BAB IV
HASIL DAN PEMBAHASAN

4.1 Hasil Penelitian

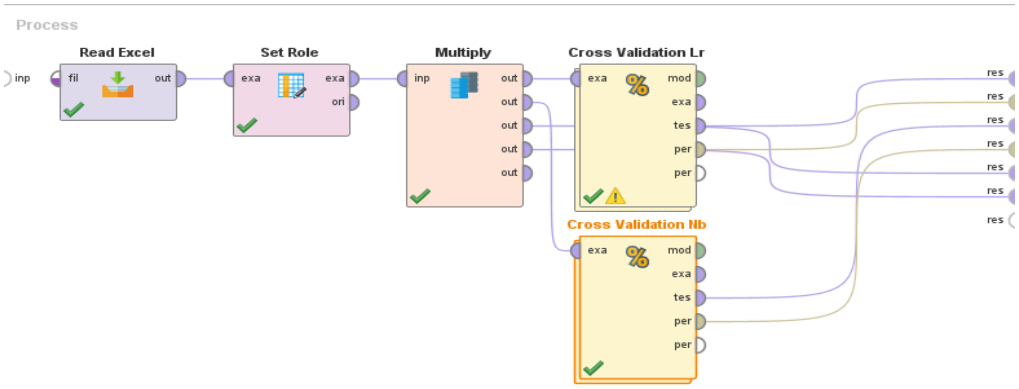
Tujuan dari penelitian ini adalah agar dapat mengetahui tingkat akurasi pengelompokkan (klasifikasi) dari metode Naïve Bayes , Logistic Regression menggunakan Cross-validation (CV) dan metode feature forward selection untuk meningkatkan kinerja algoritma Logistic Regression dan Naïve Bayes. Dan penelitian ini juga melihat perbandingan klasifikasi metode Naïve Bayes , Logistic Regression dengan menggunakan feature forward selection sebagai langkah untuk mengoptimalkan kinerja dari metode Logistic Regression dan Naïve Bayes. Perbandingan dapat dilihat pada proses klasifikasi data dengan nilai akurasi, precision, recall, auc(optimistic), auc(pessimistic), auc.

Dataset yang digunakan pada penelitian ini berjumlah 1183 data yang diproses pada penelitian ini. Setelah itu data akan dibagi menjadi 2 data, Training dan data Testing, pembagian data tersebut akan dilakukan dengan menggunakan proses Cross-validation. Yang diikuti oleh proses masing-masing algoritma yaitu Naïve Bayes , Logistic Regression, dan algoritma feature forward selection.

4.1.1 Implementasi Algoritma Logistic Regression dan Naïve Bayes Dengan Cross-validation

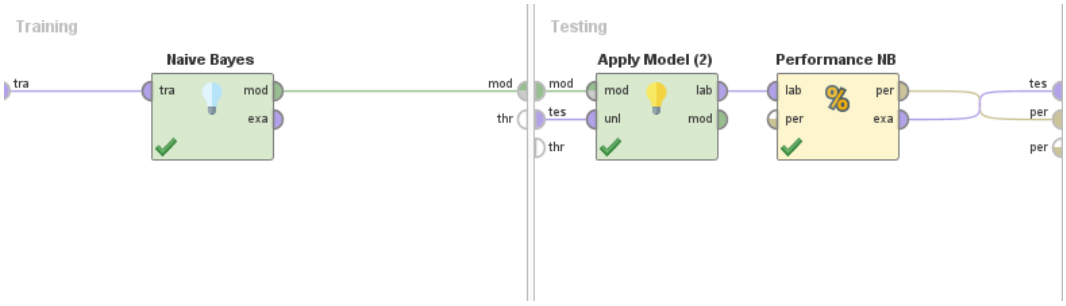
Desain awal dari penggunaan algoritma Logistic Regression, Naïve Bayes dengan Cross-validation pada Tools RapidMiner sebelum menggunakan optimaliasasi Feature Forward Selection, di lakukan dengan tahapan-tahapan sebagai berikut :

