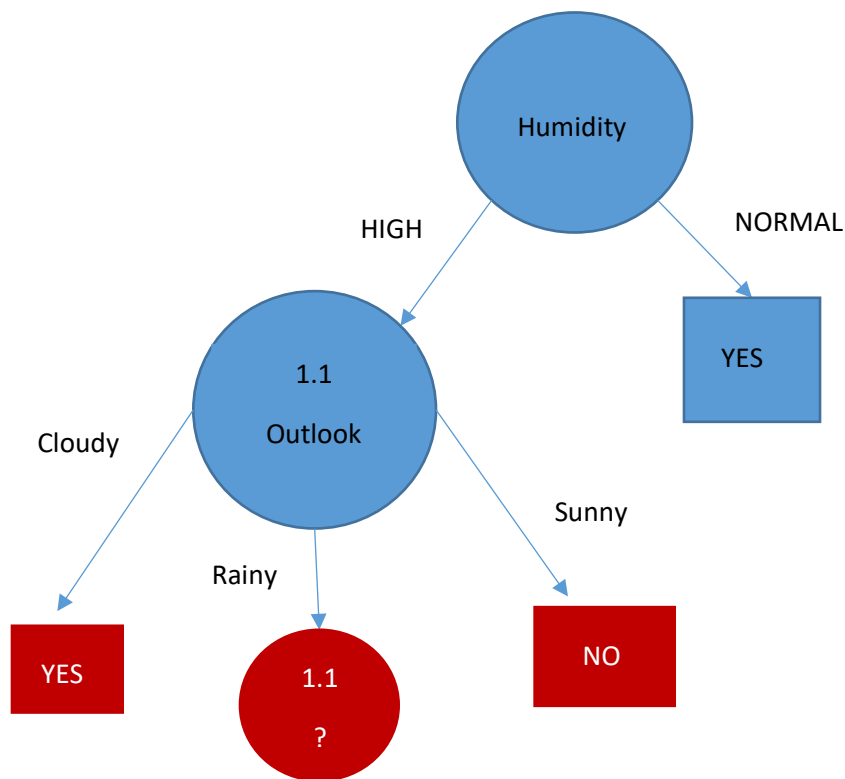


Dari proses diatas maka didapatkan hasil perhitungan Langkah kedua yang akan ditunjukkan pada tabel dibawah ini.

Node			Jumlah Kasus (S)	No (S_1)	Yes (S_2)	Entropy	Gain
1.1	Humidity High		7	4	3	0,985228136	
	Outlook						0,69951385
		Cloudy	2	0	2	0	

		Rainy	2	1	1	1	
		Sunny	3	3	0	0	
	Temperature						0,020244207
		Cool	0	0	0	0	
		Hot	3	2	1	1	
		Mild	4	2	2	2	
	Windy						0,020244207
		False	4	2	2	1	
		True	3	2	1	0,918295834	

Maka didapatkan *gain* tertinggi pada atribut *outlook* 0,69. Dan *outlook* menjadi *node* cabang dari atribut *humidity* yang bernilai “*High*”. Berdasarkan atribut “*Outlook*” terdapat 3 nilai yaitu *cloudy*, *rainy* dan *sunny*. Karena *cloudy* pasti bernilai “*Yes*” dan “*Sunny*” pasti bernilai “*No*”, maka tidak perlu dilakukan perhitungan lagi. Sedangkan “*Rainy*” bernilai “*Yes*” dan “*No*”, maka masih perlu dilakukan perhitungan lagi.



Pohon Keputusan 1.1

Perincian Algoritma (Langkah 3)

- Pilih node akar “*Rainy*” kemudian hitung:
 - Jumlah kasus
 - Jumlah kasus keputusan “*Yes*”
 - Jumlah kasus keputusan “*No*”
 - *Entropy*

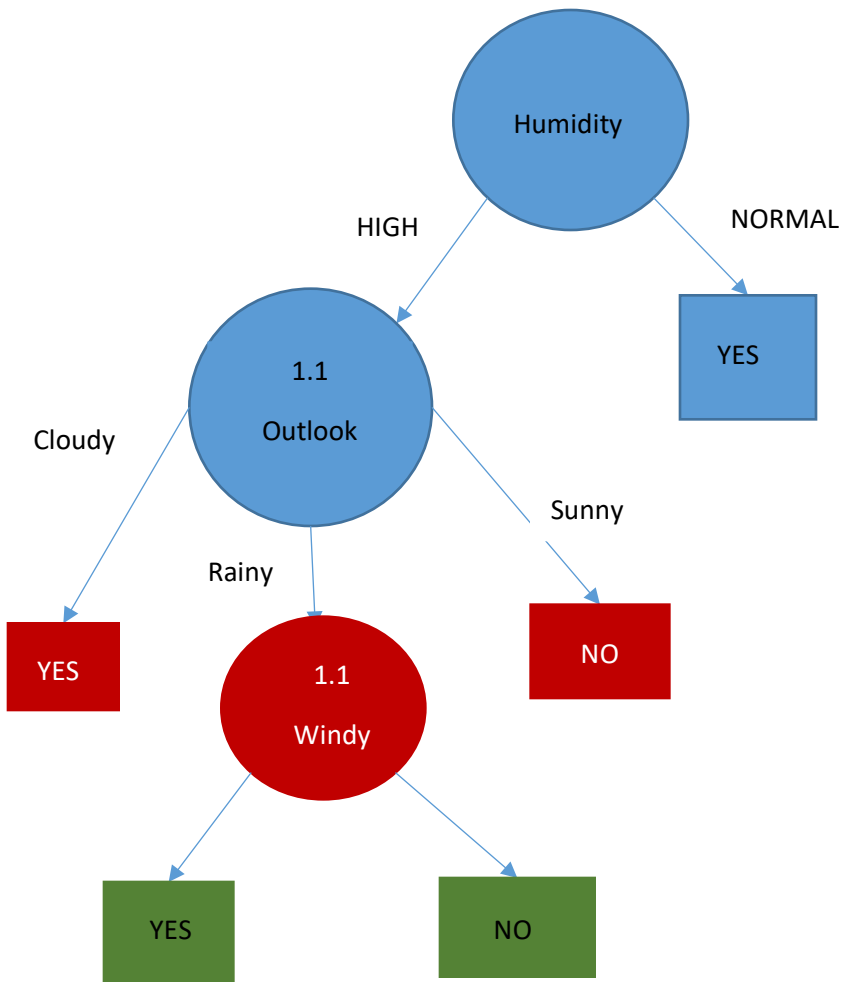


Atribut:
-Temperature
-Windy

Hasil perhitungan (Langkah 3)

Node			Jumlah Kasus (S)	No (S_1)	Yes (S_2)	<i>Entropy</i>	<i>Gain</i>
1.1	Humidity “High” dan Outlook “Rainy”		2	1	1	1	
	Temperature						0
		Cool	0	0	0	0	
		Hot	0	0	0	0	
		Mild	2	1	1	1	
	Windy						1
		False	1	0	1	0	
		True	1	1	0	0	

Maka hasil perhitungan dari langkah ketiga didapatkan *gain* tertinggi pada atribut *Windy* yaitu bernilai 1. Maka *windy* menjadi node cabang dari atribut *humidity* yang bernilai *high* dan *outlook* yang bernilai *rainy*. Berdasarkan atribut *windy* terdapat dua nilai yaitu *True* dan *False* dikarenakan nilai *True* telah diklasifikasikan dengan jelas yaitu bernilai *No* dan *False* jelas bernilai *Yes*, maka tidak perlu dilanjutkan perhitungan.



Pohon Keputusan node 1.1.2

Hasil dari perhitungan langkah ketiga yaitu, berdasarkan node 1.1.2 maka semua kasus yang ada telah mendapatkan kelas sehingga pohon keputusan diatas merupakan pohon keputusan terakhir yang terbentuk.

Nama : Efrik Kartono Ahsa
NIM : 202420030

Tugas 05

Dari dataset berikut ini, silahkan buat perhitungan manual dari algoritma decision tree C45 secara lengkap dan gambarkan pohon keputusan yang diperoleh.

No	Outlook	Temperature	Humidity	Windy	Play
1	Sunny	Hot	High	FALSE	No
2	Sunny	Hot	High	TRUE	No
3	Cloudy	Hot	High	FALSE	Yes
4	Rainy	Mild	High	FALSE	Yes
5	Rainy	Cool	Normal	FALSE	Yes
6	Rainy	Cool	Normal	TRUE	Yes
7	Cloudy	Cool	Normal	TRUE	Yes
8	Sunny	Mild	High	FALSE	No
9	Sunny	Cool	Normal	FALSE	Yes
10	Rainy	Mild	Normal	FALSE	Yes
11	Sunny	Mild	Normal	TRUE	Yes
12	Cloudy	Mild	High	TRUE	Yes
13	Cloudy	Hot	Normal	FALSE	Yes
14	Rainy	Mild	High	TRUE	No

Kumpulkan dalam format MS Word. Sertakan penjelasan dan pendapat anda tentang algoritma ini, kelebihan dan kekurangannya berbanding metode lain yang sejenis (tentukan sendiri metode pembandingnya).

Penyelesaian:

Perhitungan Gain

$$\text{Gain}(S,A) = \text{Entropy}(S) - \sum_{i=1}^n \frac{|S_i|}{|S|} * \text{Entropy}(S_i)$$

Keterangan:

- S : himpunan
- A : atribut
- n : jumlah partisi atribut A
- | S_i | : jumlah kasus pada partisi ke-i
- | S | : jumlah kasus dalam S

Nama : Efrik Kartono Ahsa
NIM : 202420030

Menghitung Nilai Entropy

$$\text{Entropy (S)} = \sum_{i=1}^n - p_i * \log_2 p_i$$

Keterangan:

S : himpunan kasus

A : fitur

n : jumlah partisi S

pi : proporsi dari Si terhadap S

Perincian algoritma (langkah 1)

- Menghitung jumlah kasus seluruhnya, jumlah berkeputusan “Yes” maupun “No”.
- Menghitung Entropy dari semua kasus yg terbagi berdasarkan atribut “Outlook”, “Temperature”, “Humidity”, “Windy”.
- Lakukan penghitungan Gain utk setiap atributnya

Perhitungan

Node			Jumlah Kasus (S)	Tidak (S ₁)	Ya (S ₂)	Entropy	Gain
1	TOTAL		14		10	0,863120569	
	OUTLOOK						0,258521037
		CLOUDY	4	0	4	0	
		RAINY	5	1	4	0,721928095	
		SUNNY	5	3	2	0,970950594	
	TEMPERATURE						0,183850925
		COOL	4	0	4	0	
		HOT	4	2	2	1	
		MILD	6	2	4	0,918295834	
	HUMIDITY						0,370506501
		HIGH	7	4	3	0,985228136	
		NORMAL	7	0	7	0	
							0,005977711
	WINDY						
		FALSE	8	2	6	0,811278124	
		TRUE	6	2	4	0,918295834	

Perhitungan Total Entropy

Node			Jumlah Kasus (S)	Tidak (S ₁)	Ya (S ₂)	Entropy	Gain
1	TOTAL		14		10	0,863120569	
...	...	...	...	...	...	...	...

