SISTEM PENDUKUNG KEPUTUSAN KLASIFIKASI TINGKAT KEGANASAN KANKER PAYUDARA DENGAN METODE *NAÏVE BAYES CLASSIFIER*

[1)]Yisti Vita Via, [2)]Budi Nugroho, [3)]Alfian Syafrizal
[1,2,3)]Program Studi Teknik Informatika Fakultas Teknologi Industri
Universitas Pembangunan Nasional "Veteran" Jawa Timur
Jl. Raya Rungkut Madya, Gunung Anyar, Surabaya, Jawa Timur 60294
[1)]yistivita@gmail.com, [2)]budinug@gmail.com, [3)]alfian.syafrizal@gmail.com

Abstrak. *Kanker payudara merupakan salah satu jenis kanker yang sering ditemukan pada kebanyakan wanita. Kanker ini ditandai dengan sel-sel abnormal yang tumbuh di luar kendali pada payudara. Hal ini menunjukkan bahwa kanker payudara adalah penyakit yang sangat ganas dan karenanya memerlukan pemeriksaan intensif dengan mendeteksi dini tingkat keganasan kanker payudara. Penelitian ini menganalisis tentang pengelompokan data kanker payudara untuk mengetahui kanker tersebut termasuk kanker jinak atau kanker ganas. Penelitian ini menggunakan 9 atribut sebagai masukan sistem dan data set yang digunakan adalah data set publik Breast Cancer Wisconsin Original (WBCO) yang diambil dari UCI Machine Learning. Untuk mengklasifikasi tingkat keganasan dapat dilakukan dengan pemanfaatan bioinformatic dengan menggunakan teknik data mining salah satuya adalah algoritma Naive Bayes Classifier (NBC). Dari hasil pengujian dengan confusion matrix diketahui bahwa NBC yang diterapkan untuk melakukan klasifikasi tingkat keganasan kanker payudara memiliki akurasi pola yang cukup besar yaitu 97,82%, sedangkan error rate yang dihasilkan sebesar 2,18%. Hasil penelitian ini menunjukan bahwa dengan error rate yang cukup kecil maka algoritma Naïve Bayes Classifier terbukti cukup bagus untuk melakukan klasifikasi pada data WBCO.*

Kata Kunci : Breast Cancer, Metode Naïve Bayes Classifier, Sistem Pendukung Keputusan

Perkembangan teknologi informasi dan komunikasi saat ini berkembang dengan begitu cepat, sehingga merambah dalam kehidupan manusia tidak terkecuali bidang kesehatan dan kedokteran. Kanker payudara atau *Breast Cancer* merupakan salah satu jenis kanker yang sering ditemukan pada kebanyakan wanita. Kanker payudara terjadi karena pertumbuhan berlebih atau perkembangan yang tidak terkendali dari sel-sel jaringan payudara [1].

Berdasarkan uraian di atas maka penulis membuat sebuah aplikasi yang bisa diterapkan dalam bidang kedokteran yaitu sistem pendukung keputusan untuk mengklasifikasi tingkat keganasan kanker payudara. Dengan sistem yang akan dibangun ini diharapkan para tenaga medis bisa lebih mudah dalam melakukan klasifikasi tingkat keganasan kanker payudara.

Metode yang digunakan untuk mengklasifikasi tingkat keganasan kanker payudara yang merupakan keluaran sistem yaitu menggunakan metode *Naïve Bayes Classifier*. Disebut juga *Bayesian Classification* karena merupakan metode pengklasifikasian statistik yang dapat digunakan untuk memprediksi probabilitas keanggotaan dari suatu *class. Naïve Bayes Classifier* didasarkan pada teorema *Bayes* yang memiliki kemampuan klasifikasi serupa *decision tree* dan *neural network*. Selain itu, *Naïve Bayes Classifier* terbukti memiliki akurasi dan kecepatan yang tinggi saat diaplikasikan ke dalam *database* yang besar [2]. Dari hasil penelitian Bellaachia,dkk algoritma *Naïve Bayes Classifier* untuk penentuan tingkat keganasan kanker payudara hasil akurasinya masih kurang dibanding menggunakan algoritma C4.5. Namun, *Naïve Bayes Classifier* mempunyai akurasi dan kecepatan yang tinggi saat diterapkan pada data yang besar. NBC dapat menangani data yang tidak lengkap (*missing value)* serta kuat terhadap atribut yang tidak relevan dan *noise* pada data . *Naïve Bayes Classifier* akan bekerja lebih efektif jika dikombinasikan dengan beberapa prosedur pemilihan atribut [3].