1. Tahapan desain area kerja atau workflow diagram pada Tools RapidMiner, seperti terlihat pada gambar 4.1 berikut :

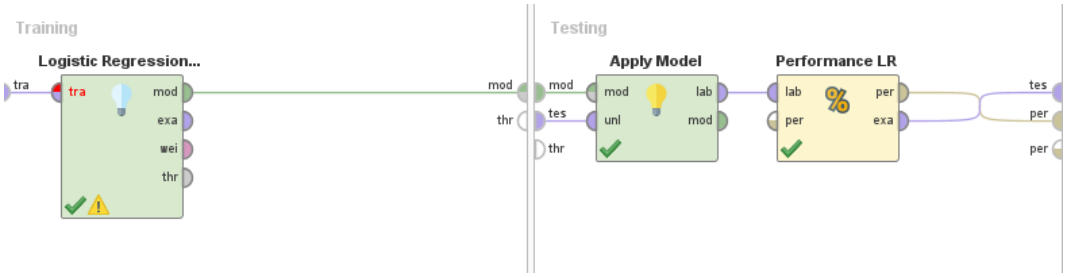


Gambar 4.1 Workflow diagram utama LR, NB dengan CV

2. Tahapan proses data training dan testing pada algoritma Naïve Bayes dan Logistic Regression seperti pada gambar 4.2 dan 4.3 berikut :



Gambar 4.2 Proses data Training dan Testing Algoritma NB



Gambar 4.3 Proses data Training dan Testing Algoritma LR

3. Tahapan hasil dari algoritma Logistic Regression dan Naïve Bayes, dengan menggunakan pengujian number of folds, 10, 5 , 15, hasilnya tampak pada tabel 4.1 berikut :

Tabel 4.1 Nilai Akurasi Logistic Regression dan Naïve Bayes

No	Algoritma	Number Of Folds	Akurasi
1	Logistic Regression	10	86.38 %
2	Naïve Bayes	10	60.92 %
3	Logistic Regression	5	86.38 %
4	Naïve Bayes	5	59.64 %
5	Logistic Regression	15	86.38 %
6	Naïve Bayes	15	60.06 %

Hasil diatas menggunakan number of folds 10 sebagai default atau standard yang diberikan oleh component Cross-validation dengan confusion matrix sebagai berikut : nilai akurasi Logistic Regression 86.38 % dengan standar deviasi +/- 0.44 % dan nilai akurasi Naïve Bayes 60.92 % dengan standar deviasi +/- 5.70 % .

4. Tahapan Proses mengukur performa klasifikasi dari algoritma Naïve Bayes dan Logistic Regression menggunakan ConfusionMatrix seperti tampak pada tabel 4.2 dan tabel 4.3

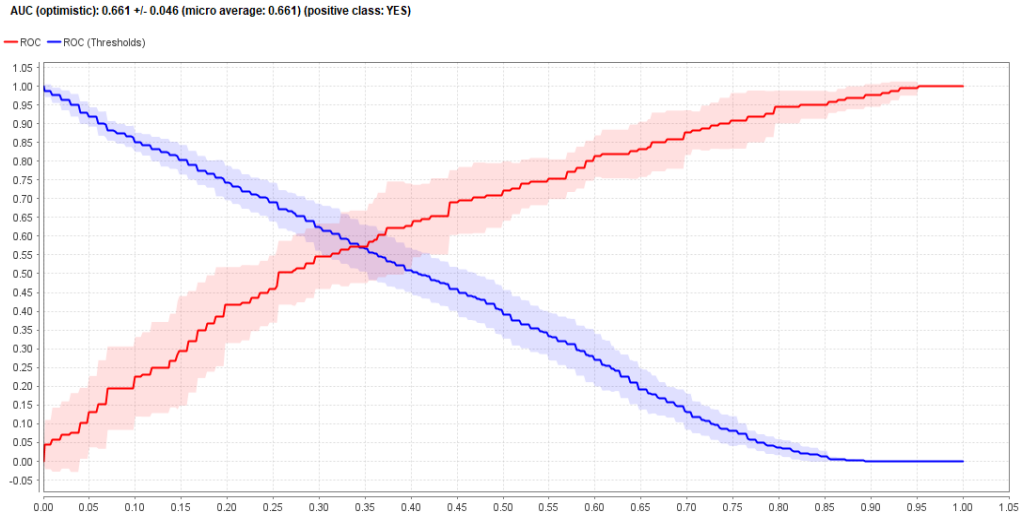
Tabel 4.2 ConfusionMatrix Logistic Regression

