PERBAIKAN ALGORITMA *NAIVE BAYES CLASSIFIER* MENGGUNAKAN TEKNIK *LAPLACIAN CORRECTION*

Muhammad Rizki[1], Muhammad Arhami*[2], dan Huzeni[3]
[1.2,3] Jurusan Teknologi Informasi dan Komputer Politeknik Negeri Lhokseumawe
Jln. B.Aceh Medan Km.280 Buketrata 24301 INDONESIA
**email : abi_hakan@yahoo.com*

Abstract

Naïve Bayes Classifier is one of the classification algorithms in Data Mining with a good processing speed and a fairly high level of accuracy. In the classification process the Naïve Bayes Classifier adopts the Bayesian theorem to map a data against a class by taking into account the probability of the attribute data, but because the Naïve Bayes Classifier makes probability the basis for its calculations, it is certainly very risk if it is wrong. If one class that is contained in the attribute has a value of 0, this will reduce the level of accuracy of the classification process carried out by the Naïve Bayes Classifier algorithm itself, therefore in this study the Laplacian Correction technique is used as an alternative to fix the problems that are owned by the Naïve Bayes Classifier Algorithm. The result of this research is that the Laplace Correction technique has succeeded in improving the performance of the Naïve Bayes Classifier by fixing the 0 value for each attribute. The level of accuracy that is owned by the Naïve Bayes Classifier after experiencing improvements with the Laplacian correction technique is 94.44%.

Keywords*: Data mining, naïve bayes classifier, laplacian correction.*

PENDAHULUAN

Naïve Bayes Classifier merupakan algoritma klasifikasi sederhana namun sangat efisien. Algoritma ini didasarkan pada formulasi Bayesian dari masalah klasifikasi yang menggunakan asumsi penyederhanaan independensi atribut[1, 2]. Naïve Bayes Classifier melakukan pengklasifikasian dengan cukup baik dibandingkan dengan Algoritma pengklasifikasian lainnya, hal tersebut dibuktikan pada jurnal Xhemali, Daniela, Chris J. Hinde, and Roger G. Stone. *Naive Bayes vs. decision trees vs. neural networks in the classification of training web pages*.[3], berikut persamaan *Naïve Bayes Classifier* yang dapat dilihat pada persamaan 1 dibawah ini.

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)} \quad (1)$$

Dimana :

X : Data dengan kelas yang belum diketahui.
c : Hipotesis dari suatu kelas yang spesifik.
P(c|x) : Probabilitas hipotesis berdasarkan kondisi (*posterior probability*).
P(c) : Probabilitas Hipotesis (*prior probability*)
P(x|c) : Probabilitas berdasarkan kondisi pada hipotesis
P(x) : Probabilitas c

Namun dikarenakan basis perhitungannya bertumpu pada probabilitas tentu akan sangat beresiko apabila nilai yang dikandung oleh *attribute* yang akan dihitung bernilai 0, oleh sebab itu

diperlukan teknik *Laplacian Correction* untuk menangani hal tersebut. Laplace Correction merupakan teknik yang digunakan untuk menyiasati supaya probabilitas pada perhitungan *Naïve Bayes Classifier* tidak menghasilkan nilai 0 dikarenakan tidak adanya data untuk kategori tertentu dalam suatu *Class*[4, 5]. Persamaan dari *Laplacian Correction* dapat dilihat pada persamaan 2 berikut:

$$P(c,t) = \frac{1 + N(ct, D)}{|V| + N(c, D)} \quad (2)$$

Dimana :

P ($_{c,t}$) : probabilitas nilai yang akan di temukan
N(c_t,D) : jumlah Sample pada atribut P
V : jumlah nilai pada atribut
N(c,D) : jumlah sampel

METODE

Data untuk Bahan Penelitian

Data dosen Jurusan Teknologi Informasi dan Komputer

Data dosen jurusan Teknologi Informasi dan Komputer digunakan untuk mencari dan menentukan jumlah responden yang nantinya akan dijadikan sebagai narasumber pengumpulan variable dependent dan variable independent. Dari total 24 dosen dari berbagai bidang keahlian yang ada pada jurusan Teknologi Informasi dan Komputer di Politeknik Negeri Lhokseumawe maka jumlah dosen yang akan dijadikan responden dan narasumber penentu variable independent dan variable dependen dapat dihitung menggunakan *Slovin Formula*.

