Analisis Pola Penyebaran Penyakit dengan Menggunakan Algoritma C4.5

[1]Syarif Mayron Turnip, [2]Parasian Silitonga
[1]Teknik Informatika Unika St. Thomas S.U; Jln. Setia Budi No.479-F Medan, 061-8210161
[2]Teknik Informatika Unika St. Thomas S.U; Jln. Setia Budi No.479-F Medan, 061-8210161
e-mail : [1]silobakcina@gmail.com; [2]parasianirene@gmail.com

Abstrak

Data Pasien pada rumah sakit tersimpan dalam sebuah system baik secara manual maupun secara komputerisasi hanya tersimpan sebagai database pasien. Dengan semakin bertambahnya volume data, maka diperlukan Algoritma Machine Learning untuk mengklasifikasi dan menganalisa pola penyebaran penyakit. Salah satu metode Machine Learning Klasifikasi adalah C4.5. Proses akhir klasifikasi dilakukan dengan menggunakan perangkat lunak WEKA.

Kata kunci : Algoritma C4.5,Weka, *Data Mining*

Abstract

Data Patients at hospital stored in a system either manually or computerized only stored as a patient database. With the increasing volume of data, it is necessary for Machine Learning Algorithm to classify and analyze patterns of disease spread. One method of Machine Learning Classification is C4.5. The final process of classification is done using WEKA software.

Keywords : C4.5 Algorithm, Weka, Data Mining

1. PENDAHULUAN

Seiring dengan meningkatnya jumlah pasien rumah sakit, diharapkan pihak rumah sakit dapat mengetahui klasifikasi penyakit yang ada di masyarakat serta faktor penyebab penyakit tersebut. Dengan demikian pihak rumah sakit dapat memberikan masukan bagi pemerintah berkaitan dengan usaha pencegahan penyakit dan penyuluhan kesehatan ke daerah-daerah.

Saat ini tumpukan data pasien yang ada di rumah sakit pada umumnya hanya sebatas statistik dan grafik pasien rumah sakit, data penyakit serta biaya perawatan pasien. Tumpukan data yang ada belum menyajikan pola penyebaran penyakit yang ada. Dengan diketahuinya pola penyebaran penyakit maka secara tidak langsung pihak rumah sakit dapat melakukan penyuluhan kesehatan ataupun pencegahan ke daerah-daerah. Selain itu rumah sakit dapat melakukan antisipasi prioritas pelayanan jika diketahui pola penyakit dengan kencederungan tertinggi (Silitonga, Parasian., 2017).

Data mining adalah solusi dalam dunia teknologi untuk mengatasi masalah yang dihadapi rumah sakit dalam memberikan informasi yang tepat dan akurat serta yang efesien kepada yang membutuhkan informasi yang tersebut, dimana informasi tersebut terdapat dalam media penyimpanan data yang memang khusus dipersiapkan oleh rumah sakit tersebut. Terlebih lagi apabila rumah sakit itu melayani pasien dalam jumlah banyak maka sudah pasti memerlukan media penyimpanan data dalam kapasitas yang besar dalam gudang data rumah sakit.

Data mining dalam prosesnya menggunakan teknik statistik, matematika, kecerdasan buatan, dan machine learning untuk mengektraksi dan mengidentifikasi informasi yang

bermanfaat serta pengetahuan yang terkait dari berbagai database yang besar (Turban,dkk. 2005). Dengan adanya masalah tersebut serta ada solusi untuk mengatasi keadaaan seperti itu, maka penulis tertarik untuk melakukan penerapan data mining dengan menggunakan algoritma C4.5 terhadap data pasien.

2. METODOLOGI PENELITIAN

II.1. Data Mining

Data mining merupakan istilah yang sering dikatakan sebagai suatu cara untuk menguraikan serta mencari penemuan berupa pengetahuan didalam suatu *database*. Salah satu kesulitan untuk mendefenisikan *data mining* adalah kenyataan bahwa *data mining* mewarisi banyak aspek dan teknik dari berbagai bidang ilmu yang sudah mapan terlebih dahulu.