Model/Logistic Regression	True/NO	True/YES	Class Precision
Pred. No	1021	161	86.38 %
Pred. Yes	0	0	0.00 %
Class Recall	100.00 %	0.00 %	

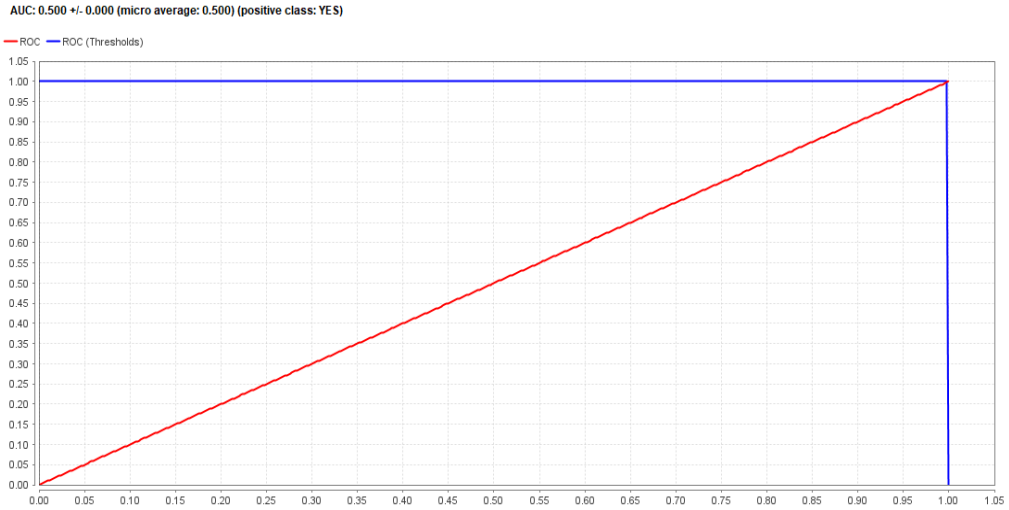
Tabel 4.3 ConfusionMatrix Naïve Bayes

Model/Naïve Bayes	True/NO	True/YES	Class Precision
Pred. No	618	59	91.29 %
Pred. Yes	403	102	20.20 %
Class Recall	60.53 %	63.35 %	

5. Tahapan pada wilayah yang menunjukkan tingkat akurasi menggunakan AUC (Area Under Curve) pada algoritma Naïve Bayes dan Logistic Regression, seperti pada gambar 4.4 dan 4.5



Gambar 4.4 AUC (Area Under Curve) Naïve Bayes



Gambar 4.5 AUC (Area Under Curve) Logistic Regression

6. Tahapan evaluasi hasil penilaian AUC dari penerapan algoritma Logistic Regression dan Naïve Bayes, berdasarkan acuan pada tabel 4.4 berikut :

Tabel 4.4 Acuan Penilaian AUC

Nilai Range	Keterangan
0.90-1.00	Klasifikasi sangat baik
0.80-0.90	Klasifikasi baik
0.70-0.80	Klasifikasi cukup
0.60-0.70	Klasifikasi buruk
0.50-0.60	Klasifikasi salah

a. Logistic Regression

Perhitungan akurasi dari model Logistic Regression yaitu :

Akurasi $= \frac{1021+0}{1021+161+0+0} = 0{,}86379 = 86\ \%$

Error $= \frac{0+0}{1021+0+0+161} = 0$

AUC = 1.00

b. Naïve Bayes

Perhitungan akurasi dari model Naïve Bayes yaitu :

