PERBANDINGAN ALGORITMA DATA MINING NAIVE BAYES DAN BAYES NETWORK UNTUK MENGIDENTIFIKASI PENYAKIT TIROID

Bambang Wijonarko
Teknik Komputer
AMIK BSI Jakarta
http://www.bsi.ac.id
bambang.bwo@bsi.ac.id

Abstract— *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*: Naive Bayes, Bayes Network, Confusion Matix, ROC Curve, AUC.*

Intisari—Dalam data mining dikenal Salah satu model Klasifikasi yang dapat digunakan untuk mengidentifikasi penyakit tiroid, yaitu dengan metode Naïve bayes dan Bayes Network . Dalam penelitian ini dilakukan pembuatan model menggunakan algoritma naïve bayes dan Bayes Network menggunakan data Pasien Penderita Tiroid dengan menggunakan tools KNIME. Model yang dihasilkan, dikomparasi untuk mengetahui algoritma yang paling baik dalam penentuan identidfikasifikasi penyakit. Untuk mengukur kinerja kedua algoritma tersebut digunakan metode pengujian Cross Validation, dan split percentace, Dengan menggunakan confusion matrix, Bayes Network memiliki akurasi yang lebih tinggi dengan nilai 98.491% dibandingan Naïve Bayes dengan nilai 91.803%. Dengan menggunakan Kurva ROC, Bayes Network juga memiliki akurasi yang lebih tinggi Pada kurva ROC – negative (0.9337), ROC – hypertiroid (0,9933) dan ROC – hypotiroid (0,9977). dibandingkan Naïve Bayes Pada kurva ROC – negative (0.8760), ROC – hypertiroid (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.

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

PENDAHULUAN

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 & Sampepajung, 2010). Fungsi kelenjar tiroid adalah menghasilkan hormon tiroid yang berguna untuk menjaga metabolisme tubuh. untuk meningkatkan jumlah oksigen pada sel dan rangsangan jaringan tubuh dalam menghasilkan protein. Penyakit gondok disebabkan oleh gangguan pada kelenjar tiroid (Werner, 2010) Ada dua jenis gangguan tiroid yang dapat muncul yaitu hipertiroid dan hypotiroid. Hipertirod adalah kondisi dimana kelenjar tersebut bekerja secara berlebihan, sedangnkan hipotiroid adalah kebalikannya (Tandra, 2011).

Menurut (Who, 2007) 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 naik menjadi 73% dan proporsi kesakitan menjadi 60%.

Kesalahan dalam menganalisis data dapat menyebabkan hal yang berbahaya dan Teknik analisis data manual secara konvensional tidak lagi efektif digunakan karena membutuhkan waktu yang lama (Neshat & 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 dihandalkan. Oleh karena itu, *metode soft computing* menunjukkan potensi yang menjanjikan untuk dikembangkan guna menghasilkan diagnosa medis yang tepat. (Hannan, 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.

Dari penelitian yang dilakukan (Sarwar, 2012). menyatakan bahwa algoritma Naive Bayes tercatat mengungguli algoritma lain di kedua masalah diagnosa medis maupun non-medis. Menurut penelitian lain menyatakan (Panda & Patra, 2007). menyatakan Naïve Bayes juga dikenal sebagai teknik yang paling baik dalam hal waktu komputasi dibandingkan teknik algoritma data mining lainnya. Didalam algoritma Naïve Bayes asumsi attributnya bersifat idependen sedangkan algoritma attributnya bersifat saling dependent. 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 ?.

