

Klasifikasi Jenis Penyakit dengan Algoritma Decision Tree Menggunakan Rapid Miner

Iswadi Hamzah¹, Muhammad Iqbal², Rian Farta Wijaya³

¹ Magister Teknologi Informasi,
Universitas Pembangunan Panca Budi, Kota Medan, Indonesia
E-mail: iswadi.unpab@gmail.com,

Corresponding Author: Iswadi Hamzah

ABSTRACT

Each disease has certain indicators that can provide clues that there are several factors that can cause a particular disease. So that it can be used as a guideline in carrying out analysis as early as possible, the data taken is sourced from Kaggle which consists of training data and testing data with several attributes such as, itching, skin rash, nodal skin eruptions, continuous sneezing, shivering, chills, joint pain, stomach pain, acidity, ulcers on tongue, muscle wasting, vomiting, burning micturition, spotting urination, fatigue, weight gain, anxiety, cold hands and feets, mood swings, weight loss, restlessness, lethargy, patches in throat, irregular sugar level, cough, high fever, sunken eyes, breathlessness, sweating, dehydration, indigestion, headache, yellowish skin, dark urine, nausea, loss of appetite, pain behind the eyes, back pain, constipation, abdominal pain, diarrhea, mild fever, yellow urine, yellowing of eyes, acute liver failure, fluid overload, swelling of stomach, swelled lymph nodes, malaise, blurred and distorted vision, phlegm, throat irritation, redness of eyes, sinus pressure, runny nose, congestion, chest pain, weakness in limbs, fast heart rate, pain during bowel movements, pain in anal region, bloody stool, irritation in anus, neck pain, dizziness, cramps, bruising, obesity, swollen legs, swollen blood vessels, puffy face and eyes, enlarged thyroid, brittle nails, swollen extremities, excessive hunger, extra marital contacts, drying and tingling lips, slurred speech, knee pain, hip joint pain, muscle weakness, stiff neck, swollen joints, movement stiffness, spinning movements, loss of balance, unsteadiness, weakness of one body side, loss of smell, bladder discomfort, foul smell of urine, continuous feeling of urine, passage of gases, internal itching, toxic look (typhos), depression, irritability, muscle pain, altered sensorium, red spots over body, belly pain, abnormal menstruation, dyschromic patches, watering from eyes, increased appetite, polyuria, family history, mucoid sputum, rusty sputum, lack of concentration, visual disturbances, receiving blood transfusion, receiving unsterile injections, coma, stomach bleeding, distention of abdomen, history of alcohol consumption, fluid overload, blood in sputum, prominent veins on calf, palpitations, painful walking, pus filled pimples, blackheads, scurring, skin peeling, silver like dusting, small dents in nails, inflammatory nails, blisters, red sore around nose and yellow crust ooze. These attributes will later be given a value and analyzed using a decision tree algorithm using a rapid miner so that they can identify a particular disease.

Keywords: Decision Tree, Rapid Miner

ABSTRAK

Setiap penyakit memiliki indikator tertentu yang dapat memberikan petunjuk bahwasanya terdapat beberapa faktor yang dapat menjadi penyebab penyakit tertentu. Sehingga dapat dijadikan sebuah pedoman dalam melakukan analisa sedini mungkin, data yang diambil bersumber dari Kaggle yang terdiri dari data training dan data testing dengan beberapa atribut seperti, *itching, skin rash, nodal skin eruptions, continuous sneezing, shivering, chills, joint pain, stomach pain, acidity, ulcers on tongue, muscle wasting, vomiting, burning micturition, spotting urination, fatigue, weight gain, anxiety, cold hands and feets, mood swings, weight loss, restlessness, lethargy, patches in throat, irregular sugar level, cough, high fever, sunken eyes, breathlessness, sweating, dehydration, indigestion, headache, yellowish skin, dark urine, nausea, loss of appetite, pain behind the eyes, back pain, constipation, abdominal pain, diarrhoea, mild fever, yellow urine, yellowing of eyes, acute liver failure, fluid overload, swelling of stomach, swelled lymph nodes, malaise, blurred and distorted vision, phlegm, throat irritation, redness of eyes, sinus pressure, runny nose, congestion, chest pain, weakness in limbs, fast heart rate, pain during bowel movements, pain in anal region, bloody stool, irritation in anus, neck pain, dizziness, cramps, bruising, obesity, swollen legs, swollen blood vessels, puffy face and eyes, enlarged thyroid, brittle nails, swollen extremities, excessive hunger, extra marital contacts, drying and tingling lips, slurred speech, knee pain, hip joint pain, muscle weakness, stiff neck, swelling joints, movement stiffness, spinning movements, loss of balance, unsteadiness, weakness of one body side, loss of smell, bladder discomfort, foul smell of urine, continuous feel of urine, passage of gases, internal itching, toxic look (typhos), depression, irritability, muscle pain, altered sensorium, red spots over body, belly pain, abnormal menstruation, dischromic patches, watering from eyes, increased appetite, polyuria, family history, mucoid sputum, rusty sputum, lack of concentration, visual disturbances, receiving blood transfusion, receiving unsterile injections, coma, stomach bleeding, distention of abdomen, history of alcohol consumption, fluid overload, blood in sputum, prominent veins on calf, palpitations, painful walking, pus filled pimples, blackheads, scurring, skin peeling, silver like dusting, small dents in nails, inflammatory nails, blister, red sore around nose and yellow crust ooze.* Dari atribut tersebut nantinya akan diberi nilai dan dianalisa dengan algoritma decision tree menggunakan rapid miner sehingga dapat melakukan identifikasi terhadap suatu penyakit tertentu.