Sedangkan implementasi pembuatan aplikasinya menggunakan bahasa pemprograman PHP. Dan penyimpanan datanya menggunakan MySQL. Dengan adanya program ini diharapkan bisa membantu dokter dan tenaga medis dalam melakukan klasifikasi.

I. Metodologi

Sistem Pendukung Keputusan

Sistem Pendukung Keputusan (SPK) didefinisikan sebagai suatu sistem berbasis komputer yang interaktif, membantu pengambilan keputusan dengan menggunakan analisa data – data dan model – model, guna menyelesaikan permasalahan yang semi terstruktur maupun masalah yang tak terstruktur. Menurut konsep yang dikemukakan oleh Michaet S. Scoot Morton dengan istilah *"Management Decision System"*. Karakteristik sistem pendukung keputusan adalah sebagai sistem berbasis komputer yang interaktif yang mendukung manajemen pengambilan keputusan melalui pemanfaatan data dan model untuk mengambil keputusan mengenai masalah yang semi terstruktur [4].

Dengan pengertian di atas dapat dijelaskan bahwa sistem pendukung keputusan bukan alat pengambil keputusan melainkan sistem yang membantu pengambil keputusan dengan mengungkapkan informasi dari data yang sudah diolah secara relevan dan diperlukan untuk membuat keputusan tentang suatu masalah dengan lebih cepat dan lebih akurat, sehingga sistem ini tidak dimaksudkan untuk menggantikan pengambilan keputusan dan proses pembuatan keputusan.

Dapat juga dikatakana sebagai sistem komputer yang mengolah data menjadi informasi untuk mengambil keputusan dari masalah semi terstruktur yang spesifik.

SPK terdiri dari kombinasi *relational database*, *knowledge base* dan *multidimensional database*. Ketiga jenis *database* ini saling berkolaborasi untuk menghasilkan sebuah keputusan yang digunakan oleh *manager*.

Tahapan SPK:

a. Definisi masalah.
b. Pengumpulan data atau elemen informasi yang relevan.
c. Pengolahan data menjadi informasi baik dalam bentuk laporan grafik maupun tulisan.
d. Menentukan alternatif-alternatif solusi (bisa dalam persentase.)

Tujuan dari SPK antara lain:

a. Membantu menyelesaikan masalah semi-terstruktur.
b. Mendukung *manager* dalam mengambil keputusan.
c. Meningkatkan efektifitas dalam pengambilan keputusan.

Dalam pemrosesannya, SPK dapat menggunakan bantuan dari sistem lain seperti *Artificial Intelligence, Expert* Sistems, *Fuzzy Logic*, dll (Hermawan, 2002).

Metode Naïve Bayes Classifier

Bayes merupakan teknik klasifikasi berbasis *probabilistic* sederhana yang berdasar pada penerapan teorema *Bayes* (aturan *Bayes*) dengan asumsi independensi yang kuat (*Naïve*) dengan kata lain *Naïve Bayes Classifier*. Model yang digunakan adalah model fitur independen.

Dalam *Bayes* terutama *Naïve Bayes Classifier*, maksud independen yang kuat dalam fitur adalah bahwa sebuah fitur pada data tidak berkaitan dangan ada atau tidaknya fitur lain pada data yang sama [5].

Klasifikasi *Bayes* didasarkan pada teorema *Bayes* dengan formula umum sebagai berikut :

$$P(B|A) = \frac{P(B|A)P(A)}{P(B)}$$

Keterangan :

P(B|A) = Peluang B jika diketahui kejadian A
P(A) = Peluang kejadian A
P(B) = Peluang kejadian B

Ide dasar dari aturan *Bayes* adalah bahwa hasil dari hipotesis atau peristiwa dapat diperkirakan berdasarkan pada beberapa bukti yang diamati. Ada beberapa hal penting dari aturan *Bayes* tersebut, yaitu :

a. Sebuah probabilitas awal / prior H atau P(H) adalah probabilitas dari suatu hipotesis sebelum bukti diamati.
b. Sebuah probabilitas akhir H atau P(H|E) adalah probabilitas dari suatu hipotesis setelah bukti diamati.

