STRING (Satuan Tulisan Riset dan Inovasi Teknologi) p-ISSN: 2527 - 9661
Vol. 5 No. 2 Desember 2020 e-ISSN: 2549 - 2837

KLASIFIKASI PENELITIAN DOSEN MENGGUNAKAN NAÏVE BAYES CLASSIFIER DAN ALGORITMA GENETIKA

Muhammad Yusuf Bakhtiar
Program Studi Teknik Informatika, Universitas Indraprasta PGRI Jakarta
bakhtiar.yusuf.by@gmail.com

Submitted August 3, 2020; Revised November 11, 2020; Accepted November 14, 2020

Abstrak

Penelitian merupakan kewajiban bagi dosen guna mengembangkan ilmunya selain mengajar. Sampai saat ini semua penelitian dosen dikelola oleh Lembaga Penelitian dan Pengabdian kepada Masyarakat yang sering disebut LPPM. Lembaga penelitian perguruan tinggi merupakan tempat semua informasi penelitian dosen bisa didapat, masalah yang ditemui di lembaga penelitian perguruan tinggi yaitu proses pengelompokan pada bidang keilmuan penelitian dosen yang dilakukan oleh bagian data dan sistem informasi LPPM, ini juga merupakan salah satu masalah yang berkaitan dengan *text classification*. Klasifikasi merupakan proses penemuan model yang menggambarkan dan membedakan kelas atau konsep yang bertujuan agar bisa digunakan untuk memprediksi kelas dari objek yang label kelasnya belum diketahui. *Naive Bayes* merupakan teknik prediksi berbasis probabilistik sederhana yang berdasarkan pada penerapan teorema *Bayes* dengan asumsi independensi yang kuat. Pada metode ini terdapat kekurangan yang dapat berpengaruh pada akurasi yang disebabkan oleh fitur *Naive Bayes* yang tidak selalu dapat diterapkan. Untuk menangani masalah tersebut, peneliti melakukan proses seleksi fitur menggunakan Algoritma Genetika. *Dataset* yang digunakan yaitu data penelitian dosen sebanyak 275 dari semua bidang keilmuan. Hasil dari eksperimen pada penelitian ini menunjukan bahwa nilai akurasi meningkat sebesar 26,06 % dengan digunakannya Algoritma Genetika pada proses seleksi fitur.

Kata Kunci : klasifikasi, naïve bayes, algoritma genetika, seleksi fitur

Abstract

Research is an obligation for lecturers to develop their knowledge besides teaching. Until now all lecturer's research is managed by the Institute for Research and Community Service, often called LPPM. College research institute is a place where all lecturer research information can be obtained, problems encountered at tertiary research institutions are the grouping process in the field of lecturer research carried out by the LPPM data and information system section, this is also one of the problems related to text classification . Classification is the process of finding a model that describes and distinguishes classes or concepts that aim to be used to predict classes from objects whose class labels are unknown. Naive Bayes is a simple probabilistic based prediction technique based on the application of the Bayes theorem with strong independence assumptions. In this method there are deficiencies that can affect the accuracy caused by the Naive Bayes feature which is not always applicable. To deal with these problems, researchers conducted a feature selection process using Genetic Algorithms. The dataset used is 275 lecturer research data from all scientific fields. The results of experiments in this study indicate that the accuracy value increased by 26.06% with the use of Genetic Algorithms in the feature selection process.

Keywords: *classification, naïve bayes, genetic algorithm, feature selection*

1. PENDAHULUAN

Lembaga penelitian perguruan tinggi merupakan tempat semua informasi penelitian dosen bisa didapat, dimasa yang akan datang semua perguruan tinggi memungkinkan diarahkan pada perguruan tinggi berbasis riset, Lembaga Penelitian dan Pengabdian Kepada Masyarakat Institut Pertanian Bogor (LPPM IPB) menjadi salah satunya yang mempunyai tujuan tersebut. Untuk itu lembaga penelitian harus menjadi tempat sebagai

mendorong, memfasilitasi, meningkatkan dan mengembangkan kerjasama kemitraan dan jaringan kerjasama PPM baik internal maupun eksternal (Nasional - Internasional) secara efektif, efisien dan terbuka.

