Perbandingan Algoritma Naive Bayes dan Algoritma Decision Tree Learning Untuk Analisa Prediksi Penyakit Jantung

Mata Kuliah Manajemen Sains

Dosen Pengampu : Wiji Lestari, M.Kom

Oleh :

Rafel Fernando

220101031

Sistem Informasi 22A1

Fakultas Ilmu Komputer

Universitas Duta Bangsa Surakarta

Tahun Pelajaran 2023/2024

PENDAHULUAN

Penyakit jantung (Heart Disease) adalah kondisi ketika bagian jantung yang meliputi pembuluh darah jantung, selaput jantung, katup jantung, dan otot jantung mengalami gangguan. Penyakit jantung bisa disebabkan oleh berbagai hal, seperti sumbatan pada pembuluh darah jantung, peradangan, infeksi, atau kelainan bawaan. Banyak penelitian yang telah dilakukan untuk membangun model prediktif yang dapat mengidentifikasi resiko yang tinggi yaitu dengan menggunakan machine learning.

Machine Learning adalah Pembelajaran mesin yang ditujukan untuk memahami dan membangun suatu metode yang memanfaatkan data untuk meningkatkan kinerja dan dikembangkan untuk bisa mendapatkan berbagai informasi secara mandiri. Jadi pengembang hanya perlu memprogramnya sekali dengan algoritma tertentu. Selanjutnya mesin akan bekerja sesuai dengan sistem algoritma yang ditanam padanya. Tentunya hal ini menjadi sebuah penemuan besar demi kemudahan berbagai kegiatan. Melakukan dengan metode klasifikasi.

Klasifikasi dalam machine learning adalah suatu pengelompokan data dimana data yang digunakan tersebut mempunyai kelas label atau target. Algoritma yang paling akurat dan paling populer dalam metode klasifikasi adalah algoritma Naive Bayes dan Decision Tree dan melakukan perbandingan untuk melihat akurasi tertinggi pada algoritma tersebut.

Pada studi ini, saya akan menjelaskan bagaimana algoritma Naive Bayes dan Decision Tree Learning dapat digunakan untuk mendeteksi kemungkinan orang memiliki penyakit jantung. Saya akan menggunakan dataset yang mencakup atribut age, sex, cp, trestbps, chol, fbs, restecg, thalach, exang, oldpeak, slope, ca, thal, dan target (class). Melalui pendekatan ini diharapkan dapat meningkatkan kemampuan dalam mengidentifikasi individu yang memiliki resiko tinggi dan dapat melakukan langkah-langkah pencegahan yang tepat.

Studi ini dilakukan untuk memberikan kontribusi dalam pengembangan sistem pendukung keputusan di bidang kesehatan. Dengan menggunakan algoritma Naive Bayes dan Decision Tree, kita memperoleh informasi yang lebih dalam faktor-faktor yang mempengaruhi seseorang kemungkinan mengalami penyakit jantung. Dan dapat mengetahui mana algoritma yang lebih akurat dalam mendeteksi seseorang memiliki penyakit jantung.

DESKRIPSI DATASET

Menurut Organisasi Kesehatan Dunia (WHO), Penyakit jantung merupakan satu dari tiga penyakit dengan tingkat kematian tertinggi dan lebih dari 17,3 juta orang didunia meninggal dunia karena penyakit jantung.

Row No.	target	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang
1	0	52	1	0	125	212	0	1	168	0
2	0	53	1	0	140	203	1	0	155	1
3	0	70	1	0	145	174	0	1	125	1
4	0	61	1	0	148	203	0	1	161	0
5	0	62	0	0	138	294	1	1	106	0
6	1	58	0	0	100	248	0	0	122	0
7	0	58	1	0	114	318	0	2	140	0
8	0	55	1	0	160	289	0	0	145	1
9	0	46	1	0	120	249	0	0	144	0
10	0	54	1	0	122	286	0	0	116	1
11	1	71	0	0	112	149	0	1	125	0
12	0	43	0	0	132	341	1	0	136	1
13	1	34	0	1	118	210	0	1	192	0
14	0	51	1	0	140	298	0	1	122	1

