Penerapan Algoritma C4.5 Untuk Menentukan Kesesuaian Lensa Kontak dengan Mata Pasien

Ketut Wisnu Antara[1], Gede Thadeo Angga Kusuma[2]
Jurusan Pendidikan Teknik Informatika
Universitas Pendidikan Ganesha
Singaraja, Bali

***Abstrak* —banyak kasus dalam bidang kesehatan yang dapat diselesaikan dengan data mining. Salah satunya adalah masalah kesesauian lensa kontak dengan mata pasien. Data mining sendiri merupakan suatu proses mendapatkan informasi dari sebuah data. Dalam pengambilan informasi kesesuaian lensa konta dengan mata akan digunakan algoritma C4.5 dalam perhitungannya. Data training di dapatkan dari UCI Machine Learning Repository "http:// archive.ics.uci.edu/ ml/ machine-learning-databases/ lenses/lenses.data, data ini merupakan kumpulan data set dari tahun 1990. Data set ini sudah bersih jadi tidak perlu dilakukan proses cleaning lagi dan masuk ke proses perhitungan. Hasil perhitungan dengan menggunakan algoritma C4.5 menunjukkan keakuratan yang tinggi terhadap evaluasi training set dengan nilai sebesar 91,6667 % dan 100% dengan supplied test set.**

Kata Kunci – Klasifikasi, Data Mining, AlgoritmaC4.5, Contact Lense Data Set

I. PENDAHULUAN

Perkembangan teknologi setiap tahunnya semakin pesat, segala bidang tak bisa lepas dari komputer dan teknologinya. Teknolgi komputer menjangkau segala bidang aktivitas kehidupan manusia, mulai dari kesehatan, bisnis, pendidikan, organisasi dan masyarakat. Dalam bidang kesehatan penggunaan teknologi juga sangat dibutuhkan untuk mendaptkan informasi-informasi penting terkait kesehatan seseorang. Salah satu kasus dibidang kesehatan yaitu bagaimana mengetahui kesesuaian lensa kontak dengan mata pasien berdasarkan informasi-informasi yang ada.

Kesesuain lensa kontak terhadap mata seseorang sangatlah penting selain membuat si pengguna lensa kontak nyaman juga menghindari resiko-resiko yang dapat timpul dari kontak lensa itu sendiri, seperti lecet pada mata, infeksi mata, serta beberapa permasalahan yang terjadi akibat kontak lensa. Namun lensa konta merupakan salah satu alat medis yang paling aman jika digunakan dengan bertanggung jawab dan sesuai dengan kebutuhan mata pasien.

Guna mendapatkan informasi dari kasus tersebut dapat diselesaikan dengan data mining. Data mining merupakan proses yang menggunakan statistik, matematika, kecerdasan buatan, dan machine learning untuk mengekstraksi dan mengidentifikasi informasi yang bermanfaat dan pengetahuan yang terkait dari berbagai database besar [1]. Data Mining sendiri memiliki empat teknik yaitu Estimasi/Prediksi, Klasifikasi, Klastering dan Asosiasi. Teknik klasifikasi terdiri beberapa metode, dan decision tree adalah bagian dari metode klasifikasi. Kemudian metode dicision tree memiliki algoritma, algoritman C4.5 adalah salah satu dari algoritma yang memiliki decision tree.

Data testing yang digunakan oleh peneliti adalah data fitting contact lenses tahun 1990. Data set ini diperoleh dari "http://archive.ics.uci.edu/ml /machine-learning-databases/lenses/ lenses.data " yang dapat digunakan sebagai data testing untuk berbagai penelitian.

Berdasarkan data testing serta pengetahuan yang ingin di dapatkan, pendekatan data mining dengan penerapan metode algoritma C4.5 akan digunakan untuk menentukan kesesuaian lensa kontak dengan mata pasisen.

II. METODOLOGI PENELITIAN

A. Pengumpulan Data

Data sekunder adalah data yang diperoleh secara tidak langsung bersumber dari dokumentasi, literatur, buku, jurnal dan informasilainnya yang ada hubungannya dengan masalah yang diteliti. Data sekunder pada penelitian ini adalah : buku-buku, jurnal tentang algoritma C4.5 dan data mining. Sedangkan Data primer adalah data set yang diperoleh dari "http://archive.ics.uci.edu/ml/ machine-learning-databases/lenses/lenses.data ".Data primer dalam penelitian ini adalah data hasil fitting lensa kontak pada tahun 1990.