Dalam mendukung tujuan lembaga penelitian perguruan tinggi , maka dalam memfasilitasi data dan informasi agar bisa dengan mudah digunakan ataupun dimanfaatkan harus adanya manajemen data yang baik. Salah satu dalam pengelolaan data penelitian adanya pengelompokan bidang keilmuan di LPPM IPB ada 5 bidang keilmuan yang telah disetujui oleh rektor dan sampai sekarang masih diterapkan. Proses pengelompokan bidang keilmuan di LPPM IPB dilakukan oleh tim fasilitator yang beranggotakan 9 dosen dari 9 fakultas yang ada di IPB, masalah yang dihadapi dalam proses pengelompokan bidang keilmuan sering ditemukan ketidak sesuaian atau terkait antara judul penelitian dosen yang dibahas dengan bidang keilmuan, ini menyebabkan proses pembuatan laporan akhir penelitian terlambat dan informasi yang dihasilkan tidak tepat.

Untuk mengatasi masalah dalam pengelompokan bidang keilmuan pada penelitian ini digunakan klasifikasi *text mining* menggunakan metode *Naive Bayes Classifier*. *Naive Bayes* yang sifat independesi antar atribut tidak selalu dapat diterapkan karea ada keterkaitan antara antribut [1] dan tidak sempurnanya performa *Naive Bayes classifier* disebabkan karena sifat independent tersebut [2]. Optimasi seleksi fitur diperlukan agar meningkatkan kinerja *Naïve Bayes* [3].
Agar hasil pengelompokan bidang keilmuan lebih akurat Metode *Naive Bayes* (NB) pada seleksi fitur akan dilakukan dengan Algoritma Genetika (GA), ini untuk mengatasi kelemahan metode *Naïve Bayes*.

Penelitian sebelumnya membandingkan metode Naive Bayes dan SVM pada sebuah text judul skripsi menggunakan teknik seleksi fitur dengan *N-gram* dan *Term Frecuancy* dan menghasilkan bahwa Navie Bayes mempunyai tingkat akurasi yang lebih baik dari SVM. Hal tersebut kemungkinan disebabkan algoritma SVM yang digunakan dalam penelitian ini adalah algoritma SVM dengan linear karnel [4]. Algoritma Genetika menggunakan kombinasi linear dari probabilitas posteriori dan tingkat kesalahan *Bayes* sebagai fungsi *fitness* untuk optimasi memperoleh hasil yang lebih baik [5] dan penggunaan Algoritma Genetika untuk meminimalkan kesalahan rata-rata kuadrat dari jalur lintasan pesawat menghasilkan jalur yang akurat [6].

Dengan adanya penelitian ini diharapkan akan mengatasi masalah klasifikasi penelitian dosen sehingga hasil dari pengelompokan bidang keilmuan akan lebih tepat.

2. METODE PENELITIAN

Metode yang digunakan adalah metode optimasi penentuan seleksi fitur pada model Naïve Bayes menggunakan algoritma Genetika.
Kaitan antara *Naive Bayes* dengan klasifikasi, korelasi hipotesis, dan bukti dengan klasifikasi adalah bahwa hipotesis dalam teorema *Bayes* merupakan lebel kelas yang menjadi target pemetaan dalam klasifikasi, sedangkan bukti merupakan fitur-fitur yang menjadi masukan dalam model klasifikasi. Jika *X* dalam vektor masukan yang berisi fitur dan *Y* adalah lebel klas, *Naive Bayes* dituliskan dengan *P(Y|X)*. Notasi tersebut berarti probabilitas lebel kelas *Y* didapatkan setelah fitur-fitur *X* diamati. Notasi ini disebut juga probalitas akhir (*posterior probability*) untuk *Y*, sedangkan *P(Y)* disebut probabilitas awal *(proir probability) Y* [7]. Perhitungan Probabilitas kategori :

$$P(C_i) = \frac{fd(C_i)}{|D|} \quad (1)$$

keterangan :

$P(C_i)$ = probabilitas dokumen kategori C_i

$fd(C_i)$ = jumlah document yang dimiliki kategori C_i

$|D|$ = jumlah seluruh training dokumen

Perhitungan jumlah kasus/kelas:

$$P(W_k|C_i) = \frac{n_k+1}{n+|vocabulary|} \quad (2)$$

Keterangan :

D = dokumen

W_k = kata

Nk = nilai kemunculan C_i

N = jumlah keseluruhan kata pada kategori C_i

$|vocabulary|$ = jumlah keseluruhan kata

Algoritma Genetika melakukan "*searching*" dengan melakukan simulasi proses evolusi makhluk hidup. Prinsip utamanya yaitu meniru proses seleksi alam dan prinsip-prinsip ilmu genetika. Dalam seleksi alam, individu-individu bersaing untuk mempertahankan hidup dan melakukan reproduksi. Individu-individu yang lebih "*fit*" akan mempunyai peluang untuk terus bertahan hidup (*survive*) dan melakukan reproduksi (menghasilkan keturunan). Sebaliknya, individu-individu yang kurang "*fit*" akan mati dan punah (prinsip ini dinamakan juga "*survival of the fittnest*"). Selanjutnya, dalam proses seleksi alam ini, beberapa individu baru yang lebih "*fit*" dari kedua orang tuanya akan "dilahirkan" melalui proses yang disebut penyilangan (*crossover*) dan mutasi. Kedua proses ini terjadi pada kromosom-kromosom individu yang melakukan reproduksi. Proses seleksi dan reproduksi (penyilangan dan mutasi) ini berlangsung berulang kali sampai individu yang paling "*fit*" dihasilkan [8]. Tahapan yang dilakukan sebagai berikut:

1. Pengumpulan Data : Data yang didapat dari dataset yang telah digunakan para peneliti lain (*data public*). Dataset yang digunakan adalah abstrak dari judul penelitian dosen yang terdiri dari enam bidang keilmuan yaitu: 1) Pangan; 2) Sumberdaya alam dan lingkungan; 3) Kesehatan; 4) Sosial, ekonomi dan budaya; 5) Tekonologi dan rekayasa; sebanyak 275 data. Dari model klasifikasi yang dihasilkan, kemudian diuji dengan 275 data yang diambil random, dimana masing-masing bidang akan diambil data secara *random* untuk dijadikan data uji.

Tabel 1. Dataset

No.	Class Label	Data
1.	Pangan	54
2.	Sumberdaya alam dan lingkungan	62
3.	Kesehatan	52
4.	Sosial, ekonomi dan budaya	53
5.	Tekonologi dan rekayasa	54
Jumlah		275

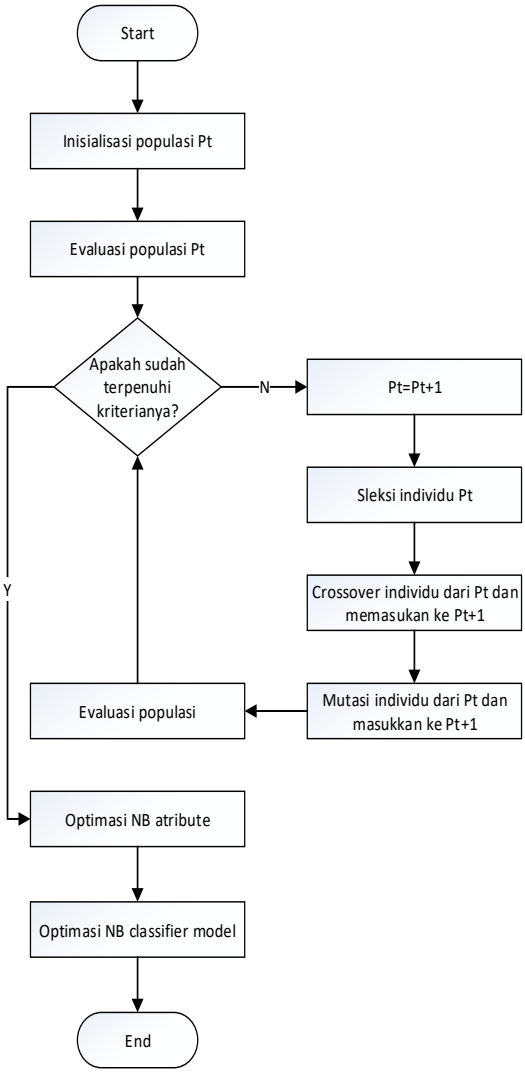
Pada Tabel 1. dataset jurnal penelitian dihubungkan dengan operator cross validation yang didalamnya terdapat proses *cross validation* yang digunakan dalam penelitian ini adalah 10-fold validation. Dataset yang berisi 275 data dengan ribuan atribut akan dipecah menjadi 10 bagian. Dimana setiap bagian akan dibentuk secara random. Prinsip 10-fold validation adalah 1:9, 1 bagian menjadi data testing, data lainnya menjadi data training. Demikian sehingga 10 bagian tersebut berkesempatan menjadi data testing. Setelah dilakukan training dan testing maka dapat diukur akurasinya.

Data yang didapat berupa judul dan abstrak penelitian dosen yang termasuk *dataset public* yang diambil atau *download* langsung pada website *https://www.lppm.ipb.ac.id/hasil-penelitian/*.

2. Pengolahan Data Awal : Pengolahan data awal meliputi penyaringan data kedalam bentuk

yang dibutuhkan, serta pengelompokan dan penentuan atribut data dengan mengambil data judul dan abstrak untuk sempel data dari tahun 2010 sampai 2018.

