THE EFFECT OF NAIVE BAYES CLASSIFIER USING DUMMY VARIABLE AND FEATURE BACKWARD SELECTION WITH PEARSON CORRELATION IN DIAGNOSING GYNECOLOGY

Febrianti
Program Studi Informatika
Fakultas MIPA
Universitas Sebelas Maret
febrianti.frumos@gmail.com

Ristu Saptono
Program Studi Informatika
Fakultas MIPA
Universitas Sebelas Maret
ristu.saptono@staff.uns.ac.id

Rini Anggrainingsih
Program Studi Informatika
Fakultas MIPA
Universitas Sebelas Maret
rini.anggrainingsih@staff.uns.ac.id

ABSTRACT

The use of pearson correlation and backward selection as feature selection can be combined to improve the accuracy of the Naive Bayes Classifier. Feature selection is done as preprocessing of data on the process Classfier Naive Bayes algorithm. Pearson Correlation as a preprocessing data will work to sort the parameters that influence the classification process of the gynecological disease and Backward Selection will select those parameters in sequence on the preprocessing data in the Naive Bayes Classfier process. Previously, the parameter would be converted to dummy variables because the parameter has the possibility of a value that appears more than two (non-binary). This study discusses the effect of Naive Bayes Classfier using dummy variables and feature backward selection with pearson correlation in the diagnosis of gynecological disease. The results obtained in this study prove that the use of dummy variables increases the accuracy value from 88% to 88.8% and the use of pearson correlation as preprocessing data increases the accuracy value of Naive Bayes Classifier from 88.8% with 24 parameters to 89.6% with 20 parameters. The use of pearson correlation not only improves accuracy but also increases the effectiveness of the features used in the Naive Bayes Classifier process. This can be seen from the increase of accuracy results with the use of the number of parameters that decreased.

Keywords : *Naive Bayes Classifier, Dummy Variable, Pearson Correlation*

1. PENDAHULUAN

Naive Bayes Classifier merupakan salah satu algoritma paling sederhana untuk klasifikasi data[1]. Meskipun sederhana, algoritma ini mampu memberikan hasil perhitungan yang baik. Kesederhanaannya dalam segi konseptual, implementasi, dan komputasi membuat algoritma ini banyak dipelajari. Namun ada hal yang harus diperhatikan dalam penggunaan *Naive Bayes Classifier*, saat variabel yang digunakan tidak independen hasil akurasi yang didapat akan menurun[2]. Sementara sebagian besar variabel memiliki setidaknya korelasi dengan variabel lainnya. Oleh sebab itu perlu dilakukan seleksi fitur pada variabelnya terlebih dahulu untuk mengatasi hal tersebut.

Feature Selection merupakan suatu pendekatan yang digunakan untuk menyeleksi variabel-variabel yang tidak berpengaruh atau kurang berpengaruh terhadap suatu kasus klasifikasi sehingga dapat menemukan variabel-variabel minimal yang dapat memberikan hasil klasifikasi terbaik[3]. *Feature selection* terdiri dari metode *filter, wrapper,* dan *embedded*[4]. Pendekatan *filter* bekerja dengan baik pada pola pengklasifikasi yang dilakukan dalam pengenalan pola, terlepas dari spesifik dan parameternya. Pemilihan atribut dapat dilakukan berdasarkan pada beberapa algoritma maupun ukuran kualitas terlepas dari apapun prediktornya. Dalam pendekatan *wrapper* dikondisikan dengan kinerja variabel itu sendiri dan karakteristiknya. Biasanya, akurasi prediktif dianggap paling penting dalam menentukan faktor. Ketergantungan pada beberapa variabel dapat berarti hilangnya generalitas dan bias, namun pada saat yang sama dapat menutup penyesuaian terhadap requirement yang ada[4].

Pendekatan *filter* dan *wrapper* dapat dikombinasikan sebagai aturan dalam proses klasifikasi[4]. Fitur-fitur ini akan dilakukan dalam dua konteks: pertama dengan mengurutkan ranking menggunakan *Pearson Correlation*, kedua dengan mengamati proses *Naive Bayes Classifier* dalam *Backward Selection* saat mengikuti ranking tersebut. Namun karena variabel yang digunakan merupakan variabel ordinal maka variabel tersebut harus diubah menjadi *dummy variable* terlebih dahulu. Dummy variable adalah variabel yang dibuat untuk mewakili atribut dengan dua kategori atau kategori yang berbeda. *Dummy variable* mengubah nilai non *binary* menjadi *binary* yang terdiri 0 dan 1[5].

