E-ISSN 2797-4278
P-ISSN 2797-4294

KLASIFIKASI KANKER PARU-PARU MENGGUNAKAN METODE *NAIVE BAYES*

[1]Eva Wulandari [2]Andreas Perdana
[1]Teknik Informatika STMIK Dharma Wacana Metro,
[2]STMIK Dharma Wacana Metro
[1]wulandarieva22@gmail.com , [2] andreas.perdana@dharmawacana.ac.id

ABSTRAK

*Kanker paru-paru merupakan penyebab utama kematian akibat kanker dibandingkan dengan jenis kanker lainnya. Penyakit ini disebabkan oleh perubahan sel yang terus tumbuh di luar kendali. Melalui klasifikasi, pola kanker paru-paru ditemukan. Metode klasifikasi yang umum digunakan adalah metode Naive Bayes. Naive Bayes adalah algoritma klasifikasi sederhana dengan kinerja dan akurasi tinggi. Proses klasifikasi naif Bayesian terdiri dari menghitung probabilitas sebelumnya dengan menjumlahkan kasus secara rinci dengan atribut data. Kemudian menghitung probabilitas posterior untuk menentukan kelas kasus baru saat memproses data. Selanjutnya hitung peluang keseluruhan untuk setiap kelas**. Dan akurasi dapat diperoleh dengan perhitungan menggunakan confusion matrix.Confusion matrik yang akan dihitung adalah accuracy, precision dan recall.*

Keyword: Algoritma, Kematian,, Penyakit

1.PENDAHULUAN

Kanker paru-paru merupakan penyebab utama kematian akibat kanker dibandingkan dengan jenis kanker lainnya. Data yang diperoleh menunjukkan bahwa 32% kematian akibat kanker pada pria dan 25% pada wanita disebabkan oleh kanker paru-paru. Sebagian besar kasus kanker paru-paru menyerang orang berusia antara 35 dan 75 tahun, dengan insiden tertinggi antara 55 dan 65 tahun (Aliya dkk., 2016). Dalam sebuah penelitian, Tommy dkk. (2022) menjelaskan bahwa kanker adalah penyakit yang disebabkan oleh perubahan sel yang menyebabkan pertumbuhan dan pembelahan sel tidak terkendali. Berdasarkan permasalahan tersebut, penulis tertarik untuk menerapkan metode klasifikasi Naive Bayes pada data mining dalam penelitian ini.Memuat latar belakang atau konteks penelitian,landasan teori jika diperlukan) atau hasil kajian pustaka yang menunjukan adanya kesenjangan temuan penelitian, wawasan rencana pemecahan masalah dan potensi kontribusinya bagi bidang ilmu dan juga memuat rumusan tujuan penelitian.

Berdasarkan latar belakang tersebut, maka rumusan masalah dalam penelitian ini adalah penerapan metode Naive Bayes untuk klasifikasi kanker paru-paru. Penerapan metode Naive Bayes bertujuan untuk mengklasifikasikan data secara akurat dan dengan akurasi yang tinggi.

Data mining adalah teknologi yang memungkinkan pengguna untuk menggunakan hubungan data sebagai dasar pengambilan keputusan dengan menemukan hubungan dari data yang tidak diketahui pengguna dan menyajikannya dalam format yang mudah dipahami. Data mining dapat dibagi menjadi beberapa kelompok berdasarkan tugas yang dapat mereka lakukan. mendeskripsikan, memperkirakan, memprediksi, mengklasifikasikan, mengelompokkan, dan mengasosiasikan (Larose, 2005).Berisi tentang teori utama yang menjadi fokus utamapenelitian anda.

