LAPORAN

PENERAPAN DECISION TREE PADA STUDI KASUS DATA HIPERTENSI

Nama : Iyan Zuli Armanda

NIM : 23051204165

Kelas : TI 23 E

Teknik

S1 - Teknik Informatika

Universitas Negeri Surabaya

2024

PENDAHULUAN

Decision Tree adalah salah satu algoritma dalam machine learning yang sering digunakan untuk pemodelan prediktif. Algoritma ini berbasis pada konsep pembagian data menjadi cabang-cabang yang merepresentasikan pilihan atau kondisi, hingga mencapai hasil akhir berupa keputusan. Decision tree bersifat intuitif karena menyerupai cara manusia membuat keputusan dengan memecah masalah kompleks menjadi bagian-bagian yang lebih sederhana.

Konsep Decision Tree

Decision tree terdiri dari beberapa elemen utama, yaitu:

1. Root Node: Node awal yang merepresentasikan seluruh dataset. Atribut dengan nilai paling informatif dipilih sebagai root.
2. Internal Node: Node bercabang yang mewakili atribut tertentu. Setiap cabang menggambarkan nilai dari atribut tersebut.
3. Leaf Node: Node terminal yang memberikan hasil akhir, seperti kategori atau nilai prediksi.
4. Cabang (Branch): Menghubungkan antar node untuk merepresentasikan jalur keputusan berdasarkan kondisi.

Tujuan utama dari decision tree adalah meminimalkan ketidakpastian atau entropi pada setiap langkah hingga menghasilkan keputusan yang paling optimal.

Proses Pembuatan Decision Tree

- Menghitung Entropy Awal

Hitung entropy untuk seluruh dataset untuk menentukan tingkat ketidakpastian awal.

- Menghitung Information Gain untuk Setiap Atribut
    - Bagi data berdasarkan nilai atribut.
    - Hitung entropy untuk setiap subset data.
    - Gunakan formula Information Gain untuk menentukan atribut paling informatif.
- Menentukan Root Node

Pilih atribut dengan nilai Information Gain tertinggi sebagai root node.

- Pembagian Data

Ulangi proses dengan dataset subset pada setiap cabang hingga semua data berada dalam kategori tertentu (leaf node) atau kriteria berhenti terpenuhi.

- Validasi dan Pruning

Validasi model untuk menghindari overfitting. Proses pruning dapat dilakukan untuk mengurangi kompleksitas tree.

STUDI KASUS (PENENTUAN HIPERTENSI)

- Data diambil 8 sample, dengan pemikiran bahwa yang mempengaruhi seseorang menderita hipertensi atau tidak adalah Usia, Berat Badan, dan Jenis Kelamin (Atribut)
- Usia mempunyai instance : muda & tua
- Berat badan mempunya instance : uderweight, average & over weight
- Jenis kelamin mempunyai instance : pria & wanita

Nama	Usia	Berat	Kelamin	Hipertensi
Ali	muda	overweight	pria	ya
Edi	muda	underweight	pria	tidak
Annie	tua	average	wanita	tidak
Budiman	tua	overweight	pria	tidak
Herman	tua	overweight	pria	ya
Didi	tua	underweight	pria	tidak
Rina	tua	overweight	wanita	ya
Gatot	tua	average	pria	tidak

Menentukan Node Terpilih

- Untuk menentukan node terpilih gunakan nilai Entropy dari setiap kriteria dengan data sample yang ditentuka
- Node terpilih adalah kriteria dengan Entropy yang paling kecil

Atribut Usia

Usia	Hipertensi	Jumlah
muda	ya	1
	tidak	3
tua	ya	2
	tidak	2

Usia = muda

$$q_1 = -\left(\frac{1}{4}\log 2\frac{1}{4}\right) - \left(\frac{3}{4}\log 2\frac{3}{4}\right) = 0.81$$

Usia = tua

$$q_2 = -\left(\frac{2}{4}\log 2\frac{2}{4}\right) - \left(\frac{2}{4}\log 2\frac{2}{4}\right) = 1$$

