PENERAPAN DECISION TREE C4.5 SEBAGAI SELEKSI FITUR DAN SUPPORT VECTOR MACHINE (SVM) UNTUK DIAGNOSA KANKER PAYUDARA

Pakarti Riswanto[1], RZ. Abdul Aziz[2], Sriyanto[3]

[123]Fakultas Ilmu Komputer Informatics & Business Institute Darmajaya
Jl. Z.A. Pagar Alam No. 93, Bandar Lampung - Indonesia 35142
Telp. (0721) 787214 Fax. (0721) 700261
e-mail : tutiriswanto@gmail.com, RZ.Aziz@gmail.com, sriyanto@darmajaya.ac.id

ABSTRACT

In the field of medicine, the use of data mining has a quite important and evolutionary role that can change the perspective of doctors, practitioners and health researchers in the process of detecting breast cancer in a patient. There are 2 classification applications in it, namely the process of diagnosing (diagnosing) cancer cells that distinguishes between tumors (benign cancer) or malignant cancer and prognosis (prognosis) to determine the possibility of reappearance of cancer cells in patients who have been operated on in the future. Data mining aims to describe new findings in the dataset and explain a process that uses statistical, mathematical, artificial intelligence, and machine learning techniques to extract and identify useful information and related knowledge from the database.

Classification with data mining can be done using several methods, namely Decision Tree, K-Nearest Neighbor, Naive Bayes, ID3, CART, Linear Discriminant Analysis, etc., which certainly have advantages and disadvantages of each. But in this study, the author focuses on the classification of data mining using the Support Vector Mechine and Deccision Tree algorithms.

This study will analyze the Breast Cancer Wisconsin Original data set obtained from the UCI Machine Learning Repository (repository of research data) to classify breast cancer malignancies. This time the author correlates between the Decision Tree classifier algorithm which has good ability to process large databases as a feature selection, then with a proper and relevant SVM Method used in analyzing and diagnosing breast breast cancer patients because it has accurate results for existing problems and several bases .

Keywords— *Data Mining, diagnosis, Decision Tree, SVM Method*

ABSTRAK

Dalam bidang kedokteran, penggunaan data mining mempunyai peranan yang cukup penting dan evolusioner yang dapat mengubah cara pandang para dokter, praktisi dan peneliti kesehatan dalam melakukan proses deteksi penyakit kanker payudara pada seorang pasien. Terdapat 2 aplikasi klasifikasi di dalamnya, yaitu proses diagnose (diagnosis) sel kanker yang membedakan antara tumor (kanker jinak) atau kanker ganas dan proses prognosa (prognosis) untuk mengetahui kemungkinan munculnya kembali sel kanker pada pasien yang telah dioperasi di masa yang akan datang. data mining bertujuan untuk menguraikan temuan baru di dalam dataset dan menjelaskan suatu proses yang

menggunakan teknik statistik, matematis, artificial intelligence, dan machine learning untuk melakukan ekstrak dan identifikasi informasi yang bermanfaat dan pengetahuan yang terkait dari database tersebut.

Klasifikasi dengan data mining dapat dilakukan menggunakan beberapa metode yaitu Decision Tree, K-Nearest Neighbor, Naive Bayes, ID3, CART, Linear Discriminant Analityc dan lain sebagainya yang tentunya memiliki kelebihan dan kekurangan masing - masing. Namun pada penelitian kali ini, penulis berfokus pada klasifikasi data mining dengan menggunakan algoritma Support Vector Mechine dan Deccision Tree.

Penelitian ini akan menganalisis Breast Cancer Winconsin Original data set yang diperoleh dari UCI Machine Learning Repository (repositori data penelitian) untuk melakukan klasifikasi keganasan kanker payudara. Kali ini penulis mengkobinasikan antara algoritma Decision Tree classifier yang memiliki kemampuan baik untuk mengolah database yang besar sebagai feature selection kemudian dengan Metode SVM yang layak dan relevan digunakan dalam menganalisis dan mengdiaknosa Pasien Kanker Payudara payudara karena memiliki hasil yang akurat atas permasalahan yang ada dan beberapa landasan.