Nama : Efrik Kartono Ahsa
NIM : 202420030

$$\text{Entropy (S)} = \sum_{i=1}^n - p_i * \log_2 p_i$$

$$\text{Entropy (Total)} = - 4/14 * \log_2 (4/14) + - 10/14 * \log_2 (10/14)$$

Menghitung gain pada baris Outlook

Gain (Total, Outlook)

$$= \text{Entropy}(\text{Total}) - \sum_{i=1}^n \frac{|\text{Outlook}_i|}{|\text{Total}|} * \text{Entropy}(\text{Outlook}_i)$$
$$= 0,863120569 - ((4/14 * 0) + (5/14 * 0,72) + (5/14 * 0,97))$$
$$= 0,258521037$$

Lakukan Hitung Gain untuk temperature, humidity dan windy

Sepert yg terlihat pd tabel, diperoleh bhw atribut dgn Gain tertinggi adalah Humidity -> 0,37

Maka Humidity menjadi node akar

Humidity memiliki dua nilai yaitu "High" dan "Normal"

Humidity -> "Normal" sdh mengklasifikasikan kasus menjadi 1 yaitu keputusannya "yes"

Untuk humidity -> "High" msh perlu dilakukan perhitungan lagi (karena masih terdapat "yes" dan "no")

Pohon Keputusan Node 1

Perincian Algoritma (Langkah 2)

- Pilih node akar "High" dan hitung:
 - Jumlah kasus
 - Jumlah kasus keputusan "Yes"
 - Jumlah kasus keputusan "No"
 - Entropy



Atribut:
• Outlook
• Temperature
• Windy

Nama : Efrik Kartono Ahsa
NIM : 202420030

Hasil perhitungan (Langkah 2)

Node			Jumlah Kasus (S)	Tidak (S ₁)	Ya (S ₂)	Entropy	Gain
1.1	HUMIDITY HIGH		7	4	3	0,985228136	
	OUTLOOK						0,69951385
		CLOUDY	2	0	2	0	
		RAINY	2	1	1	1	
		SUNNY	3	3	0	0	
	TEMPERATURE						0,020244207
		COOL	0	0	0	0	
		HOT	3	2	1	1	
		MILD	4	2	2	2	
	WINDY						
		FALSE	4	2	2	1	
		TRUE	3	2	1	0,918295834	

Hasil perhitungan (Langkah 2)

Didapat Gain tertinggi -> outlook -> 0,69

Maka "Outlook" menjadi node cabang dari atribut humidity yg bernilai "High"

Berdasarkan atribut "Outlook" terdpt 3 nilai

Cloudy

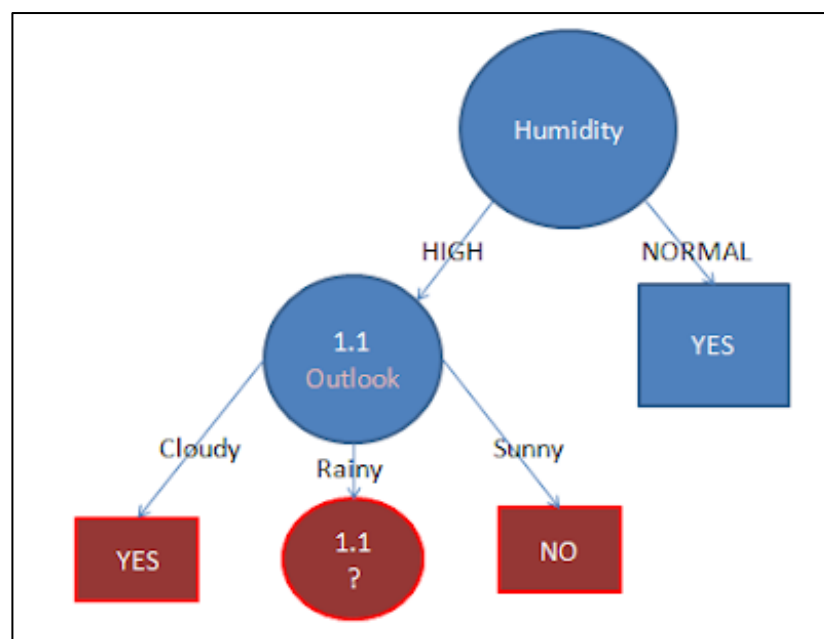
Rainy

Sunny

Krn "Cloudy" pasti bernilai "Yes" dan "Sunny" pasti bernilai "No", maka tdk perlu dilakukan perhitungan lagi

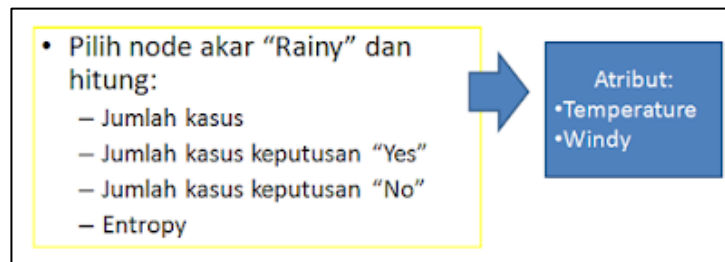
Sedangkan "Rainy" bernilai "yes" dan "No", maka masih perlu dilakukan perhitungan lagi

Pohon keputusan node 1.1



Nama : Efrik Kartono Ahsa
NIM : 202420030

Perincian algoritma (Langkah 3)



Hasil perhitungan (Langkah 3)

Node			Jumlah Kasus (S)	Tidak (S ₁)	Ya (S ₂)	Entropy	Gain
1.1	HUMIDITY "HIGH" DAN OUTLOOK "RAINY"		2	1	1	1	
	TEMPERATURE						0
		COOL	0	0	0	0	
		HOT	0	0	0	0	
		MILD	1	1	1	1	
	WINDY						
		FALSE	1	0	1	0	
		TRUE	1	1	0	0	

Hasil perhitungan (Langkah 3)

Didapat Gain tertinggi -> Windy -> 1

Maka "Windy" menjadi node cabang dari atribut humidity yg bernilai "High" dan outlook yg bernilai "Rainy"

Berdasarkan atribut "Windy" terdpt 2 nilai

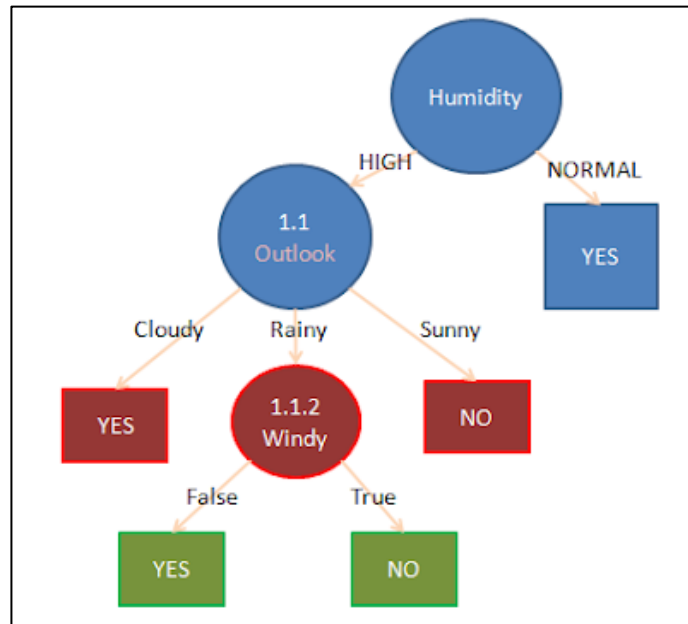
True

False

Karena "True" sdh terklasifikasi pasti bernilai "No" dan "False" pasti bernilai "Yes", maka tidak perlu dilakukan perhitungan lagi

Nama : Efrik Kartono Ahsa
NIM : 202420030

Pohon keputusan node 1.1.2



Hasil perhitungan (Langkah 3)

Berdasarkan node 1.1.2, maka:

“Semua kasus sudah masuk dapat kelas”

Sehingga pohon keputusan diatas merupakan pohon keputusan terakhir yang terbentuk

-Tugas 05

Nama : GATOT ARIFianto
Nim : 202420029
Kelas : MTI 23 Reguler A
Mk : Advanced Database

Dari dataset berikut ini, silahkan buat perhitungan manual dari algoritma decision tree C45 secara lengkap dan gambarkan pohon keputusan yang diperoleh.

No	Outlook	Temperature	Humidity	Windy	Play
1	Sunny	Hot	High	FALSE	No
2	Sunny	Hot	High	TRUE	No
3	Cloudy	Hot	High	FALSE	Yes
4	Rainy	Mild	High	FALSE	Yes
5	Rainy	Cool	Normal	FALSE	Yes
6	Rainy	Cool	Normal	TRUE	Yes
7	Cloudy	Cool	Normal	TRUE	Yes
8	Sunny	Mild	High	FALSE	No
9	Sunny	Cool	Normal	FALSE	Yes
10	Rainy	Mild	Normal	FALSE	Yes
11	Sunny	Mild	Normal	TRUE	Yes
12	Cloudy	Mild	High	TRUE	Yes
13	Cloudy	Hot	Normal	FALSE	Yes
14	Rainy	Mild	High	TRUE	No

Kumpulkan dalam format MS Word. Sertakan penjelasan dan pendapat anda tentang algoritma ini, kelebihan dan kekurangannya berbanding metode lain yang sejenis (tentukan sendiri metode pembandingnya).

TABEL PERHITUNGAN NODE 1

Rumus :

Entropy = $((-\text{don't play} / \text{jumlah total kasus}) * \log((\text{don't play} / \text{jumlah total kasus});2)) + ((-\text{play} / \text{jumlah total kasus}) * \log((\text{play} / \text{jumlah total kasus});2))$

$$1. ((-4/14) * \log(4/14);2) + ((-10/14) * \log(10/14);2) = 0,863121$$

Gain = jumlah total entropy – ((jumlah kasus variable x/jumlah total kasus)*entropy variable x)+((jumlah kasus variable x/jumlah total kasus)*entropy variable x)

1. Outlook = $0,863121 - ((4/14)*0) + ((5/14)*0,721928) + ((5/14)*0,970951) = 0,25852104$
2. Temp = $0,863121 - ((4/14)*0) + ((4/14)*1) + ((6/14)*0,918296) = 0,18385093$
3. Humidity = $0,863121 - ((7/14)*0,985228) + ((7/14)*0) = 0,3705065$
4. Windy = $0,863121 - ((8/14)*0,811278) + ((6/14)*0,918296) = 0,00597771$

		Jumlah kasus	Don't play (s1)	Play (s2)	Entropy	Gain
TOTAL		14	4	10	0,863121	
OUTLOOK						0,25852104
	Cloudy	4	0	4	0	
	Raini	5	1	4	0,721928	
	Sunny	5	3	2	0,970951	
TEMP						0,18385093
	Cool	4	0	4	0	
	Hot	4	2	2	1	
	Mild	6	2	4	0,918296	
HUMIDITY						0,3705065
	Hight	7	4	3	0,985228	
	Normal	7	0	7	0	
WINDY						0,00597771
	No	8	2	6	0,811278	
	yes	6	2	4	0,918296	

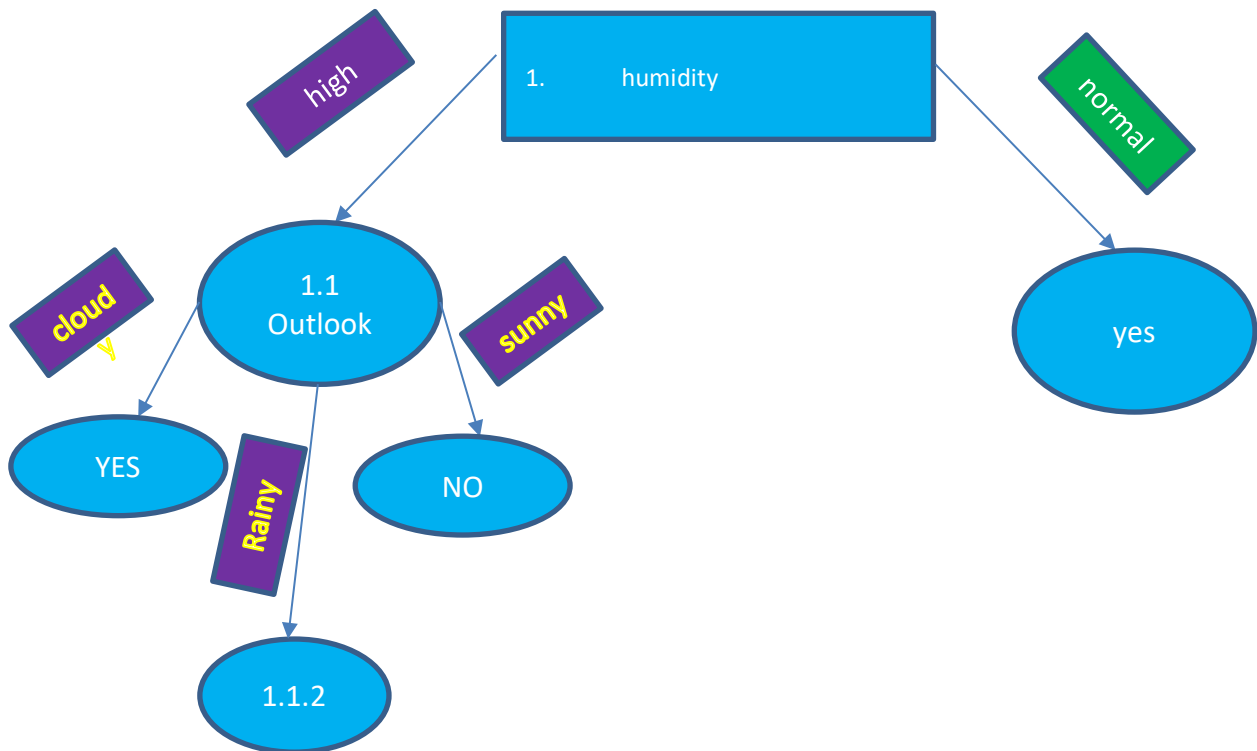
DECISION TREE



TABEL PERHITUNGAN 1.1

		Jumlah kasus	Don't play (s1)	Play (s2)	Entropy	Gain
HUMIDITY		7	4	3	0,985228136	
OUTLOOK						0,69951385
	Cloudy	2	0	2	0	
	Raini	2	1	1	1	
	Sunny	3	3	0	0	
TEMP						0,020244207
	Cool	0	0	0	0	
	Hot	3	2	1	0,918295834	
	Mild	4	2	2	1	
WINDY						0,020244207
	No	4	2	2	1	
	yes	3	2	1	0,918295834	

