Jurnal Smart Teknologi
Vol. 3, No. 4, Mei 2022, Halaman 350 – 355
ISSN: 2774-1702, http://jurnal.unmuhjember.ac.id/index.php/JST

Pengklasifikasian Kanker Payudara Dan Kanker Paru-Paru Dengan Metode Gaussian Naïve Bayes, Multinomial Naïve Bayes, Dan Bernoulli Naïve Bayes

Classification Of Breast Cancer And Lung Cancer Using The Gaussian Naïve Bayes Multinomial Nave Bayes And Bernoulli Naïve Bayes Methods

Hedva Kenang Candra Alivian Pratama[1)], Wiwik Suharso[2)*], Qurrota A'yun[3)]
[1]Mahasiswa Fakultas Teknik, Universitas Muhammadiyah Jember
Email: hkenang@gmail.com
[2]Dosen Fakultas Teknik, Universitas Muhammadiyah Jember *Koresponden Author
Email: wiwiksuharso@unmuhjember.ac.id
[3]Dosen Fakultas Teknik, Universitas Muhammadiyah Jember
Email: qurrotaayun@unmuhjember.ac.id

Abstrak

Penyakit kanker merupakan salah satu penyebab utama kematian pada seluruh dunia. Kanker payudara dan kanker paru-paru merupakan jenis kanker yang sering muncul pada kasus baru yang dimana menyebabkan kematian terbesar (setelah dikontrol dengan umur) yaitu dengan jumlah persentase sekitar 43,3% untuk kanker payudara dan 23,1% untuk kanker paru-paru. Oleh karena itu dilakukannya penelitian untuk mengetahui kinerja dari *Gaussian Naïve Bayes, Multinomial Naïve Bayes* dan *Bernoulli Naïve Bayes* dengan menggunakan pemrograman berbahasa *phyton* dengan *tools* pemrograman *google colab.*Data yang digunakan pada penelitian ini sebanyak 699 data pada *Breast Cancer* dan 309 data pada *Lung Cancer*. Hasil dari penelitian ini bahwa performara rata-rata metode *Bernoulli Naïve Bayes* lebih unggul dengan hasil rata-rata *accuracy* 93,25%, rata-rata *precesion* 94,23%, dan rata-rata recall 94,69%

Kata Kunci : *Bernoulli Naïve Bayes, Gaussian Naïve Bayes,* Kanker, Klasifikasi, *Multinomial Naïve Bayes*.

Abstract

Cancer is one of the leading causes of death worldwide. Breast cancer and lung cancer are types of cancer that often appear in new cases which cause the largest death (after controlling for age) with a total percentage of about 43.3% for breast cancer and 23.1% for lung cancer. Therefore, a study was conducted to determine the performance of Gaussian Naïve Bayes, Multinomial Nave Bayes and Bernoulli Naïve Bayes using Python language programming with Google Colab programming tools. The data used in this study were 699 data on Breast Cancer and 309 data on Lung Cancer. The results of this study show that the average performance of the Bernoulli Naïve Bayes method is superior with an average accuracy of 93.25%, an average precesion of 94.23%, and an average recall of 94.69%..

Keywords : *Bernoulli Nave Bayes, Gaussian Nave Bayes, Cancer, Classification, Multinomial Nave Bayes.*

1. PENDAHULUAN

Kanker merupakan suatu pertumbuhan sel yang terjadi secara tidak normal dari sel-sel jaringan tubuh yang berubah menjadi ganas. Sel-sel tersebut dapat tumbuh lebih lanjut serta menyebar ke bagian tubuh lainnya yang dapat menyebabkan kematian. Sel tubuh yang mengalami perubahan mutasi dan mulai tumbuh dapat membelah lebih cepat dan tidak terkendali seperti sel normal. Sel kanker tidak mati setelah

usianya melainkan dapat tumbuh terus dan bersifat *invasive* (menyerang) yang menyebabkan sel normal yang tumbuh terdesak atau malah mati [1].