3. Metode yang diusulkan : Metode yang digunakan dalam penelitian ini adalah Algoritma Genetika untuk optimasi pemilihan fitur pada *Naïve Bayes*. Proses yang dilakukan dalam tahap modeling untuk menyelesaikan klasifikasi abstrak penelitian dosen menggunakan Naïve Bayes yang dioptimasi dengan Algoritma Genetika.



Gambar. 1 Diagram Alur Algoritma Genetika dan Naïve Bayes

Pada Gambar 1. menunjukkan bahwa langkah pertama dari Algoritma Genetika adalah inisialisasi populasi pt. Kromosom pada populasi ditentukan nilai gennya. Langkah selanjutnya evaluasi populasi pt. Kromosom diseleksi menggunakan nilai *fitness*. Kromosom yang memiliki nilai *fitness* terbesar akan dipilih. Nilai kriteria akan terpenuhi jika jumlah generasi sudah maksimal, jika belum maka iterasi akan terus berjalan. Kemudian berdasarkan *probability crossover*, kromosom terpilih akan dicrossoverkan. Dan berdasarkan *probability mutation* ditentukan berapa banyak gen dalam kromosom yang akan dimutasi. Setelah mencapai generasi maksimal, maka didapat kromosom dengan nilai *fitness* tertinggi sebagai solusi dari permasalahan seleksi atribut. Kemudian data dengan atribut yang terpilih akan ditraining dan ditesting oleh Naive Bayes, sehingga naive bayes dapat melakaukan klasifikasi terhadap judul penelitan dosen.

4. *Experiment* dan Pengujian Model : Untuk pengujian model dilakukan dengan menggunakan *Rapid Miner*. Dari model yang telah dibuat maka dataset akan diolah sehingga menghasilkan model yang diharapkan.
5. Evaluasi dan Validasi Hasil : Setelah dilakukan *experiment* pada semua dataset dengan model yang diusulkan, maka akan dievaluasi dan ditarik kesimpulan dari hasil *experiment*. Metode ini merepresentasikan hasil evaluasi medel dengan menggunakan tabel matriks, jika dataset terdiri dari dua kelas, kelas pertama dianggap positif, dan kelas kedua dianggap negative [9]. Evaluasi menggunakan confusion matrix menghasilkan akurasi, persesi, recall. Precision atau confidence merupakan proporsi kasus yang diprediksi positif yang juga positif benar pada data sebenarnya. *recall*

atau *sensitivity* merupakan proporsi kasus positif yang sebenarnya yang diprediksi positif secara benar [10].

Tabel 2. *Confusion Matrix*

Correct Classification	*Classified as*	
	+	-
+	*True positive*	*False negatives*
-	*False positives*	*True negatives*

True positive (tp) merupakan jumlah *record positive* dalam *dataset* yang diklasifikasi *positive. True negative* (tn) merupakan jumlah *record negative* dalam dataset yang diklasifikasi *negative. False positive* (fp) merupakan jumlah *record negative* dalam dataset yang diklasifikasikan *positive. False negative* (fn) merupakan jumlah *record positive* dalam dataset yang diklasifikasi *negative.*

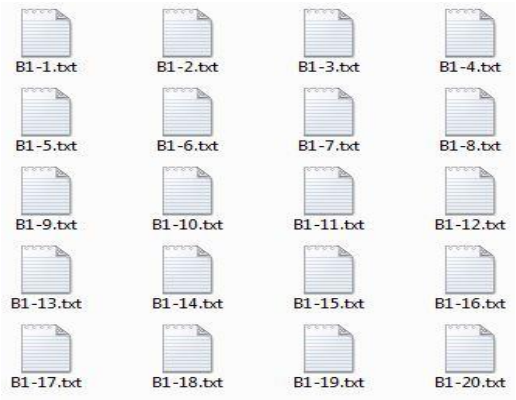
3. HASIL DAN PEMBAHASAN

Data yang digunakan adalah data *text* yang didapat dari situs Lembaga Penelitian dan Pengabdian kepada Masyarakat Institut Pertanian Bogor (LPPM-IPB). Data *text* ini merupakan kumpulan abstrak pada judul penelitian pada berbagai bidang ilmu. Pada penelitian ini jumlah judul penelitian yang digunakan 275 data latih dan data uji. Daftar judul penelitian untuk data latih dan data uji yang akan digunakan pada penelitian ini dapat dilihat pada Lampiran. Data yang telah diunduh dari situs LPPM-IPB akan diambil isi abstraknya dan diuraikan menggunakan perangkat lunak Rapid Miner.