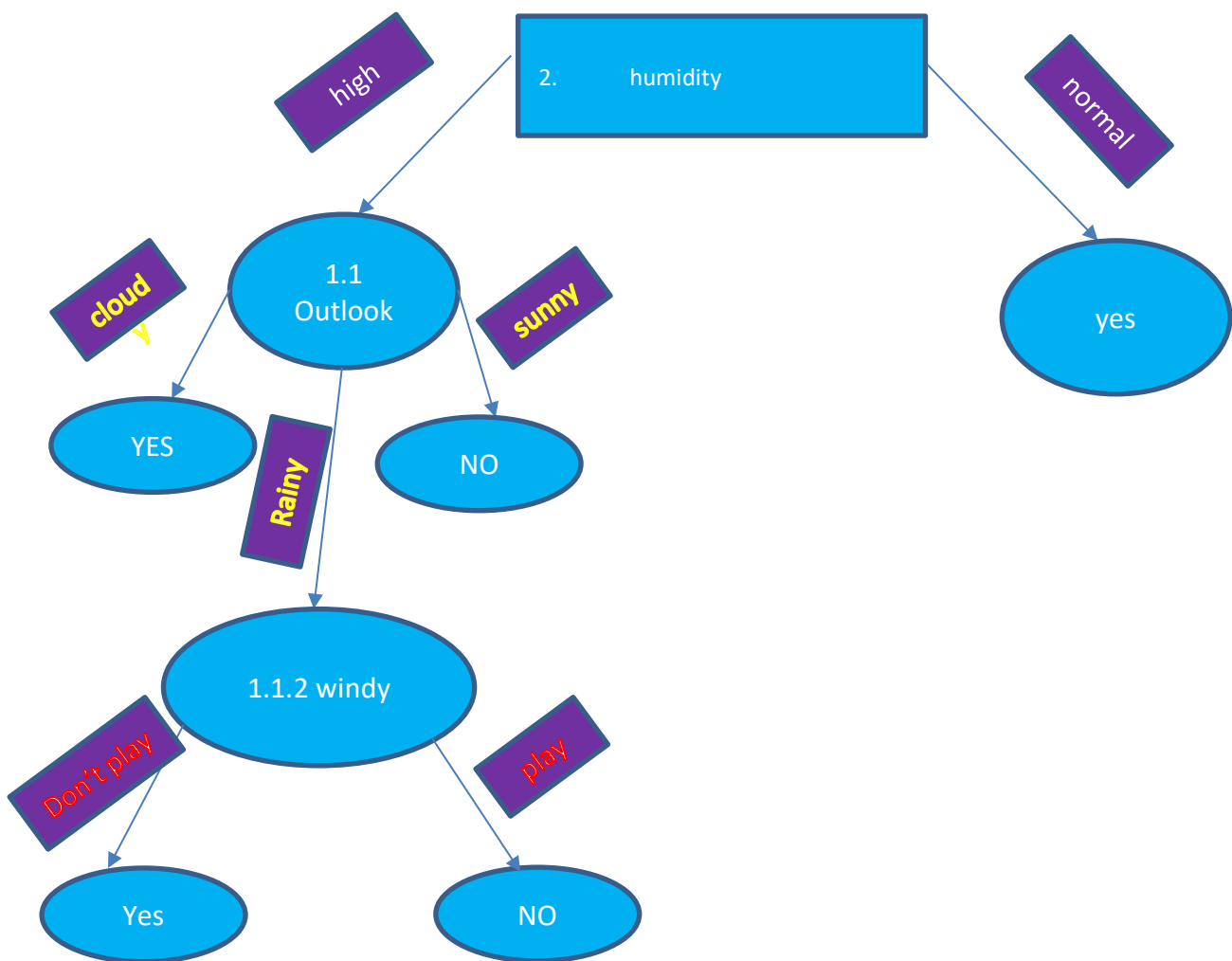
DECISION TREE



TABEL PERHITUNGAN 1.1.2

NODE 1.1.2		Jumlah kasus	Don't play (s1)	Play (s2)	Entropy	Gain
HUMIDITY HIGHT DAN OUTLOOK RAINY		2	1	1	1	
TEMP						0
	Cool	0	0	0	0	
	Hot	0	0	0	0	
	Mild	2	1	1	1	
WINDY						1
	No	1	0	1	0	
	yes	1	1	0	0	

DECISION TREE



TUGAS 05 ADVANCED DATABASE

Dari dataset berikut ini, silahkan buat perhitungan manual dari algoritma decision tree C45 secara lengkap dan gambarkan pohon keputusan yang diperoleh.

No	Outlook	Temperature	Humidity	Windy	Play
1	Sunny	Hot	High	FALSE	No
2	Sunny	Hot	High	TRUE	No
3	Cloudy	Hot	High	FALSE	Yes
4	Rainy	Mild	High	FALSE	Yes
5	Rainy	Cool	Normal	FALSE	Yes
6	Rainy	Cool	Normal	TRUE	Yes
7	Cloudy	Cool	Normal	TRUE	Yes
8	Sunny	Mild	High	FALSE	No
9	Sunny	Cool	Normal	FALSE	Yes
10	Rainy	Mild	Normal	FALSE	Yes
11	Sunny	Mild	Normal	TRUE	Yes
12	Cloudy	Mild	High	TRUE	Yes
13	Cloudy	Hot	Normal	FALSE	Yes
14	Rainy	Mild	High	TRUE	No

Kumpulkan dalam format MS Word. Sertakan penjelasan dan pendapat anda tentang algoritma ini, kelebihan dan kekurangannya berbanding metode lain yang sejenis (tentukan sendiri metode pembandingnya).

JAWAB :

Decision Tree adalah sebuah struktur pohon, dimana setiap node pohon merepresentasikan atribut yang telah diuji, setiap cabang merupakan suatu pembagian hasil uji, dan node daun merepresentasikan kelompok kelas tertentu. Level node teratas dari sebuah Decision Tree adalah node akar (root) yang biasanya berupa atribut yang paling memiliki pengaruh terbesar pada suatu kelas tertentu. Pada umumnya Decision Tree melakukan strategi pencarian secara top-down untuk solusinya. Pada proses mengklasifikasi data yang tidak diketahui, nilai atribut akan diuji dengan cara melacak jalur dari node akar (root) sampai node akhir (daun) dan kemudian akan diprediksi kelas yang dimiliki oleh suatu data baru tertentu.

Information, Entropy dan Information Gain

Entropy

Secara umum entropy adalah ukuran dari sebuah ketidakpastian dengan variabel yang random (acak). Dalam hal stored instances, maka entropy adalah ukuran ketidakpastian dari atribut.

$$H(X) = - \sum_{i=1}^n p(x_i) \log_b p(x_i)$$

dimana $p(x_i)$ adalah propabilitas mass function dari outcome x_i

Information Gain

Information gain adalah pengurangan yang diharapkan dalam entropy. Dalam machine learning, ini dapat digunakan untuk menentukan urutan atribut atau mempersempit atribut yang dipilih. Urutan inilah yang akan membentuk pohon keputusan (decision tree). Biasanya atribut dengan gain informasi terbesar yang dipilih.

$$Gain(S, A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

CONTOH

Tabel dataset diatas menunjukkan 14 instance yang tersimpan dalam database dengan 6 atribut (no, outlook, temperature, humidity, windy, play). Setiap instance menggambarkan kondisi dan aksi dari orang yang akan bermain (played atau not played).

Untuk root element, atribut dengan entropy terbesar yang dipilih. Untuk contoh di atas ini maka yang dipilih adalah **outlook** karena mempunyai parameter ketidakpastian terbesar. Setelah memilih root, maka selanjutnya dihitung setiap pasangan kombinasi yang mungkin terjadi.

$$Entropy(s) = ([10+, 4-]) = -\frac{10}{14} \log_2 \frac{10}{14} - \frac{4}{14} \log_2 \frac{4}{14}$$

Entropy(S) adalah entropy of classifier dari Play - ada 14 instance dalam dataset yang mana ada 10x (10+) keputusan untuk play dan 4x not to play (4-).

Contoh kalkulasi untuk classifier Windy :

Ada 8 instance dengan value : Weak. Jika Windy bernilai Weak maka ada 6x keputusan untuk Play dan 3x keputusan untuk not Play. [6+, 2-]. Jika strong windy (6 instances), keputusan to play 3x dan not to play 3x juga. Maka information gain attribute wind adalah 0.48.

Information gain dari semua attribut adalah :

Gain(S, Outlook) = 0.246

Gain(S, Humidity) = 0.151

Gain(S, Wind) = 0.048

Gain(S, Temperature) = 0.029

Dataset diatas :

Humidity = High: No (6.0/2.0)

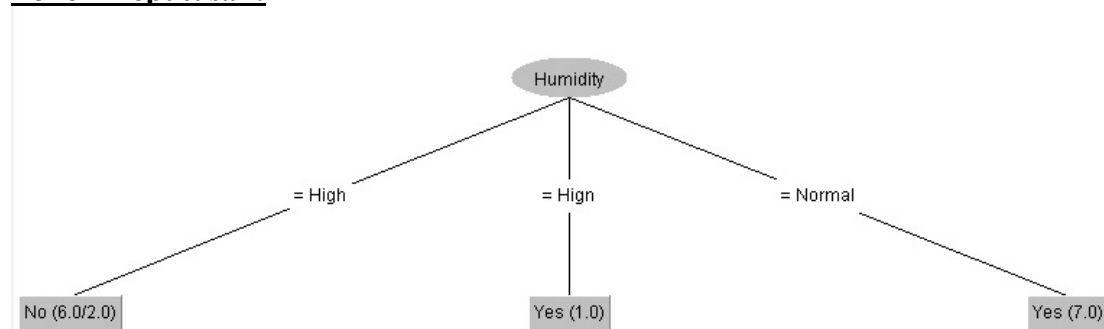
Humidity = High: Yes (1.0)

Humidity = Normal: Yes (7.0)

Number of Leaves : 3

Size of the tree : 4

Pohon Keputusan:



Naive bayesian adalah suatu klasifikasi berpeluang sederhana berdasarkan aplikasi teorema Bayes dengan asumsi antar variabel penjelas saling bebas (independen). Dalam hal ini, diasumsikan bahwa kehadiran atau ketiadaan dari suatu kejadian tertentu dari suatu kelompok tidak berhubungan dengan kehadiran atau ketiadaan dari kejadian lainnya.