Slovin Formula atau Rumus Slovin merupakan rumus yang digunakan untuk menentukan jumlah sample atau jumlah data yang dapat dijadikan model pada suatu populasi data. Rumus Slovin digunakan untuk menentukan sample acak dengan memanfaatkan estimasi suatu ukuran atau populasi sampel[6, 7]. Berikut persamaan Slovin yang dapat dilihat pada persamaan 3.

$$n = \frac{N}{1 + Ne^2} \quad (3)$$

Dimana:

n : ukuran sample
N : total populasi
e : toleransi error (dalam satuan persen / atau 1 bagi 100)

Berdasarkan persamaan 3 diatas maka jumlah dosen yang dijadikan sebagai responden dan narasumber penentu variable independent dan variable dependen adalah sebagai berikut.

$$n = \frac{24}{1 + 24 * 0.3^2}$$

$$n = \frac{24}{1 + 24 * 0.09}$$

$$n = \frac{24}{1 + 2.16}$$

$$n = \frac{24}{3{,}16}$$

$$n = 7.59$$

Hasil yang didapatkan dari perhitungan jumlah sample yang dilakukan menggunakan rumus Slovin dengan jumlah populasi sebesar 24 serta tingkat Toleransi Error 0.3 adalah 7.59 yang dibulatkan menjadi 8, maka jumlah dosen yang nantinya akan menjadi narasumber dan responden pengumpulan kuesioner adalah 8 orang dosen.

Data Variabel Bebas

Dari hasil wawancara yang telah dilakukan dengan 8 orang dosen TIK dari berbagai bidang keahlian maka terdapat beberapa topik tugas akhir yang sering diambil oleh mahasiswa jurusan TIK di Politeknik Negeri Lhokseumawe yang dapat dijadikan sebagai variable dependent atau variable hasil klasifikasi. Berdasarkan hasil wawancara maka didapati terdapat 10

topik tugas akhir mahasiswa yang dapat dijadikan variable hasil klasifikasi.

Berikut topik tugas akhir yang sering diambil oleh mahasiswa jurusan TIK yang dapat dilihat pada Tabel 1.

Tabel 1. Topik tugas akhir mahasiswa Jurusan TIK

No	Topik TGA	Frekwensi jawaban
1	Sistem pakar	4
2	Sistem informasi	5
3	Citra	3
4	Gis	3
5	Sistem pendukung keputusan	3
6	Internet of things	4
7	Jaringan	4
8	Augmented reality	3
9	Virtual reality	3
10	Animasi	1

Data Variable Terikat

Variable terikat juga diperoleh dari hasil wawancara dengan yang telah dilakukan dengan 8 orang dosen TIK yang berasal dari berbagai bidang keahlian, berikut data variable terikat yang dapat dilihat pada Tabel 2.

Tabel 2. Data variabel terikat

No	Faktor
1	Jenis Bahasa Pemograman yang di kuasai Dekstop/Web/Mobile/Hardware)
2	Jumlah Bahasa Pemrograman yang di kuasai
3	Penelitian Kuantitatif atau Kualitatif
4	Keadaan Sekitar / Lingkungan
5	Pekerjaan Orang Tua
6	Pengaruh Tempat PKL
7	Hoby
8	IPK
9	Mata Kuliah Tertentu
10	Seberapa sering mahasiswa membaca
11	Nominal IPK

HASIL DAN PEMBAHASAN

Data Testing Yang Digunakan

Berikut data testing yang digunakan untuk melakukan pengujian algoritma *Naïve Bayes*

Classifier yang telah diperbaiki dengan teknik Laplacian Correction seperti ditunjukkan dalam Tabel 3.