Entropy untuk usia

$$E = \left(\frac{4}{8}\right)q_1 + \left(\frac{4}{8}\right)q_2 = \left(\frac{4}{8}\right)0.81 + \left(\frac{4}{8}\right)1 = 0.91$$

Atribut Berat Badan

Berat	Hipertensi	Jumlah
overweight	ya	3
	tidak	1
average	ya	0
	tidak	2
underweight	ya	0
	tidak	2

Berat = overweight

$$q_1 = -\left(\frac{3}{4}\log 2\frac{4}{4}\right) - \left(\frac{1}{4}\log 2\frac{1}{4}\right) = 0.82$$

Berat = average

$$q_2 = -\left(\frac{0}{2}\log 2\frac{0}{2}\right) - \left(\frac{2}{2}\log 2\frac{2}{2}\right) = 0$$

Berat = underweight

$$q_3 = -\left(\frac{0}{2}\log 2\frac{0}{2}\right) - \left(\frac{2}{2}\log 2\frac{2}{2}\right) = 0$$

Entropy untuk Berat

$$E = \left(\frac{4}{8}\right)q_1 + \left(\frac{2}{8}\right)q_2 + \left(\frac{2}{8}\right)q_3 = \left(\frac{4}{8}\right)0.82 + 0 + 0 = 0.41$$

Atribut Jenis Kelamin

Kelamin	Hipertensi	Jumlah
pria	ya	2
	tidak	4
wanita	ya	1
	tidak	1

Kelamin = pria

$$q_1 = -\left(\frac{2}{6}\log 2\frac{2}{6}\right) - \left(\frac{4}{6}\log 2\frac{4}{6}\right) = 0.92$$

Kelamin = Wanita

$$q_1 = -\left(\frac{1}{2}\log 2\frac{1}{2}\right) - \left(\frac{1}{2}\log 2\frac{1}{2}\right) = 1$$

Entropy untuk kelamin

$$E = \left(\frac{6}{8}\right)q_1 + \left(\frac{2}{8}\right)q_2 = \left(\frac{6}{8}\right)0.92 + \left(\frac{2}{8}\right)1 = 0.94$$

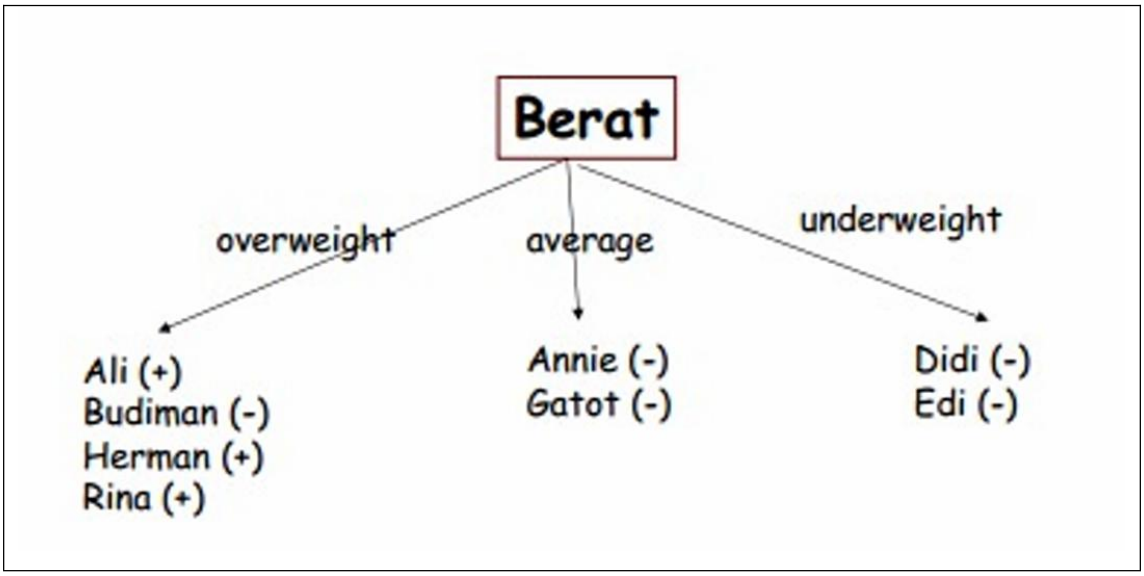
Pemilihan Entropy

- Atribut usia : 0.91
- Atribut berat : 0.41
- Atribut kelamin : 0.94

Terpilih atribut berat badan karena memiliki entropi paling kecil, 0.41

Menyusun Tree

Tree Awal :