2.2 Naive Bayes

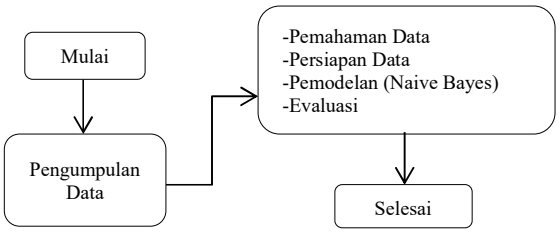
Metode Klasifikasi adalah proses menemukan model atau fungsi yang menggambarkan atau membedakan kelas data, dengan membentuk model yang dapat membagi data ke dalam kelas yang berbeda berdasarkan aturan atau fungsi tertentu, sehingga memungkinkan objek yang tidak ditentukan. (Fadlan dkk., 2018). Dalam penelitian ini akan mengklasifikasikan kanker paru-paru berdasarkan data yang diperoleh dan mencari keakuratan data. Metode yang digunakan dalam pengklasifikasian pada penelitian ini adalah Naive Bayes. Metode Naive Bayes dipilih karena menghasilkan akurasi tertinggi dengan data latih yang lebih sedikit (Handayani, 2015).

Metode Naive Bayes telah banyak digunakan oleh para peneliti sebelumnya dengan berbagai dataset. Dalam sebuah studi oleh Saptro dkk. (2019) menjelaskan algoritma pengujian untuk memprediksi waktu belajar siswa. Pengujian dalam penelitian ini dilakukan di AMIKOM Universitas Yogyakarta menggunakan metode algoritma Naive Bayes pada dataset 300 data dengan akurasi 68%, precision 61,3%, dan recall 65,3%. Sebuah studi oleh Annur (2018) menjelaskan klasifikasi penduduk miskin dengan metode Naive Bayes. Data kecamatan Tibawa menggunakan atribut umur, pendidikan, pekerjaan, pendapatan, kerabat dan status. Hasil pengujian mencapai akurasi 73%, precisioni 92%, dan recall 86%.

2.METODOLOGI

3.1 Tahapan Penelitian

Tahapan penelitian untuk klasifikasi kanker paru-paru dapat dilihat pada gambar berikut :



Gambar.1 Tahapan Penelitian

Tahapan penelitian dapat diuraikan sebagai berikut:

1. Pengumpulan Data

Dalam penelitian ini penulis menggunakan dataset lung cancer

E-ISSN 2797-4278
P-ISSN 2797-4294

yang didapat dari web kaggle, yang mana pengumpulan datanya dilakukan dengan menggunakan kuesioner. Data berupa file .csv sebanyak 306 data dan terdapat 16 atribut yaitu: *gender, anxiety, age, smoking, fatigue, yellow fingers, peer pressure, chronic disease, wheezing, alcohol consuming, coughing, shortness of breath, allergy, swallowing difficulty, chest paint, dan lung cancer.*

2. Pemahaman Data

Penelitian ini menggunakan data kaggle dengan 16 atribut. Data kaggle berupa data nominal dan data numerik. Data nominal adalah data yang mempunyai nilai berupa lambang atau nama dan tidak mempunyai urutan nilai. Dan data numerik adalah data kuantitatif yang dapat dihitung dan direpresentasikan sebagai bilangan bulat atau bilangan real. Yang termasuk data nominal yaitu *gender, smoking, anxiety, yellow fingers, peer pressure, chronic disease, wheezing, alcohol consuming, fatigue, coughing, shortness of breath, allergy, swallowing difficulty,* dan *chest paint.* Sedangkan atribut *Age* adalah numerik dan *lung cancer* atribut *class*. Penyelesaian menggunakan metode Naive Bayes dan diuji menggunakan Aplikasi RapidMiner.

3. Persiapan Data

Data yang digunakan dalam format .csv. Data yang akan diproses dimasukkan pada tabel data awal dan kemudian akan dilakukan perubahan pada beberapa jenis data.

4. Pemodelan

Langkah pertama dalam perhitungan Naive Bayes yaitu menghitung probabilitas prior, kemudian menghitung hipotesis class dengan melakukan perhitungan terhadap probabilitas pada kondisi tertentu atau probabilitas prior dengan menjumlahkan kasus secara terperinci tiap atribut data. Selanjutnya yaitu menghitungan probabilitas posterior yang berguna dalam penentuan kelas terhadap kasus baru dalam mengelola data. Setelah diketahui nilai tiap atribut probabilitasnya makan dapat dirumuskan dengan P(Ci|X). Langkah selanjutnya yaitu dilakukan perhitungan terhadap jumlah keseluruhan probabilitas tiap kelas.

