

**PERBANDINGAN ALGORITMA KLASIFIKASI DATA
MINING MODEL *NAÏVE BAYES* DAN *BAYES NETWORK*
UNTUK MENGIDENTIFIKASI PENYAKIT TIROID**



TESIS

BAMBANG WIJONARKO
14000218

PROGRAM PASCASARJANA MAGISTER ILMU KOMPUTER
SEKOLAH TINGGI MANAJEMEN INFORMATIKA DAN KOMPUTER
NUSA MANDIRI
JAKARTA
2013

Program Pascasarjana Magister Ilmu Komputer STMIK Nusa Mandiri

**PERBANDINGAN ALGORITMA KLASIFIKASI DATA
MINING MODEL *NAÏVE BAYES* DAN *BAYES NETWORK*
UNTUK MENGIDENTIFIKASI PENYAKIT TIROID**



TESIS

Diajukan sebagai salah satu syarat untuk memperoleh gelar
Magister Ilmu Komputer (M.Kom)

BAMBANG WIJONARKO
14000218

PROGRAM PASCASARJANA MAGISTER ILMU KOMPUTER
SEKOLAH TINGGI MANAJEMEN INFORMATIKA DAN KOMPUTER
NUSA MANDIRI
JAKARTA
2013

SURAT PERNYATAAN ORISINALITAS

Yang bertanda tangan di bawah ini :

Nama : Bambang Wijonarko
NIM : 14000218
Program Studi : Ilmu Komputer
Jenjang : Strata Dua (S2)
Konsentrasi : *E-Business*

Dengan ini menyatakan bahwa tesis yang telah saya buat dengan judul: “Perbandingan Algoritma Klasifikasi Data Mining Model Naïve Bayes Dan Bayes Network Untuk Mengidentifikasi Penyakit Tiroid” adalah hasil karya sendiri, dan semua sumber baik yang dikutip maupun yang dirujuk telah saya nyatakan dengan benar dan tesis belum pernah diterbitkan atau dipublikasikan dimanapun dan dalam bentuk apapun.

Demikianlah surat pernyataan ini saya buat dengan sebenar-benarnya. Apabila dikemudian hari ternyata saya memberikan keterangan palsu dan atau ada pihak lain yang mengklaim bahwa tesis yang telah saya buat adalah hasil karya milik seseorang atau badan tertentu, saya bersedia diproses baik secara pidana maupun perdata dan kelulusan saya dari Program Pascasarjana Magister Ilmu Komputer Sekolah Tinggi Manajemen Informatika dan Komputer Nusa Mandiri dicabut/dibatalkan.

Jakarta, 15 Mei 2013
Yang menyatakan,



Bambang Wijonarko

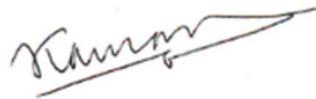
HALAMAN PENGESAHAN

Tesis ini diajukan oleh :

Nama : Bambang Wijonarko
 NIM : 14000218
 Program Studi : Ilmu Komputer
 Jenjang : Strata Dua (S2)
 Konsentrasi : *E-Business*
 Judul Tesis : “ **PERBANDINGAN ALGORITMA KLASIFIKASI DATA MINING
 MODEL NAÏVE BAYES DAN BAYES NETWORK UNTUK
 MENGIDENTIFIKASI PENYAKIT TIROID** ”

Telah berhasil dipertahankan dihadapan Dewan Penguji dan diterima sebagai bagian persyaratan yang diperlukan untuk memperoleh gelar Magister Ilmu Komputer (M.Kom) pada Program Pascasarjana Magister Ilmu Komputer Sekolah Tinggi Manajemen Informatika dan Komputer Nusa Mandiri (STMIK Nusa Mandiri).

Jakarta, 15 Mei 2013
 Pascasarjana Magister Ilmu Komputer
 STMIK Nusa Mandiri
 Direktur



Prof. Dr. Ir. Kaman Nainggolan, MS

DEWAN PENGUJI

Penguji I : Dr. Ir. Prabowo Pudjo Widodo, MS



Penguji II : Mochamad Wahyudi, M.Kom, MM, M.Pd



Penguji III / Pembimbing : Ir. Dana Indra Sensuse, MLIS, Ph.D



Program Pascasarjana Magister Ilmu Komputer STMIK Nusa Mandiri

	Lembar Konsultasi Bimbingan Tesis
	Pascasarjana Magister Ilmu Komputer STMIK Nusa Mandiri

NIM : 14000218
 Nama Lengkap : Bambang Wijonarko
 Dosen Pembimbing : Dr. Dana Indra Sensuse, MLIS, Ph.D
 Judul Tesis : Perbandingan Algoritma Klasifikasi Data Mining Model Naïve Bayes dan Bayes Network Untuk Mengidentifikasi Penyakit Tiroid



No	Tanggal Bimbingan	Materi Bimbingan	Paraf Dosen Pembimbing
1	26 Januari 2013	Pengajuan proposal	
2	9 Februari 2013	Pengajuan Bab I, Bab II,	
3	16 Februari 2013	Revisi Bab I, Bab II, dan Pengajuan Bab III	
4	23 Februari 2013	Revisi Bab III dan Pengajuan Bab IV	
5	28 Februari 2013	Revisi Bab IV dan Pengajuan Bab V	
6	03 Maret 2013	Revisi Bab V	
7	07 Maret 2013	Revisi keseluruhan	

Catatan :

Total bimbingan yang harus dilakukan adalah 6 (enam) kali pertemuan

- Bimbingan dimulai pada tanggal : 26 Januari 2013
- Bimbingan diakhiri pada tanggal : 07 Maret 2013
- Jumlah pertemuan : Tujuh Kali

Jakarta, 07 Maret 2013
Dosen Pembimbing



[Dr. Dana Indra Sensuse, MLIS, Ph.D]

KATA PENGANTAR

Puji serta syukur ke hadirat Allah SWT, karena hanya dengan rahmat dan hidayahNyalah sehingga pada akhirnya penulis dapat menyelesaikan tesis ini tepat pada waktunya meskipun terdapat banyak kekurangan. Adapun judul tesis yang penulis ambil sebagai berikut ”Perbandingan Algoritma Klasifikasi Data Mining Model Naïve Bayes Dan Bayes Network Untuk Mengidentifikasi Penyakit Tiroid”

Tujuan dari penulisan tesis ini adalah sebagai salah satu syarat untuk menyelesaikan Program Strata Dua (S-2) pada program Magister Ilmu Komputer STMIK Nusa Mandiri. Tesis ini disusun berdasarkan data sekunder yang bersumber dari UCI (University of California, Irvine) Machine Learning Repository dan diunduh melalui alamat web: <http://archive.ics.uci.edu/ml/>.

Penulis menyadari bahwa tanpa adanya dukungan dari semua pihak dalam pembuatan tesis ini, maka penulis tidak dapat menyelesaikan tesis ini tepat pada waktunya. Untuk itu, pada kesempatan ini ijinkanlah penulis menyampaikan ucapan terima kasih yang sebesar-besarnya kepada :

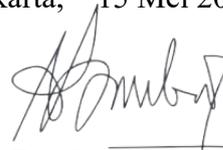
1. Bapak Prof. Dr. Ir. Kaman Nainggolan, MS, selaku Direktur Pascasarjana Magister Ilmu komputer STMIK Nusa Mandiri Jakarta.
2. Bapak Ir. Dana Indra Sensuse, MLIS, Ph.D, selaku dosen pembimbing, yang telah banyak menyediakan waktu dan pikiran untuk penulis dalam penyusunan tesis ini.
3. Seluruh karyawan dan staf pengajar Program Pascasarjana Magister Ilmu Komputer Sekolah Tinggi Manajemen Informatika dan Komputer Nusa Mandiri, Jakarta yang telah memberikan pelajaran yang berarti bagi penulis selama menempuh studi.
4. Orang tua tercinta, serta seluruh keluarga yang telah memberikan dukungan moril dan materil kepada penulis.
5. Rekan–rekan mahasiswa Program Pascasarjana Magister Ilmu Komputer, STMIK Nusa Mandiri Jakarta, atas dukungan, motivasi serta bantuan yang telah diberikan.

Program Pascasarjana Magister Ilmu Komputer STMIK Nusa Mandiri

Serta semua pihak yang tidak dapat penulis sebutkan satu persatu, sehingga atas bantuan semua pihak tersebut penulis dapat menyelesaikan penulisan tesis ini. Penulis menyadari bahwa dalam penulisan tesis ini masih terdapat banyak kekurangan dan kekeliruan, untuk itu penulis mengharapkan kritik dan saran membangun agar penulisan karya ilmiah yang penulis hasilkan di masa yang akan datang menjadi semakin lebih baik.

Akhirnya, penulis berdoa dan berharap semoga tesis ini dapat bermanfaat bagi penulis khususnya dan bagi para pembaca pada umumnya.

Jakarta, 15 Mei 2013



Bambang Wijonarko

Penulis

**SURAT PERNYATAAN PERSETUJUAN PUBLIKASI KARYA ILMIAH
UNTUK KEPENTINGAN AKADEMIS**

Yang bertanda tangan di bawah ini, saya :

Nama : Bambang Wijonarko
NIM : 14000218
Program Studi : Ilmu Komputer
Jenjang : Strata Dua (S2)
Konsentrasi : *E-Business*
Jenis Karya : Tesis

Demi pengembangan ilmu pengetahuan, dengan ini menyetujui untuk memberikan ijin kepada pihak Program Pascasarjana Magister Ilmu Komputer Sekolah Tinggi Manajemen Informatika dan Komputer Nusa Mandiri (STMIK Nusa Mandiri) **Hak Bebas Royalti Non-Eksklusif (*Non-exclusive Royalti-Free Right*)** atas karya ilmiah kami yang berjudul :“ Perbandingan Algoritma Klasifikasi Data Mining Model *Naïve Bayes* Dan *Bayes Network* Untuk Mengidentifikasi Penyakit Tiroid”.

Dengan **Hak Bebas Royalti Non-Eksklusif** ini pihak STMIK Nusa Mandiri berhak menyimpan, mengalih-media atau *bentuk*-kan, mengelolanya dalam pangkalan data (*database*), mendistribusikannya dan menampilkan atau mempublikasikannya di *internet* atau media lain untuk kepentingan akademis tanpa perlu meminta ijin dari kami selama tetap mencantumkan nama kami sebagai penulis/pencipta karya ilmiah tersebut.

Saya bersedia untuk menanggung secara pribadi, tanpa melibatkan pihak STMIK Nusa Mandiri, segala bentuk tuntutan hukum yang timbul atas pelanggaran Hak Cipta dalam karya ilmiah saya ini.

Demikian pernyataan ini saya buat dengan sebenarnya.

Jakarta, 15 Mei 2013
Yang menyatakan,



Bambang Wijonarko

ABSTRAKSI

Nama : Bambang Wijonarko
NIM : 14000218
Program Studi : Ilmu Komputer
Jenjang : Strata Dua (S2)
Konsentrasi : *E-Business*
Judul : “Perbandingan Algoritma Klasifikasi Data Mining Model Naïve Bayes Dan Bayes Network Untuk Mengidentifikasi Penyakit Tiroid”

Dalam data mining, dikenal Salah satu model Klasifikasi yang dapat digunakan untuk mengidentifikasi penyakit tiroid, adalah metode *Naïve bayes* dan *Bayes Network*. Dalam penelitian ini model dibuat menggunakan algoritma keduanya. Data yang digunakan diambil dari data pasien penderita tiroid dengan menggunakan tools KNIME. Model yang dihasilkan, kemudian dikomparasi untuk mengetahui algoritma yang paling baik dalam penentuan identifikasifikasi penyakit. Untuk mengukur kinerja kedua algoritma tersebut digunakan, metode pengujian Cross Validation dan split percentace. Hasil pengukuran ini menggunakan Confusion Matrik dan Curva ROC. Dengan menggunakan *Confusion Matrix*, *Bayes Network* memiliki akurasi yang lebih tinggi dengan nilai 98.491% dibandingkan *Naïve Bayes* dengan nilai 91.803%. Sedangkan dengan menggunakan Kurva ROC, *Bayes Network* juga memiliki akurasi yang lebih tinggi dengan curva ROC – negative (0.9337), ROC – hipertiroid (0,9933) dan ROC – hypotiroid (0,9977). Dibandingkan *Naïve Bayes* Pada curva ROC – negative (0.8760), ROC – hipertiroid (0,9789) dan ROC – hypotiroid (0,9018). Metode yang memiliki klasifikasi sangat baik secara berurut adalah *Bayes Network* dan *Naïve Bayes* berdasarkan penilaian AUC antara 0.90-1.00. Dengan demikian algoritma *Bayes Network* dapat memberikan pemecahan untuk permasalahan dalam mengidentifikasi penyakit tiroid.

Kata Kunci : *Naïve Bayes*, *Bayes Network*, *Confusion Matix*, *ROC Curve*, *AUC*

ABSTRACT

Name : Bambang Wijonarko
NIM : 14000218
Study of Program : Magister Ilmu Komputer
Levels : Strata Dua (S2)
Concentration : *E-Business*
Tittle : “Perbandingan Algoritma Klasifikasi Data Mining Model Naïve Bayes Dan Bayes Network Untuk Mengidentifikasi Penyakit Tiroid”

In data mining, known Classification model that can be used to identify thyroid disease, is Naive Bayes and Bayes Network methods. In this study, a model is made by using both algorithm. the data used are taken from the data of Patients with thyroid by using the tools KNIME. The model then compared to determine the best algorithm in determination of disease identification. To measure the performance of the two algorithms, it used methods of testing of cross validation and split percentage. The measurement results using confusion matrix and ROC curves. By using the confusion matrix, Bayes Network has higher accuracy with 98,491% compared with the Naive Bayes with 91,803%. By Using the ROC curve, Bayes Network also has higher accuracy with the ROC curve - negative (0.9337), ROC - hypertiroid (0.9933) and ROC - hypotiroid (0.9977). while Naive Bayes with ROC curve - negative (0.8760), ROC - hypertiroid (0.9789) and ROC - hypotiroid (0.9018). The method which has very good classification is sequentially bayes network and naïve bayes based on assessment AUC between 0.90-1.00. thus Bayes Network algorithm can provide solutions the problems of identifying thyroid disease.

Keywords : Naïve Bayes, Bayes Network,, Confusion Matix, , ROC Curve, AUC

DAFTAR ISI

	Halaman
HALAMAN SAMPUL	i
HALAMAN JUDUL.....	ii
SURAT PERNYATAAN ORISINALITAS	iii
HALAMAN PENGESAHAN.....	iv
LEMBAR KONSULTASI BIMBINGAN TESIS	v
KATA PENGANTAR	vi
SURAT PERNYATAAN PERSETUJUAN PUBLIKASI	
KARYA ILMIAH UNTUK KEPENTINGAN AKADEMIS.....	viii
ABSTRAK	ix
<i>ABSTRACT</i>	x
DAFTAR ISI.....	xi
DAFTAR TABEL.....	xiii
DAFTAR GAMBAR	xiv
DAFTAR LAMPIRAN.....	xv
BAB 1. PENDAHULUAN	1
1.1.Latar Belakang Penulisan	1
1.2.Masalah Penelitian.....	4
1.2.1. Identifikasi Masalah	4
1.2.2. Rumusan Masalah	5
1.2.3. Batasan Masalah.....	5
1.3.Tujuan Penelitian	5
1.4.Manfaat Penelitian.....	5
1.5. Ruang Lingkup Penelitian	6
1.6. Sistematika Penulisan.....	6
BAB 2. LANDASAN TEORI/KERANGKA PEMIKIRAN	7
2.1.Tinjauan Pustaka.....	7
2.1.1. <i>Data Mining</i>	7
2.1.2. Klasifikasi <i>Data Mining</i>	8
2.1.3. <i>Naïve Bayes</i>	11
2.1.4. Algoritma <i>Neural Network</i>	14
2.1.5. Tiroid	21
2.1.6. <i>Pengujian K-Fold Cross Validation</i>	27
2.1.7. Evaluasi dan validasi Metode prediksi data mining	28
2.1.8. Evaluasi dan Validasi Metode Prediksi <i>Data Mining</i>	20
2.2.Tinjauan Studi.....	29
2.3.Kerangka Pemikiran	30
BAB 3. METODE PENELITIAN	31
3.1.Desain Penelitian	31
3.1.1. Jenis Penelitian	31
3.1.2. Metode Pengumpulan Data	31
3.2. Tahap-tahap Penelitian	31
3.2.1. Pengumpulan Data.....	31

3.2.2. Pengolahan Awal Data	35
3.2.3. Model atau Metode yang diusulkan atau yang dikembangkan	36
3.2.4. Eksperimeny dan pengujian model atau metode	38
3.2.1. Evaluasi dan Validasi hasil.....	39
3.3. Perangkat Yang digunakan.....	39
3.3.1. Perangkat Keras (Hardware)	39
3.3.2. Perangkat Lunak (Software).....	39
BAB 4. HASIL ANALISA DAN PEMBAHASAN.....	40
4.1 Eksperimen dan Pengujian Model	40
4.1.1. Hasil Penelitian.....	40
4.2. Evaluasi dan Validasi	46
4.2.1. Pengujian Model	46
4.2.2. Kurva ROC.....	48
4.3. Implikasi Hasil Penelitian.....	50
BAB 5. KESIMPULAN.....	53
5.1 Kesimpulan	53
5.2 Saran	53
DAFTAR REFERENSI	54

DAFTAR TABEL

	Halaman
Tabel 2.1 Data pemilihan rumah	12
Tabel 2.2 Probabilitas harga tanah terhadap pemilihan rumah	13
Tabel 2.3 Probabilitas jarak terhadap pemilihan rumah	13
Tabel 2.4 Probabilitas angkutan umum terhadap pemilihan rumah	13
Tabel 2.5 Probabilitas pemilihan rumah	13
Tabel 2.6 Data pemilihan rumah nilai kontinue.....	14
Tabel 2.7 Probabilitas harga tanah terhadap pemilihan rumah nilai kontinue	15
Tabel 2.8 Probabilitas jarak terhadap pemilihan rumah nilai kontinue	15
Tabel 2.9 Probabilitas angkt umum terhadap pemiiil rumah nilai kontinue	15
Tabel 2.10 Probabilitas pemilihan rumah.....	15
Tabel 2.11 Excerize dan Diet terhadap HD=Yesaan dengan fakta	18
Tabel 2.12 HD terhadap BP=Yes	18
Tabel 2.13 Diet terhadap Hb=Yes.....	19
Tabel 2.14 HD dan Hb terhadap CP=Yes	19
Tabel 2.15 Perbandingan gejala Hipertiroid dan Hipertiroid.....	23
Tabel 2.16 K-flod Cross Validation.....	27
Tabel 2.17 Model Confusion Matrix.....	28
Tabel 2.12 HD terhadap BP=Yes	18
Tabel 2.13 Diet terhadap Hb=Yes.....	19
Tabel 2.14 HD dan Hb terhadap CP=Yes	19
Tabel 2.15 Perbandingan gejala Hipertiroid dan Hipertiroid.....	23
Tabel 2.14 HD dan Hb terhadap CP=Yes	19
Tabel 2.15 Perbandingan gejala Hipertiroid dan Hipertiroid.....	23
Tabel 3.1 Attribut yang digunakan untuk mengidentifikasi penyakit tiroid.....	33
Tabel 3.2 Data Pasien Penyakit tiroid maupun yang tidak	34
Tabel 3.3 Penderita penyakit hipertiroid.....	34
Tabel 3.4 Penderita penyakit hipotiroid	35
Tabel 3.5 Data Yang valid	36
Tabel 4.1. Attribut Kategorikal data trining.....	40
Tabel 4.2. Attribut Kontiyu.....	42
Tabel 4.3. Perhitungan untuk Menentukan Klasifikasi Data testing X.....	42
Tabel 4.4. Perhitungan untuk Menentukan Klasifikasi Data testing X menggunakan Bayes Network Menggunakan berdasarkan Tabel Pada lampiran 2	45
Tabel 4.5. Pengujian menggunakan cross 6	46
Tabel 4.6. Pengujian menggunakan cross 8	47
Tabel 4.7. Pengujian menggunakan cross 10	47
Tabel 4.8. Dengan Menggunakan Split	48
Tabel 4.9 Komparasi Nilai AUC.....	50
Lampiran 1 : Data sumber.....	
Lampiran 2 : Tabel Perhitungan Bayes Network.....	

