BAB III
METODOLOGI PENELITIAN

3.1. Sumber Data

Penggunaan data pada penelitian ini adalah, menggunakan sumber data sekunder, yaitu data yang telah tersedia yang bersumber pada alamat https://www.kaggle.com/datasets/kunal28chaturvedi/covid19-and-its-impact-on-students

Data yang tersedia merupakan data COVID-19 dan dampaknya terhadap pendidikan, kehidupan sosial, dan kesehatan mental siswa. Dalam studi ini, survei yang dilakukan dengan ukuran sampel 1183 siswa dari kelompok usia yang berbeda

3.2. Atribut Penelitian

Terdapat beberapa variabel atau atribut-atribut yang digunakan pada proses penelitian ini seperti pada table 3.1.

Tabel 3.1 Atribut COVID-19 Survey Student Respons

No	Atribut/Field	Keterangan
1	ID	Identitas wilayah
2	Region of residence	Wilayah tempat tinggal
3	Age of Subject	Usia Subyek
4	Time spent on Online Class	Waktu yang dihabiskan di Kelas Online
5	Rating of Online Class experience	Peringkat pengalaman Kelas Online
6	Medium for online class	Media untuk kelas online
7	Time spent on self study	Waktu yang dihabiskan untuk belajar mandiri
8	Time spent on fitness	Waktu yang dihabiskan untuk kebugaran
9	Time spent on sleep	Waktu yang dihabiskan untuk tidur
10	Time spent on social media	Waktu yang dihabiskan di media sosial
11	Prefered social media platform	Platform media sosial pilihan

12	Time spent on TV	Waktu yang dihabiskan di TV
13	Number of meals per day	Jumlah makan per hari
14	Change in your weight	Ubah berat badan Anda
15	Health issue during lockdown	Masalah kesehatan selama penguncian
16	Stress busters	Penghilang stres
17	Time utilized	Waktu yang digunakan
18	Do you find yourself more connected with your family, close friends , relatives ?	Apakah Anda menemukan diri Anda lebih terhubung dengan keluarga, teman dekat, kerabat Anda?
19	What you miss the most	Apa yang paling kamu rindukan

3.3. Tahapan Penelitian

Tahapan-tahapan yang dilakukan untuk pencapaian terhadap proses penelitian ini akan dilakukan dengan langkah-langkah sebagai berikut :

1. Teknik pengumpulan Data

   Langkah awal dalam penelitian ini adalah dengan melakukan pengambilan data sekunder yaitu data yang telah tersedia yang bersumber pada alamat https://www.kaggle.com/datasets/kunal28chaturvedi/covid19-and-its-impact-on-students

2. Pengujian dan Proses Data

   Pada tahapan ini yang akan dibahas adalah pengolahan data dengan pendekatan algoritma Logistic Regression dan Naïve Bayes bersama penggunaan optimasi algoritma feature forward selection

   A. Penggunaan algoritma Logistic Regression , Naïve Bayes dan forward selection

      a. Logistic Regression

         1 Data dibagi menjadi 2, yaitu data testing dan data training dengan menggunakan cross-validation (CV)

         2 Proses uji kemandirian dengan data training (data latih).

         3 Membangun model Logistic Regression menggunakan data training (data latih)

4 Uji coba keterkaitan variable secara keseluruhan dan secara individu.

5 Membuat keakuratan nilai validasi prediksi dari model dengan data testing (data uji).

6 Mengukur nilai akurasi, precision, recall, auc(optimistic), auc(pessimistic), auc

b. Naïve Bayes

1 Pembagian data menjadi 2, yaitu data testing (data uji) dan data training (data latih) dengan menggunakan Cross-validation

2 Mengukur probabilitas awal (prior probability) $((Y))$.

3 Mengukur semua fitur dalam vektor X (Π $(Xi|Y)$ ki =1) dari Nilai Probabilitas independen kelas Y.

4 Mengukur jumlah (nilai) posterior probability bagi masing-masing klasifikasi $((Y|X))$.

5 Menghitung jumlah (nilai) maksimum pada perhitungan posterior probability dari prediksi yang didapat.

6 Menghitung precision, recall, auc(optimistic), auc(pessimistic), auc dan nilai akurasi

c. Feature Forward Selection

1 Operator Forward Selection dimulai dengan pemilihan atribut yang kosong dan, di setiap putaran, ia menambahkan setiap atribut yang tidak digunakan dari ExampleSet yang diberikan