Penyakit kanker adalah salah satu penyakit yang menyebabkan kematian utama di seluruh dunia. Pada tahun 2012, sekitar 8,2 juta kematian dikarenakan oleh kanker. Disini penulis mengambil Kanker paru-paru dan payudara ke dua penyakit tersebut adalah penyebab kematian terbesar tiap tahunnya [1].

Berdasarkan Data GLOBOCAN, International Agency for Research on Cancer (IARC), diberitahukan pada tahun 2012 terdapat kasus baru yang menyebabkan 14.067.894 korban dan korban kematian kanker sekitar 8.201.575 di seluruh dunia. Kanker payudara dan kanker paru-paru merupakan jenis kanker yang sering muncul pada kasus baru yang dimana menyebabkan kematian terbesar (setelah dikontrol dengan umur) yaitu dengan jumlah persentase sekitar 43,3% untuk kanker payudara dan 23,1% untuk kanker paru-paru [1].

Naïve Bayes adalah metode klasifikasi probabilistik yang dimana digunkan untuk menghitung nilai kemunculan probabilitas dengan menambahkan frekuensi dan beberapa kombinasi nilai dari data yang ada. Algoritma yang digunakan adalah teorema *Bayes* yang dimana semua atribut independen atau tidaknya saling terhubung yang dimana diberikannya nilai pada suatu variabel kelas [2].

Penelitian terkait yang sudah pernah dilakukan diantaranya oleh [3] dengan judul "Pengklasifikasian Breast Cancer Dengan Metode Naïve Bayes" yang bertujuan untuk mendeteksi resiko terkena kanker payudara dengan data Wisconsin Breast Cancer dengan mendapatkan nilai accuracy sebesar 96% dengan data yang digunakan sebanyak 699 data, penelitian lain yang dilakukan oleh [4] dengan judul "Multinomial Naïve Bayes Untuk Klasifikasi Status Kredit Mitra Binaan Di Pt. Angkasa Pura I Program Kemitraan" yang bertujuan untuk mendetiksi kredit yang bermasalah dengan hasil yang didapatkan dengan accuracy 86%, precission 73% dan recal 73% dengan data yang digunakan sebanyakan 148 data. Penelitian lain yang dilakukan [6] dengan judul "Klasifikasi Dokumen Berkategori Menggunakan Algotirma Naive Bayes Berbasis Bernoulli" yang bertujuan untuk mengklasifikasi kategori dokumen dengan menggunakan algoritma Naive Bayes Berbasis Bernoulli. Klasifikasi ini ditekankan pada kategori dokumen diantaranya Ekonomi, Kesehatan, Hiburan dan Teknologi, hasil yang di dapatkan presicion sebesar 70%, accuracy 65% dan recall 70% dengan objek penelitian sebanayak 60 dokomen.

2. TINJAUAN PUSTAKA

A. Naïve Bayes Classifier

Menurut [2] *Naïve Bayes* adalah metode klasifikasi probabilistik yang dimana digunkan untuk menghitung nilai kemunculan probabilitas dengan menambahkan frekuensi dan beberapa kombinasi nilai dari data yang ada. Algoritma yang digunakan adalah teorema *Bayes* yang dimana semua atribut independen atau tidaknya saling terhubung yang dimana diberikannya nilai pada suatu variabel kelas. *Naïve Bayes* merupakan cara pengklasifikasian yang dimana menggunakan metode probabilitas dan statistik yang ditemukan oleh ilmuwan Inggris Thomas Bayes, yang dimana bertujuan untuk memprediksi peluang di masa yang akan datang berdasarkan data dari pengalaman dimasa sebelumnya [2]

Persamaan dari algoritma Bayes adalah [2]:

$$P(H \mid X) = (P(X \mid H) \times P(H))/P(X) \quad (1)$$

Dimana :