BAHAN DAN METODE

1. Desain 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.

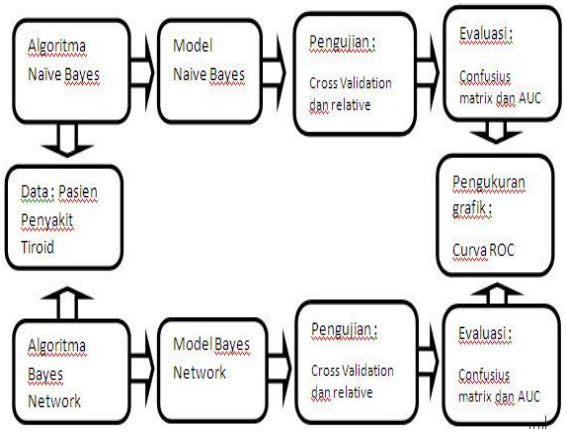
2. Pengumpulan Data

Di dalam penelitian ini menggunakan 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. Mengenai data pasien yang terkena penyakit tirod data yag di dapat sebanyak 3711 pasien yang di periksa dana sebanyak 3318 terdetksi normal, 102 pasien teridentifikasi menderita hipertiroid dan 291 pasien teridentifikasi menderita hypotiroid. 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.(Wang, Cao, Wang, Sun, & Dong, 2017) Terdapat 1 attribut special dan 21 attribut regular.

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. Model atau Metode

Pada tahap modeling ini dilakukan pemprosesan data traning 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.



Sumber: (Wijonarko, 2017)

Gambar 1. Model yang yang diusulkan

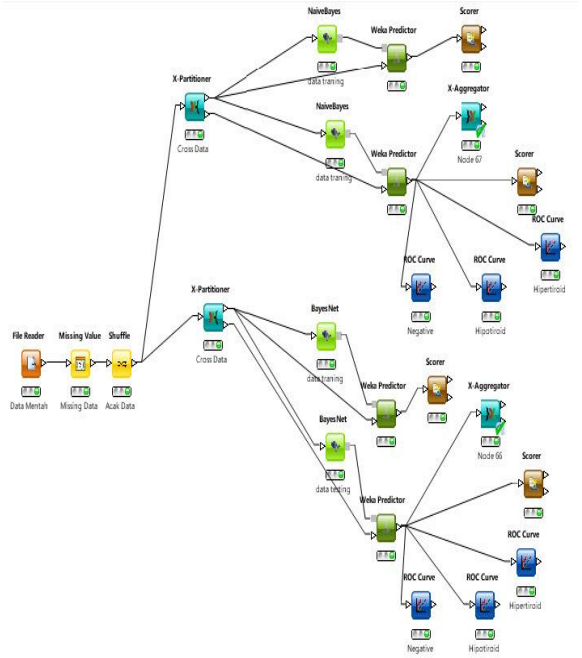
Tahap-tahap proses yang dilakukan pada tahap modeling ini adalah sebagai berikut :

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 (Kirschen, O'Higgins, & Lee, 2000). 10 folds cross validation merupakan merupakan pengujian yang memiliki tingkat akurasi yang paling baik dalam pengklasifikasian. Dan yang kedua menggunakan split data sebsar 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

4. Pengujian Model dengan Software

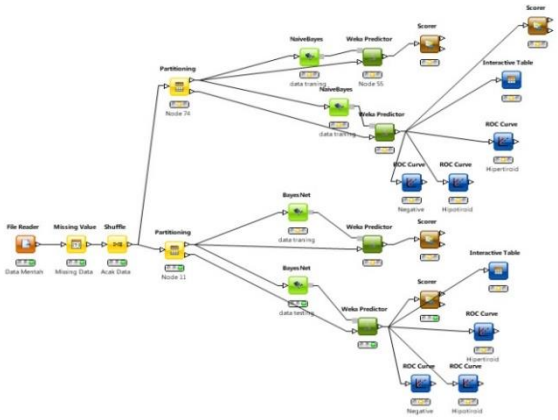
Pada tahapan ini akan di lakunakn pengujian dengan menggunakan software KNAME , yang akan mengolah data dengan membandingkan dua metode, berikut gambar di bawah ini.



Sumber: (Wijonarko, 2017)

Gambar 2. Model split persentace untuk membandingkan dua metode

Gambar di atas adalah model uji data yang menggunakan split persentace dengan menggunakan metode Naive Bayes dan Bayes Network



Sumber: (Wijonarko, 2017)

Gambar 3. Model Cross Validation untuk membandingkan dua metode

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

HASIL DAN PEMBAHASAN

Hasil 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 Traning 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.