DAFTAR GAMBAR

		Halaman
Gambar 2.1	Proses KDD	7
Gambar 2.2	Struktur Bayes Network untuk contoh kasus klasifikasi	18
Gambar 2.3	Pembentukan Hormon Tiroid	25
Gambar 2.4	Kerangka Pemikiran	30
Gambar 3.1	Relasi pengolahan data mentah ke data missing	36
Gambar 3.2	Model yang diusulkan	37
Gambar 3.3	Model Split Precentage	38
Gambar 3.4	Model Cross Validation	38
Gambar 4.1.	Model Bayes Network.....	44
Gambar 4.2.	Curva ROC pada test data Algoritma Naïve Bayes	49
Gambar 4.3.	Curva ROC pada test data Algoritma Bayes Network	49

BAB I

PENDAHULUAN

1.1. Latar Belakang penulisan

Tiroid adalah merupakan salah satu bagian tubuh yang sangat penting bagi manusia, tiroid berbentuk kelenjar dan letaknya di bawah jakun pada leher. Tiroid merupakan kelenjar endokrin terbesar dalam tubuh berbentuk kupu-kupu (Hamdani, 2010). Fungsi kelenjar tiroid adalah memproduksi hormon yang berguna untuk menjaga metabolisme dan pertumbuhan normal. Berfungsi untuk meningkatkan jumlah oksigen pada sel dan rangsangan jaringan tubuh dalam menghasilkan protein. Gangguan pada kelenjar tiroid sering disebut dengan penyakit gondok. Ada dua jenis gangguan tiroid yang dapat muncul yaitu hipertiroid dan hipotiroid. Hipertiroid adalah kondisi dimana kelenjar tersebut bekerja secara berlebihan, sedangkan hipotiroid adalah kebalikannya (Tandra, 2009).

Menurut WHO (2004), proporsi kematian di dunia yang disebabkan oleh penyakit tidak menular sebesar 60% dan proporsi kesakitan sebesar 47%, dan diperkirakan pada tahun 2020 proporsi kematian akan meningkat menjadi 73% dan proporsi kesakitan menjadi 60%.

Menurut Tahaha didalam (Bachtiar, 2009) Di Indonesia pada tahun 1982, diperkirakan terdapat 30 juta orang beresiko (Gangguan akibat kekurangan yodium) GAKY. Angka tersebut bertambah menjadi sekitar 42 juta jiwa pada tahun 1994, tersebar di sekitar 190 kabupaten dan 26 propinsi. Dari hasil survey pemetaan nasional GAKY tahun 1998, penderita GAKY terdapat di kepulauan Maluku, NTT (nusa Tenggara Timur), Sumatra barat dan propinsi Sulawesi tenggara.

Penderita penyakit tiroid lebih sering terkena pada wanita pada RSUP Perbandingan wanita dan laki-laki adalah 3,1 : 1 di RSCM Jakarta adalah 6 : 1, di RS. Dr. Soetomo 8 : 1 dan di RSHS Bandung 10 : 1.1 Sedangkan

distribusi menurut umur di RSUP Palembang yang terbanyak adalah pada usia 21 - 30 tahun (41,73%), tetapi menurut beberapa penulis lain puncaknya antara 30 - 40 tahun. Jumlah penderita penyakit ini di seluruh dunia pada tahun 1960 diperkirakan 200 juta, 12 juta di antaranya terdapat di Indonesia. Angka kejadian hipertiroidi yang didapat dari beberapa klinik di Indonesia berkisar antara 44,44% — 48,93% dari seluruh penderita dengan penyakit kelenjar gondok (Hermawan, 1990).

Di amerika serikat dengan jumlah penduduk lebih dari 275 juta, diperkirakan ada 27 juta orang yang mengidap berbagai kelaian kelenjar tiroid. Lebih dari separuh di antaranya ternyata tidak tahu mengenai penyakit tiroid yang mereka derita.(Tandra, 2009)

Penelitian H. Moayeri dan Z. Haghshenas di Iran University of Medical Sciences tahun 2004 menemukan 384 kasus yang didiagnosis menderita struma diantaranya 320 orang (83,4%) eutiroidisme, 49 orang (12,7%) hipotiroidisme dan 15 orang (3,9%) hipertiroidisme (Moayeri, 2002).

Penderita hipotiroidisme juga sering di alami oleh ibu hamil dan akan berakibat jelek pada janin yang di kandung terutama untuk perkembangan otaknya, sedangkan pada ibu hamil dengan penderita hipertiroidisme, bayi yang dikandung akan lahir memiliki berat badan yang rendah dan kemungkinan terjadinya cacat bawan (Susanto, 2006). Dan berdasarkan latar belakang tersebut maka perlu dilakukan penelitian tentang klasifikasi penderita penyakit tiroid.

Dalam bidang kesehatan, kesalahan dalam mendiagnosa penyakit yang dialami pasien adalah tanggung jawab yang paling berat untuk diemban oleh ahli kesehatan. Kesalahan dalam mendiagnosa penyakit dapat menyebabkan hal yang membahayakan bagi kesehatan pasien dan bahkan dapat menyebabkan kematian (Neshat, Mehdi, Sargolzaei, Toosi, Masoumi, 2012).

Teknik analisa konvensional secara manual yang selama ini digunakan tidak lagi efektif digunakan untuk mendiagnosa karena membutuhkan waktu yang cukup lama. penggunaan sistem pengetahuan berbasis komputer sebagai teknik analisa dalam mengidentifikasi penyakit menjadi semakin penting. (Neshat, Mehdi, Yaghobi, 2009).

Untuk mengurangi waktu yang dibutuhkan untuk mengidentifikasi dan meningkatkan tingkat akurasi diagnosa, dibutuhkan adanya pengembangan suatu sistem diagnosa medis yang baik dan dapat diandalkan. Oleh karena itu, metode *soft computing* menunjukkan potensi yang menjanjikan untuk dikembangkan guna menghasilkan diagnosa medis yang tepat. (Hanan, Shaikh Abdul, Manza, Ramteke, 2010).

Penggunaan *data mining* dengan model *Klasifikasi* sebagai salah satu pilihan untuk diagnosa penyakit tiroid dapat menjadi alternatif pilihan yang tepat. Namun sampai saat ini belum diketahui algoritma yang paling akurat dalam penentuan diagnosa untuk prediksi penyakit tiroid. Untuk itu maka dalam penelitian ini akan dilakukan komparasi algoritma yang memiliki performance lebih tinggi dalam mendeteksi penyakit tiroid

Didalam model *Klasifikasi* menggunakan teorema Bayes Menurut (Shadiq, 2009) Data latih untuk Teorema Bayes membutuhkan paling tidak perkalian kartesius dari seluruh kelompok atribut yang mungkin, jika misalkan ada 16 atribut yang masing-masingnya berjenis Boolean, maka data latih minimal yang dibutuhkan oleh Teorema Bayes untuk digunakan dalam klasifikasi adalah $2^{16} = 65.536$ data, masalah yang terjadi adalah kebanyakan data latih tidak memiliki sebanyak itu sehingga dapat mempengaruhi dalam memprediksi. Untuk mengatasi permasalahan diatas, berbagai varian dari pengklasifikasian yang menggunakan Teorema Bayes diajukan, adalah Naïve Bayes Dan Bayes Network.

Dari penelitian yang dilakukan (sarwar dan sharma, 2012) menyatakan bahwa algoritma Naive Bayes tercatat mengungguli algoritma lain di kedua masalah diagnosa medis maupun non-medis. Menurut penelitian lain menyatakan (Mrutyunjaya Panda dan Mana R. Patra 2007) menyatakan Naïve Bayes juga dikenal sebagai teknik yang paling baik dalam hal waktu komputasi dibandingkan teknik algoritma data mining lainnya.

Didalam Naïve Bayes asumsi keidependenan atribut akan menghilangkan kebutuhan banyaknya jumlah data latih dari perkalian kartesius seluruh atribut yang dibutuhkan untuk mengklasifikasikan suatu data Dampak negative dari

asumsi Naïve tersebut secara intuitif akan berpengaruh dalam pengklasifikasian, namun percobaan empiris mengatakan sebaliknya.

Sedangkan pada Bayesian Network, variabel input bisa saling dependent. Bayesian Network atau Probabilistik Network adalah model grafik untuk merepresentasikan interaksi antar variabel. Bayesian Network digambarkan seperti graf yang terdiri dari simpul (node) dan busur (arc). Pertanyaan yang muncul adalah apakah performa yang didapatkan dari pengaplikasian asumsi Naïve Bayes dengan keidependenan atributnya lebih baik dibandingkan Bayes network dalam mengidentifikasi penyakit tiroid ?.

1.2. Masalah Penelitian

1.2.1. Identifikasi Masalah

Banyak metode yang dapat digunakan dalam memprediksi suatu penyakit diantaranya menggunakan teknik data mining. Teknik data mining yang dapat digunakan dalam memprediksi salah satunya adalah metode klasifikasi banyak metode yang dapat digunakan untuk memprediksi penyakit, diantaranya adalah Metode Naïve Bayes dan Network Bayes. Dan identifikasi permasalahan adalah sebagai berikut :

1. Belum adanya Metode apa yang tepat untuk menyelesaikan permasalahan tersebut
2. Belum diketahui performance kedua metode Naïve bayes dengan Bayes network dalam mengidentifikasi penyakit tiroid.

1.2.2. Rumusan Masalah

Berdasarkan identifikasi masalah dan batasan masalah, maka rumusan masalah dalam penelitian ini adalah bersifat agregat: Apakah Performance *Naïve Bayes* lebih baik dibandingkan *Bayes Network* dalam mengidentifikasi jenis penyakit tiroid dengan menggunakan data secara keseluruhan.

1.3. Tujuan Penelitian

Tujuan dari penelitian ini adalah untuk melakukan analisis dan komparasi metode klasifikasi data mining yang paling akurat dalam mengidentifikasi penyakit pasien penderita tiroid ?.

1.4. Manfaat Penelitian

Manfaat dari penelitian ini adalah:

- a. Manfaat praktis dari penelitian ini adalah diharapkan agar dapat digunakan untuk membantu dalam mengidentifikasi jenis penyakit tiroid, yang didapat melalui hasil percobaan sebagai bukti empiris.
- b. Manfaat teoritis dari hasil penelitian ini diharapkan dapat memberikan sumbangan bagi pengembangan teori yang berkaitan dengan komparasi klasifikasi data mining khususnya dalam mengidentifikasi suatu penyakit di bidang kedokteran.

1.5 Ruang Lingkup Penelitian

Ruang lingkup pembahasan dalam dibatasi pada perbandingan metode klasifikasi data mining dengan algoritma Naive Bayes dan Bayes Network dengan cara menganalisis atribut yang menjadi parameter.

1.6. Sistematika Penulisan

Sistematika penulisan penelitian ini disusun dalam 5 (lima) bab, yang terdiri dari :

BAB 1. PENDAHULUAN

Bab ini terdiri dari beberapa sub bab, antara lain : latar belakang penulisan, masalah penelitian, tujuan, manfaat penelitian, Ruang lingkup dan sistematika penulisan.

BAB 2. LANDASAN TEORI/KERANGKA PEMIKIRAN

Bab ini berisi tentang obyek penelitian, tinjauan studi, kerangka pemikiran yang melandasi penelitian ini.

BAB 3. METODE PENELITIAN

Bab ini membahas tentang desain penelitian yang terdiri dari jenis penelitian yang dilakukan dan metode pengumpulan data, serta eksperimen yang dilakukan.

BAB 4. HASIL PENELITIAN DAN PEMBAHASAN

Bab ini berisi hasil penelitian, pembahasan dan implikasi penelitian.

BAB 5. PENUTUP

Bab ini mencakup kesimpulan yang dibuat berdasarkan pembahasan bab-bab sebelumnya dan saran-saran untuk kepentingan penelitian selanjutnya.

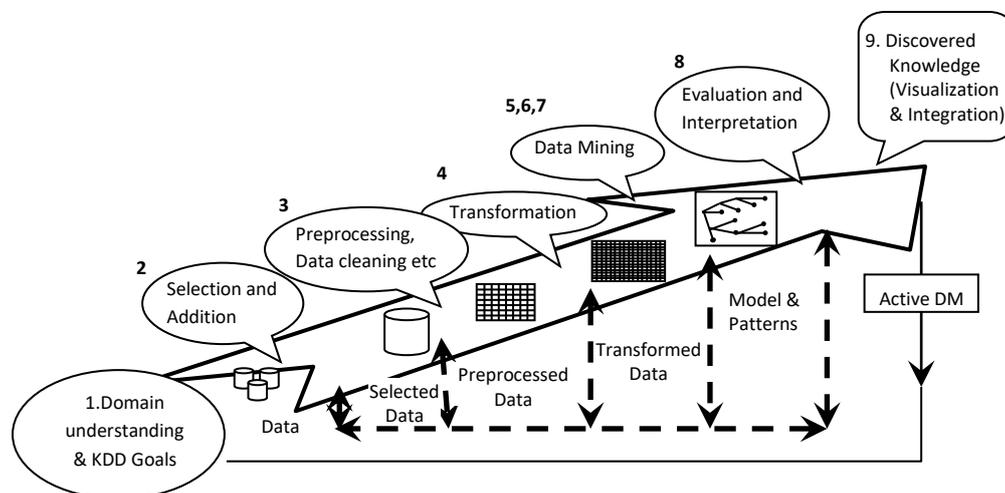
BAB 2

LANDASAN TEORI/KERANGKA PEMIKIRAN

2.1. Tinjauan Pustaka

2.1.1. Data mining

Menurut Daryl Pregibons dalam (Gorunescu, 2011) *Data mining* adalah perpaduan dari ilmu statistik, kecerdasan buatan, dan penelitian bidang *database*. Nama *data mining* berasal dari kemiripan antara pencarian informasi yang bernilai dari *database* yang besar dengan menambang sebuah gunung untuk sesuatu yang bernilai (Sumathi, 2006). Menurut (Witten, 2011) data mining didefinisikan sebagai proses pengenalan pola dalam data. Data mining, sering di sebut *knowledge discovery in database (KDD)*, adalah suatu kegiatan yang meliputi pengumpulan, pemakaian data historis untuk menentukan keteraturan, pola atau hubungan dalam set data berukuran besar. Keluaran dari data mining ini bisa dipakai untuk memperbaiki pengambilan keputusan di masa depan (Santoso, 2007).



Gambar 2.1 Proses KDD (Maimon, 2010)

Banyak kasus-kasus dalam kehidupan sehari-hari yang memakai teknik data mining contoh berikut ini penerapan data mining dalam kehidupan sehari-hari:

1. memprediksi harga suatu saham dalam beberapa bulan kedepan berdasarkan performasi perusahaan dan data-data ekonomi

2. memprediksi apakah seorang pasien yang di opname akan mendapat serangan jantung berikutnya berdasarkan catatan sebelumnya dan pola makanan.
3. Identifikasi apakah sudah terjadi penipuan kartu kredit dengan melihat catatan transaksi yang tersimpan dalam database perusahaan kartu kredit
4. Memprediksi permintaan barang dalam beberapa tahun berdasarkan data permintaan barang di tahun-tahun sebelumnya.

Dan masih banyak lagi contoh-contoh dari berbagai bidang yang bisa diselesaikan dengan teknik data mining (Santoso, 2007).

2.1.2. Klasifikasi *Data Mining*

Klasifikasi (Han, 2006) adalah proses penemuan model (atau fungsi) yang menggambarkan dan membedakan kelas data yang bertujuan agar bisa digunakan untuk memprediksi kelas dari objek yang label kelasnya tidak diketahui.

Klasifikasi data terdiri dari 2 langkah proses. Pertama adalah *learning* (fase *training*), dimana algoritma klasifikasi dibuat untuk menganalisa data *training* lalu direpresentasikan dalam bentuk *rule* klasifikasi. Proses kedua adalah klasifikasi, dimana data tes digunakan untuk memperkirakan akurasi dari *rule* klasifikasi. (Han, 2006).

Proses klasifikasi didasarkan pada empat komponen: (Gorunescu, 2011)

a. Kelas

Variabel dependen yang berupa kategorikal yang merepresentasikan 'label' yang terdapat pada objek. Contohnya: resiko penyakit jantung, resiko kredit, *customer loyalty*, jenis gempu.

b. *Predictor*

Variabel independen yang direpresentasikan oleh karakteristik (atribut) data. Contohnya: merokok, minum alkohol, tekanan darah, tabungan, aset, gaji.

c. *Training dataset*

Satu set data yang berisi nilai dari kedua komponen di atas yang digunakan untuk menentukan kelas yang cocok berdasarkan *predictor*.

d. *Testing dataset*

Berisi data baru yang akan diklasifikasikan oleh model yang telah dibuat dan akurasi klasifikasi dievaluasi.

Berikut ini adalah algoritma klasifikasi yang banyak digunakan secara luas: (Gorunescu, 2011)

1. *Decision/classification trees*

Decision tree digunakan untuk memprediksi keanggotaan suatu objek ke dalam kategori (kelas) yang berbeda, berdasarkan variabel prediktor. Algoritma *decision tree* yang dikenal luas antara lain *Hunt*, CART (C&RT), ID3, C4.5&C5.0, SLIQ, SPRINT, QUEST, DTREG, THAID, CHAID (Gorunescu, 2011).

2. *Bayesian classifiers/ Naïve Bayes classifiers*

Klasifikasi Bayes (Kusrini, 2009) adalah pengklasifikasian statistik yang dapat digunakan untuk memprediksi probabilitas keanggotaan suatu kelas. Klasifikasi Bayes juga dikenal dengan *Naïve Bayes*, *idiot's Bayes*, *simple Bayes*, dan *independence Bayes* (Wu, 2009). Klasifikasi Bayes didasarkan pada teorema *Bayes*, diambil dari nama seorang ahli matematika yang juga menteri *Prebysterian* Inggris, Thomas Bayes (1702-1761).

3. *Neural networks*

Neural network adalah (Han, 2006) satu set unit *input/output* yang terhubung dimana tiap relasinya memiliki bobot. *Neural Network* dimaksudkan untuk mensimulasikan perilaku sistem biologi susunan syaraf manusia (Alpaydin, 2010).

4. Analisa Statistik

Analisa statistik selama ini juga melakukan hal yang sama seperti *data mining*. Beberapa metode estimasi dan prediksi tradisional merupakan turunan dari analisa statistik, antara lain *linear regression* dan *multiple regression* (Larose, 2005).