Kesimpulan :

Setelah melakukan percobaan dengan dua metode yaitu naive bayes dan decision tree pada dataset diatas, dapat disimpulkan bahwa lebih baik menggunakan **metode decision tree** pada dataset diatas, karena tingkat keberhasilannya 100%, sedangkan jika menggunakan **naive bayes** tingkat keberhasilannya hanya 30%.

Kelebihan

1. Sebagai alat edukasi yang bagus untuk para pekerja. Dapat digunakan untuk mengetahui bagaimana tindakan dan keputusan mereka mempengaruhi sebuah situasi.
2. Sangat berguna sebagai *human factor analysis* pada situasi tertentu.
3. Cara yang baik bagi [manajer](#) untuk melihat seberapa besar simpati seorang pekerja atas terjadinya insiden.

Kekurangan

1. Bisa menjadi kurang jelas apabila setiap *point* pada *decision tree* tidak dilakukan [pengambilan keputusan](#) yang teliti.
2. Sulit untuk memperhitungkan elemen lain.

S E L E S A I

Nama : Hairun Anisyah

NIM : 202420039

Kelas : MTI 23

Tugas 04

Dari dataset berikut ini, silahkan buat perhitungan manual dari algoritma decision tree C45 secara lengkap dan gambarkan pohon keputusan yang diperoleh.

No	Outlook	Temperature	Humidity	Windy	Play
1	Sunny	Hot	High	FALSE	No
2	Sunny	Hot	High	TRUE	No
3	Cloudy	Hot	High	FALSE	Yes
4	Rainy	Mild	High	FALSE	Yes
5	Rainy	Cool	Normal	FALSE	Yes
6	Rainy	Cool	Normal	TRUE	Yes
7	Cloudy	Cool	Normal	TRUE	Yes
8	Sunny	Mild	High	FALSE	No
9	Sunny	Cool	Normal	FALSE	Yes
10	Rainy	Mild	Normal	FALSE	Yes
11	Sunny	Mild	Normal	TRUE	Yes
12	Cloudy	Mild	High	TRUE	Yes
13	Cloudy	Hot	Normal	FALSE	Yes
14	Rainy	Mild	High	TRUE	No

Kumpulkan dalam format MS Word. Sertakan penjelasan dan pendapat anda tentang algoritma ini, kelebihan dan kekurangannya berbanding metode lain yang sejenis (tentukan sendiri metode pembandingnya).

Jawab:

Algoritma C4.5

Algoritma C4.5 yaitu sebuah algoritma yang digunakan untuk membangun decision tree (pengambilan keputusan). Secara Definisi Algoritma C.45 adalah salah satu algoritma induksi pohon keputusan yaitu ID3 (Iterative Dichotomiser 3). ID3 dikembangkan oleh J. Ross Quinlan.

Dalam prosedur algoritma ID3, input berupa sampel training, label training dan atribut. Algoritma C4.5 merupakan pengembangan dari ID3. Beberapa pengembangan yang dilakukan pada C4.5 adalah sebagai antara lain bisa mengatasi missing value, bisa mengatasi kontinu data, dan pruning.

Untuk memilih atribut akar, didasarkan pada nilai gain tertinggi dari atribut-atribut yang ada. Untuk menghitung gain digunakan rumus seperti yang tertera dalam persamaan berikut:

$$Gain(S, A) = Entropy(S) - \sum_{i=1}^n \frac{|S_i|}{|S|} * Entropy(S_i)$$

Rumus 1. Menghitung Gain

Di mana :

S : Himpunan kasus

A : Atribut

N : Jumlah partisi atribut A

|S_i| : Jumlah kasus pada partisi ke-i

|S| : Jumlah kasus dalam S

Sementara itu, perhitungan nilai entropi dapat dilihat pada persamaan 2 berikut:

$$Entropy(S) = \sum_{i=1}^n -p_i * \log_2 p_i$$

Rumus 2. Menghitung Entropy

Di mana :

S : Himpunan kasus

A : Fitur

N : Jumlah partisi S

p_i : Proporsi dari S_i terhadap S

Secara umum algoritma C4.5 untuk membangun pohon keputusan adalah sebagai berikut :

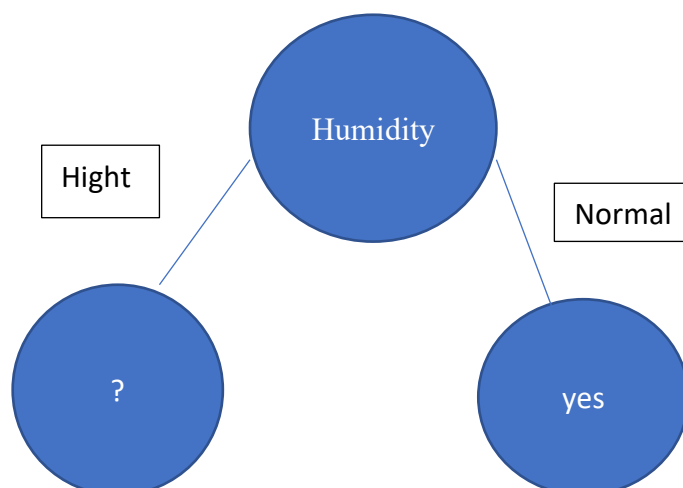
- (a) Pilih atribut sebagai akar.
- (b) Buat cabang untuk tiap-tiap nilai.
- (c) Bagi kasus dalam cabang.

(d) Ulangi proses untuk setiap cabang sampai semua kasus pada cabang memiliki kelas yang sama.

Table 1 Perhitungn Nude 1

NO	Deskripsi	Jumlah Kasus (s)	Jumlah Play "Yes"	Jumlah Play "No"	Entropy	Gain
	Total	14	10	4	0.8631	
1	OUTLOOK					0.2585
	Cloudy	4	4			
	Rainy	5	4	1	0.7219	
	Sunny	5	2	3	0.09709	
2	Temperature					0.1838
	Cool	4	4			
	Hot	4	2	2	1	
	Mild	6	4	2	0.9182	
3	Humidity					0.370
	High	7	3	4	0.9852	
	Normal	7	7			
4	Windy					0.0059
	False	8	6	2	0.8112	
	True	6	4	2	0.9182	

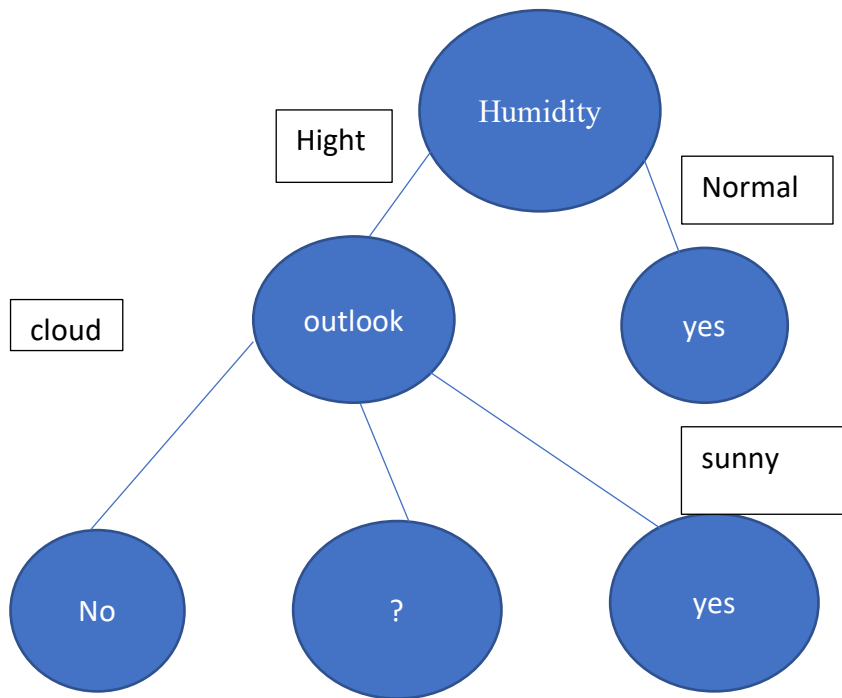
Pada tabel 1, terlihat bahwa gain tertinggi berada pada atribut Humidity yaitu yang bernilai 0.370 dengan nilai entropi tertinggi. Berada di nilai atribut High dan Normal Dari hasil perhitungan di atas, pohon keputusan yang terbentuk :



Untuk menghitung jumlah untuk keputusan Yes, jumlah untuk keputusan No , dan Entropy dari semua kasus yang dibagi berdasarkan atribut OUTLOOK, TEMPERATURE, dan WINDY yang dapat menjadi node akar dari nilai atribut HIGH . Kemudian lakukan perhitungan Gain untuk tiap atribut. Dengan menghasilkan perhitungan sebagai berikut :

NO	Deskripsi	Jumlah Kasus (s)	Jumlah Play "Yes"	Jumlah Play "No"	Entropy	Gain
	Total	7	3	4	0.9851	
1	OUTLOOK					0.6995
	Cloudy	4	4			
	Rainy	5	4	1	1	
	Sunny	5	2	3	0	
2	Temperature					0.1838
	Cool	0	0	0		
	Hot	3	1	2	0,9182	
	Mild	6	4	2	1	
3	Humidity					0
	High	7	3	4	0.9852	
	Normal	7	7			
4	Windy					0.0059
	False	8	6	2	0.8112	
	True	6	4	2	0.9182	

Dilihat dari table, diketahui bahwa atribut dengan Gain tertinggi adalah OUTLOOK , yaitu sebesar 0.69. Jadi OUTLOOK dapat menjadi node cabang dari nilai atribut High . Ada tiga nilai atribut dari OUTLOOK, yaitu CLOUDY, RAINY dan SUNNY. Dari ketiga nilai atribut tersebut, nilai atribut CLOUDY adalah 1, yaitu keputusannya Yes dan nilai atribut SUNNY menjadi keputusan No, sehingga tidak perlu dilakukan perhitungan lebih lanjut, tetapi nilai atribut RAINY masih perlu dilakukan perhitungan lagi.

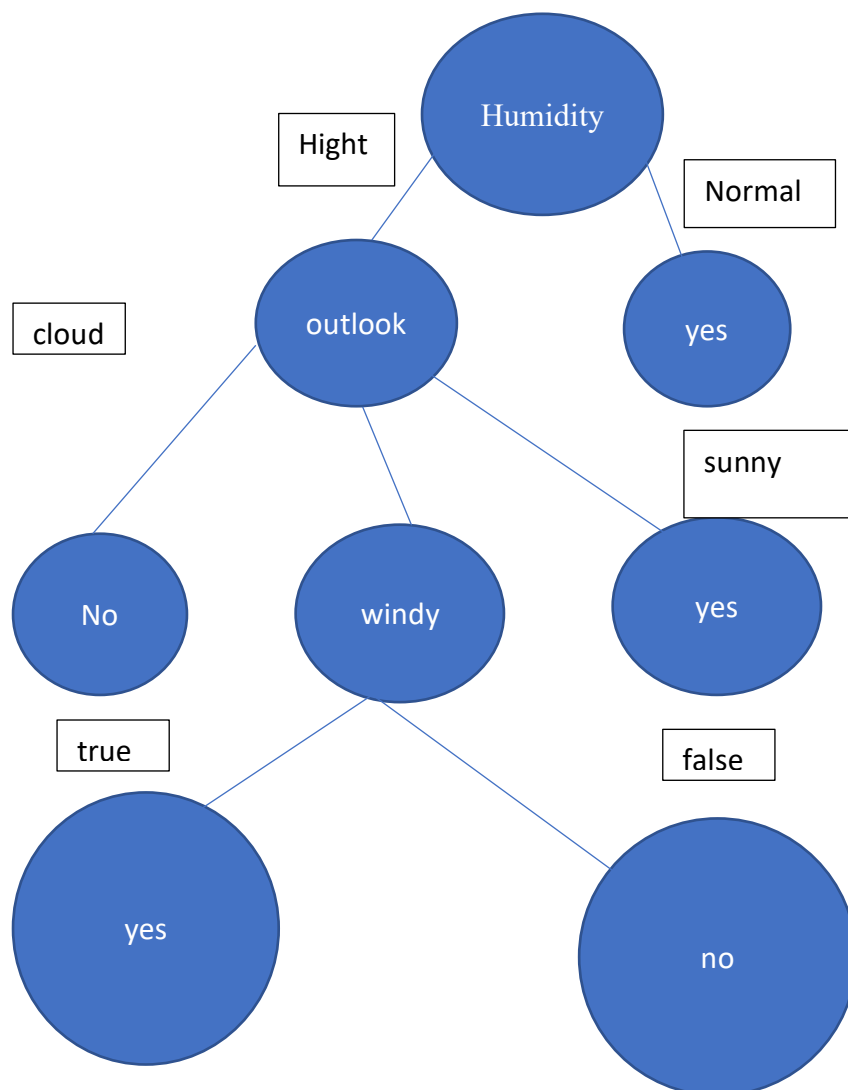


Dari menghitung jumlah untuk keputusan Yes, jumlah keputusan No, dan Entropy dari semua kasus yang dibagi berdasarkan atribut TEMPERATURE dan WINDY yang dapat menjadi node cabang dari nilai atribut RAINY. Kemudian lakukan perhitungan Gain untuk tiap-tiap atribut. Hasil perhitungan seperti yang terdapat pada table dibawah ini :