X	: Data dengan class yang belum diketahui
H	: Hipotesis data merupakan suatu class spesifik
P(H|X)	: Probabilitas hipotesis H berdasar kondisi X (posteriori probabilitas)
P(H)	: Probabilitas hipotesis H (prior probabilitas)
P(X|H)	: Probabilitas X berdasarkan kondisi pada hipotesis H
P(X)	: Probabilitas X

Jurnal Smart Teknologi
Vol. 3, No. 4, Mei 2022, Halaman 350 – 355
ISSN: 2774-1702, http://jurnal.unmuhjember.ac.id/index.php/JST

a. Naïve Bayes Gaussian

Untuk fitur bertipe numerik (kontinu), distribusi *Gauss* biasanya dipilih untuk merepresentasikan probabilitas bersyarat dari fitur kontinu pada sebuah kelas, P(Xi|C) sedangkan distribusi *Gauss* dikarakteristikkan dengan dua parameter: *mean*, μ, dan variansi, σ^2. Untuk setiap kelas cj, probabilitas bersyarat kelas yj [7].

untuk fitur Fi adalah Persamaan dari teorima Naïve Bayes Gaussian [2]:

$$P(Xi = xi|Y = yj) = \frac{1}{\sqrt{2\pi\sigma ij}} e^{-\frac{(Xi-\mu ij)^2}{2\sigma 2ij}} \quad (2)$$

Dimana :
P : Peluang
Xi : Atribut ke i
xi : Nilai atribut ke i
Y : Kelas yang dicari
μ : Mean, menyatakan rata-rata dari seluruh atribut.
σ : Standar deviasi, menyatakan varian dari seluruh atribut

b. Naïve Bayes Bernoulli

Algoritma *Bernoulli Naive Bayes* mengimplementasikan klasifikasi untuk data yang didistribusikan sesuai dengan distribusi *Bernoulli multivariat*; yaitu, mungkin terdapat beberapa fitur tetapi masing-masing dianggap sebagai variabel bernilai biner (*Bernoulli, boolean*). Oleh karena itu, kelas ini membutuhkan sampel untuk direpresentasikan sebagai vektor fitur bernilai biner [8].

Aturan keputusan untuk algoritma Bernoulli Naive Bayes diberikan pada persamaan berikut [8]:

$$P(xi \mid y) = P(i \mid y)\, x\_i + (1 - P(i \mid y))(1 - x\_i) \quad (3)$$

c. Naïve Bayes Multinomial

Multinomial Naive Bayes merupakan suatu kondisi probabilitas yang dilakukan tanpa memperhitungkan urutan pada kata dan informasi yang telah ada pada dokumen atau kalimat pada umumnya. Dalam algoritma tersebut juga menghitung jumlah kata yang muncul pada dokumen [9]

Menurut [10], model *Multinomial Naïve Bayes* menggunakan rumus sebagai berikut

$$P(c \mid \text{term dok d}) = P(c) \times P(t1 \mid c) \times P(t2 \mid c) \times P(tn \mid c) \quad (4)$$

Dimana :
P(c|term dok d) : Probabilitas suatu dokumen dalam kelas c
P(c) : Probabilitas prior dari kelas c
P(tn|c) : Probabilitas kata ke-n pada kelas c
tn : kata ke n pada dokumen

3. METODE PENELITIAN

Pada penelitian ini akan dirancang tahapan-tahapan yang meliputi Pengumpulan Dataset, *Preprocessing*, Pengklasifikasian..