Pada penelitian yang dilakukan oleh Nugraha[6] disarankan untuk menggunakan *Dummy Variable* pada Naive Bayes Clasiffier agar dapat meningkatkan hasil akurasi klasifikasi penyakit kandungan. Sedangkan pada penelitian Shofieyuddin[7] terbukti bahwa penggunaan *dummy* dapat meningkatkan hasil akurasi klasifikasi penyakit kandungan.

Berdasarkan penjelasan di atas, pada penelitian ini akan dianalisis pengaruh penerapan *Dummy Variable* dan *Feature Backward Selection* dengan *Pearson Corellation* pada klasifikasi penyakit kandungan dengan menggunakan data set dari penelitian Prabawaningrum[8].

2. LANDASAN TEORI

2.1 Naive Bayes Classifier

Naive Bayes adalah algoritma pembelajaran sederhana yang menggunakan Teorema Bayes, bersama dengan asumsi bahawa atribut yang diberikan pada kelas bersifat independen[9]. Teorema Naive Bayes dapat dinyatakan dalam persamaan berikut:

$$P(X_k|Y) = \frac{P(Y|X_k)}{\Sigma_i P(Y|X_i)} \quad (1)$$

Dimana, keadan posterior (probabilitas Xk di dalam Y) dapat dihitung dari keadaan prior (probabilitas Y di dalam Xk dibagi dengan jumlah dari semua probabilitas di dalam semua Xi)

2.2 *Feature Selection*

Istilah *feature selection* digunakan dalam *machine learning* untuk proses pemilihan subset dari fitur (dimensi) yang digunakan dalam mewakili data[10]. Pada umumnya *feature selection* dikelompokkan menjadi tiga pendekatan, yaitu *filter, wrapper,* dan *embedded*[4]. Pendekatan filter memanfaatkan karakteristik intrinsik dalam data untuk peringkat fitur. Pendekatan *filter* yang pada umumnya digunakan adalah korelasi Pearson[11]. Pengujian korelasi *Pearson* digunakan untuk mengetahui kuat tidaknya hubungan antara variabel X dan Y[12]. Cara menghitung korelasi *Pearson* dapat dilihat pada persamaan berikut:

$$r = \frac{n(\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{\{n(\Sigma x^2) - (\Sigma x)^2\}\{n(\Sigma y^2) - (\Sigma y)^2\}}} \quad (2)$$

dimana,
r = koefisien korelasi *Pearson* $(-1 \leq r \leq 1)$,
x = variabel bebas
y = variabel terikat
n = jumlah sampel

Pendekatan *wrapper* digunakan untuk menghitung bobot atribut dengan menggunakan model klasifikasi untuk mengukur kinerja atribut[13].

Salah satu pendekatan *wrapper* yang biasa digunakan adalah *backward elimination*[14]. Proses pada *backward elimination*, dimulai dengan menghimpun semua variabel dan secara progresif menghilangkan yang paling sedikit menjanjikan. Proses ini diulangi sampai tidak ada variabel yang dapat dihapus tanpa kehilangan nilai penurunan yang signifikan.

2.3 *Dummy Variable*

Dummy variable merupakan pengkodean ulang dari *categorical variables* yang mempunyai lebih dari dua kategori yang diubah menjadi beberapa variabel biner[7]. Contoh : Status Pernikahan, jika data asli dilabeli dengan 1 = Menikah, 2 = Belum Menikah, 3 = Cerai/Janda/Duda/Berpisah, dapat diubah menjadi dua variable sebagai berikut : var_1 : 1 = Belum Menikah, 0 = Lain, var_2 : 1 = Cerai/Janda/ Duda/Berpisah. Untuk kasus diatas jika ada seorang yang sudah menikah, maka kedua var_1 dan var_2 akan memiliki nilai 0. Umumnya, *categorical variable* dengan kategori (k) akan dikodekan mennjadi (k -1) untuk *dummy variable*. Contoh penggunaan *dummy variable* dapat dilihat pada gambar Gambar 1.

2.4 *Confusion Matrix*

Confusion Matrix adalah matriks yang digunakan untuk menganalisis seberapa baik classifier mengenali data kelas yang berbeda[15]. Tabel tentang *Confusion Matrix* dapat dilihat pada tabel 1.

Tabel 1. Tabel *Confusion Matrix*

Actual Class		*Assigned Class*	
		Positive	*Negative*
	Positive	*True Positive*	*False Negative*
	Negative	*False Positive*	*True Negative*

Dimana, *True Positive* dan *True Negative* adalah keadaan pada saat hasil *outcome* sesuai dengan kondisi sebenarnya yang terjadi. *False Positive* dan *False Negative* adalah keadaan dimana hasil *outcome* tidak sesuai dengan kondisi yang sebenarnya terjadi.