Tabel 3. Input data testing tahapan perhitungan manual

No	Variabel independent	Kode	Nilai
1	Jenis bahasa pemrograman yang dikuasai	jns_bp	Pemrograman web
2	Jumlah bahasa pemrograman yang di kuasai	jml_bp	Lebih dari 3 kurang dari 6
3	Jenis penelitian	jns_pen	Kuantitatif
4	Pengaruh lingkungan/pertemanan	ling_per	Tidak berpengaruh
5	Pengaruh pekerjaan orang tua	p_ortu	Tidak berpengaruh
6	Pengaruh hobi	hobi	Tidak berpengaruh
7	Pengaruh ipk	ipk	Tidak berpengaruh
8	Pengaruh matakuliah tertentu	mk	Sangat berpengaruh
9	Pengaruh tempat pkl	tmp_pkl	Tidak berpengaruh
10	Sering membaca	baca	Sangat sering
11	Nominal ipk	jml ipk	3.4

Menghitung Probabilitas Kelas

Probabilitas kelas dihitung dengan cara menggunakan persamaan 4 di bawah ini.

$$P = \frac{Data\ PerKelas}{total\ data} \quad (4)$$

Dengan menggunakan persamaan tersebut maka probabilitas kelas berdasarkan 54 rekord data training dapat dihitung seperti yang dapat dilihat pada Tabel 4.

Tabel 4. Probabilitas Kelas

No	Kelas	Jumlah Data	Probabilitas
1	Animasi	1	0.0185185
2	VR	1	0.0185185
3	AR	1	0.0185185
4	Jaringan	1	0.0185185
5	IOT	4	0.0740740
6	SPK	10	0.1851851
7	GIS	1	0.0185185
8	Citra	1	0.0185185
9	Sistem Informasi	27	0.5
10	Sistem Pakar	7	0.1296296

Menghitung Jumlah Kejadian Perkelas

Jumlah kejadian perkelas merupakan jumlah kejadian dari setiap variable independent yang telah diinputkan pada tahapan input data testing, berikut perhitungan manual hitung jumlah kejadian perkelas yang dapat dilihat pada Tabel 5.

Berdasarkan perhitungan jumlah kejadian yang telah dilakukan di atas, maka didapatkan hasil probabilitas kejadian sementara yang dapat dilihat pada Tabel 6.

Menghitung *Laplacian Correction*

Perhitungan manual Laplacian correction berdasarkan jumlah kejadian perkelas yang sudah dihitung sebelumnya dapat dilihat Tabel 7

Tabel 5. Probabilitas Kejadian

no	variable independen	anm	vr	ar	jrg	iot	spk	gis	ctr	si	sp
1	jns_bp	0	0	0	0	1	10	1	0	25	6
2	jml_bp	0	1	0	1	2	1	1	0	7	1
3	jns_pen	1	0	0	1	3	9	1	1	1	4
4	ling_per	0	0	0	0	1	1	1	0	5	0
5	p_ortu	1	0	0	0	1	1	0	0	4	0
6	hobi	0	0	0	0	0	4	1	0	9	4
7	ipk	1	1	1	1	2	5	1	1	12	5
8	mk	0	0	1	1	2	2	0	1	9	2
9	tmp_pkl	0	0	0	1	0	4	1	1	11	4
10	baca	1	0	0	1	0	2	0	0	5	0
11	jml ipk	0	1	0	0	0	2	0	0	6	0