Naive Bayes Classifier disebut juga *Bayesian Classification* merupakan metode pengklasifikasian statistik yang dapat digunakan untuk mengklasifikasi probabilitas keanggotaan dari suatu *class*. *Naïve Bayes Classifier* didasarkan pada teorema *Bayes* yang memiliki kemampuan klasifikasi serupa *decision tree* dan *neural network*. Selain itu, *Naïve Bayes Classifier* terbukti memiliki akurasi dan kecepatan yang tinggi saat diaplikasikan ke dalam *database* yang besar. Berikut penjelasan mengenai metode *Naïve Bayes Classifier* :

Pendekatan *Bayes* pada saat klasifikasi adalah mencari probabilitas tertinggi (*Vmap*) dengan masukan atribut (a1,a2,a3,a4,…,an).

$$Vmap = \arg\max P(Vj|a1, a2, a3, \dots, an)$$

Keterangan :
Vmap = Probabilitas tertinggi
A1,a2,a3,…,an = Atribut (*input*)

Menggunakan teorema *Bayes* ini persamaan di atas dapat ditulis menjadi :

$$Vmap = \arg\max \frac{P(a1, a2, a3, \dots, an|Vj)|P(Vj)}{P(a1,a2,a3\dots,an)}$$

Keterangan :
Vmap = Probabilitas tertinggi
P(Vj) = Peluang *class* ke j
P(a1,a2,a3…,an)= Peluang atribut *input*
P(a1,a2,a3,…,an|Vj) = Peluang atribut *input* jika diketahui keadaan Vj ke j

Karena nilai P(a1,a2,a3,…,an) konstan untuk semua Vj, maka persamaan ini dapat ditulis menjadi :

$$Vmap = \arg\max P(a1, a2, a3, \dots, an|Vj)\, P(Vj)$$

Keterangan :
Vmap = Probabilitas tertinggi
P(Vj) = Peluang *class* ke j
P(a1,a2,a3,…,an|Vj) = Peluang atribut *input* jika diketahui keadaan Vj

Diberikan data dengan banyak atribut, ini akan menjadi komputasi yang kompleks untuk mengomputasi P(a1,a2,a3,…,an | Vj). Untuk mengurangi komputasi pada saat mengevaluasi P(a1,a2,a3,…,an | Vj), maka dapat dihitung menggunakan perhitungan .

$$P(a1, a2, a3, \dots, an|Vj) = \prod P(ai|Vj)$$

Keterangan :
P(a1,a2,a3,…,an|Vj) = Peluang atribut *input* jika diketahui keadaan Vj ke j
P(ai|Vj)= Peluang atribut ke i jika diketahui keadaan j

Dari rumus di atas sehinggan menghasilkan rumus seperti di bawah :

$$Vmap = \arg\max P(Vj) \prod P(ai|Vj)$$

Jika P(ai|Vj) sama dengan nol, maka menggunakan pendekatan estimasi sebagai berikut [6] :

$$P(ai|Vj) = \frac{nc+m.p}{n+m}$$

Keterangan :
nc = Kemunculan ai (atribut) terhadap vj
n = Jumlah kemunculan Vj pada dataset
p = Jumlah tiap Vj
m = Nilai konstan dari ukuran sampel yang equivalen.