1. Pengumpulan Dataset: Dataset yang dikumpulkan berupa dataset publik yang tersedia pada [11] dan [12]
2. *Preprocessing* Data: Mengolah data seperti pembersihan data, pemilihan data, dan perubahan data.
3. Pengklasifikasian: Proses pelatihan data yang dimana nanti akan di ujikan kepada data uji yang telah tersedia

A. PreProcessing Data

Sebelum data digunakan ke dalam sistem, data perlu dilakukan pengolahan dengan cara sebagai berikut:

a. Pembersihan Data

Pembersihan data dilakukan untuk menghilangkan beberapa missing value yang terdapat pada data yang kita kumpulkan. Pada 2 data yang memiliki missing value terdapat pada *Breast Cancer* yang berjumlah 16 data, sehingga data yang semulanya 699 hanya tersisa 683 data..

b. Pemilihan Data

Pada tahap ini data yang akan di ambil, merupakan data yang relevan untuk dijadikan fitur. Pada kasus Breast Cancer atribut yang dihilangkan adalah atribut *Sample Code Number*, sedangkan pada Lung Cancer atribut

Jurnal Smart Teknologi
Vol. 3, No. 4, Mei 2022, Halaman 350 – 355
ISSN: 2774-1702, http://jurnal.unmuhjember.ac.id/index.php/JST

yang dihilangkan adalah *GENDER, FATIGUE, WHEEZING, COUGHING,* dan *CHEST PAIN*.

c. Perubahan Data

Pada tahap ini dilakukan perubahan data terhadapap data *Breast Cancer* yaitu di mana merubah nilai dari setiap atribute yang berupa angka akan di golongkan menjadi beberapa text.

B. Pengklasifikasian

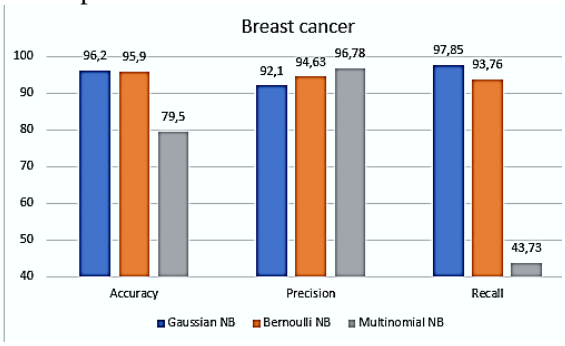
Pada proses ini ada beberapa tahapan seperti berikut :

1. Pada klasifikasi jenis data pada umumnya di pisahkan menjadi data training dan data uji.
2. Clean dataset yang sudah siap, akan di pisah menjadi data training dan data testing.Pada bagian ini dataset akan dibagi dengan menggunakan teknik k-fold cross validation, karena tidak ada aturan formal dalam pemilihan nilai pada K-Fold Cross Validation [13], maka pemilihan nilai fold K tersebut diambil nilai yang habis dibagi atau tidak menyisahkan nilai, sehingga pada setiap partisi akan memiliki nilai yang seimbang. Dimana rasio k-fold yang digunakan adalah k-fold 5 dengan rasionya 80:20, untuk data breast cancer terdapat 546 data training dan 137 untuk data uji, sedangkan data lung cancer terdapat 247 data training dan 62 data uji. Pada k-fold 10 rasionya 90 : 10 untuk data breast cancer terdapat 615 data training dan 68 untuk data uji, sedangkan data lung cancer terdapat 278 data training dan 31 data uji. Setelah pembagian data maka data testing akan di proses kedalam 3 metode, yaitu Naïve Bayes Gaussian, Bernoulli, dan Multinomial
3. Setelah dilakukan data *training* ke 3 metode yang digunakan, selanjutnya dilakukan proses klasifikasi dengan data *testing*.
4. Tahapan selanjutnya setelah proses klasifikasi adalah proses evaluasi menggunakan *confusion matrix*

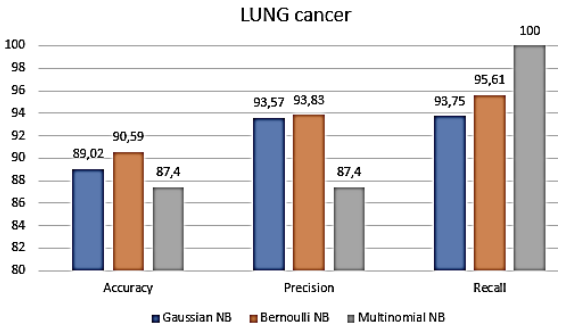
4. HASIL DAN PEMBAHASAN

A. Hasil Klasifikasi

Berdasarkan hasil uji coba dari 2 dataset yang telah dilakukan proses klasifikasi dapat dilihat pada **Gambar 1 dan 2**.



