Jurnal Teknik Komputer,
Vol. III, No. 2, Agustus 2017

Penerapan Algoritma Naïve Bayes Untuk Klasifikasi Bakteri Gram-Negatif

Evy Priyanti
Program Studi Komputerisasi Akuntansi
Akademik Manajemen Informatika dan Komputer Bina Sarana Informatika
Jl. Rs Fatmawati 24 Jakarta Selatan
http://www.bsi.ac.id
evy.evp@bsi.ac.id

Abstract *- The classification depends on the variety of bacteria that exist. The important feature to identify an organism of bacterial phenotype is the scheme that utilizes the morphology and staining properties of the bacteria itself, to classify the phenotype scheme is used Naïve Bayes algorithm that has proven to have a high degree of accuracy and high rate of speed when applied into E.coli dataset in E. coli dataset consisting of seven features are: mcg, gvh, lips, chg, aac, alm1, alm2, and proteins are classified into 8 classes: cytoplasmic (cp), an inner membrane without the signal sequence (im), perisplasm (pp), in the membrane with uncleavable signal sequence (IMU), outer membrane (oM), outer membrane lipoprotein (OML), the membrane lipoprotein (IML), an inner membrane with cleavable signal sequence (IMS) with an accuracy of 80.93%, with Naïve Bayes algorithm so it can be ascertained that the classification of gram-negative bacteria with E. coli phenotype datasets prove to be accurate.*

Keyword : Bacteria Gram-Negative, Naïve Bayes, Ecoli

Abstrak - Klasifikasi bakteri tergantung dari varietas yang ada. Fitur penting untuk mengidentifikasi suatu organisme dari bakteri adalah dengan skema fenotipe yang memanfaatkan morfologi dan pewarnaan sifat dari bakteri itu sendiri, untuk mengklasifikasikan skema fenotipe tersebut digunakanlah algoritma Naïve Bayes yang sudah terbukti memiliki tingkat akurasi dan tingkat kecepatan yang tinggi saat diaplikasikan kedalam dataset E.coli Dalam dataset E.coli yang terdiri dari tujuh fitur yaitu : mcg, gvh, bibir, chg, aac, alm1, alm2. Dan protein diklasifikasikan ke dalam 8 kelas: sitoplasma (cp), membran dalam tanpa urutan sinyal (im), perisplasm (pp), dalam membran dengan uncleavable sinyal urutan (IMU), luar membran (om), luar membran lipoprotein (OML), dalam membran lipoprotein (IML), membran dalam dengan cleavable urutan sinyal (IMS) dengan tingkat akurasi sebesar 80.93%, dengan demikian algoritma Naïve Bayes sudah dapat dipastikan bahwa klasifikasi bakteri gram-negatif dengan fenotipe dataset E.coli terbukti akurat.

Kata Kunci : Bakteri Gram-Negatif, Naïve Bayes, Ecoli

I. PENDAHULUAN

Bakteri Gram-negatif merupakan bakteri yang ada disekitar kita, sebagian merugikan dan sebagian bermafaat bagi kehidupan manusia.

Berikut penjelasan bakteri Gram-Negatif menurut Chatterjee dan Chaudhuri, Selama proses pewarnaan Gram, bakteri Gram negatif tidak mempertahankan zat warna kristal violet. Bakteri gram negatif berwarna merah atau merah muda di bawah mikroskop, dengan penambahan counter-noda.

Mereka memiliki membran sitoplasma dan membran luar yang mengandung lipopolisakarida. Selain itu, ada lapisan-S yang melekat pada membran luar. Lipopolisakarida yang di membran luar bakteri Gram negatif merupakan endotoksin, yang memicu respons dari sistem kekebalan tubuh bawaan. Peradangan adalah gejala umum dari infeksi dan dapat menyebabkan keracunan.

Bakteri gram negatif menyebabkan infeksi, seperti kolera, tipus, meningitis dan berbagai macam kesesakan gastrointestinal.Infeksi sekunder di rumah sakit biasanya akibat dari infeksi oleh bakteri Gram negatif. Berikut adalah beberapa contoh bakteri Gram negatif, termasuk penjelasan singkat dari struktur, fungsi dan signifikansi medis.

a. Salmonella

Salmonella adalah genus bakteri berbentuk batang.Mereka tidak membentuk spora enterobacteria dengan flagela. Mereka mengoksidasi dan mengurangi zat organik dan, dalam proses, menghasilkan hidrogen sulfida. Salmonella ditemukan di seluruh kerajaan hewan. Mereka biasanya hidup dalam saluran usus burung dan hewan lain dan dapat menyebar dari hewan ke manusia melalui konsumsi susu, telur, unggas dan daging sapi yang tercemar. Mual, muntah, diare dan demam adalah gejala umum pada manusia yang terinfeksi Salmonella.

 p-ISSN: 2442-2436, e-ISSN: 2550-0120

b. Shigella

Shigella adalah genus bakteri Gram negatif berbentuk batang.Mirip dengan Salmonella, mereka tidak membentuk spora.Mereka hanya mempengaruhi primata dan tidak pada mamalia lainnya.Shigella adalah penyebab Shigellosis pada manusia.Mereka juga menyebabkan diare dan disentri dan dapat menyebar dari orang ke orang melalui kontak dan melalui konsumsi makanan dan air yang terkontaminasi.Mereka menghancurkan sel-sel yang melapisi usus besar, mengakibatkan ulserasi dan diare berdarah.

c. Escherichia Coli

E. Coli, seperti yang umum dikenal, adalah bakteri Gram negatif berbentuk batang. E.coli yang non-bersporulasi.Mereka dapat tumbuh aerobik atau anaerobik dan menyebabkan pengurangan substrat, seperti oksigen dan nitrat. Meskipun sebagian besar strain E. Coli yang tidak berbahaya dan yang hadir dalam usus manusia beberapa jam setelah melahirkan, strain E. Coli tertentu dapat menghasilkan racun yang mematikan dan bisa berbahaya. Mereka bisa menyebabkan infeksi saluran kemih, meningitis neonatal, keracunan makanan dan komplikasi serius, seperti sindrom hemolitik uremik, pada manusia. Konsumsi sayuran yang tidak dicuci dengan benar dan daging yang belum dimasak benar-benar dapat mengakibatkan infeksi E. Coli. Infeksi E. Coli juga diketahui terjadi dari makan hazelnut.

d. Bakteri Asam asetat

Bakteri asam asetat adalah bakteri Gram negatif berbentuk batang. Bakteri ini dinamakan demikian karena mereka mengoksidasi etanol menjadi asam asetat selama proses fermentasi, dari mana mereka memperoleh energi mereka. Mereka hadir di alam, di bunga dan buah-buahan, dan merupakan bagian penting dari industri makanan.Fermentasi anggur juga memanfaatkan bakteri asam asetat, yang umumnya tidak berbahaya bagi manusia.