Nilai akurasi *classifier* pada set test yang diberikan dapat dihitung persentase dari test set data yang diklasifikasikan dengan benar oleh *classifier*. Pengukuran presisi dan *recall* juga banyak digunakan dalam klasifikasi. Presisi dapat dianggap sebagai ukuran ketepatan untuk mencocokkan informasi jawaban dengan permintaan, sedangkan *recall* sebagai ukuran kelengkapan untuk mengukur persentase kasus positif yang diidentifikasi dengan benar. Akurasi, presisi, dan *recall* dapat dihitung menggunakan rumus:

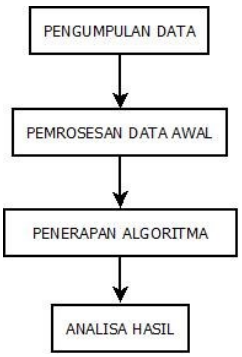
$$\text{Akurasi} = \frac{True\ positive + True\ Negative}{True\ Positive + True\ Negative + False\ Positive + False\ Negative}$$

$$Recall = \frac{True\ positive}{True\ Positive + False\ Negative}$$

$$\text{Presisi} = \frac{True\ positive}{True\ Positive + False\ Positive}$$

3. METODOLOGY PENELITIAN

Pelaksanaan penelitian ini dibagi menjadi beberapa bagian yaitu pengumpulan data, transformasi data, penerapan algoritma dan analisa hasil. Gambar 2 menunjukkan tahapan penelitian.



Gambar 2 Tahapan Metodologi penelitian

3.1 Pengumpulan Data

Pengumpulan data dilakukan dengan mengambil data pada penelitian sebelumnya, yaitu penelitian oleh Prabhawaningrum[9] yang merupakan data rekam medik pasien RSUD Dr. Moewardi Solo. Data yang digunakan pada penelitian ini adalah 125 data dengan 18 gejala penyakit yang dikategorikan menjadi 5 kelas penyakit. Rincian data kelas penyakit dapat dilihat pada tabel 2 dan data gejala penyakit dapat dilihat pada tabel 3.

Tabel 2. Data kelas penyakit

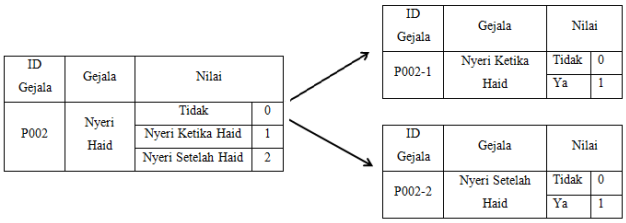
No.	Kelas	Kategori Penyakit
1.	C001	Tidak Termasuk 4 Jenis Penyakit Kandungan
2.	C002	Radang Panggul
3.	C003	Mioma Uteri
4.	C004	Kanker Serviks
5.	C005	Kanker Ovarium

Tabel 3. Data gejala penyakit

No.	ID Gejala	Gejala	Nilai
1.	P001	Anemia	Tidak
			Ya
2.	P002	Nyeri Haid	Tidak
			Nyeri Setelah Haid
			Nyeri Ketika Haid
3.	P003	Susah Hamil	Tidak
			Ya
4.	P004	Benjolan di Perut	Tidak
			Ya
5.	P005	Pendarahan	Tidak
			Pendarahan Menstruasi Abnormal
			Pendarahan Tiba-tiba
6.	P006	Nyeri Berhubungan Seksual	Tidak
			Ya
7.	P007	Cepat Lelah	Tidak
			Ya
8.	P008	Penurunan Berat Badan	Tidak
			Ya
9.	P009	Nyeri Panggul	Tidak
			Ya
10.	P010	Gangguan Pencernaan	Tidak
			Sembelit
			Diare
11.	P011	Nyeri Perut	Tidak
			Nyeri pada Rongga Perut
			Nyeri pada Bagian Bawah
			Nyeri pada Bagian Bawah dan Pinggul
12.	P012	Nyeri Punggung	Tidak
			Ya
13.	P013	Penurunan Nafsu Makan	Tidak
			Ya
14.	P014	Demam	Tidak
			Ya
15.	P015	Sakit Kepala	Tidak
			Ya
16.	P016	Kembung	Tidak
			Ya
17.	P017	Keputihan	Tidak
			Ya
18.	P018	Gangguan BAK	Tidak
			Sering BAK
			Nyeri BAK
			Nyeri dan Sering BAK

3.2 Transformasi Data

Transformasi data akan mengubah data asli menjadi data dummy. Proses transformasi dari data asli menjadi *Dummy Variable* dapat dilihat pada gambar 3.