Gambar 1. Persentase hasil dataset *Breast Cancer*
Sumber: Hasil Penelitian



Gambar 2. Persentase hasil dataset *Lung Cancer*
Sumber: Hasil Penelitian

Berdasarkan Gambar1 hasil terbaik pada dataset *Breast Cancer* dalam hal *accuracy* dan *recall* metode *Gaussian NB* mendapatkan hasil tertinggi dengan nilai 96,2% dan 97,85 , untuk hasil *precision* metode *Multinomial NB* mendapatkan hasil tertinggi dengan nilai 96,78%. Untuk Gambar 2 hasil terbaik pada dataset *Lung Cancer* dalam hal *accuracy* dan *precision* metode *Gaussian NB* mendapatkan hasil tertinggi dengan nilai 90,95% dan 93,81%, dan untuk hasil *recall* metode *Multinomial NB* mendapatkan hasil tertinggi dengan nilai 100%.

Tabel 1. Tabel Tabulasi

Dataset	*Accuracy(%)*			*Precision(%)*			*Recall(%)*		
	GNB	BNB	MNB	GNB	BNB	MNB	GNB	BNB	MNB
Breast Cancer	96,2	95,9	79,5	92,1	94,63	96,78	97,85	93,76	43,73
Lung Cancer	89,02	90,59	87,4	93,57	93,83	87,4	93,75	95,61	100
Rata-rata	92,61	93,25	83,45	92,84	94,23	92,09	95,8	94,69	71,87

Sumber: Hasil Perhitungan

Dari tabel 1. diatas dapat dilihat kinerja hasil dari setiap metode sebagai berikut:

1. *Accuracy*: Pada dataset *Breast Cancer* metode *Gaussian NB* lebih unggul dengan nilai 96,2% dan pada dataset *Lung Cancer* metode *Gaussian NB* lebih unggul dengan nilai 90,59%. Untuk performa rata-rata dalam kedua dataset *Gaussian NB* memiliki nilai rata-rata lebih unggul dengan hasil 93,58%.
2. *Precision*: Pada dataset *Breast Cancer* metode *Multinomial NB* lebih unggul dengan nilai 96,2% dan pada dataset *Lung Cancer* metode *Bernoulli NB* lebih unggul dengan nilai 93,83%. Untuk performa rata-rata dalam kedua dataset *Bernoulli NB* memiliki nilai rata-rata lebih unggul dengan hasil 94,23%.
3. *Recall*: Pada dataset *Breast Cancer* metode *Gaussian NB* lebih unggul dengan nilai 96,2% dan pada dataset *Lung Cancer* metode *Multinomial NB* lebih unggul dengan nilai 100%. Untuk performa rata-rata dalam kedua dataset *Gaussian NB* memiliki nilai rata-rata lebih unggul dengan hasil 95,8%.

5. KESIMPULAN

Berdasarkan dari hasil penelitian yang sudah dilakukan maka dapat disimpulkan :

1. Dari hasil yang didapatkan dalam pembuatan model prediksi dengan 2 dataset menggunakan metode *Gaussian Naïve Bayes, Bernoulii Naïve Bayes, dan Multinomial Naïve Bayes*. Didapatkan bahwa metode *Bernoulli Naïve Bayes* memiliki performa rata-rata lebih unggul dengan hasil rata-rata *accuracy* 93,25%, rata-rata *precesion* 94,23%, dan rata-rata *recall* 94,69%
2. Pada hasil performa metode *Gaussian Naïve Bayes* memiliki hasil yang hampir sama atau tidak memiliki selisih hasil yang terlalu jauh dengan metode *Bernoulli Naïve Bayes*