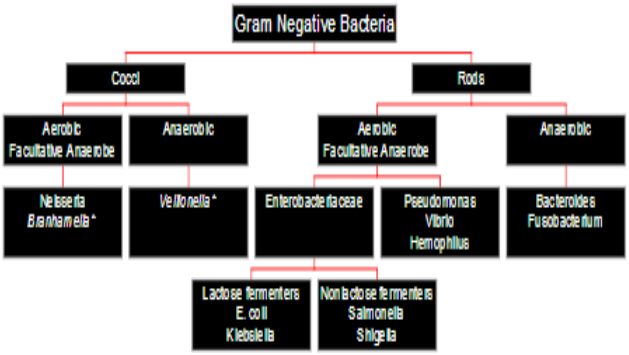
e. Legionella

Bakteri Legionella adalah berbentuk batang. Komposisi kimia dari dinding sel bakteri sisi-rantai ‘, serta gula yang berbeda, bertanggung jawab untuk mengklasifikasikan jenis Legionella.Legionella yang paling sering diketahui menyebabkan Legionellosis, atau penyakit Legionairre ini, dan demam Pontiac. Legionella biasanya ditemukan di sumber air masyarakat, seperti kolam renang dan air mancur, kolam dan sungai.

f. Cyanobacteria

Juga dikenal sebagai ganggang biru-hijau, cyanobacteria datang dalam segala bentuk, dari batang dan kokus sampai spiriila.Cyanobacteria bertanggung jawab untuk mengubah atmosfer dari atmosfer yang kurang oksigen menjadi atmosfer yang kaya oksigen.Sebagian besar energi mereka berasal dari fotosintesis.

Cyanobacteria yang umum ditemukan di sistem air tawar, lingkungan laut dan sumber terestrial.Mereka juga dapat ditemukan di lingkungan yang ekstrim, seperti mata air panas.Beberapa spesies Cyanobacteria menghasilkan cyanotoxin, yang bisa berbahaya bagi manusia dan spesies lainnya.Pada manusia, Cyanobacteria dapat menyebabkan keracunan, dan bukti terbaru menunjukkan bahwa mereka juga dapat menyebabkan Amyotrophic Lateral Sclerosis (ALS).



Sumber : Chatterjee (2012)

Gambar 1.1 Spesies Bakteri Gram-Negatif

Bakteri gram negatif penting bagi ekosistem. Mereka adalah bagian dari banyak hewan dan manusia, dan cyanobacteria bertanggung jawab untuk mengubah atmosfer bumi. Banyak bakteri Gram negatif juga digunakan untuk terapi medis dan pengobatan infeksi.

Teknik klasifikasi bakteri Gram-negatif dengan cara lokalisasi protein menjadi bagian-bagian berdasarkan urutan asam amino mereka. lokalisasi protein dengan skema fenotipe berperan dalam penentuan obat yang akan diberikan pada pasien nantinya. Lokalisasi protein digunakan untuk memeriksa metode yang cocok untuk sebuah penelitian dengan algoritma yang sesuai.

Selanjutnya akan dievaluasi dengan menganalisa tingkat akurasi pada dataset. Klasifikasi ini menggunakan dataset ecoli dari uci dataset yang terdiri dari 8 kelas.

Teknik klasifikasi adalah salah satu dari teknik data mining yang termasuk supervised learning. Supervised learning artinya proses pembentukan sebuah korespondensi (fungsi) menggunakan sebuah training dataset, dilihat sebagai sebuah "pengalaman masa lalu" dari sebuah model. Tujuannya adalah untuk memprediksi dari sebuah nilai (*output*) dari sebuah fungsi untuk setiap objek baru (*input)* setelah menyelesaikan proses training (Borunescu, 2011).

Klasifikasi merupakan suatu pekerjaan menilai objek data untuk memasukkannya kedalam kelas tertentu dari sejumlah kelas yang tersedia. Dalam klasifikasi ada dua pekerjaan utama yang dilakukan yaitu pembangunan model sebagai prototype untuk disimpan sebagai memori dan penggunaan model tersebut untuk melakukan pengenalan/klasifikasi/ prediksi pada unsur objek data lain agar diketahui di kelas mana objek data tersebut dalam model yang sudah disimpannya (prasetyo,2012)

Naïve bayes merupakan algoritma yang sesuai untuk klasifikasi pada data ecoli.

Berikut beberapa teori pendukung didalam penelitian ini :

a. Knowledge Discovery in Database (KDD)

Knowledge Discovery in Database (KDD) adalah keseluruhan proses non-trivial untuk mencari dan mengidentifikasi pola (pattern) dalam data, dimana pola yang ditemukanbersifat sah, baru, dapat bermanfaat dan dapat dimengerti (Maimoon & Rokach, 2010).

Tahapan Proses KDD

1. Data Selection

Menciptakan himpunan data target, pemilihan himpunan data, atau memfokuskan pada subset variabel atau sampel data, dimana penemuan (*discovery*) akan dilakukan. Pemilihan data dari sekumpulan data operasional perlu dilakukan sebelum tahap penggalian informasi dalam KDD dimulai. Data hasil seleksi yang akan digunakan untuk proses data mining, disimpan dalam suatu berkas, terpisah dari basis data operasional.

2. Pre-processing/ Cleaning

Pemprosesan pendahuluan dan pembersihan data merupakan operasi dasar seperti penghapusan noise dilakukan. Sebelum proses data mining dapat dilaksanakan, perlu dilakukan proses cleaning pada data yang menjadi fokus KDD. Proses cleaning mencakup antara lain membuang duplikasi data, memeriksa data yang inkonsisten, dan memperbaiki kesalahan pada data, seperti kesalahan cetak (tipografi). Dilakukan proses *enrichment*, yaitu proses "memperkaya" data yang sudah ada dengan data atau informasi lain yang relevan dan diperlukan untuk KDD, seperti data atau informasi eksternal.

3. *Transformation*

Pencarian fitur-fitur yang berguna untuk mempresentasikan data bergantung kepada goal yang ingin dicapai. Merupakan proses transformasi pada data yang telah dipilih, sehingga data tersebut sesuai untuk proses data mining. Proses ini merupakan proses kreatif dan sangat tergantung pada jenis atau pola informasi yang akan dicari dalam basis data

4. Data mining

Pemilihan tugas data mining; pemilihan goal dari proses KDD misalnya klasifikasi, regresi, clustering, dll. Pemilihan algoritma data mining untuk pencarian (*searching*) Proses Data mining yaitu proses mencari pola atau informasi menarik dalam data terpilih dengan menggunakan teknik atau metode tertentu. Teknik, metode, atau algoritma dalam data mining sangat bervariasi. Pemilihan metode atau algoritma yang tepat sangat bergantung pada tujuan dan proses KDD secara keseluruhan.

5. *Interpretation/ Evaluation*

Penerjemahan pola-pola yang dihasilkan dari data mining. Pola informasi yang dihasilkan dari proses data mining perlu ditampilkan dalam bentuk yang mudah mimengerti oleh pihak yang berkepentingan. Tahap ini merupakan bagian dari proses KDD yang mencakup pemeriksaan apakah pola atau informasi yang ditemukan bertentangan dengan fakta atau hipotesa yang ada sebelumnya.

b. Data mining

Data mining adalah proses yang menggunakan teknik statistik, matematika, kecerdasan buatan, dan machine learninguntuk mengekstrasi dan mengidentifikasi informasi yang bermanfaat dan pengetahuan yang terkait dari berbagai database besar (Maimoon & Rocach, 2010) .

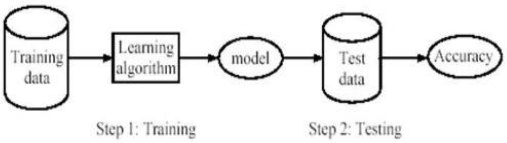
Pengelompokan Data Mining Menurut Larose, data mining dibagi menjadi beberapa kelompok berdasarkan tugas yang dapat di lakukan, yaitu (Larose, 2005):

a.Deskripsi, terkadang peneliti dan analisis secara sederhana ingin mencoba mencari cara untuk menggambarkan pola dan kecendrungan yang terdapat dalam data.

b.Estimasi hampir sama dengan klasifikasi, kecuali variabel target estimasi lebih ke arah numerik dari pada ke arah kategori.

c.Prediksi hampir sama dengan klasifikasi dan estimasi, kecuali bahwa dalam prediksi nilai dari hasil akan ada di masa mendatang.

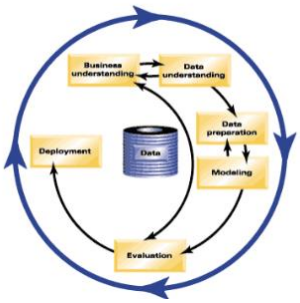
d.Klasifikasi, terdapat target variabel kategori.



Gambar 1.1 Tahapan Proses Klasifikasi

e.Pengklusteran merupakan suatu metode untuk mencari dan mengelompokkan data yang memiliki kemiripan karakteriktik (*similarity*) antara satu data dengan data yang lain. Clustering merupakan salah satu metode data mining yang bersifat tanpa arahan (*unsupervised*).

f.Asosiasi, tugas asosiasi dalam data mining adalah menemukan atribut yang muncul dalam suatu waktu. Dalam dunia bisnis lebih umum disebut analisis keranjang belanja.



Sumber : Larose, Daniel T, Data Mining Methods and Models

Gambar 1.2 Proses CRISP-DM

Metode Data Mining Menurut Larose, data mining memeliki enam fase CRISP-DM (Cross Industry Standard Process for Data Mining).

a.Fase Pemahaman Bisnis (Business Understanding Phase)

b.Fase Pemahaman Data (Data Understanding Phase)

c.Fase Pengolahan Data (Data Preparation Phase)

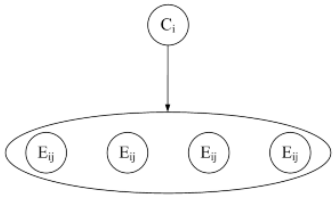
d.Fase Pemodelan (Modeling Phase)

p-ISSN: 2442-2436, e-ISSN: 2550-0120

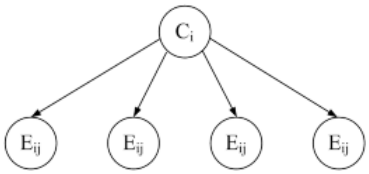
e.Fase Evaluasi (Evaluation Phase)
f.Fase Penyebaran (Deployment Phase)

c. Naïve bayes

Klasifikasi adalah salah satu tugas yang penting dalam data mining, dalam klasifikasi sebuah pengklasifikasi dibuat dari sekumpulan data latih dengan kelas yang telah di tentukan sebelumnya. Performa pengklasifikasi biasanya diukur dengan ketepatan (atau tingkat galat) (Prasetyo, 2012).
Teorema Bayes adalah teorema yang digunakan dalam statistika untuk menghitung peluang untuk suatu hipotesis, Bayes Optimal Classifier. Menghitung peluang dari suatu kelas dari masing-masing kelompok atribut yang ada, dan menentukan kelas mana yang paling optimal. Umumnya kelompok atribut *E* direpresentasikan dengan sekumpulan nilai atribut (*x1,x2,x3,….,xn*) dimana *xi* adalah nilai atribut *Xi*. *C* adalah variable klasifikasi dan *c* adalah nilai dari *C*. Pengklasifikasian adalah sebuah fungsi yang menugaskan data tertentu kedalam
sebuah kelas. Dari sudut pandang peluang, berdasarkan aturan Bayes kedalam kelas *c* adalah :



Sumber : Kusrini & Luthfi, 2009
Gambar 2.3 Teorema Bayes



Sumber : Kusrini & Luthfi, 2009
Gambar 1.4 Teorema Naive Bayes

Naive Bayes adalah pengklasifikasian statistik yang dapat digunakan untuk memprediksi probabilitas keanggotaan suatu class. Naive Bayes didasarkan pada teorema Bayes yang memiliki kemampuan klasifikasi serupa dengan decision treedan neural network. Naive Bayes terbukti memiliki akurasi dan kecepatan yang tinggi saat diaplikasikan kedalam database dengan data yang besar (Kusrini & Luthfi, 2009).

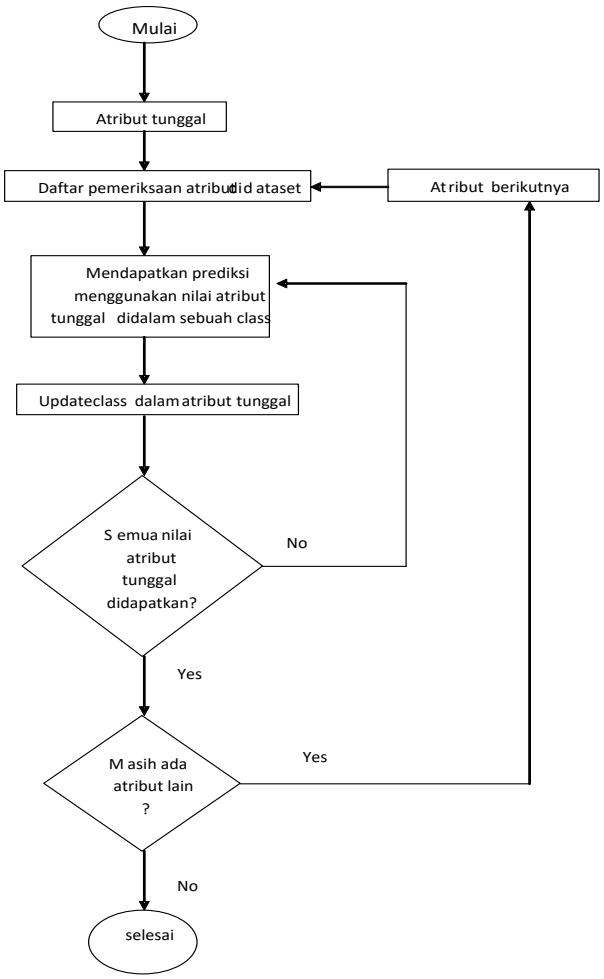
Naïve bayes classifier adalah dasar dari teorema dan ini dapat menjadikan ferforma yang lebih baik didalam menyelesaikan tugas. Kondisi probabilitas adalah P(xj|ci) dan prior probabilitas adalah P(cj).P(ci) adalah dikalkulasikan by counting the training sample dan kemudian dividing hasil penjumlahan hasil berdasarkan training set s size. Klasifikasi Naïve bayes didefinisikan dengan

Fi(X)-P(xj|ci)P(ci)

Klasifikasi naïve bayes berhubungan dengan teori probabilitas sederhana, yang merupakan cabang dari matematika probabilitas dapat digunakan untuk menentukan model dengan data yang tidak pasti dengan tujuan dan hasil yang menarik dengan menggabungkan pengetahuan dari hasil eksperimental dan bukti-bukti pengamatan.

II. METODOLOGI PENELITIAN

Menurut Berson dan Smith metode Naive Bayes merupakan metode yang digunakan memprediksi probabilitas. sedangkan klasifikasi Bayes adalah klasifikasi statistik yang dapat memprediksi kelas suatu anggota probabilitas. Untuk klasifikasi Bayes sederhana yang lebih dikenal sebagai naïve Bayesian Classifier dapat diasumsikan bahwa efek dari suatu nilai atribut sebuah kelas yang diberikan adalah bebas dari atribut-atribut lain. Naïve Bayes Classifier merupakan sebuah metoda klasifikasi yang berakar pada teorema Bayes. Ciri utama dari Naïve Bayes Classifier ini adalah asumsi yang sangat kuat (naif) akan independensi dari masing-masing kondisi/kejadian, dimana diasumsikan bahwa setiap atribut contoh (data sampel) bersifat saling lepas satu sama lain berdasarkan atribut kelas.



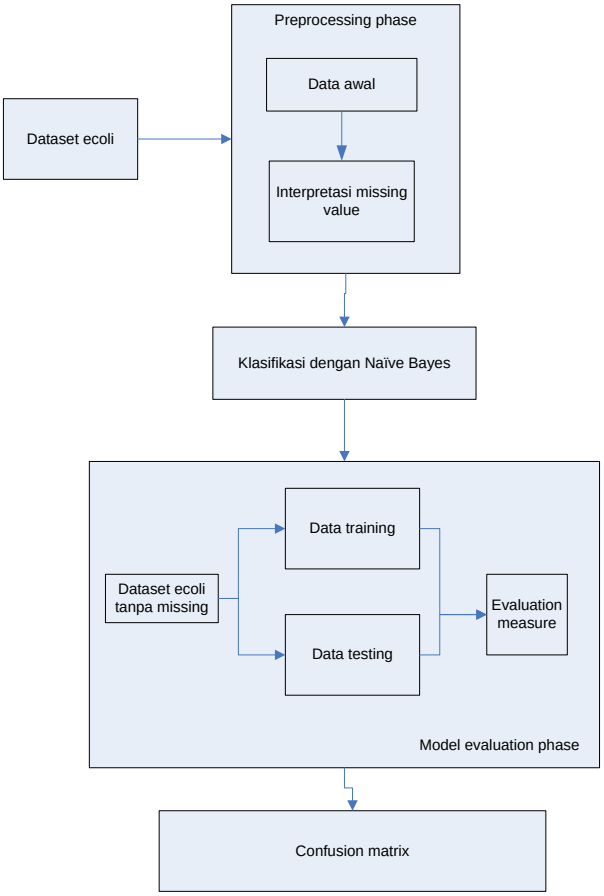
Gambar 2.1 algoritma Naïve bayes

Pada gambar 2.1 bagaimana algoritma naïve bayes bekerja. Pertama atribut akan dianalisa contoh atribut A lalu atribut A akan diperiksa didataset dengan nilai probabilitas, selanjutnya atribut A akan mendapatkan nilai prediksi didalam sebuah class. Atribut A dalam class akan diperbaharui. Jika atribut A sudah tidak

ditemukan maka akan dianalisa atribut lain sampai semua atribut dalam satu class selesai.

Naïve bayes kategorial adalah naive bayes dengan data statik berupa kategori atau merupakan data pasti, sehingga dalam pengerjaannya sudah didapatkan hasil yang pasti juga. Naive bayes merupakan metode dengan rumus dasar bayesian, Pada teorema Bayes, bila terdapat dua kejadian yangterpisah (misalkan A dan B), Dimana probabilitas dari A dengan ketentuan B didapat dari probabilitas data B terhadap A dikali kemudian dibagi peluang B. Untuk memperkirakan parameter pada distribusi fitur ini, seseorang harus mengasumsikan distribusi atau menghasilkan model nonparametric untuk fitur dari training data set (Maimoon & Rokach, 2010). Jika berhadapan dengan atribut bertipe data kontinu, sebuah asumsi yang khas adalah bahwa nilai-nilai terus menerus berhubungan dengan kelasnya masingmasing yang didistribusikan menurut metode distribusi Gaussian. Distribusi ini dikarakterisasi dengan dua parameter yaitu mean (μ), dan variansi(s2), untuk setiap kelas yj, peluang kelas bersyarat.

Naive Bayes merupakan teknik prediksi berbasis probabilistik sederhana yang berdasarpada penerapan teorema atau aturan bayesdengan asumsi independensi yang kuat padafitur, artinya bahwa sebuah fitur pada sebuahdata tidak berkaitan dengan ada atau tidaknya fitur lain dalam data yang sama. Namun Naive Bayes memiliki kelemahan yaitu attribut atau fitur independen sering salah dan hasil estimasi probabilitas tidak dapat berjalan optimal. Untuk mengatasi kelemahan tersebut salah satu caranya dengan metodepembobotan attribut untuk meningkatkanakurasi dari Naive Bayes tersebut dan nantinya hasil bobot atribut tersebut akandigunakan untuk menseleksi fitur dan atribut yang ada.Naive Bayes adalah pengklasifikasian statistik yang dapat digunakan untuk memprediksi probabilitas keanggotaan suatu class (Ridwan et. al, 2013).



Gambar 2.2 Model Pemikiran Penelitian

Pada model pemikiran penelitian dataset ecoli akan diperiksa apakah terdapat missing value atau tidak, jika ada maka pada fase preprocessing akan menghilangkan missing value. Selanjutnya dataset akan diklasifikan dengan algoritma naïve bayes dan kemudian dataset ecoli tanpa missing value akan dibagi menjadata training dan data testing yang nantinya akan menghasilakan evaluation measure berupa confusion matrix.

Pada fase evaluasi data training maka akan dilakukan beberapa tahapan diantaranya :

1. Baca *data training*
2. Hitung Jumlah dan probabilitas, namun apabila data numerik maka:
   a. Cari nilai mean dan standar deviasi dari masing-masing parameter yang merupakan data numerik. Adapun persamaan yang digunakan untuk menghitung nilai rata – rata hitung (mean) dapat dilihat sebagai berikut :

$$\mu = \sum xini = 1n$$

atau

$$\mu = x1 + x2 + x3 + ... + xnn$$

di mana :
μ : rata – rata hitung (*mean*)
xi : nilai sample ke -*i*
n : jumlah sampel

p-ISSN: 2442-2436, e-ISSN: 2550-0120

Dan persamaan untuk menghitung nilai simpangan baku (standar deviasi) dapat
dilihat sebagai berikut:

$\sigma = \text{v}\sum (xi\text{- } \mu)2ni{=}1n \text{ - } 1$
di mana :
s : standar deviasi
xi
: nilai x ke -*i*
μ : rata-rata hitung
n : jumlah sampel

b. Cari nilai probabilistik dengan cara menghitung jumlah data yang sesuai dari kategori yang sama dibagi dengan jumlah data pada kategori tersebut.

3. Mendapatkan nilai dalam tabel *mean*, standard deviasi dan probabilitas.
4. Solusi kemudian dihasilkan.

Naive Bayes didasarkan pada asumsi penyederhanaan bahwa nilai atribut secara kondisional saling bebas jika diberikan nilai output. Dengan kata lain, diberikan nilai output, probabilitas mengamati secara bersama adalah produk dari probabilitas individu. Keuntungan penggunaan Naive Bayes adalah bahwa metode ini hanya membutuhkan jumlah data pelatihan yang kecil untuk menentukan estimasi paremeter yang diperlukan dalam proses pengklasifikasian. Naive Bayes sering bekerja jauh lebih baik dalam kebanyakan situasi dunia nyata yang kompleks dari pada yang diharapkan (Keogh, 2006).

Persamaan dari teorema Bayes menurut Bustami adalah :

$$P(H|X) = \frac{P(X|H).\,P(H)}{P(X)} \qquad (1)$$

Di mana :
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*

Untuk menjelaskan metode *Naive Bayes*, perlu diketahui bahwa proses klasifikasi memerlukan sejumlah petunjuk untuk menentukan kelas apa yang cocok bagi sampel yang dianalisis tersebut. Karena itu, metode *Naive Bayes* di atasdisesuaikan sebagai berikut:

$$P(C|F1\ldots Fn) = \frac{P(C)P(F1\ldots Fn|C)}{P(F1\ldots Fn)}$$

Di mana Variabel *C* merepresentasikan kelas, sementara variabel *F1 ... Fn* merepresentasikan karakteristik petunjuk yang dibutuhkan untuk melakukan klasifikasi. Maka rumus tersebut menjelaskan bahwa peluang masuknya sampel karakteristik tertentu dalam kelas *C* (*Posterior*) adalah peluang munculnya kelas *C* (sebelum masuknya sampel tersebut, seringkali disebut *prior*), dikali dengan peluang kemunculan karakteristik-karakteristik sampel pada kelas *C* (disebut juga *likelihood*), dibagi dengan peluang kemunculan karakteristik-karakteristik sampel secara global (disebut juga *evidence*). Karena itu, rumus di atas dapat pula ditulis secara sederhana sebagai berikut:

$$Posterior = \frac{prior\ x\ likelihood}{evidence}$$

Nilai *Evidence* selalu tetap untuk setiap kelas pada satu sampel. Nilai dari *posterior* tersebut nantinya akan dibandingkan dengan nilai-nilai *posterior* kelas lainnya untuk menentukan ke kelas apa suatu sampel akan diklasifikasikan. Penjabaran lebih lanjut rumus *Bayes* tersebut dilakukan dengan menjabarkan $(C|F1, \ldots, Fn)$ menggunakan aturan perkalian sebagai berikut:

$P(C|F1, \ldots, Fn = P(C)P(F1, \ldots, Fn|C)$
$= P(C)P(F1|C)P(F2, \ldots, Fn|C, F1)$
$= P(C)P(F1|C)P(F2|C, F1)P(F3, \ldots, Fn|C, F1,F2$
$= (C)P(F1|C)P(F2|C, F1)P(F3|C, F1,F2)P(F4, \ldots, Fn|C, F1, F2, F3)$
$= P(C)P(F1|C)P(F2|C, F1)P(F3|C, F1,F2)\ldots P(Fn|C, F1, F2, F3, \ldots, Fn\text{-}1)$

Dapat dilihat bahwa hasil penjabaran tersebut menyebabkan semakin banyak dan semakin kompleksnya faktor-faktor syarat yang mempengaruhi nilai probabilitas, yang hampir mustahil untuk dianalisa satu persatu. Akibatnya, perhitungan tersebut menjadi sulit untuk dilakukan. Disinilah digunakan asumsi independensi yang sangat tinggi (*naif*), bahwa masing-masing petunjuk *(F1,F2...Fn)* saling bebas *(independen)* satu sama lain. Dengan asumsi tersebut, maka berlaku suatu kesamaan sebagai berikut:

$$P(Fi|Fj) = \frac{P(Fi \cap Fj)P(Fj)}{P(Fj)} = \frac{P(Fi)P(Fj)}{P(Fj)} = P(Fi)$$

Untuk i≠j , sehingga

$P(Fi|C, Fj) = P(Fi|C)$

Atau

$$P(X_k \mid Y) = \frac{P(Y \mid X_k)}{\sum_i P(Y \mid X_i)}$$

Jika P(Yes|X)>P(No|X), maka *record* diklasifikasikan sebagai *Yes*
Jika P(Yes|X)<P(No|X), maka *record* diklasifikasikan sebagai No

III. HASIL DAN PEMBAHASAN

Naïve bayes akan memprediksi kelas dari atribut ecoli yang ada di uci dataset. Dataset E.coli, terdiri dari

tujuh fitur atau atribut sebagai berikut: mcg, gvh, lip, chg, aac, alm1, alm2, dan protein diklasifikasikan ke dalam 8 kelas: sitoplasma (cp), membran dalam tanpa urutan sinyal (im), perisplasm (pp), dalam membran dengan uncleavable sinyal urutan (IMU), luar membran (om), luar membran lipoprotein (OML), dalam membran lipoprotein (IML), membran dalam dengan cleavable urutan sinyal (IMS).

Table 4.1 dataset Ecoli

Sequence Name	mcg	gvh	lip	chg	aac	alm1	alm2	Class Distribution
AAT_ECOLI	0,49	0,29	0,48	0,50	0,56	0,24	0,35	cp
ACEA_ECOLI	0,07	0,40	0,48	0,50	0,54	0,35	0,44	cp
ACEK_ECOLI	0,56	0,40	0,48	0,50	0,49	0,37	0,46	cp
ACKA_ECOLI	0,59	0,49	0,48	0,50	0,52	0,45	0,36	cp
ADI_ECOLI	0,23	0,32	0,48	0,50	0,55	0,25	0,35	cp
ALKH_ECOLI	0,67	0,39	0,48	0,50	0,36	0,38	0,46	cp
AMPD_ECOLI	0,29	0,28	0,48	0,50	0,44	0,23	0,34	cp
AMY2_ECOLI	0,21	0,34	0,48	0,50	0,51	0,28	0,39	cp
APT_ECOLI	0,20	0,44	0,48	0,50	0,46	0,51	0,57	cp
ARAC_ECOLI	0,42	0,40	0,48	0,50	0,56	0,18	0,30	cp
ASG1_ECOLI	0,42	0,24	0,48	0,50	0,57	0,27	0,37	cp
BTUR_ECOLI	0,25	0,48	0,48	0,50	0,44	0,17	0,29	cp
CAFA_ECOLI	0,39	0,32	0,48	0,50	0,46	0,24	0,35	cp
CAIB_ECOLI	0,51	0,50	0,48	0,50	0,46	0,32	0,35	cp
CFA_ECOLI	0,22	0,43	0,48	0,50	0,48	0,16	0,28	cp
CHEA_ECOLI	0,25	0,40	0,48	0,50	0,46	0,44	0,52	cp
CHEB_ECOLI	0,34	0,45	0,48	0,50	0,38	0,24	0,35	cp
CHEW_ECOLI	0,44	0,27	0,48	0,50	0,55	0,52	0,58	cp
CHEY_ECOLI	0,23	0,40	0,48	0,50	0,39	0,28	0,38	cp
CHEZ_ECOLI	0,41	0,57	0,48	0,50	0,39	0,21	0,32	cp

Berikut penjelasan dari setiap atribut :

Sequence Name: Accession number for the SWISS-PROT database adalah nomor akses untuk database SWISS-PROT.

1. mcg: *McGeoch's method for signal sequence recognition* disebut juga microgram yaitu alat ukur untuk satu per satujuta gram.
2. Gvh (*graft versus host*) : *von Heijne's method for signal sequence recognition* merupakan konsekuensi patologis dari respons umumnya diprakarsai oleh limfosit T immunocompetent ditransplantasikan ke host alogenik, imunologis tidak kompeten. Host tidak mampu menolak sel T dicangkokkan dan menjadi target mereka.
3. lip: *von Heijne's Signal Peptidase II consensus sequence score*. Binary attribute merupakan signal von Heijne peptidase II urutan konsensus skor dengan atribut berupa biner
4. chg: *Presence of charge on N-terminus of predicted lipoproteins*. Binary attribute merupakan prediksi banyaknya N-terminus lipoprotein, atribut berupa biner.
5. aac: *score of discriminant analysis of the amino acid content of outer membrane and periplasmic proteins* merupakan skor analisis diskriminan dari kandungan asam amino dari membran luar dan protein periplasmic.
6. alm1: *score of the ALOM membrane spanning region prediction program* merupakan skor/nilai dari membrane ALOM mencangkup daerah prediksi program.
7. alm2: *score of ALOM program after excluding putative cleavable signal regions from the sequence* merupakan skor program Alom setelah tidak termasuk didalam prediksi daerah alm1.

Algoritma naïve bayes dapat memprediksi secara cepat dan akurat data ecoli dengan perhitungan sebagai berikut :

$$Posterior = \frac{prior\ x\ likelihood}{evidence}$$

P sequence name (Yes) = 1/336 = 0,002976

P Sequence name (No)=335/336= 0,997024

p-ISSN: 2442-2436, e-ISSN: 2550-0120

Pmcg (Yes)=4/336=0,0119048
P mcg (No)=332/336=0,9880952

Pgvc (Yes)=4/336=0,0119048
P gvc (No)=332/336= 0,9880952

P Lip (Yes)= 326/336=0,9702381
Plip (No)= 10/336=0,02976

P chg (Yes)=335/336=0,9970238
P chg (No)= 1/336=0,0029762

P aac (Yes)= 11/336=0,0327381
P aac (No)= 325/336=0,9672619

P alm1 (Yes)=5/336=0,014881
P alm1 (No)= 331/336=0,9851

P alm2 (Yes)=12/336 =0,0357143
P alm2 (No)= 324/336=0,9642857

P (Yes) = 0,002976 x 0,0119048 x 0,0119048 x 0,9702381 x 0,9970238 x 0,0327381 x 0,014881 x 0,0357143 = 7,09924E-12

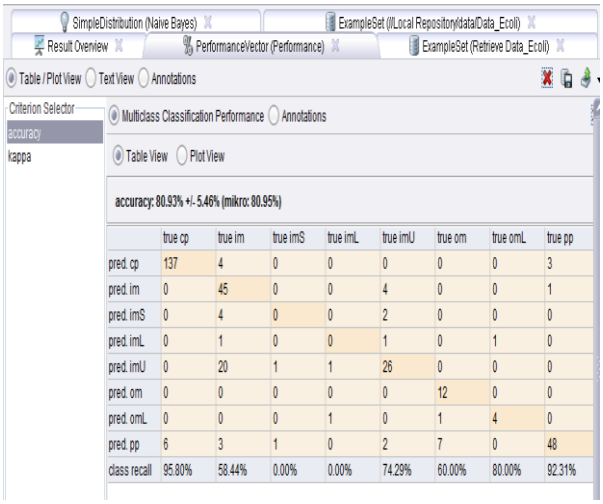
P (No) = 0,997024 x 0,9880952 x 0,9880952 x 0,0297619 x 0,0029762 x 0,9672619 x 0,985119 x 0,9642857 = 7,92252E-05

P (Yes) = 7,09924E-12
P (No) = 7,92252E-05

Maka P (Yes) < P (No) jadi record dengan sequence name AAT_ECOLI bernilai no

Evaluasi terhadap model yang terbentuk akan dilakukan dengan pengukuran akurasi. Akurasi diukur dengan menggunakan confusion matrix. Confusion matrix akan menggambarkan hasil akurasi mulai dari prediksi positif yang benar, prediksi positif yang salah, prediksi negatif yang benar, dan prediksi negatif yang salah (Han & Kamber, 2007 : p360).

Cara kerja Algoritma Naïve Bayes Classifier (NBC) Yaitu melalui dua tahapan yaitu: Learning Naïve Bayes adalah algoritma yang termasuk kedalam supervised learning, maka akan dibutuhkan pengetahuan awal untuk dapat mengambil keputusan.



	true cp	true im	true imS	true imL	true imU	true om	true omL	true pp
pred. cp	137	4	0	0	0	0	0	3
pred. im	0	45	0	0	4	0	0	1
pred. imS	0	4	0	0	2	0	0	0
pred. imL	0	1	0	0	1	0	1	0
pred. imU	0	20	1	1	26	0	0	0
pred. om	0	0	0	0	0	12	0	0
pred. omL	0	0	0	1	0	1	4	0
pred. pp	6	3	1	0	2	7	0	48
class recall	95.80%	58.44%	0.00%	0.00%	74.29%	60.00%	80.00%	92.31%

Gambar 4.1 nilai akurasi Naïve Bayes

Gambar 4.1 menunjukkan nilai akurasi yang dilakukan algoritma Naïve bayes untuk dataset Ecoli dengan tujuh fitur atau atribut dalam delapan class.

IV. KESIMPULAN

Penelitian ini dilakukan untuk mengetahui jenis bakteri gram-negatif berupa bakteri ecoli yang dibedakan dengan skema fenotip berdasarkan algoritma Naïve bayes yang memiliki tingkat akurasi dan kecepatan yang maksimal.

Pada penelitian ini secara umum mendapatkan nilai akurasi yang baik yaitu 80.93%, akan tetapi karena keterbatasan penelitian ini perlu disarankan untuk melakukan penelitian lanjutan yang berkaitan dengan klasifikasi untuk mendapatkan akurasi yang lebih baik. Adapun saran-saran yang perlu diberikan yaitu:

1. Perlu dilakukan penelitian lanjutan dengan penambahan fitur.
2. Perlu dilakukan penelitian yang sejenis dengan variasi proses model misalnya dengan penambahan Cross validation
3. Perlu dilakukan penelitian yang sejenis dengan variasi metode seleksi dalam pengklasifikasian misalnya feature selection.

REFERENSI

Alpaydin, Ethem. (2010). *Introduction to Machine Learning*. The MIT Press, London UK.

Asliyan, Rifat. (2011). *Syllable Based Speech Recognition*. Computer and Information Science. Diambil dari: http://www.intechopen.com/books/speech-technologies/syllable-based-speech-recognition. (3 Desember 2014).

Berson, A., and Smith S. J. (2001). Data Warehousing, Data Mining, & OLAP. New York, NY : McGraw-Hill.

Bevan, Nigel. (1997). *Quality and Usability: A New Framework*. National Physical Laboratory. UK.

Bustami. (2013). Penerapan Algoritma Naive Bayes Untuk Mengklasifikasi Data Nasabah Asuransi, TECHSI : Jurnal Penelitian Teknik Informatika, Vol. 3, No.2, Hal. 127-146.

Pattern Classification by R. O. Duda, P. E. Hart, D. Stork, Wiley and Sons.

E. Prasetyo. (2012). Data Mining : Konsep dan Aplikasi menggunakan MATLAB, 1st ed. Yogyakarta,Indonesia: Andi.

F. Gorunescu. (2011). Data Mining Concepts, Models and Techniques. Chennai, India: Springer.

Gorunescu, Florin. (2011). *Data Mining: Concepts, Models and Techniques*. Verlag Berlin Heidelberg, Springer. Jerman.

Guillet, Fabrice. Hamilton, Howard J. (2007). Quality Measures in Data Mining. Verlag Berlin Heidelberg, Springer. Jerman.

Han,J & Kamber, Micheline. (2007). *Data Mining Concepts, Models and Techniques*. Second Edition, Morgan Kaufmann Publisher. Elsevier.

Kadhim, Jehan & Abdulrazzaq, Mohammad (2015). Forecasting USD/IQD Future Values According to Minimum RMSE Rate. Thi_Qar University. pg.271–285.

Keogh, Eamonn, UCR. (2006). Pattern Recognition and Machine Learning, Christopher Bishop, Springer-Verlag.

Kusrini and E. T. Luthfi. (2009). Algoritma Data Mining, 1st ed. Yogyakarta, Indonesia: Andi.

Larose, D. (2005). Discovering Knowledge in Data. New Jersey, John Willey & Sons.Inc.

Larose, Daniel T. (2006). Data Mining Methods and Models. Hoboken New Jersey : Jhon Wiley & Sons, Inc.

Liao, Warren. T. & Triantaphyllou. Evangelos. (2007). *Recent Advances in Data Mining of Enterprise Data: Algorithms and Applications*. Series: Computer and Operation Research. 6. 190.

Lim TS, Loh WY, Shih YS. (1999). A comparison of prediction accuracy, complexity, and training time of thirty-three old and new classification algorithms. Kluwer Academic Publishers: Boston.

Maimon, Oded & Rokach, Lior. (2010). *Data Mining and Knowledge Discovery Handbook*, Springer, New York.

Myatt, Glenn J. (2007). *Making sense of data : A Practical Guide to Exploratory data analysis and Data Mining*. John Wiley & Sons Inc, New Jersey.

Patil, T. R., Sherekar, M. S. (2013). Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data Classification, International Journal of Computer Science and Applications, Vol. 6, No. 2, Hal 256-261.

Ridwan, M., Suyono, H., Sarosa, M. (2013). Penerapan Data Mining untuk Evaluasi Kinerja Akademik Mahasiswa Menggunakan Algoritma Naive Bayes Classifier, Jurnal EECCIS, Vol 1, No. 7, Hal. 59-64.

Shukla, Anupam. Tiwari, Ritu. & Kala, Rahul. (2010). *Real Life Application of Soft Computing*. New York: Taylor and Francis Groups, LLC.

S. N. Chatterjee and K. Chaudhuri. (2012). Outer Membrane Vesicles of Bacteria, SpringerBriefs in Microbiology, DOI: 10.1007/978-3-642-30526-9_2.

Sudjana. (1996). Metoda Statistika, Edisi ke-6. Bandung.

Vercellis,C. (2009). Business Intelligence: *Data Mining and Optimization for Decision Making*. Wiley.

Witten,I. Frank, E., & Hall. (2011*). Data Mining: Practical Machine Learning and tools*. Morgan Kaufmann Publisher, Burlington.

PROFIL PENULIS

Nama : Evy Priyanti
Tempat Lahir : Jakarta
Tanggal Lahir : 1 Februari 1986
Kuliah DIII di AMIK BSI lulus tahun 2007
Kuliah S1 di STMIK Kuwera lulus tahun 2008
Kuliah S2 di STMIK Nusa Mandiri lulus Tahun 2015
Paper yang pernah dipublikasi

1. Jurnal PARADIGMA Volume : XVII Nomor 2 Bulan September Tahun 2015. Judul "Peningkatan Backward Elimination Dengan Windowed Momentum Untuk Prediksi Kontrasepsi"
2. Jurnal Swabumi Volume IV no 1 maret 2016. Judul "Peningkatan Neural Network Dengan Feature Selection Untuk Prediksi Kanker Payudara".
3. Jurnal Pilar Nusa Mandiri Vol.XII, No.2 September 2016. Judul "Peningkatan Feature Selection Denganwindowed Momentum Untuk Prediksi Kanker Payudara".

 p-ISSN: 2442-2436, e-ISSN: 2550-0120