

Analisis Penerapan Algoritma ID3 dalam Mendiagnosis Kesuburan Pria

Khaerul Ma'mur

Teknik Informatika, Fakultas Teknik, Universitas Pamulang, Tangerang Selatan, Banten, Indonesia
Email: kemunk.alfatih@gmail.com

Abstract

Men's fertility disorder is one of difficult thing to be diagnosed, if fertility has a problem, the fertilization process will be disrupted. That's all to be the most predominant cause of discordant household. Numbers of data on men's fertility are obtained by detailed enough. Secondary data from this dataset, made a reference in making decision tree using the ID3 algorithm in software Rapid Miner. Iterative Dichotomizer (ID3) is used because it is one of the commonly used classification algorithm in Data Mining to implement decisions of some attributes that have been determined in accordance with the data customized with their needs. ID3 algorithm generally calculates the Entropy value of each attribute and the value of the Information Gain of each attribute that are able to form a decision tree that is expected to assist in decision making. The results of the tests and the method can be used to diagnose men's fertility appropriately with a fairly high degree of accuracy.

Keywords: Decision Tree, Data Mining, Iterative Dichotomizer 3, Men's Fertility

1. Pendahuluan

Masalah ketidsuburan pada seseorang menjadi salah satu penyebab renggangnya suatu hubungan dalam rumah tangga. Terlebih apabila ada rumah tangga yang belum dikaruniai keturunan. Seperti yang diketahui pada umumnya bahwa penilaian di tengah-tengah masyarakat, jika ada seorang wanita yang susah hamil maka cenderung hanya wanitanya saja yang dituntut untuk mencari tahu penyebabnya.

Gaya hidup yang kurang sehat menjadi salah satu faktor penyebabnya. Adanya pola serta gaya hidup di masyarakat yang kurang sehat, secara tak langsung, punya peran memengaruhi kualitas kesuburan sperma pada pria yang dapat mengakibatkan sulitnya pasangan suami istri memiliki anak (Irvine DS, 1998).

Untuk pemeriksaan, pria yang mempunyai masalah dengan kesuburan dapat berkonsultasi kepada seorang dokter spesialis, khususnya spesialis urologi. Dokter akan melakukan identifikasi masalah atau mendiagnosis sekaligus mengambil tindakan yang tepat guna membantu mengatasi permasalahan tersebut.

Mengidentifikasi kesuburan pria dapat dilakukan salah satunya menggunakan *data mining*. Data mining dan machine learning dapat digunakan untuk memprediksi berdasarkan data-data masa lalu (Saifudin, 2018). Sebab metode, model, teknik, atau algoritma dalam *data mining* cukup banyak variasinya. Model yang digunakan

untuk membantu pemrosesan datanya menggunakan pohon keputusan (*decision tree*).

Pohon keputusan membangun model klasifikasi atau regresi dalam bentuk struktur pohon. Pembangunan model tersebut digunakan untuk memecah dataset menjadi himpunan bagian yang lebih kecil sementara pada saat yang sama pohon keputusan terkait, dikembangkan secara bertahap. *Node* daun mewakili klasifikasi atau keputusan. *Node* keputusan teratas dalam pohon yang sesuai dengan prediktor terbaik disebut simpul akar. Pohon keputusan dapat menangani data kategorikal dan numerik. Oleh karena itu digunakan model berupa pohon keputusan yang dapat dijadikan rujukan alternatif dengan menggunakan bantuan algoritma ID3.

Algoritma ID3 (*Iterative Dichotomiser Tree*) digunakan pada penelitian ini karena algoritma ID3 menentukan nilai bobot dari tiap atribut, yang kemudian dilanjutkan dengan seleksi alternatif. Selain itu algoritma ID3 juga dapat dijadikan pilihan kepada pria yang akan memeriksakan kesuburannya ketika enggan ke rumah sakit.

2. Landasan Teori Kesuburan

Dalam dunia kesehatan, permasalahan kesuburan terjadi akibat adanya gangguan sistem reproduksi pada wanita dan adanya penurunan kualitas serta kuantitas sperma pada pria. Terjadinya masalah kesuburan atau infertilitas dapat disimpulkan apabila sepasang suami istri

dalam tenggat waktu dua tahun atau lebih belum dikaruniai kehamilan sedangkan setiap kali 'berhubungan' tidak pernah menggunakan alat pencegah kehamilan.