Akurasi $= \frac{618+102}{618+102+403+59} = 0{,}609137 = 61\ \%$

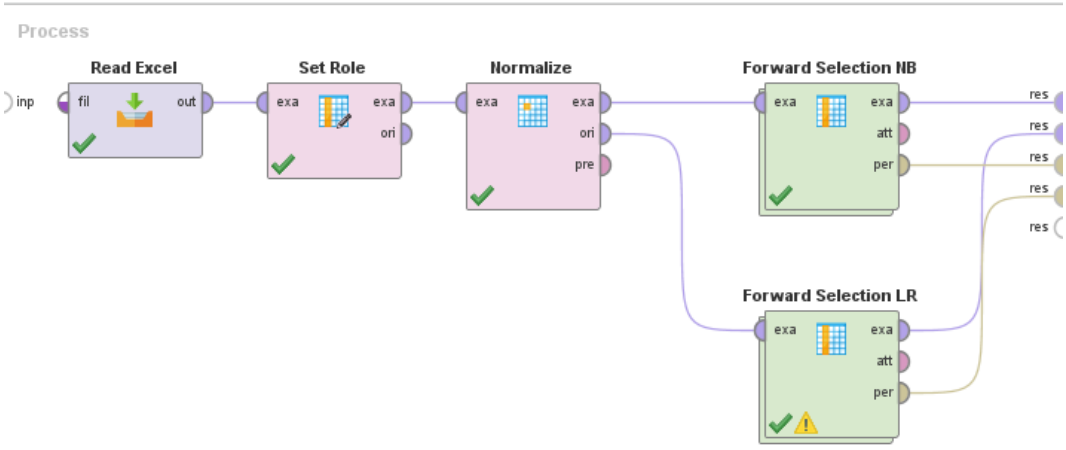
Errore $= \frac{403+59}{618+102+403+59} = 0{,}390863 = 39\ \%$

AUC = 0.661

4.1.2 Implementasi Optimalisasi Feature Forward Selection pada Logistic Regression dan Naïve Bayes

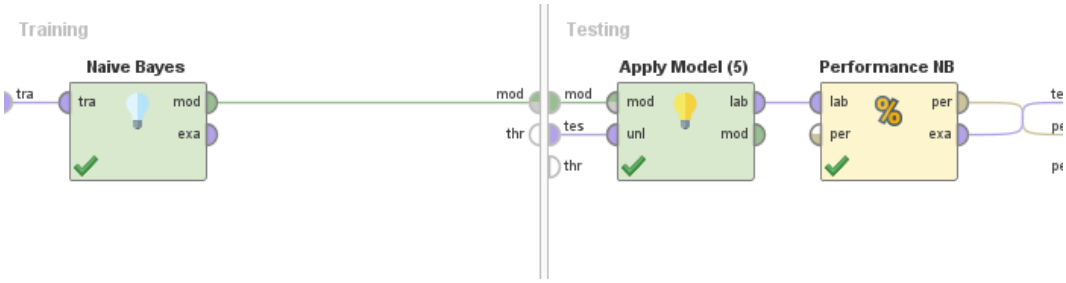
Penerapan optimalisasi Feature Forward Selection pada algoritma Logistic Regression dan Naïve Bayes, bertujuan untuk meningkatkan performa dari kinerja masing-masing algoritma dengan tahapan sebagai berikut :

1. Tahapan desain area kerja atau workflow diagram pada Tools RapidMiner, seperti terlihat pada gambar 4.6 berikut :

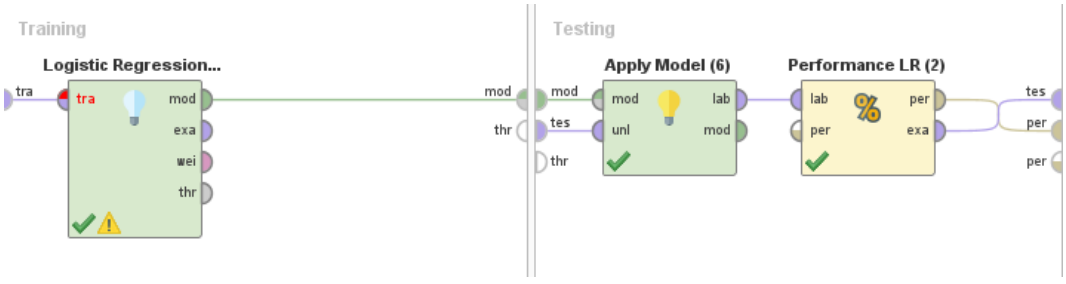


Gambar 4.6 Optimalisasi Feature Forward Selection pada algoritma LR dan NB

2. Tahapan proses data training dan testing optimalisasi Feature Forward Selection pada algoritma Naïve Bayes dan Logistic Regression seperti pada gambar 4.7 dan 4.8 berikut :



Gambar 4.7 Proses optimalisasi Featuare Forward Selection data Training dan Testing Algoritma NB



Gambar 4.8 Proses optimalisasi Featuare Forward Selection data Training dan Testing Algoritma LR

3. Tahapan hasil dari Optimalisasi Feature Forward Selection pada algoritma Logistic Regression dan Naïve Bayes, dengan menggunakan pengujian number of folds, 10 hasilnya tampak pada tabel 4.5 berikut :