Teorema bayes dapat ditulis menggunakan persamaan sebagai berikut :

P(Ci|X)=(P(X|Ci)*P(Ci))/(P(X)) .....Persamaan (1)

Keterangan :

P(Ci|X) : Probabilitas posterior dari Ci pada kondisi X (posterior probability)
P(X|Ci) : Probabilitas posterior dari X pada kondisi Ci
P(Ci) : Probabilitas prior Ci (prior probability)
P(X) : Probabilitas prior X
X : Data dengan class yang belum diketahui
Ci : Hipotesis data X merupakan class spesifik

5. Evaluasi

Evaluasi performa menggunakan *confusion matrix* untuk menghitung persentase *accuracy, precision* dan *recall*.*Confusion matrix* yaitu tabel yang menampilkan hasil klasifikasi jumlah data uji yang terprediksi benar dan jumlah data uji yang terprediksi salah dapat ditulisdengan rumus sebagai berikut:

$$Accuracy = \frac{TP+TN}{TP+TN+FN+FP}$$ .....Persamaan (2)

$$Recall = \frac{TP}{TP+FP}$$.....Persamaan (3)

$$Presisi = \frac{TP}{TP+FN}$$ .....Persamaan (4)

Keterangan :
TP = *True Positive*
TN = *True Negative*
FN = *False Negative*
FP = *False Positif*

3.HASIL DAN PEMBAHASAN

Dari 309 data kanker paru-paru akan dibagi menjadi dua data, yaitu: persentase 70% data train dan persentase 30% data test, dapat dilihat pada tabel 1 dan tabel 2 dibawah ini:

Tabel.1 Data Train

A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	A12	A13	A14	A15	*Predic. A*
M	1	69	1	2	1	2	2	2	2	2	1	2	2	2	*YES*
M	2	74	2	1	1	1	2	1	1	1	2	2	2	2	*YES*
F	1	59	1	1	2	2	2	1	2	1	1	2	2	1	*NO*
M	1	63	2	2	1	1	1	2	1	2	1	2	1	2	*NO*
F	1	63	1	2	1	2	1	1	2	1	1	1	2	1	*NO*
F	2	75	1	2	1	2	2	1	2	1	2	1	2	1	*YES*
M	1	52	2	1	1	2	2	1	2	2	1	2	2	1	*YES*
F	1	51	2	2	2	1	2	2	1	1	2	1	2	2	*YES*
F	1	68	2	1	1	1	2	2	1	1	1	1	1	1	*NO*
M	2	53	2	2	2	1	1	2	1	2	2	2	1	2	*YES*
F	2	61	2	2	2	2	2	2	2	1	1	1	2	2	*YES*
M	2	72	1	1	1	2	2	1	2	2	2	2	2	1	*YES*
F	1	60	2	1	1	1	2	1	1	1	1	1	2	1	*NO*
M	1	58	2	1	1	2	2	1	2	2	2	2	2	1	*YES*
M	1	69	2	1	1	2	1	1	2	2	2	2	1	1	*NO*
F	2	48	1	2	2	2	2	2	2	1	2	1	2	2	*YES*
M	2	75	2	1	1	2	1	1	2	2	2	2	2	1	*YES*
M	2	57	2	2	2	1	1	2	1	2	1	2	1	2	*YES*
F	2	68	2	2	2	2	2	2	1	1	1	1	2	1	*YES*
F	2	61	1	1	1	1	2	1	1	1	1	1	2	1	*NO*

Keterangan:

A1 = *Gender*
A2 = *Chronic Disease*
A3 = *Age*
A4 = *Smoking*
A5 = *Yellow Fingers*
A6 = *Peer Pressure*
A7 = *Coughing*
A8 = *Fatigue*
A9 = *Anxiety*
A10 = *Wheezing*
A11 = *Alcohol Consuming*
A12 = *Allergy*
A13 = *Chest Pain*
A14 = *Shortness Of Breath*
A15 = *Swallowing Difficulty*

E-ISSN 2797-4278
P-ISSN 2797-4294

Predic. A = *Lung Cancer*

Data train berguna untuk mengklasifikasikan kanker paru-paru pada kelas sentimennya (probabilitas class) dan jumlah datanya lebih banyak. Pada penelitian ini menggunakan data train sebanyak 216 data (70%) dari total data keseluruhan.