5. Algoritma Genetika

Algoritma genetika merepresentasikan teknik identifikasi solusi perkiraan untuk masalah optimasi dan pencarian. Algoritma genetika merupakan tipe algoritma evolusioner yang paling populer, berbasis pada populasi solusi potensial menggunakan mekanisme spesifik yang diinspirasi dari evolusi biologi (genetika alamiah), seperti kromosom, reproduksi, rekombinasi, seleksi, *survival of fittest* (Gorunescu, 2011).

6. *Rough sets*

Rough Set dibuat oleh Zdzislaw Pawlak pada awal tahun 1980, dalam rangka mengungkap secara matematis konsep kesamaran, tujuan utamanya adalah menjadi proses otomatisasi transformasi data ke dalam pengetahuan. *Rough set* merupakan pendekatan matematis untuk pengetahuan yang tidak sempurna, hal ini penting dalam logika *fuzzy*. Konsep *rough set* secara umum bisa didefinisikan dengan 2 topologi, yaitu *interior* dan *closure*. Ide dasar *rough set* terdapat pada kenyataan bahwa, berdasarkan satu set objek, satu set atribut dan nilai keputusan, seseorang dapat membuat *rule* untuk menemukan perkiraan *upper* dan *lower*, dan daerah batasan (*boundary region*) dari set objek. Setelah *rule* dibuat, maka objek baru bisa dengan mudah diklasifikasikan ke dalam salah satu daerah (*region*) (Gorunescu, 2011).

7. *k-nearest neighbor*

K- nearest neighbor (kNN) adalah klasifikasi yang menyimpan semua data *training* dan melakukan klasifikasi dengan cara membandingkan antara atribut data baru yang paling cocok dengan atribut *record* yang terdapat pada data *training*. kNN adalah kasus khusus dalam *instance-based learning*. Ini termasuk case-based reasoning, yang menangani data simbol. kNN juga merupakan contoh teknik *lazy learning*, yaitu teknik yang menunggu sampai pertanyaan (*query*) datang agar sama dengan *data training* (Wu, 2009).

8. Metode *Rule Based*

Menurut (Gorunescu, 2011) Klasifikasi *rule based* artinya proses penggunaan data *training* dengan objek berlabel dimana *rule* akan diekstrak untuk menghasilkan klasifikasi (Berdasarkan proses pembuatan *rule*, ada dua pendekatan yang mungkin, yaitu :

a. Metode langsung

Metode ini mengekstrak *rule* langsung dari data. Contohnya algoritma RIPPER, CN2, *Holte's 1R*, *Boolean reasoning*.

b. Metode tidak langsung

Metode ini mengekstrak *rule* dari metode klasifikasi lainnya, contohnya, mengekstrak *rule* dari *decision trees* atau *neural networks*.

9. *Support vector machines* (SVM)

Support vector machines (SVM) adalah suatu teknik yang relatif baru (1995) untuk melakukan prediksi, baik dalam kasus klasifikasi maupun regresi, yang sangat populer belakangan ini. SVM berada dalam satu kelas dengan ANN dalam fungsi dan kondisi permasalahan yang bisa diselesaikan. Keduanya masuk dalam kelas Supervised learning (Santoso, 2007). SVM cukup populer untuk penggunaan klasifikasi karena kelebihanannya antara lain dari segi cara kerja, SVM baik untuk klasifikasi, tidak tergantung pada jumlah fitur dan bisa mengatasi masalah dimensi. Dari segi komputasi, SVM dapat melakukan proses *training* dengan cepat dan ini berguna dalam teknik *learning* ketika menghadapi masalah ketidaktegasan (Maimon, 2010).

2.1.3. *Naive Bayes*

Naive Bayesian merupakan salah satu metode pengklasifikasi berpeluang sederhana yang berdasarkan pada penerapan Teorema Bayes dengan asumsi antar variabel penjelas saling bebas (independen). Klasifikasi Bayes juga dikenal dengan *Naïve Bayes*, memiliki kemampuan sebanding dengan dengan pohon keputusan dan *neural network* (Han, 2006). Algoritma ini memanfaatkan metode probabilitas dan statistik yang dikemukakan oleh ilmuwan Inggris Thomas Bayes, yaitu memprediksi probabilitas di masa depan berdasarkan pengalaman di masa sebelumnya. Teknik *Naïve Bayes* (NB) adalah salah satu bentuk sederhana dari *Bayesian* yang jaringan untuk klasifikasi. Sebuah jaringan *Bayes* dapat dilihat sebagai diarahkan sebagai tabel dengan distribusi probabilitas gabungan lebih dari satu set diskrit dan variabel stokastik (liao, 2007). Klasifikasi Bayes didasarkan pada teorema Bayes, diambil dari nama seorang ahli matematika yang juga menteri *Prebysterian* Inggris, Thomas Bayes (1702-1761), yaitu (Bramer, 2007):

$$P(x|y) = \frac{P(y|x) P(x)}{P(y)}$$

keterangan :

y = data dengan kelas yang belum diketahui

x = hipotesis data y merupakan suatu kelas spesifik

$P(x | y)$ = probabilitas hipotesis x berdasar kondisi y (*posteriori probability*)

$P(x)$ = probabilitas hipotesis x (*prior probability*)

$P(y | x)$ = probabilitas y berdasarkan kondisi pada hipotesis x

$P(y)$ = probabilitas dari y

1. Jika atribut ke- i bersifat diskret, maka $P(x_i|y)$ diestimasi sebagai frekwensi relatif dari sampel yang memiliki nilai x_i sebagai atribut ke i dalam kelas C .
2. Namun jika atribut ke- i bersifat kontinu, maka $P(x_i|y)$ diestimasi dengan fungsi densitas Gauss.

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

dengan μ = mean, dan σ = deviasi standar $\pi = 3,1416$ dengan $e = 2,7183$

Berikut ini adalah contoh perhitungan Algoritma Naïve Bayes.

Misalnya diketahui apakah suatu objek masuk dalam ketegori dipilih untuk perumahan atau tidak dengan algoritma Naive Bayes Classifier. Untuk menetapkan suatu daerah akan dipilih sebagai lokasi untuk mendirikan perumahan, telah dihimpun 10 aturan.

Ada 4 atribut yang digunakan, yaitu:

- a. Harga tanah per meter persegi (C1),
- b. Jarak daerah tersebut dari pusat kota (C2),
- c. Ada atau tidaknya angkutan umum di daerah tersebut (C3), dan
- d. Keputusan untuk memilih daerah tersebut sebagai lokasi perumahan (C4).

1. Probabilitas Atribut Data Kategori

Tabel 2.1 Data Pemilihan Rumah

Aturan ke-	Harga tanah (C1)	Jarak dari pusat kota (C2)	Ada Angkutan Umum (C3)	Dipilih untuk Perumahan (C4)
1	Murah	Dekat	Tidak	Ya
2	Sedang	Dekat	Tidak	Ya
3	Mahal	Dekat	Tidak	Ya
4	Mahal	Jauh	Tidak	Tidak
5	Mahal	Sedang	Tidak	Tidak
6	Sedang	Jauh	Ada	Tidak
7	Murah	Jauh	Ada	Tidak
8	Murah	Sedang	Tidak	Ya
9	Mahal	Jauh	Ada	Tidak
10	Sedang	Sedang	Ada	Ya

1. Probabilitas kemunculan setiap nilai untuk attribut Harga Tanah (C1)

Tabel 2.2 Probabilitas Harga Tanah terhadap Pemilihan Rumah

Harga tanah	Jumlah kejadian "Dipilih"		Probabilitas	
	Ya	Tidak	Ya	Tidak
Murah	2	1	2/5	1/5
Sedang	2	1	2/5	1/5
Mahal	1	3	1/5	3/5
Jumlah	5	5	1	1

2. Probabilitas kemunculan setiap nilai untuk attribut Jarak dari Pusat Kota (C2)

Tabel 2.3 Probabilitas Jarak terhadap Pemilihan Rumah

Jarak dari pusat kota	Jumlah kejadian "Dipilih"		Probabilitas	
	Ya	Tidak	Ya	Tidak
Dekat	3	0	3/5	0
Sedang	2	1	2/5	1/5
Jauh	0	4	0	4/5
Jumlah	5	5	1	1

3. Probabilitas kemunculan setiap nilai untuk attribut Ada Angkutan Umum (C3)

Tabel.2.4 Probabilitas Angkutan umum terhadap Pemilihan Rumah

Angkutan umum	Jumlah kejadian "Dipilih"		Probabilitas	
	Ya	Tidak	Ya	Tidak
Ada	1	3	1/5	3/5
Tidak	4	2	4/5	2/5
Jumlah	5	5	1	1

4. Probabilitas kemunculan setiap nilai untuk attribut Dipilih untuk perumahan (C4)

Tabel.2.5 Probabilitas Pemilihan Rumah

Dipilih Untuk Perumahan	Jumlah kejadian "Dipilih"		Probabilitas	
	Ya	Tidak	Ya	Tidak
Jumlah	5	5	1/2	1/2

5. Menghitung probabilitas setiap kejadian :

- a. Berdasarkan data tersebut, apabila diketahui suatu daerah dengan harga tanah MAHAL, jarak dari pusat kota SEDANG, dan ADA angkutan umum, maka dapat dihitung:

$$YA = P(Ya|Tanah=MAHAL).P(Ya|Jarak=SEDANG).P(Ya|Angkutan=ADA).P(Ya) \\ = 1/5 \times 2/5 \times 1/5 \times 5/10 = 2/125 = 0,008$$

$$TIDAK = P(Tidak|Tanah=MAHAL).P(Tidak|Jarak=SEDANG).P(Tidak|Angkutan=ADA).P(Ya) \\ = 3/5 \times 1/5 \times 3/5 \times 5/10 = 2/125 = 0,036$$

Nilai probabilitas dapat dihitung dengan melakukan normalisasi terhadap likelihood tersebut sehingga jumlah nilai yang diperoleh = 1

$$\text{Probabilitas YA} = \frac{0,008}{0,008+0,036} = 0,182$$

$$\text{Probabilitas TIDAK} = \frac{0,036}{0,008+0,036} = 0,818$$

Maka hasil yang di dapat adalah TIDAK , karena $P(\text{TIDAK}) >$ dari $P(\text{YA})$,

2. Probabilitas Atribut Data Kontinyu

Untuk jenis data harga tanah dan jarak pusat kota yang kontinue, misalnya :

Tabel 2.6 Data Pemilihan Rumah nilai kontinyu

Aturan ke-	Harga tanah (C1)	Jarak dari pusat kota (C2)	Ada Angkutan Umum (C3)	Dipilih untuk Perumahan (C4)
1	100	2	Tidak	Ya
2	200	1	Tidak	Ya
3	500	3	Tidak	Ya
4	6001	20	Tidak	Tidak
5	550	8	Tidak	Tidak
6	250	25	Ada	Tidak
7	75	15	Ada	Tidak
8	80	10	Tidak	Ya
9	700	18	Ada	Tidak
10	100	8	Ada	Ya

Probabilitas kemunculan setiap nilai untuk atribut Harga Tanah (C1)

Tabel 2.7 Probabilitas Harga tanah terhadap Pemilihan Rumah nilai kontinyu

	Ya	Tidak
1	100	600
2	200	550
3	500	250
4	80	75
5	180	700
Mean (μ)	212	435
Deviasi Standar (σ)	168,8787	261,9637

Probabilitas kemunculan setiap nilai untuk atribut Jarak dari Pusat Kota (C2)

Tabel 2.8 Probabilitas Jarak terhadap Pemilihan Rumah nilai kontinyu

	Ya	Tidak
1	2	20
2	1	8
3	3	25
4	10	15
5	8	18
Mean (μ)	4,8	17,2
Deviasi Standar (σ)	3,9623	6,3008

Probabilitas kemunculan setiap nilai untuk atribut Ada Angkutan Umum (C3)

Tabel.2.9 Peluang angkutan umum terhadap Pemilihan Rumah

Angkutan umum	Jumlah kejadian "Dipilih"		Probabilitas	
	Ya	Tidak	Ya	Tidak
Ada	1	3	1/5	3/5
Tidak	4	2	4/5	2/5
Jumlah	5	5	1	1

Probabilitas kemunculan setiap nilai untuk atribut Dipilih untuk perumahan (C4)

Tabel.2.10 Peluang Pemilihan Rumah

Dipilih Untuk Perumahan	Jumlah kejadian "Dipilih"		Probabilitas	
	Ya	Tidak	Ya	Tidak
Jumlah	5	5	1/2	1/2

Apabila diberikan $C1 = 300$, $C2 = 17$, $C3 = \text{Tidak}$, maka:

$$f(C1 = 300|YA) = \frac{1}{\sqrt{2\pi(168,8787)}} e^{-\frac{(300-212)^2}{2(168,8787)^2}} = 0,0021$$

$$f(C1 = 300|TIDAK) = \frac{1}{\sqrt{2\pi(261,9637)}} e^{-\frac{(300-435)^2}{2(261,9637)^2}} = 0,0013$$

$$f(C2 = 17|YA) = \frac{1}{\sqrt{2\pi(3,9623)}} e^{-\frac{(17-4,8)^2}{2(3,9623)^2}} = 0,0009$$

$$f(C2 = 17|TIDAK) = \frac{1}{\sqrt{2\pi(6,3008)}} e^{-\frac{(17-17,2)^2}{2(6,3008)^2}} = 0,0633$$

Sehingga :

$$\begin{aligned} \text{Likelihood Ya} &= (0,0021) \times (0,0009) \times 4/5 \times 5/10 \\ &= 0,000000756. \end{aligned}$$

$$\begin{aligned} \text{Likelihood Tidak} &= (0,0013) \times (0,0633) \times 2/5 \times 5/10 \\ &= 0,000016458. \end{aligned}$$

Nilai probabilitas dapat dihitung dengan melakukan normalisasi terhadap likelihood tersebut sehingga jumlah nilai yang diperoleh = 1

$$\text{Probabilitas Ya} = \frac{0,000000756}{0,000016458+0,000000756} = 0,0439$$

$$\text{Probabilitas Tidak} = \frac{0,000016458}{0,000016458+0,000000756} = 0,9561$$

Maka hasil yang di dapat adalah TIDAK , karena $P(\text{TIDAK}) >$ dari $P(\text{YA})$.

2.1.4. Bayesian Network (BN)

Menurut (Han, 2006) Bayesian Network dapat digunakan untuk menghitung probabilitas keanggotaan dari suatu kelas. Dengan begitu Bayesian Network dapat dikatakan sebagai salah satu penerapan metode klasifikasi pada fungsi prediktif pada teknologi data mining.

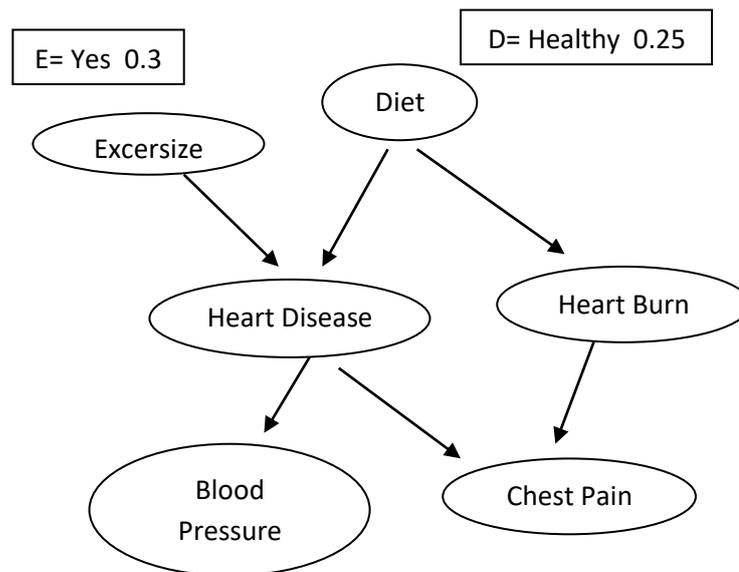
Bayesian Network menyatakan representasi grafis hubungan kausalitas yang berada dalam himpunan variabel acak (Cheng, J,1998). Terdapat 2 (dua) elemen kunci dari Bayesian Network yaitu:

1. Merupakan sebuah directed acyclic graph dimana masing-masing node merepresentasikan sebuah variabel acak, dan masing-masing garis menggambarkan probabilitas ketergantungan dari node sebelumnya (node parent - nya).
2. Memiliki sebuah conditional probability table (CPT) untuk masing-masing node ke node parent -nya .

Asumsi independen bersyarat dibuat oleh *Naive bayes classifier* mungkin terlalu rapuh, khususnya untuk masalah klasifikasi dengan atribut yang dihubungkan dengan sesuatu. Bagian ini menghadirkan pendekatan lebih fleksibel untuk memodelkan peluang kelas bersyarat $P(X|Y)$. Sebagai ganti mensyaratkan seluruh atribut untuk independen secara bersyarat dengan kelas yang diberikan, pendekatan ini menspesifikasi pasangan atribut yang independen secara bersyarat.

Algoritma yang dapat digunakan dalam membangun struktur Bayesian Network dapat dibagi menjadi dua kategori yaitu :

1. algoritma yang menggunakan suatu metode pencarian untuk membangun model dan mengevaluasi model tersebut berdasarkan nilai suatu skor. Proses ini berlanjut dan berhenti jika suatu nilai model yang ke n tidak mempunyai perbedaan yang lebih signifikan yang dibandingkan model yang ke n-1, algoritma ini elisebut juga dengan "*Scoring Based Algorithm*" ;
2. algoritma yang membangun struktur *Bayesian Network* dengan menganalisa hubungan ketergantungan yang terdapat di antar node-node/ variabel *I* atribut. Algoritma ini membangun struktur Bayesian Network dengan mengidentifikasi hubungan bebas bersyarat yang terdapat di antara node-node. Algoritma ini disebut juga dengan "*Constraint Based Algorithm*", contoh algoritma yang termasuk dalam kategori ini adalah TPDA (*Three phase Dependency Analysis*) dan SLA (*Simple Learning Algorithm*) yang dikembangkan oleh Jie Cheng , Russel Greiner dan Jonathan Kelly (2002). Dalam makalah ini penelitian dibatasi hanya untuk mengeksplorasi secara studi literatur algoritma TPDA.



Gambar 2.2. Struktur Bayesian Network Untuk contoh kasus Klasifikasi

Pada gambar di atas tiap-tiap node mempunyai hubungan dengan node yang lain. Dan pada node excersize dan Diet mempunyai idenpendent dan node yang lain tidak

Tabel 2.11. Excerize dan Diet Terhadap HD=Yes

Excersize	Diet	HD=Yes
E= NO	D= Healthy	0.25
E= NO	D=Unhealth	0.45
E= Yes	D= Healthy	0.55
E=Yes	D=Unhealth	0.75

Dapat kita lihat pada struktur di atas untuk mencari P (HD=Yes) Node yang terhubung adalah Excerize dan Diet.

Tabel 2.12. HD terhadap BP=Yes

	BP=High
HD=Yes	0.85
HD= No	0.2

Untuk mencari P (BP=High) Node yang terlibat hanya Heart Disease .

Tabel 2.13. Diet terhadap Hb=Yes

	Bb=Yes
D= Healthy	0.2
D= Unhealth	0.85

Untuk mencari P (Hb=Yes) struktur yang terlibat hanya Melibatkan node Diet.

Table 2.14. HD dan Hb Terhadap CP=Yes

Heart Disease	Heart Burn	CP=Yes
HD= Yes	Hb=Yes	0.2
HD= Yes	Hb=No	0.5
HD= No	Hb=Yes	0.4
HD=No	Hb=No	0.1

Untuk mencari P (CP=Yes) Node yang terhubung adalah Heart Disease dan Heart Burn.