Ada banyak faktor yang dapat memengaruhi kesuburan sperma pada pria, salah satunya yaitu kebiasaan gaya hidup dan lingkungan (David dkk, 2012). Berikut ini beberapa penyebab terjadinya permasalahan yang mengganggu kesuburan pada seorang pria antara lain:

- 1) Adanya kelainan genetik
- 2) Gangguan hormonal
- 3) Kelainan skrotum
- 4) Sumbatan pada saluran sperma
- 5) Disfungsi ereksi
- 6) Banyak konsumsi rokok
- 7) Sering minum alkohol
- 8) Paparan radiasi
- 9) Konsumsi obat-obatan

Data Mining

Data mining (penambangan data) adalah proses berbantuan komputer untuk mencari dan menganalisis set data yang sangat besar dan mencoba memahami makna data. Penambangan data adalah metode analisis yang memungkinkan komputer mengekstrak informasi yang dapat digunakan dari kumpulan data mentah yang sangat besar.

Data mining memproses informasi terstruktur melalui aplikasi kecerdasan buatan, jaringan saraf, dan alat statistik canggih untuk mendeteksi pola dan merangkum data ke dalam format yang dapat dipahami.

Mulanya *data mining* menyertakan proses otomatisasi sebagai fokus utamanya. *Data mining* dijadikan salah satu proses penjabaran dan analisis dengan cara otomatis ataupun semi otomatis kepada data dengan jumlah yang besar untuk menentukan pola atau *rule* yang akan digunakan (Berry dan Linoff, 2004).

Data mining juga merupakan proses menganalisis sejumlah besar data untuk menemukan pola dan informasi lainnya. Ini biasanya dilakukan pada database, yang menyimpan data dalam format terstruktur. Dengan 'menambang' sejumlah besar data, informasi tersembunyi dapat ditemukan dan digunakan untuk tujuan lain.

Teknik Data Mining

Sebagai acuan dalam penelitian ini, digunakan teknik data mining. Macam teknik *data mining* yang umumnya diterapkan seperti di bawah ini (Dunham H. Margareth, 2002).

1. *Clustering*

Analisis *Clustering* merupakan bagian dari proses mengidentifikasi data set yang mirip antara satu sama lain, untuk memahami perbedaan serta kesamaan dalam data. *Cluster* memiliki kesamaan sifat-sifat tertentu yang dapat digunakan untuk meningkatkan algoritma penargetan. Misalnya, kelompok pelanggan dengan perilaku pembelian yang sama dapat ditargetkan dengan produk dan layanan serupa untuk meningkatkan tingkat percakapan. Tujuan penggunaan teknik *clustering* yaitu guna dibuat kelompok data atau objek ke dalam kumpulan *cluster* sehingga tiap *cluster* akan berisi data-data yang cukup mirip (Larose, D. T., 2005).

2. *Association Rule (Market Basket Analysis)*

Kegunaan teknik ini memungkinkan penemuan hubungan yang menarik (interdependensi) antara berbagai variabel dalam database besar. *Association rule learning* mengungkap pola yang belum terlihat pada data yang bisa diterapkan untuk mengidentifikasi variabel dalam data dan kejadian bersama dari berbagai variabel yang muncul dengan frekuensi terbesar.

3. *Classification*

Analisis klasifikasi yaitu proses sistematis guna memperoleh informasi dan sekelompok model yang relevan dari data set. Analisis klasifikasi membantu mengidentifikasi kategori atau kelas-kelas yang dimiliki data. Analisis klasifikasi terkait erat dengan analisis cluster karena klasifikasi dapat digunakan untuk data cluster.

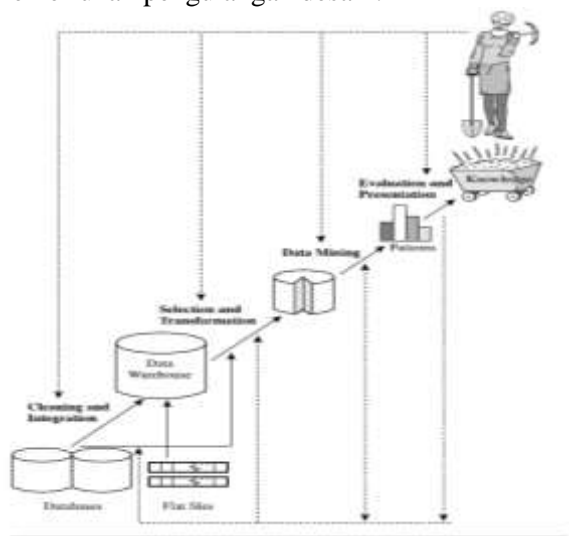
4. *Regression (Predictive)*

Analisis regresi merupakan teknik dalam mendefinisikan ketergantungan antar variabel. Analisis regresi dapat menunjukkan bahwa satu variabel tergantung pada variabel lain tetapi tidak sebaliknya.