Tabel.2 Data Test

A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	A12	A13	A14	A15	*Predic.A*
F	2	67	2	2	2	1	2	2	1	1	1	1	2	2	*YES*
F	2	56	2	2	2	1	1	2	2	2	2	2	1	2	*YES*
F	2	70	1	1	2	2	1	1	1	2	2	1	2	1	*YES*
M	2	70	1	1	1	2	2	1	1	2	2	2	2	1	*YES*
F	2	57	1	1	2	1	2	2	2	2	2	2	2	2	*YES*
M	2	61	1	1	2	1	1	2	1	2	1	1	1	1	*NO*
F	2	77	1	1	2	1	2	1	2	2	2	2	1	1	*YES*
M	2	63	2	2	1	1	2	2	2	1	2	1	2	1	*YES*
F	1	62	2	1	2	2	2	1	2	2	2	2	1	2	*YES*
M	2	59	2	1	2	2	2	1	2	2	2	2	2	2	*YES*
F	2	70	1	2	2	2	2	1	1	2	2	1	1	1	*YES*
M	2	71	1	2	1	2	1	2	2	2	2	1	1	2	*YES*
F	1	56	1	2	2	2	2	1	1	1	1	2	2	1	*YES*
M	2	57	1	1	1	2	1	1	1	2	2	2	2	2	*YES*
M	2	78	1	2	2	1	2	1	2	1	1	2	2	2	*YES*
M	1	64	2	2	2	2	2	2	1	2	2	2	2	1	*YES*
M	2	62	1	2	2	2	2	1	2	2	2	2	1	2	*YES*
F	1	49	1	2	1	2	1	2	1	1	1	1	1	1	*YES*
M	1	77	1	2	2	2	2	1	2	2	2	1	1	1	*YES*
F	2	64	1	1	1	1	1	2	2	2	2	1	1	2	*YES*

Data test atau data uji berguna untuk mengetahui seberapa baik clssifier melakukan pengklasifikasian dengan benar. Jumlah data yang digunakan dalam penelitian ini sebanyak 93 data (30%) dari data keseluruhan.

Tabel.3 Probabilitas

KATEGORI/ATRIBUT	SUBSET	YES	NO
GENDER	M	0,540	0,407
	F	0,460	0,444
SMOKING	1	0,418	0,519
	2	0,582	0,481
YELLOW_FINGERS	1	0,402	0,593
	2	0,598	0,407
ANXIETY	1	0,476	0,704
	2	0,524	0,296
PEER_PRESSURE	1	0,476	0,704
	2	0,524	0,296
CHRONIC DISEASE	1	0,476	0,630
	2	0,524	0,370
FATIGUE	1	0,312	0,593
	2	0,688	0,407
ALLERGY	1	0,402	0,926
	2	0,598	0,074
WHEEZING	1	0,418	0,778
	2	0,582	0,222
ALCOHOL CONSUMING	1	0,386	0,815
	2	0,614	0,185
COUGHING	1	0,386	0,704
	2	0,614	0,296
SHORTNESS OF BREATH	1	0,365	0,444
	2	0,635	0,556
SWALLOWING DIFFICULTY	1	0,481	0,852
	2	0,519	0,148
CHEST PAINT	1	0,392	0,704
	2	0,608	0,296

E-ISSN 2797-4278
P-ISSN 2797-4294

Nilai probabilitas tiap class diperoleh dengan membagi jumlah class yes tiap atribut dibagi jumlah yes keseluruhan data. Dan probabilitas class no tiap atribut dibagi jumlah no keseluruhan data.