	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal
	125	212	0	1	168	0	1	2	2	3
	140	203	1	0	155	1	3.100	0	0	3
	145	174	0	1	125	1	2.600	0	0	3
	148	203	0	1	161	0	0	2	1	3
	138	294	1	1	106	0	1.900	1	3	2
	100	248	0	0	122	0	1	1	0	2
	114	318	0	2	140	0	4.400	0	3	1
	160	289	0	0	145	1	0.800	1	1	3
	120	249	0	0	144	0	0.800	2	0	3
	122	286	0	0	116	1	3.200	1	2	2
	112	149	0	1	125	0	1.600	1	0	2
	132	341	1	0	136	1	3	1	0	3
	118	210	0	1	192	0	0.700	2	0	2
	140	298	0	1	122	1	4.200	1	3	3

Informasi Atribut pada Dataset diatas adalah sebagai berikut

1. Age (Umur)
   Merupakan atribut usia dari pasien dalam dataset.
2. Sex (Jenis Kelamin)
   Merupakan atribut yang mencakup jenis kelamin dari pasien dimana 0 = Perempuan dan 1 = Laki-Laki dalam dataset.
3. Cp (Chest Pain)
   Merupakan atribut yang mencakup chest pain atau nyeri dada pasien dalam dataset.
4. Threstbps
   Merupakan atribut yang mencakup tekanan darah pasien dalam dataset.
5. Chol
   Merupakan atribut yang mencakup kolestrol pasien.

6. Fbs
   Merupakan atribut yang mencakup kadar gula darah puasa pada pasien
7. Restecg
   Merupakan atribut yang mencakup aktivitas listrik jantung pada saat istirahat pada pasien.
8. Thalach
   Merupakan atribut yang mencakup penyakit genetik yang mempengaruhi produksi hemoglobin.
9. Exang
   Merupakan atribut yang mencakup Angina Pectoris atau nyeri dada yang terjadi ketika otot jantung tidak mendapat cukup oksigen pada pasien.
10. Oldpeak
    Merupakan atribut yang mencakup tes latihan jantung untuk menilai aktivitas jantung dan respons jantung selama latihan pada pasien.
11. Slope
    Merupakan atribut yang mencakup tes latihan jantung mengacu kemiringan segmen ST pada EKG selama latihan pada pasien.
12. Ca
    Merupakan atribut yang mencakup kalsium pada pasien.
13. Thal
    Merupakan atribut yang mencakup jenis thalasemia yang dimiliki oleh pasien.

Dataset ini memiliki class pada atribut sebagai berikut :

1. Target
   Merupakan atribut yang mencakup class untuk 0 = normal dan 1 = penyakit jantung pada pasien

ALGORITMA

Naive bayes merupakan metode pengklasifikasian berdasarkan probabilitas sederhana dan dirancang agar dapat dipergunakan dengan asumsi antar variabel penjelas saling bebas (independen). Pada algoritma ini pembelajaran lebih ditekankan pada pengestimasian probabilitas. Keuntungan algoritma naive bayes adalah tingkat nilai error yang didapat lebih rendah ketika dataset berjumlah besar, selain itu akurasi naive bayes dan kecepatannya lebih tinggi pada saat diaplikasikan ke dalam dataset yang jumlahnya lebih besar.

Probabilitas X di dalam Y adalah probabilitas interseksi X dan Y dari probabilitas Y, atau dengan bahasa lain. P(X|Y) adalah prosentase banyaknya X di dalam Y

$$P(X \mid Y) = \frac{P(X \cap Y)}{P(Y)}$$

Decision tree adalah metode pemodelan prediktif dalam analisis data yang menggunakan struktur pohon. Tujuan decision tree adalah untuk menggambarkan serta membuat keputusan berdasarkan serangkaian aturan dan kondisi.

Entropy merupakan suatu parameter untuk mengukur heterogenitas (keberagaman) dari suatu kumpulan sampel data. Jika kumpulan data semakin heterogen maka maka nilai entropinya semakin besar.