2 Untuk setiap atribut yang ditambahkan, kinerja diperkirakan menggunakan operator dalam, mis. validasi silang

3 Hanya atribut yang memberikan peningkatan kinerja tertinggi yang ditambahkan ke pilihan

4 Dimulai dengan seleksi yang dimodifikasi. Implementasi ini menghindari konsumsi memori tambahan selain memori

yang awalnya digunakan untuk menyimpan data dan memori yang mungkin diperlukan untuk menerapkan operator dalam

5 Parameter berhenti menentukan kapan iterasi harus dibatalkan

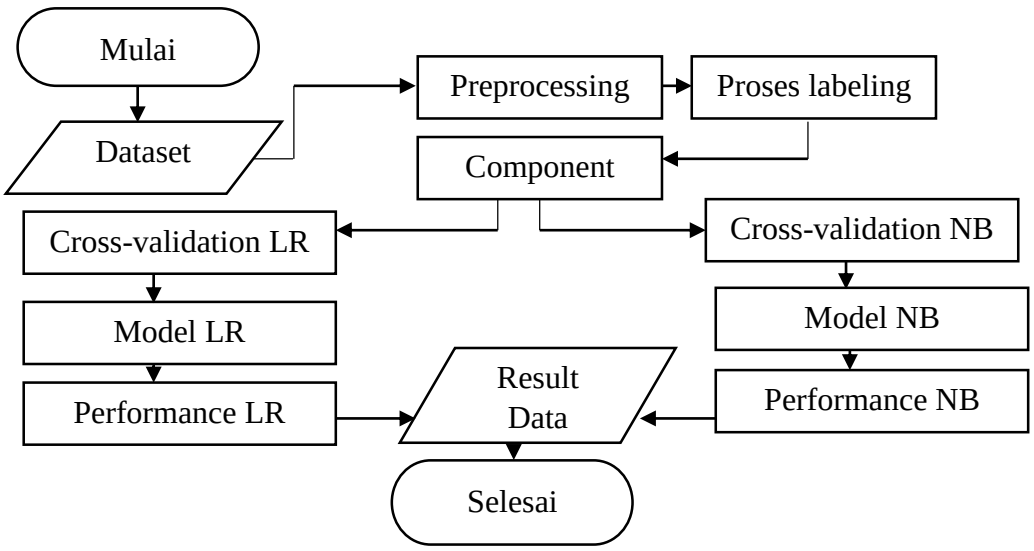
B. Menentukan pengklasifkasikan metode terbaik dari algoritma yang diterapkan dengan menggunakan akurasi, precision, recall, auc(optimistic), auc(pessimistic), auc