Kata Kunci: Decision Tree, Rapid Miner



Lisensi

Lisensi Internasional Creative Commons Attribution-ShareAlike 4.0.

pal pita tion s	pal ntu _ali ng	pus _fill ed _pi mpl es	bla _ck _ea _st	acu _rin g	ski _r _pe lin g	sili _ke _di _usi _ng	sm _ile _ms _in _n _alis	infl _am _mar _bit _ter	red _so _ari _ust _no _ce	prognosis
0	0	0	0	0	0	0	0	0	0	Fungal Infection
0	0	0	0	0	0	0	0	0	0	Fungal Infection
0	0	0	0	0	0	0	0	0	0	Fungal Infection
0	0	0	0	0	0	0	0	0	0	Fungal Infection
0	0	0	0	0	0	0	0	0	0	Fungal Infection
0	0	0	0	0	0	0	0	0	0	Fungal Infection
0	0	0	0	0	0	0	0	0	0	Fungal Infection
0	0	0	0	0	0	0	0	0	0	Fungal Infection
0	0	0	0	0	0	0	0	0	0	Fungal Infection
0	0	0	0	0	0	0	0	0	0	Fungal Infection
0	0	0	0	0	0	0	0	0	0	Fungal Infection
0	0	0	0	0	0	0	0	0	0	Fungal Infection
0	0	0	0	0	0	0	0	0	0	Fungal Infection
0	0	0	0	0	0	0	0	0	0	Allergy
0	0	0	0	0	0	0	0	0	0	Allergy
0	0	0	0	0	0	0	0	0	0	Allergy
0	0	0	0	0	0	0	0	0	0	Allergy
0	0	0	0	0	0	0	0	0	0	Allergy
0	0	0	0	0	0	0	0	0	0	Allergy
0	0	0	0	0	0	0	0	0	0	Allergy
0	0	0	0	0	0	0	0	0	0	Allergy
0	0	0	0	0	0	0	0	0	0	Allergy
0	0	0	0	0	0	0	0	0	0	Allergy
0	0	0	0	0	0	0	0	0	0	Allergy
0	0	0	0	0	0	0	0	0	0	Allergy
0	0	0	0	0	0	0	0	0	0	Allergy
0	0	0	0	0	0	0	0	0	0	Allergy
0	0	0	0	0	0	0	0	0	0	GERD
0	0	0	0	0	0	0	0	0	0	GERD
0	0	0	0	0	0	0	0	0	0	GERD
0	0	0	0	0	0	0	0	0	0	GERD
0	0	0	0	0	0	0	0	0	0	GERD
0	0	0	0	0	0	0	0	0	0	GERD
0	0	0	0	0	0	0	0	0	0	GERD
0	0	0	0	0	0	0	0	0	0	GERD
0	0	0	0	0	0	0	0	0	0	GERD

Metodologi

Berisi mengenai tahapan-tahapan penelitian dimulai dari penelitian awal berupa pengumpulan data sampai kepada penerapan algoritma kepada aplikasi rapid miner sehingga ditemukan kecocokan antara hasil akhir terhadap masing-masing perhitungan. Output dari penelitian ini adalah analisa terhadap hasil faktor jenis penyakit.