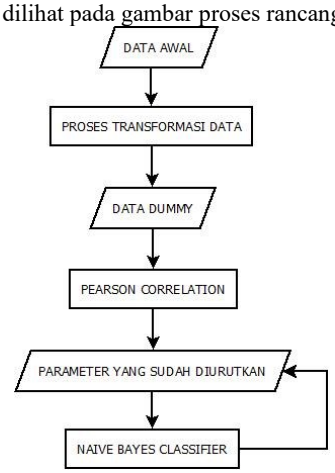


Gambar 3. Proses transformasi data

3.3 Penerapan Algoritma

Tahap yang akan dilakukan dimulai dengan *Feature Selection* dilanjutkan dengan penerapan *Naive Bayes Classifier. Feature Selection* dilakukan terhadap data yang telah ditransformasi. *Filter* dilakukan dengan melakukan *Pearson Corellation* untuk mengurutkan ranking dari parameter memiliki nilai *Pearson Corellation* paling besar ke parameter yang memiliki nilai *Pearson Corellation* paling kecil.

Proses selanjutnya akan menghitung nilai *Naive Bayes Classifier* dengan semua parameter. Parameter yang memiliki nilai *Pearson Corellation* terendah akan dieliminasi, kemudian dihitung kembali dengan *Naive Bayes Classifier*. Proses ini akan diulangi sampai semua parameter dihilangkan Secara sederhana penerapan algoritma dapat dilihat pada gambar proses rancangan penelitian.



Gambar 3. Proses Rancangan Penelitian

3.4 Analisa Hasil

Analisa hasil pengujian merupakan hasil evaluasi metode yang telah dilakukan sebelumnya.

4. Hasil dan Pembahasan

4.1 Deskripsi Data

Data penelitian ini merupakan data pada penelitian sebelumnya[]. Data pada penelitian ini terdiri dari 18 gejala penyakit dengan 5 kelas penyakit. Data gejala penyakit dapat dilihat pada tabel 4 dan data kelas penyakit dapat dilihat pada tabel 4.

Tabel 4. Data gejala penyakit

No.	Gejala	Deskripsi	Nilai	
1.	Anemia	Keadaan saat jumlah sel darah merah berada di bawah normal karena pendarahan.	Tidak	0
			Ya	1
2.	Nyeri Haid	Nyeri yang dirasakan pada bagian perut bawah pada saat masa menstruasi.	Tidak	0
			Nyeri Setelah Haid	1
			Nyeri Ketika Haid	2
3.	Susah Hamil	Keadaan dimana wanita susah mengalami kehamilan dikarenakan gangguan transportasi sperma untuk pembuahan sel telur.	Tidak	0
			Ya	1
4.	Benjolan di Perut	Munculnya benjolan di bagian perut atas atau bawah.	Tidak	0
			Ya	1
5.	Pendarahan	Pendarahan yang terjadi pada uterus wanita.	Tidak	0
			Pendarahan Menstruasi Abnormal	1
			Pendarahan Tiba-tiba	2
6.	Nyeri Hubungan Seksual	Nyeri yang terjadi saat berhubungan seksual karena adanya penekanan tumor pada daerah panggul.	Tidak	0
			Ya	1
7.	Cepat Lelah	Kondisi dimana kondisi badan cepat mengalami kelelahan dalam melakukan aktifitas sehari-hari.	Tidak	0
			Ya	1
8.	Penurunan Berat Badan	Menurunnya berat badan karena nafsu makan menurun.	Tidak	0
			Ya	1
9.	Nyeri Panggul	Nyeri yang disebakan tumor yang membesar pada rongga pelvik yang menekan saraf.	Tidak	0
			Ya	1
10.	Gangguan Pencernaan	Gangguan pencernaan yang biasanya meliputi diare ataupun sembelit.	Tidak	0
			Sembelit	1
			Diare	2
11.	Nyeri Perut	Nyeri yang terjadi di area perut.	Tidak	0
			Nyeri pada Rongga Perut	1
			Nyeri pada Bagian Bawah	2
			Nyeri pada Bagian Bawah dan Pinggul	3
12.	Nyeri Punggung	Nyeri yang disebakan tumor yang membesar pada rongga pelvik yang menekan saraf sampai bagian punggung.	Tidak	0
			Ya	1
13.	Penurunan Nafsu Makan	Penurunan nafsu makan akibat perut terasa penuh, mual, dan kembung.	Tidak	0
			Ya	1
14.	Demam	Keadaan dimana temperatur diatas 38.3 derajat celcius.	Tidak	0
			Ya	1
15.	Sakit Kepala	Kondisi dimana kepala mengalami nyeri.	Tidak	0
			Ya	1
16.	Kembung	Kondisi dimana perut terasa penuh dan kencang.	Tidak	0
			Ya	1
17.	Keputihan	Keluarnya cairan bukan darah yang keluar melalui organ genital wanita.	Tidak	0
			Ya	1
18.	Gangguan BAK	Gangguan buang air kecil karena penekanan terhadap kandung kemih oleh tumor.	Tidak	0
			Sering BAK	1
			Nyeri BAK	2
			Nyeri dan Sering BAK	3