B. Pengolahan Awal Data

Data yang akan digunakan seagai data testing adalah data fitting lensa kontak. Data set ini bersumber dari tanggal 1 August 1990, data set ini berjumlah total 24 record. Data set ini memiliki 4 atribut dan 3 class atau label yaitu age, spectacle prescription, astigmatic, dan tear production rate. 3 class atau label pada data set meliputi fitted with hard contact lenses, fitted with soft contact lenses, dan not be fitted with contact lenses. Data set ini sudah bersih dari noise jadi tidak perlu dilakukan proses cleaning.

C. Model yang Diusulkan

Tan mendefinisikan data mining sebagai proses untuk mendapatkan informasi yang berguna dari gudang basis data yang besar [2]. Data mining juga dapat diartikan sebagai pengekstrakan informasi baru yang diambil dari bongkahan data besar yang membantu dalam pengambilan keputusan. Istilah data mining kadang disebut juga knowledge discovery.

Salah satu teknik yang dibuat dalam data mining adalah adalah bagaimanana menelusuri data yang ada untuk membangun sebuah model, kemudian menggunakan model tersebut agar dapat mengenali pola data yang lain yang tidak berada dalam basis data yang tersimpan. Kebutuhan untuk prediksi juga dapa memanfaatkan teknik ini [3].

Dalam data mining, pengelompokan data juga bisa dilakukan. Tujuannya adalah agar kita dapat mengetahui pola universal data-data yang ada. Anomali data transaksi juga perlu dideteksi untuk dapat mengetahui tindak lanjut berikutnya yang dapat diambil. Semua hal tersebut bertujuan mendukung kegiatan operasional perusahaan sehingga tujuan akhir perusahaan diharapkan dapat tercapai. Data mining merupakan bagian dari proses Knowledge Discovery from Data (KDD). Dibawah ini digambarkan skema dari proses KDD.

Model yang diusulkan untuk memprediksi kesesuaian lensa kontak dengan mata pasien adalah algoritma C4.5. Algoritma C4.5 merupakan salah satu algoritma yang digunakan untuk melakukan klasifikasi atau segmentasi atau pengelompokan dan bersifat prediktif. Dasar algoritma C4.5 adalah pembentukan pohon keputusan (*decision tree*). Cabang-cabang pohon keputusan merupakan pertanyaan klasifikasi dan daun-daunnya merupakan kelas-kelas atau segmen-segmennya.

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

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

Keterangan:

- S : himpunan
- A : atribut
- n : jumlah partisi atribut A

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

Menghitung Nilai Entropy

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

Keterangan:
- S : himpunan kasus
- A : fitur
- n : jumlah partisi S
- pi : proporsi dari Si terhadap S

III. HASIL DAN PEMBAHASAN

Data set yang digunakan adalah *Contact Lenses* yang bersumber dari UCI Machine Learning Repository. *Data set LENSES* terdiri dari 24 *instance* dengan atribut yang berjumlah 5. Pertama adalah atribut *Recommended Lenses* terdiri dari 3 kelas yaitu : *Hard, Soft,* dan *None*. Kedua adalah atribut *Age* yang terdiri dari 3 kelas yaitu : *Young, Pre – Presbyopic, dan Presbyopic*. Ketiga adalah atribut *Spectacle Prescription* yang terdiri dari 2 kelas yaitu : *Myope* dan *Hypermetrope*. Keempat adalah atribut *Atigmatic* yang terdiri dari 2 kelas yaitu : *No* dan *Yes*, dan yang terakhir adalah atribut *Tear Production* yang trediri dari 2 kelas yaitu : *Reduced* dan *Normal*.

A. Algoritma Decision Tree J48

J48 adalah salah satu algoritma yang sama persis dengan C45 namun terdapat dalam software WEKA. ALgoritma J48 membangun sebuah pohon keputusan berdasarkan pada seperangkat *input data* yang berlabel. Pohon keputusan adalah model prediksi menggunakan struktur pohon atau struktur berhirarki. Konsep dari pohon keputusan adalah mengubah data menjadi pohon keputusan dan aturan-aturan keputusan. Berikut ini tahapan algoritme J48 :

1) Menyiapkan *data training*
2) Menentukan akar dari pohon.
3) Menghitung nilai Gain melalui Persamaan (1).