Hasil 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 persentace untuk menguji tingkat akurasi. Untuk pengujian menggunakan metode cross validation.

Tabel 1. Pengujian menggunakan cross 6

	Naive Bayes			Bayes Network		
row ID	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

Sumber : (Wijonarko, 2017)

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 1.

Tabel 2. Penggujian menggunakan cross 8

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

Sumber : (Wijonarko, 2017)

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 7.5757 % untuk lebih jelasnya dapat di lihat pada Tabel 2.

Tabel 3. Penggujian menggunakan cross 10

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

Sumber : (Wijonarko, 2017)

Didalam pengujian menggunakan Cross 10 data yang di gunakan untuk testing 265 data secara acak maka di dapatkan hasil terbaik pada metode Bayes Network pada fold ke-0 dengan error in sebesar 7.5471 % dengan salah klasifikasi sebesar 1 dan dengan metode Naive Bayes pada fold ke-7 dengan error in sebesar 5.6818 % untuk lebih jelasnya dapat di lihat pada tabel 3.

Tabel 4. Dengan Menggunakan Split

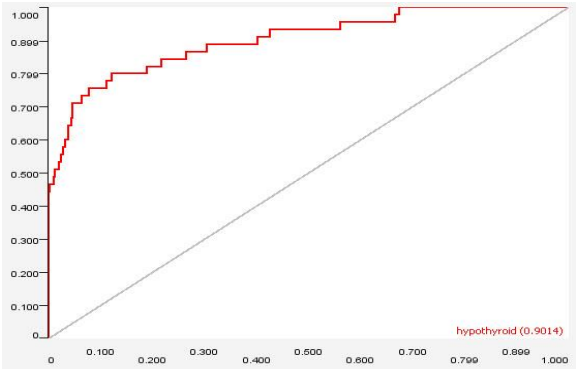
Data		Naive Bayes		Bayes Network	
Tranning (%)	Testing (%)	Accuracy (%)	Error (%)	Accuracy (%)	Error (%)
40	60	91.425	8.575	97.793	2.207
50	50	90.537	9.463	96.745	3.255
60	40	90.539	9.461	97.256	2.744
70	30	91.803	8.197	95.352	2.648
80	20	89.603	10.397	97.353	2.647
90	10	90.566	9.434	98.491	1.509

Sumber : (Wijonarko, 2017)

Didalam penggujian menggunakan Split percentace maka didapat hasil pada Naive Bayes dengan menggunakan data tranning sebesar 40 % dan data testing 60 % dan data secara acak memperoleh accuracy sebesar 91.803 % sedangkan dengan menggunakan Bayes Network diperoleh accuracy sebesar 97.793 % untuk lebih jelasnya dapat di lihat pada tabel 4.

Kurva ROC

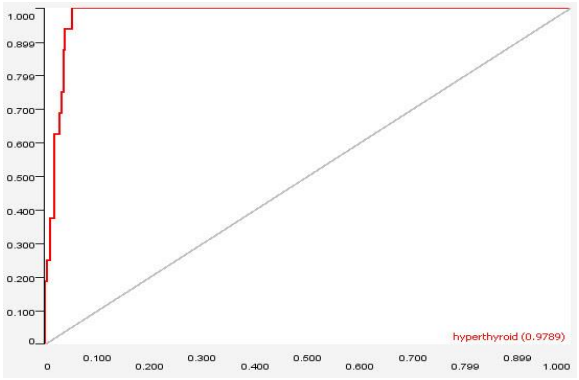
Hasil perhitungan dapat kita lihat melalui kurva ROC. Perbandingan kedua metode, True Positif Untuk Posisi Atas, False Negatif Untuk Posisi Bawah, Baseline untuk posisi Tengah-tenah (garis abu-abu).