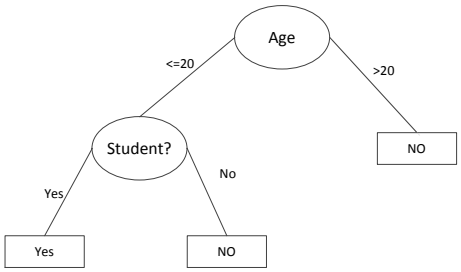
Data mining adalah proses yang menggunakan teknik statistic, matematika, kecerdasan buatan dan *machine learning* untuk mengekstraksi dan mengidentifikasi informasi yang bermanfaat dan pengetahuan yang terkait dari berbagai *database* yang besar (Silitonga, Parasian., Irene Sri Morina., 2018). Menurut Partner Group *data mining* adalah suatu proses menemukan hubungan yang berarti, pola dan kecenderungan dengan memeriksa dalam sekumpulan besar data yang tersimpan dalam penyimpanan dengan menggunakan teknik pengenalan pola seperti teknik statistic dan matematika (Larose, 2005).

Salah satu fungsi *data mining* adalah *clustering. Clustering* merupakan teknik pengelompokkan *record* pada basis data berdasarkan kriteria tertentu. Hasil *clustering* diberikan kepada pengguna akhir untuk memberikan gambaran tentang apa yang terjadi pada basis data (Jiawei Han, Micheline, 2006).

Clustering melakukan pengelompokkan data tanpa berdasarkan kelas data tertentu. Bahkan clustering dapat dipakai untuk memberikan label pada kelas data yang belum diketahui itu. Karena itu clustering sering digolongkan sebagai metode *unsupervised learning* (Ian H, Eibe., 2005). Analisa *clustering* mengidentifikasi kumpulan objek yang memiliki kemiripan satu dengan yang lainnya. Metode *clustering* yang baik dapat menghasilkan *cluster* yang berkualitas untuk memastikan kesamaan data-data yang ada pada sebuah *cluster*.

II.2. Algoritma C4.5

Algoritma C4.5 merupakan *algoritma yang digunakan untuk membentuk pohon keputusan* yang dapat digunakan *untuk memprediksi sebuah* keputusan dengan menerapkan serangkaian aturan keputusan (Larose, Daniel T., 2005). Pohon keputusan merupakan sebuah diagram alir dimana setiap internal node menotasikan atribut yang diuji, setiap cabangnya mempresentasikan kelas-kelas tertentu atau distribusi kelas-kelas. Pohon keputusan menerapkan prediksi dengan menggunakan struktur pohon atau struktur berjenjang (H. Li., X. M. Hu, 2008) . Konsep dari pohon keputusan adalah mengubah data menjadi pohon keputusan dan aturan-aturan keputusan, seperti tersaji pada Gambar 1.



Gambar 1. Ilustrasi Pohon Keputusan

Manfaat utama dari penggunaan pohon keputusan adalah kemampuannya untuk mem-*break down* proses pengambilan keputusan yang kompleks menjadi lebih sederhana sehingga pengambil keputusan akan lebih menginterpretasikan solusi dari permasalahan (Dai, W., Ji, W., 2014). Pohon Keputusan juga berguna untuk mengeksplorasi data, menemukan hubungan tersembunyi antara sejumlah calon variabel input dengan sebuah variabel target. Pohon keputusan memadukan antara eksplorasi data dan pemodelan, sehingga sangat bagus sebagai langkah awal dalam proses pemodelan bahkan ketika dijadikan sebagai model akhir dari beberapa teknik lain.