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

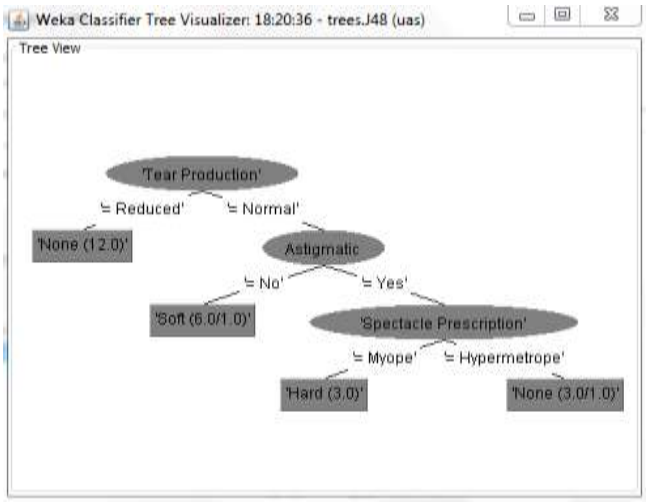
4) Ulangi langkah ke-2 hingga semua tupel terpartisi dengan menggunakan Persamaan (2).

$$\text{Gain(S,A)} = S - \sum_{i=1}^{n} \frac{|S_i|}{|S|} * S_i \quad (2)$$

5) Proses partisi pohon keputusan akan berhenti saat semua tupel dalam node N mendapat kelas yang sama dan atau tidak ada atribut di dalam tupel yang dipartisi lagi dan atau tidak ada tupel di dalam cabang yang kosong.

B. Pengujian dan Hasil Klasifikasi dengan Algoritma J48

Dalam pengujian data set yang telah disediakan sebelumnya menggunakan sebuah software yang bernama WEKA. Dari hasil pengujian tersebut menghasilkan sebanyak 4 *leaf node* dan 7 *tree.* Berikut gambar dari hasil pengujian menggunakan algoritma J48.



Gambar 1. Decision Tree Data Set Contact Lense

Pada software WEKA terdapat berbagai macam test option diantaranya yaitu : 1) *Use Trainig Set*, 2) *Supplied Test Set*, 3) *Cross Validation*, 4) *Percentage Split*.

1) Use Training Set

Pengolahan klasifikasi *data set Contact Lenses* dengan pilihan '*use training set*' dengan menggunakan *training data* yang

berjumlah 24 *instance* pada WEKA menghasilkan data yang dapat dilihat pada Tabel 1.

Tabel 1. Hasil Evaluasi *Training Set*

No	*Spesifikasi Pengukuran*	*Nilai*
1	Correctly Classified Instances	22 atau 91,6667 %
2	Incorrectly Classified Instances	2 atau 8,3333%
3	Kappa statistic	0,8447 %
4	Mean absolute error	0,0833 %
5	Root mean squared error	0,2041 %
6	Relative absolute error	22,6257 %
7	Root relative squared error	48,1223 %
8	Total Number of Instances	24

2) Supplied Test Set

Classifier yang telah terbentuk pada tahap *training set* selanjutnya diuji dengan menggunakan data *test data* dengan 50% data dari data set, yang kurang lebih sebanyak 15 Instance menghasilkan data yang dapat dilihat pada Tabel 2.

Tabel 2. Hasil Evaluasi *Supplied Test Set*

No	*Spesifikasi Pengukuran*	*Nilai*
1	Correctly Classified Instances	15 atau 100 %
2	Incorrectly Classified Instances	0
3	Kappa statistic	1
4	Mean absolute error	0,037 %
5	Root mean squared error	0,093 %
6	Relative absolute error	9,8684 %
7	Root relative squared error	21,5018 %
8	Total Number of Instances	15

3) Cross Validation

Pilihan tes untuk *cross validation* dengan jumlah *folds* sebanyak 10 dengan menggunakan 24 *instance* menghasilkan sebanyak 4 *leaf node* dan 7 *tree* dan data lengkapnya dapat dilihat pada Tabel 3.

Tabel 3. Hasil Evaluasi *Cross Validation*

No	*Spesifikasi Pengukuran*	*Nilai*
1	Correctly Classified Instances	20 atau 83,3333 %
2	Incorrectly Classified Instances	4 atau 16,6667%
3	Kappa statistic	0,71 %
4	Mean absolute error	0,15 %
5	Root mean squared error	0,3249 %
6	Relative absolute error	39,7059 %
7	Root relative squared error	74,3898 %
8	Total Number of Instances	24

4) Percentage Split

Pengolahan klasifikasi data dengan menggunakan test *percentage split* sebesar 50% memiliki hasil yang dapat dilihat pada table 4.