Kata Kunci— Data Mining, diagnosis, Decision Tree, Metode SVM

I. PENDAHULUAN

Dalam era yang semakin berkembang ini, penggunaan data mining semakin banyak dalam berbagai bidang dan menjadi bagian dari perkembangan teknologi informasi yang tak terhindarkan. Setiap hari selalu terkumpul sejumlah besar data, di mana data - data tersebut perlu untuk dianalisa. Data tersebut berasal dari berbagai bidang seperti bisnis, keuangan, kesehatan, ilmu pengetahuan dan teknologi, dan hampir semua aspek kehidupan manusia.

Dengan banyaknya data yang tersedia tersebut, dibutuhkan metode untuk mengetahui pola dan informasi bermakna yang terkandung di dalam data tersebut secara cepat, efisien , dan mudah dipahami. Proses demikian menjadi cikal bakal dari data mining. Data mining merupakan proses penggalian informasi yang bermakna dari sejumlah besar data, dengan melalui berbagai prosedur dan metode.

Disebutkan Gupta dkk . (2011) dalam bidang kedokteran, penggunaan data mining mempunyai peranan yang cukup penting dan evolusioner yang dapat mengubah cara pandang para dokter, praktisi dan peneliti kesehatan dalam melakukan proses deteksi penyakit kanker payudara pada seorang pasien. Terdapat 2 aplikasi klasifikasi di dalamnya, yaitu proses diagnose (diagnosis) sel kanker yang membedakan antara tumor (kanker jinak) atau kanker ganas dan proses prognosa (prognosis) untuk mengetahui kemungkinan munculnya kembali sel

kanker pada pasien yang telah dioperasi di masa yang akan datang. Penelitian dalam bidang ini,

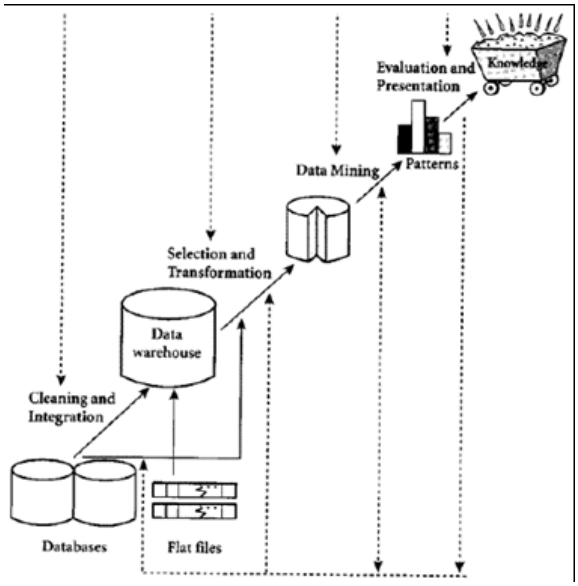
Pada tahun 2012 kasus kanker payudara di Indonesia mencapai kurang lebih 40 kasus untuk setiap 100.000 penduduk dan kejadian itu meningkat pada setiap tahunnya (Menurut Data WHO). Data terbaru yang didapatkan oleh Riset Kesehatan Dasar 2013, kasus kematian pasien penyakit kanker payudara meningkat menjadi kasus kematian tertinggi dengan angka 21,5 pada setiap 100.000 penduduk. Dibandingkan dengan malaysia, di Indonesia penderita kanker payudara lebih dominan diderita oleh wanita berusia muda dan pada tahap yang lebih lanjut. Angka kematian yang terus meningkat dari kasus Kanker Payudara membutuhkan perhatian khusus untuk upaya pencegahan dini dan penanggulanan dengan diagnosa gejala awal kanker payudara. Diagnosis dini penyakit kanker payudara atau breast cancer dapat dilakukan dengan metode data mining. Dimana proses data mining bertujuan untuk menguraikan temuan baru di dalam dataset dan menjelaskan suatu proses yang menggunakan teknik statistik, matematis, artificial intelligence, dan machine learning untuk melakukan ekstrak dan identifikasi informasi yang bermanfaat dan pengetahuan yang terkait dari database tersebut. Klasifikasi dengan data mining dapat dilakukan menggunakan beberapa metode yaitu Decision Tree, K-Nearest Neighbor, Naive Bayes, ID3, CART, Linear Discriminant Analityc dan lain sebagainya yang tentunya memiliki kelebihan dan kekurangan masing - masing. Namun pada penelitian kali ini, penulis berfokus pada klasifikasi data mining dengan menggunakan algoritma Support Vector Mechine dan Deccision Tree.