Tabel 4.5 Nilai Akurasi Optimalisasi Feature Forward Selection pada Algoritma LR dan NB

No	Algoritma	Number Of Folds	Akurasi
1	Logistic Regression	10	86.47 %
2	Naïve Bayes	10	86.38 %

Hasil diatas menggunakan number of folds 10 sebagai default atau standard yang diberikan dengan menggunakan feature forward selection maka diperoleh confusion matrix sebagai berikut : nilai akurasi Logistic Regression 86.47 % dengan standar deviasi +/- 0.76 % dan nilai akurasi Naïve Bayes 86.38 % dengan standar deviasi +/- 0.44 %.

4. Tahapan Proses mengukur performa klasifikasi optimalisasi feature forward Selection pada algoritma Naïve Bayes dan Logistic Regression menggunakan ConfusionMatrix seperti tampak pada tabel 4.6 dan tabel 4.7

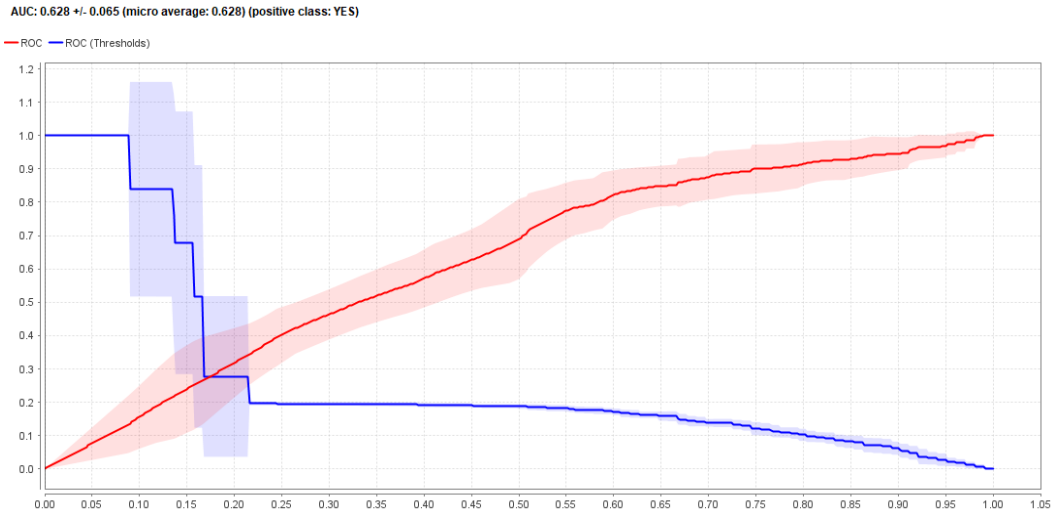
Tabel 4.6 ConfusionMatrix Naïve Bayes dengan metode Feature Forward Selection

Model/Naïve Bayes	True/NO	True/YES	Class Precision
Pred. No	1021	161	86.38 %
Pred. Yes	0	0	0.00 %
Class Recall	100.00 %	0.00 %	

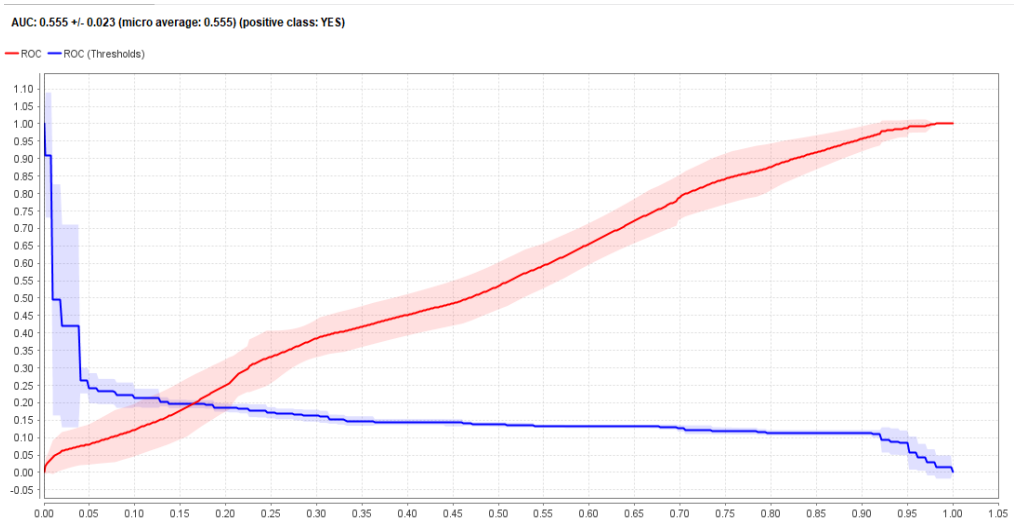
Tabel 4.7 ConfusionMatrix Logistic Regression dengan metode Feature Forward Selection