3. Pembahasan dan Uraian Penelitian

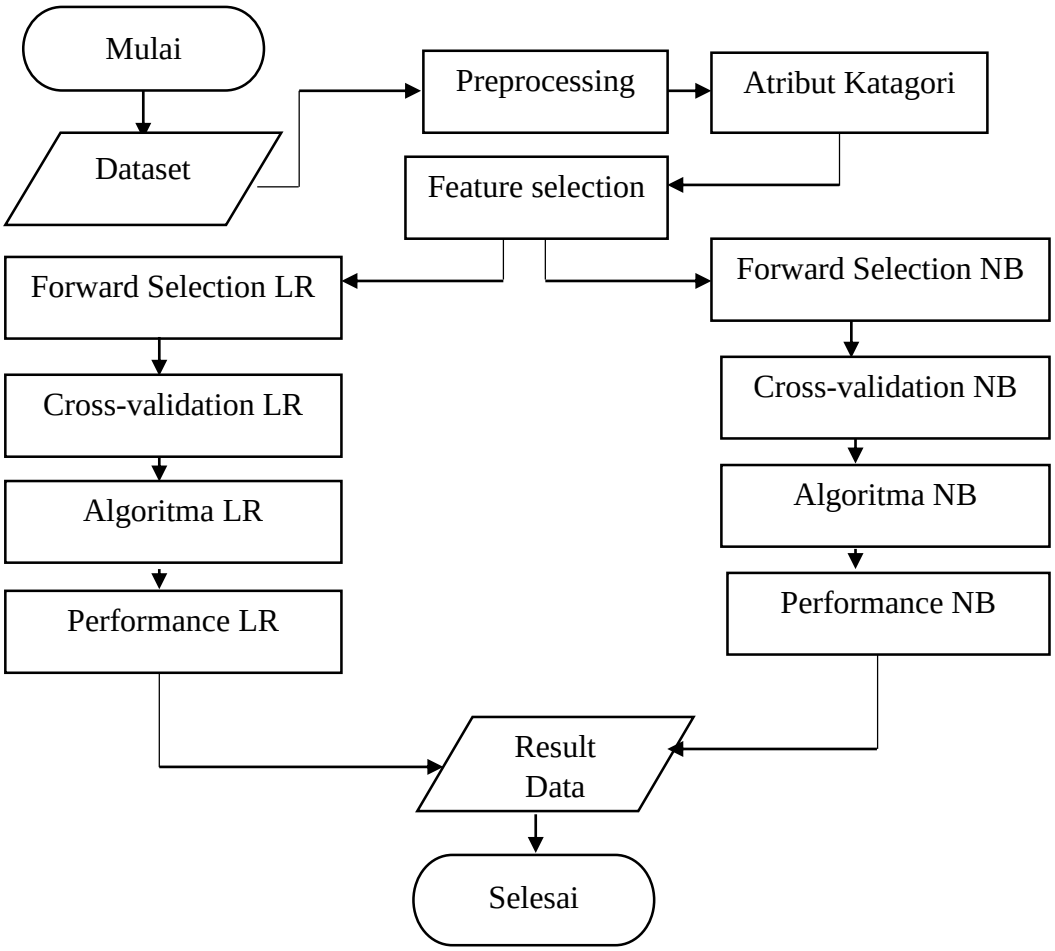
Langkah-langkah Pada tahapan hasil penelitian ini akan membahas pencapaian yang telah dilakukan dan kemudian disimpulkan.

Hasil pencapaian di atas dapat digambarkan menggunakan flow chart (diagram alir), sebagai berikut :

1. Diagram alir Cross-validation Logistic Regression dan Naïve Bayes pada gambar 3.1
2. Diagram alir Optimalisasi Feature Forward Selection pada algoritma Logistic Regression dan Naïve Bayes pada gambar 3.2



Gambar 3.1 Cross-validation Logistic Regression dan Naïve Bayes



Gambar 3.2 Featrue Forward Selecion dengan metode Logistic Regression dan Naïve Bayes

3.4 Analisis Data

Berdasarkan penerapan algoritma Regression Logistic dan Naïve Bayes dengan metode optimalisasi Feature Forward Selection, maka dapat diperoleh gambaran secara umum terhadap perhitungan dari masing-masing algoritma yang digunakan. Adapun penerapan perhitungannya adalah sebagai berikut :

A. Perhitungan menggunakan algoritma Naïve Bayes

Pada penulisan ini, penulis menggunakan 1.183 jumlah data dan 19 atribut, adapun yang menjadi atribut pilihan adalah : Health issue during lockdown (Masalah kesehatan selama penguncian) dan Age of Subject (Usia Subyek), dengan penggunaan rumus Probabilitas Naïve Bayes adalah sebagai berikut :

$$P(H/X) = \frac{P(H/X)\ .\ P(H)}{P(X)} \quad (3.1)$$

1. Atribut Health Issue During Lockdown (Masalah kesehatan selama penguncian).

Diketahui : jumlah data adalah : 1183 dan jumlah atribuut adalah :19, pada contoh ini penulis menggunakan 23 record dan 2 atribut dengan pengambilan data secara acak, seperti terlihat pada tabel 3.2

Tabel 3.2
Data usia pelajar

Age of Subject	Health issue during lockdown
15	NO
19	NO
20	NO
17	NO
18	NO
20	NO
18	NO
22	NO
18	NO
15	NO
22	NO
19	NO
20	YES
20	NO
20	NO
18	NO
19	NO
20	YES
20	NO
19	YES
19	NO
18	NO
15	NO

Berdasarkan data pada tabel 1 diatas maka akan dilakukan perhitungan probabilitas terhadap pembelajaran daring, klasifikasi akan dilakukan dengan kriteria YES dan NO.