Penelitian ini akan menganalisis Breast Cancer Winconsin Original data set yang diperoleh dari UCI Machine Learning Repository (repositori data penelitian) untuk melakukan klasifikasi keganasan kanker payudara. Kali ini penulis mengkobinasikan antara algoritma Decision Tree classifier yang memiliki kemampuan baik untuk mengolah database yang besar sebagai feature selection kemudian dengan Metode SVM yang layak dan relevan digunakan dalam menganalisis dan mengdiaknosa Pasien Kanker Payudara payudara karena memiliki hasil yang akurat atas permasalahan yang ada dan beberapa landasan yang telah diterangkan diatas.

II. METODE PENELITIAN

Data mining terbagi beberapa kelompok sesuai dengan tugas yang

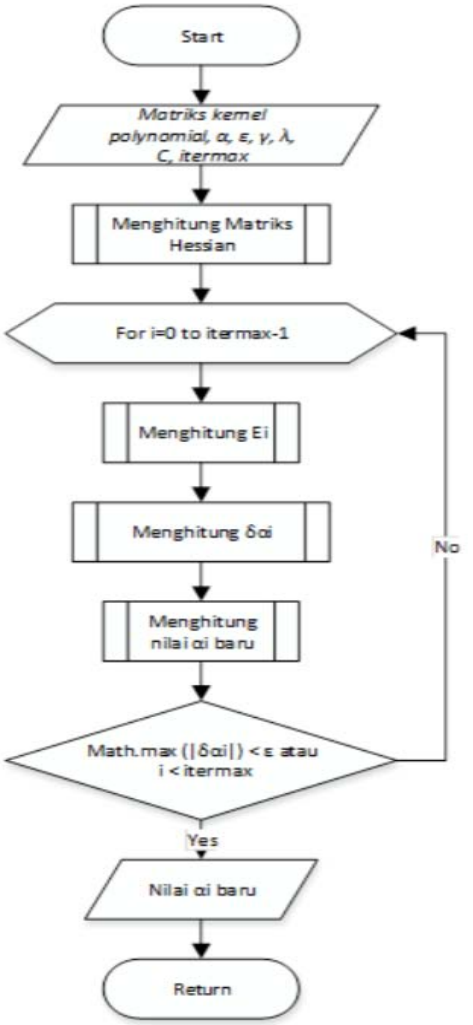
dilakukannya, yaitu, yaitu : Deskripsi, Estimasi, Prediksi, Klasifikasi, Pengklusteran, Asosiasi.



Gambar 1. Tahapan Data Mining

Untuk menghilangkan data yang tidak diperlukan, data yang diperoleh dari tahap pengambilan dataset akan disaring untuk menghasilkan data yang benar-benar dibutuhkan. umumnya data tersebut memiliki nilai yang tidak sempurna seperti data yang hilang. Selain itu, ada juga atribut-atribut data yang tidak sesuai dengan pemrosesan data mining yang akan digunakan. Data-data yang tidak relevan itu juga lebih baik dibuang karena keberadaannya bisa mengurangi mutu atau akurasi dari hasil data mining nantinya. Pembersihan data juga akan mempengaruhi performasi dari sistem data mining karena data yang ditangani akan berkurang jumlah dan kompleksitasnya.