Penentuan Leaf Node untuk Berat = Overweight

- Leaf Node berikutnya dapat dipilih yang memiliki nilai + dan

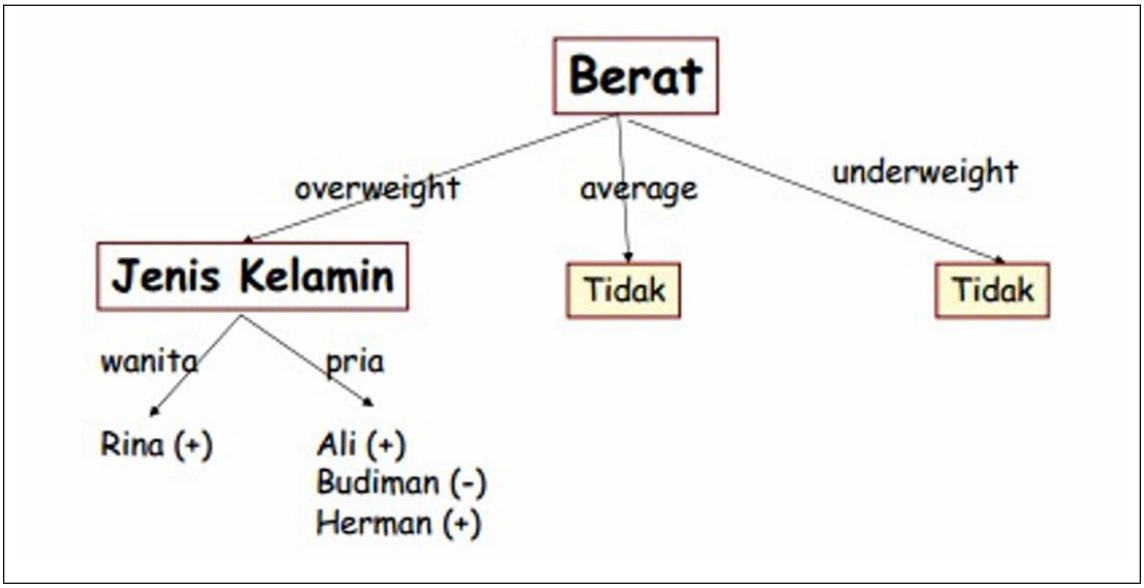
- Hanya overweight yang memiliki nilai + dan -

Nama	Usia	Kelamin	Hipertensi
Ali	muda	pria	ya
Budiman	tua	pria	tidak
Herman	tua	pria	ya
Rina	tua	wanita	ya