Pengolahan Data Awal

Sebelum data set dimasukan ke dalam model yang diusulkan, terlebih dahulu dilakukan *preprocessing* data. Pada tahapan ini, dilakukan beberapa hal, yaitu konversi data text menjadi *.txt, tokenized, transform cases, filter stopword*.

a. *text converting* : agar teks yang dimasukan kedalam model dapat dibaca oleh model, dilakukan konversi tipe format teks. Tahapan ini adalah dilakukan sebuah konversi format *teks* yang diambil dari sumbernya menjadi format *.txt*.



Gambar 2. File txt

b. *Tokenized* : *Tokenized* merupakan proses untuk memisah-misahkan kata. Hasil dari pemisahan tersebut dinamakan *token*.
c. *Transform Cases* : Proses ini dilakukan untuk mengubah bentuk kata-kata. Pada proses ini, karakter dijadikan huruf kecil (*lower case*) semua.
d. *Filter Stopword* : Proses ini menghilangkan kata-kata yang sering muncul tapi tidak memiliki pengaruh apapun dalam ekstrksi klasifikasi teks. Pada proses ini, kata yang termasuk dalam penunjukan waktu, kata tanya, dan kata sambung.

Seleksi fitur Algoritma Genetika

Ada beberapa tahapan yang ada pada algoritma genetika dalam mengoptimasi seleksi fitur pada Naive Bayes. Pada tahap pertama adalah pengkodean yaitu proses kodefikasi atas solusi dari permasalahnnya. Hasil dari pengkodean berbentuk string yang merupakan representasi dari suatu kromosom. Tahap kedua *selection* yaitu menentukan kromosom mana yang tetap tinggal pada generasi berikutnya. Tahap ketiga *crossover* yang akan menghasilkan kromosom baru yang akan menggantikan kromosom lama. Tahap keempat mutasi yang memungkinkan terjadinya kromosom baru secara *unpredictable.* Proses terakhir

adalah *decoding* yaitu mengambil hasil kromosom terbaik untuk memberikan nilai yang optimal.

a. Inisialisasi Populasi

Inisialisasikan populasi dengan membangkitkan bilangan biner sebanyak jumlah data latih, sehingga pada masalah ini menggunakan pengkodean dalam bilangan biner.

Tabel. 3. Pembangkitan Populasi Awal

	W_1	W_2	W_3	W_4	dst	W_{210}
K_1	1	1	0	0	…	0
K_2	0	0	1	1	…	0
↓	↓	↓	↓	↓	↓	↓
K_{50}	0	0	0	0	…	1

Ukuran populasi atau *popsize* yang telah ditentukan dalam proses manualisasi ini adalah 5 dari 275 individu. Pada proses manualisasi ini kromosom terdiri 3126 gen yang didapat dari representasi jumlah kata pada bidang ilmu. Pada Tabel Pembangkitan Populasi Awal terlihat bahwa kode Wn merupakan banyaknya kata yang terdapat dari tiap kata pada bidang ilmu B1, Kn merupakan jumlah kromosom atau tiap dokumen.

b. Evaluasi Populasi

Proses selanjutnya yaitu evaluasi populasi dengan ditentukannya parameter Algoritma Genetika sebagai berikut:

- Probabilitas Penyilangan (Crossover) (Pc)=1,00
- Probabilitas Mutasi (Pm)= 0,5

Evaluasi populasi yang terbentuk dengan menghitung nilai *fitness* dengan menggunakan fungsi objektif yang dibentuk dalam studi kasus. Dimana nilai bobot ditentukan sebagai berikut:

Tabel 4. Nilai Bobot pada setiap Gen

K	Gen	Bobot
1	anjing	0,003
2	Antera	0,003
3	arsen	0,003
4	biomursa	0,003
5	adas	0,003
↓	↓	↓
275	udang	0,003

Sehingga nilai *fitness* kromosom ditentukan sebagai berikut:

Tabel 5. Nilai Fitness pada setiap Kromosom

Kromosom	Bi					Nilai *fitness*
	Anjing	Antera	Arsen	Biomursa	Adas	
	0.003	0.003	0.003	0.003	0.003	
K1	1	1	0	0	0	
Bi*Si	0.003	0.003	0	0	0	0.006
K2	0	0	1	1	0	
Bi*Si	0	0	0.003	0.003	0	0.006
K3	0	0	0	0	1	
Bi*Si	0	0	0	0	0.003	0.003
K4	0	0	1	2	0	
Bi*Si	0	0	0.003	0.006	0	0.009
↓	↓	↓	↓	↓	↓	↓
K50	0	1	0	0	0	
Bi*Si	0	0.003	0	0	0	0.003

Keterangan :
Bi = bobot pada geni
Si = nilai gen pada kromosom i di populasi awal

c. Penyilangan pada Kromosom

Model penyilangan yang dipakai pada penelitian ini adalah metode penyilangan satu titik (*one point crossover*). Pertama pilih bangun bilangan acak sebanyak jumlah kromosom dalam populasi. Memilih bilangan acak lagi mulai dari 0 sampai dengan panjang kromosom / gen-1 disebut posisi *cut-point crossover* dimana posisi itu akan menentukan posisi gen yang akan disilangkan antara kromosom induk yang telah dipilih sebelumnya.