II.3. WEKA

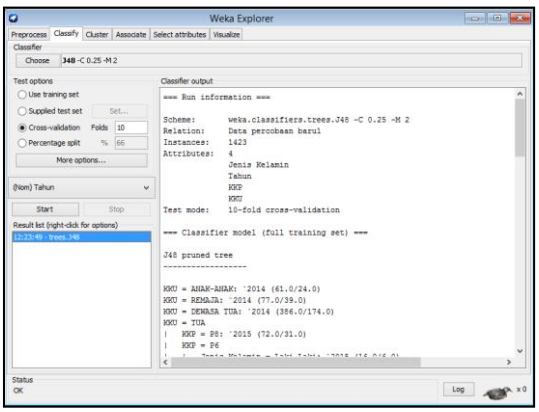
Weka adalah aplikasi data mining open source yang berbasis java. Aplikasi ini dikembangkan tahun 1994 dan pertama kali di sebuah universitas di selandia baru yang bernama universitas Waikato. Aplikasi ini mulai menjadi aplikasi data mining open source yang sangat terkenal pada awal perkembangannya. Hal itu dikarenakan aplikasi weka memliki kelebihan yang tidak dipakai oleh aplikasi data mining lainnya yaitu pada aplikasi weka terdapat algoritma yang terdapat didalam weka dan disertai juga machine learning, kemudian dalam penggunaannya tidak terlalu rumit sehingga tidak menyulitkan penggunanya, dan ditambah dengan kelebihan lainnya bahwa algoritma-algoritma yang terdapat pada aplikasi weka selalu baru dan terupdate, sehingga dengan beberapa kelebihan aplikasi weka tersebut banyak perusahaan dalam dunia bisnis untuk membantu usaha bisnisnya, akademik juga tidak ketinggalan untuk menggunakan aplikasi weka ini, serta instansi dalam bidang kesehatan yaiut rumah sakit juga saat ini menggunakan aplikasi ini.

Aplikasi weka merupakan software yang tediri dari koleksi algoritma machine learning yang dapat digunakan untuk melakukan generalisasi atau formulasi dari sekumpulan data sampling.inti dari kekuatan aplikasi weka terletak pada algoritma yang semakin lengkap dan canggih, namun walaupun begitu canggihnya aplikasi weka tersebut, letak keberhasilan data mining tetap ditentukan oleh manusia itu sendiri sebagai penggunanya/user. Keberhasilan data mining itu berdasarkan pengumpulan data yang berkualitas tinggi, penggunaan model dan algoritma yang tepat.

3. HASIL DAN PEMBAHASAN

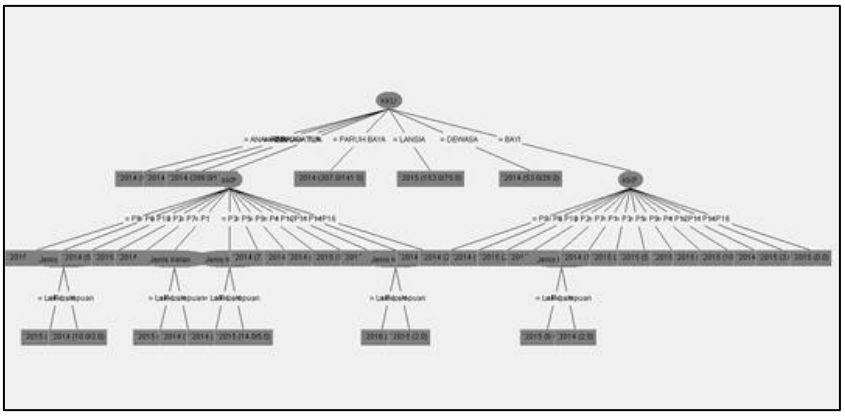
III.1. Hasil

Form hasil classify menggunakan Algoritma C.45 merupakan form untuk menampilkan hasil classify menggunakan Algoritma C.45. Bentuk form hasil classify menggunakan Algoritma C.45 dapat dilihat pada Gambar 2.



Gambar 2. Form Clasify Algoritma C4.5

Form pohon keputusan merupakan Form untuk menampilkan pohon keputusan menggunakan Algoritma C.45. Bentuk form pohon keputusan menggunakan Algoritma C.45 dapat dilihat pada Gambar 3.



Gambar 3. Form Pohon Keputusan C4.5

III.2. Pembahasan

Secara umum algoritma C4.5 untuk membangun pohon keputusan dilakukan dengan mengikuti langkah-langkah :

1. Pilih atribut sebagai akar.
2. Buat cabang untuk masing-masing nilai.
3. Bagi kasus dalam cabang.
4. Ulangi proses untuk masing-masing cabang sampai semua kasus pada cabang memiliki kelas yang sama.

Untuk memilih atribut sebagai akar, didasarkan pada nilai gain tertinggi dari atribut-atribut yang ada. Proses perhitungan gain dan entropy algoritma C4.5 dilakukan dengan menggunakan Persamaan 1 dan Persamaan 2.