Model/Logistic Regression	True/NO	True/YES	Class Precision
Pred. No	1020	159	86.51 %
Pred. Yes	1	2	66.67 %
Class Recall	99.90 %	1.24 %	

5. Tahapan pada wilayah yang menunjukkan tingkat akurasi menggunakan AUC (Area Under Curve) Optimalisasi Feature Forward Selection pada algoritma Naïve Bayes dan Logistic Regression, seperti pada gambar 4.9 dan 4.10



Gambar 4.9 AUC (Area Under Curve) Optimalisasi Feature Forward Selection pada algoritma Naïve Bayes



Gambar 4.10 AUC (Area Under Curve) Optimalisasi Feature Forward Selection pada algoritma Logistic Regression

6. Tahapan evaluasi hasil penilaian AUC dari penerapan optimalisasi Feature Forward Selection pada algoritma Logistic Regression dan Naïve Bayes, sebagai berikut :

a. Logistic Regression Feature Forward Selection

Perhitungan akurasi dari model regresi logistic dengan metode Feature Forward Selection yaitu :

Akurasi $= \frac{1021+159}{159+1+1021+2} = 0{,}997462 = 100\ \%$

Error $= \frac{1+2}{159+1+1021+2} = 0{,}002538 = 0\ \%$

AUC = 0.625

b. Naïve Bayes Feature Forward Selection

Perhitungan akurasi dari model Naïve Bayes dengan metode Feature Forward Selection yaitu :

Akurasi $= \frac{1021+0}{1021+161+0+0} = 0{,}86379 = 86\ \%$

Error $= \frac{0+0}{1021+0+0+161} = 0$

AUC = 0.682

4.2 Perbandingan Nilai Akurasi

Berdasarkan hasil pengolahan data yang diproses menggunakan rapidminer pada studi ini, survei yang dilakukan dengan ukuran sampel 1183 siswa dari kelompok usia yang berbeda dengan 19 atribut, maka diperoleh perbandingan nilai akurasi dari algoritma Logistic Regression dan Naive bayes sebelum menggunakan Feature Forward Selection dan sesudah menggunakan Feature Forward Selection seperti terlihat pada tabel 4.8 sebagai berikut :

a. Perbandingan Nilai Akurasi

Tabel 4.8 Perbandingan nilai akurasi Algoritma sebelum menggunakan Feature Forward Selection dan sesudah menggunakan Feature Forward Selection

Algoritma	Nilai akurasi sebelum menggunakan Feature Forward Selection	Nilai akurasi sesudah menggunakan Feature Forward Selection	Selisih
Logistic Regression	86.38%	86.47%	0.09
Naïve Bayes	60.92%	86.38%	25.46

Nialai akurasi dari kedua metode memperlihat perbedaan sebelum menggunakan Feature Forward Selection dan sesudah menggunakan Feature Forward Selection, artinya penggunaan Feature Forward Selection sangat membantu kinerja dari masing algoritma untuk mengoptimalisasi nialai akurasi yang lebih baik. Berdasarkan data pada tabel 4.8, nilai akurasi Algoritma Regression Logistic memiliki akurasi yang tinggi yaitu sebesar 86.47% sedangkan Algoritma Naïve Bayes memiliki nilai akurasi sebesar 86.36%

b. Perbandingan ConfusionMatrix

Perbandingan ConfusionMatrix sebelum dan sesudah menggunakan Feature Forward selection dapat dilihat pada tabel 4.9 dan tabel 4.10 dari ke dua algoritma, yaitu Logistic Regression dan algoritma Naïve sebagai berikut :