6. SARAN

Saran pengembangan dari sistem deteksi *Breast Cancer* dan *Lung Cancer* untuk penelitian kedepannya adalah:

1. Gunakan metode lain untuk melakukan klasifikasi pada dataset *breast cancer* dan *lung cancer,* seperti *k-means* dan *fuzzy c-means*
2. Gunakan metode *balance dataset*, seperti *oversampling, under sampling, class weight* atau *threshold*
3. Gunakan teknik train, test, validasi

DAFTAR PUSTAKA

[1] Kementerian Kesehatan Republik Indonesia. 2019. *INFODATIN (Pusat Data dan Informasi Kementerian Kesehatan RI)*. Kementerian Kesehatan Republik Indonesia. Jakarta Selatan, Indonesia.

[2] Alfa Saleh. 2014. *Klasifikasi Metode Naïve Bayes dalam Data Mining untuk Menetukan Konsentrasi Siswa (Studi Kasus di MAS PAB 2 Medan)*. KeTIK (Konferensi Nasional Pengembangan Teknologi Informasi dan Komunikasi).

[3] Hasbi, W. 2018. Pengklasifikasian Breast Cancer dengan Metode Naïve Bayes. Tesis. Fakultas Teknik Universitas Hasanudin, Makasar.

[4] Bunga, M.T.H., Djahi, B.S., dan Nabuasa Y.Y. 2018. Multinomial Naïve Bayes Untuk Klasifikasi Status Kredit Mitra Binaan Di PT. Angkasa Pura I Program Kemitraan. *J-ICON*. 6(2): 30-34

[5] Susanti, M.A. 2016. Klasifikasi Dokumen Berkatergori menggunakan Algoritma Naïve Bayes berbasi Bernoulli. Tesis. Fakultas Teknik, Jurusan Teknik Informatika, Universitas Muhammadiyah Jember.
[6] Prasetyo, E. (2012). *Data Mining Konsep dan Aplikasi menggunakan MATLAB* Edisi 1. Yogyakarta: ANDI
[7] Saraswati, M. dan Riminarsih, D. 2020. Analisis Sentimen terhadap Pelayanan KRL Cummuterline berdasarkan Data Twitter menggunakan Algoritma Bernoulli Naïve Bayes. *Jurnal Ilmiah Informatika Komputer*. 25(1): 225-238.
[8] Shofiya Feni. (2020). *Perbandingan Algoritma Support Vector Machine (SVM) Dan Multinomial Naive Bayes (MNB) Dalam Klasifikasi Abstrak Tugas Akhir (Studi Kasus: Fakultas Teknik Universitas Muhammadiyah Jember)*. Fakultas Teknik, Jurusan Teknik Informatika, Universitas Muhammadiyah Jember
[9] Hamdan Ashari. (2020). *Perbandingan Kinerja Algoritma Multinomial Naïve Bayes (MNB), Multivariate Bernoulli Dan Rocchio Algorithm Dalam Klasifikasi Konten Berita Hoax Berbahasa Indonesia Pada Media Sosial*. Fakultas Teknik, Jurusan Teknik Informatika, Universitas Muhammadiyah Jember.
[10] Wolberg, W.H, Street, W.N., dan Mangasarian, O.L. 1995. Breast Cancer Wisconsin (Diagnostic) https://data.world/uci/breast-cancer-wisconsin-diagnostic. Diakses pada tanggal 14 Mei 2011
[11] Staceyinrobert. 2017. Survey Lung Cancer https://data.world/sta427ceyin/survey-lung-cancer. Diakses pada tanggal 15 Mei 2011
[12] Kuhn, M., dan Johnson, K. 2013. Applied Predictive Modeling. New York: Springer.