Pada penelitian ini, peneliti menemukan beberapa keadaan yang membuat peneliti tertarik untuk melakukan suatu penganalisaan terhadap kemungkinan – kemungkinan yang akan membuat seseorang terkena penyakit. Tahapan dari penelitian ini dilakukan dengan beberapa metode tahapan yang saling berhubungan antara satu dengan yang lainnya. Metode yang dilakukan berupa:

1. Mengambil dataset dari Kaggle dalam format csv.
2. Melakukan pengolahan data dengan menggunakan software rapid miner.
3. Setelah data diklasifikasikan ke dalam .csv, maka data tersebut kemudian akan diproses dengan menggunakan software rapid miner. Dengan menggunakan software tersebut maka data akan dapat dianalisa untuk mendapatkan sebuah keputusan.

KDD (Knowledge Discovery In Database) Istilah data mining dan knowledge discovery in database (KDD) sering kali digunakan secara bergantian untuk menjelaskan proses penggalian informasi tersembunyi dalam suatu basis data yang besar. [4] Proses KDD secara garis besar sebagai berikut :

1. Data Selection Pemilihan (Seleksi) data dari sekumpulan data operasional perlu dilakukan sebelum tahap penggalian informasi dalam KDD dimulai.
2. Pre Processing / Pembersihan Data Sebelum proses data mining dapat dilaksanakan, perlu dilakukan proses pembersihan pada data yang menjadi fokus KDD.
3. Transformation pada data yang telah dipilih, sehingga data tersebut sesuai proses data mining.
4. Data mining adalah proses mencari pola atau informasi menarik dalam data terpilih dengan menggunakan teknik atau metode tertentu
5. Evaluasi Pola informasi yang dihasilkan dari proses data mining perlu ditampilkan dalam bentuk yang mudah dimengerti oleh pihak yang berkepentingan



Decision Tree C45

Decision tree adalah salah satu metode klasifikasi yang paling populer dan banyak digunakan karena bersifat praktis. salah satu metode pohon keputusan yang paling banyak digunakan adalah Iterative Dychotomizer Version 3. Metode ID3 berusaha membangun model klasifikasi yang berupa pohon keputusan secara top-down. Caranya adalah dengan mengevaluasi semua atribut menggunakan suatu ukuran statistik, biasanya berupa information gain, untuk mengukur efektifitas suatu atribut dalam mengklasifikasikan himpunan sampel data. Dengan tujuan untuk mendefinisikan gain, pertama-tama digunakanlah ide dari teori informasi yang disebut entropy. Entropy mengukur jumlah dari informasi yang ada pada atribut dengan rumus :

$$Entropy(S) = -P_+ \log_2 P_+ - P_- \log_2 P_-$$

- S = ruang data sample
 P₊ = jumlah bersolusi positif (mendukung)
 P₋ = jumlah bersolusi negatif (tidak mendukung)

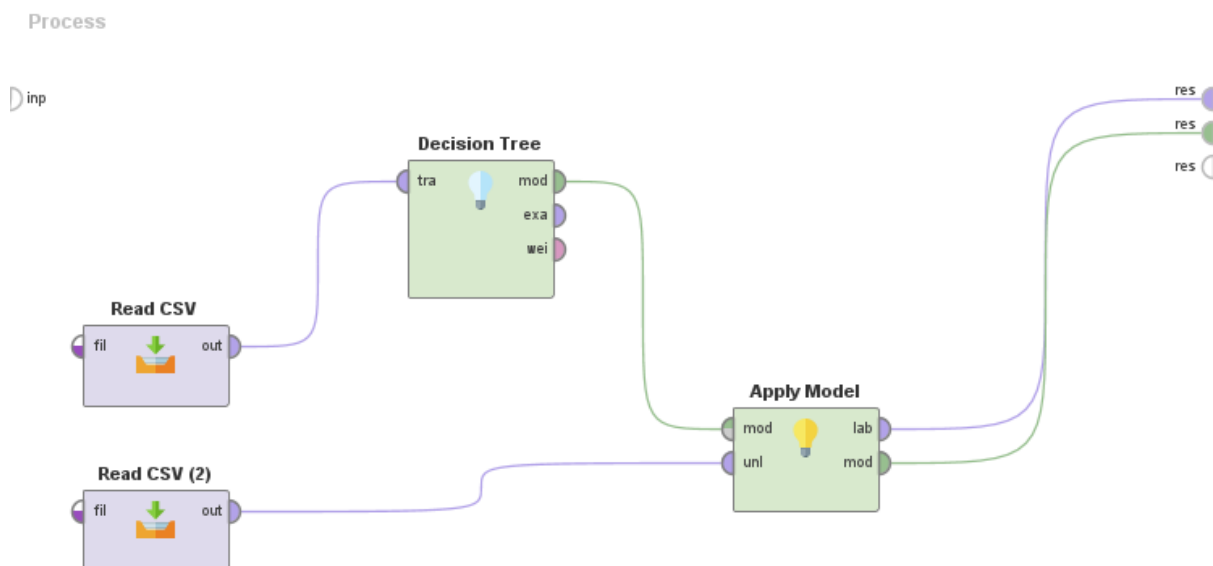
Dari rumus entropy diatas dapat disimpulkan bahwa definisi entropy (S) adalah jumlah bit yang diperkirakan dibutuhkan untuk dapat mengekstrak suatu kelas (+ atau-) dari sejumlah data acak pada suatu ruang sampel S. Entropy bisa dikatakan sebagai kebutuhan bit untuk menyatakan suatu kelas. Semakin kecil nilai entropy maka semakin baik digunakan dalam mengekstraksi suatu kelas. Pada algoritma ID3 pengurangan entropy disebut dengan informasi gain. Pembagian sample S terhadap atribut A dapat dihitung information gain dengan rumus:

$$Gain(S, A) = Entropy(S) - \sum_{v=nilai(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

Pilih atribut yang memiliki nilai information gain terbesar, ulangi proses perhitungan information gain akan terus dilaksanakan sampai semua data telah masuk dalam kelas yang sama. Atribut yang telah dipilih tidak diikuti lagi dalam perhitungan nilai information gain. [5]

Hasil dan Pembahasan

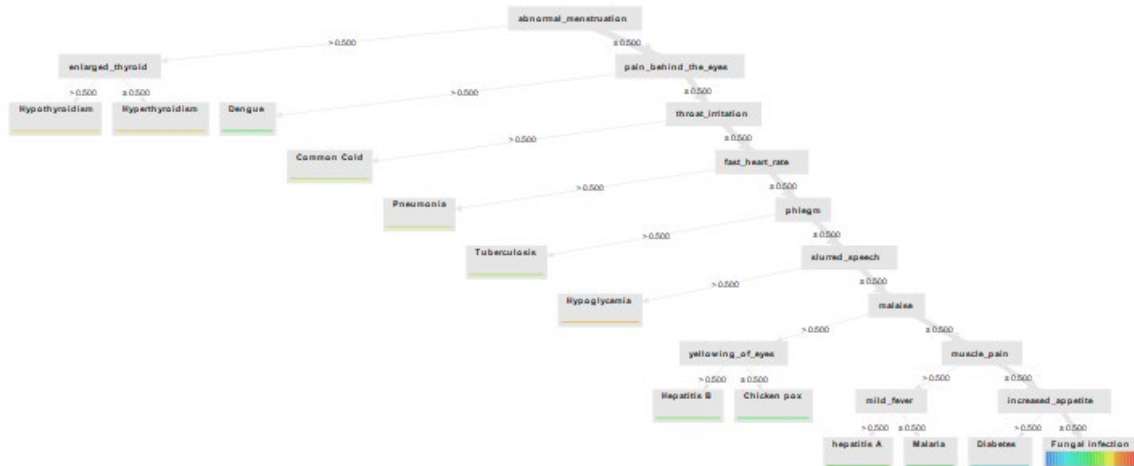
Pengujian Analisa ini menggunakan Rapidminer seperti berikut ini :



Gambar 1 : Pengujian Decision Tree



Data Training yang digunakan dalam penelitian ini adalah data indikator diabetes sebanyak 47.276 dari 6.714 dinyatakan terindikasi diabetes dan 40.562 negatif diabetes.



Gambar 2 : Pohon Keputusan

Evaluasi dan Pengujian Hasil

Setelah kedua data diolah maka dapat diuji tingkat akurasi untuk metode Decision Tree. Penelitian ini adalah untuk menentukan data yang paling akurat dengan sebagai acuan diagnosis pada penyakit diabetes. Pengujian tingkat akurasi dilakukan menggunakan confusion matrix dan kurva ROC/AUC (Area Under Cover). Hasil perhitungan data training menggunakan Algoritma Decision Tree (C4.5). Diketahui tingkat akurasi adalah 85,78%. Tabel Confussion Matrix disajikan pada tabel 2 dan gambar 5 merupakan grafik ROC/AUC (Area Under Cover) dari model Confussion Matrix yaitu 0.784.