Maka perhitungan probabilitas dari kriteria YES dan NO adalah sebagai berikut :

$$\boldsymbol{Probabilitas\ (YES) = \frac{3}{23} = 0.13043 = 13\%}$$

$$\boldsymbol{Probabilitas\ (NO) = \frac{20}{23} = \quad 0.86957 = \quad 87\%}$$

Tabel 3.3
probabilitas Kriteria YES dan NO

Atribut	Kriteria	Jumlah	persen	Klasifikasi
Health issue during lockdown	YES	0,13043	13%	3
	NO	0,86957	87%	20

Hasil perhitungan atribut Health issue during lockdown berdasarkan tabel 2 dengan kriteria YES dan NO, 13% menyatakan setuju melakukan pembelajaran jarak jauh, sedangkan 87% memilih pembelajaran tatap muka, hal ini diperkuat dari hasil klasifikasi data yaitu jumlah kriteria YES sebanyak 3 data setuju dengan pembelajaran daring dan jumlah kriteria NO sebanyak 20 data menyatakan tidak setuju dengan pembelajaran daring .

2. Atribut Age of Subject (subjek usia)

Atribut berikut adalah pengelompokkan data dengan pengambilan data pada tabel 1 diatas, pengelompokkan ini dilakukan berdasarkan usia pelajar, seperti terlihat pada tabel 3 berikut ini :

Tabel 3.4
Age Of Subject

Atribut Age Of Subject	Klasifikasi
15	Klasifikasi kriteria YES = 0 dan Klasifikasi kriteria NO = 3
19	Klasifikasi kriteria YES = 1 dan Klasifikasi kriteria NO = 4

20	Klasifikasi kriteria YES = 2 dan Klasifikasi kriteria NO = 5
22	Klasifikasi kriteria YES = 0 dan Klasifikasi kriteria NO = 2
17	Klasifikasi kriteria YES = 0 dan Klasifikasi kriteria NO = 1
18	Klasifikasi kriteria YES = 0 dan Klasifikasi kriteria NO = 5

Adapun perhitungan probibilitasnya atribut age of subject adalah sebagai berikut :

$$\boldsymbol{P(X|Ci)\ p\ usia\ "15\ Tahun"|Setuju\ "YES" = \frac{0}{3} = \ 0}$$

$$\boldsymbol{P(X|Ci)\ p\ usia\ "15\ Tahun"|Tidak\ Setuju\ "NO" = \frac{3}{20} = \ 0.15 = 15\%}$$

$$\boldsymbol{P(X|Ci)\ p\ usia\ "19\ Tahun"|Setuju\ "YES" = \frac{1}{3} = \ 0.33333 = 33\%}$$

$$\boldsymbol{P(X|Ci)\ p\ usia\ "19\ Tahun"|Tidak\ Setuju\ "NO" = \frac{4}{20} = \ 0.2 = 20\%}$$

$$\boldsymbol{P(X|Ci)\ p\ usia\ "20\ Tahun"|Setuju\ "YES" = \frac{2}{3} = \ 0.66667 = 67\%}$$

$$\boldsymbol{P(X|Ci)\ p\ usia\ "20\ Tahun"|Tidak\ Setuju\ "NO" = \frac{5}{20} = \ 0.25 = 25\%}$$

$$\boldsymbol{P(X|Ci)\ p\ usia\ "22\ Tahun"|Setuju\ "YES" = \frac{0}{3} = \ 0}$$

$$\boldsymbol{P(X|Ci)\ p\ usia\ "22\ Tahun"|Tidak\ Setuju\ "NO" = \frac{2}{20} = \ 0.1 = 10\%}$$

$$\boldsymbol{P(X|Ci)\ p\ usia\ "17\ Tahun"|Setuju\ "YES" = \frac{0}{3} = \ 0}$$

$$\boldsymbol{P(X|Ci)\ p\ usia\ "17\ Tahun"|Tidak\ Setuju\ "NO" = \frac{1}{20} = \ 0.05 = 5\%}$$