Proses Sequential Training SVM adalah proses yang digunakan untuk mencari nilai hyperplane terbaik. Proses ini dimulai dari inisialisasi nilai konstanta, perhitungan matriks hessian, perhitungan Ei , $\delta\alpha i$ hingga menghitung nilai αi yang baru. Hasil akhir dari proses training ini adalah berupa nilai αi yang akan digunakan untuk proses pengujian. Gambar 1 akan memperlihatkan diagram alir proses SVM.



Gambar 2. Diagram Alir Squential Training SVM

III. HASIL DAN PEMBAHASAN

A. Evaluasi dan Validasi

Tahap pengujian data yang dilakukan menggunakan RapidMiner 7. dengan tujuan untuk melihat nilai akurasi, pohon keputusan, dan rule sebagai seleksi fitur. Pada model klasifikasi dapat diketahui hasil evaluasi berdasarkan pada banyaknya dataset record yang diklasifikasi secara benar atau tidak benar pada model klasifikasi tersebut. Dari 683 record akan dilakukan pengujian sebanyak 1 kali. Pembagian pengujian dengan data training dan testing yang berbeda. Data training 70% dan data testing 30% Data training digunakan untuk membentuk model, sedangkan data testing digunakan untuk menguji ketepatan klasifikasi dari model yang telah dibentuk. Berikut merupakan tampilan proses import Data diagnosa didapat dari Universitas California Irvene (UCI) Machine Learning Repository yaitu Breast Cancer Winconsin Original Data Set yang berjumlah 683 record. Dalam bentuk file Microsoft Excel dalam rapidminer 5.3:

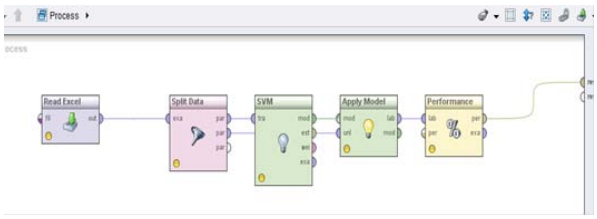
Ketebalan Gumpal	Keseragaman Ukuran S	Keseragaman bentuk S	Marginal Adhesi
5	1	1	1
5	4	4	5
3	1	1	1
6	8	8	1
4	1	1	3
8	10	10	8
1	1	1	1
2	1	2	1
2	1	1	1
4	2	1	1
1	1	1	1
2	1	1	1
5	3	3	3
1	1	1	1

Ukuran Sel Epitel Tungg	Ukuran Asli Nuclei	Kromatin	Keadaan Nucleoli Norm	Mitosis	Cl
2	1	3	1	1	B
7	10	3	2	1	B
2	2	3	1	1	B
3	4	3	7	1	B
2	1	3	1	1	B
7	10	9	7	1	M
2	10	3	1	1	B
2	1	3	1	1	B
2	1	1	1	5	B
2	1	2	1	1	B
1	1	3	1	1	B
2	1	2	1	1	B
2	3	4	4	1	M
2	3	3	1	1	B

Gambar. 3 Proses import Data diagnose

B. Pengujian

Pengujian yang pertama data set dari jumlah data 683 record dengan 9 atribut dan 1 sebagai Class. Berikut ini merupakan model algoritma menggunakan RapidMiner 5.3 sebelum dilakukan Feature Selection oleh Deccision Tree dengan algoritma SVM sebagai Model klasifikasi :



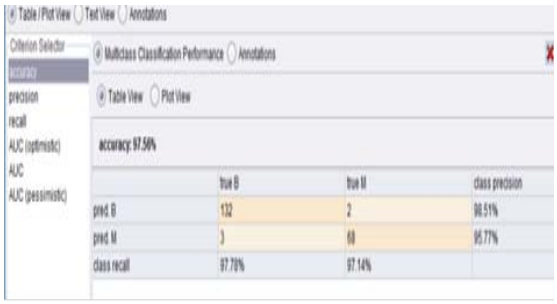
Gambar 4 Model Algoritma Klasifikasi SVM

Split Data digunakan sebagai ratio pembanding antara data training dan data testing. Data training digunakan untuk membentuk model, sedangkan data testing digunakan untuk menguji ketepatan klasifikasi dari model yang telah dibentuk.