Contoh Inference Menggunakan BN

Kasus 1 : tidak ada informasi terdahulu

Tanpa informasi sebelumnya, dapat ditentukan apakah seseorang memiliki *heart disease* dengan menghitung peluang *prior* $P(HD=Yes)$ dan $P(HD=No)$. Untuk menyederhanakan notasi, $\alpha \in \{Yes, No\}$ melambangkan nilai biner dari *Exercise* dan $\beta \in \{Healthy, Unhealthy\}$ melambangkan nilai biner dari *Diet*.

$$\begin{aligned}
 P(HD = Yes) &= \sum_{\alpha} \sum_{\beta} P(HD = Yes | E = \alpha, D = \beta) P(E = \alpha, D = \beta) \\
 &= \sum_{\alpha} \sum_{\beta} P(HD = Yes | E = \alpha, D = \beta) P(E = \alpha) P(D = \beta) \\
 &= 0.25 \times 0.7 \times 0.25 + 0.45 \times 0.7 \times 0.75 + 0.55 \times 0.3 \times 0.25 + \\
 &\quad 0.75 \times 0.3 \times 0.75
 \end{aligned}$$

$$= 0.49$$

Karena $P(HD=no) = 1 - P(HD=yes)=0.51$, orang tersebut besar kemungkinan tidak terkena penyakit tersebut.

Kasus 2 : tekanan darah tinggi

Jika seseorang memiliki tekanan darah tinggi, dapat dilakukan diagnosa penyakit hati dengan membandingkan peluang posterior $P(HD = Yes|BP=High)$ dengan $P(HD = No|BP=High)$. Untuk melakukan ini, harus dihitung $P(BP=High)$.

$$\begin{aligned} P(BP = High) &= \sum_{\gamma} P(BP = High|HD = \gamma)P(HD = \gamma) \\ &= 0.85 \times 0.49 + 0.2 \times 0.51 = 0.5185. \end{aligned}$$

dengan $\gamma \in \{Yes, No\}$. Oleh karena itu, peluang *posterior* seseorang memiliki penyakit hati adalah :

$$\begin{aligned} P(HD = Yes|BP = High) &= \frac{P(BP = High|HD = Yes)P(HD = Yes)}{P(BP = High)} \\ &= \frac{0.85 \times 0.49}{0.5185} = 0.8033. \end{aligned}$$

Dengan cara yang sama, $P(HD = No|BP=High) = 1 - 0.8033 = 0.1967$. Oleh karena itu, ketika seseorang memiliki tekanan darah tinggi, maka resiko terkena penyakit hati akan meningkat.

Kasus 3 : tekanan darah tinggi, diet sehat dan olahraga teratur

Jika diberitahu bahwa orang tersebut melakukan olahraga teratur dan makan dengan pola diet yang sehat. Bagaimana informasi baru mempengaruhi diagnosa? Dengan informasi baru tersebut, peluang *posterior* bahwa seseorang terkena penyakit hati adalah :

$$P(HD = Yes|BP = High, D = Healthy, E = Yes)$$

$$= \left[\frac{P(BP = High | HD = Yes, D = Healthy, E = Yes)}{P(BP = High | D = Healthy, E = Yes)} \right]$$

$$xP(HD = Yes | D = Healthy, E = Yes)$$

$$= \frac{P(BP = High | HD = Yes)P(HD = Yes | D = Healthy, E = Yes)}{\sum_{\gamma} P(BP = High | HD = \gamma)P(HD = \gamma | D = Healthy, E = Yes)}$$

$$= \frac{0.85 \times 0.25}{0.85 \times 0.25 + 0.2 \times 0.75}$$

$$= 0.5862$$

sedang peluang bahwa seseorang tidak terkena penyakit hati adalah :

$$P(HD=No/BP=High, D=Healthy, E=Yes) = 1 - 0.5862 = 0.4138.$$

Model tersebut selanjutnya menyatakan bahwa dengan pola makan sehat dan melakukan olahraga teratur akan mengurangi resiko penyakit hati.

2.1.5. Tiroid

Tubuh kita terdiri dari sistem-sistem yang memiliki peran yang sangat penting. Salah satunya sistem endokrin. Sistem ini meliputi sistem dalam tubuh manusia yang terdiri dari beberapa kelenjar penghasil hormone, disebut dengan sistem atau kelenjar endokrin. Kelenjar tiroid termasuk dalam sistem endokrin

Tiroid bisanya kita kenal sebagai kelenjar gondok (thyroid glan) letaknya di bagian depan di bawah jakun. Kelenjar kecil ini bisanya tidak terlihat karena tertutup oleh kulit luar. Kelenjar tiroid merupakan pengendali utama metabolisme tubuh. Kelenjar ini bertugas menghasilkan, menyimpan, dan melepaskan hormone tiroid ke dalam peredaran darah. kelenjar tiroid diatur oleh hormon perangsang tiroid (TRH) yang disekresikan oleh hipofisis anterior. Hormon yang disekresikan kelenjar tiroid yaitu tiroksin (sekitar 90%) dan triiodotironin (sekitar 10%).(heryati dan faizah, 2008)

Kelenjar tiroid berfungsi mengatur proses oksidasi, pengeluaran karbondioksida, pada anak mempengaruhi perkembangan fisik dan mental, kelenjar ini juga mensekresi hormon tiroksin (T₄) dan triiodotironin (T₃) yang berasal dari yodium yang masuk ke dalam tubuh dari makanan dan air minum (syarifudin, 2006)

1. Sejarah penyakit tiroid

Di mulai tahun 1800-an, pembesaran tiroid dianggap sebagai suatu pembengkakan dari saluran napas di leher. Bahasa latin menamakan tumid guttur, yang berarti swollen throat, atau pembekakan dari tenggorokan. Istilah ini selanjutnya di terjemahkan ke dalam bahasa Prancis goiter, bahasa Inggris goiter juga, dan orang Amerika menyebutnya goiter. Di Indonesia, para dokter ikut memanggil semua pembesaran kelenjar gondok sebagai goiter (Tandra : 2009).

Ada juga yang memberi nama struman, yang artinya tumefaction atau pembengkakan. Pada tahun 1896, Baumann memastikan ada bahan penting yang terkandung di dalam tiroid. Lalu Kendall (1915) berhasil menemukan hormon tiroksin (T₄) dari tiroid, dan Harington dan Barger (1926) mendapatkan susunan kimia dari hormon T₄ itu. Penelitian-penelitian terus dilakukan, bermacam-macam kelainan tiroid ditemukan. Tahun 1825 ditemukan oleh Parry, tahun 1835 oleh Graves, dan tahun 1840 oleh Von Basedow, yang melaporkan penyakit yang ternyata sama, yaitu yang dikenal sebagai hipertiroid.

Kebanyakan dari kita pasti pernah mendengar atau melihatnya, yaitu penderita akan mengalami jantung berdebar-debar, badan semakin kurus, gemetaran, serta mata terbelalak. Penyakit ini disebut Parry's disease, Graves' disease, atau Von Basedow's disease. Ada juga radang tiroid (tiroiditis) yang ditemukan oleh Riedel (1896), Hashimoto (1912), dan De Quervain (1936). Berbagai kanker tiroid juga mulai diketahui pada akhir abad 19. Di tahun 1909, Theodor Kocher, seorang ahli bedah di Swiss, menjadi dokter pertama yang berhasil mengoperasi tiroid pada pasien goiter. Dia membuktikan peran hormon tiroid pada pertumbuhan dan berbagai fungsi organ tubuh. Prestasi hebat ini membuat Dr. Kocher dianugerahi hadiah Nobel di bidang kedokteran (Tandra : 2009)

Tabel 2.15.. Perbandingan Gejala Hipotiroid dan Hipertiroid

Hipotiroid	Hipertiroid
Berat badan bertambah	Berat Badan menurun
Denyut nadi lambat	Denyut nadi cepat
Kulit kering	Kulit basah
Menstruasi deras Kolesterol tinggi	Menstruasi jarang Kolesterol rendah
Suhu tubuh turun (dingin)	Suhu tubuh naik (panas)

2. Anatomi tiroid yang normal

Kelenjar tiroid pada orang dewasa ukurannya sekitar 5 cm, beratnya kira-kira 10-20 gram. Letak kelenjar ini di leher bagian depan, tepat di bawah jakun, di depan trakea. Jakun adalah tempatnya pita suara. Trakea adalah jalan napas. Bentuk kelenjar ini menyerupai huruf H atau seperti dasi kupu-kupu. Dalam keadaan normal, tiroid ini tidak terlihat dan hampir tidak teraba. Namun, jika membesar, dokter bisa merabanya dengan mudah, tampak ada benjolan di bawah jakun, yang bergerak naik ketika kita menelan sesuatu.

Anda bisa memeriksanya sendiri. Berdirilah di hadapan cermin, minum segelas air, kemudian lihat gerakan leher ketika Anda menelan air itu. Benjolan tiroid akan terlihat jelas bergerak ke atas pada saat menelan.

Tiroid terdiri dari dua bagian, yaitu kedua sayap kupu-kupu, yang dinamakan lobus (lobe). Lobus kanan (right lobe) dan lobus kiri (left lobe), yang dihubungkan oleh istmus (isthmus). Untuk membuat hormon tiroid diperlukan mineral yodium. Yodium bersumber dari makanan dan air yang kita konsumsi setiap hari, kemudian diserap masuk ke dalam peredaran darah. Sel folikel menangkap yodium, lalu menggandengkan yodium dengan asam amino tirosin, sehingga terbentuklah yodotirosin.

Mulanya terjadi mono-yodo-tirosin (satu yodium plus tirosin), disingkat T1. Selanjutnya terbentuk di-yo-do-tirosin (dua yodium plus tirosin), atau T2. Proses selanjutnya adalah penggabungan T1 dan T2 menjadi T3, atau tri-yodo-tirosin. Bisa

juga dua T₂ menyatu menjadi T₄, atau tetra-yodo-tirosin. Nama lain dari T₄ adalah ti-roksin (thyroxine).

Akhirnya, T₃ dan T₄ akan diikat oleh protein tiroglobulin (thyroglobulin), dan ditimbun sebagai koloid di dalam folikel. Ketika tubuh memerlukan hormon, maka tiroglobulin akan dilepas dari ikatannya, sehingga T₃ dan T₄ masuk ke dalam peredaran darah, menunaikan tugasnya sebagai hormon yang penting bagi tubuh. Setelah hormon tiroid digunakan, beberapa yodium akan kembali ke tiroid untuk didaur-ulang menghasilkan hormon tiroid kembali. Sumber yodium banyak terdapat di laut, misalnya ikan dan hasil laut lainnya. Telur, susu, dan produk susu biasanya juga diberi yodium dalam pengolahannya. Setiap hari Anda membutuhkan yodium sebanyak 150 mg. Wanita hamil perlu lebih banyak, yaitu 220 mg. Ibu yang menyusui bahkan harus mendapatkan yodium 290 mg setiap harinya.

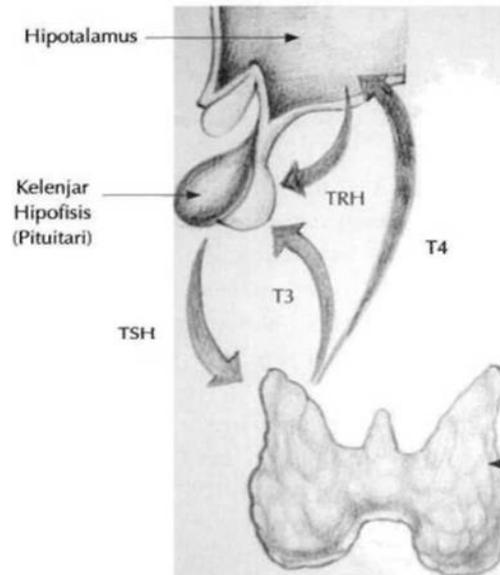
Ada Kelenjar Lain yang Mengatur

Siklus produksi hormon dikendalikan oleh beberapa kelenjar:

1. Kelenjar hipotalamus
2. Kelenjar hipofisis (kelenjar pituitari)
3. Kelenjar tiroid.

Kelenjar hipotalamus terletak di dalam otak, tepat di atas kelenjar hipofisis. Kelenjar ini menghasilkan thyrotrophin-releasing hormone (TRH). Hormon TRH ini me-rangsang kelenjar hipofisis untuk memproduksi thyroid-stimulating hormone (TSH). Sesuai dengan namanya, TSH merangsang kelenjar tiroid untuk mengeluarkan T₃ dan T₄. Jika tubuh kekurangan T₃ dan T₄, timbul umpan-balik yang memicu hipotalamus untuk mengeluarkan TRH, sehingga hipofisis melepaskan lebih banyak TSH dengan akibat T₃ dan T₄ dikeluarkan. Sebaliknya, kalau T₃ dan T₄ di dalam tubuh sudah berlebihan, TRH dan TSH akan ditekan, dan pembentukan T₃ dan T₄ menjadi berkurang.

Sebenarnya hormon yang paling banyak dikeluarkan oleh tiroid adalah T₄. Namun, hormon tiroid yang paling menentukan metabolisme tubuh adalah T₃. T₃ ini adalah bentuk hormon yang aktif. Kurang lebih 80 persen dari T₃ diperoleh melalui konversi (perubahan) T₄ menjadi T₃. Perubahan ini berlangsung di hati, ginjal, dan otot. Hanya 20 persen dari T₃ yang dikeluarkan oleh tiroid.



Gambar 2.3 . Pembentukan Hormon Tiroid

1. Hipotalamus mengirim TRH ke hipofisis
2. Hipofisis membuat TSH
3. TSH merangsang tiroid membentuk T3 dan T4
4. T3 dan T4 dalam peredaran darah, melapor kembali untuk hipotalamus dan hipofisis.

3. Hormon Berikatan dengan Protein

Di dalam peredaran darah biasanya hormon berikatan dengan protein. Bentuk yang berikatan ini sebenarnya adalah bentuk hormon yang tidak aktif. Mengapa harus berikatan? Ada yang berpendapat bahwa molekul ikatan ini lebih stabil, tidak gampang rusak, dan ikatannya hanya dilepaskan jika hormon betul-betul diperlukan. Pendapat lain adalah karena molekul ikatan ini lebih besar, sehingga tidak bisa dibuang lewat air kencing.

Hormon T3 dan T4 yang diikat protein jumlahnya sebanyak 99,97 persen, sisanya 0,03 persen adalah hormon yang aktif, yang disebut hormon tiroid yang bebas, yaitu Free T3 (FT3) dan Free T4 (FT4). Jika dokter menyuruh Anda memeriksakan T3 dan T4, maksudnya adalah hormon yang terikat dengan protein. Anda juga bisa langsung memeriksakan FT3 dan FT4, jika ingin tahu berapa kadar hormon yang aktif (yang bebas).

Ada tiga macam protein yang bisa mengikat hormon. Sebanyak 75 persen hormon diikat oleh globulin, yang di-namakan thyroxine-binding globulin (TBG). Sisanya adalah prealbumin (thyroxine-binding prealbumin) mengikat kurang lebih 20 persen, dan albumin (thyroxine-binding albumin) yang bersenyawa dengan 5 persen hormon.

4. T3 dan T4

Beberapa tahun yang lalu, hanya ada pemeriksaan T3 total dan T4 total. Artinya yang dihitung adalah T3 yang bebas (FT3) dan T3 yang berikatan dengan protein, serta T4 yang bebas (FT4) dan yang berikatan. Kadang hormon yang bebas masih normal, tetapi kadar hormon total di dalam darah tinggi. Keadaan ini bisa disebabkan oleh lebih banyaknya hormon yang berikatan dengan protein. Sehingga, pengukuran hormon lewat pemeriksaan T3 dan T4 total menjadi tidak akurat lagi, dan cara kuno ini telah mulai ditinggalkan.

T3 dan T4 total akan meningkat akibat beberapa keadaan yang membuat ikatan dengan protein bertambah, atau lantaran mengonsumsi obat-obatan tertentu. Misalnya:

1. Kehamilan, hormon estrogen meningkat
2. Penyakit-penyakit akut, seperti hepatitis atau AIDS
3. Penyakit jiwa yang akut
4. Pemakaian hormon estrogen, pil KB
5. Amiodarone, obat untuk gangguan irama jantung
6. Obat amphetamine
7. Obat methadone.

Sebaliknya, keadaan-keadaan tertentu bisa membuat kadar T3 dan T4 total menjadi rendah. tiroid, sedangkan jika FT4 rendah menandakan Anda terkena hipotiroid.

Pada keadaan tertentu, hipertiroid bisa disebabkan oleh T3 yang tinggi, sehingga FT4 masih normal bahkan rendah. Sekalipun demikian, pemeriksaan FT4 ini bukanlah seratus persen sempurna, karena masih ada beberapa keadaan yang bisa membuat hasil FT4 berubah. Misalnya penderita penyakit yang menahun cenderung memiliki FT4 yang lebih rendah. Pasien dengan penyakit akut akan membuat FT4 naik sesaat.

Pemakai obat heparin untuk mencegah pembekuan darah bisa sedikit meningkatkan FT4. Akan tetapi, keadaan-keadaan seperti itu dapat se-gera diketahui oleh dokter yang telah berpengalaman. Pemeriksaan Free T3 (FT3) tidak selalu diperlukan. Tes darah ini biasanya ditujukan untuk pasien hipertiroid yang FT4-nya normal. Kadar normal FT3 adalah 0,25-0,65 ng/dl. Pasien hipertiroid mempunyai FT3 yang tinggi. Pada penderita hipotiroid akan ditemukan FT3 yang rendah.

5. TSH

Thyroid-Stimulating Hormone (TSH) adalah hormon yang diproduksi oleh kelenjar hipofisis atau pituitari. Ketika hormon tiroid yang beredar di dalam darah me-nurun, TSH akan banyak dikeluarkan. Sebaliknya, jika kebanyakan hormon tiroid, pembentukan TSH akan di-kurangi.

2.1.6 Pengujian K-Fold Cross Validation

Dalam penelitian ini dilakukan pengujian metode dengan dengan menggunakan cross validation. Cross validaton (Redaelzadeh, Tang, & Liu, 2008) adalah metode statistic untuk mengevaluasi dan membandingkan belajar algoritma dengan membagi data menjadi dua segmen, satu segmen digunakan untuk belajar atau melatih data dan yang lain digunakn untuk memvalidasi model. Dalam cross validation set pelatihan dan validasi harus crossover berturut-turut sehingga setiap data memiliki kesempatan tervalidasi.

Tabel 2.16. K-flod cross validation

Data Set									
split 1	split 2	split 3	split 4	split 5	split 6	split 7	split 8	split 9	split 10
Traning									test
Traning								test	
Traning						test			
Traning					test				
Traning				test					
			test	Traning					
		test	Traning						
	test	Traning							
test	Traning								

2.1.7 Evaluasi dan Validasi Metode Prediksi *Data Mining*

Untuk menguji model, pada penelitian ini, digunakan *Confusion Matrix*, dan kurva ROC (*Receiver Operating Characteristic*).

1. *Confusion matrix*

Confusion matrix memberikan keputusan yang diperoleh dalam *training* dan *testing* (Bramer, 2007), *confusion matrix* memberikan penilaian *performance* klasifikasi berdasarkan objek dengan benar atau salah (Gorunescu, 2011)

Metode ini menggunakan tabel matriks seperti pada Tabel 2.2, jika data set hanya terdiri dari dua kelas, kelas yang satu dianggap sebagai positif dan yang lainnya negatif (Bramer, 2007).