NO	Deskripsi	Jumlah Kasus (s)	Jumlah Play "Yes"	Jumlah Play "No"	Entropy	Gain
	Total	2	1	1	1	
1	OUTLOOK					0
	Cloudy	2	2		0	
	Rainy	2	4	1	1	
	Sunny	3		3	0	
2	Temperature					0
	Cool	0	0			
	Hot	0	0	0	1	
	Mild	2	1	1	0.9182	
3	Humidity					-2.442
	High	7	3	4	0.9852	

	Normal	7	7			
4	Windy					1
	False	8	6	2	0	
	True	6	4	2	0	

Dapat diketahui bahwa atribut dengan Gain tertinggi adalah Windy yaitu sebesar 1. Jadi Windy dapat menjadi node cabang dari nilai atribut RAINY. Ada dua nilai atribut dari Windy, yaitu FALSE dan TRUE. Dari kedua nilai atribut tersebut, nilai atribut FALSE adalah 1, yaitu keputusannya Yes dan nilai atribut TRUE menjadi keputusan No, sehingga tidak perlu dilakukan perhitungan lebih lanjut untuk nilai atribut ini. Dapat digambarkan untuk pohon keputusannya sebagai berikut :



Dari dataset berikut ini, silahkan buat perhitungan manual dari algoritma decision tree C45 secara lengkap dan gambarkan pohon keputusan yang diperoleh.

No	Outlook	Temperature	Humidity	Windy	Play
1	Sunny	Hot	High	FALSE	No
2	Sunny	Hot	High	TRUE	No
3	Cloudy	Hot	High	FALSE	Yes
4	Rainy	Mild	High	FALSE	Yes
5	Rainy	Cool	Normal	FALSE	Yes
6	Rainy	Cool	Normal	TRUE	Yes
7	Cloudy	Cool	Normal	TRUE	Yes
8	Sunny	Mild	High	FALSE	No
9	Sunny	Cool	Normal	FALSE	Yes
10	Rainy	Mild	Normal	FALSE	Yes
11	Sunny	Mild	Normal	TRUE	Yes
12	Cloudy	Mild	High	TRUE	Yes
13	Cloudy	Hot	Normal	FALSE	Yes
14	Rainy	Mild	High	TRUE	No

Jawaban

Pohon keputusan untuk atribut simpul sebagai akar (root) berdasarkan nilai gain tertinggi
Dengan menggunakan rumus

$$Gain(S, A) = Entropy(S) - \sum_{i=1}^n \frac{[Si]}{[S]} * Entropi(Si)$$

Keterangan :

S : Himpunan kasus

A : Atribut

N : jumlah partisi atribut A

[Si] : jumlah kasus pada partisi ke-i

[S] : jumlah kasus pada S

Untuk menghitung Entropi menggunakan rumus

$$Entropy(S) = \sum_{i=1}^n - p_i * \log_2 p_i$$

Keterangan :

S : Himpunan kasus

N : jumlah partisi S

Pi : proporsi dari Si terhadap S

Perhitungan Gain untuk setiap atribut

No			Jlh kasus (S)	NO (S1)	YES (S2)	Entropy	Gain
1	TOTAL		14	4	10	0.8631205	
	OUTLOOK						0.2585210
		Cloudy	4	0	4	0	
		Rainy	5	1	4	0.721982	
		Sunny	5	3	2	0.970950	
	TEMPERATURE						0.1838509
		Cool	4	0	4	0	
		Hot	4	2	2	1	
		Mild	6	2	4	0.9182958	
	HUMIDITY						0.3705065
		High	7	4	3	0.9852281	
		Normal	7	0	7	0	
	WINDY						0.0059777
		False	8	2	6	0.8112781	
		True	6	4	2	0.9182958	

$$Entropy (Total) = \left(-\frac{4}{14} * \log_2 \left(\frac{4}{14} \right) \right) + \left(-\frac{10}{14} * \log_2 \left(\frac{10}{14} \right) \right)$$

$$= 0.8631205$$

Nilai Gain pada baris OUTLOOK dihitung dengan persamaan

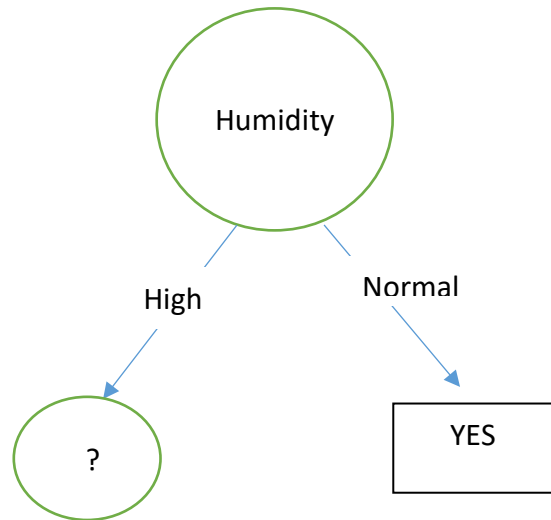
$$Gain (total, OUTLOOK) = entropy (total) - \sum_{i=1}^n \frac{|outlook|}{|total|} * entropy (outlook)$$

$$Gain (total, OUTLOOK)$$

$$= 0.8631205 - \left(\left(\frac{4}{14} * 0 \right) + \left(\frac{5}{14} * 0.7219928 \right) + \left(\frac{5}{14} * 0.970950 \right) \right)$$

$$= 0.2585210$$

Dari tabel diatas dapat diketahui bahwa atribut dengan **Gain tertinggi** adalah **HUMIDITY** dan Ada dua nilai atribut dari HUMIDITY, yaitu HIGH dan NORMAL. Dari kedua atribut, nilai atribut NORMAL adalah 1, yaitu keputusan Yes, sehingga tidak perlu perhitungan lebih lanjut, tetapi untuk HIGH masih perlu dilakukan perhitungan lagi. Dari hasil tersebut dapat digambarkan pohon keputusan sementara

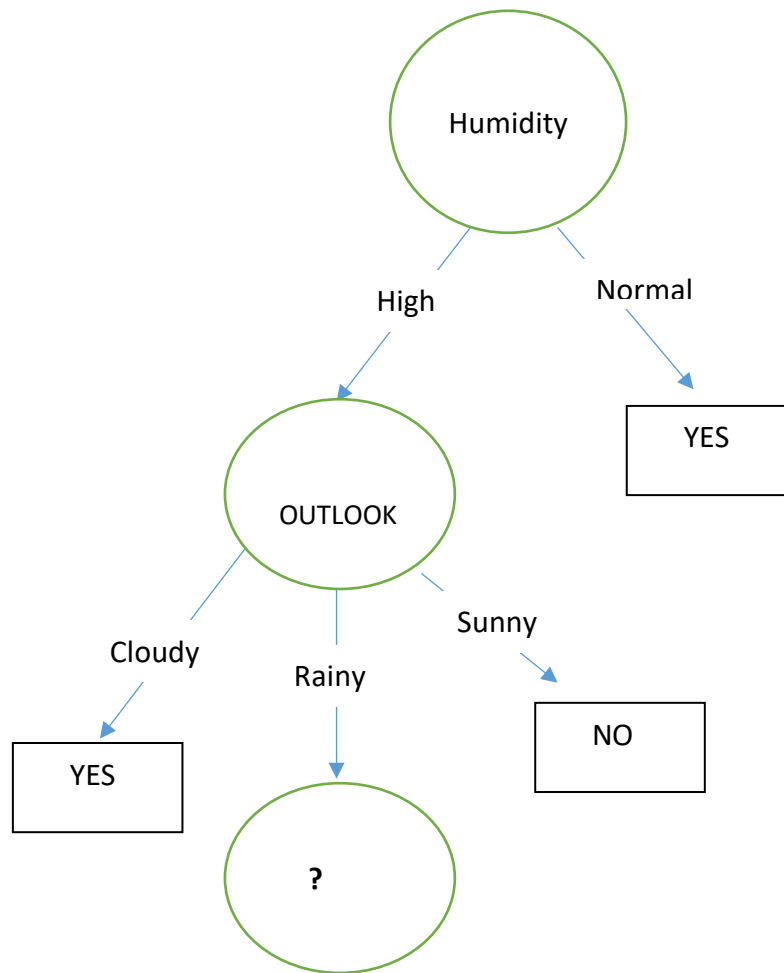


Menghitung jumlah kasus, jumlah kasus untuk keputusan Yes, jumlah kasus untuk keputusan No, dan Entropy dari semua kasus dan kasus yang dibagi berdasarkan atribut OUTLOOK, TEMPERATURE, dan WINDY yang dapat menjadi node akardari nilai atribut HIGH. Kemudian lakukan perhitungan Gain KE 2 untuk tiap-tiap atribut tanpa atribut HUMIDITY.

Perhitungan Gian tanpa atribut Humidity

No			Jlh kasus (S)	NO (S1)	YES (S2)	Entropy	Gain
2	TOTAL		7	4	3	0.9852281	
	OUTLOOK						0.69951385
		Cloudy	2	0	2	0	
		Rainy	2	1	1	1	
		Sunny	3	3	0	0	
	TEMPERATURE						0.02024420
		Cool	0	0	0	0	
		Hot	4	2	1	0.9182958	
		Mild	3	2	2	1	
	WINDY						0.02024420
		False	4	2	2	1	
		True	3	2	1	0.9182958	

atribut dengan Gain tertinggi adalah OUTLOOK, yaitu sebesar 0.699 . Jadi OUTLOOK dapat menjadi cabang dari nilai atribut HIGH. Ada tiga nilai atribut dari OUTLOOK, yaitu CLOUDY, RAINY dan SUNNY. Dari ketiga nilai atribut tersebut, nilai atribut CLOUDY adalah 1, yaitu keputusannya Yes dan nilai atribut SUNNY menjadi keputusan No, sehingga tidak perlu dilakukan perhitungan lebih lanjut, tetapi nilai atribut RAINY masih perlu dilakukan perhitungan lagi. Pohon keputusan yang terbentuk sampai tahap ini adalah