C. Analisis Hasil

Berikut merupakan analisis hasil Pengujian berdasarkan jumlah data yang diuji dari sumber data training dan data testing. Hasil dari data keseluruhan 683

record dari dataset kanker payudara. Dengan 9 Atribut dan 1 atribut sebagai Class dapat dilihat melalui gambar berikut dibawah ini.



Gambar 5. Analisis Hasil akurasi SVM Sebelum Feature Selection

Klasifikasi SVM pada penelitian ini menggunakan aplikasi program rapidminer dengan Data training dan testing sebesar 70:30. Ketepatan klasifikasi terbesar yang dihasilkan oleh metode SVM dari partisi data Training dan testing 70:30 yaitu sebesar 99,02 %. Itu artinya tingkat keakurasian meningkat dan dengan adanya seleksi fitur menggunakan algoritma Decision Tree maka dapat diimprove tingkat keakurasianya meskipun dengan metode yang sama yaitu Support Vector Mechine (SVM).

Gambar 6. Nilai AUC

Nilai AUC sebelum dilakukan seleksi fitur menunjukan angka 1.000 itu artinya bisa dikatakan excellent classification.

IV. Simpulan

Berdasarkan hasil dan pembahasan yang telah dilakukan, dapat disimpulkan bahwa performansi akurasi klasifikasi terbaik dimiliki oleh SVM yaitu sebesar 99,02% dengan menggunakan partisi 30:70 dengan menggunakan algoritma decision tree sebagai seleksi fitur. Dataset public masih terdapat eror sehingga diharapkan nantinya akan diperoleh analisis yang lebih tepat. Selain itu untuk metode Support Vector Machine dalam pengimprove akurasi dapat mengubah parameter sampling linier disesuaikan dengan dataset. Parameter SVM sebaiknya tidak menggunakan trial and erroragar efisien dan menghasilkan akurasi yang optimum. Namun apabila data missing value tersebut tidak dapat dihindarkan maka untuk penelitian selanjutnya dapat dilakukan pengembangan metode SVM untuk data missing value dan penentuan parameter SVM tanpa trial and error yang diharapkan nantinya akan memberikan akurasi yang lebih tinggi.

DAFTAR PUSTAKA

[1]. Algoritma Data Mining.Yogyakarta: Andi Publishing.D. T. Larose,

Discovering Knowledge in Data: An Introduction to Data Mining. United States of America: John Wiley & Sons, Inc, 2005.

[2]. Technical and P. Series, Guidelines for management of breast cancer. World Health Organization, 2006.Kusrini, & Luthfi, E. T. (2009).

[3]. Gorunesco, Data Mining Concept Model Technique. Romania: Springer, 2011.

[4]. Laily Hermawanti, "Penerapan Algoritma Klasifikasi C4.5 Untuk Diagnosis Penyakit Kanker Payudara," *JURNAL SAINS DAN SENI ITS Vol. 7 No. 2, Hal 57-64 Maret 2012 Vol. 7 No. 2, Hal 57-64*

[5]. Larose, D., 2006, Data Mining Mathod And Model, Canada, Inc. Hoboken, New Jersey.

[6]. J. Han and M. Kamber, Data Mining Concept dan Techniques, 2nd ed. United States of America: Diane Cerra, 2006.