Tabel 5. Data kelas penyakit

No.	Kelas	Kategori Penyakit	Deskripsi
1	C1	Tidak mempunyai penyakit kandungan	Pasien tidak menderita salah satu dari 4 penyakit kandungan
2	C2	Radang Panggul	Penyakit infeksi traktus genital atas yang merupakan salah satu komplikasi dari infeksi menular seksual
3	C3	Mioma Uteri	Tumor jinak otot polos yang terdiri dari sel-sel jaringan otot polos, jaringan fibroid dan kolagen
4	C4	Kanker Serviks	Merupakan kelainan yang terjadi pada sel-sel serviks yang berkembang dengan cepat dan tidak terkontrol
5	C5	Kanker Ovarium	Kanker yang bermula pada indung telur (ovarium) wanita

4.2 Transformasi Data

Data asli dari gejala penyakit ditransformasikan menjadi data *dummy*. Pada penelitian ini ada 5 data asli yang perlu ditransformasikan menjadi data *dummy*. Transformasi data dapat dilihat pada table 6.

Tabel 6. Transformasi data asli menjadi *Dummy Variable*

Data Awal	Data Dummy	Nilai
Nyeri Haid	Nyeri Setelah Haid	Tidak
		Ya
	Nyeri Ketika Haid	Tidak
		Ya
Pendarahan	Pendarahan Menstruasi Abnormal	Tidak
		Ya
	Pendarahan Tiba-tiba	Tidak
		Ya
Gangguan Pencernaan	Sembelit	Tidak
		Ya
	Diare	Tidak
		Ya
Nyeri Perut	Nyeri pada Rongga Perut	Tidak
		Ya
	Nyeri pada Bagian Bawah	Tidak
		Ya
	Nyeri pada Bagian Bawah dan Pinggul	Tidak
		Ya
Gangguan BAK	Sering BAK	Tidak
		Ya
	Nyeri BAK	Tidak
		Ya

4.3 Penerapan Algoritma

4.3.1 *Preprocessing Data*

Preprocessing data akan mengurutkan peringkat parameter dari yang memiliki korelasi terbesar hingga terkecil pada data asli dan *dummy variable* dengan melakukan uji korelasi pearson. Hasil *preprocessing data* untuk data asli dapat dilihat pada tabel 7 sedangkan hasil *preprocessing data* pada *dummy variable* dapat dilihat pada tabel 8.

Tabel 7. Hasil *Preprocessing Data* pada data asli

Peringkat	Id Gejala	Parameter	Nilai
1	P05	Pendarahan	0,81
2	P01	Anemia	0,68
3	P11	Nyeri Perut	0,63
4	P18	Gangguan BAK	0,45
5	P09	Nyeri Panggul	0,45
6	P13	Penurunan Nafsu Makan	0,41
7	P08	Penurunan Berat Badan	0,40
8	P15	Sakit Kepala	0,40
9	P16	Kembung	0,40
10	P07	Cepat Lelah	0,39
11	P04	Benjolan di Perut	0,31
12	P02	Nyeri Haid	0,13
13	P12	Nyeri Punggung	0,12
14	P03	Susah Hamil	0,1
15	P10	Gangguan Pencernaan	0,09
16	P17	Keputihan	0,08
17	P14	Demam	0,07
18	P06	Nyeri Berhubungan Seksual	0,05

Tabel 8. Hasil Preprocessing Data pada Dummy Variable

Peringkat	Id Gejala	Parameter	Nilai
1	P05-2	Pendarahan Tiba-tiba	0,78
2	P01	Anemia	0,67
6	P01	Anemia	0,67
3	P09	Nyeri Panggul	0,47
4	P13	Penurunan Nafsu Makan	0,4
5	P08	Penurunan Berat Badan	0,39
7	P15	Sakit Kepala	0,38
8	P07	Cepat Lelah	0,37
9	P16	Kembung	0,35
10	P11-3	Nyeri pada Bagian Bawah dan Pinggul	0,29
11	P04	Benjolan di Perut	0,29
12	P05-1	Pendarahan Menstruasi Abnormal	0,26
13	P11-2	Nyeri pada Bagian Bawah	0,24
14	P10-2	Diare	0,24
15	P18-1	Sering BAK	0,18
16	P10-1	Sembelit	0,14
17	P02-2	Nyeri Ketika Haid	0,14
18	P12	Nyeri Punggung	0,13
19	P03	Susah Hamil	0,11
20	P17	Keputihan	0,1
21	P02-1	Nyeri Setelah Haid	0,1
22	P14	Demam	0,05
23	P06	Nyeri Berhubungan Seksual	0,01
24	P11-1	Nyeri pada Rongga Perut	0