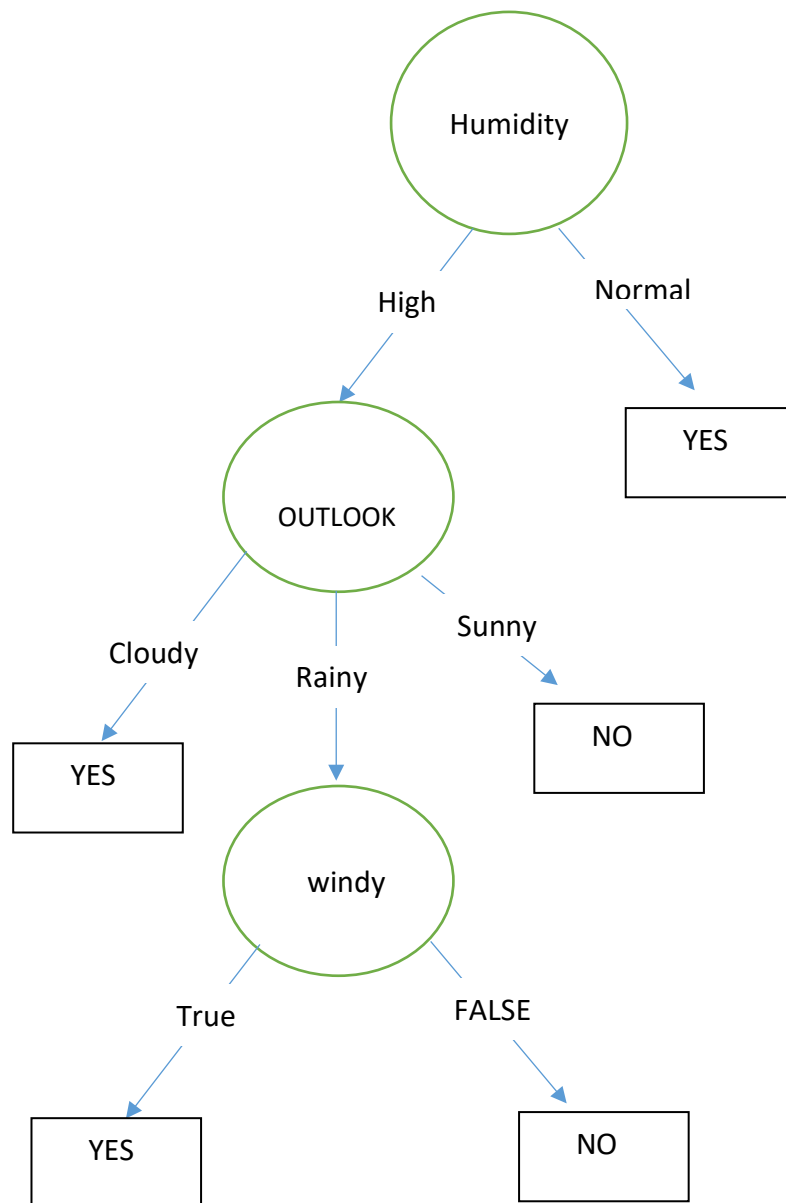
jumlah kasus untuk keputusan Yes, jumlah kasus untuk keputusan No, dan Entropy dari semua kasus dan kasus yang dibagi berdasarkan atribut tinggal menyisakan atribut TEMPERATURE dan WINDY yang dapat menjadi cabang dari nilai atribut RAINY. Kemudian lakukan perhitungan Gain untuk tiap-tiap atribut.

Hasil perhitungan atribut TEMPERATURE dan WINDY

No			Jlh kasus (S)	NO (S1)	YES (S2)	Entropy	Gain
3	TOTAL		2	1	1	1	
	TEMPERATURE						0
		Cool	0	0	0	0	
		Hot	0	0	0	0	
		Mild	2	1	1	1	
	WINDY						1
		False	1	0	1	0	
		True	1	1	0	0	

Dari hasil tabel dapat diketahui bahwa atribut dengan Gaint tertinggi adalah WINDY, yaitu sebesar 1. Jadi WINDY dapat menjadi cabang dari nilai atribut RAINY. Ada dua nilai atribut dari

WINDY, yaitu FALSE dan TRUE. Dari kedua nilai atribut tersebut, nilai atribut FALSE adalah 1, yaitu keputusannya Yes dan nilai atribut TRUE menjadi keputusan No, sehingga tidak perlu dilakukan perhitungan lebih lanjut untuk nilai atribut ini. Pohon keputusan yang terbentuk sampai tahap ini adalah



Berdasarkan gambar Pohon Keputusan di atas dapat ditarik kesimpulan bahwa jika humidity normal maka keputusannya adalah yes, jika humidity adalah high, outlook adalah cloudy maka keputusannya adalah bermain (yes), selanjutnya jika humidity adalah high, outlook adalah rainy dan windy adalah true, maka keputusannya adalah bermain (yes). Dan jika humidity adalah high, outlook adalah rainy dan windy adalah false, maka keputusannya adalah tidak bermain (no). Terakhir jika humidity adalah high, outlook adalah no, maka keputusannya adalah tidak bermain (no).

Tugas 05

Dari dataset berikut ini, silahkan buat perhitungan manual dari algoritma decision tree C45 secara lengkap dan gambarkan pohon keputusan yang diperoleh.

No	Outlook	Temperature	Humidity	Windy	Play
1	Sunny	Hot	High	FALSE	No
2	Sunny	Hot	High	TRUE	No
3	Cloudy	Hot	High	FALSE	Yes
4	Rainy	Mild	High	FALSE	Yes
5	Rainy	Cool	Normal	FALSE	Yes
6	Rainy	Cool	Normal	TRUE	Yes
7	Cloudy	Cool	Normal	TRUE	Yes
8	Sunny	Mild	High	FALSE	No
9	Sunny	Cool	Normal	FALSE	Yes
10	Rainy	Mild	Normal	FALSE	Yes
11	Sunny	Mild	Normal	TRUE	Yes
12	Cloudy	Mild	High	TRUE	Yes
13	Cloudy	Hot	Normal	FALSE	Yes
14	Rainy	Mild	High	TRUE	No

Kumpulkan dalam format MS Word. Sertakan penjelasan dan pendapat anda tentang algoritma ini, kelebihan dan kekurangannya berbanding metode lain yang sejenis (tentukan sendiri metode pembandingnya).

Perhitungan Gain

$$\text{Gain}(S,A) = \text{Entropy}(S) - \sum_{i=1}^n \frac{|S_i|}{|S|} * \text{Entropy}(S_i)$$

Keterangan:

- S : himpunan
- A : atribut
- n : jumlah partisi atribut A
- | S_i | : jumlah kasus pada partisi ke-i
- | S | : jumlah kasus dalam S

Menghitung Nilai Entropy

$$\text{Entropy}(S) = - \sum_{i=1}^n p_i * \log_2 p_i$$

Keterangan:

S : himpunan kasus

A : fitur

n : jumlah partisi S

p_i : proporsi dari S_i terhadap S

Perincian algoritma (langkah 1)

- Menghitung jumlah kasus seluruhnya, jumlah berkeputusan “Yes” maupun “No”.
- Menghitung Entropy dari semua kasus yang terbagi berdasarkan atribut “Outlook”, “Temperature”, “Humidity”, “Windy”.
- Lakukan penghitungan Gain untuk setiap atributnya

Perhitungan

Node			Jumlah Kasus (S)	Tidak (S ₁)	Ya (S ₂)	Entropy	Gain
1	TOTAL		14	4	10	0,863120569	
	OUTLOOK						0,258521037
		CLOUDY	4	0	4	0	
		RAINY	5	1	4	0,721928095	
		SUNNY	5	3	2	0,970950594	
	TEMPERATURE						0,183850925
		COOL	4	0	4	0	
		HOT	4	2	2	1	
		MILD	6	2	4	0,918295834	
	HUMIDITY						0,370506501
		HIGH	7	4	3	0,985228136	
		NORMAL	7	0	7	0	
	WINDY						0,005977711
		FALSE	8	2	6	0,811278124	
		TRUE	6	2	4	0,918295834	

Perhitungan Total Entropy

Node			Jumlah Kasus (S)	Tidak (S ₁)	Ya (S ₂)	Entropy	Gain
1	TOTAL		14	4	10	0,863120569	
...	...	...	...	...	...	...	...

$$\text{Entropy (S)} = \sum_{i=1}^n - p_i * \log_2 p_i$$

$$\begin{aligned} \text{Entropy (Total)} = & - 4/14 * \log_2 (4/14) + \\ & - 10/14 * \log_2 (10/14) \end{aligned}$$

Menghitung gain pada baris Outlook

Gain (Total, Outlook)

$$\begin{aligned} & n \\ = & \text{Entropy(Total)} - \sum_{i=1}^n \\ = & 0,863120569 - ((4/14*0) + (5/14*0,72) + (5/14*0,97)) \\ = & 0,258521037 \end{aligned}$$

Hitung Gain untuk temperature, humidity dan windy

Sepert yg terlihat pada tabel, diperoleh bahwa atribut degan Gain tertinggi adalah Humidity - > 0,37

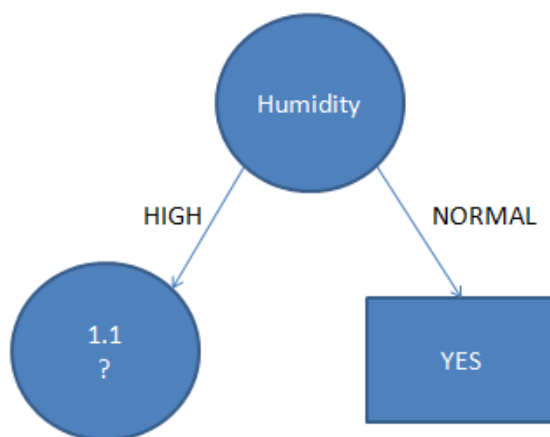
Maka Humidity menjadi node akar

Humidity memiliki dua nilai yaitu “High” dan “Normal”

Humidity -> “Normal” sudah mengklasifikasikan kasus menjadi 1 yaitu keputusannya “yes”

Untuk humidity -> “High” masih perlu dilakukan perhitunngn lagi (karena masih terdapat “yes” dan “no”)

Pohon Keputusan Node 1



Perincian Algoritma (Langkah 2)

Pilih node akar "High" dan hitung

- Jumlah kasus
- Jumlah kasus keputusan "yes"
- Jumlah kasus keputusan "no"
- Entropy

Atribut:

- Outlook
- Temperature
- windy

Hasil perhitungan (Langkah 2)

Node			Jumlah Kasus (S)	Tidak (S ₁)	Ya (S ₂)	Entropy	Gain
1.1	HUMIDITY HIGH		7	4	3	0,985228136	
	OUTLOOK						0,69951385
		CLOUDY	2	0	2	0	
		RAINY	2	1	1	1	
		SUNNY	3	3	0	0	
	TEMPERATURE						0,020244207
		COOL	0	0	0	0	
		HOT	3	2	1	1	
		MILD	4	2	2	2	
	WINDY						0,020244207
		FALSE	4	2	2	1	
		TRUE	3	2	1	0,918295834	

Hasil perhitungan (Langkah 2)

Didapat Gain tertinggi -> outlook -> 0,69

Maka "Outlook" menjadi node cabang dari atribut humidity yg bernilai "High"

Berdasarkan atribut "Outlook" terdpt 3 nilai

Cloudy

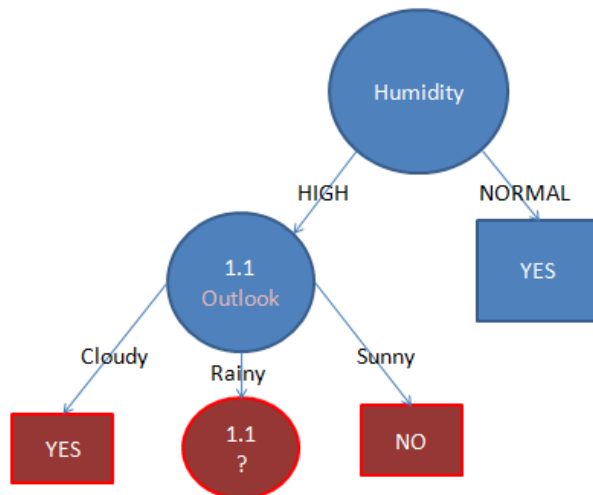
Rainy

Sunny

Krn "Cloudy" pasti bernilai "Yes" dan "Sunny" pasti bernilai "No", maka tidak perlu dilakukan perhitungan lagi

Sedangkan "Rainy" bernilai "yes" dan "No", maka masih perlu dilakukan perhitungan lagi

Pohon keputusan node 1.1



Perincian algoritma (Langkah 3)

Pilih node akar "Rainy"

- Jumlah kasus
- Jumlah kasus keputusan "yes"
- Jumlah kasus keputusan "no"
- Entropy

Atribut:

- Temperature
- windy

Hasil perhitungan (Langkah 3)