Tabel 4.9 Perbandingan ConfusionMatrix Logistic Regression

Metode	ConfusionMatrix sebelum		Class Precision	ConfusionMatrix Sesudah		Class Precision
Logistic Regression	True/No	True/Yes		True/No	True/Yes	
Pred. No	1021	161	86.38 %	1020	159	86.51 %
Pred. Yes	0	0	0.00 %	1	2	66.67 %
Class Recall	100.00 %	0.00 %		99.90 %	1.24 %	

Tabel 4.10 Perbandingan ConfusionMatrix Naïve Bayes

Metode	ConfusionMatrix sebelum		Class Precision	ConfusionMatrix Sesudah		Class Precision
Naïve Bayes	True/No	True/Yes		True/No	True/Yes	
Pred. No	618	59	91.29 %	1021	161	86.38 %
Pred. Yes	403	102	20.20 %	0	0	0.00 %
Class Recall	60.53 %	63.35 %		100.00 %	0.00 %	

Berdasarkan perbandingan ConfusionMatrix dari ke dua algoritma diatas dapat diambil kesimpulan bahwa : nilai Class Precision Logistic Regression

sebelum menggunakan Feature Forward Selection adalah 86.38%, setelah menggunakan Feature Forward Selection menjadi 86.51%, sedangkan nilai Class Precision Naïve Bayes sebelum menggunakan Feature Forward Selectio adalah 91.29%, setelah menggunakan Feature Forward Selection menjadi 86.38%. Dari perbandingan diatas, maka dapat dihitung nilai akurasi sebagai berikut :

- ConfusionMatrix Logistic Regression sebelum menggunakan Feature Forward Selection

  Akurasi = (TP +TN)/(TP+FP+FN+TN)
  = (1021+0)/(1021+161+0+0)
  = 1021/1182 = 0.863
  = 0.863 * 100% = 86.38

- ConfusionMatrix Logistic Regression setelah menggunakan Feature Forward Selection

  Akurasi = (TP+TN)/(TP+FP+FN+TN)
  = (1020+2)/(1020+159+1+2)
  = 1022/1182 = 0.8646
  = 0.8646 * 100% = 86.4636 = 86.51

Hal yang sama berlaku juga untuk perhitungan ConfusionMatrix algoritma Naïve Bayes sebelum menggunakan Feature Forward Selection dan sesudah menggunakan Feature Forward Selection.

c. Perbandingan Nilai AUC (Area Under Curve)

Perbandingan Nilai AUC (Area Under Curve) dari algoritma Logistic Regression dan Naïve Bayes sebelum menggunakan Feature Forward Selection dan sesudah menngunakan Feature Forward Selection dapat dilihat pada Tabel 4.11 berikut ini

Tabel 4.11 Perbandingan Nilai AUC

Metode	AUC Sebelum FS			AUC Sesudah FS		
	Akurasi	Error	Nialai AUC	Akurasi	Error	Nilai AUC
Logistic Regression	0.863	0	1.00	0.997	0	0.625
Naïve Bayes	0.609	0.390	0.661	0.863	0	0.682

d. Perbandingan ExampleSet (data tabel)

Perbandingan ExampleSet dari kedua metode memperlihatkan perbedaan pembacaan atribut dari penggunaan data set, baik sebelum menggunakan Feature Forward Selection dan sesudah menggunakan Feature Forward Selection dapat dilihat pada gambar 4.11, 4.12, 4.13, dan 4.14 berikut :

Row No.	Health issue...	prediction(H...	confidence(NO)	confidence(YES)	ID	Region of residence	Age of Subject	Time spent on Online Class
1	NO	NO	1.000	0.000	R4	Delhi-NCR	20	3
2	NO	NO	1.000	0.000	R19	Delhi-NCR	21	0
3	NO	NO	1.000	0.000	R26	Delhi-NCR	20	5
4	NO	NO	1.000	0.000	R38	Outside Delhi-NCR	20	4
5	NO	NO	1.000	0.000	R44	Delhi-NCR	21	5
6	YES	NO	1.000	0.000	R51	Delhi-NCR	19	0
7	YES	NO	1.000	0.000	R59	Delhi-NCR	20	5
8	NO	NO	1.000	0.000	R104	Outside Delhi-NCR	16	3
9	NO	NO	1.000	0.000	R109	Delhi-NCR	18	2
10	NO	NO	1.000	0.000	R110	Outside Delhi-NCR	10	3
11	YES	NO	1.000	0.000	R125	Delhi-NCR	19	1
12	NO	NO	1.000	0.000	R135	Delhi-NCR	18	5
13	NO	NO	1.000	0.000	R137	Delhi-NCR	18	3