Tahapan Data Mining

Sebuah keputusan yang tepat dalam penggunaan metodologi sistematis tak hanya saat melakukan analisis saja namun saat mempersiapkan data dan saat melakukan interpretasi dari hasil penerapan *data mining* yang mempunyai skala besar.

Secara umum proses *data mining* begitu interaktif karena terkadang hasil yang didapat pada mulanya kurang sesuai terhadap harapan penganalisanya, sehingga pada prosesnya memerlukan pengulangan desain.



Gambar 1. *Data mining* sebagai bagian dari proses *knowledge discovery*

Tahapan-tahapan yang bersifat interaktif dalam proses *data mining* di mana pengguna terlibat secara langsung atau dengan perantara *knowledge base* (Tan, dkk, 2004) antara lain:

1. Pembersihan data. Berfungsi memilah atau membersihkan data yang kerap berubah-ubah serta mempunyai noise.
2. Intergrasi data. Diperlukan untuk *data mining* yang bukan saja bersumber dari satu database namun dapat pula dari banyak database atau *textfile*. Hasil dari pengintegrasian data akan ditampilkan berupa data *warehouse*, sebab data dikonsolidasikan melalui struktur khusus yang efektif dan efisien.
3. Transformasi data. Guna menentukan seberapa berkualitasnya hasil dari *data mining* yang diperoleh, agar data dapat diubah ke dalam bentuk yang diinginkan.
4. Aplikasi teknik *data mining*. Salah satu bagian dari tahapan proses *data mining*. Digunakan untuk memproses atau mengolah data yang besar.
5. Evaluasi pola yang diperoleh. Pada tahap yang ke 5, hasil yang diperoleh dari penggunaan *data mining* berbentuk pola-pola unik ataupun model prediksi dilihat kembali untuk memberikan penilaian mengenai hipotesa yang telah dirumuskan sudah sesuai atau belum sesuai rencana.

6. Presentasi pengetahuan. Tahap ini mempresentasi pola yang diperoleh untuk memperoleh tindakan tahap akhir dari pemrosesan *data mining* yaitu seperti apa cara memformulasikan keputusan dari hasil analisa yang telah didapat.

Pohon Keputusan

Pohon keputusan (*decision tree*) merupakan struktur *flowchart* yang memiliki *tree* (pohon), yang mana tiap simpul internal menandakan suatu tes atribut, setiap cabang merepresentasikan hasil tes, dan simpul daunnya merepresentasikan kelas atau distribusi kelas (Santosa Budi, 2007). Pohon keputusan juga bisa diartikan sebagai diagram yang menyerupai diagram alur yang menunjukkan berbagai hasil dari serangkaian keputusan. Pohon keputusan membuat model prediksi berdasarkan serangkaian percabangan yang menggunakan fakta-fakta spesifik untuk membuat kesimpulan yang lebih umum.

Pohon keputusan memiliki tiga bagian utama diantaranya simpul akar, simpul daun, dan cabang. Simpul akar adalah titik awal dari pohon, dan simpul akar dan daun berisi pertanyaan atau kriteria yang harus dijawab. Cabang adalah panah yang menghubungkan node, menunjukkan aliran dari pertanyaan ke jawaban. Setiap node biasanya memiliki dua atau lebih node yang memanjang darinya. Misalnya, jika pertanyaan di simpul pertama membutuhkan jawaban "ya" atau "tidak", akan ada satu simpul daun untuk respons "ya", dan simpul lain untuk "tidak".