Node			Jumlah Kasus (S)	Tidak (S _n)	Ya (S _y)	Entropy	Gain
1.1	HUMIDITY "HIGH" dan OUTLOOK "RAINY"		2	1	1	1	
	TEMPERATURE						0
		COOL	0	0	0	0	
		HOT	0	0	0	0	
		MILD	2	1	1	1	
	WINDY						1
		FALSE	1	0	1	0	
		TRUE	1	1	0	0	

Hasil perhitungan (Langkah 3)

Didapat Gain tertinggi -> Windy -> 1

Maka "Windy" menjadi node cabang dari atribut humidity yang bernilai "High" dan outlook yang bernilai "Rainy"

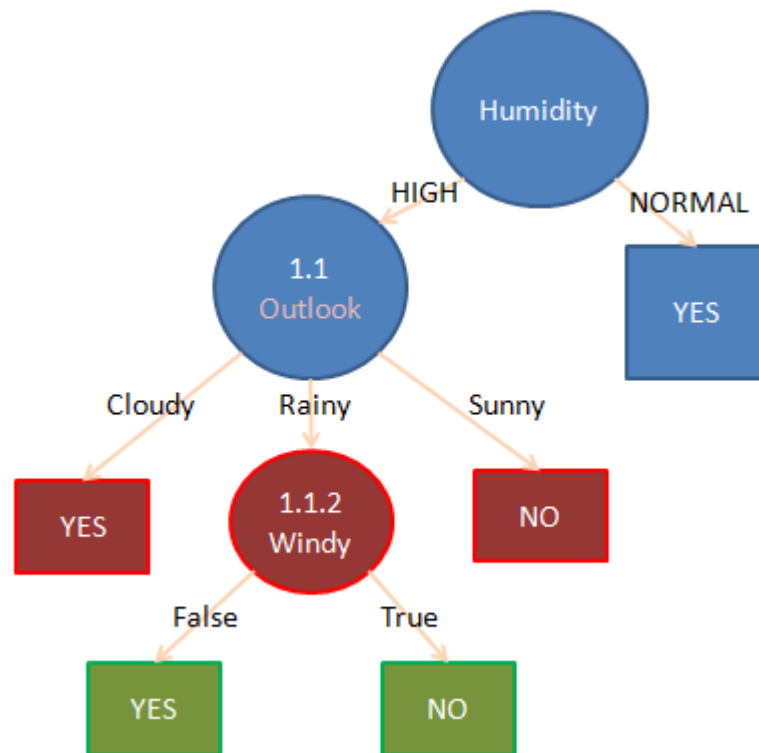
Berdasarkan atribut “Windy” terdpt 2 nilai

True

False

Karena “True” sudah terklasifikasi pasti bernilai “No” dan “False” pasti bernilai “Yes”, maka tidak perlu dilakukan perhitungan lagi

Pohon keputusan node 1.1.2



Hasil perhitungan (Langkah 3)

Berdasarkan node 1.1.2, maka:

“Semua kasus sudah masuk dapat kelas”

Sehingga pohon keputusan diatas merupakan pohon keputusan terakhir yang terbentuk

Diketahui Dataset adalah sebagai berikut :

No	Outlook	Temperature	Humidity	Windy	Play
1	Sunny	Hot	High	FALSE	No
2	Sunny	Hot	High	TRUE	No
3	Cloudy	Hot	High	FALSE	Yes
4	Rainy	Mild	High	FALSE	Yes
5	Rainy	Cool	Normal	FALSE	Yes
6	Rainy	Cool	Normal	TRUE	Yes
7	Cloudy	Cool	Normal	TRUE	Yes
8	Sunny	Mild	High	FALSE	No
9	Sunny	Cool	Normal	FALSE	Yes
10	Rainy	Mild	Normal	FALSE	Yes
11	Sunny	Mild	Normal	TRUE	Yes
12	Cloudy	Mild	High	TRUE	Yes
13	Cloudy	Hot	Normal	FALSE	Yes
14	Rainy	Mild	High	TRUE	No

Attribute : Outlook, Temperature, Humidity, Windy

Kelas : Play = Yes atau No

Jumlah data ada 12, terdiri dari :

Ya = 10

Tidak = 4

1. Menghitung Entropy Total Dataset

$$Entropi (S) = \sum_{j=1}^k - p_j \log_2 p_j$$

Entropy [Total]

$$S = \left\{ -\frac{10}{14} * (\log_2 \left(\frac{10}{14} \right)) \right\} + \left\{ -\frac{4}{14} * (\log_2 \left(\frac{4}{14} \right)) \right\} = 0.863120569$$

2. Menghitung Entropy dan Gain tiap Atribut

$$\text{Gain (S,A)} = \text{Entropy (s)} - \sum_{i=1}^n \frac{|s_i|}{|S|} * \text{Entropy (s}_i\text{)}$$

- Menghitung Gain pada baris Outlook

$$= 0.863120569 - ((5/14)*0.97) + (4/14*0) + (5/14*0.72)) \\ = 0.258521037$$

- Menghitung Gain pada baris Temperature

$$= 0.863120569 - (4/14 * 1) + (4/14 * 0) + (6/14 * 0.91)$$

$$= 0.183850925$$
- Menghitung Gain pada baris Humidity

$$= 0.863120569 - ((7/14 * 0.98) + (7/14 * 0))$$

$$= 0.370506501$$
- Menghitung Gain pada baris Windy

$$= 0.863120569 - ((6/14 * 0.91) + (8/14 * 0.81))$$

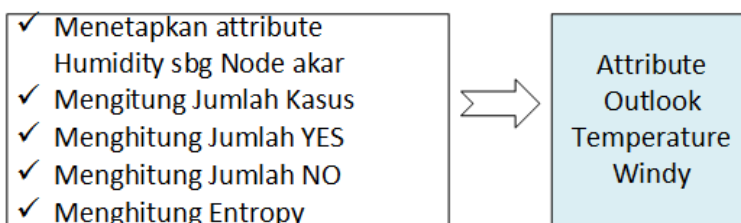
$$= 0.005977711$$

3. Perhitungan Langkah 1 (ref excel terlampir)

No	Attributte	Jumlah Kasus (s)	Play No (s2)	Play Yes (s1)	Entropy	Gain
1	Total	14	4	10	0,863120569	
	Outlook					0,258521037
	Sunny	5	3	2	0,970950594	
	Cloudy	4	0	4	0	
	Rainy	5	1	4	0,721928095	
	Tempeature					0,183850925
	Hot	4	2	2	1	
	Cool	4	0	4	0	
	Mild	6	2	4	0,918295834	
	Humidity					0,370506501
	High	7	4	3	0,985228136	
	Normal	7	0	7	0	
	Windy					0,005977711
	TRUE	6	2	4	0,918295834	
	FALSE	8	2	6	0,811278124	

- Dari hasil Perhitungan, gain tertinggi di attribute Humidity (Gain>0.37), maka Attribute Humidity menjadi root node (node akar).
- Humidity memiliki dua nilai yaitu "High" dan "Normal"
- Nilai "Normal" hasil keputusannya "YES"
- Attribute value "High" Masih memiliki YES dan NO, perlu di hitung kembali untuk keputusan di cabang bawahnya

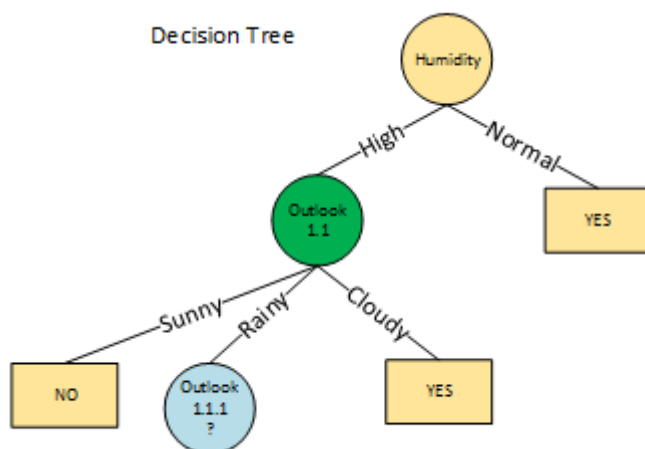
4. Perincian Algoritma Langkah 2 (no#3)



No	Attribute	Jumlah Kasus (s)	Play No (s2)	Play Yes (s1)	Entropy	Gain
1.1	Total	7	4	3	0,985228136	
Humidity - High						
	High	7	4	3	0,985228136	
	Normal	7	0	7	0	
Outlook						0,69951385
	Sunny	3	3	0	0	
	Cloudy	2	0	2	0	
	Rainy	2	1	1	1	
Temperature						0,020244207
	Hot	3	2	1	0,918295834	
	Cool	0	0	4	0	
	Mild	4	2	2	1	
Windy						0,020244207
	TRUE	3	2	1	0,918295834	
	FALSE	4	2	2	1	

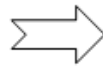
- Dari hasil perhitungan, gain tertinggi di attribute outlook
- Maka "Outlook" menjadi node cabang dari atribut humidity yg bernilai "High"
- Attribute "Outlook" memiliki 3 nilai (Sunny, Cloudy & Rainy)
- Dari perhitungan, didapat keputusan "Cloudy" bernilai "Yes" dan "Sunny" bernilai "No"
- Value "Rainy" bernilai "yes" dan "No", perlu dilakukan perhitungan lagi

5. Pohon Keputusan Node 1.1



6. Perincian Algoritma langkah 3

- ✓ Menetapkan attribute Outlook Node cabang
- ✓ Mengitung Jumlah Kasus
- ✓ Menghitung Jumlah YES
- ✓ Menghitung Jumlah NO
- ✓ Menghitung Entropy



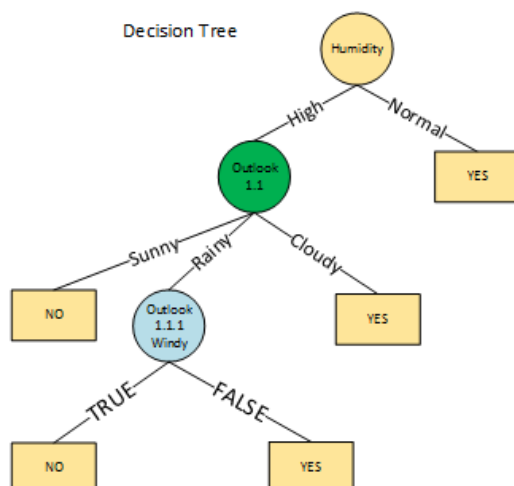
Attribute
Temperature
Windy

7. Hasil Perhitungan langkah 3

No	Attributte	Jumlah Kasus (s)	Play No (s2)	Play Yes (s1)	Entropy	Gain
1.1.1	Total	Humidity HIGH & Outlook RAINY	2	1	1	1
	Temperature					0
	Hot	0	0	0	0	0
	Cool	0	0	0	0	0
	Mild	2	1	1	1	1
	Windy					1
	TRUE	1	1	0	0	0
	FALSE	1	0	1	0	0

- Dari perhitungan didapat Gain tertinggi -> Windy -> 1
- Keputusan sudah di dapat Value attribute TRUE bernilai Value “No” dan attribute False pasti bernilai “Yes” **sehingga perhitungan di atas merupakan akhir dari cabang pohon keputusan.**

8. Pohon Keputusan Node 1.1.1



Pendapat saya tentang Algoritma ini :

Algoritma C4.5 merupakan salah satu algoritma kelompok decision tree, selain Hunt's Algorithm (one of the earliest) ,CART, ID3, & SLIQ,SPRINT

Algoritma ini mempunyai input berupa training samples dan samples. Training samples merupakan data contoh yang digunakan untuk membangun sebuah tree yang telah diuji kebenarannya. Sedangkan samples merupakan field-field data yang digunakan sebagai parameter dalam klasifikasi data.

Kelebihan Algoritma C4.5 :

- C4.5 mampu menangani atribut yang kosong (Missing Value).
- C4.5 mampu menangani atribut dengan kontinu.
- C4.5 memangkas pohon keputusan untuk mengatasi overfitting.