[7]. Fitria -, Hariyanto Wibowo, Feven Indriyani, (2018), K-Nearest Neighbor Method For Monitoring of Production And Preservation Information (Treatmeant) of Rubber Tree Plant, prosiding International Conference on Information Technology And Business (ICTB) 4, th 2018 pp.29 - 44

[8]. Jiawei Han and Micheline Kamber, Data Mining: Concepts and Techniques Third Edition, *Elsevier*, 2012

[9]. Ian H. Witten, Frank Eibe, Mark A. Hall, Data mining: Practical Machine Learning Tools and Techniques 3rd Edition, *Elsevier*, 2011

[10]. Markus Hofmann and Ralf Klinkenberg, RapidMiner: Data Mining Use Cases and Business Analytics Applications, *CRC Press Taylor & Francis Group*, 2014

[11]. Daniel T. Larose, Discovering Knowledge in Data: an Introduction to Data Mining, *John Wiley & Sons*, 2005

[12]. Ethem Alpaydin, Introduction to Machine Learning, 3rd ed., *MIT Press*, 2014

[13]. Florin Gorunescu, Data Mining: Concepts, Models and Techniques, *Springer*, 2011

[14]. WHO. (2005). Data penderita kanker payudara di dunia. Dikases pada tanggal 3 Februari 2012 dari [http://www.who.int/cancer/dete-ction/braestcancer/en/index1.html].

[15]. Dinas Kesehatan Nasional.(2007). Data penderita kanker payudara di Indonesia. Diakses pada tanggal 31 januari 2011 dari

[http://www.depkes.go.id/index.php/berita/press-release/1060-jika-tidak-dikendalikan-26-juta-orang-di-dunia-menderita-kanker-.html]

[16]. Keles, A., Keles, A., dan Yavuz, U. (2011). Expert System Based On Neuro-Fuzzy Rules For Diagnosis Breast Cancer.

[17]. Expert Systems with Applications. 38. 5719–5726. [4] Purwantaka, R. I. (2010).[Tugas Akhir]

[18]. Faktor-Faktor Yang Mempengaruhi Resiko Penyebab Penderita Kanker Payudara Dengan Menggunakan Pendekatan Regresi Logistik. Surabaya: Institut Teknologi Sepuluh Nopember Surabaya.

[19]. Purnami, S. W., dan Embong, A. (2008). Smooth Support Vector Machine For Breast Cancer Classification.

[20]. The 4th IMT-GT 2008 Conference on Mathematics, Statistics, and Their Applications (ICMSA08), Banda Aceh, Indonesia.

[21]. Wang, D., Shi, L., dan Heng, P. A. (2009). Automatic Detectiom of Breast Cancer in Mammogrmas using Support Vector Machines. Neurocomputing 72.3296-3302.

[22]. Huang, C-L., Liao, H-C., dan Chen, M-C. (2008). Prediction Model Building and Feature Selection With Support Vector Machine. Expert System with Application 34. 578-587.

[23]. Ellis, E.O., Schnitt, S.J., S.-Garau, X., Bussolati, G., Tavassaoli, F.A., Eusebi, V. Pathology and Genetic of Tumours of The Breast and Female Genital Organs / WHO Classification of Tumours. Washington: IARC Press; 2003. P.10, 34-6.

[24]. Kardinah (2002). Penatalaksanaan Kanker Payudara Terkini oleh Penanggulangan & Pelayanan Kanker Payudara Terpadu Paripurna R.S. Kanker Dharmais. Jakarta: Pustaka Populer Obor.

[25]. Hosmer, D. W., dan Lemeshow, S. (2000). Applied Logistic Regression. New York: John Wiley & Sons, Inc.

[26]. Agresti, A. (2002). Categorical Data Analysis, Second Edition. John Willey & Sons, New York.

[27]. Santosa, B. (2006). Data Mining: Teknik Pemanfaatan Data Untuk Keperluan Bisnis. Yogyakarta: Graha Ilmu.