Algoritma ID3

ID3 merupakan pendahulu dari algoritma C4.5. Secara sederhana, ID3 membangun pohon keputusan dari serangkaian contoh yang tetap. Pohon keputusan yang dihasilkan digunakan untuk mengklasifikasikan sampel untuk dijadikan pedoman di masa yang akan datang. Contoh-contoh dari data set yang diberikan memiliki beberapa atribut dan setiap contoh milik kelas (seperti ya atau tidak). Node daun pohon keputusan berisi nama kelas sedangkan node bukan daun adalah node keputusan. Node keputusan adalah atribut dengan setiap cabang (ke pohon keputusan lain) menjadi nilai yang mungkin dari tiap atribut. ID3 menggunakan heuristik pemilihan fitur untuk membantunya memutuskan atribut mana yang masuk ke simpul keputusan. Heuristik yang diperlukan dapat dipilih oleh parameter kriteria.

Algoritma ID3 berupaya menghasilkan pohon keputusan dengan urutan *top-down* (atas ke

bawah), melalui pertanyaan “mana atribut yang harus diperiksa terlebih dahulu dan ditempatkan pada *root*?”. Untuk menjawabnya dengan cara mengevaluasi seluruh atribut yang tersedia melalui penggunaan sebuah ukuran statistik (seringnya menggunakan *information gain*) untuk melihat sejauh mana tingkat keefektifan suatu atribut ketika mengelompokkan sekumpulan sampel data (David, Mcg, 2004).

Rapid Miner

Rapid miner dibuat oleh Dr. Markus Hofmann dari Institute of Technology Blanchardstown dan Ralf Klinkenberg dari rapid-i.com. Menggunakan GUI (*Graphical User Interface*) yang apik sangat mempermudah *user* saat mengoperasikan rapid miner. Aplikasi ini dibuat dengan bahasa pemrograman Java di bawah lisensi GNU *Public License*, sehingga berlisensi *open source*. Untuk menjalankan perangkat lunak ini dapat berjalan di semua *operating system*.

Untuk mengoperasikan aplikasi Rapid Miner, tidak diperlukan kemampuan khusus, sebab fiturnya sudah cukup lengkap. Rapid Miner ditujukan khusus untuk pengelolaan data dalam jumlah yang banyak. Model Bayesian, Neural Network, Tree Induction, dan lain sebagainya disediakan dalam aplikasi guna membentuk pemodelan. Metode yang tersedia pada rapid miner juga cukup banyak, diantaranya klustering, asosiasi, klasifikasi, dan lain-lain.

3. Hasil dan Pembahasan Pengolahan data

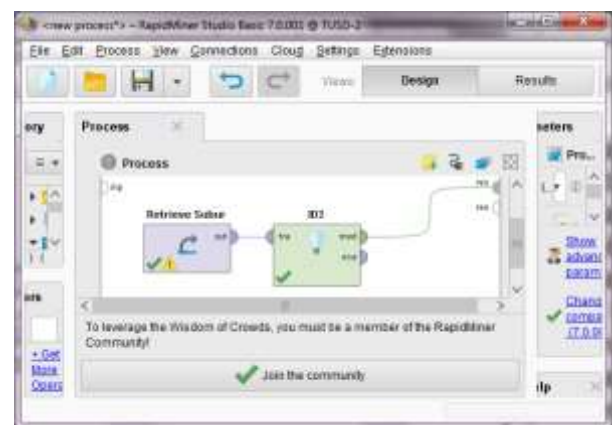
Data yang dijadikan sebagai bahan penelitian bukan merupakan data primer, di mana data tersebut diambil dari WHO (*World Health Organization*) pada tahun 2010 yang dipublikasikan oleh UCI *dataset machine learning*. Dengan rincian 100 relawan memberikan sampel sperma (air mani) yang dianalisis menurut kriteria WHO. Konsentrasi kesuburan terkait dengan data sosio-demografis, faktor lingkungan, status kesehatan, dan kebiasaan hidup.

Data mentah yang telah didapat kemudian diolah untuk diubah ke dalam data kategorial karena ID3 tidak dapat membaca data dalam bentuk numerik. Berikut disajikan dalam bentuk tabel.