Tabel 6. Kromosom dikawin-silang

Kromosom		
K1	><	K3
K3	><	K5
K5	><	K1

Tabel 7. Point Crossover

Kromosom Orangtua 1	Kromosom Orangtua 2	Keturunan
11001011	1101**1111**	**11001111**

d. Mutasi Algoritma Genetika

Proses mutasi akan melakukan pergantian 1 gen acak dengan nilai baru sehingga tidak selalu menjamin bahwa setelah proses mutasi akan diperoleh kromosom dengan *fitness* yang lebih baik. Proses yang dilakukan adalah menginversi nilai bit pada posisi tertentu yang terpilih secara acak (atau menggunakan skema tertentu) pada kromosom, yang disebut *inversi* bit.

Tabel. 8 Mutasi pada Pengkodean Biner

Kromosom sebelum mutase	Kromosom setelah mutase
10010 1 11	10010 **0** 11

Proses mutasi dengan pengkodean bilangan biner dilakukan dengan memilih gen secara acak, kemudian apabila gen tersebut bernilai 0 maka akan diganti 1, dan sebagainya.

e. Evaluasi Populasi Baru

Setelah melewati proses mutasi, telah dihasilkan populasi baru yang disebut generasi 1, pada penelitian ini akan ditentukan jumlah generasi maksimum atau keturunan yaitu 250, maka proses algoritma genetika akan berhenti ketika sudah mencapai 250 generasi. Misal sudah didapat hasil akhir pada proses evaluasi populasi baru hanya saja sudah dipilih yang mempunyai nilai fitness tertinggi yang mempunyai nilai fitness dibawah rata-rata akan dihilangkan.

Tabel 9. Kombinasi Pengujian GA

Kombinasi	Pc (Probabilitas Crossover)	Pm (Probabilitas Mutasi)
1	0.1	0.5
2	0.3	0.5
3	0.8	1
4	1	1
5	0.6	0.1
6	1	0.5

Pada parameter GA terdapat nilai probabilitas crossover (Pc) dan probabilitas mutasi (Pc) yang akan digunakan pada sistem. Nilai Pc yang akan digunakan adalah 0.6 dan nilai Pm yang adalah digunakan yaitu 0.1

Tabel 10. Performansi dari pengaruh Pc dan Pm

No	Pc	Pm	Rata-rata Performansi					Jumlah rata-rata
			Pangan	SDM	Kesehatan	Sosekbud	Teknologi	
1	0.1	0.5	48.15	85.48	88.46	54.72	44.44	64.67
2	0.3	0.5	48.15	85.48	88.46	47.17	42.59	62.88
3	0.8	0.5	55.56	88.71	96.15	50.94	48.15	68.23
4	1	1	96.30	85.48	94.23	32.08	42.59	70.58
5	0.6	0.1	96.30	85.84	92.31	60.36	57.41	78.61
6	1	0.5	96.30	85.48	96.15	58.49	44.44	76.42

Pada Tabel 10, diperoleh perhitungan rata-rata performansi dari kombinasi pengujian GA menunjukan kombinasi antara Pc dan Pm untuk menghasilkan populasi yang terbaik ditunjukan dengan nilai rata-rata terbesar pada kombinasi nomer lima. Kombinasi ini akan digunakan untuk proses GA dalam untuk seleksi fitur Naïve Bayes.

1. **Model Naïve Bayes**

Kata kunci yang didapat dari proses seleksi fitur menggunakan algoritma genetika akan diklasifikasi menggunakan model Naïve Bayes, berikut proses yang dilakukan:

a. Dataset

Dataset berupa dokumen exel yang berisi nama bidang ilmu sebagai label seperti :

B1, B2, B3, B4, B5 dan kata sebagai kasus per kelas dan berjumlah 275 data.

Tabel 11. Dataset yang digunakan

Dataset	Jumlah record	Jumlah kelas
Penelitian dosen	275	5

Setelah ditentukan dataset yang akan digunakan selanjutnya dataset tersebut akan dihitung jumlah kelas/lebelnya.