Tabel 6. Probabilitas Kejadian

no	varibale independen	anm	vr	ar	jrg	iot	spk	gis	ctr	si	sp
1	jns_bp	0	0	0	0	0.25	1	1	0	0.92	0.85
2	jml_bp	0	1	0	1	0.5	0.1	1	0	0.25	0.14
3	jns_pen	1	0	0	1	0.75	0.9	1	1	0.03	0.57
4	ling_per	0	0	0	0	0.25	0.1	1	0	0.18	0
5	p_ortu	1	0	0	0	0.25	0.1	0	0	0.14	0
6	hobi	0	0	0	0	0	0.4	1	0	0.33	0.57
7	ipk	1	1	1	1	0.5	0.5	1	1	0.44	0.71
8	mk	0	0	1	1	0.5	0.2	0	1	0.33	0.28
9	tmp_pkl	0	0	0	1	0	0.4	1	1	0.40	0.57
10	baca	1	0	0	1	0	0.2	0	0	0.18	0
11	jml ipk	0	1	0	0	0	0.2	0	0	0.22	0

Tabel 7. Perbaikan jumlah kejadian dengan Laplacian Correction

No	Varibale Independen	ANM	VR	AR	JRG	IOT	SPK	GIS	CTR	SI	SP
1	JNS_BP	1	1	1	1	1	10	2	1	25	7
2	JML_BP	1	2	1	2	3	1	2	1	7	2
3	JNS_PEN	2	1	1	2	4	9	2	2	1	5
4	LING_PER	1	1	1	1	2	1	2	1	5	1
5	P_ORTU	2	1	1	1	2	1	1	1	4	1
6	HOBI	1	1	1	1	1	4	2	1	9	5
7	IPK	2	2	1	2	3	5	2	2	12	6
8	MK	1	1	2	2	3	2	1	2	9	3
9	TMP_PKL	1	1	1	2	1	4	2	2	11	5
10	BACA	2	1	1	2	1	2	1	1	5	1
11	NML IPK	1	1	1	1	1	2	1	1	6	1

Berdasarkan Tabel 7 maka nilai probabilitas jumlah kejadian baru dapat dihitung sepeti di bawah ini.

$$P(Jns\ BP = WEB \mid Kelas = ANM) = \frac{0+1}{1+10} = 0.090909091$$

$$P(Jml\ BP = Lebih\ dari\ 3\ Kurang\ dari\ 6 \mid Kelas = ANM) = \frac{0+1}{1+10} = 0.090909091$$

$$P(Jns\ Pen = Kuantitatif \mid Kelas = ANM) = \frac{1+1}{1+10} = 0.181818182$$

$$P(Ling\ Per = Tidak\ Berpengaruh \mid Kelas = ANM) = \frac{0+1}{1+10} = 0.090909091$$

$$P(Ortu = Tidak\ Berpengaruh \mid Kelas = ANM) = \frac{1+1}{1+10} = 0.181818182$$

$$P(Hobi = Tidak\ Berpengaruh \mid Kelas = ANM) = \frac{0+1}{1+10} = 0.090909091$$

$$P(IPK = Tidak\ Berpengaruh \mid Kelas = ANM) = \frac{1+1}{1+10} = 0.181818182$$

$$P(mk = Sangat\ Berpengaruh \mid Kelas = ANM) = \frac{0+1}{1+10} = 0.090909091$$

$$P(PKL = Tidak\ Berpengaruh \mid Kelas = ANM) = \frac{0+1}{1+10} = 0.090909091$$

$$P(Membaca = Sangat\ Sering \mid Kelas = ANM) = \frac{1+1}{1+10} = 0.181818182$$

$$P(Nominal\ IPK = 3.4 \mid Kelas = ANM) = \frac{0+1}{1+10} = 0.090909091$$

Setelah melalui proses perbaikan probabilitas yang telah dihitung dengan menggunakan teknik Laplacian correction maka probabilitas kejadian yang baru dapat dihasilkan, berikut hasil probabilitas kejadian yang dihasilkan yang dapat dilihat pada Tabel 8.