Tabel 1 Kategori Atribut

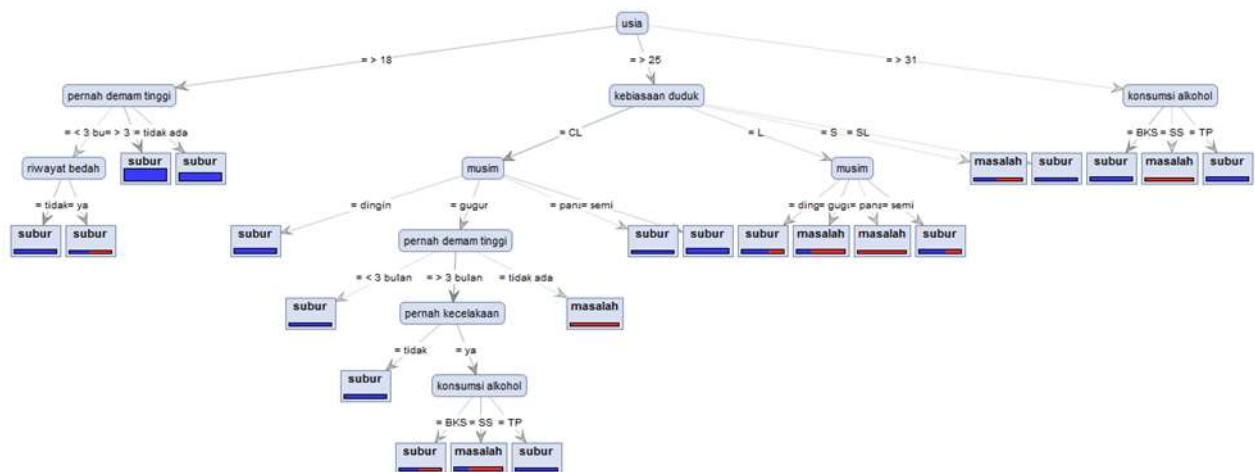
Atribut	Nilai Angka	Kategori
Musim	-1	Dingin
	-0,33	Semi
	0,33	Panas
	1	Gugur
Usia	18-25	>18
	26-30	>25
	31-36	>31
Penyakit masa kecil	0	Ya
	1	Tidak
Kecelakaan atau Trauma serius	0	Ya
	1	Tidak
Riwayat Operasi	0	Ya
	1	Tidak
Demam tahun lalu	-1	>3 bulan
	0	<3bulan
	1	Tidak
Konsumsi Alkohol	0,2	Beberapa kali sehari
	0,4	Tiap Hari
	0,6	Beberapa kali seminggu
	0,8	Sesekali
	1	Tidak Pernah
Merokok	-1	Sesekali
	0	Tiap hari
	1	Tidak pernah
Duduk (dalam sehari/jam)	1-4	Sebentar
	5-8	Cukup lama
	9-12	Normal
	13-16	Sangat lama

Data yang telah tersusun seperti di atas, kemudian dimasukkan ke dalam aplikasi Rapid Miner untuk membuat pohon keputusan yang sistematis. Berikut penggambaran untuk menjadikan pohon keputusannya.



Gambar 2 Relasi antara data training dan model

Ikuti alur seperti pada gambar di atas. Setelah di eksekusi, dengan mengklik tanda *Run*, maka hasilnya terlihat seperti pada gambar yang ditampilkan di bawah ini.



Gambar 3. Pohon keputusan hasil ekstraksi dari tools Rapid Miner

Dan didapat pula *rule* (aturan-aturan) yang didapatkan dari pohon keputusan tersebut. Berikut aturan-aturanya.

- usia = > 18
 - | pernah demam tinggi = < 3 bulan
 - | | riwayat bedah = tidak: subur {subur=3, masalah=0}
 - | | riwayat bedah = ya: subur {subur=1, masalah=1}
 - | pernah demam tinggi = > 3 bulan: subur {subur=27, masalah=0}
 - | pernah demam tinggi = tidak ada: subur {subur=14, masalah=0}
- usia = > 25
 - | kebiasaan duduk = CL
 - | | musim = dingin: subur {subur=8, masalah=0}
 - | | musim = gugur
 - | | | pernah demam tinggi = < 3 bulan: subur {subur=1, masalah=0}
 - | | | pernah demam tinggi = > 3 bulan
 - | | | | pernah kecelakaan = tidak: subur {subur=4, masalah=0}
 - | | | | pernah kecelakaan = ya
 - | | | | | konsumsi alkohol = BKS: subur {subur=1, masalah=1}
 - | | | | | konsumsi alkohol = SS: masalah {subur=1, masalah=2}
 - | | | | | konsumsi alkohol = TP: subur {subur=2, masalah=0}
 - | | | pernah demam tinggi = tidak ada: masalah {subur=0, masalah=1}
 - | | musim = panas: subur {subur=2, masalah=0}