4.3.2 Penerapan *Naive Bayes Classifier* dengan Backward Elimination

Perhitungan pengujian *Naive Bayes Classifier* pada data asli dan *dummy variable* akan dihitung dengan semua parameter yang kemudian dieliminasi satu persatu parameternya sesuai urutan peringkat dari hasil *preprocessing data.* Hasil pengujian *Naive Bayes Classifier* untuk data asli dapat dilihat pada tabel 9 sedangkan hasil pengujian *Naive Bayes Classifier* pada *dummy variable* dapat dilihat pada tabel 10.

Dimana pada tabel 9 dan tabel 10 dapat dilihat bahwa:

- NBC xx merupakan hasil dari NBC dengan xx parameter,
- $NBC_{xx-1} = NBC_{xx}|P_{yy} \ = NBC_{xx} - P_{yy}$ atau NBC xx yang mengeliminasi Parameter yy).

Tabel 9. Hasil Pengujian *Naive Bayes Classifier* pada data asli

Parameter	Presisi	*Recall*	Akurasi
NBC 18: P05, P01, P11, P18, P09, P13, P08, P15, P16, P07, P04, P02, P12, P03, P10, P17, P1, P06	88,08%	88,02%	88%
NBC 17 = NBC18|P06: P05, P01, P11, P18, P09, P13, P08, P15, P16, P07, P04, P02, P12, P03, P10, P17, P1	88,02%	87,99%	88%
NBC 16 = NBC17|P01: P05, P01, P11, P18, P09, P13, P08, P15, P16, P07, P04, P02, P12, P03, P10, P17	83,99%	84,17%	84%
NBC 15 = NBC16|P17: P05, P01, P11, P18, P09, P13, P08, P15, P16, P07, P04, P02, P12, P03, P10	80,75%	81,07%	80.8%
NBC 14 = NBC15|P10: P05, P01, P11, P18, P09, P13, P08, P15, P16, P07, P04, P02, P12, P03	80,75%	81,07%	80.8%
NBC 13 = NBC14|P03: P05, P01, P11, P18, P09, P13, P08, P15, P16, P07, P04, P02, P12	77,47%	77,93%	77.6%
NBC 12 = NBC13|P12: P05, P01, P11, P18, P09, P13, P08, P15, P16, P07, P04, P02	79,18%	78,44%	78.4%
NBC 11 = NBC12|P02: P05, P01, P11, P18, P09, P13, P08, P15, P16, P07, P04	79,18%	78,44%	78.4%
NBC 10 = NBC11|P04: P05, P01, P11, P18, P09, P13, P08, P15, P16, P07	64,80%	64,67%	64.8%
NBC 9 = NBC 10|P07: P05, P01, P11, P18, P09, P13, P08, P15, P16	65,98%	65,53%	65.6%
NBC 8 = NBC 9|P16: P05, P01, P11, P18, P09, P13, P08, P15	63,22%	63,10%	63.2%
NBC 7 = NBC 8|P15: P05, P01, P11, P18, P09, P13, P08	52,40%	57,59%	57.6%
NBC 6 = NBC 7|P08: P05, P01, P11, P18, P09, P13	49,96%	50,93%	51.2%
NBC 5 = NBC 6|P13: P05, P01, P11, P18, P09	40,76%	44,80%	45.6%
NBC 4 = NBC 5|P09: P05, P01, P11, P18	16,66%	39,23%	40.8%
NBC 3 = NBC 4|P18: P05, P01, P11	16,66%	39,23%	40.8%
NBC 2 = NBC 3|P11: P05, P01	16,66%	39,23%	40.8%
NBC 1 = NBC 2|P01: P05	4%	20%	20%