usia	hipertensi	jumlah
muda	ya	1
	tidak	0
tua	ya	2
	tidak	1
Entropy = 0.69		

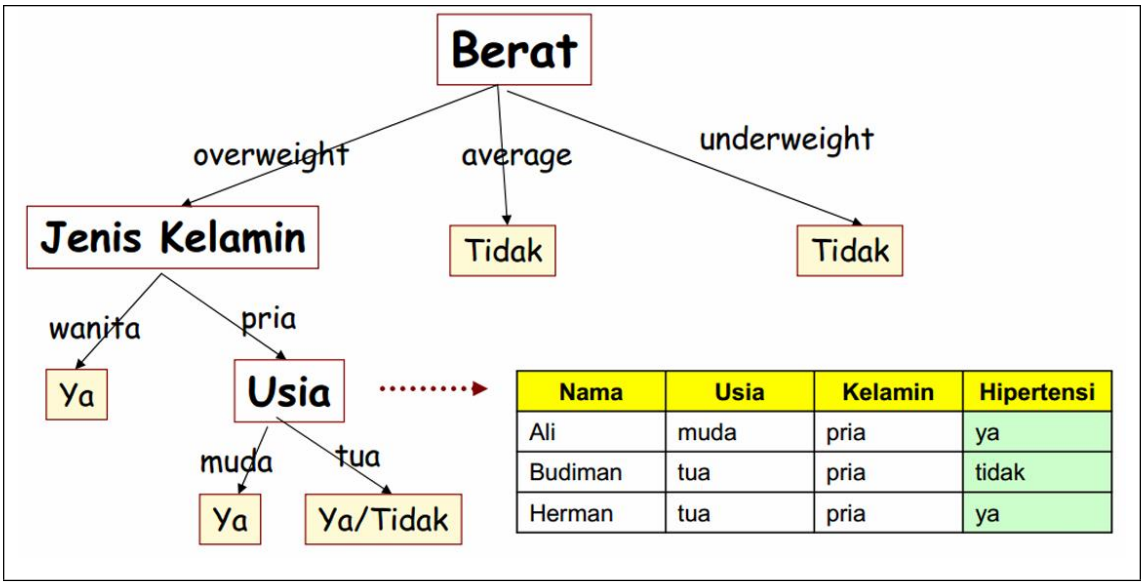
Kelamin	Hipertensi	Jumlah
pria	ya	2
	tidak	
wanita	ya	1
	tidak	0
Entropy = 0.69		

Pengeliminasian average dan underweight dikarenakan tidak terdapat +



Leaf Node usia & jenis kelamin memiliki entropy sama, maka diperlukan pengetahuan pakar atau percaya saja pada hasil aca

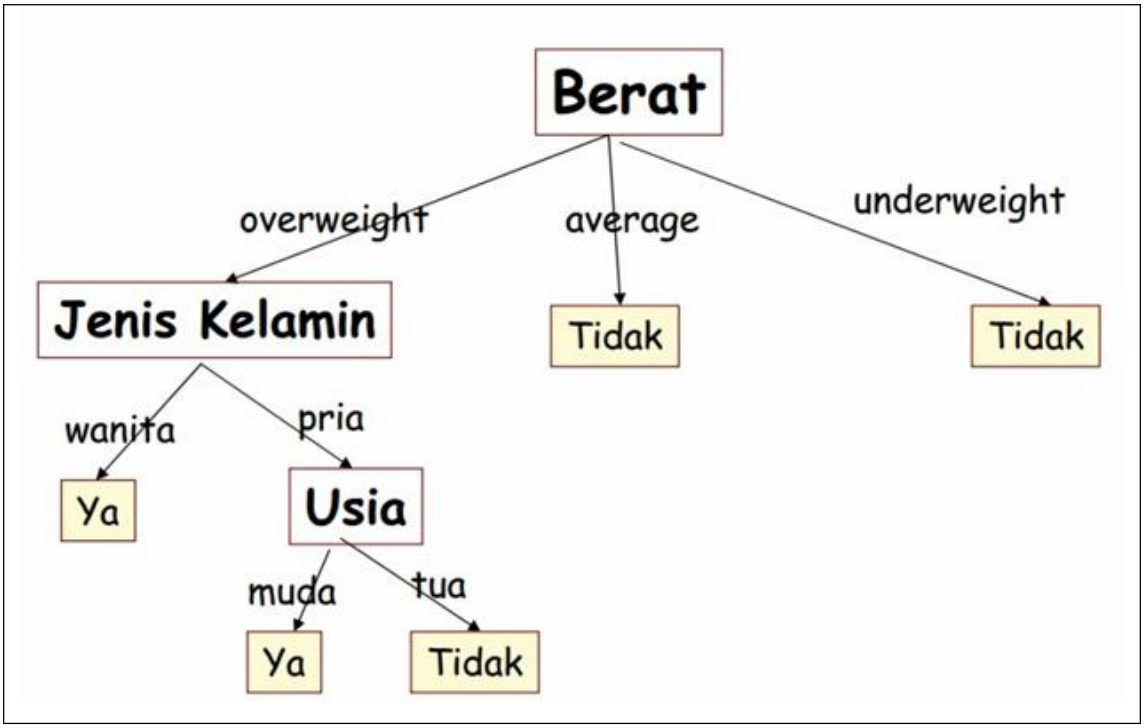
Hasil Tree :