- | | musim = semi: subur {subur=8, masalah=0}
 - | kebiasaan duduk = L
 - | | musim = dingin: subur {subur=2, masalah=1}
 - | | musim = gugur: masalah {subur=1, masalah=2}
 - | | musim = panas: masalah {subur=0, masalah=1}
 - | | musim = semi: subur {subur=2, masalah=1}
 - | kebiasaan duduk = S: masalah {subur=1, masalah=1}
 - | kebiasaan duduk = SL: subur {subur=2, masalah=0}
- usia = > 31
- | konsumsi alkohol = BKS: subur {subur=5, masalah=0}
 - | konsumsi alkohol = SS: masalah {subur=0, masalah=1}
 - | konsumsi alkohol = TP: subur {subur=3, masalah=0}

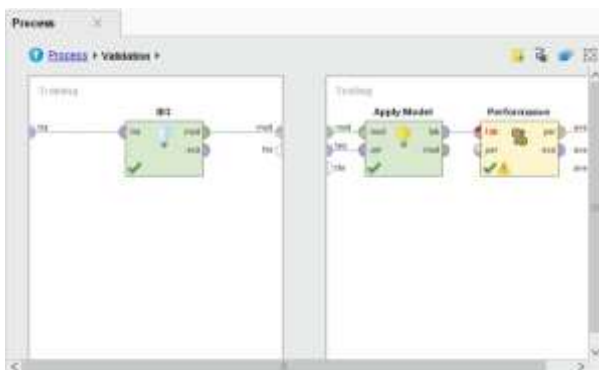
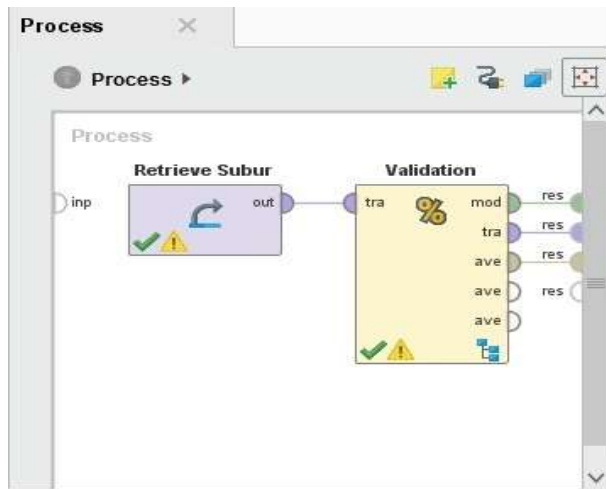
Pengujian

Guna mengetahui tingkat ketepatan dan keberhasilan dari penelitian yang dilakukan, bisa dilihat dari hasil kerja model yang terbentuk. Pengujian penelitian ini diharapkan dapat mendapatkan model yang tepat dalam mendiagnosis kesuburan seorang pria menggunakan algoritma ID3 dengan baik.

Model yang telah dikembangkan akan diuji keakuratannya dengan memasukkan beberapa data uji ke dalam pemodelan. Guna menghitung keakuratan model secara tepat, data uji sebaiknya

bukan data yang berasal dari data training (Han, J., & Kamber, M. 2006). Namun, dikarenakan data sampel hanya terdapat sejumlah 100, maka data training yang digunakan juga akan dijadikan data testing.

Guna memperoleh hasil prediksi dari penggunaan ID3, sampel yang tersedia diuji ke dalam data training. Dalam pengujian ini, ditambahkan metode untuk menyeleksi atribut lain berupa *gini index* dan *gain ratio* dengan maksud mencari tahu nilai akurasi dari masing-masing metode penyeleksian atribut melalui algoritma yang sama yaitu ID3.



Gambar 4 Pola Pencarian dan Pengujian Model

Dari hasil penerapan seperti gambar di atas, berikutnya akan dilihat kembali tingkat akurasinya dengan menggunakan dua model yaitu *confusion matrix* dan ROC (*Receiver Operating Characteristic*).