Tabel 10. Hasil Pengujian *Naive Bayes Classifier* data dummy

Parameter	Presisi	*Recall*	Akurasi
NBC 24: P05-2, P01, P09, P13, P08, P18-2,P15, P07, P16,P11-3, P04, P05-1, P11-2, P10-2, P18-2, P10-1, P02-2, P12, P03, P17, P02-1, P14, P06, P11-1	89,21%	88,92%	88,8%
NBC23=NBC24|P11-1: P05-2, P01, P09, P13, P08, P18-2,P15, P07, P16,P11-3, P04, P05-1, P11-2, P10-2, P18-2, P10-1, P02-2, P12, P03, P17, P02-1, P14, P06	89,21%	88,92%	88,8%
NBC 22 = NBC23|P06: P05-2, P01, P09, P13, P08, P18-2,P15, P07, P16,P11-3, P04, P05-1, P11-2, P10-2, P18-2, P10-1, P02-2, P12, P03, P17, P02-1, P14	88,41%	88,08%	88%
NBC 21 = NBC22|P14: P05-2, P01, P09, P13, P08, P18-2,P15, P07, P16,P11-3, P04, P05-1, P11-2, P10-2, P18-2, P10-1, P02-2, P12, P03, P17, P02-1	89,21%	88,92%	88,8%
NBC 20=NBC21|P021: P05-2, P01, P09, P13, P08, P18-2,P15, P07, P16,P11-3, P04, P05-1, P11-2, P10-2, P18-2, P10-1, P02-2, P12, P03, P17	89,82%	89,75%	89,6%
NBC19 =NBC20|P17: P05-2, P01, P09, P13, P08, P18-2,P15, P07, P16,P11-3, P04, P05-1, P11-2, P10-2, P18-2, P10-1, P02-2, P12, P03	86,30%	85,84%	85,6%
NBC 18 = NBC19|P03: P05-2, P01, P09, P13, P08, P18-2,P15, P07, P16,P11-3, P04, P05-1, P11-2, P10-2, P18-2, P10-1, P02-2, P12	86,30%	85,84%	85,6%
NBC 17 = NBC18|P12: P05-2, P01, P09, P13, P08, P18-2,P15, P07, P16,P11-3, P04, P05-1, P11-2, P10-2, P18-2, P10-1, P02-2,	84,69%	84,3%	84%
NBC16=NBC17|P02-2: P05-2, P01, P09, P13, P08, P18-2,P15, P07, P16,P11-3, P04, P05-1, P11-2, P10-2, P18-2, P10-1	86,05%	85,84%	85,6%
NBC 15=NBC16|P101: P05-2, P01, P09, P13, P08, P18-2,P15, P07, P16,P11-3, P04, P05-1, P11-2, P10-2, P18-2	85,09%	85,01%	84,8%
NBC 14=NBC15|P182: P05-2, P01, P09, P13, P08, P18-2,P15, P07, P16,P11-3, P04, P05-1, P11-2, P10-2	85,88%	85,71%	85,6%
NBC13=NBC14|P10-2: P05-2, P01, P09, P13, P08, P18-2,P15, P07, P16,P11-3, P04, P05-1, P11-2	85,88%	85,71%	85,6%
NBC12=NBC13|P11-2: P05-2, P01, P09, P13, P08, P18-2,P15, P07, P16,P11-3, P04, P05-1	86,64%	86,57%	86,4%
NBC11=NBC12|P05-1: P05-2, P01, P09, P13, P08, P18-2,P15, P07, P16,P11-3, P04	76,38%	75,93%	76%
NBC 10 = NBC11|P04 : P05-2, P01, P09, P13, P08, P18-2, P15, P07, P16,P11-3	76,06%	75,81%	76%
NBC 9=NBC10|P11-3: P05-2, P01, P09, P13, P08, P18-2, P15, P07, P16	69,50%	68,50%	68,8%
NBC 8 = NBC 9|P16 : P05-2, P01, P09, P13, P08, P18-2, P15, P07	72,05%	71,06%	71,2%
NBC 7 = NBC 8|P07 : P05-2, P01, P09, P13, P08, P18-2, P15	74,73%	72,73%	72,8%
NBC 6 = NBC 7|P15: P05-2, P01, P09, P13, P08, P18-2	68,08%	66,35%	66,4%
NBC 5 = NBC 6|P18-2: P05-2, P01, P09, P13, P08	71,48%	69,49%	69,6%
NBC 4 = NBC 5|P08 : P05-2, P01, P09, P13	66,46%	64,23%	64,8%
NBC 3 = NBC 4|P13 : P05-2, P01, P09	58,72%	53,49%	53,6%
NBC 2 = NBC 3|P09 : P05-2, P01	49,46%	46,56%	46,4%
NBC 1 = NBC 2|P01 : P05-2	21,19%	38,11%	37,6%

4.4 Analisa Hasil

Pada percobaan yang telah dibuat untuk data asli dapat memberikan kesimpulan bahwa N*aive Bayes Classifier* bekerja efektif pada 18 parameter dengan hasil akurasi 88%, Sedangkan untuk data dummy*, Naive Bayes Classifier* akan bekerja lebih

efektif bekerja pada 20 parameter dengan akurasi 89,6% bila dibandingkan dengan 24 parameter yang mempunyai hasil akurasi 88,8%,