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

n : Jumlah partisi S

pi : Proporsi dari terhadap S

Di mana c adalah jumlah nilai yang ada pada atribut target (jumlah kelas klasifikasi). Sedangkan pi adalah menyatakan jumlah sampel pada kelas i. Pada data prediksi penyakit jantung , jumlah kelas adalah 2, yaitu “0” atau “1” (c = 2).

Pemilihan atribut sebagai simpul, baik simpul akar (root) atau simpul internal didasarkan pada nilai Gain tertinggi dari atribut-atribut yang ada. Penghitungan nilai Gain digunakan rumus

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

S : Himpunan kasus

A : Atribut

n : Jumlah partisi himpunan atribut

A ||: Jumlah kasus pada partisi ke- i

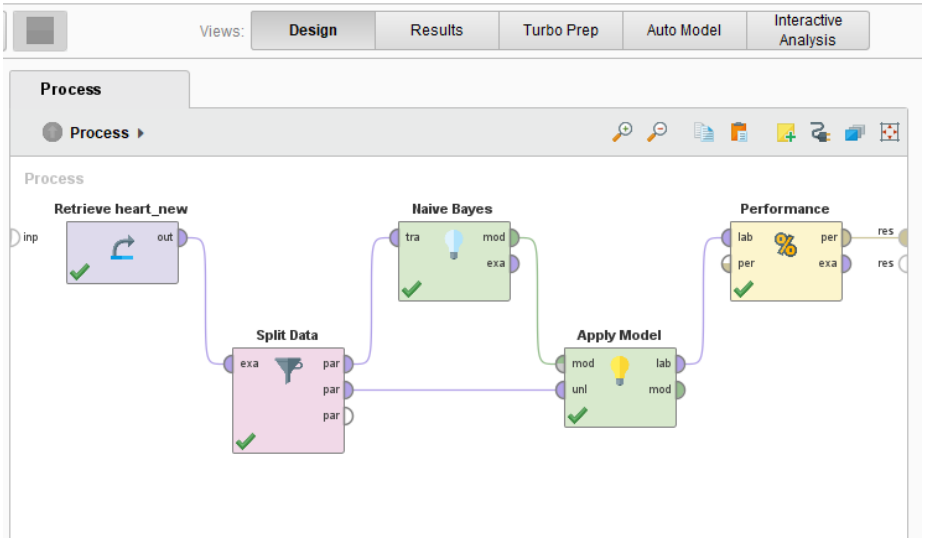
| S | : Jumlah kasus dalam S

Pada tabel atribut target =“1” dikatakan sebagai sampel positif (+), dan atribut target = “0” dikatakan sebagai sampel negatif (-).

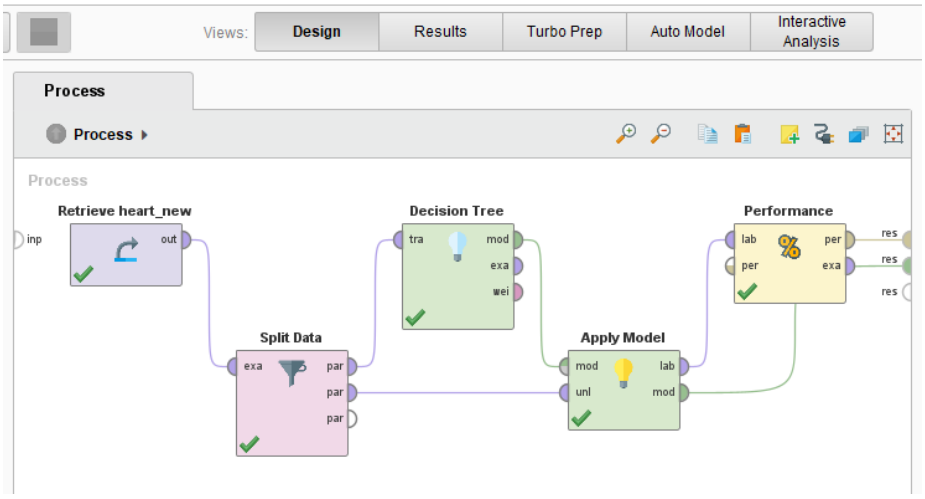
PEMBANGUNAN MODEL

Penelitian ini bertujuan untuk menerapkan algoritma Naive Bayes dan Algoritma Decision Tree Learning untuk mendapatkan tingkat akurasi Penyakit Jantung. Pada tahap ini dataset akan dibagi menjadi Data Training dan Data Testing.