Tabel 8. Probabilitas Kejadian yang telah diperbaiki dengan Laplacian correction

No	VARIABEL INDEPENDEN	ANM	VR	AR	JRG	IOT	SPK	GIS	CTR	SI	SP
1	JNS_BP	0.09	0.09	0.09	0.09	0.14	1	0.18	0.09	0.92	0.41
2	JML_BP	0.09	0.18	0.09	0.18	0.21	0.1	0.18	0.09	0.25	0.11
3	JNS_PEN	0.18	0.09	0.09	0.18	0.28	0.9	0.18	0.18	0.03	0.29
4	LING_PER	0.09	0.09	0.09	0.09	0.14	0.1	0.18	0.09	0.18	0.05
5	P_ORTU	0.18	0.09	0.09	0.09	0.14	0.1	0.09	0.09	0.14	0.05
6	HOBI	0.09	0.09	0.09	0.09	0.07	0.4	0.18	0.09	0.33	0.29
7	IPK	0.18	0.18	0.18	0.18	0.21	0.5	0.18	0.18	0.44	0.35
8	MK	0.09	0.09	0.18	0.18	0.21	0.2	0.09	0.18	0.33	0.17
9	TMP_PKL	0.09	0.09	0.09	0.18	0.07	0.4	0.18	0.18	0.40	0.29
10	BACA	0.18	0.09	0.09	0.18	0.07	0.2	0.09	0.09	0.18	0.05
11	Nomnial IPK	0.09	0.18	0.09	0.09	0.07	0.2	0.09	0.09	0,22	0.05

Perhitungan Probabilitas Kondisi

Perhitungan probabilitas kondisi dilakukan untuk menghitung dan menggabungkan seluruh probabilitas kejadian yang telah diperbaiki oleh Laplacian correction sebelumnya. Setelah melakukan perhitungan probabilitas kondisi di atas maka didapati hasil probabilitas kondisi sebagaimana yang ditunjukan pada Tabel 9.

Tabel 9. Hasil Perhitungan Probabilitas Kondisi

KELAS	PROBABILITAS
Animasi	5.6079023917023E-11
Virtual Reality	2.8039511958511E-11
Augmented Reality	1.4019755979256E-11
Jaringan	2.2431609566809E-10
Internet Of Things	2.1335624043598E-10
Sistem Pendukung Keputusan	5.76E-7
Geographic Information System	4.4863219133618E-10
Citra	5.6079023917023E-11
Sistem Informasi	2.0195138846915E-7
Sistem Pakar	2.0195138846915E-7

Tabel 9 merupakan hasil perhitungan probabilitas kondisi, data – data tersebut nantinya akan dijadikan modal untuk mendapatkan hasil akhir dari klasifikasi.

Menghitung Probabilitas Akhir Naïve Bayes Classifier

Perhitungan probabilitas akhir naïve bayes Classifier merupakan proses perhitungan data klasifikasi terakhir sebelum data berhasil diklasifikasikan berikut perhitungan manual probabilitas akhir naïve bayes Classifier.

$$P(ANM) = \frac{5.6079023917023E\_11 * 1}{54} = 1.0385004429078E - 12$$

Berdasarkan hasil perhitungan manual yang probabilitas akhir naïve bayes maka hasil akhir dari klasifikasi dapat dilihat pad Tabel 10.

Tabel 10. Probabilitas Akhir Naive Bayes Classifier .

Kelas	Probabilitas Akhir
Animasi	1.0385004429078E-12
Virtual Reality	5.1925022145391E-13
Augmented Reality	2.5962511072696E-13
Jaringan	4.1540017716313E-12
Internet Of Things	1.5804165958221E-11
Sistem Pendukung Keputusan	**1.0666666666667E-7**
Geographic Information System	8.3080035432626E-12
Citra	1.0385004429078E-12
Sistem Informasi	1.0097569423458E-7
Sistem Pakar	1.1914524065667E-10

Berdasarkan hasil akhir dari probabilitas naïve bayes Classifier maka algoritma naïve bayes Classifier mengklasifikasikan data testing yang telah diinputkan sebelumnya dengan kelas **Sistem Pendukung Keputusan**, hal tersebut dikarenakan kelas Sistem Pendukung Keputusan memiliki nilai probabilitas tertinggi dibanding kelas lainnya yaitu **1.0666666666667E-7**.