1. Evaluasi model *confusion matrix*.

Evaluasi ini digunakan guna mengukur keakuratan model yang terbentuk yaitu menggunakan *confusion matrix*. Pada evaluasi ini

digunakan perbandingan seperti pada tabel di bawah ini:

Tabel 2 Model *Confusion Matrix*

Correct Classification	Classified as	
	Positive	Negative
Positive	True Positives	False Negatives
Negative	False Negatives	True Negatives

True positives adalah tupel positif pada data set yang dikategorikan positif. Sedangkan *true negatives* adalah tupel negatif pada data set yang dikategorikan negatif. *False positives* adalah tupel positif pada data set yang dikategorikan negatif sedangkan untuk *false negatives* adalah jumlah tupel negatif yang dikategorikan positif.

Sesudah memasukkan data pengujian ke dalam model *confusion matrix*, akan dihitung nilai yang sudah dimasukkan sebelumnya untuk diketahui jumlah *sensitivity*, *specificity*, *precision* dan *accuracy*. Jumlah *sensitivity* dipakai guna membuat perbandingan jumlah *true positives* kepada jumlah tupel yang *positives* sedangkan *specificity* ialah untuk membandingkan jumlah *true negatives* terhadap jumlah tupel yang *negatives*. Dalam mengevaluasinya digunakan perhitungan seperti di bawah ini.

$$sensitivity = \frac{t_pos}{pos}$$

$$specificity = \frac{t_neg}{neg}$$

$$precision = \frac{t_pos}{(t_pos + f_pos)}$$

$$accuracy = sensitivity \frac{pos}{pos + neg} + specificity \frac{neg}{pos + neg}$$

di mana:

t_pos = jumlah *true positives*

t_neg = jumlah *true negatives*

pos = jumlah tupel positif

neg = jumlah tupel negatif

f_pos = jumlah *false positives*

Nilai *sensitivity*, *specificity*, *accuracy*, *precision*, dan *recall* dari data training yang telah dihitung menggunakan *tools rapid miner* diuji menggunakan perhitungan *split-validation* dengan kriteria *information gain*. Dengan mengikuti prosedur dari *tools rapid miner* hasil pengukuran yang telah otomatis di *split* (pembagian) sebesar 0,75 guna mencari nilai akurasi tertinggi dari data

training. Hasil pembagiannya seperti tabel di bawah ini.

Tabel 3 Konversi ke *confusion matrix*

	true Subur	true Bermasalah
pred. Subur	21	1
pred. Bermasalah	1	2

Dari pembagian seperti tabel di atas, nilai masing-masingnya menghasilkan angka dalam persen seperti di bawah ini.

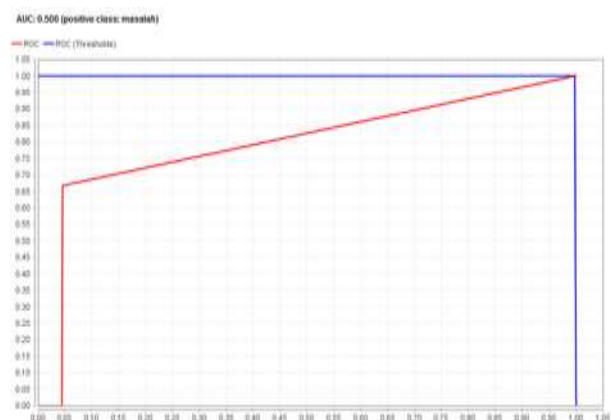
Tabel 3 Nilai *sensitivity, specificity, precision, recall, dan accuracy* dalam persentase.

Nilai	Angka (%)
<i>Sensitivity</i>	67
<i>Precision</i>	67
<i>Recall</i>	67
<i>Specificity</i>	95
<i>Accuracy</i>	92

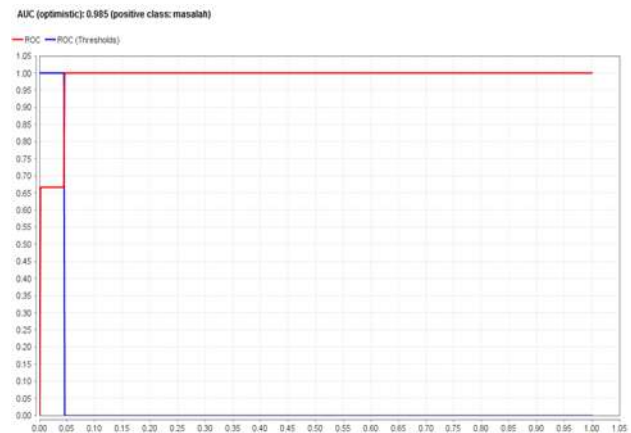
2. Evaluasi model menggunakan kurva ROC (*Receiver Operating Characteristic*)

Hasil yang ditampilkan pada gambar 5 memperlihatkan grafik ROC beserta nilai AUC (*Area Under Curve*) umum. Sedangkan gambar 6 menunjukkan grafik ROC dengan asumsi optimis. Dan pada gambar 7 menggambarkan grafik ROC dengan nilai AUC pesimis.