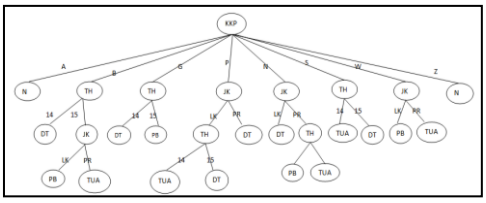
$Entropy\ (S) = \sum_{i-1}^{n} - pi * log\ 2\ pi$ …… ….(1)

Gains (S,A)=*Entropy* (S)- $\sum_{i=1}^{n}$ **Entropy* (Si) ……..(2)

Tabel 1. Perhitungan *Entropy* dan Gain kelas Atribut

		Kasus	Kelompok Usia									
			Bayi	Anak-Anak	Remaja	Dewasa	Dewasa Tua	Paruh Baya	Tua	Lansia	*ENTROPY*	GAIN
A		116	12	6	7	7	27	23	20	14	2.806277	
	2014	58	6	3	1	1	15	14	10	8	2.592631	2.866231
Tahun	2015	58	6	3	6	6	12	9	10	6	2.900013	
	Laki-Laki	72	10	4	4	6	17	15	9	7	2.822701	2.855914
JK	Perempuan	44	2	2	3	1	10	8	11	7	2.648536	

Dari seluruh tabel diatas yang sudah dilakukan perhitungan untuk menentukan cabang kode penyakit, maka dapat ditentukan bahwa atribut Tahun dan Jenis Kelamin memiliki nilai tinggi pada beberapa kode penyakit, sehingga pohon keputusannya dapat terlihat seperti pada Gambar 4.



Gambar 4. Pohon Keputusan Klasifikasi Penyakit Pasien

4. KESIMPULAN

Berdasarkan hasil-hasil analisis yang dilakukan dengan menggunakan algoritma C4.5, maka kesimpulan yang dapat diambil adalah sebagai berikut :

1. Secara keseluruhan dari 1424 record data pasien yang lakukan percobaan, persentase kemunculan penyakit kelompok usia Bayi sampai Remaja dari jumlah data sangat kecil.
2. Kode Penyakit yang didominan pada tahun 2014 dan 2015 adalah Kode G,P,J dengan kelompok usia Dewasa Tua Sampai Paruh Baya
3. Pada Tahun 2014 penderita penyakit dengan Kelompok Usia Dewasa Tua dan Paruh Baya adalah Laki-Laki

5. SARAN

Saran yang diusulkan untuk penelitian selanjutnya dapat dikembangkan dengan menambah atau menggunkan beberapa teknik untuk meningkatkan akurasi sebuah algoritma dalam proses klasifikasi, seperti teknik *bagging* dan *Boosting.*

DAFTAR PUSTAKA

[1] Dai, W. and Ji, W., 2014. A mapreduce implementation of C4. 5 decision tree algorithm. International Journal of Database Theory and Application, 7(1), pp.49-60.

[2] H. Li and X. M. Hu, "Analysis and Comparison between ID3 Algorithm and C4. 5 Algorithm in Decision Tree", Water Resources and Power, vol. 26, no. 2, (2008), pp. 129-132.

[3] Larose, Daniel T. 2005. Discovering Knowledge ini Data: An Introduction to Data Mining. Wiley.

[4] Saputra, Rizal Amegia. (2014), Komparasi Algoritma Klasifikasi Data Mining Untuk Memprediksi Penyakit TuberCulosis (TBC) : Studi Kasus Puskesmas Karawang SukaBumi, Seminar Nasional Inovasi dan Tren (SNIT)

[5] Syahril, Muhammad, (2011), Konversi Data Training Tentang Penyakit HiperTensi Menjadi Bentuk Pohon Keputusan Dengan Teknik Klasifikasi Menggunakan Tools RapidMiner 4.1. Jurnal SAINTIKOM, Vol.10 / No.2

[6] Silitonga, Parasian, 2017, *Clustering of Patient Disease Data by Using K-Means Clustering, International Journal of Computer Science and Information Security (IJCSIS), Vol. 15, No. 7, July 2017- ISSN 1947-5500*

[7] Silitonga, Parasian., Irene, Sri, Morina., 2018., "Implementation of K-Means Clustering on Patient Data Of National Social and Healthcare Security by Using Java"., International Journal of Computer Science Engineering (IJCSE), ISSN : 2319-7323 Vol. 7 No.1 Jan-Feb 2018.