Tabel 2.17. Model *Confusion Matrix*

Klasifikasi yang benar	Diklasifikasikan sebagai	
	+	-
+	<i>True positives</i>	<i>False negatives</i>
-	<i>False positives</i>	<i>True negatives</i>

True positives adalah jumlah *record* positif yang diklasifikasikan sebagai positif, *false positives* adalah jumlah *record* negatif yang diklasifikasikan sebagai positif, *false negatives* adalah jumlah *record* positif yang diklasifikasikan sebagai negatif, *true negatives* adalah jumlah *record* negatif yang diklasifikasikan sebagai *negative*.

Setelah melakukan *Confusion matrix*, selanjutnya menghitung nilai akurasi, yaitu:

$$\text{True Positif} + \text{True Negatif}$$

$$\text{Accuracy} = \frac{\text{True Positif} + \text{True Negatif}}{\text{True Positif} + \text{True Negatif} + \text{False Positif} + \text{False negatif}}$$

$$\text{True Positif} + \text{True Negatif} + \text{False Positif} + \text{False negatif}$$

2. Kurva ROC

Untuk dapat melihat akurasi secara manual dilakukan perbandingan klasifikasi menggunakan kurva ROC hasil eksperisi dari *confusion matrix*. Kurva ROC menunjukkan akurasi dan membandingkan klasifikasi secara visual. ROC mengekspresikan *confusion matrix*. ROC adalah grafik dua dimensi dengan *false positives* sebagai garis horisontal dan *true positives* sebagai garis vertikal (Vecellis, 2009). *The area under curve* (AUC) dihitung untuk mengukur perbedaan performansi metode yang digunakan.

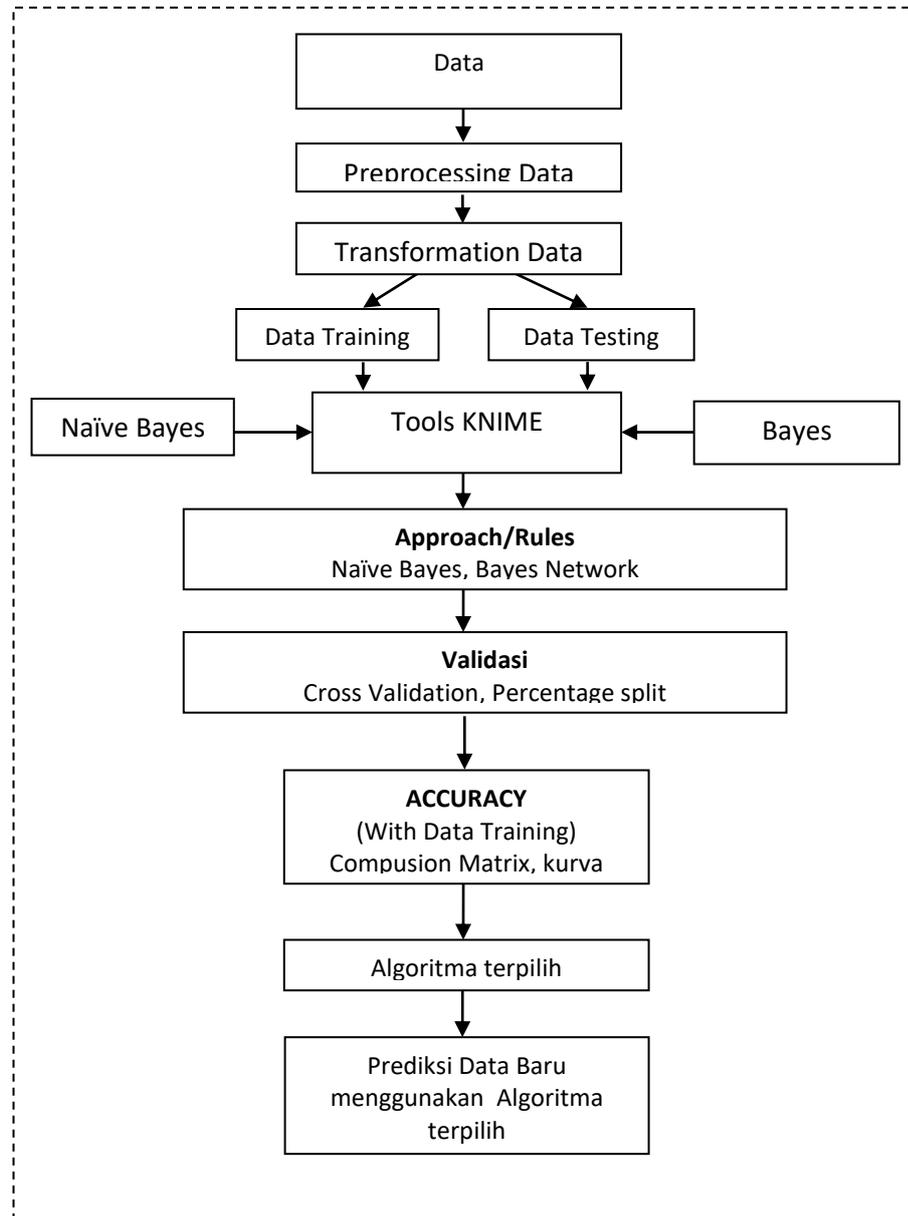
2.2. Tinjauan Studi

Penelitian sebelumnya sebagai bahan untuk tinjauan studi penulis diantaranya:

1. Banyak penelitian prediksi penyakit tiroid dengan teknik klasifikasi data mining diantaranya dilakukan oleh Aswani Kumar dengan menggunakan metode Decision Tree pada 2012 yang melakukan evaluasi dan komporasi dalam mendeteksi penyakit Tiroid dengan menggunakan decision tree, dengan membedakan splitting rule (gini index, information gain, chi square, gain ratio, Measure distance)
2. Penelitian lain oleh Dr. Sarah Behnam Aziz dengan menggunakan metode Genetic Algorithm dan NN untuk memprediksi penyakit tiroid dengan melakukan splitting random (10% to 90%, 30% to 70% , and 50%).
3. Penelitian lain tentang prediksi penyakit tiroid juga di lakukan oleh M. R. Nazari Kousarrizi, F.Seiti, and M. Teshnehlab pada tahun 2012 menggunakan metode SVM , dengan menggabungkan data lama penelitian dengan data medis rumah sakit baru
4. Penelitian lain dilakukan oleh (Mamani, 2012) dengan 340 data dan 7 attribut untuk memprediksi penyakit liver dengan hasil J48 64.49% LMT 71.00% Bayes Net 65.08% Naïve Bayes 65.58% MLP (2 hidden layer) 65.68% RBF 65.68% Rough Set 65.08%

2.3. Kerangka Pemikiran

Penelitian ini terdiri dari beberapa tahap seperti terlihat pada kerangka pemikiran Gambar 2.6.



Gambar 2.4. Kerangka pemikiran

BAB 3

METODOLOGI PENELITIAN

3.1. Desain Penelitian

3.1.1. Jenis Penelitian

Desain penelitian yang digunakan dalam penelitian ini adalah dengan menggunakan metode eksperimen komparatif yaitu membandingkan dua objek yang berbeda, misalnya membandingkan dua algoritma yang berbeda dengan melihat hasil statistik masing-masing mana yang lebih baik. Yang bertujuan untuk menguji kebenaran sebuah hipotesa dengan perhitungan statistik dapat memberikan solusi terhadap permasalahan dalam penelitian ini (Kothari, 2004). Penelitian ini bertujuan untuk melakukan komparasi dan evaluasi metode Naïve Bayes dan Bayes Network untuk mengetahui performance mana yang lebih baik dalam mengidentifikasi penyakit tiroid.

3.1.2. Metode Pengumpulan Data

Ada dua tipe dalam pengumpulan data, yaitu pengumpulan data primer dan pengumpulan data sekunder. Data primer adalah data yang dikumpulkan pertama kali dan untuk melihat apa yang sesungguhnya terjadi. Data sekunder adalah data yang sebelumnya pernah dibuat oleh seseorang baik di terbitkan atau tidak (C.R.Kothari, 2004). Data yang digunakan dalam penelitian ini adalah data sekunder yang didapat dari UCI (University of California, Irvine) *Machine Learning Repository*. dan sumber data dari the Garavan Institute and J. Ross Quinlan, New South Wales Institute, Sydney, Australia. Dalam pengumpulan data diperoleh dari internet untuk di jadikan objek penelitian dan mencari data tambahan melalui buku-buku, jurnal, publikasi dan lain-lain untuk di jadikan rujukan penulisan dan penelitian.

3.2. Tahap-tahap Penelitian

Pada tahapan penelitian dibagi menjadi 5 tahapan antara lain pengumpulan data, pengolahan awal data, model atau metode yang diusulkan

atau dikembangkan, eksperimen dan pengujian model atau metode dan yang terakhir evaluasi dan validasi hasil.

1. Pengumpulan Data

Ada dua tipe dalam pengumpulan data, yaitu pengumpulan data primer dan pengumpulan data sekunder. Data primer adalah data yang dikumpulkan pertama kali dan untuk melihat apa yang sesungguhnya terjadi. Data sekunder adalah data yang sebelumnya pernah dibuat oleh seseorang baik di terbitkan atau tidak (C.R.Kothari, 2004). Dalam pengumpulan data sumber dapat di peroleh dari internet untuk di jadikan objek penelitian dan mencari data tambahan melalui buku-buku, jurnal, publikasi dan lain-lain untuk di jadikan rujukan penulisan dan penelitian.

2. Pengolahan awal data.

Dilakukan penyeleksian data, data dibersihkan dan ditransformasikan bentuk yang diinginkan sebelum dilakukan pembuatan model

3. Model atau metode yang diusulkan atau dikembangkan.

Data yang diteliti dianalisa kemudian di kelompokkan variabel mana yang berhubungan dengan satu sama lainnya, lalu dibuatkan model yang sesuai dengan jenis data. Pembagian data kedalam data latihan (training data) dan data uji (testing data) juga diperlukan untuk pembuatan model.

4. eksperimen dan pengujian model atau metode

pengujian model diusulkan pada model yang akan diuji untuk melihat hasil berupa rule yang akan dimanfaatkan dalam mengambil keputusan hasil penelitian.

5. Evaluasi dan validasi hasil

Pada penelitian ini dilakukan evaluasi terhadap model yang ditetapkan untuk mengetahui tingkat keakurasian model.

3.2.1. Pengumpulan Data

Dalam penelitian ini, pengumpulan data dari Institut Garavan dan J. Ross, Quinlan, New South Wales Institute, Sydney, Australia. Mengenai data pasien yang terkena penyakit tiroid data yang di dapat sebanyak 3711 pasien yang di periksa dan sebanyak 3318 terdeteksi normal, 102 pasien teridentifikasi menderita hipertiroid dan 291 pasien teridentifikasi menderita hipotiroid. Dengan

atribut dari setiap penyakit Tiroid yang diperiksa adalah. Age, sex, onthyroxine, query onthyroxine, on antithyroid medication, sick, pregnant, thyroid surgery, I131 treatment, query hypothyroid, query hyperthyroid, lithium, goiter, tumor, hypopituitary, psych, TSH, T3, TT4, T4U, FTI. Terdapat 1 atribut special dan 21 atribut regular. Terdapat pada table 3.1.

Table 3.1. Atribut yang digunakan untuk mengidentifikasi penyakit tiroid

Atribut	Tipe
1. Age (Umur) :	continuous.
2. Sex (Jenis Kelamin):	M (male), F (female).
3. On thyroxine (pemberian thyroxine secara berkala) :	f (false) ,t (true)
4. query on thyroxine (Jarang memberikan thyroxine):	f (false) ,t (true)
5. on antithyroid medication (pemberian obat thyroid):	f (false) ,t (true)
6. sick (Sakit pada leher):	f (false) ,t (true)
7. pregnant (Apakah Pasien wanita Hamil):	f (false) ,t (true)
8. thyroid surgery (operasi tiroid):	f (false) ,t (true)
9. I131 treatment (pengobatan I131):	f (false) ,t (true)
10. query hypothyroid (dugaan hypotiroid):	f (false) ,t (true)
11. query hyperthyroid (dugaan hipertiroid):	f (false) ,t (true)
12. lithium (pemberian lithium):	f (false) ,t (true)
13. goiter (terdapat gondok):	f (false) ,t (true)
14. tumor (terdapat tumor):	f (false) ,t (true)
15. hypopituitary (terdapat Gejala Hypopituitarism) :	f (false) ,t (true)
16. psych (perubahan fisik):	f (false) ,t (true)
17. TSH (<i>Thyroid-stimulating hormone</i>):	f (false) ,t (true)
18. T3 (Triiodothyronine):	continuous.
19. TT4 Total thyroxine):	continuous.
20. T4U <i>T4 uptake</i>):	continuous.
21. FTI (Free Thyroxine Index):	continuous.
22. Class	Negative, hipertiroid, hypotiroid

Tabel 3.2. Data pasien penyakit tiroid maupun yang tidak.

No	age	sex	on	tl	quer	on	ar	sick	pregn	thyr	c131	tquery	quer	lithiu	goitr	tumo	hypo	psycl	TSH	T3	TT4	T4U	FTI	Class
1	66	F	f	f	f	f	f	f	f	f	f	f	f	f	f	t	f	f	0.6	2.2	123	0.93	132	negative
2	68	M	f	f	f	f	f	f	f	f	f	f	f	f	f	f	f	f	2.4	1.6	83	0.89	93	negative
3	62	M	t	f	f	f	f	f	f	f	t	t	f	f	f	f	f	f	13	1.1	87	0.81	107	negative
4	70	F	f	f	f	f	f	f	f	f	f	f	f	f	f	f	f	f	0.005	2.6	160	0.88	180	hyperthyroid
4	21	M	f	t	f	f	f	f	f	f	f	f	f	f	f	t	f	f	0.25	1.3	54	1.07	51	negative
5	90	F	f	t	f	f	f	f	f	f	f	f	f	f	f	f	f	f	0.03	1.9	203	0.9	224	negative
6	60	M	t	f	f	f	f	f	f	f	f	f	f	f	f	f	f	f	13	1.4	57	0.62	92	negative
7	73	F	f	f	f	f	f	f	f	f	f	f	f	f	f	f	f	f	1.9	1.5	113	1.06	106	negative
8	20	F	f	f	t	f	f	f	f	f	f	f	f	f	f	f	f	f	0.02	2.6	106	0.91	116	negative
9	20	F	f	f	t	f	f	f	f	f	f	f	f	f	f	f	f	f	0.15	2.5	93	1	93	negative
10	63	M	f	f	f	t	f	f	f	f	f	f	f	f	f	f	f	f	1.4	1.4	90	0.91	100	negative
11	72	F	t	f	f	t	f	f	f	f	f	f	f	f	f	f	f	f	6.3	1.5	94	0.92	102	negative
12	34	F	f	f	f	f	t	f	f	f	t	f	f	f	f	f	f	f	0.02	3.7	196	1.65	119	negative
13	28	F	t	f	f	f	t	f	f	f	f	f	f	f	f	f	f	f	2.5	2.6	179	1.55	115	negative
14	28	F	f	f	f	f	f	t	f	f	f	f	f	f	f	f	f	f	103	2.1	65	1.16	56	hypothyroid
15	32	F	f	f	f	f	f	t	f	f	f	f	f	f	f	f	f	f	1.8	2.2	101	1.13	90	negative
16	59	M	f	f	f	f	f	f	f	t	f	f	f	f	f	f	f	f	0.3	1.6	112	0.75	148	negative
17	63	F	f	f	f	f	f	t	f	f	f	f	f	f	f	f	f	f	42	1.6	60	0.93	64	hypothyroid
18	34	F	t	f	f	f	f	f	f	f	t	f	f	f	f	f	f	f	0.25	6.7	230	1.73	133	negative
19	35	F	f	f	f	f	f	f	f	t	f	f	f	f	f	f	f	f	2.3	2.4	130	1.11	117	negative
20	66	F	f	f	f	f	f	f	f	f	t	f	f	f	f	f	f	f	0.005	1.5	76	0.85	89	negative
21	53	M	f	f	f	f	f	f	f	f	t	f	f	f	f	f	f	f	3.8	2.5	105	1.04	101	negative
22	44	M	f	f	f	f	f	f	f	f	f	t	f	f	f	f	f	f	3	1.4	103	0.88	118	negative
23	23	F	f	f	f	f	f	f	f	f	f	t	f	f	f	f	t	f	2.1	3.1	116	1.2	96	negative
25	33	F	f	f	f	f	f	f	f	f	f	f	f	f	f	f	f	f	45	1.4	39	1.16	33	hypothyroid
26	25	F	f	f	f	f	f	f	f	f	f	f	f	f	f	f	f	f	0.15	1.7	162	0.87	186	hyperthyroid
27	21	M	f	f	f	f	f	f	f	f	f	f	f	f	f	f	f	f	14.8	1.5	61	0.85	72	hypothyroid
28	33	M	f	f	f	f	f	f	f	f	f	f	f	f	f	f	t	f	1.3	2.2	120	0.96	126	negative
29	64	M	f	f	f	f	f	f	f	f	f	f	f	f	f	f	t	f	1.3	2.4	111	1.15	97	negative

Selengkapnya terdapat pada lampiran

Data yang digunakan dalam penelitian ini sebanyak 3711 Data pasien yang memeriksakan penyakit Tiroid yaitu terdiri dari 102 penderita hipertiroid 291 orang 3318 Orang negative terhadap penyakit tiroid.

Table 3.3. Penderita penyakit hyperteroid

No	age	sex	on	tl	quer	on	ar	sick	pregn	thyr	c131	tquery	quer	lithiu	goitr	tumo	hypo	psycl	TSH	T3	TT4	T4U	FTI	Class
26	63	F	f	f	f	f	f	f	f	f	f	f	f	f	f	f	f	f	0.03	5.5	199	1.05	190	hyperthyroid
49	87	F	f	f	f	f	f	f	f	f	f	f	f	f	f	f	f	f	0.15	1.7	162	0.87	186	hyperthyroid
76	70	F	f	f	f	f	f	f	f	f	f	f	f	f	f	f	f	f	0.005	2.6	160	0.88	180	hyperthyroid
107	58	?	t	f	f	f	f	f	f	f	f	t	f	f	f	f	f	f	0.01	3.3	129	1.23	105	hyperthyroid
115	65	F	f	f	f	f	f	f	f	f	f	f	t	f	f	f	f	f	0.15	3.5	217	0.98	221	hyperthyroid
152	55	F	f	f	f	f	f	f	f	f	f	f	t	f	f	f	f	f	0.01	5.5	219	0.8	274	hyperthyroid
238	52	F	f	f	f	f	f	f	f	f	f	f	f	f	f	f	f	f	0.03	?	163	0.95	172	hyperthyroid
239	19	F	f	f	f	f	t	f	f	f	f	f	f	f	f	t	f	f	0.45	3.2	130	1.83	71	hyperthyroid
341	87	F	f	f	f	f	f	f	f	f	f	t	f	f	f	f	f	f	0.2	3.4	164	0.82	200	hyperthyroid
393	54	?	f	f	f	f	f	f	f	f	t	f	f	f	f	f	f	f	0.01	3.6	142	1.23	115	hyperthyroid
431	49	F	f	f	f	f	f	f	f	f	f	f	f	f	f	f	f	f	0.035	2.1	183	1.02	179	hyperthyroid
445	59	F	f	f	f	f	f	f	f	f	f	f	f	f	f	f	f	f	0.02	1.8	162	0.99	164	hyperthyroid
453	33	?	f	f	f	f	f	f	f	f	f	t	f	f	f	f	f	f	0.015	5	244	0.67	362	hyperthyroid
454	58	F	t	f	f	f	f	f	f	f	f	f	f	f	f	f	f	f	0.065	3.4	129	0.95	136	hyperthyroid
464	73	F	f	f	f	f	f	f	f	f	f	f	f	f	f	t	f	f	0.005	3.6	119	0.9	133	hyperthyroid
539	15	F	f	f	f	f	f	f	f	f	f	f	f	f	f	f	f	f	0.015	6	250	0.8	312	hyperthyroid

selengkapnya terdapat pada lampiran

Telah disebutkan sebelumnya bahwa terdapat 102 penderita hipertiroid , data ini telah di pisahkan dari data pasien yang menderita hypoteroid maupun pasia yang negatif terkena penyakit tiroid .