Selain itu, penggunaan *dummy*, terbukti dapat meningkatkan hasil akurasi pada nilai *Naive Bayes Classifier,* Analisa hasil penerapan *Naive Bayes Classifier* dengan *Backward Elimination* dapat dilihat pada tabel 11,

Tabel 11 Analisa hasil penerapan *Naive Bayes Classifier* dengan *Backward Elimination*

	Data Asli		Data Dummy	
	18 parameter awal	18 parameter efektif	24 parameter awal	20 parameter efektif
Akurasi (%)	88	88	88,8	89,6
Presisi (%)	88,08	88,08	89,21	89,82
Recall (%)	88,02	88,02	88,9	89,75

5. Kesimpulan dan Saran

5.1 Kesimpulan

Penelitian yang menerapkan *dummy variable* pada metode *Naive Bayes* dengan *backward feature selection* dengan *preprocessing* data menggunakan *pearson corellation* pada kedua data, baik data asli maupun data dummy telah dilakukan, Penelitian ini membandingkan hasil *naive bayes classifier* pada data asli dan pada data dummy dengan 125 data asli,

Dari 18 percobaan diperoleh akurasi terbaik pada asli sebesar 88% dengan 18 parameter, Sedang dari 24 percobaan untuk data dummy diperoleh nilai akurasi terbaik sebesar 89,6% dengan 20 parameter,

Dengan begitu hasil penelitian dapat disimpulkan bahwa nilai akurasi klasifikasi data menggunakan *dummy variables* lebih tinggi daripada nilai akurasi pada data asli, Selain itu, penggunaan *dummy variables* dapat lebih meningkatkan efektivitas dari beberapa kemungkinan parameter yang memiliki beberapa nilai, seperti misalnya nyeri haid diubah menjadi nyeri ketika haid dan nyeri setelah haid,

5.2 Saran

Saran untuk penelitian selanjutnya adalah melanjutkan penelitian pengaruh penggunaan dummy variable pada klasifikasi penyakit kandungan dengan mengganti metode *prepocessing* data awal yang menggunakan *pearson correlation* menjadi *pearson partial correlation* sebagai efektivitas fitur dan untuk meningkatkan hasil akurasi,

6. Daftar Pustaka

[1] P, Cichosz, “Data mining Algorithms: Explained Using R,” *Wiley*, 2015,

[2] M, Martinez-Arroyo and L, E, Sucar, “Learning an Optimal Naive Bayes Classifier,” *18th International Conference on Pattern Recognition (ICPR'06)*, Hong Kong, 2006, pp, 1236-1239,

[3] R, Nilsson, J,M Pena, J, Bjokergen, and J, Tegner, “Detecting multivariate differentially expressed genes,” BMC Bioinform, 8:150, 2007,

[4] U, Stanczyk, “Feature Selection for Data and Pattern Recognition,” Springer, 2014,

[5] J, M, Wooldridge, “Introductory Econometrics: A Modern Approach,” Cencage Learning US, 2012,

[6] P, A, Nugraha, R, Saptono, dan M, E, Sulistyo, "Perbandingan Metode Probablistik Naive Bayes Classifier Dan Jaringan Syaraf Tiruan Learning Vector Quantization Dalam Kasus Klasifikasi Penyakit Kandungan," ITSMART, 2013,

[7] M, Shofieyuddin, R, Saptono, and A, Doewes, “The Effect of Using Dummy Variable on Classification of Womb Disease with C4,5 Method,” ITSMART, 2016

[8] A Prabhawaningrum, "Perbandingan Algoritma Levenberg-Marquadt Dengan Backpropagation Untuk Mendiagnosa Jenis Penyakit Kandungan," Surakarta, 2013,

[9] G, I, Webb, “Naive Bayes,” Encyclopedia of Machine Learning, pp 713-714, 2017,

[10] D, Mladenic, “Feature Selection in Text Mining,” Encyclopedia of Machine Learning, pp 406-410, 2010

[11] A, M, De Silva and P,H,W, Leong, “Grammar-Based Feature Generation for Time-Series Prediction,” Springer, 2015,

[12] Narimawati, Umi, "Metodologi Penelitian: Dasar Penyusun Penelitian Ekonomi," *Jakarta: Genesis* (2010),

[13] R, Panthong and A, Srvihok, “Wrapper Feature Subset Selection for Dimension Reduction Based on Ensemble Learning Algorithm,” Procedia Computer Science 72, 2015,

[14] I, Guyon and A, Elisseeff, “An Introduction to Variable and Fature Selection,” Journal of Machine Learning Research 3, 2003,

[15] J, Han, M, Kamber, and J, Pie, “Data Mining: Concept and Technique,” Morgan Kaufmann Publisher, 2012