P(Vj) dapat diartikan peluang diagnosa kategori j. dapat ditulis seperti : |diagnosa j| adalah jumlah diagnosa pada kategori j dan |tot_diagnosa| adalah jumlah diagnosa yang digunakan dalam data *training*. Sedangkan P(ai|Vj) adalah peluang *ai* jika diketahui keadaan Vj dalam artian, peluang ke i, dalam diagnosa kategori j. dapat ditulis seperti ini

$$P(Vj) = \frac{|Diagnosa\ j|}{|tot_{diagnosa}|}$$

Dimana *ni* adalah jumlah kemunculan atribut pada kategori Vj. Dan |Diagnosa_Vj| adalah jumlah gejala yang ada pada kategori Vj (Nirmala, 2010).

$$P(ai|Vj) = \frac{ni}{|Diagnosa_{Vj}|}$$

Skenario Uji Coba

Untuk memastikan bahwa aplikasi ini berjalan dengan lancar, penulis akan menyusun skenario uji coba, di antaranya :

a. Membagi keseluruhan dataset yang berjumlah 683 *record* menjadi dua bagian yaitu 546 *record* sebagai data *training* dan 137 *record* sebagai data *testing*.
b. Cara pembagian data yaitu dengan mengambil 80% data yang mempunyai *class* jinak untuk dijadikan data *training* dan 20% sisanya menjadi data *testing*, 80% data yang mempunyai *class* ganas dijadikan data *training* dan 20% sisanya dijadikan data *testing.*
c. Memasukkan 546 *record* ke dalam data *training.*
d. Melakukan *learning* terhadap seluruh data *training* untuk memperoleh pola.
e. Melakukan *testing* dengan menggunakan data *training* sebanyak 546 *record.*

f. Melakukan *testing* dengan menggunakan data *Testing* sebanyak 137 *record*.
g. Membandingkan hasil dari data *testing* dengan hasil dari perhitungan *Naïve Bayes*.
h. Setelah seluruh data *testing* dimasukan selanjutnya adalah menghitung nilai presentase akurasi sistem dan *error* sistem.

II. Hasil dan Pembahasan

Perhitungan Dengan Metode Naïve Bayes

Diketahui sebuah tuple x tanpa class sebagai berikut: Tuple x= (5,1,3,1,2,1,3,1,1), Class= ? . Dari tuple tersebut maka akan di cari classnya dengan menggunakan metode naïve bayes sebagai berikut:

- Menghitung jumlah data jinak dan ganas, menghitung jumlah atribut 1-9 dengan class jinak, menghitu- ng jumlah atribut 1-9 dengan class ganas pada data *training*.
  Jumlah jinak = 355
  Jumlah ganas = 191
- Menghitung probabilitas jinak dan ganas.
  P(jinak) = 355 / 191 + 355 = 0,650
  P(ganas) = 191 / 191 + 355 = 0,3498
- Menghitung probabilitas masing – masing atribut yang mempunyai class jinak.
  P(*Clumpthicnkes* = 5 |Jinak) = 63 / 355 = 0,1775
  P(*Uniformity of cell size* = 1 |Jinak) = 291 / 355 = 0,8197
  P(*Uniformity of cell shape* = 3 |Jinak) = 275 / 355 = 0,0648
  P(*Marginal adhesion* = 1 |Jinak) = 289 / 355 = 0,8141
  P(*Single epithelial cell size* = 2 |Jinak) = 277 / 355 = 0,7803
  P(*Bare nuclei* = 1 |Jinak) = 304 / 355 = 0,8563
  P(*Bland chromatin* = 3 |Jinak) = 177 / 355 = 0,3296
  P(*Normal nucleoli* = 1 |Jinak) = 307 / 355 = 0,8648
  P(*Mitosis* = 1 |Jinak) = 345 / 355 = 0,9718
- Menghitung probabilitas masing – masing atribut yang mempunyai class ganas.

  P(*Clumpthicnkes* = 5 |Ganas) = 31 / 191 = 0,1623
  P(*Uniformity of cell size* = 1 | Ganas) = 4 / 191 = 0,0209
  P(*Uniformity of cell shape* = 3 | Ganas) = 2 / 191 = 0,1099
  P(*Marginal adhesion* = 1 | Ganas) = 27 / 191 = 0,1414
  P(*Single epithelial cell size* = 2 | Ganas) = 21 / 191 = 0,1099
  P(*Bare nuclei* = 1 | Ganas) = 11 / 191 = 0,0576
  P(*Bland chromatin* = 3 | Ganas) = 36 / 191 = 0,1885
  P(*Normal nucleoli* = 1 | Ganas) = 31 / 191 = 0,1623
  P(*Mitosis* = 1 | Ganas) = 102 / 191 = 0,5340