Hasil Apply model algoritma memiliki tingkat akurasi sebesar 33,33 % pada tabel 2 berikut ini :

Tabel 2. Akurasi Decision Tree

accuracy: 33.33%

	true Fun...	true Aller...	true GERD	true Chr...	true Dru...	true Pept...	true AIDS	true Diab...	true Gast...	true Bron...	true Hyp...	true Migr...	true Cen...
pred. Fu...	2	1	1	1	1	1	1	0	1	1	1	1	1
pred. All...	0	0	0	0	0	0	0	0	0	0	0	0	0
pred. GE...	0	0	0	0	0	0	0	0	0	0	0	0	0
pred. Ch...	0	0	0	0	0	0	0	0	0	0	0	0	0
pred. Dr...	0	0	0	0	0	0	0	0	0	0	0	0	0
pred. Pe...	0	0	0	0	0	0	0	0	0	0	0	0	0
pred. AIDS	0	0	0	0	0	0	0	0	0	0	0	0	0
pred. Dia...	0	0	0	0	0	0	0	1	0	0	0	0	0
pred. Ga...	0	0	0	0	0	0	0	0	0	0	0	0	0
pred. Bro...	0	0	0	0	0	0	0	0	0	0	0	0	0
pred. Hy...	0	0	0	0	0	0	0	0	0	0	0	0	0
pred. Mig...	0	0	0	0	0	0	0	0	0	0	0	0	0
pred. Ce...	0	0	0	0	0	0	0	0	0	0	0	0	0
pred. Par...	0	0	0	0	0	0	0	0	0	0	0	0	0



Kesimpulan dan Saran

Pada kesimpulan Dalam penelitian ini telah dilakukan pengujian pada data penyakit menggunakan metode Decision Tree (C4.5). Untuk mengetahui komparasi algoritma yang paling baik dalam menentukan jenis penyakit. Dapat disimpulkan bahwa metode Decision Tree (C4.5) dalam klasifikasinya menghasilkan akurasi 33,33%. Dari tingkat akurasi yang dihasilkan ini, maka perlu dilakukannya kembali pengujian dengan menggunakan algoritma dan model yang lainnya, sehingga dapat memberikan tingkat keakurasian yang lebih baik lagi.

Daftar Pustaka

- [1] M. d. M. N. Arhami, *Data Mining Algoritma dan Implementasi*, Yogyakarta: CV Andi Offset (Penerbit ANDI, 2020).
- [2] d. E. T. L. Kusriani, *Algoritma Data Mining.*, Yogyakarta: Andi, 2009.
- [3] A. Kurniawan, “Sitem Pakar Untuk Mendiagnosa Penyakit Tuna Grahita Dengan Metode Naive Bayes,” in *Jurnal Ilmiah SINUS*, 2013.
- [4] E. Kusriani dan Taufiq, *Algoritma Data Mining*, Jogjakarta: CV. Andi Offset, 2009.
- [5] A. Kristanto, *Perancangan Sistem Informasi dan Aplikasinya*, Yogyakarta: Gava Media, 2018.
- [6] A. Kurniawan, “Sitem Pakar Untuk Mendiagnosa Penyakit Tuna Grahita Dengan Metode Naive Bayes,” in *Jurnal Ilmiah SINUS*, 2013.
- [7] Y. N. B. & S. F. Via, “Sistem Pendukung Keputusan Klasifikasi Tingkat Keganasan Kanker Payudara Dengan Metode Naive Bayes Classifier,” in *Scan*, 2 (X), pp.64-65, 2015.
- [8] M. A. a. N. Fitrianiingsih, “Implementasi Algoritma Naive Bayes Classifier sebagai Sistem Rekomendasi Pembimbing Skripsi,” *InfoTekJar J. Nas. Inform. dan Teknol. Jar*, vol. vol. 5 no. 1, pp. pp. 44–50,, 2020.
- [9] F. L. a. H. Nugroho, “Klasifikasi Daun Herbal Menggunakan Metode Naive Bayes Classifier Dan Knearest Neighbor,” *J. Simantec.*, Vols. vol. 5, no. 1,, p. pp. 9–16, 2015.