Tabel.4 Hasil Klasifikasi

A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	A12	A13	A14	A15	*Predic. A*	*YES*	*NO*	*Predic. B*
F	2	67	2	2	2	1	2	2	1	1	1	1	2	2	*YES*	1	0	*YES*
F	2	56	2	2	2	1	1	2	2	2	2	2	1	2	*YES*	1	0	*YES*
F	2	70	1	1	2	2	1	1	1	2	2	1	2	1	*YES*	1	0	*YES*
M	2	70	1	1	1	2	2	1	1	2	2	2	2	1	*YES*	1	0	*YES*
F	2	57	1	1	2	1	2	2	2	2	2	2	2	2	*YES*	1	0	*YES*
M	2	61	1	1	2	1	1	2	1	2	1	1	1	1	*NO*	0	1	*NO*
F	2	77	1	1	2	1	2	1	2	2	2	2	1	1	*YES*	1	0	*YES*
M	2	63	2	2	1	1	2	2	2	1	2	1	2	1	*YES*	1	0	*YES*
F	1	62	2	1	2	2	2	1	2	2	2	2	1	2	*YES*	1	0	*YES*
M	2	59	2	1	2	2	2	1	2	2	2	2	2	2	*YES*	1	0	*YES*
F	2	70	1	2	2	2	2	1	1	2	2	1	1	1	*YES*	1	0	*YES*
M	2	71	1	2	1	2	1	2	2	2	2	1	1	2	*YES*	1	0	*YES*

Dari hasil klasifikasi diatas dapat diketahui bahwa pada data test pertama kelas prediksi sesuai dengan kelas sebenarnya (P) yaitu yes. Kelas prediksi adalah kelas hasil prediksi dari proses klasifikasi menggunakan Naive Bayes, sedangkan kelas sebenarnya adalah kelas dengan sentimen yang telah ditentukan sebelumnya.

Tabel.5 Hasil *Confusion Matrix*

	True Yes	*True No*	*Class precision*
Predic. Yes	80	4	95,24%
Predic. No	1	8	88,89%
Class recall	98,77%	66,67%	

Kemudian menghitung dan mengevaluasi dengan menghitung confusion matrixnya. Setelah menghitung confusion matrixnya, akan terlihat tingkat akurasi sebesar 94,62% dan klasifikasi Yes memprediksi "Yes" sebanyak 80 dan hasil prediksi "Yes" adalah 4 yang merupakan klasifikasi "No"'. Dan hasil prediksi "No" maka ternyata klasifikasi "Yes" yaitu 1, dan hasil prediksi "No" sesuai dengan klasifikasi "No" yaitu 8. Perhitungan precision dan recall juga digunakan untuk mengukur kinerja klasifikasi. Precision yaitu hasil dari menjumlahkan data yang relevan kemudian dibandingkan dengan jumlah data yang ditemukan. Precision dihitung dengan cara membagi seluruh data benar yang rnilainya positif dibagi dengan seluruh data benar yang nilainya positif dan seluruh data salah yang nilainya positif. Dari angka data salah nilainya positif dapat diambil dari total nilai selain kolom true positive yang sesuai tiap kelasnya. Dalam penelitian ini diperoleh precision sebesar 95,24%. Sedangkan recall adalah hasil dari membandingkan total materi relevan yang ditemukan terhadap jumlah materi yang relevan. Cara menghitung recall yaitu dengan membagi data benar yan nilainya positif dengan total jumlah dari data benar yang nilainya positif dan data salah yang nilainys negatif. Hasil dari data salah yang nilainya negatif diperoleh dari jumlah total baris tiap kelas selain true positivenya. Jumlah persentasi recall yang diperoleh dalam penelitian ini sebesar 98,77%.