Sumber: (Wijonarko, 2017)

Gambar 4. kurva ROC Class Hypothyroid Algoritma Naive Bayes

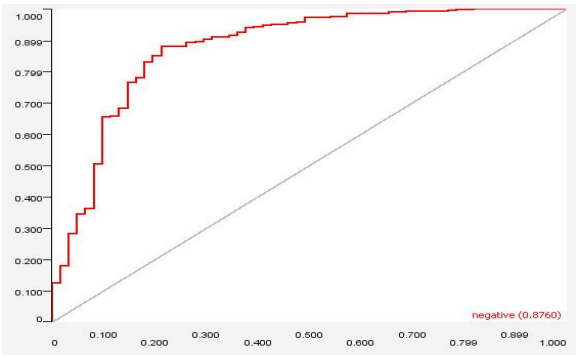
Pada kurva di atas ROC berada pada posisi true positif sebesar 0.9014 untuk class hypothyroid pada test data Algoritma Naive Bayes.



Sumber: (Wijonarko, 2017)

Gambar 5. kurva ROC Class Hyperthyroid Algoritma Naive Bayes

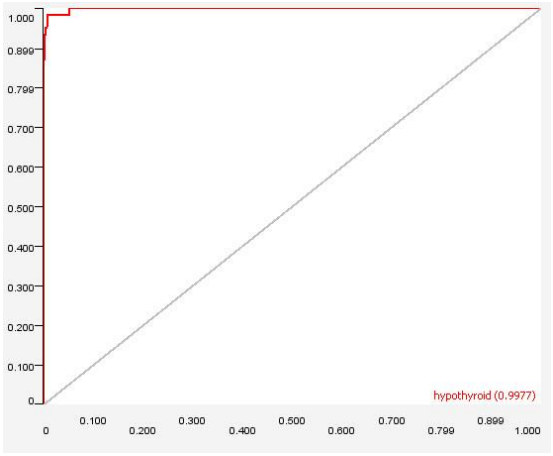
Pada kurva di atas ROC berada pada posisi true positif sebesar 0.9789 untuk class hyperthyroid pada test data Algoritma Naive Bayes.



Sumber: (Wijonarko, 2017)

Gambar 6. kurva ROC Class Negative Algoritma Naive Bayes

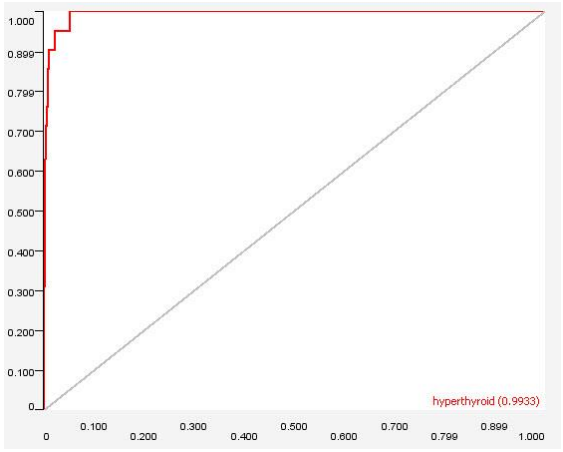
Pada kurva di atas ROC berada pada posisi true positif sebesar 08760 untuk class negative pada test data Algoritma Naive Bayes.



Sumber: (Wijonarko, 2017)

Gambar 7. kurva ROC Class Hypothyroid Algoritma Bayes Network

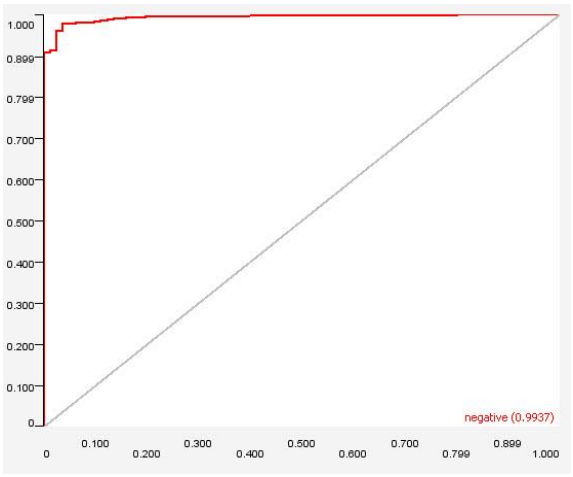
Pada kurva di atas ROC berada pada posisi true positif sebesar 0.9977 untuk class hypothyroid pada test data Algoritma Bayes Network.