Perhitungan Tingkat Akurasi

Tingkat Akurasi Algoritma dihitung dengan cara melakukan klasifikasi pada data training yang sudah ada dan mencocokan kelas yang diklasifikasikan oleh sistem dengan kelas yang dimiliki oleh data training itu sendiri, setiap kelas yang benar akan bernilai 1, sehingga total akurasi akan dihitung dengan cara menggunakan persamaan 5.

$$akurasi = \left(\frac{100}{totalData}\right) * hasil\ KlasifikasiBenar \quad (5)$$

Persamaan tersebut akan menghasilkan tingkat akurasi yang dikonversikan ke dalam bentuk persen (%). Hasil akhir dari proses pengujian tingkat akurasi algoritma adalah sebesar 94,44%.

KESIMPULAN

1. Algoritma *Naïve Bayes Classifier* terbukti memiliki tingkat akurasi sebesar 94.44% setelah mengalami perbaikan dengan menggunakan teknik *Laplacian Correction*.
2. Hasil penerapan Teknik *Laplacian Correction* pada Algoritma *Naïve Bayes Classifier* terbukti dapat memperbaiki dan membangkitkan nilai probabilitas akhir yang pada awalnya bernilai 0.

DAFTAR PUSTAKA

[1] Muktamar, A., Setiawan, A., and Adji, B., 2015. *Pembobotan Korelasi Pada Naive Bayes Classifier, Stmik Amikom Yogyakarta,* in *Seminar Nasional Teknologi Informasi dan Multimedia*.

[2] Kantarcıoglu, M., Vaidya, J., and Clifton, C., 2003. *Privacy Preserving Naive Bayes Classifier for Horizontally Partitioned Data,* in *IEEE ICDM workshop on privacy preserving data mining,* 3-9.

[3] Xhemali, D., J HINDE, C., and G STONE, R., 2009. *Naïve Bayes Vs. Decision Trees Vs. Neural Networks in the Classification of Training Web Pages,* D. XHEMALI, CJ HINDE and Roger G. STONE," Naive Bayes vs. Decision Trees vs. Neural Networks in the Classification of Training Web Pages", International Journal of Computer Science Issues, IJCSI, Volume 4, Issue 1, 16-23.

[4] Syahputra, I.K., Bachtiar, F.A., and Wicaksono, S.A., 2018. *Implementasi Data Mining Untuk Prediksi Mahasiswa Pengambil Mata Kuliah Dengan Algoritme Naive Bayes,* Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer e-ISSN, Vol. 2548, 964X.

[5] Mussa, H.Y., Mitchell, J.B., and Glen, R.C., 2013. *Full "Laplacianised" Posterior Naive Bayesian Algorithm,* Journal of cheminformatics, Vol. 5, No. 1, 1-6.

[6] Nyambura, M.M. and Simon, K., 2019. *Effect of Safety Awareness Campaigns on Employee Performance in Power Transmission Companies in Kenya,* International Journal of Business Management and Finance, Vol. 2, No. 1.

[7] Salim, E., Puspa, D.F., and Darmayanti, Y., 2014. *Faktor-Faktor Yang Mempengaruhi Penggunaan Fasilitas E-Filling Oleh Wajib Pajak Sebagai Sarana Penyampaian Spt Masa Secara Online Dan Realtime (Studi Empiris Pada Wajib Pajak Badan Di Kpp Madya Jakarta Pusat),* Jurnal Fakultas Ekonomi, Vol. 4, No. 1.