Tabel 12. Data Penelitian LPPM IPB

Judul Penelitian	Kata Kunci	Bidang Ilmu (B)
Deteksi Bakteri Patogen dan Fermentatif dari Pangan menggunakan Real-time Polymerase Chain Reaction (PCR)	bakteri patogen, Cronobacter sakazakii, Staphylococcus aureus, Lactobacillus plantarum, real-time PCR	B1
Karakteristik Biofisik Agroforestry Di Hutan Rakyat	agroforestry, karakteristik biofisik, hutan rakyat, sengon (F. moluccana) porang (A. onchophyllus)	B2
Studi Lintas Konversi Asam Lemak Menjadi Senyawa Amina Dengan Teknik Termodegradasi	asam lemak, amina, termodegradasi, oleokimia, fatty amine, fatty nitril, sintesis satu tahap	B3
IbM Petani Biofarmaka Desa Sukaluyu Kec. Nanggung Kab. Bogor	biofarmaka, petani, bogor, kec nanggung	B4
Rekonstruksi Alat Tangkap Garuk untuk Meningkatkan Hasil Tangkapan Kerang	alat tangkap garuk, kerang	B5

b. Hitung Jumlah Kelas/Label

Kelas yang dimaksud adalah bidang ilmu (B) yang telah ditentukan pada tiap judul penelitian. Jumlah kelas pada bidang ilmu ada 5, yaitu : B1, B2, B3, B4, B5

Perhitungan probabilitas kategori

$$P(B1) = \frac{54}{275} = 0{,}2$$
$$P(B2) = \frac{62}{275} = 0{,}2$$
$$P(B3) = \frac{52}{275} = 0{,}2$$
$$P(B4) = \frac{53}{275} = 0{,}2$$
$$P(B5) = \frac{54}{275} = 0{,}2$$

c. Hitung Jumlah Kasus/Kelas

Pada dataset yang digunkan terdapat kata pada B1, B2, B3, B4, dan B5, jadi jumlah semua kata ada 1076 kata.

d. Bandingkan Hasil/Kelas

Dari hasil probabilitas pada class B1 sampai dengan B5 didapat kesimpulan bahwa D6 termasuk pada class B1 karena mempunyai nilai probabilitas yang paling besar.

2. **Evaluasi dan Validasi**

Pengujian model pada penelitiian ini menggunakan teknik *10-flod cross validation*, dimana setiap *record* digunakan beberapa kali dalam jumlah yang sama untuk *training* dan tepat satu kali untuk *testing*. Selama proses, salah satu dari partisi dipilih untuk *training*, sedangkan sisanya untuk *testing*. Pengulangan dilakukan 10 kali sehingga setiap partisi digunakan untuk *testing* tepat satu kali.

Pada tahap evaluasi dan validasi hasil evaluasi, dilakukan pengukuran dengan menggunakan *confusion matrix*, hasil akurasi akan ditampilkan dengan menggunakan metode *Naive Bayes Classifier*.

accuracy: 52.55% +/- 26.40% (micro average: 52.73%)

	true pangan	true sdm	true kesehatan	true sosekbud	true teknologi	class precision
pred. pangan	28	3	5	0	0	77.78%
pred. sdm	1	34	0	0	0	97.14%
pred. kesehatan	25	25	47	0	0	48.45%
pred. sosekbud	0	0	0	16	34	32.00%
pred. teknologi	0	0	0	37	20	35.09%
class recall	51.85%	54.84%	90.38%	30.19%	37.04%	

Gambar 3. Confusion Matrix Naïve Bayes

Pada Gambar 3, menunjukan hasil dari klasifikasi menggunakan Naïve Bayes dengan akurasi 52,55 %, angka ini termasuk dalam akurasi yang rendah. Salah satu penyebabnya yaitu isi abstrak yang berupa *text* sangat bermacam-macam atau unik dalam pengembangan kalimat yang digunakan, ini bertujuan untuk menghindari isi dari jurnal penelitian terdeteksi plagiat.

accuracy: 78.61% +/- 17.81% (micro average: 78.55%)

	true pangan	true sdm	true kesehatan	true sosekbud	true teknologi	class precision
pred. pangan	52	3	4	0	0	88.14%
pred. sdm	1	53	0	0	0	98.15%
pred. kesehatan	1	6	48	0	0	87.27%
pred. sosekbud	0	0	0	32	23	58.18%
pred. teknologi	0	0	0	21	31	59.62%
class recall	96.30%	85.48%	92.31%	60.38%	57.41%	

Gambar 4. Confusion Matrix GA dan NB

Hasil dari pengujian pertama klasifikasi Naïve Bayes didapat akurasi sebesar 52,55%. Pada pengujian kedua algoritma Naive Bayes akan digabungkan dengan Algoritma Genetika yang berfungsi untuk mengoptimasi proses seleksi fitur sehingga meningkatkan akurasi pada Naive Bayes. Hasil pengujian kedua yang menggabungkan antara Naive Bayes dengan algoritma genetika diliah pada Gambar 4, menunjukan akurasi 78,61% ketika seleksi fitur pada Naïve Bayes dioptimasi dengan Algoritma Genetika meningkatkan akurasi sebesar 26,06%.