ExampleSet (1,182 examples, 4 special attributes, 18 regular attributes)

Gambar 4.11 AUC ExempleSet Logistic Regression sebelum menggunakan Forward Selection

Row No.	Health issue...	Prefered social media platform	Do you find yourself more connected with your family, close friends , relatives ?
1	NO	Linkedin	YES
2	NO	Youtube	NO
3	NO	Linkedin	YES
4	NO	Instagram	NO
5	NO	Instagram	NO
6	YES	Youtube	YES
7	NO	Instagram	YES
8	YES	Instagram	YES
9	NO	Whatsapp	NO
10	YES	Instagram	NO
11	NO	Instagram	NO
12	YES	Instagram	YES
13	YES	Instagram	NO
14	NO	None	YES

ExampleSet (1,182 examples, 1 special attribute, 2 regular attributes)

Gambar 4.12 ExampelSet Logistic Regression setelah menggunakan Forward Selection

Row No.	Health issue...	prediction(H...	confidence(NO)	confidence(YES)	ID	Region of residence	Age of Subject	Time spent on Online Class
1	YES	YES	0.258	0.742	R13	Delhi-NCR	21	3
2	NO	YES	0.160	0.840	R18	Delhi-NCR	20	1
3	NO	YES	0.208	0.792	R23	Delhi-NCR	21	4
4	NO	NO	0.912	0.088	R41	Delhi-NCR	24	4
5	NO	YES	0.393	0.607	R66	Delhi-NCR	20	4
6	NO	NO	1.000	0.000	R67	Outside Delhi-NCR	19	0
7	NO	NO	0.760	0.240	R82	Delhi-NCR	21	3
8	NO	YES	0.089	0.911	R100	Delhi-NCR	21	1
9	NO	YES	0.043	0.957	R104	Outside Delhi-NCR	16	3
10	NO	NO	0.993	0.007	R110	Outside Delhi-NCR	10	3
11	NO	NO	0.647	0.353	R123	Outside Delhi-NCR	21	4
12	NO	YES	0.403	0.597	R137	Delhi-NCR	18	3
13	NO	NO	0.975	0.025	R139	Outside Delhi-NCR	22	1

ExampleSet (1,182 examples, 4 special attributes, 18 regular attributes)

Gambar 4.13 ExampleSet Naïve Bayes sebelum menggunakan Forward Selection

Row No.	Health issue...	Age of Subject
1	NO	0.151
2	NO	0.151
3	NO	-0.030
4	NO	-0.030
5	NO	0.151
6	YES	0.151
7	NO	-0.211
8	YES	-0.211
9	NO	0.151
10	YES	-0.030
11	NO	0.151
12	YES	0.151
13	YES	0.151
14	NO	0.332

ExampleSet (1,182 examples, 1 special attribute, 1 regular attribute)

Gambar 4.14 ExampleSet Naïve Bayes sesudah menggunakan Forward Selection

Berdasarkan perbandingan ExampleSet yang dibaca oleh masing-masing algoritma pada saat melakukan proses klasifikasi terhadap dataset, kemampuan optimalisasi metode feature forward selection bertujuan untuk memilih variable yang digunakan oleh algorithma Logistic Regression dan Naïve Bayes adalah seperti tampak pada tabel 4.12 berikut :

Tabel 4.12 perbandingan ExampleSet

Algoritma	ExampleSet Sebelum Menggunakan FS		ExampleSet Sesudah Menggunakan FS	
	Jumlah atribut yang terbaca	Keterangan	Jumlah atribut yang terbaca	Keterangan
Logistic Regression	19	Atribut terbaca semua	3	- health issue during lock down - preferred social media platform - Do you find yourself more connected with your family, close friends , relatives
Naïve Bayes	19	Atribut terbaca semua	2	- health issue during lock down - age of subject

Logistic Regression mampu menghasilkan 3 variable yaitu : health issue during lock down, preferred social media platform dan Do you find yourself more connected with your family, close friends , relatives dan untuk Naïve Bayes mampu menghasilkan 2 variabel yaitu : health issue during lock down dan age of subject.