Nama	Usia	Kelamin	Hipertensi
Ali	muda	pria	ya
Budiman	tua	pria	tidak
Herman	tua	pria	ya

- Pada usia = tua ternyata ada 1 data menyatakan ya dan 1 data menyatakan tidak, keadaan ini perlu dicermati.
- Pilihan hanya dapat ditentukan dengan campur tangan seorang pakar

Mengubah Tree Menjadi Rule



R1:IF berat = average OR berat = underweight THEN hipertensi = tidak

R2:IF berat = overweight AND j.kelamin = wanita THEN hipertensi = ya

R3:IF berat = overweight AND j.kelamin = pria AND usia = muda THEN hipertensi = ya

R4:berat = overweight AND j.kelamin = pria AND usia = tua THEN hipertensi = tidak

Hasil Prediksi

Nama	Usia	Berat	Kelamin	Hipertensi	Prediksi
Ali	muda	overweight	pria	ya	ya
Edi	muda	underweight	pria	tidak	tidak
Annie	tua	average	wanita	tidak	tidak
Budiman	tua	overweight	pria	tidak	tidak
Herman	tua	overweight	pria	ya	tidak
Didi	tua	underweight	pria	tidak	tidak
Rina	tua	overweight	wanita	ya	ya
Gatot	tua	average	pria	tidak	tidak

Kesalahan 12.5% (1 dari 8 data)

Implementasi Program dalam Python

```
from sklearn.tree import DecisionTreeClassifier, export_text
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score

data = [
    ["muda", "overweight", "pria", 1, "Ali"],
    ["muda", "underweight", "pria", 0, "Edi"],
    ["tua", "average", "wanita", 0, "Annie"],
    ["tua", "overweight", "pria", 0, "Budiman"],
    ["tua", "overweight", "pria", 1, "Herman"],
    ["tua", "underweight", "pria", 0, "Didi"],
    ["tua", "overweight", "wanita", 1, "Rina"],
    ["tua", "average", "pria", 0, "Gatot"],
]

map_usia = {"muda": 0, "tua": 1}
map_berat = {"underweight": 0, "average": 1, "overweight": 2}
map_kelamin = {"pria": 0, "wanita": 1}

X = [[map_usia[row[0]], map_berat[row[1]], map_kelamin[row[2]]] for row in data]
y = [row[3] for row in data]

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.25, random_state=42)

model = DecisionTreeClassifier(criterion="entropy", random_state=42)
model.fit(X_train, y_train)

y_pred = model.predict(X_test)
accuracy = accuracy_score(y_test, y_pred)

print("Accuracy:", accuracy * 100, "%")
print("\nDecision Tree Rules:")
tree_rules = export_text(model, feature_names=["Usia", "Berat Badan", "Jenis Kelamin"])
print(tree_rules)
```

KESIMPULAN

Berdasarkan analisis data hipertensi menggunakan decision tree, atribut Berat Badan terpilih sebagai root node karena memiliki pengaruh terbesar terhadap risiko hipertensi. Model ini menghasilkan peluang kesalahan sebesar 12.5% (1 dari 8 data), sehingga memiliki akurasi sekitar 87.5%. Individu dengan kondisi overweight memiliki risiko lebih tinggi terkena hipertensi, sementara atribut Usia memberikan kontribusi tambahan, dan Jenis Kelamin kurang signifikan dalam dataset ini. Model ini memberikan gambaran awal yang baik, namun akurasinya dapat ditingkatkan dengan dataset yang lebih besar dan penambahan faktor risiko lainnya.