Table 3.4. Pasien penderita penyakit hypoteroid

No	age	sex	on	tf	quen	on	arsick	pregn	thyr	c1131	t	query	quen	lithiu	goitr	tumo	hypo	psycl	TSH	T3	TT4	T4U	FTI	Class
16	65	F	f	f	f	f	f	f	f	t	f	f	f	f	f	f	f	f	12	?	99	1.14	87	hypothyroid
41	44	M	f	f	f	f	f	f	f	f	f	f	f	f	f	f	f	f	45	1.4	39	1.16	33	hypothyroid
58	65	M	f	f	f	f	f	f	f	f	f	f	f	f	f	f	f	f	14.8	1.5	61	0.85	72	hypothyroid
59	27	F	f	f	f	f	f	f	f	f	f	f	f	f	f	f	f	f	15	1.6	82	0.82	100	hypothyroid
60	54	F	f	f	f	f	f	f	f	f	f	f	f	f	f	f	f	f	19	2.2	83	1.03	81	hypothyroid
86	34	?	f	f	f	f	f	f	f	f	f	f	f	f	t	f	f	f	7.8	2	95	0.99	96	hypothyroid
89	39	F	f	f	f	f	f	f	f	f	f	f	f	f	f	f	f	f	160	0.4	11	1.24	8.9	hypothyroid
102	62	M	f	f	f	t	f	f	f	f	f	f	f	f	f	f	f	f	12	1.5	105	0.86	122	hypothyroid
116	1	?	f	f	f	f	f	f	f	f	f	f	f	f	f	f	f	f	8.8	1.9	113	0.86	131	hypothyroid
117	50	M	f	f	f	f	f	f	f	f	f	f	f	f	f	f	f	f	151	1	32	1.16	28	hypothyroid
122	26	F	f	t	f	f	f	f	f	f	f	f	f	f	f	f	f	f	9.4	?	70	0.95	74	hypothyroid
189	42	F	f	f	f	f	f	f	f	f	f	f	f	f	f	f	f	f	6.1	2.2	74	?	?	hypothyroid
219	59	F	f	f	f	f	f	f	f	f	f	f	f	f	f	f	f	f	11.4	?	99	0.97	102	hypothyroid
223	79	F	f	f	f	f	f	f	f	f	f	f	f	f	f	f	f	f	143	0.7	73	1.01	72	hypothyroid
249	33	F	f	f	f	f	f	f	f	t	f	f	f	f	f	f	f	f	16	1.7	75	?	?	hypothyroid
258	63	M	f	f	f	f	f	f	f	f	f	f	f	f	f	f	f	f	108	0.4	14	0.98	14	hypothyroid
259	19	M	f	f	f	f	f	f	f	t	f	f	f	f	f	f	f	f	9	2.2	117	?	?	hypothyroid
265	55	F	f	f	f	f	f	f	f	f	f	f	f	f	f	f	f	t	9.2	2.5	106	1.13	94	hypothyroid
281	70	M	f	f	f	f	f	f	f	f	f	f	f	f	f	f	f	f	86	0.5	27	0.84	32	hypothyroid
289	74	F	f	f	f	f	f	f	f	f	f	f	f	f	f	f	f	f	9.1	2	72	0.9	80	hypothyroid
293	67	F	f	f	f	f	f	f	f	f	f	f	f	f	f	f	f	f	52	1.2	65	0.92	71	hypothyroid
302	80	M	f	f	f	f	f	f	f	f	f	f	f	f	f	f	f	f	6.1	1.3	83	0.88	95	hypothyroid
304	42	F	f	f	f	f	f	f	f	f	f	f	f	f	f	f	f	f	31	0.8	13	1.19	11	hypothyroid
328	67	F	f	f	f	f	f	f	f	f	f	f	f	f	f	f	f	f	51	0.4	12	1.26	10	hypothyroid

Selengkapnya terdapat pada lampiran

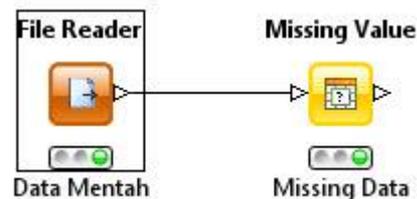
Dari table di atas yang sudah di filter terdapat sebanyak 291 orang yang telah teidentifikasi terkena hypoteroid.

3.2.2. Pengolahan Awal Data

Data yang diperoleh untuk penelitian ini sebanyak 3711 record pemeriksaan penyakit tiroid baik menderita atau tidak menderita. Tetapi dalam data tersebut masih mengandung duplikasi dan alomali atau inkonsisten data maka dengan ini dilakukan replace missing. Untuk mendapatkan data yang berkualitas, beberapa teknik preprocessing digunakan, yaitu: (Vecellis, 2009)

1. data validation, untuk mengidentifikasi dan menghapus data yang ganjil (outlier/noise), data yang tidak konsisten, dan data yang tidak lengkap (missing value)
2. data integration and transformation, untuk meningkatkan akurasi dan efisiensi algoritma.

3. data size reduction and discretization, untuk memperoleh data set dengan jumlah atribut dan record yang lebih sedikit tetapi bersifat informatif. Di dalam data training yang digunakan dalam penelitian ini, dilakukan seleksi atribut dan penghapusan data duplikasi menggunakan software KNIME.



Gambar 3.1. Relasi pengolahan data mentah ke data missing

Dari gambar di atas data yang utuh akan di filter ke dalam data yang siap di olah untuk melakukan training maupun testing

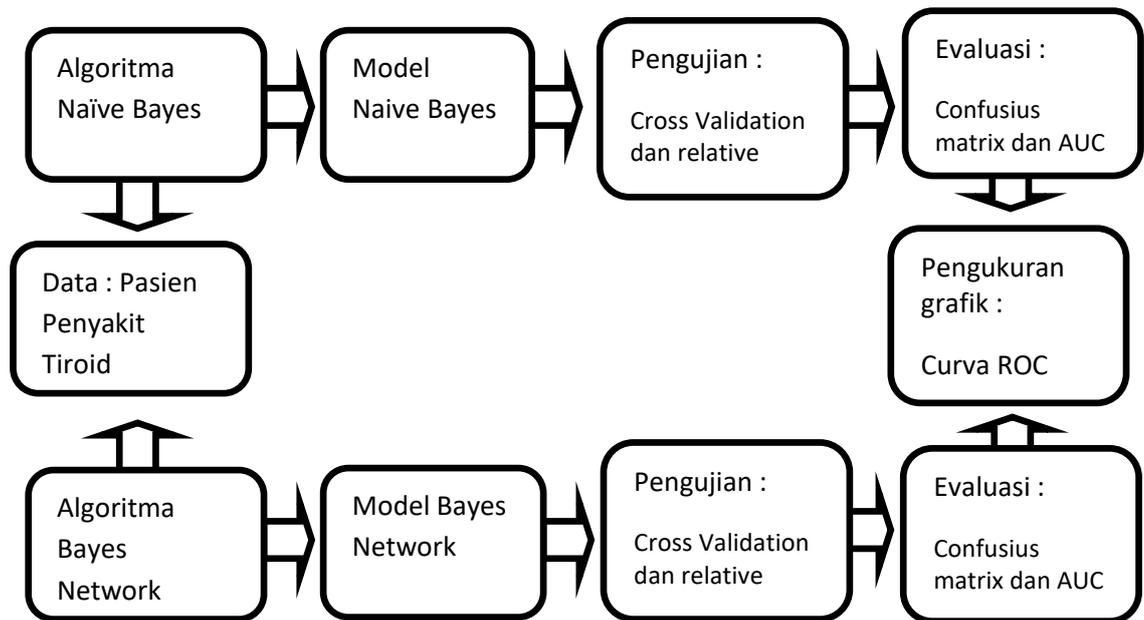
Tabel 3.5. Data yang telah valid

No	age	sex	on	th	query	on	ar	sick	preg	thyro	l131	l	query	query	lithi	u	goitr	ε	tumo	hypo	psych	TSH	T3	TT4	T4U	FTI	Class
1	41	F	f	f	f	f	f	f	f	f	f	f	f	f	f	f	f	f	f	f	f	1.3	2.5	125	1.1	109	negative
2	70	F	f	f	f	f	f	f	f	f	f	f	f	f	f	f	f	f	f	f	f	0.7	1.2	61	0.9	70	negative
3	80	F	f	f	f	f	f	f	f	f	f	f	f	f	f	f	f	f	f	f	f	2.2	0.6	80	0.7	115	negative
4	66	F	f	f	f	f	f	f	f	f	f	f	f	f	f	f	t	f	f	f	f	0.6	2.2	123	0.9	132	negative
5	68	M	f	f	f	f	f	f	f	f	f	f	f	f	f	f	f	f	f	f	f	2.4	1.6	83	0.9	93	negative
6	84	F	f	f	f	f	f	f	f	f	f	f	f	f	f	f	t	f	f	f	f	1.1	2.2	115	1	121	negative
7	71	F	f	f	f	f	t	f	f	f	f	f	t	f	f	f	f	f	f	f	f	0	3.8	171	1.1	151	negative
8	59	F	f	f	f	f	f	f	f	f	f	f	f	f	f	f	f	f	f	f	f	2.8	1.7	97	0.9	107	negative
9	28	M	f	f	f	f	f	f	f	f	f	f	f	f	f	f	f	f	f	f	f	3.3	1.8	109	0.9	119	negative
10	63	F	f	f	f	f	f	f	f	f	f	f	f	f	f	f	f	f	f	f	f	1.5	1.2	117	1	121	negative
11	55	M	f	f	f	f	f	f	f	f	f	f	f	f	f	f	f	f	f	f	f	0.2	1.8	134	1	131	negative
12	63	F	f	f	f	f	f	f	f	f	f	f	f	f	f	f	f	f	f	f	f	0	5.5	199	1.1	190	hyperthyroid
13	60	M	t	f	f	f	f	f	f	f	f	f	f	f	f	f	f	f	f	f	f	13	1.4	57	0.6	92	negative
14	58	F	f	f	f	f	f	f	f	f	f	f	f	f	f	f	f	f	f	f	f	0.3	2.7	120	1.3	95	negative
15	44	M	f	f	f	f	f	f	f	f	f	f	f	f	f	f	f	f	f	f	f	45	1.4	39	1.2	33	hypothyroid
16	48	M	f	t	f	f	f	f	f	f	f	t	f	f	f	f	f	f	f	f	f	5.4	1.9	87	1	87	negative
17	61	F	f	f	f	f	f	f	f	f	f	f	f	f	f	t	f	f	f	f	f	1	1.5	63	0.6	113	negative
18	83	F	f	f	f	f	f	f	f	f	f	f	f	f	f	f	f	f	f	f	f	0.9	0.9	86	0.8	113	negative
19	87	F	f	f	f	f	f	f	f	f	f	f	f	f	f	f	f	f	f	f	f	0.2	1.7	162	0.9	186	hyperthyroid
20	78	F	f	f	f	f	f	f	f	f	t	f	f	f	f	f	f	f	f	f	f	0.6	1.7	113	1	115	negative

Setelah dilakukan preprocessing data yang didapat data sebanyak 3711 mengalami missing value sebanyak 1062 record dan menjadi data yang valid sebanyak 2642 record.

3.2.3. Model atau metode yang diusulkan atau dikembangkan.

Pada tahap modeling ini dilakukan pemrosesan data training sehingga akan membahas metode algoritma yang diujikan dengan memasukan data pasien penyakit tiroid kemudian di analisa dan dikomparasi. Berikut ini bentuk gambaran metode algoritma yang akan diuji.



Gambar 3.2. Model yang yang diusulkan

Tahap-tahap proses yang dilakukan pada tahap modeling ini adalah sebagai berikut :

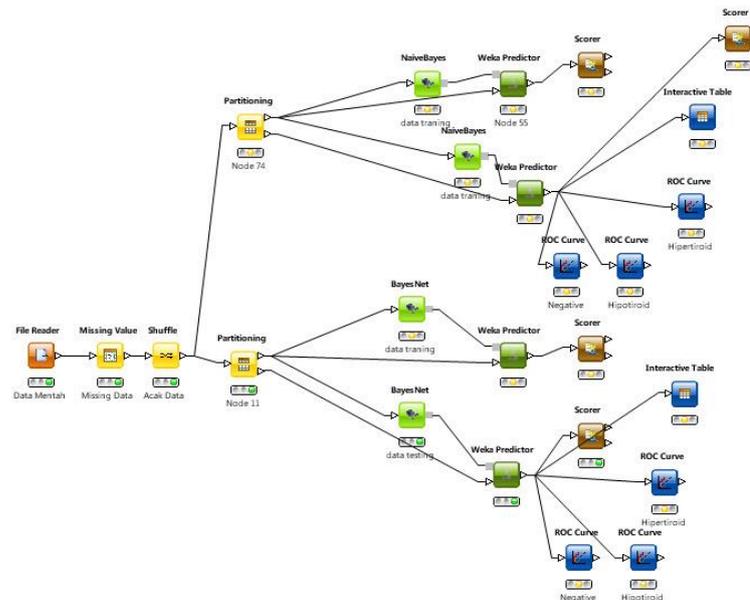
1. Melakukan prediksi penyakit tiroid dengan menggunakan dua metode yaitu algoritma Naive Bayes dan algoritma Bayes Network
2. Melakukan Pengujian yang akan dilakukan melalui dua jenis pengujian, yang pertama dengan uji coba sebanyak sepuluh kali untuk masing-masing teknik atau disebut 10 folds cross validation. 10 folds cross validation digunakan karena pengujian dengan cara ini sudah sering dipakai oleh banyak penelitian dalam bidang data mining. Menurut Kohavi (1995) 10 folds cross validation merupakan merupakan pengujian yang memiliki tingkat akurasi yang paling baik dalam pengklasifikasian. Dan yang kedua menggunakan split data sebesar 60 % - 40% dengan data acak.
3. Melakukan evaluasi.

Pada tahap ini dilakukan pengujian terhadap model-model yang dikomparasi untuk mendapatkan informasi model yang paling akurat. Evaluasi dan validasi menggunakan metode cross validation, confusion matrix, dan, kurva ROC.

4. Melakukan pengukuran dengan grafik dengan menggunakan Kurva ROC

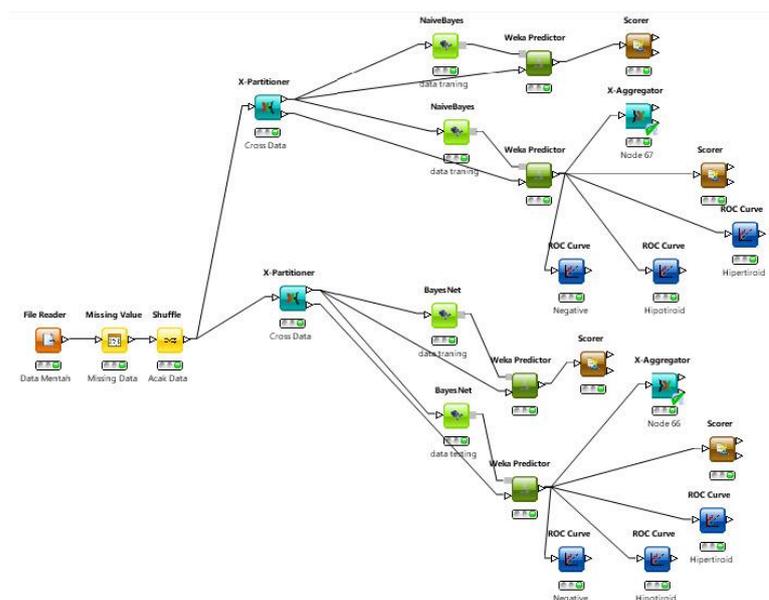
3.2.4. Eksperimen dan pengujian model atau metode

Pada tahapan ini akan dilakukan pengujian dengan menggunakan software KNAME, yang akan mengolah data dengan membandingkan dua metode, berikut gambar di bawah ini.



Gambar 3.3. Model split persentase untuk membandingkan dua metode

Gambar di atas adalah model uji data yang menggunakan split persentase dengan menggunakan metode Naïve Bayes dan Bayes Network.



Gambar 3.4. Model Cross Validation untuk membandingkan dua metode

Gambar di atas adalah model uji data yang menggunakan Cross Validation dengan menggunakan metode Naïve Bayes dan Bayes Network.

3.2.5. Evaluasi dan validasi hasil

Pada tahap ini dilakukan pengujian terhadap model-model yang dikomparasi untuk mendapatkan informasi model yang paling akurat. Evaluasi dan validasi menggunakan metode cross validation, confusion matrix, dan, kurva ROC.

3.3 Perangkat yang digunakan

Perangkat adalah alat yang digunakan dalam penelitian, terdiri dari perangkat keras (*hardware*) dan perangkat lunak (*software*).

3.3.1 Perangkat Keras (*Hardware*)

Spesifikasi perangkat keras yang digunakan untuk mendukung penelitian ini dapat dilihat sebagai berikut.

1. Processor : Pentium Dual Core
2. Memori : 2Gbyte.
4. Hardisk : 250 Gbyte.

3.3.2 Perangkat Lunak (*Software*)

Spesifikasi perangkat lunak yang digunakan untuk mendukung penelitian ini dapat dilihat sebagai berikut.

1. Sistem Operasi : Windows XP
2. Aplikasi : - Microsoft Excel
- Tools KNIME

BAB IV

HASIL ANALISA DAN PEMBAHASAN

4.1. Ekperimen dan Pengujian Model

4.1.1 Hasil Penelitian

Tujuan dari penelitian ini menguji keakuratan analisa identifikasi penyakit tiroid dengan menggunakan algoritma *Naïve Bayes* dan *Bayes Network*. Data yang dianalisa adalah data Pasien yang telah melakukan pemeriksaan kesehatan yang berhubungan dengan penyakit

Data Training yang digunakan adalah sebanyak 1585 data diantaranya 44 data pasien penderita hipertiroid dan 135 data pasien penderita hipotiroid dan 1406 data pasien dalam keadaan negative, Terdapat pada lampiran 1.

1. Naïve Bayes

Data training pada lampiran 1 menentukan apakah seorang Pasien terkena tiroid atau tidak. Kolom 4,5 dan 6 berikutnya adalah hasil perhitungan nilai probabilitas prior, yaitu untuk menentukan nilai probabilitas dari masing-masing kelas kita dapat nilai yang bernilai negative dan hipertiroid dan hipotiroid pada masing-masing atribut seluruh data.