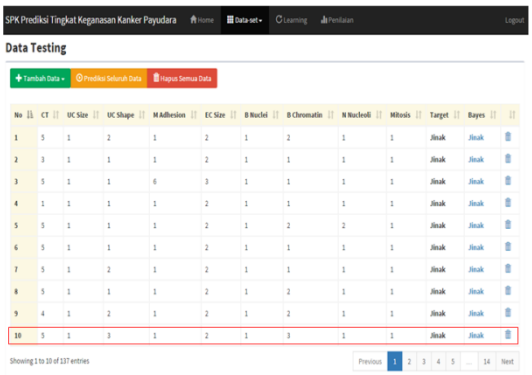
- Menghitung probabilitas dari ke 9 atribut yang mempunyai *class* Jinak.
  P(P(*clump thickness*|jinak), P(*Uniformity of cell size*|jinak), … ,P(*Mitosis*|jinak) | P(Jinak)) = (0,1775 * 0,8197 * 0, 0648 * 0,8141 * 0,7803 * 0,8563 * 0,3296 * 0,8648 * 0,9718) * 0,6502 = 0,00092367746422448

- mencari probabilitas dari ke 9 atribut yang mempunyai *class* Ganas.
  P(P(*clump thickness*|ganas), P(*Uniformity of cell size*|ganas), … ,P(*Mitosis*|ganas) | P(ganas)) = (0,1623 * 0,0209 * 0,1099 * 0,1414 * 0,1099 * 0,0576 * 0,1885 * 0,1623 * 0,5340) * 0,3498 = 1,906878491614E-9

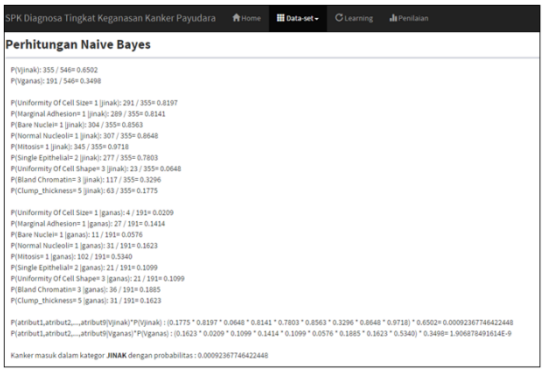
Dari perhitungan di atas diketahui bahwa probabilitas class jinak lebih besar daripada probabilitas class ganas sehingga tuple x termasuk dalam class JINAK.

Hasil Perhitungan Program

Uji coba akan dilakukan dengan mengambil salah satu *record* secara acak yang ada pada data *testing*. Dan di sini *record* yang akan digunakan untuk uji coba adalah *record* ke-10 dengan nilai record di dalamnya= (5,1,3,1,2,1,3,1,1).



Gambar 1. Uji Coba Record ke-10



Gambar 2. Hasil Perhitungan Sistem Record ke-10

Pada gambar 2 telah di hasilkan nilai probabilitas dari masing – masing class dengan nilai probabilitas class jinak sebesar 0,000923677464224 48 dan nilai probabilitas dari class ganas sebesar 1,906878491614E-9. Dengan hasil tersebut sistem mengklasifikasi bahwa *record* 10 adalah termasuk kanker jinak karena probabilitas dari kanker jinak lebih besar dari probabilitas dari kanker ganas.

Akurasi Data Training

Dari hasil pengujian yang dilakukan terhadap data *training* yang berjumlah 546 data. Dengan keputusan target ganas berjumlah 191 data pada tabel 4.1 dapat dilihat 189 data diidentifikasi dengan benar dan 2 data yang diidentifikasi salah. Dan dengan keputusan target jinak berjumlah 355 data pada tabel 4.1 dapat dilihat 11 data yang diidentifikasi salah dan 344 data yang diidentifikasi benar. Sehingga akurasi pola data *training* yang diperoleh sebesar 97,62% dan kesalahan sebesar 2,38%.