4. SIMPULAN

Penelitian dengan menggunakan algoritma genetika untuk mengoptimalkan seleksi fitur pada algoritma Naive Bayes sehingga akurasinya meningkat. Hasil penelitian mendapatkan kenaikan 26.06 % menjadi 78.61 % pada nilai akurasinya.

Dari hasil penelitian ini bisa disimpulkan bahwa algoritma genetika mampu menaikan akurasi pada Naive Bayes dengan mengoptimalkan proses seleksi fitur berupa kata yang ada pada abstrak tiap penelitian.

Namun algoritma genetika membutuhkan menentukan fitur yang optimal dikarnakan jumlah kata, sehingga proses klasifikasi berjalan lambat. Untuk penelitian kedepannya disarankan antara lain, 1) mengatasi masalah waktu yang lama pada waktu yang lama dalam saat proses klasifikasi, 2) mencoba algoritma lain untuk melakukan optimasi sehingga hasil yang dicapai bisa lebih baik dari penelitian-penelitian yang ada.

DAFTAR PUSTAKA

[1] Socrates, I. G. A., Akbar, A. L., Akbar, M. S., Arifin, A. Z., & Herumurti, D. (2016). Optimasi Naive Bayes Dengan Pemilihan Fitur Dan Pembobotan Gain Ratio. *Lontar Komputer : Jurnal Ilmiah Teknologi Informasi*, *7*(1), 22. https://doi.org/10.24843/LKJITI.2016.v07.i01.p03

[2] Li, J., Ding, L., & Li, B. (2014). A novel naive bayes classification algorithm based on particle swarm optimization. *Open Automation and Control Systems Journal*, *6*(1), 747–753.https://doi.org/10.2174/1874444301406010747

[3] Muhamad, H., Prasojo, C. A., Sugianto, N. A., Surtiningsih, L., & Cholissodin, I. (2017). Optimasi Naïve Bayes Classifier Dengan Menggunakan Particle Swarm Optimization Pada Data Iris. *Jurnal Teknologi Informasi Dan Ilmu Komputer*, *4*(3), 180. https://doi.org/10.25126/jtiik.201743251

[4] Fathan Hidayatullah, A., Rifqi Ma, M., & Program Studi Manajemen Informatika STMIK Jenderal Achmad Yani Yogyakarta Jl Ringroad Barat, arif. (2016). Penerapan Text Mining dalam Klasifikasi Judul Skripsi. *Seminar Nasional Aplikasi Teknologi Informasi (SNATi) Agustus*, 1907–5022.

[5] Rammal, A., Perrin, E., Vrabie, V., Assaf, R., & Fenniri, H. (2017). Selection of discriminant mid-infrared wavenumbers by combining a naïve Bayesian classifier and a genetic algorithm: Application to the evaluation of lignocellulosic biomass biodegradation. *Mathematical Biosciences*, *289*, 153–161. https://doi.org/10.1016/j.mbs.2017.05.002

[6] Gupta, D. K., Vasudev, K. L., & Bhattacharyya, S. K. (2018). Genetic algorithm optimization based nonlinear ship maneuvering control. *Applied Ocean Research*, *74*, 142–153. https://doi.org/10.1016/j.apor.2018.03.001

[7] Prasetyo, E. (2012). *Data Mining : Konsep dan Aplikasi menggunakan MATLAB*. ANDI.

[8] Yandra Arkeman, Yeni Herdiyeni, Irman Hermadi, G. F. L. (2014). *Algoritma Genetika Tujuan Jamak (Multi-Objective Genetic Algorithms): Teori dan Aplikasinya untuk bisnis dan Agrondustri*. IPB Press.

[9] Bramer, M. (2007). Principles of Data Mining. In *Principles of Data Mining*. https://doi.org/10.1007/978-1-84628-766-4

[10] Han, J., Kamber, M., & Pei, J. (2012). Introduction. In *Data Mining* (3th ed., pp. 1–38). Elsevier. https://doi.org/10.1016/b978-0-12-381479-1.00001-0