Kekurangan Algoritma C4.5 :

- Algoritma C4.5 hanya dapat digunakan untuk menangani sampel-sampel yang dapat disimpan secara keseluruhan dan pada waktu yang bersamaan di memori.

No	Outlook	Temperature	Humadity	Windy	Play
1	Sunny	Hot	High	FALSE	NO
2	Sunny	Hot	High	TRUE	NO
3	Cloudy	Hot	High	FALSE	YES
4	Rainy	Mild	High	FALSE	YES
5	Rainy	Cool	Normal	FALSE	YES
6	Rainy	Cool	Normal	TRUE	YES
7	Cloudy	Cool	Normal	TRUE	YES
8	Sunny	Mild	High	FALSE	NO
9	Sunny	Cool	Normal	FALSE	YES
10	Rainy	Mild	Normal	FALSE	YES
11	Sunny	Mild	Normal	TRUE	YES
12	Cloudy	Mild	High	TRUE	YES
13	Cloudy	Hot	Normal	FALSE	YES
14	Rainy	Mild	High	TRUE	NO

No	Outlook	Temperature	Humidity	Windy	Play	Sum NO
1	Sunny	Hot	High	FALSE	NO	1
2	Sunny	Hot	High	TRUE	NO	1
3	Cloudy	Hot	High	FALSE	YES	0
4	Rainy	Mild	High	FALSE	YES	0
5	Rainy	Cool	Normal	FALSE	YES	0
6	Rainy	Cool	Normal	TRUE	YES	0
7	Cloudy	Cool	Normal	TRUE	YES	0
8	Sunny	Mild	High	FALSE	NO	1
9	Sunny	Cool	Normal	FALSE	YES	0
10	Rainy	Mild	Normal	FALSE	YES	0
11	Sunny	Mild	Normal	TRUE	YES	0
12	Cloudy	Mild	High	TRUE	YES	0
13	Cloudy	Hot	Normal	FALSE	YES	0
14	Rainy	Mild	High	TRUE	NO	1
						4

No	Attributte	Jumlah Kasus (s)	Play No (s2)	Play Yes (s1)	Entropy
1	Total	14	4	10	0,863120569

Outlook					
	Sunny	5	3	2	0,970950594
	Cloudy	4	0	4	0
	Rainy	5	1	4	0,721928095
Tempeature					
	Hot	4	2	2	1
	Cool	4	0	4	0
	Mild	6	2	4	0,918295834
Humidity					
	High	7	4	3	0,985228136
	Normal	7	0	7	0
Windy					
	TRUE	6	2	4	0,918295834
	FALSE	8	2	6	0,811278124

Sum YES
0
0
1
1
1
1
1
0
1
1
1
1
0
10

Gain

0,258521037

0,183850925

0,370506501

0,005977711

No	Outlook	Temperature	Humidity	Windy	Play	Sum NO
1	Sunny	Hot	High	FALSE	NO	1
2	Sunny	Hot	High	TRUE	NO	1
3	Cloudy	Hot	High	FALSE	YES	0
4	Rainy	Mild	High	FALSE	YES	0
5	Rainy	Cool	Normal	FALSE	YES	0
6	Rainy	Cool	Normal	TRUE	YES	0
7	Cloudy	Cool	Normal	TRUE	YES	0
8	Sunny	Mild	High	FALSE	NO	1
9	Sunny	Cool	Normal	FALSE	YES	0
10	Rainy	Mild	Normal	FALSE	YES	0
11	Sunny	Mild	Normal	TRUE	YES	0
12	Cloudy	Mild	High	TRUE	YES	0
13	Cloudy	Hot	Normal	FALSE	YES	0
14	Rainy	Mild	High	TRUE	NO	1

No	Attributte	Jumlah Kasus (s)	Play No (s2)	Play Yes (s1)	Entropy
1.1	Total	7	4	3	0,985228136

Humidity - High					
	High	7	4	3	0,985228136
	Normal	7	0	7	0

Outlook					
	Sunny	3	3	0	0
	Cloudy	2	0	2	0
	Rainy	2	1	1	1

Tempeature					
	Hot	3	2	1	0,918295834
	Cool	0	0	4	0
	Mild	4	2	2	1

Windy					
	TRUE	3	2	1	0,918295834
	FALSE	4	2	2	1

Sum YES
0
0
1
1
1
1
1
1
0
1
1
1
1
1
0

Gain

0,69951385

0,020244207

0,020244207

No	Outlook	Temperature	Humidity	Windy	Play	Sum NO
1	Sunny	Hot	High	FALSE	NO	1
2	Sunny	Hot	High	TRUE	NO	1
3	Cloudy	Hot	High	FALSE	YES	0
4	Rainy	Mild	High	FALSE	YES	0
5	Rainy	Cool	Normal	FALSE	YES	0
6	Rainy	Cool	Normal	TRUE	YES	0
7	Cloudy	Cool	Normal	TRUE	YES	0
8	Sunny	Mild	High	FALSE	NO	1
9	Sunny	Cool	Normal	FALSE	YES	0
10	Rainy	Mild	Normal	FALSE	YES	0
11	Sunny	Mild	Normal	TRUE	YES	0
12	Cloudy	Mild	High	TRUE	YES	0
13	Cloudy	Hot	Normal	FALSE	YES	0
14	Rainy	Mild	High	TRUE	NO	1
						4

No	Attributte	Jumlah Kasus (s)	Play No (s2)	Play Yes (s1)	Entropy	
1.1.1	Total	Humidity HIGH & Outlook RAINY	2	1	1	1
	Tempeature					
	Hot		0	0	0	0
	Cool		0	0	0	0
	Mild		2	1	1	1
	Windy					
	TRUE		1	1	0	0
	FALSE		1	0	1	0

Sum YES
0
0
1
1
1
1
1
1
0
1
1
1
1
1
1
0
10

Gain

0

1

No	Outlook	Temperature	Humidity	Windy	Play
1	Sunny	Hot	High	FALSE	No
2	Sunny	Hot	High	TRUE	No
3	Cloudy	Hot	High	FALSE	Yes
4	Rainy	Mild	High	FALSE	Yes
5	Rainy	Cool	Normal	FALSE	Yes
6	Rainy	Cool	Normal	TRUE	Yes
7	Cloudy	Cool	Normal	TRUE	Yes
8	Sunny	Mild	High	FALSE	No
9	Sunny	Cool	Normal	FALSE	Yes
10	Rainy	Mild	Normal	FALSE	Yes
11	Sunny	Mild	Normal	TRUE	Yes
12	Cloudy	Mild	High	TRUE	Yes
13	Cloudy	Hot	Normal	FALSE	Yes
14	Rainy	Mild	High	TRUE	No

KETERANGAN :

Outlook, temperature, humidity dan windy adalah Attribute

Dan Play adalah keputusan

Dari data diatas terdapat 14 kasus

Maka :

Setelah kita ketahui data atribut, keputusan dan jumlah kasusnya maka kita cari entropy-nya dengan rumus sebagai berikut :

$$\text{Entropy (S)} = \sum_{i=1}^n - p_i * \log_2 p_i$$

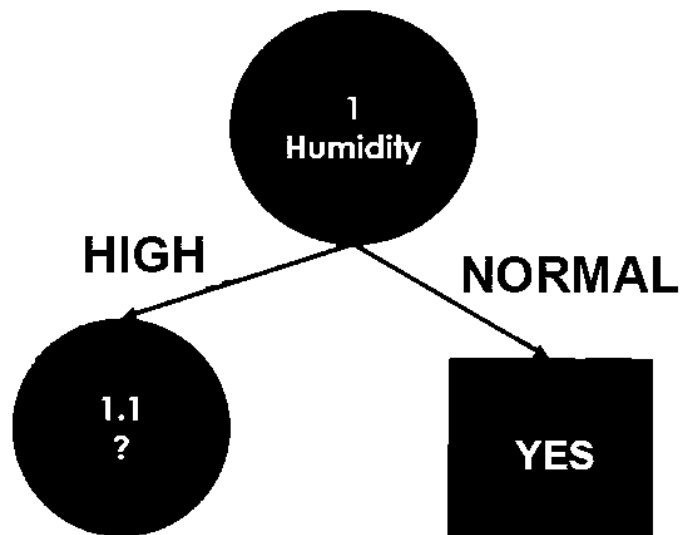
Lalu kita hitung gain-nya

$$\text{Gain (S,A)} = \text{Entropy (S)} - \sum_{i=1}^n p_i * \text{Entropy (S}_i\text{)}$$

Hasil perhitungan 1

		Jumlah kasus(S)	No (S1)	Yes (S2)	Entropy	Gain
Total		14	4	10	0,86312	
outlook						0,258512
	Cloudy	4	0	4	0	
	Rainy	5	1	4	0,721928	
	Sunny	5	3	2	0,970950	
Temperatur						0,183850
	Cool	4	0	4	0	
	Hot	4	2	2	1	
	Mild	6	2	4	0,918295	
Humidity						0,370506
	High	7	4	3	0,985228	
	Normal	7	0	7	0	
Windy						0,005977
	False	8	2	6	0,811278	
	True	6	4	2	0,918295	

Maka dapat digambarkan dalam pohon data :

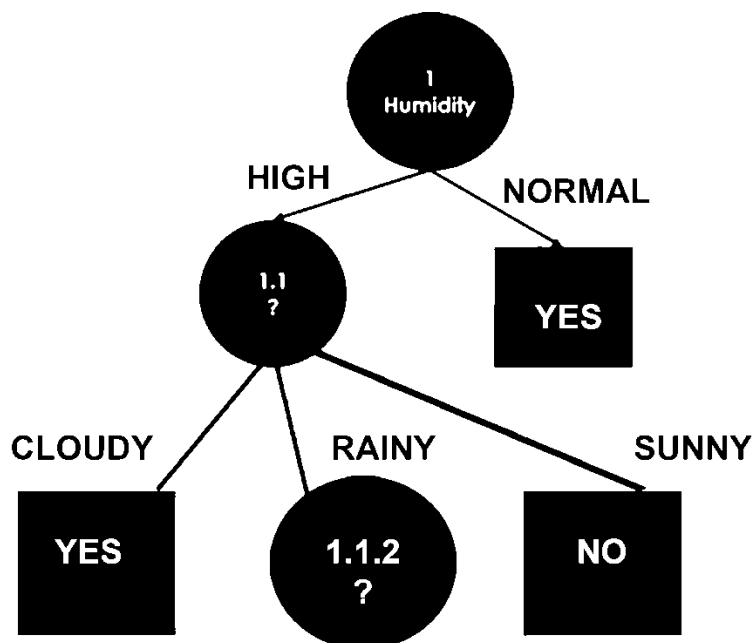


Karena nilai Humadity High masih mempunyai nilai semua maka masih perlu diperhitungkan lebih lanjut ke perhitungan 1.1

Hasil perhitungan 1.1

		Jumlah kasus(S)	No (S1)	Yes (S2)	Entropy	Gain
Humadity Hlgh		7	4	3	0,985228	
outlook						0,699513
	Cloudy	2	0	2	0	
	Rainy	2	1	1	1	
	Sunny	3	3	0	0	
Temperatur						0,020244
	Cool	0	0	0	0	
	Hot	3	2	1	0,918295	
	Mild	4	2	2	1	
Windy						0,020244
	False	4	2	2	1	
	True	3	2	1	0,918295	

Karena Outlook mempunyai 3 nilai Cloudy, Rainy, dan Sunny maka ia mempunyai 3 cabang pohon data yang dapat digambarkan sebagai berikut :



Karena rainy masih mempunyai nilai yang belum memiliki klasifikasi maka diperlukan perhitungan lebih lanjut

Hasil perhitungan Humidity High and Outlook Rainy

		Jumlah kasus(S)	No (S1)	Yes (S2)	Entropy	Gain
Humadity High and outlook Rainy		2	1	1	1	
Temperatur						0
	Cool	0	0	0	0	
	Hot	0	0	0	0	
	Mild	2	1	1	1	
Windy						1
	False	1	0	1	0	
	True	1	1	0	0	

Dari data tersebut dapat digambarkan dalam pohon data sebagai berikut:

