JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Journal homepage: https://jurnal.stkippgritulungagung.ac.id/index.php/jipi
ISSN: 2540-8984
Vol. 8, No. 3, September 2023, Pp. 985-994

IMPLEMENTASI METODE NAÏVE BAYES UNTUK KLASIFIKASI DATA BLOGGER

Nur Widiastuti*[1)], Arief Hermawan[2)], Donny Avianto[3)]
1. Universitas Teknologi Yogyakarta, Indonesia
2. Universitas Teknologi Yogyakarta, Indonesia
3. Universitas Teknologi Yogyakarta, Indonesia

Article Info
Kata Kunci: Data mining; Klasifikasi; Naïve Bayes; algoritma; Blogger

Keywords: *Data mining; Classification; Algorithm; Naïve Bayes; Blogger*

Article history:
Received 17 June 2023
Revised 1 July2023
Accepted 15 July 2023
Available online 1 September 2023

DOI :
https://doi.org/10.29100/jipi.v8i3.3713

* Corresponding author.
Corresponding Author
E-mail address:
widy.jogja@gmail.com

ABSTRAK
Di era teknologi yang modern seperti saat ini peluang kerja sebagai blog-ger cukup banyak diminati. Para blogger me-manfaatkan situs blog baik yang gratis maupun berbayar untuk menulis artikel. Hal tersebut me-nyebabkan pengguna situs blog semakin meningkat. Diantara para blog-ger ada yang menjadi blogger professional dan ada juga yang menjadi blog-ger musiman untuk menulis artikel pada blog. Penelitian ini meneliti blogger mana yang masuk dalam kategori blogger professional atau blog-ger musiman. Penelitian ini mengklasifikasi data blogger yang diambil dari UCI Machine Learning dengan jumlah data sebanyak 100 data kemudian diuji menggunakan Metode Naïve Bayes. Adapun tool yang digunakan untuk penelitian adalah Rapidminer untuk mengklasifikasi blogger professional atau blogger musiman. Penelitian ini menghasilkan akurasi sebesar 76,27% atau meningkat 1,27 % dibandingkan penelitian sebelumnya dan hasil classification error sebesar 23,73%. Sedangkan class recall sebanyak 12 fold, hal ini dapat diartikan penelitian menggunakan correla-tion matrix dan cross validation dengan number of fold 12 menghasilkan nilai akurasi yang lebih baik dari penelitian sebe-lumnya.

ABSTRACT
In the era of modern technology like today, job opportunities as bloggers are quite popular. Bloggers take advantage of both free and paid blog sites to write articles. This causes blog site users to increase. Among the blog-gers there are those who become professional bloggers and there are also those who become seasonal bloggers to write articles on blogs. This re-search examines which bloggers fall into the category of professional bloggers or seasonal bloggers. This study classifies blogger data taken from UCI Machine Learning with a total of 100 data then tested using the Naïve Bayes Method. The tool used for research is Rapidminer to classify professional bloggers or seasonal bloggers. This study resulted in an ac-curacy of 76.27% or an increase of 1.27% compared to previous studies and a classification error of 23.73%. While the class recall is 12 folds, this means that research using a correlation matrix and cross validation with a number of folds of 12 produces a better accuracy value than previ-ous studies.

I. PENDAHULUAN

SEORANG blogger adalah seseorang yang melakukan kegiatan blogging atau tulis menulis dengan menggunakan berbagai platform blog. Blogger merupakan salah satu profesi yang paling banyak diminati di era teknolo-gi yang modern seperti saat ini. Peluang bisnis yang mendapatkan cukup banyak penghasilan jika ditekuni secara mendalam. Profesi blogger mengandalkan kegiatan dengan domain yang gratis hingga berbayar un-tuk menulis opini atau review cerita yang menarik perhatian para pembacanya. Blogger profesional harus mem-iliki kemampuan menulis yang baik dilihat dari tata bahasa, alur cerita tulisan, dan kemudahan tu-lisannya di-mengerti pembaca. Kemampuan ini tidak harus sebaik sastrawan atau pujangga, minimal tu-lisannya dipahami pembaca. Pada penelitian sebelumnya tentang pengguna situs blogger yang professional maupun yang tidak, kemudian dilakukan pengklasifikasian data untuk mengetahui pengguna tersebut masuk dalam kategori pengguna blogger professional atau bukan. Sebagai referensi terkait penelitian ini adalah penelitian yang sudah dilakukan oleh peneliti sebelumnya. Teknik pengklasifikasian pemodelan deskriptif dan prediktif dengan algoritma data

mining yaitu menggunakan metode Naïve Bayes. Untuk mengelola data digunakan software rapid miner studio 6.0, dataset blogger diperoleh dari website UCI Machine learning Repository, Perhitungan performance vector menunjukkan akurasi klasifikasi metode Naive Bayes diperoleh sebesar 75%. Sedangkan class precision dan class recall untuk prediksi yes menunjukkan tingkat precision sebesar 77% dan Sedangkan *class precision* dan *class recall* untuk prediksi *yes* menunjukkan tingkat *precision* sebesar 76,54% dan untuk prediksi *no* sebesar 68,42%. Berdasarkan penelitian menggunakan algoritma Naïve Bayes memiliki keuntungan dapat mengklasifikasi data blogger yang professional dan musiman, namun nilai akurasi yang dihasilkan masih kurang tinggi yakni 75% [1].

Penelitian lainnya terkait klasifikasi penyakit ISPA yang diderita oleh masyarakat menggunakan metode Naive Bayes Classifier. Data diambil dari puskesmas karena puskesmas merupakan salah satu yang menjadi rujukan pengobatan untuk masyarakat. Puskesmas harus mengidentifikasi jenis ISPA yang tepat sehingga pengobatan bagi penderita ISPA dapat diberikan secara optimal. Ini penelitian mengklasifikasikan data pasien ISPA di puskesmas berdasarkan faktor penentu yaitu penyakit yang diderita, umur, dan jangka waktu tinggal. Klasifikasi dilakukan dengan menggunakan Naive Bayes Classifier dengan metode pengujian Confusion Matrix. Hasil dari penerapan metode Naive Bayes Classifier pada data pasien menghasilkan tiga jenis ISPA yaitu ringan, sedang dan berat. Paling atas jumlah penderita ISPA adalah ISPA berat. Hasil dari Kebingungan Uji matriks yang telah dilakukan membuktikan bahwa metode ini memiliki akurasi sebesar 93,33% sehingga cocok digunakan untuk mengklasifikasikan penyakit ISPA. Penelitian tentang klasifikasi penyakit ISPA ini menghasilkan tingkat akurasi yang sangat baik atau akurasi yang baik yakni 93,33%[2].

Penelitian selanjutnya tentang Klasifikasi Kualitas Produk Kelapa Sawit Menggunakan Metode Naïve Bayes. Pulau Sumatera Dan Kalimantan Adalah Pulau Yang Memiliki Perkebunan Terbesar Khususnya Kelapa Sawit Di Indonesia. Propinsi Riau Yang Berada Di Pulau Sumatera Dapat Menghasilkan Kelapa Sawit Tertinggi Di Pulau Sumatera. Mutu Merupakan Komponen Penting Dalam Kelangsungan Bisnis Pada Industri Minyak Kelapa Sawit. Kualitas Produk Mentah Kelapa Sawit Sebuah Perusahaan Di Kecamatan Kerumutan Kabupaten Pelalawan Riau Tergantung Pada Kandungan Dari Hasil Akhir Pengolahan. Kandungan Tersebut Terdiri Dari Kadar Kotoran Pada Crude Palm Oil (Cpo), Kadar Moisture Cpo, Kadar Free Fatty Acid Cpo, Deteration Of Bleachability Index Cpo, Carotin Cpo, Dirt Kernel, Moisture Kernel, Dan Broken Kernel. Kualitas Akhir Produk Kelapa Sawit Ditentukan Dari Hasil Gabungan Kualitas Cpo Dan Kualitas Kernel. Bahan Mentah Yang Berkualitas Baik Akan Mempengaruhi Harga Jual Bahan Mentah Tersebut Untuk Menghasilkan Kualitas Akhir Produk Yang Baik. Untuk Menentukan Kualitas Produk Kelapa Sawit Menimbulkan Masalah Dalam Segi Waktu Karena Harus Dicek Satu Per Satu Melalui Pengolahan Di Laboratorium. Pembangun Aplikasi Yang Dapat Menentukan Klasifikasi Kualitas Produk Kelapa Sawit Merupakan Tujuan Dari Penelitian Ini. Aplikasi Ini Diharapkan Dapat Membantu Petugas Labor Dalam Proses Klasifikasi Kualitas Produk Mentah Kelapa Sawit Dengan Lebih Cepat, Tepat Dan Akurat. Penelitian Ini Menggunakan Algoritma Naïve Bayes Karena Memerlukan Data Latih Dalam Jumlah Yang Lebih Kecil Pada Proses Klasifikasi Data. Tingkat Akurasi Metode Naïve Bayes Dalam Menentukan Kualitas Produk Kelapa Sawit Adalah Sebesar 82,05%. Penelitian tentang Klasifikasi Kualitas Produk Kelapa Sawit menggunakan metode Naïve Bayes ini menghasilkan nilai akurasi yang baik yakni sebesar 82,05%, nuamun jika dilakukan pengujian ulang dengan metode lain kemungkinan dapat menghasilkan akurasi yang lebih tinggi[3].

Penelitian selanjutnya tentang membangun model ketenagakerjaan lulusan menggunakan tugas klasifikasi dalam penambangan data, membandingkan beberapa pendekatan penambangan data seperti metode Bayesian dan metode Pohon dengan visualisasi, dan mengeksplorasi Aturan Asosiasi menggunakan[4]. Penelitian lain yakni tentang Data Mining Model For Designing Diagnostic Applications Inflammatory Liver Disease. Pada penelitian ini membandingkan beberapa metode klasifikasi data mining, antara lain algoritma C4.5, Naïve Bayes, dan k-Nearest Neighbor untuk mendiagnosis penyakit radang hati, kemudian membandingkan mana dari ketiga metode tersebut yang paling akurat. Berdasarkan hasil pengukuran performansi ketiga model menggunakan metode Cross Validation, Confusion Matrix dan ROC Curve, diketahui bahwa metode C4.5 merupakan metode terbaik dengan akurasi 70,99% dan under the curva (AUC).) nilai 0,950, kemudian metode k-Nearest Neighbor dengan akurasi 67,19% dan nilai under the curve (AUC) 0,873, kemudian metode Naïve Bayes dengan tingkat akurasi 66,14% dan nilai under the curve (AUC) sebesar 0,742. Selanjutnya penelitian yang dilakukan oleh (Prayoga, 2018) dan kawan-kawan yang meneliti tentang Diagnosis penyakit hepatitis dengan metode Naïve Bayes. Dari pengujian tingkat akurasi yang dilakukan, diketahui bahwa nilai akurasi yang dihasilkan dengan metode Naïve Bayes cukup baik atau cukup tinggi yakni sebesar 87,50%, namun jika pengujiannya ditambahkan dengan feature selection kemungkinan dapat menghasilkan akurasi yang lebih tinggi[5].

Setelah mempelajari beberapa penelitian sebelumnya, maka penelitian ini akan menguji kembali klasifikasi data blogger professional dan musiman dengan algoritma Naïve Bayes dengan menerapkan correlation matrix dan feature selection Principal component analysis (PCA). Naïve bayes dipilih dalam penelitian ini karena memiliki keunggulan antara lain Naive Bayes sederhana dan komputasinya tidak tinggi dan tidak membutuhkan biaya yang

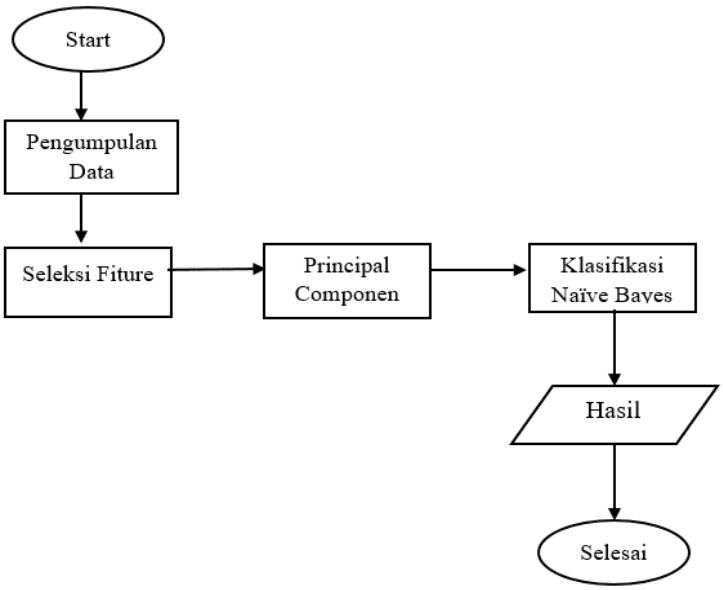
banyak untuk menjalakannya, selain itu mudah diimplementasikan dan cepat menyatu dari pada model diskriminatif seperti regresi logistik, serta membutuhkan lebih sedikit data pelatihan.

Kelebihan PCA pada penelitian sebelumnya adalah untuk mengurangi dimensi dengan membentuk variabel-variabel baru yang disebut Principal Components. Sedangkan PCA dipilih pada pada penelitian ini karena untuk mengenerate atau menyederhanakan fiture dalam jumlah yang banyak menjadi beberapa fiture yang lebih distinguisble atau dapat di bedakan oleh klasifier sehingga dapat meningkatkan akurasi. Adapun penelitian yang digunakan sebagai referensi panduan terkait penelitian ini adalah penelitian yang dilakukan oleh Recha Abriana Anggraini, Galih Widagdo, Arief Setya Budi, M. Qomaruddin, 2019 dengan judul Penerapan Data Mining Classification untuk Data Blogger Menggunakan Metode Naïve Bayes.

I. METODE PENELITIAN

A. Alur Penelitian

Berikut ini adalah alur penelitian yang akan dilakukan pada penelitian ini, adapun alurnya meliputi pengumpulan data, seleksi fiture yang akan digunakan, Principal component analysis (PCA), kemudian pengujian dengan lagoritma Naïve Bayes hingga di dapatkan hasil. Cara kerja dari metode Naïve Bayes adalah dengan menghitung peluang dari satu kelas dari masing-masing kelompok atribut yang ada dan menentukan kelas mana yang paling optimal, artinya pengelompokan dapat dilakukan berdasarkan atribut yang yang akan digunakan sebagai label. Berikut ini adalah alur penelitian yang dapat dilihat pada Gambar 1.



Gambar 1. Alur Penelitian

B. Pengumpulan Data

Penelitian ini menggunakan dataset dari UCI Machine Learning yang terdiri dari 100 baris dan 6 kolom atau atribut. Atribute tersebut adalah Local Political Social Space (LPSS), Local Media Turnover (LMT), topics, Caprice, *dan* Degree. Sedangkan untuk class atau label yang terdapat pada dataset tersebut yaitu Professional Blogger (PB) dengan nilai yes dan no. Teknik pengolahan dataset dalam penelitian ini menggunakan teknik klasifikasi dengan permodelan deskriptif dan prediktif menggunakan salah satu algoritma data mining yaitu metode Naïve Bayes. Penelitian sebelumnya yang mengambil datasetd ari UCI Machine Learning yakni Analytic Predictive of Hepatitis using The Regression Logic Algorithm. Seperti pada penelitian sebelumnya, teknik regresi dalam machine learning yang digunakan adalah Logistic Regression untuk prediksi hepatitis dengan menggunakan dataset yang diperoleh dari UCI Machine Learning Repository. Dimana hasil akhir penelitian ini adalah penerapan algoritma Logistic Regression sebagai metode prediksi cukup baik dengan memperoleh tingkat akurasi sebesar 83,33% [1].

C. Data Mining

Data mining adalah suatu proses yang memiliki tujuan untuk menemukan suatu pola otomatis atau semi-otomatis dari data yang sudah kita dapat/miliki di dalam basis data yang dimanfaatkan untuk penyelesaian suatu masalah. Data mining memiliki beberapa teknik, diantaranya klasifikasi dan clustering. Teknik klasifikasi adalah teknik

pembelajaran yang digunakan untuk memprediksi nilai dari atribut kategori target. Metode yang paling populer digunakan untuk teknik klasifikasi adalah Decision Trees, Naïve Bayes , Statistical analysis, dan lain lain. Data mining *berisi pencarian trend atau pola yang diinginkan dalam* database *besar untuk membantupengambilan keputusan di waktu yang akan datang.*

D. Seleksi Fitur

Seleksi fitur bertujuan guna mengurangi atribut, sebab banyaknya fitur pada dataset yang digunakan dapat mengakibatkan overfitting dan perlu menentukan fitur yang dibutuhkan dalam pengolahan data. Seleksi fitur menggunakan heatmap untuk menyeleksi fitur dari dataset dilihat dari nilai hubungan antar fitur-fitur tersebut terhadap variabel outcome sebagai target value yang akan diprediksi. Fitur yang terpilih merupakan fitur yang akan dipergunakan karena mempunyai hubungan tertinggi terhadap variabel outcome.

*E.*Correlation Matrix

Operator ini menentukan korelasi antara semua Atribut dan dapat menghasilkan vektor bobot berdasarkan korelasi ini. Korelasi adalah teknik statistik yang dapat menunjukkan apakah dan seberapa kuat pasangan Atribut terkait. Korelasi adalah angka antara -1 dan +1 yang mengukur tingkat hubungan antara dua Atribut (sebut saja X dan Y). Nilai positif untuk korelasi menyiratkan asosiasi positif. Dalam hal ini nilai X yang besar cenderung diasosiasikan dengan nilai Y yang besar dan nilai X yang kecil cenderung diasosiasikan dengan nilai Y yang kecil. Nilai negatif untuk korelasi mengimplikasikan asosiasi negatif atau terbalik. Dalam hal ini nilai X yang besar cenderung diasosiasikan dengan nilai Y yang kecil dan sebaliknya.

F.Klasifikasi

Klasifikasi merupakan salah satu teknik dalam pengolahan data yang bekerja dengan cara objek yang dipergunakan dibagi menjadi kelas-kelas dengan jumlah kelas sesuai dengan yang diinginkan. Klasifikasi dapat menciptakan suatu pola yang dapat memisahkan tiap-tiap kelas data yang bertujuan guna menentukan objek yang tergolong ke dalam kategori tertentu dilihat dari perilaku serta atribut dari kelompok yang telah didefinisikan [6]. Klasifikasi yang dilakukan dalam penelitian ini bertujuan untuk menggolongkan data termasuk ke dalam kelas blogger professional dan blogger amatir. Salah satu penelitian sebelumnya tentang Klasifikasi Penelitian Klasifikasi data stunting di Kabupaten/Kota di Indonesia berdasarkan faktor penyebab stunting pada balita, yaitu menggunakan metode clustering dengan algoritma K-Means. Tujuannya adalah untuk membantu pemerintah dalam mengambil kebijakan yang sesuai terkait penurunan prevalensi stunting pada balita berdasarkan karakteristik dan permasalahan masing-masing cluster. Hasil penelitian menunjukan bahwa dengan bantuan metode elbow menghasilkan 2 cluster sebagai cluster terbaik dengan nilai selisih Sum of Square Error(SSE) sebesar 1401.5156, dimana cluster 1 merupakan cluster dengan faktor penyebab stunting tinggi yang terdiri dari 324kabupaten/kota, dan cluster 2 merupakan cluster dengan faktor penyebab stunting rendah yang terdiri dari 49 kabupaten/kota. Data latih dan data uji yang digunakan untuk klasifikasi berjumlah 75 data latih dan 25 data uji. Berdasarkan master pelanggan yang dijadikan data latih, telah berhasil mengklasifikasikan 23 data dari 25 data yang diuji. Sehingga berhasil memprediksi pelanggannya dengan nilai *precision* mencapai 100%, nilai *recall* mencapai 91%, nilai *accuracy* mencapai 92%.[7]

G. Naïve Bayes

Naïve Bayes merupakan salah satu dari metode pengklasifikasian. NBC dipilih karena merupakan metode klasifikasi yang simpel dan efisien. Naïve Bayes dapat diterapkan pada data yang lumayan besar/banyak jumlahnya[9], dan dapat menangani data yang tidak lengkap (memiliki missing value). Teori Naïve Bayes memiliki kemampuan klasifikasi yang serupa dengan decision tree dan neural network bahkan algoritma naïve bayes memiliki akurasi dan kecepatan yang tinggi saat diaplikasikan ke dalam database dengan data yang besar. Teorema bayes dapat ditulis menggunakan persamaan 1[8]:

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

Dimana:
$P(AA|BB)$ = Probabilitas posterior dari A pada kondisi B (posterior probability).
$P(BB|AA)$ = Probabilitas posterior dari B pada kondisi A(likelihood).
$P(AA)$ = Probabilitas prior dari A (class prior probability).
$P(BB)$ = Probabilitas prior dari B (predictor prior probability).

JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Journal homepage: https://jurnal.stkippgritulungagung.ac.id/index.php/jipi
ISSN: 2540-8984
Vol. 8, No. 3, September 2023, Pp. 985-994

Salah satu penelitian sebelumnya tentang Naïve Bayes adalah Penelitian tentang perbandingan data metode klasifikasi pertambangan dibuat untuk mendeteksi keaslian lowongan kerja di media sosial. Metode yang digunakan adalah Naif Bayes, KNN, dan pohon keputusan. Untuk mengetahui yang mana metode memiliki nilai akurasi tertinggi dan dapat digunakan untuk mengklasifikasikan keaslian lowongan kerja, dan penipuan di media sosial dapat terjadi dicegah. Berdasarkan penelitian ini, metode yang memiliki nilai tertinggi nilai akurasi adalah metode KNN. Nilai akurasi sebesar 94,93%, sedangkan model Decision Tree memiliki nilai akurasi sebesar 91,57% dan model Naive Bayes memiliki akurasi sebesar 84,35%. KNN metode adalah metode terbaik untuk mengklasifikasikan keaslian pekerjaan Lowongan[8].

H. Langkah-langkah Analisis Naïve Bayes

Setelah hasil Naïve Bayes didapatkan peneliti akan melakukan beberapa langkah-langkah analisis terhadap hasil Naïve Bayes. Langkah yang pertama adalah mencari nilai akurasi dari algoritma Naïve Bayes tanpa PCA. Kemudian peneliti membuat table Data simple distribution untuk memetakan berapa data yang masuk ke dalam kelas Yes dan berapa yang masuk dalam kelas No. Selanjutnya penulis mencari hasil Class Precision dan Class Recall. Setelah itu peneliti akan membandingkan hasil penelitian yang sudah silakukan dengan hasil penelitian sebelumnya kemudian menyimpulkan hasil analisanya.

II.HASIL DAN PEMBAHASAN

A.Dataset

Dataset adalah kumpulan daya atau dokumen yang berisi satu atau lebih catatan record. Setiap kelompok record ini tadi disebut sebagai dataset dan memiliki peran untuk menyimpan informasi pada data. Dalam permodelan algoritma Naïve Bayes pada penelitian ini, data training diolah menggunakan model Naïve Bayes serta cross validation 12 fold. Jumlah responden dalam penelitian ini sebesar 100 responden, dengan dataset blogger ini peneliti akan mengklasifikasikan jenis blogger kedalam 2 kelompok yaitu Blogger Professional (BP) dan Blogger Musiman (BM). Data training dari dataset blogger dapat dilihat pada Tabel I.

TABEL I
DATA TRAINING

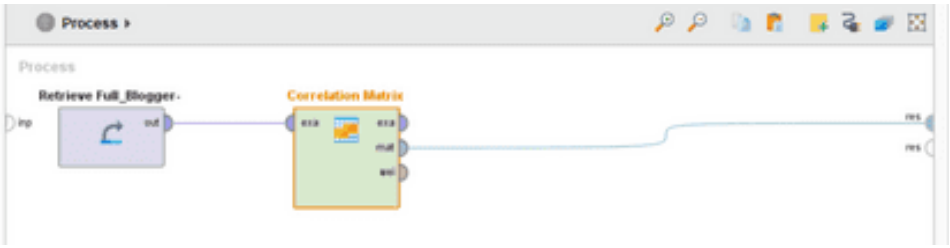
Degree	caprice	Topic	lmt	lpss	pb
high	left	impression	yes	yes	yes
high	left	political	yes	yes	yes
medium	middle	Tourism	yes	yes	yes
high	left	political	yes	yes	yes
medium	middle	news	yes	yes	yes
medium	middle	news	yes	yes	yes
high	left	political	yes	yes	yes
high	right	political	yes	no	yes
high	right	political	yes	no	no
medium	right	tourism	yes	no	yes
high	right	tourism	yes	yes	yes
medium	left	news	yes	no	yes
high	left	political	yes	yes	no
low	right	news	no	yes	No
high	left	political	yes	yes	Yes
medium	left	impression	yes	yes	Yes
medium	left	political	yes	yes	Yes
high	right	political	yes	yes	Yes
medium	left	impression	yes	yes	yes

JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Journal homepage: https://jurnal.stkippgritulungagung.ac.id/index.php/jipi
ISSN: 2540-8984
Vol. 8, No. 3, September 2023, Pp. 985-994

Tabel 1 merupakan *dataset blogger* yang dipakai sebagai *data training* dalam penelitian ini. Data tersebut berjumlah 100 data yang direpresentasikan dalam bentuk tabel. Atribut yang terdapat pada *dataset blogger* yaitu *Local Political Social Space (LPSS)* atau diartikan kecenderungan dalam membahas Ruang Sosial Politik Lokal, *Local Media Turnover (LMT)* atau bisa diartikan kecenderungan membahas pergantian media lokal, *topics, professional Blogger (PB)* atau blogger yang profesional, *Caprice* (perilaku), dan *Degree* (derajat/level). Berdasarkan pengujian menggunakan correlation matrix *Local Political Social Space (LPSS).*

B.Data Training

Data training adalah dataset yang digunakan untuk pembelajaran, sedangkan data testing adalah dataset yang digunakan untuk pengujian, sehingga menghasilkan output berupa prediksi [10]. Pada penelitian ini akan dilakukan menghitung hubungan antar variabel menggunakan operator Correlation Matrix, yakni menghitung korelasi yang berisi nilai-nilai korelasi antara variabel-variabel yang akan dianalisis. Proses Corelation matrix dapat dilihat pada Gambar 2.



Gambar 2. Proses Correlation Matrix

Setelah dilakukan proses correlation matrix maka muncul hasil, bahwa yang kekuatan hubungannya semakin besar nilai correlation-nya maka semakin kuat/banyak hubungangannya begitu juga sebaliknya semakin kecil nilai correlationnya maka semakin lemah/sedikit hubungannya contohnya adalah hubungan antara keterangan (penun-daan) dengan visibility memiliki hubungan yang sangat kuat karena blogger profesional pada visibility sehingga nilai correlationnya besar, dapat dilihat pada Gambar 3.

Attribut...	Degree	caprice	topic	lmt	lpss
Degree	1	?	?	?	?
caprice	?	1	?	?	?
topic	?	?	1	?	?
lmt	?	?	?	1	0.134
lpss	?	?	?	0.134	1

Gambar. 3. Hasil Correlation Matrix

Berbeda dengan penelitian sebelumnya yang tidak melalui proses correlatin matrix yakni menghitung korelasi antar table atau variable yang di analisa. Pada penelitian ini melalui proses Correlation Matrix dengan hasil bahwa faktor yang paling berpengaruh dalam menentukan blogger profesional adalah *Local Political Social Space (LPSS)*, maka attribute LPSS tidak akan digunakan dalam pengujian.

C.PERHITUNGAN PROBABILITAS NAÏVE BAYES

Perhitungan Class Probabilitas (P(PB) dan P(BM))
Dari 100 data latih yang digunakan, diketahui kelas Profesional Blogger (PB) sebanyak 68 data, dan kelas Blogger Musiman (BM) sebanyak 32 data. Perhitungan probabilitas untuk kemungkinan kelas Profesional Blogger dan Blog-ger Musiman sebagai berikut :
Perhitungan probabilitas untuk kemungkinan kelas Blogger Profesional (BP) dapat dilihat pada persamaan (1) :

$$P(PB) \frac{68}{100} = 0{,}68 \qquad (1)$$

Pada perhitungan PB mendapatkan hasil sebesar 0,68, sedangkan Perhitungan Probabilitas prior untuk kemungkinan kelas Blogger Musiman (BM) dapat dilihat pada persamaan (2) :

JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Journal homepage: https://jurnal.stkippgritulungagung.ac.id/index.php/jipi
ISSN: 2540-8984
Vol. 8, No. 3, September 2023, Pp. 985-994

$$P(BM) \frac{32}{100} 0{,}32 \qquad (2)$$

Pada perhitungan BM mendapatkan hasil sebesar 0,32, selanjutnya menghitung Probabilitas Degree (D) dengan menggunakan D dengan menghitung banyaknya jumlah data low, medium dan high pada atribut Degree mendapatkan hasil Degree low No sebesar 10, degree low yes sebesar 4. Kemudian degree medium No sebesar 13, degree medium Yes sebesar 34 dan degree high No sebesar 9 dan degree high Yes sebesar 30, hasil dapat dilihat pada Tabel II.

TABEL II
PROBABILITAS PRIOR

Degree	No	Yes
low	10	4
medium	13	34
high	9	30

Langkah selanjutnya adalah menghitung Conditional Probabilities, yaitu probililitas setiap nilai input terhadapat nilai class, pada Tabel III.

TABEL III
CONDITIONAL PROBABILITIES

Degree	No	Yes
Low	$P(low\|No) = \frac{jumlah\ low\ yang\ no}{jumlah\ no} = \frac{10}{32} = 0.3125$	$P(low\|Yes) = \frac{jumlah\ low\ yang\ yes}{jumlah\ yes} = \frac{4}{68} = 0.058824$
medium	$P(medium\|No) = \frac{jumlah\ medium\ yang\ no}{jumlah\ no} = \frac{13}{32} = 0.40625$	$P(medium\|No) = \frac{jumlah\ medium\ yang\ yes}{jumlah\ yes} = \frac{34}{68} = 0.5$
high	$P(high\|No) = \frac{jumlah\ high\ yang\ no}{jumlah\ no} = \frac{9}{32} = 0.28125$	$P(high\|No) = \frac{jumlah\ high\ yang\ yes}{jumlah\ yes} = \frac{30}{68} = 0.441176$

Peluang yang didapatkan adalah :
P(low|No) = 0,3125
P(medium|No) = 0,40625
P(high|No) = 0,28125
P(low|Yes) = 0,058824
P(medium|Yes) = 0,5
P(High|Yes) = 0,441176

Selanjutnya adalah menentukan Probabilitas Caprice, didapatkan hasil pada TabeL IV.

TABEL IV
PROBABILITAS PRIOR

Caprice (CP)	No	Yes
left	10	42
middle	4	10
right	18	16

Pada perhitungan Probabilitas Caprice (CP) dengan menghitung banyaknya jumlah data left, middle dan right pada atribut left mendapatkan hasil CP left No sebesar 10, CP middle No sebesar 4 dan CP right No sebesar 18. Kemudian CP left Yes sebesar 42, CP middle No sebesar 10, CP right Yes sebesar 16. Hasil dapat dilihat pada Tabl VI.

TABEL V
CONDITIONAL PROBABILITIES

Degree	No	Yes
Left	$P(left\|No) = \frac{jumlah\ left\ yang\ no}{jumlah\ no} = \frac{10}{32} = 0.3125$	$P(left\|Yes) = \frac{jumlah\ left\ yang\ yes}{jumlah\ yes} = \frac{42}{68} = 0.617647$
Middle	$P(middle\|No) = \frac{jumlah\ middleyang\ no}{jumlah\ no} = \frac{4}{32} = 0.125$	$P(middle\|No) = \frac{jumlah\ middle\ yang\ yes}{jumlah\ yes} = \frac{10}{68} = 0.147059$
Right	$P(right\|No) = \frac{jumlah\ right\ yang\ no}{jumlah\ no} = \frac{18}{32} = 0.5625$	$P(right\|No) = \frac{jumlah\ right\ yang\ yes}{jumlah\ yes} = \frac{16}{68} = 0.23529$

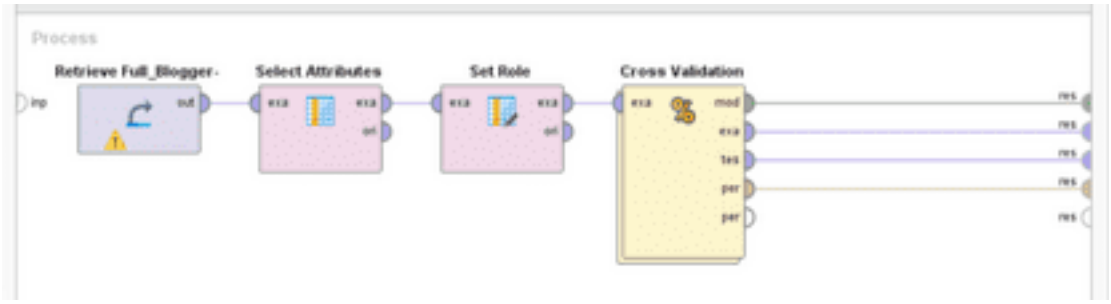
Peluang yang didapatkan adalah :

JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Journal homepage: https://jurnal.stkippgritulungagung.ac.id/index.php/jipi
ISSN: 2540-8984
Vol. 8, No. 3, September 2023, Pp. 985-994

P(left|No) = 0,3125
P(middle|No) = 0,125
P(right|No) = 0,5625
P(left|Yes) = 0,617647
P(middle|Yes) = 0, 147059
P(right|Yes) = 0, 23529

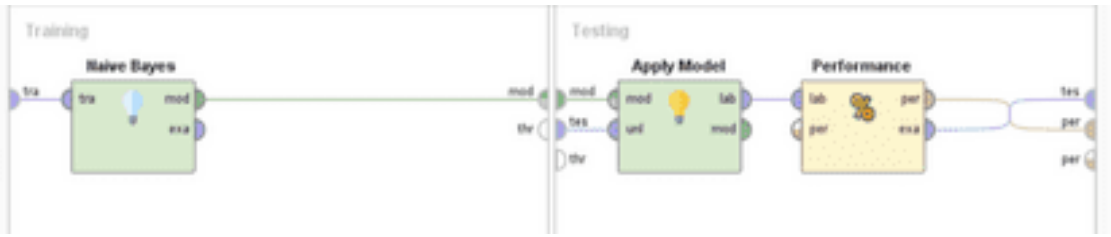
B. PENGUJIAN ALGORITMA NAÏVE BAYES

Pada proses ini akan dilakukan pengujian dengan algoritma klasifikasi Naïve Bayes mengunakan tool Rapidminer, meliputi pemilihan atribut yang akan digunakan untuk pengujian menggunakan operator select Attributes, dilanjutkan penentuan label menggunakan operator set role dan cross validation dengan number of fold 12 kali, dapat dilihat pada Gambar 4.



Gambar 4. Proses pengujian mengggunakan Cross Validation

Pada pengujian menggunakan algortima Naïve Bayes dataset yang semula berjumlah 6 atributte hanya digunakan 5 atributte, yakni *Local Media Turnover (LMT), topics, Caprice, Degree dan professional Blogger (PB),* sedangkan attribute *Local Political Social Space (LPSS)* tidak digunakan, dapat dilihat pada Gambar 5.



Gambar 5. Proses pengujian mengggunakan Algoritma Klasifikasi Naïve Bayes

Gambar diatas adalah proses pengujian menggunakan algoritman Naïve Bayes, kemudian apply model dan performance dengan memilih option accuracy dan classification error agar muncul hasil accuracy performancenya.

D. HASIL

E. Simple Distribution Models

Dengan menggunakan software rapid miner studio untuk menganalisis tabel data blogger dengan menggunakan metode Naïve Bayes dapat menghasilkan beberapa kelas utama pembagian dibandingkan dengan hasil penelitian sebelumya dapat dilihat pada Tabel VI.

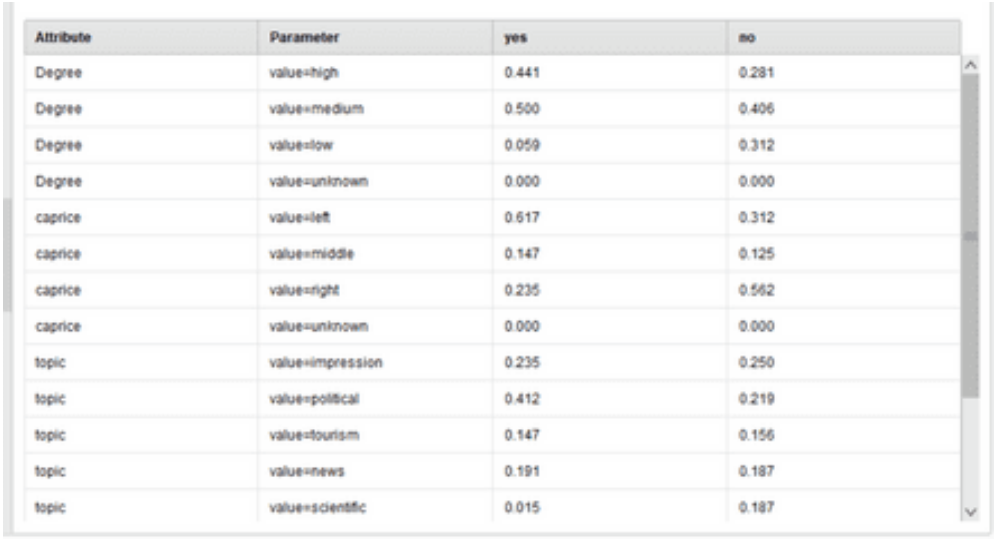
TABEL VI
DATA SIMPLE DISTRIBUTION

	Penelitian ini		Penelitian sebelumnya [1]	
		Distribution		Distribution
Class. Yes	0,680	4	0,680	3
Class. No	0,320	4	0,320	3

Tabel diatas merupakan hasil data simple distribution perbandingan penelitian sebelumnya dan penelitian ini menggunakan metode Naïve Bayes. Data simple distribution dengan metode Naïve Bayes membagi 2 kelas klasifikasi PB yaitu class yes dan class no. Untuk nilai class yes pada penelitian ini yaitu 0.680 dan nilai class no yaitu

JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Journal homepage: https://jurnal.stkippgritulungagung.ac.id/index.php/jipi
ISSN: 2540-8984
Vol. 8, No. 3, September 2023, Pp. 985-994

0.320 dan distribution 4, sedangkan pada penelitian sebelumnya untuk nilai class yes pada penelitian ini yaitu 0.680 dan nilai class no yaitu 0.320 dan distribution 3, dapat dilihat pada Gambar 6.



Attribute	Parameter	yes	no
Degree	value=high	0.441	0.281
Degree	value=medium	0.500	0.406
Degree	value=low	0.059	0.312
Degree	value=unknown	0.000	0.000
caprice	value=left	0.617	0.312
caprice	value=middle	0.147	0.125
caprice	value=right	0.235	0.562
caprice	value=unknown	0.000	0.000
topic	value=impression	0.235	0.250
topic	value=political	0.412	0.219
topic	value=tourism	0.147	0.156
topic	value=news	0.191	0.187
topic	value=scientific	0.015	0.187

Gambar 6. Simple Distribution Naïve Bayes

Gambar diatas merupakan distribution table hasil dari olah data rapid miner. Dari distribution table yang ada dapat diketahui hasil klasifikasi data blogger professional dan blogger musiman.

F. Hasil Akurasi

Proses klasifikasi dengan metode Naïve Bayes yang digunakan untuk mengklasifikasikan data blogger *pada penelitian ini menghasilkan* accuracy, precision, *dan* recall *dapat dilihat pada Table VII :*

TABEL VII
DATA PERBANDINGAN ACCURACY PERFORMANCE

	Penelitian ini		Penelitian sebelumnya [1]	
	Accuracy : 76.27% +/- 7,78%		Accuracy : 75,00% +/- 10,25%	
	True yes	True no	True yes	True no
Pred. Yes	62	18	62	19
Pred. No	6	14	6	13

Hasil pengujian ini menunjukan accuracy performance menggunakan algoritma klasifikasi Naïve Bayes sebesar 76.27% +/- 7,78%. Kemudian hasil *class precision* dan *class recall* untuk prediksi *yes* menunjukkan tingkat *precision* sebesar 77,50% dan untuk prediksi *no* sebesar 70,00%. Sedangkan hasil pengujian pada penelitian sebelumnya menunjukkan akurasi klasifikasi menggunakan metode Naive bayes diperoleh sebesar 75,00% +/- 10,25%, dapat dilihat pada Tabel VIII.

TABEL VIII
DATA PERBANDINGAN CLASS RECALL

	Penelitian ini		Penelitian sebelumnya [1]	
	True yes	True no	True yes	True no
Class Precision	77,50 %	70.00 %	76,54%	68,42%.
Class recall	91.18 %	43.75 %	91.18 %	40,62 %

Tabel diatas menunjukkan hasil perbandingan class precision dan class recall pada penelitian sebelumnya untuk class precision true yes sebesar 76,54% dan untuk true no sebesar 68,42%, sedangkan pada penelitian ini precision true yes sebesar 77,50 % dan untuk true no sebesar 70,00%. Selanjutnya untuk hasil perbandingan class precision dan class recall pada penelitian sebelumnya untuk class recall true yes sebesar 91,18 % dan untuk true no sebesar 40,62%, sedangkan pada penelitian ini Class recall true yes sebesar 91,18 % dan untuk true no sebesar 43,75%.

III. KESIMPULAN

Pada penelitian yang sudah dilakukan oleh peneliti sebelumnya menggunakan algoritma klasifikasi Naïve Bayes dengan cross validation dan number of fold sebanyak 10 kali menghasilkan accuracy performance sebesar 75 %. Pada penelitian validasi yang dilakukan menggunakan confusion matrix dengan algoritma klasification Naïve Bayes dan cross validation dan number of fold sebanyak 10 kali menghasilkan accuracy performance yang lebih tinggi

yakni 76.27% atau meningkat 1,27% dari penelitian yang dilakukan oleh peneliti sebelumnya, sedangkan *class precision* dan *class recall* untuk prediksi *yes* menunjukkan tingkat *precision* sebesar 77,50% dan untuk prediksi *no* sebesar 70,00% dan classification error adalah 23.73% +/-7.78% atau dapat diartikan prosentase errornya kecil. Sedangkan hasil pengujian pada penelitian sebelumnya menunjukkan akurasi klasifikasi menggunakan metode Naive bayes diperoleh sebesar 75,00% +/- 10,25%. Sedangkan class precision dan class recall untuk prediksi yes menunjukkan tingkat precision sebesar 76,54% dan untuk prediksi no sebesar 68,42%.

DAFTAR PUSTAKA

[1] R. A. Anggraini, G. Widagdo, A. S. Budi, and M. Qomaruddin, "Penerapan Data Mining Classification untuk Data Blogger Menggunakan Metode Naïve Bayes," *J. Sist. dan Teknol. Inf.*, vol. 7, no. 1, p. 47, 2019, doi: 10.26418/justin.v7i1.30211.

[2] M. S. Syarah, M. Wati, and N. Puspitasari, "Klasifikasi Penderita ISPA Menggunakan Metode Naive Bayes Classifier," vol. 1, pp. 8–15, 2022.

[3] D. Suryani, A. Yulianti, E. L. Maghfiroh, and J. Alber, "Quality Classification of Palm Oil Products Using Naïve Bayes Method," *Sistemasi*, vol. 11, no. 1, p. 251, 2022, doi: 10.32520/stmsi.v11i1.1713.

[4] F. D. S. Harahap, "Dampak Pandemi Covid-19 Terhadap Masyarakat Khususnya Dunia Ketenagakerjaan," vol. 2019, 2020.

[5] O. Pahlevi and A. Amrin, "Data Mining Model For Designing Diagnostic Applications Inflammatory Liver Disease," *SinkrOn*, vol. 5, no. 1, p. 51, 2020, doi: 10.33395/sinkron.v5i1.10589.

[6] A. Fadilah, M. N. Pangestu, S. Lumbanbatu, and S. Defiyanti, "Pengelompokan Kabupaten/Kota Di Indonesia Berdasarkan Faktor Penyebab Stunting Pada Balita Menggunakan Algoritma K-Means," *JIKO (Jurnal Inform. dan Komputer)*, vol. 6, no. 2, p. 223, 2022, doi: 10.26798/jiko.v6i2.581.

[7] H. F. Putro, R. T. Vulandari, and W. L. Y. Saptomo, "Penerapan Metode Naive Bayes Untuk Klasifikasi Pelanggan," *J. Teknol. Inf. dan Komun.*, vol. 8, no. 2, 2020, doi: 10.30646/tikomsin.v8i2.500.

[8] M. M. Fajar, A. R. Putri, and K. F. H. Holle, "Perbandingan Metode Klasifikasi Data Mining Untuk Deteksi Keaslian Lowongan Pekerjaan di Medsos," *J. Ilm. Inform.*, vol. 7, no. 1, pp. 41–48, 2022, doi: 10.35316/jimi.v7i1.41-48.

[9] L. N. Rani, "Klasifikasi Nasabah Menggunakan Algoritma C4.5 Sebagai Dasar Pemberian Kredit," *INOVTEK Polbeng - Seri Inform.*, vol. 1, no. 2, p. 126, 2016, doi: 10.35314/isi.v1i2.131.

[10] A. L. Maukar, F. Marisa, and A. A. Widodo, "Analisis Data Penerimaan Mahasiswa Baru Berbasis K-Means," *JIKO (Jurnal Inform. dan Komputer)*, vol. 6, no. 2, p. 142, 2022, doi: 10.26798/jiko.v6i2.558.