Tabel 1. Akurasi Data Training

	Naïve Bayes=Ganas	Naïve Bayes=Jinak
Ganas=191	189	2
Jinak= 355	11	344

- Akurasi = (189 + 344) / (191 + 355) * 100% = 97.62 %
- Kesalahan = (11 + 2)/(191 + 355) * 100% = 2.38 %

Akurasi Data Testing

Dan dari perhitungan akurasi dan kesalahan data *testing* yang berjumlah 137 data. Dari tabel 4.2 dapat dilihat perhitungan yang diperoleh mempunyai akurasi yang baik dalam melakukan *class* tingkat keganas kanker payudara. Pada tabel di atas *class* ganas memiliki data sebanyak 48, diidentifikasi benar oleh sistem sebanyak 47 data dan diidentifikasi salah oleh sistem sebanyak 1 data. Dan sesuai dengan tabel di atas *class* jinak memiliki data sebanyak 89, diidentifikasi salah oleh sistem sebanyak 2 data dan diidentifikasi benar oleh sistem sebanyak 87 data. Sehingga akurasi yang dihasilkan dari data *testing* ini cukup tinggi yaitu sebesar 97,81% dan kesalahan sebesar 2,19%. Dari akurasi sebesar 97,81% maka sistem dan pola yang dibangun sudah cukup bagus dalam melakukan pengklasifikasian atau *class* terhadap keganasan kanker payudara.

Tabel 2. Akurasi Data Testing

	Naïve Bayes=Ganas	Naïve Bayes=Jinak
Ganas=48	47	1
Jinak= 89	2	87

- Akurasi = (47 + 87) / (48 + 89) * 100% = 97.81 %
- Kesalahan = (2 + 1)/(48 + 89) * 100% = 2.19 %

III. Simpulan

1. Dari uji coba perbandingan perhitungan yang dilakukan oleh sistem dengan perhitungan manual, maka bisa disimpulkan bahwa Metode *Naïve Bayes Classifier* telah dapat digunakan untuk mengklasifikasi tingkat keganasan kanker payudara dengan masukan 9 atribut yang berasal dari *UCI Machine Learning*.
2. Akurasi yang dihasilkan dari pemodelan metode *Naïve Bayes Classifier* yaitu sebesar 97,82% dari total 137 *record* data *testing*. Sedangkan akurasi yang dihasilkan dari pembuatan pola data *training* dengan data sebanyak 546 *record* sebesar 97,62%.
3. Dengan menggunakan optimasi pada atribut yang memiliki nilai probabilitas 0, bisa membantu sistem dalam melakukan klasifikasi dengan lebih akurat. Karena dengan optimasi tersebut atribut yang bernilai 0 akan tetap diikutkan dalam proses perhitungan sehingga atribut lainya tidak dianggap 0. 90

IV. Daftar Pustaka

[1] Febrida, Melly,2013, WHO: Jumlah Kematian akibat Kanker di Dunia Meningkat. Liputan6. [Online] ,(http://health.liputan6.com

/read/776217/who-jumlah-kematian-akibat-kanker-di-dunia-meningkat,diakses tanggal 03 Mei 2015).

[2] Ammar,S.Muhammad.,2009,. "KEOPTIMALAN NAÏVE BAYES DALAM KLASIFIKASI", Program Studi Ilmu Komputer. Fakultas Pendidikan Matematika Dan Ilmu Pengetahuan Alam. Universitas Pendidikan Indonesia.

[3] Bellaachia, Abdelghani, dkk, 2006, Predicting Breast Cancer SurvivabilityUsing Data Mining Techniques.

[4] Turban, E., 1995. Decision Support and Expert System: Management Support System. Forth Edition. Prentice Hall International Inc. New Jersey.

[5] Nirmala, M., 2010. "Sistem Pakar Untuk Menentukan Makanan Diet Sehat Pada Penyakit Jantung Berdasarkan Golongan Darah Dengan Menggunakan Naive Bayes" Undergraduate thesis, Faculty of Industrial Technology. Universitas Pembangungan Nasional "Veteran" Jawa Timur.

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

[7] William, H. W., 1992, Uci Machine Learning Repository, [online], (http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Original%29, diakses tanggal 26 Maret 2015).