Tabel 4.1. Atribut Kategori data training

Attribute	Class			P(X Ci)		
	hyperthyroid	hypothyroid	negative	hyperthyroid	hypothyroid	negative
Total	44	135	1406	0.0278	0.0876	0.9999
sex						
F	41	102	902	0.9318	0.7556	0.6415
M	3	33	504	0.0682	0.2444	0.3585
onthyroxine						
f	43	132	1260	0.9773	0.9778	0.8962
t	1	3	146	0.0227	0.0222	0.1038
queryonthyroxine						
f	44	133	1391	1.0000	0.9852	0.9893
t	0	2	15	0.0000	0.0148	0.0107
onantithyroidmedication						
f	44	134	1386	1.0000	0.9926	0.9858
t	0	1	20	0.0000	0.0074	0.0142
sick						
f	44	126	1340	1.0000	0.9333	0.9531

Tabel 4.1. Atribut Kategori data training (lanjutan)

t	0	9	66	0.0000	0.0667	0.0469
pregnant						
f	43	135	1382	0.9773	1.0000	0.9829
t	1	0	24	0.0227	0.0000	0.0171
thyroidsurgery						
f	44	134	1388	1.0000	0.9926	0.9872
t	0	1	18	0.0000	0.0074	0.0128
l131treatment						
f	43	134	1383	0.9773	0.9926	0.9836
t	1	1	23	0.0227	0.0074	0.0164
queryhypothyroid						
f	44	126	1346	1.0000	0.9333	0.9573
t	0	9	60	0.0000	0.0667	0.0427
queryhyperthyroid						
f	31	129	1328	0.7045	0.9556	0.9445
t	13	6	78	0.2955	0.0444	0.0555
lithium						
f	44	134	1398	1.0000	0.9926	0.9943
t	0	1	8	0.0000	0.0074	0.0057
goitre						
f	44	135	1394	1.0000	1.0000	0.9915
t	0	0	12	0.0000	0.0000	0.0085
tumor						
f	37	131	1381	0.8409	0.9704	0.9822
t	7	4	25	0.1591	0.0296	0.0178
hypopituitary						
f	44	135	1405	1.0000	1.0000	0.9993
t	0	0	1	0.0000	0.0000	0.0007
psych						
f	44	130	1308	1.0000	0.9630	0.9303
t	0	5	98	0.0000	0.0370	0.0697

Dari atribut yang bernilai Kategori memiliki nilai Probabilitas di masing-masing kelas yang sudah dilakukan perhitungan.

Tabel 4.2. Atribut Kontiyu

Program Pascasarjana Magister Ilmu Komputer STMIK Nusa Mandiri

Attribute	Class		
	hyperthyroid	hypothyroid	negative
age			
mean	50.4184	53.2342	53.1247
std. dev.	20.5148	17.101	21.8205
TSH			
mean	0.0198	30.3406	1.8263
std. dev.	0.1455	41.6664	4.9036
T3			
mean	3.7067	1.522	2.0062
std. dev.	1.7151	0.7974	0.7005
TT4			
mean	187.8547	75.2166	109.4425
std. dev.	52.29	33.0315	29.867
T4U			
mean	0.9622	1.0088	0.9991
std. dev.	0.1849	0.1913	0.2002
FTI			
mean	201.0879	75.9752	110.3967
std. dev.	59.4688	32.5479	25.0307

Dari atribut yang bernilai kontinyu memiliki nilai mean dan standar deviasi yang juga akan dilakukan perhitungan, untuk melakukan prediksi.

Dari hasil pengolahan data training maka di setiap atribut memiliki nilai probabilitas yang akan di hitung, data yang di tampilkan akan diujiakan menggunakan data testing untuk melakukan prediksi. Misalkan diambil sebuah data testing X dengan nilai seperti pada Tabel 4.3, untuk menentukan kelas mana, dilakukan perhitungan probabilitas *posterior* yang hasilnya terdapat pada Tabel 4.3 kolom 4,5,dan 6.

Tabel 4.3. Perhitungan untuk Menentukan Klasifikasi Data *testing X*

No	Atribut	Nilai	P (X Ci)		
			Hipertiroid	hipotiroid	Negative
1	age	31	0.0123	0.01	0.011
2	sex	F	0.9318	0.7556	0.6415
3	on thyroxine	t	0.0227	0.0222	0.1038
4	query on thyroxine	f	1	0.9852	0.9893

Tabel 4.3. Perhitungan untuk Menentukan Klasifikasi Data *testing X*
(Lanjutan)

5	on antithyroid medication	f	1	0.9926	0.9858
6	sick	f	1	0.9333	0.9531
7	pregnant	f	0.9773	1	0.9829
8	thyroid surgery	f	1	0.9926	0.9872
9	I131 treatment	f	0.9773	0.9926	0.9836
10	query hypothyroid	t	0	0.0667	0.0427
11	query hyperthyroid	f	0.7045	0.9556	0.9445
12	lithium	f	1	0.9926	0.9943
13	goitre	f	1	1	0.9915
14	tumor	f	0.8409	0.9704	0.9822
15	hypopituitary	f	1	1	0.9993
16	psych	f	1	0.963	0.9303
17	TSH	0.045	2.7013	0.0073	0.0761
18	T3	2.7	0.1958	0.1679	0.3488
19	TT4	163	0.0067	3.5412	0.0026
20	T4U	0.94	2.1424	1.955	1.9079
21	FTI	172	0.0059	1.5714	0.0007

Dari hasil yang telah di dapat kita bisa hitung probabilitas semuanya

Sehingga :

1. Likelihood Hipertiroid = $0.0123 \times 0.9318 \times 0.0227 \times 1 \times 1 \times 1 \times 0.9773 \times 1 \times 0.9773 \times 0 \times 0.7045 \times 1 \times 1 \times 0.8409 \times 1 \times 1 \times 2.7013 \times 0.1958 \times 0.0067 \times 2.1424 \times 0.0059 = 0$
2. Likelihood hipotiroid = $0.01 \times 0.7556 \times 0.0222 \times 0.9852 \times 0.9926 \times 0.9333 \times 1 \times 0.9926 \times 0.9926 \times 0.0667 \times 0.9556 \times 0.9926 \times 1 \times 0.9704 \times 1 \times 0.963 \times 0.0073 \times 0.1679 \times 3.5412 \times 1.955 \times 1.5714 = 1.19E-07$
3. Likelihood Negative = $0.011 \times 0.6415 \times 0.1038 \times 0.9893 \times 0.9858 \times 0.9531 \times 0.9829 \times 0.9872 \times 0.9836 \times 0.0427 \times 0.9445 \times 0.9943 \times 0.9915 \times 0.9822 \times 0.9993 \times 0.9303 \times 0.0761 \times 0.3488 \times 0.0026 \times 1.9079 \times 0.0007 = 2.17431E-12$

Nilai probabilitas dapat dihitung dengan melakukan normalisasi terhadap likelihood tersebut sehingga jumlah nilai yang diperoleh = 1

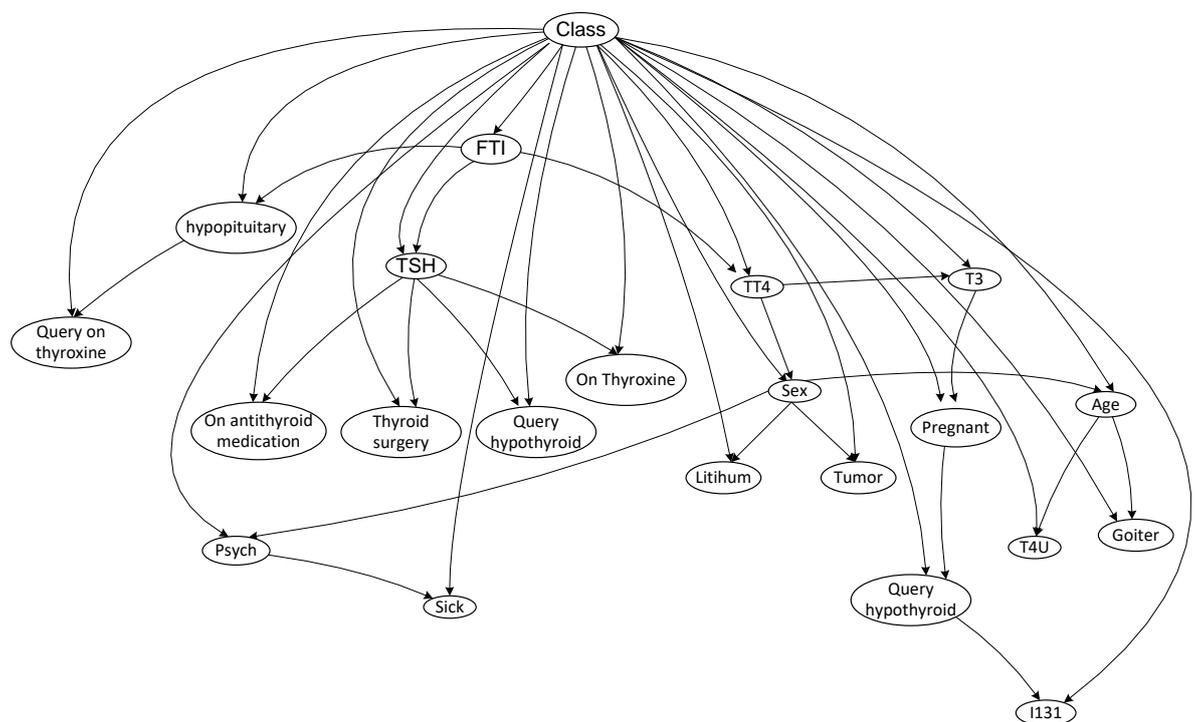
$$\text{Probabilitas Hipertiroid} = \frac{0}{0+1.19E-07+2.17431E-12} = 0$$

$$\text{Probabilitas Hipotiroid} = \frac{1.19E-07}{0+1.19E-07+2.17431E-12} = 0.999982$$

$$\text{Probabilitas Negative} = \frac{2.17431E-12}{0+1.19E-07+2.17431E-12} = 1.82848E-05$$

Maka hasil yang di dapat adalah Pasien menderita Hipotiroid , karena $P(\text{Hipotiroid}) > P(\text{Negative})$ dan $P(\text{Hipertiroid})$.

2. Bayes Network



Gambar 4.1. Model Bayes Network

Didalam gambar di atas terdapat node-node yang saling berhubungan satu sama lain yang dapat sangat berpengaruh untuk menentukan probailitas suatu attribut dan penglompokan berupa tabel antara node satu dengan node yang lain dapat di lihat di lampiran 2.

Tabel 4.4. Perhitungan untuk menentukan Klasifikasi Data *testing* X menggunakan Bayes Network Menggunakan berdasarkan Tabel pada lampiran 2

No	Attribut	Nilai	P (X Ci)		
			Hipertiroid	hipotiroid	Negative
1	age	31	1	1	1
2	sex	F	0.9125	0.5000	0.8850
3	on thyroxine	t	0.0714	0.5000	0.1565
4	query on thyroxine	f	0.9889	0.9816	0.9897
5	on antithyroid medication	f	0.9762	0.5000	0.9838
6	sick	f	0.9889	0.9351	0.9492
7	pregnant	f	0.9722	0.9938	0.9920
8	thyroid surgery	f	0.9762	0.5000	0.9874
9	I131 treatment	f	0.9531	0.9885	0.9868
10	query hypothyroid	t	0.0238	0.5000	0.0306
11	query hyperthyroid	f	0.6932	0.9522	0.9505
12	lithium	f	0.9881	0.9854	0.9906
13	goitre	f	0.9889	0.9963	0.9911
14	tumor	f	0.8214	0.9563	0.9751
15	hypopituitary	f	0.9167	0.7500	0.9979
16	psych	f	0.9881	0.9757	0.9496
17	TSH	0.045	0.3571	0.1667	0.3650
18	T3	2.7	0.4321	0.3333	0.5522
19	TT4	163	0.4667	0.1429	0.1916
20	T4U	0.94	1	1	1
21	FTI	172	0.1170	0.0109	0.1671

Dari hasil yang telah di dapat kita bisa hitung probabilitas semuanya

Sehingga :

Likelihood Hipertiroid = $1 \times 0.9125 \times 0.0714 \times 0.9889 \times 0.9762 \times 0.9889 \times 0.9722 \times 0.9762 \times 0.9531 \times 0.0238 \times 0.6932 \times 0.9881 \times 0.9889 \times 0.8214 \times 0.9167 \times 0.9881 \times 0.3571 \times 0.4321 \times 0.4667 \times 1 \times 0.1170 = 5.69122E-06$

Likelihood hipotiroid = $1 \times 0.5000 \times 0.5000 \times 0.9816 \times 0.5000 \times 0.9351 \times 0.9938 \times 0.5000 \times 0.9885 \times 0.5000 \times 0.9522 \times 0.9854 \times 0.9963 \times 0.9563 \times 0.7500 \times 0.9757 \times 0.1667 \times 0.3333 \times 0.1429 \times 1 \times 0.0109 = 1.59032E-06$

Likelihood Negative = $1 \times 0.8850 \times 0.1565 \times 0.9897 \times 0.9838 \times 0.9492 \times 0.9920 \times 0.9874 \times 0.9868 \times 0.0306 \times 0.9505 \times 0.9906 \times 0.9911 \times 0.9751 \times 0.9979 \times 0.9496 \times 0.3650 \times 0.5522 \times 0.1916 \times 1 \times 0.1671 = 2.10488E-05$

Nilai probabilitas dapat dihitung dengan melakukan normalisasi terhadap likelihood tersebut sehingga jumlah nilai yang diperoleh = 1

$$\text{Probabilitas Hipertiroid} = \frac{5.69122E-06}{5.69122E-06+1.59032E-06+2.10488E-05} = 0.200887595$$

$$\text{Probabilitas Hipotiroid} = \frac{1.59032E-06}{5.69122E-06 + 1.59032E-06 + 2.10488E-05} = 0.056134881$$

$$\text{Probabilitas Negative} = \frac{2.10488E-05}{5.69122E-06 + 1.59032E-06 + 2.10488E-05} = 0.742977524$$

Maka hasil yang di dapat adalah Pasien tidak mendrerita (Negative) , karena $P(\text{Negative}) >$ dari $P(\text{Hipertiroid})$ dan $P(\text{Hipotiroid})$.

4.2. Evaluasi dan Validasi

untuk membuat model klasifikasi, bisa digunakan banyak metode. penulisan menggunakan metode yang digunakan, yaitu Algoritma Naïve Bayes dan Bayes Network, kemudian dilakukan komparasi keduanya dan mengukur metode mana yang memiliki performance paling baik. Metode klasifikasi bisa dievaluasi berdasarkan beberapa kriteria seperti tingkat akurasi, kecepatan, kehandalan, skalabilitas, dan interpretabilitas (Vecellis, 2009)

4.2.1. Pengujian Model

Model yang telah dibentuk diuji tingkat akurasinya dengan memasukan data uji yang berasal dari data training. data maka digunakan metode cross validation dan split persentase untuk menguji tingkat akurasi. Untuk pengujian menggunakan metode cross validation

Tabel 4.5. Pengujian menggunakan cross 6

row ID	Naïve Bayes			Bayes Network		
	Error in %	Size of Test Set	Error Count	Error in %	Size of Test Set	Error Count
fold 0	8.843537	441	39	4.761905	441	21
fold 1	12.5	440	55	2.272727	440	10
fold 2	8.863636	440	39	2.045455	440	9
fold 3	8.61678	441	38	3.854875	441	17
fold 4	5.681818	440	25	2.272727	440	10
fold 5	9.772727	440	43	2.727273	440	12

Didalam pengujian menggunakan Cross 6 data yang di gunakan untuk testing 441 data secara acak maka di dapatkan hasil terbaik untuk metode Bayes Network pada fold ke -2 dengan error in sebesar 2.0454 % sedangkan kesalahan terkecil dengan menggunakan Metode Naive Bayes adalah pada fold ke-4 dengan error sebesar 5.6818% dan untuk lebih jelasnya dapat di lihat pada tabel 4.5.

Tabel 4.6. Pengujian menggunakan cross 8

row ID	Naïve Bayes			Bayes Network		
	Error in %	Size of Test Set	Error Count	Error in %	Size of Test Set	Error Count
fold 0	8.459215	331	28	4.833837	331	16
fold 1	12.42424	330	41	1.818182	330	6
fold 2	10.60606	330	35	1.818182	330	6
fold 3	8.484848	330	28	2.424242	330	8
fold 4	9.063444	331	30	3.625378	331	12
fold 5	7.272727	330	24	2.424242	330	8
fold 6	7.575758	330	25	0.909091	330	3
fold 7	9.090909	330	30	1.818182	330	6

Didalam pengujian menggunakan Cross 8 data yang di gunakan untuk testing 330 data secara acak maka di dapatkan hasil terbaik pada fold ke-5 menggunakan metode Nive Bayes dengan error sebesar 7.2727% sedangkan dengan menggunakan metode Bayes Network didapatkan nilai terbaik pada fold ke-6 dengan error in sebesar 0.9090 % untuk lebih jelasnya dapat di lihat pada Tabel 4.6.

Tabel 4.7. Pengujian menggunakan cross 10

row ID	Naïve Bayes			Bayes Network		
	Error in %	Size of Test Set	Error Count	Error in %	Size of Test Set	Error Count
fold 0	7.54717	265	20	0.377358	265	1
fold 1	10.98485	264	29	3.787879	264	10
fold 2	12.87879	264	34	3.787879	264	10
fold 3	10.98485	264	29	1.893939	264	5
fold 4	7.954545	264	21	3.030303	264	8
fold 5	8.301887	265	22	4.528302	265	12
fold 6	9.090909	264	24	2.272727	264	6
fold 7	5.681818	264	15	1.893939	264	5
fold 8	8.712121	264	23	1.136364	264	3
fold 9	9.090909	264	24	3.409091	264	9

Didalam pengujian menggunakan Cross 10 data yang di gunakan untuk testing 365 data secara acak maka di dapatkan hasil terbaik pada metode Byes Network pada fold ke-0 dengan error in sebesar 0.3773 % dengan salah klasifikasi sebesar 1 dan dengan metode Naive Bayes pada fold ke-7 untuk lebih jelasnya dapat di lihat pada tabel 4.7.