Sumber: (Wijonarko, 2017)

Gambar 8. kurva ROC Class Hyperthyroid Bayes Network

Pada kurva di atas ROC berada pada posisi true positif sebesar 0.9933 untuk class hyperthyroid pada test data Algoritma Bayes Network.



Sumber: (Wijonarko, 2017)

Gambar 9. kurva ROC Class Negative Algoritma Bayes Network

Pada kurva di atas ROC berada pada posisi true positif sebesar 0.9937 untuk class negative pada test data Algoritma Bayes Network.

Komparasi Nilai AUC

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

Tabel 5. Komparasi Nilai AUC

	Data	Naïve Bayes	*Bayes Network*
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

Sumber: (Wijonarko, 2017)

Membandingkan accuracy dan AUC dari tiap metode. Terlihat bahwa nilai accuracy Bayes Network paling tinggi dibandingkan dengan Naive 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 pengelompokkan di atas dan Tabel 5 maka dapat disimpukan bahwa metode yang memiliki klasifikasi sangat baik secara berurut adalah, Bayes Network dan Naive Bayes termasuk karena memiliki nilai AUC antara 0.90-1.00

KESIMPULAN

Dalam penelitian ini dilakukan pembuatan model menggunakan algoritma Naive 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% dibandingan Naive Bayes dengan nilai 91.803%.

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

REFERENSI

Gorunescu, F. (2011). *Data Mining: Concepts, Models and Techniques*. *Data mining - Concepts, Models and Technique*. https://doi.org/10.1007/978-3-642-19721-5

Hamdani, W., & Sampepajung, D. (2010). Thyroid Cancer : the Diagnose and the Management.

Hannan, S. A., Manza, R. R., & Ramteke, R. J. (2010). Generalized Regression Neural Network and Radial Basis Function for Heart Disease Diagnosis. *International Journal of Computer Applications*, *7*(13), 975–8887. https://doi.org/10.5120/1325-1799

Kirschen, R. H., O'Higgins, E. A., & Lee, R. T. (2000). The Royal London Space Planning: An integration of space analysis and treatment planning. *American Journal of Orthodontics and Dentofacial Orthopedics*, *118*(4), 448–455. https://doi.org/10.1067/mod.2000.109031

Kothari, C. (2004). *Research methodology: methods and techniques*. *New Age International*.

Neshat, M., & Yaghobi, M. (2009). Designing a Fuzzy Expert System of Diagnosing the Hepatitis B Intensity Rate and Comparing it with Adaptive Neural Network Fuzzy System, *II*.

Panda, M., & Patra, M. (2007). Network intrusion detection using naive bayes. *... Journal of Computer Science and Network Security*, *7*(12), 258–263.

Sarwar, A. (2012). abid savar-Intelligent Naïve Bayes Approach to Diagnose-2012.pdf, (November), 14–16.

Tandra, H. (2011). *Mencegah Dan Mengatasi Penyakit Tiroid*. Jakarta: Gramedia.

Wang, L., Cao, F., Wang, S., Sun, M., & Dong, L. (2017). Using k-dependence causal forest to mine the most significant dependency relationships among clinical variables for thyroid disease diagnosis. *Plos ONE*, *12*(8), e0182070. https://doi.org/10.1371/journal.pone.0182070

Werner, D. (2010). *Apa yang Anda Kerjakan Bila Tidak Ada Dokter*. (Andi Offset, Ed.). Yogyakarta.

Who, M. (2007). Scaling up prevention and control of noncommunicable diseases : The SEANET-NCD meeting, (October), 22–26.

Wijonarko, B. (2017). *Laporan Akhir Penelitian Mandiri*. Jakarta.