Tabel 4. Hasil Evaluasi *Percentage Split*

No	*Spesifikasi Pengukuran*	*Nilai*
1	Correctly Classified Instances	5 atau 41,6667 %
2	Incorrectly Classified Instances	7 atau 58,3333%
3	Kappa statistic	0,134%

4	Mean absolute error	0,3889%
5	Root mean squared error	0,6236%
6	Relative absolute error	94,5946%
7	Root relative squared error	124,4457 %
8	Total Number of Instances	12

C. Rekomendasi

Hasil yang diperoleh dari keseluruhan tes yang dilakukan baik dengan '*use training set*', '*supplied test set*', *cross validation* dan *percentage split* dari Algoritma J48 dapat dilihat pada Tabel 5.

Tabel 5. Hasil Klasifikasi Algoritma J48

No	*Metode*	*Spesifikasi Pengukuran*	*Nilai*
1	*Use training set*	Incorrectly Classified	2
		Correctly Classified	22
		Akurasi (%)	91,6667%
		Error Mean (%)	0,0833%
2	*Supplied test set*	Incorrectly Classified	0
		Correctly Classified	15
		Akurasi (%)	100%
		Error Mean (%)	-
3	*Cross Validation*	Incorrectly Classified	4
		Correctly Classified	20
		Akurasi (%)	83,3333%
		Error Mean (%)	0,15%
4	*Percentage Split*	Incorrectly Classified	7
		Correctly Classified	5
		Akurasi (%)	41,6667%
		Error Mean (%)	0,3889%

Hasil dari percobaan klasifikasi pada *data set Contact Lenses* menunjukkan bahwa untuk evaluasi dengan menggunakan *supplied test set* memiliki akurasi yang lebih tinggi dibandingkan dengan menggunakan training lainnya, yaitu tingkat akurasi *supplied test set* adalah 100% dengan jumlah data yang benar adalah 15 dan data yang salah adalah 0 dari 15 jumlah data. Sedangkan untuk evaluasi yang menggunakan semua data set sebanyak 24 data di dapatkan hasil evaluasi yang terbaik dengan menggunakan *option test Use Training Test* dibandingkan dengan menggunakan *Cross Validation* dengan tingkat akurasi adalah 91,6667% dengan jumlah data yang benar sebanyak 22 dan data yang salah sebanyak 2 dari 24 data.

IV. KESIMPULAN

Melalui percobaan yang telah dilakukan terhadap *data set Contact Lenses* dengan algoritma klasifikasi yang terdapat pada WEKA yaitu algoritma J48 atau yang pada umumnya disebut algoritma C45 dapat disimpulkan bahwa kinerja dari algoritma J48 sangat baik. Dari analisis data set menggunakan algoritma J48 ini didapatkan pengetahuan ataupun aturan yaitu

1 Jika tear production (Reduced) Maka penggunaan Soft Lenses (None)

2 Jika tear production (Normal) dan Astigmatic (Yes) dan Spectacle Prescription (Hypermetrope) Maka penggunaan Soft Lenses (None)
3 Jika tear production (Normal) dan Astigmatic (Yes) dan Spectacle Prescription (Hypermetrope) Maka penggunaan Soft Lenses (Hard)
4 Jika tear production (Normal) dan Astigmatic (No) dan Age (Young) Maka penggunaan Soft Lenses (Soft)
5 Jika tear production (Normal) dan Astigmatic (No) dan Age (presbyopic) Maka penggunaan Soft Lenses (None)
6 Jika tear production (Normal) dan Astigmatic (No) dan Age (pre-presbyopic) Maka penggunaan Soft Lenses (Soft)

Dengan Algoritma J48 hasil yang di dapatkan dihasilkan memiliki akurasi yang sangat tinggi dengan berbagai macam pilihan evaluasi option test yang telah disediakan pada software WEKA. Selain itu jika dilihat dari lama waktu komputasi, untuk algoritma J48 rata-rata menghabiskan waktu sekitar 0.01 – 0.04 detik.

DAFTAR PUSTAKA

[1] Turban Efraim, Aronson Jay E, and Liang, *Decision Support Systems and Intelligent Systems*, 7th ed.: Prentice Hall, Upper Saddle River, NJ, 2005.

[2] Han & Kamber, (2006). *Data Mining Concept and Thenique*. Morgan Kaufffman.San Fransisco.

[3] Prasetyo, Eko, (2012). *Data Mining*, Andi Yogyakarta, 356 Halaman.

[4] Larose Daniel T, *Discovering knowledge in data: An Introduction to Data Mining*.: Wiley Interscience, 2005.