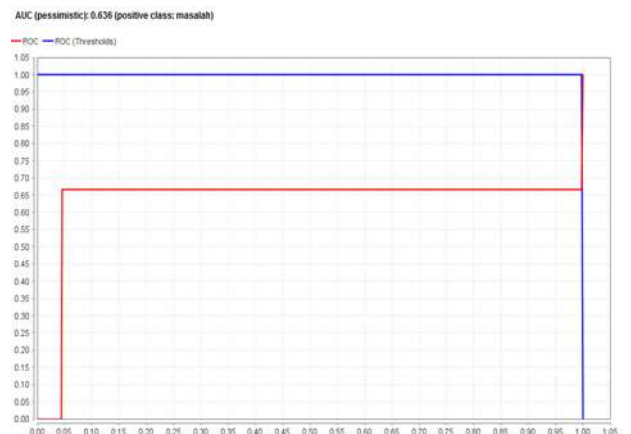
Untuk penilaian keakuratannya digunakan skala dari 0 hingga 1. Tingkat akurasi pada AUC dianggap sempurna jika nilai AUC yang dihasilkan menempati angka 1,000 dan tingkat akurasinya dianggap kurang tepat apabila nilai AUC di bawah angka 0,500.



Gambar 5 Kurva ROC dengan AUC secara umum



Gambar 6 Kurva ROC dengan AUC *Optimistic*



Gambar 7 Kurva ROC dengan AUC *Pesimistic*

4. Kesimpulan

Setelah dilakukan pengujian dan validasi terhadap data yang ada, didapatkan akurasi rata-rata sebesar 92%. Pengukuran kinerja pada hasil yang terbentuk dengan menerapkan metode pengukuran *Confusion Matrix* dan Kurva ROC.

Sehingga dapat disimpulkan analisis metode algoritma ID3 terkait permasalahan dalam mendiagnosis kesuburan pada pria merupakan salah satu metode yang tepat dalam menentukan sebuah konklusi dengan data kategorial.

Referensi

David, G., Jose Luis Girela, Joaquin De Juan, M. Jose Gomez-Torrez, Magnus Johnsson, 2012, "Predicting Seminal Quality with Artificial Intelligence Methods". Expert Systems with Applications, http://www.researchgate.net/publication/230868076_Predicting_seminal_quality_with_artificial_intelligence_methods/file/79e415058f10cc3081.pdf, (Diakses 10 April 2016)

David, Mcg. 2004. Tutorial: The ID3 Decision Tree Algorithm, Monash University Faculty of Information Technology.

- Dunham H. Margareth. 2002. *Data Mining Introductory and Advantaced Topics*, Southern Methodist University.
- Han, J., & Kamber, M. 2006. *Data Mining Concept and Tehniques*. San Fransisco: Morgan Kauffman.
- Hand, D., Mannila, H. and Smyth, P.; 2001. *Principles of Data Mining*. MIT Press.
- Irvine DS. 1998. Epidemiology and aetiology of male infertility. *Hum. Reprod.*; Vol 13(1):33-44.
- Kusrini & Luthfi, E. T. 2009. "Algoritma Data Mining. Yogyakarta: Andi Publishing.
- Larose, D. T. 2005. *Discovering Knowledge in Data An Introduction to Data Mining*. New Jersey: John Willey and Sons.
- Saifudin, A. (2018). Metode Data Mining untuk Seleksi Calon Mahasiswa pada Penerimaan Mahasiswa Baru di Universitas Pamulang. *Jurnal Teknologi*, 10(1), 25-36.
- Santosa Budi. 2007. *Data Mining: Teknik Pemanfaatan Data Untuk Keperluan Bisnis*, Yogyakarta.
- Tan, Pang-Ning, Michael Steinbach, and Vipin Kumar. 2004. *Introduction to Data Mining*. Boston: Pearson.