1. Pembangunan Model Menggunakan Algoritma Naive Bayes menggunakan Rapidminer



2. Pembangunan Model Menggunakan Algoritma Decesion Tree menggunakan Rapidminer



PENGUJIAN MODEL

1. NAIVE BAYES

a. Menghitung akurasi secara manual

Accuracy = $\frac{tp+tn}{(tp+tn+fp+fn)} x\ 100\ \%$

= $\frac{131+120}{(131+120+30+27)}\ x\ 100\ \%$

= 0,8149 = 81,49 %

b. Menghitung akurasi menggunakan Rapidminer

Table View Plot View

accuracy: 81.49%

	true 0	true 1	class precision
pred. 0	120	27	81.63%
pred. 1	30	131	81.37%
class recall	80.00%	82.91%	

2. DECISION TREE LEARNING

a. Menghitung akurasi secara manual

Accuracy = $\frac{tp+tn}{(tp+tn+fp+fn)} x\ 100\ \%$

= $\frac{132+138}{(132+138+12+26)} x\ 100\%$

= 0,8766 = 87,66%

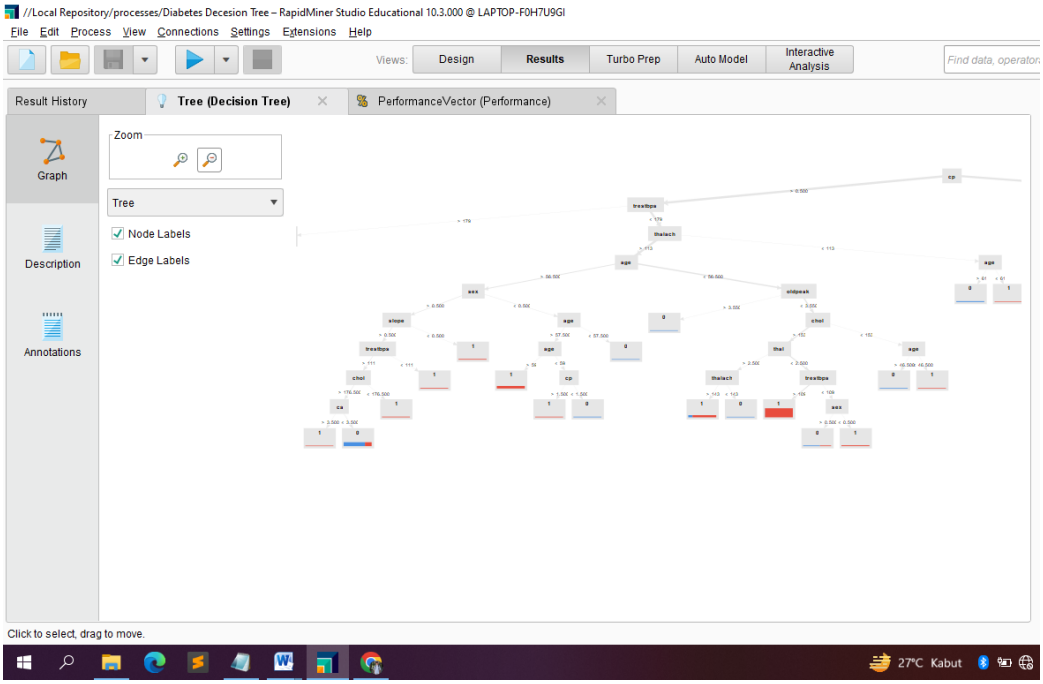
b. Menghitung akurasi menggunakan Rapidminer