5.KESIMPULAN

Dari hasil perhitungan diatas,peneliti menyimpulkan bahwa penerapan data mining menggunakan Naive Bayes dapat menemukan pola klasifikasi untuk kanker paru-paru. Dan perhitungan menggunakan Naive Bayes menghasilkan akurasi data test dengan tingkat persentasi recall 98,77%, persentasi precision 95,24% dan tingkat persentasi akurasinya sebesar 94,62%. Dari 93 data, ada 80 data diprediksikan "Yes" sesuai dengan klasifikasi "Yes" dan 4 data yang diprediksikan "Yes" tapi ternyata klasifikasinya "No". Dan sebanyak 1 data diprediksi "No" ternyata termasuk klasifikasi "Yes" dan 8 data diprediksi "No" sesuai dengan perhitungan klasifikasinya yaitu "No" dengan tingkat persentasi recall 98,77% dan tingkat persentasi precision 95,24%.

REFERENCES

1) Bustami, B. (2013). PENERAPAN ALGORITMA NAIVE BAYES UNTUK MENGKLASIFIKASI DATA NASABAH ASURANSI. TECHSI - Jurnal Teknik Informatika, 5(2). https://doi.org/10.29103/TECHSI.V5I2.154

2) Damanik, A. R., Sumijan, S., & Nurcahyo, G. W. (2021). Prediksi Tingkat Kepuasan dalam Pembelajaran Daring Menggunakan Algoritma Naïve Bayes. Jurnal Sistim Informasi dan Teknologi, 88–94. https://doi.org/10.37034/JSISFOTEK.V3I3.49

3) Fadlan, C., Ningsih, S., Perdana Windarto STMIK-MUSIRAWAS Lubuklinggau, A., Fadlan, C., Perdana Windarto, A., Studi Sistem Informasi-STIKOM Tunas Bangsa Pematangsiantar Jalan, P., & Sudirman Blok, J. A. (2018). PENERAPAN METODE NAÏVE BAYES DALAM KLASIFIKASI KELAYAKAN KELUARGA PENERIMA BERAS RASTRA. JUTIM, 3(1).

4) Firdaus, D. (2017). Penggunaan Data Mining dalam Kegiatan Sistem Pembelajaran Berbantuan Komputer. Jurnal, 6, 2089–5615.

5) Jefi, J., Hendri, H., Afni, N., Salim, A., & Maulana, Y.

E-ISSN 2797-4278
P-ISSN 2797-4294

I. (2021). Penerapan Algoritma Naive Bayes Dalam Memprediksi Penyakit Lambung. Journal of Information System, Informatics and Computing, 5(2), 524–531. https://doi.org/10.52362/JISICOM.V5I2.659

6) Liantoni, F., & Nugroho, H. (2015). KLASIFIKASI DAUN HERBAL MENGGUNAKAN METODE NAÏVE BAYES CLASSIFIER DAN KNEAREST NEIGHBOR. Jurnal Simantec, 5(1). https://doi.org/10.21107/SIMANTEC.V5I1.1009

7) Novianti, D., Nusa, S., Jakarta, M., & Sitasi, C. (2019). Implementasi Algoritma Naïve Bayes Pada Data Set Hepatitis Menggunakan Rapid Miner. 21(1). https://doi.org/10.31294/p.v20i2

8) Purnama, I., Saputra, R., & Wibowo, A. (n.d.). IMPLEMENTASI DATA MINING MENGGUNAKAN CRISP-DM PADA SISTEM INFORMASI EKSEKUTIF DINAS KELAUTAN DAN PERIKANAN PROVINSI JAWA TENGAH.

9) Sari, R. W., Wanto, A., & Windarto, A. P. (2018). IMPLEMENTASI RAPIDMINER DENGAN METODE K-MEANS (STUDY KASUS: IMUNISASI CAMPAK PADA BALITA BERDASARKAN PROVINSI). KOMIK (Konferensi Nasional Teknologi Informasi dan Komputer), 2(1). https://doi.org/10.30865/KOMIK.V2I1.930

10) Utomo, D. P., & Purba, B. (2019). Penerapan Datamining pada Data Gempa Bumi Terhadap Potensi Tsunami di Indonesia. Prosiding Seminar Nasional Riset Information Science (SENARIS), 1(0), 846–853. https://doi.org/10.30645/SENARIS.V1I0.91