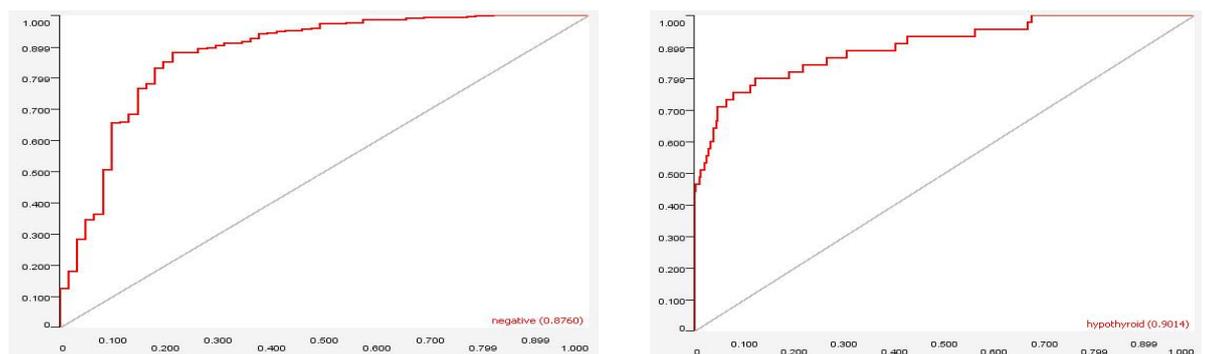
Tabel 4.8. Dengan Menggunakan Split

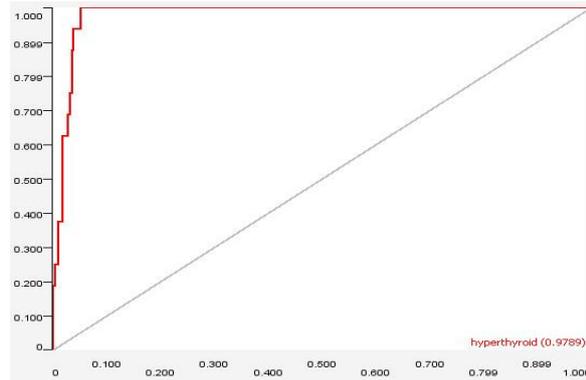
Data		Naive Bayes		Bayes Network	
Tranning (%)	Testing (%)	Accuracy (%)	Error (%)	Accuracy (%)	Error (%)
40	60	91.425	8.575	97.793	2.207
50	50	90.537	9.463	96.745	3.255
60	40	90.539	9.461	97.256	2.744
70	30	91.803	8.197	95.352	2.648
80	20	89.603	10.397	97.353	2.647
90	10	90.566	9.434	98.491	1.509

Didalam pengujian menggunakan Split percentase maka didapat hasil pada Naive Bayes dengan menggunakan data tranning sebesar 40 % dan data testing 60 % dan data secara acak memperoleh accuracy sebesar 91.425 % sedangkan dengan menggunakan Bayes Network diperoleh accuracy sebesar 97.793 % untuk lebih jelasnya dapat di lihat pada tabel 4.8.

4.2.2. Kurva ROC

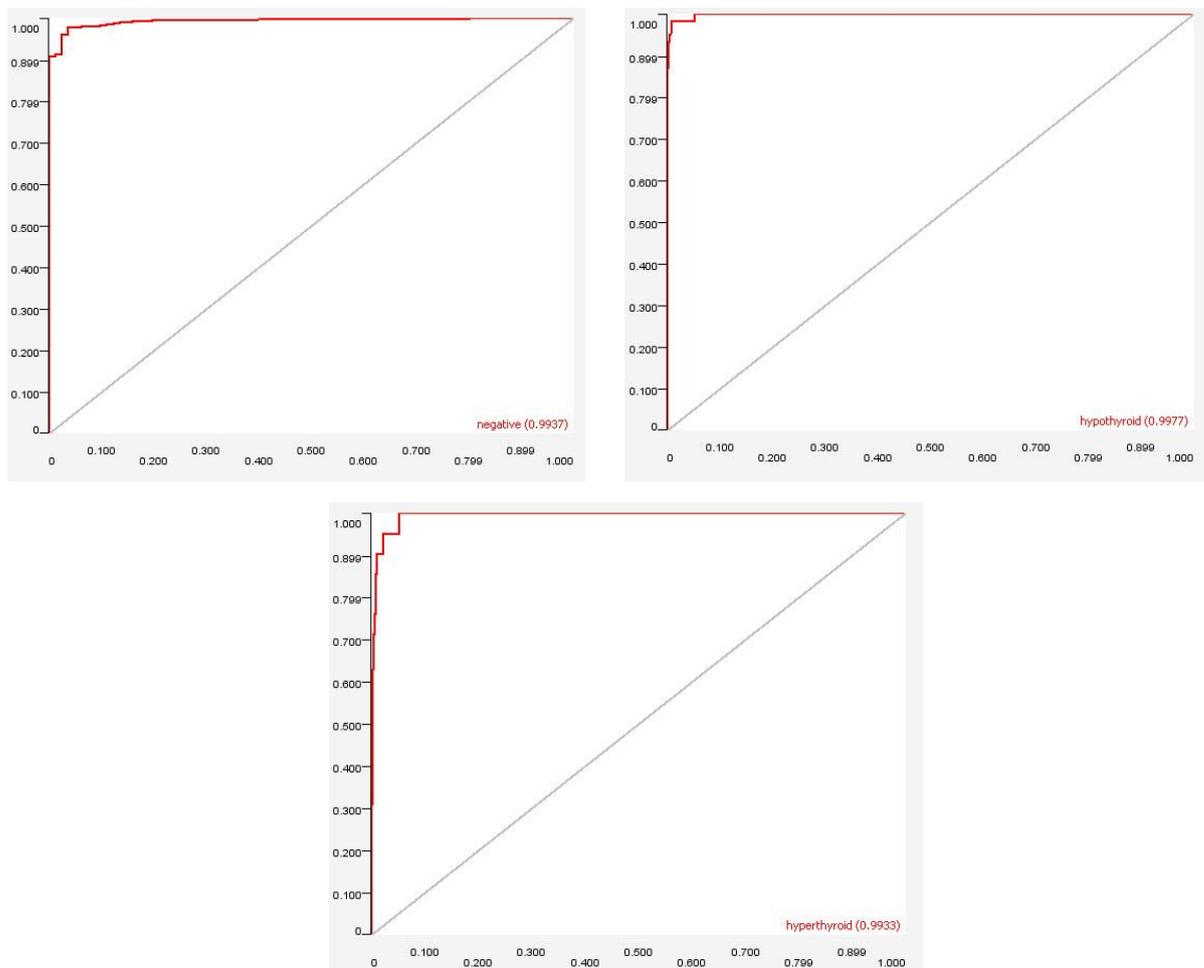
Hasil perhitungan divisualisasikan dengan kurva ROC. Perbandingan kedua metode komparasi bisa dilihat pada Gambar 4.2 yang merupakan kurva ROC untuk Algoritma Naïve Bayes.





Gambar 4.2. Curva ROC pada test data Algoritma Naïve Bayes

Hasil perhitungan kurva ROC mengekspresikan confusion matrix. Garis horizontal adalah false positives dan garis vertikal true positives Pada curva ROC – negative (0,8760), ROC – hypotiroid (0,9014) dan ROC – hipertiroid (0,9789).



Gambar 4.3. Curva ROC pada test data Algoritma Bayes Network

Hasil perhitungan divisualisasikan dengan Kurva ROC mengekspresikan confusion matrix. Garis horizontal adalah false positives dan garis vertikal true positives Pada curva ROC – negative (0,9937), ROC – hypotiroid (0,9977) dan ROC - hypertiroid (0,9933).

4.2.3. Komparasi Nilai AUC

Model yang dihasilkan dengan metode Naïve Bayes dan Bayes Network diuji menggunakan metode split percentase dan Cross Validation, terlihat perbandingan nilai accuracy, untuk metode Bayes Network memiliki nilai *accuracy* yang paling tinggi, diikuti dengan metode Naïve Bayes

Tabel 4.9. Komparasi Nilai AUC

	Data	Naïve Bayes	<i>Bayes Network</i>
AUC	40-60	0.914	0.978
	50-50	0.905	0.967
	60-40	0.905	0.973
	70-30	0.918	0.974
	80-20	0.896	0.974
	90-10	0.906	0.985

Tabel 4.9 membandingkan *accuracy* dan AUC dari tiap metode. Terlihat bahwa nilai *accuracy* Bayes Network paling tinggi dibandingkan dengan Naïve Bayes Untuk klasifikasi *data mining*, nilai AUC dapat dibagi menjadi beberapa kelompok (Gorunescu, 2011).

- a. 0.90-1.00 = klasifikasi sangat baik
- b. 0.80-0.90 = klasifikasi baik
- c. 0.70-0.80 = klasifikasi cukup
- d. 0.60-0.70 = klasifikasi buruk
- e. 0.50-0.60 = klasifikasi salah

Berdasarkan pengelompokan di atas dan Tabel 4.9 maka dapat disimpulkan bahwa metode yang memiliki klasifikasi sangat baik secara berurut adalah, *Bayes Network* dan *Naïve Bayes* termasuk karena memiliki nilai AUC antara 0.90-1.00.

4.3. Implikasi Hasil Penelitian

Tidak semua permasalahan dapat di selesaikan oleh salah satu algoritma saja karena belum tentu algoritma yang digunakan memiliki performance yang

baik. Oleh sebab itu untuk mengetahui algoritma mana yang paling akurat untuk digunakan pada permasalahan atau kasus tertentu maka perlu dilakukan komparasi beberapa algoritma.

Pada kasus Diagnosa penyakit tiroid menggunakan Algoritma *Naive Bayes* dan *Bayes Network* dapat diterapkan pada data pasien yang memeriksakan kesehatan untuk mengetahui identifikasi suatu penyakit. Dari penelitian ini juga terbukti bahwa algoritma *Bayes Network* ternyata lebih akurat bila dibandingkan dengan algoritma *Naive Bayes*. Hal ini terlihat dari hasil evaluasi yang telah dilakukan. Dengan hasil ini, menunjukkan bahwa *Bayes Network* merupakan metode yang cukup baik dalam prediksi data sehingga dapat memberikan pemecahan untuk permasalahan identifikasi penyakit Tiroid.

Untuk mendukung pengambilan keputusan dan pengembangan sistem pengetahuan berbasis komputer, hasil penelitian ini memiliki implikasi pada beberapa aspek, yaitu :

1. Aspek Sistem

Dari hasil penelitian yang telah penulis lakukan menunjukkan bahwa metode *Bayes Network* memiliki tingkat keakuratan yang lebih tinggi dibandingkan dengan *Naive Bayes*. Hal ini mengandung implikasi bahwa metode *Bayes Network* lebih cocok diterapkan untuk menyelesaikan permasalahan identifikasi penyakit tiroid, sehingga dengan melihat hasil perbandingan tersebut, maka kedepannya dibutuhkan adanya program pengolahan data mining sebagai langkah strategik sebagai alat bantu pertimbangan dokter dalam memutuskan suatu penyakit apa yang diderita pasien agar lebih akurat.

2. Aspek Manajerial

Setelah penulis menyelesaikan penelitian ini, penulis dapat menyimpulkan bahwa ternyata adanya permasalahan ketika pasien memberikan rekam medis yang salah itu akan sangat berpengaruh dalam hal mengidentifikasi suatu penyakit. Oleh karena itu penelitian ini penulis harapkan bahwa perlu adanya pengecekan ulang secara bertahap dan factor dominan apa saja yang sangat mempengaruhi dalam mengidentifikasi suatu penyakit tiroid. Sehingga faktor yang lemah bisa diabaikan

Agar hasil penelitian ini dapat diterapkan dengan baik, perlu adanya langkah-langkah yang harus dilakukan, yaitu :

- a. Pembuatan SOP perlu dilakukan agar penerapan software hasil penelitian ini sesuai dengan tujuan yang diharapkan.
- b. Penyuluhan (*desimination*)
Berdasarkan SOP yang dibuat, perlu adanya penyuluhan (desiminasi) agar hasil dari penelitian ini dapat tersosialisasi dengan baik.
- c. Pelatihan (*training*)
Pelatihan terhadap calon pengguna diperlukan, agar hasil penelitian ini dapat dimengerti dan diterapkan.
- d. Penerapan (*Implementation*)
Penerapan hasil penelitian ini merupakan langkah yang terpenting. Jika langkah ini tidak dilakukan, maka hasil penelitian yang telah dilakukan tidaklah berguna.

3. Aspek Penelitian Lanjutan

Penelitian semacam ini perlu terus dikembangkan, baik dari sisi bidang kesehatan ataupun dari sisi *data mining*nya, diantaranya :

a. *Extrapolasi*

Hasil penelitian ini dapat dikembangkan pada penyakit-penyakit lain, khususnya untuk penyakit yang selama ini susah terdeteksi melalui penilaian dokter.

b. *Extention*

Penelitian selanjutnya dapat mengembangkan untuk prediksi penyakit tiroid dengan menggunakan variabel-variabel atau menghapus variable yang tidak signifikan untuk prediksi penyakit tiroid. Selain itu, penelitian ini juga dapat dikembangkan dengan menggunakan algoritma yang lain.

BAB V

PENUTUP

5.1 Kesimpulan

Dalam penelitian ini dilakukan pembuatan model menggunakan algoritma Naïve bayes dan Bayes Network menggunakan data Pasien Penderita Tiroid. Model yang dihasilkan, dikomparasi untuk mengetahui algoritma yang paling baik dalam penentuan Identifikasifikasi penyakit. Untuk mengukur kinerja kedua algoritma tersebut digunakan metode pengujian Cross Validation, dan Split Percentace, dan pengukuranya Dengan menggunakan *confusion matrix*, Bayes Network memiliki Akurasi yang lebih tinggi dengan nilai 98.491% dibandingkan Naïve Bayes dengan nilai 91.803%.

Dengan menggunakan Kurva ROC, Bayes Network juga memiliki akurasi yang lebih tinggi Pada Curva ROC – negative (0.9337), ROC – hipertiroid (0,9933) dan ROC – hypotiroid (0,9977). dibandingkan Naïve Bayes Pada curva ROC – negative (0.8760), ROC – hipertiroid (0,9789) dan ROC – hypotiroid (0,9018). dan dapat disimpulkan pula bahwa metode yang memiliki klasifikasi sangat baik secara berurut adalah, dan *Bayes Network* dan *Naïve bayes* berdasarkan penilaian AUC antara 0.90-1.00 dengan demikian algoritma Bayes Network dapat memberikan pemecahan untuk permasalahan dalam mengidentifikasi penyakit tiroid.

5.2 Saran

Agar penelitian ini bisa ditingkatkan, berikut adalah saran-saran yang diusulkan:

1. Hasil penelitian ini diharapkan bisa digunakan pada Rumah Sakit sebagai alat bantu dalam mengidentifikasi penyakit tiroid oleh dokter

2. Dalam Penelitian berikutnya disarankan menggunakan metode algoritma Bayes Network dengan mengurangi beberapa atribut yang tidak berpengaruh atau lemah agar pengujian dapat memperoleh hasil yang lebih baik.
3. Dalam Penelitian berikutnya disarankan menggunakan metode komparasi yang lainnya agar dapat terus dikembangkan secara periodik.

DAFTAR REFERENSI

- Alpaydin, Ethem. (2010). *Introduction to Machine Learning*. London: The MIT Press
- Aziz , Behnam (2012) *Thyroid Disease Diagnosis using Genetic Algorithm and Neural Network*
- Bachtiar, Hafni (2009). Faktor Determinan Kejadian Gondok di daerah Pantai Jawa Timur. *Jurnal kesehatan Masyarakat*.
- Barnaghi, Mamani (2012) *A Comparative Study for Various Methods of Classification*. Department of Artificial Intelligence, National University of Malaysia, Malaysia.
- C.R.Kothari. (2004). *Research Methology Methods and Techniques*. India: New Age International Limited.
- Cheng, J., Bell, D., and Liu, W. (1998). *Learning Bayesian Networks from Data : An Efficient Approach Based on Information Theory* . Faculty of Informatics, University of Ulster, U.K.
- Gorunescu, F. (2011). *Data Mining Concepts, Models and Techniques*. Berlin Heidelberg: Springer Verlag.
- Hamdani , William (2010) *Thyroid Cancer: The Diagnose And The Management*
- Han, J., and Kember, M. (2006). *Data Mining Concepts adn Techniques*. San Fransisco: Morgan Kauffman.
- Hannan, Shaikh Abdul, R. R. Manza, R. J. Ramteke. (2010). *Generalized Regression Neural Network and Radial Basis Function for Heart Disease diagnosis*. Maharashtra, India. *International Journal of Computer Application* (0975-8887) Volume 7- No.13.
- Hermawan, Guntur. (1990) *Pengelolaan dan Pengobatan Hipertiroidi*. Surakarta Fakultas Kedokteran Universitas Sebelas Maret
- Hamdani , William (2010) *Thyroid Cancer: The Diagnose And The Management*
- Kousarrizi, Nazari (2012) *An Experimental Comparative Study on Thyroid Disease Diagnosis Based on Feature Subset Selection and classification*
- Kumar, Aswani. (2012) *Diagnosis of Thyroid Disorders using Decision Tree Splitting Rules*.

- Kusrini, dan Luthfi, E. T. (2009). *Algoritma Data Mining*. Yogyakarta: Andi Publishing.
- Larose, D. T. (2005). *Discovering Knowledge in Data*. New Jersey: John Wiley & Sons, Inc.
- Liao. (2007). *Recent Advances in Data Mining of Enterprise Data: Algorithms and Application*. Singapore: World Scientific Publishing
- Maimon, Oded and Rokach, Lior.(2010). *Data Mining and Knowledge Discovey Handbook*. New York: Springer
- Moayeri, H, Haghshenas, Z., 2002. *Prevalensi tirotoksikosis: Klinis. Acta Medica Iranica*, Vol. 40, No.4. http://journals.tums.ac.ir/upload_files/pdf/1108.pdf
- Neshat, Mehdi, Mehdi Sargolzaei, Adel Nadjaran Toosi, dan Azra Masoumi. (2012). *Hepatitis Disease Diagnosis using hybrid case based reasoning and paticle swarm optimization*. International Scholarly Research Network: ISRN Artificial Intelligence Volume 2012
- Neshat, Mehdi, Yaghoobi, Mehdi. (2009). *Designing a Fuzzy Expert System of Diagnosing the Hepatitis B Intensity Rate and Comparing It with Adaptive Neural Netwrk Fuzy System*. Proceeding of the world congress on engineering and computer science 2009 Vol II, WCECS 2009, October 20-22, 2009, San Fransisco, USA. ISBN:978-988-18210-2-7.
- Reena, Sophia (2010), *Analysis of Liver Disorder Using Data mining Algorithm*
- Shadiq, (2009), *Keoptimalan Naïve Bayes Dalam Klasifikasi*. Program Ilmu Komputer FPMIPA Universitas Pendidikan Indonesia
- Santoso, B. (2007). *Data Mining Teknik Pemanfaatan Data untuk Keperluan Bisnis*. Yogyakarta: Graha Ilmu.
- Shukla, A., Tiwari, R., and Kala, R. (2010). *Real Life Applications of Soft Computing*. United States of America on: Taylor and Francis Group, LLC.
- Setiyani (2009) . *Gangguan pada kelenjar tiroid*
- Sumathi, S., and Sivanandam, S. (2006). *Introduction to Data Mining and its Applications*. Verlag Berlin Heidelberg: Springer.
- Susanto, rudi (2006). *Outcome bayi dari ibu hipo atau hipertiroidisme*. Semarang , Fakultas UNDIP, Bagian ilmu kesehatan anak.
- Tandra, hans (2011) *Mencegah Dan Mengatasi Penyakit Tiroid*. Jakarta : Gramedia

UCI repository of machine learning databases, University of California at Irvine, Department of Computer Science, <http://archive.ics.uci.edu/ml/datasets/Thyroid> Disease, diakses Februari (2013)

Vercellis, C. (2009). *Business Intelligent: Data Mining and Optimizzation for Decision Making*. Southern Gate, Chichester, West Sussex, United Kingdom : John Wiley & Sons Ltd.

WHO, 2007. *Scaling Up Prevention and Control of Non -Communicable Disease*. The SEANET- NCD Meeting, Oktober 2007. Phucket. Thailand. <http://www.searo.who.int/> diakses Februari (2013)

Witten, I. H., Frank, E., and Hall, M. A. (2011). *Data Mining: Practical Machine Learning and Tools*. Burlington: Morgan Kaufmann Publisher.

Wu, X., and Kumar, V. (2009). *The Top Ten Algorithms in Data Mining*. Boca Raton, London, New York: Taylor & Francis Group, LLC.