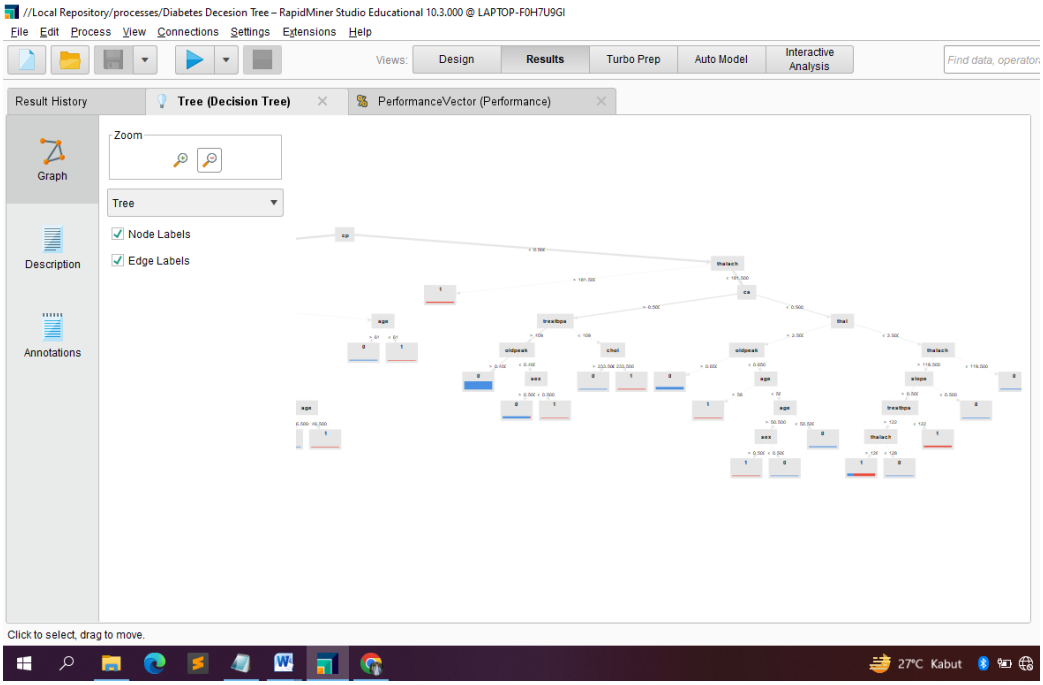
Table View Plot View

accuracy: 87.66%

	true 0	true 1	class precision
pred. 0	138	26	84.15%
pred. 1	12	132	91.67%
class recall	92.00%	83.54%	

c. Pohon keputusan





3. Tabel Performance algoritma Machine Learning (Naive Bayes dan Decesion Tree)

	Algoritma	
	Naive Bayes	Decision Tree Learning
Accuracy	81.49 %	87.66 %
Classification error	18.51 %	12.34 %
Weighted mean recall	81.46 %	87.77 %
Weighted mean precision	81.50 %	87.91 %

PENUTUP

Kesimpulan

Berdasarkan klasifikasi menggunakan algoritma diatas, diperoleh bahwa nilai akurasi algoritma Decesion Tree lebih akurat dibandingkan algoritma Naive Bayes. Dapat dikatakan penggunaan algoritma Decesion Tree lebih baik untuk jumlah data yang banyak mencapai ribuan. Hasil akhir menyatakan bahwa penggunaan algoritma Naive Bayes dalam klasifikasi penyakit jantung pada penelitian ini mencapai tingkat akurasi 81.49 % dan penggunaan algoritma Decision Tree Learning dalam klasifikasi penyakit jantung pada penelitian ini mencapai tingkat akurasi 87,66%. Dalam penghitungan akurasi secara manual dan menggunakan rapidminer didapatkan hasil yang sama. Hasil pengklasifikasian dengan algoritma Decision Tree Learning mampu menganalisi penyakit jantung.