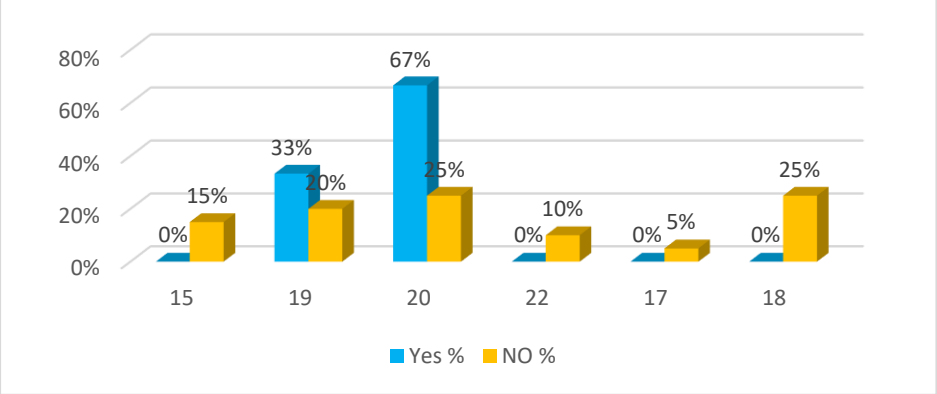
$$\boldsymbol{P(X|Ci)\ p\ usia\ "18\ Tahun"|Setuju\ "YES" = \frac{0}{3} = \ 0}$$

$$\boldsymbol{P(X|Ci)\ p\ usia\ "18\ Tahun"|Tidak\ Setuju\ "NO" = \frac{5}{20} = \ 0.25 = 25\%}$$

Tabel 3.5
Rekapitulasi atribut Age Of Subject

Atribut Age Of Subject	YES	NO
15	0	15%
19	33%	20%
20	67%	25%
22	0	10%
17	0	5%
18	0	25%

Berdasarkan tabel 4 diatas, diperoleh hasil perhitungan atribut Age Of Subject, yaitu usia 15, 22, 17, dan 18, lebih memilih kriteria NO, artinya pelajar tidak setuju dengan pembelajaran daring, berbeda dengan usia 19 dan 20 lebih memilih kriteria YES, artinya pelajar lebih memilih pembelajaran daring, seperti terlihat pada grafik metode pembelajaran berikut ini :



Gambar 3.3 Metode Klasifikasi pembelajaran daring dan tidak daring

B. Perhitungan menggunakan algoritma Regression Logistic

Pada perhitungan ini penulis menggunakan 1.183 jumlah data dan 19 atribut, adapun yang menjadi atribut pilihan adalah : Health issue during lockdown (Masalah kesehatan selama penguncian) dan Age of Subject (Usia Subyek), dengan penggunaan persamaan Log likelihood pada algoritma Regression Logistic, data awal terlihat seperti pada tabel 5 berikut :

Tabel 3.6
Atribut Usia

Age of Subject	Health issue during lockdown
15	NO
19	NO
20	NO
17	NO
18	NO
20	NO
18	NO
22	NO
18	NO
15	NO
22	NO
19	NO
20	YES
20	NO
20	NO
18	NO
19	NO
20	YES
20	NO
19	YES
19	NO
18	NO
15	NO

Berdasarkan data pada tabel diatas akan dikembangkan menjadi bilangan biner pada atribut Health issue during lockdown data seperti terlihat pada tabel 6 berikut ini

Tabel 3.7

Perubahan Atribut

Age of Subject	Health issue during lockdown	X
15	NO	0
19	NO	0
20	NO	0
17	NO	0
18	NO	0

20	NO	0
18	NO	0
22	NO	0
18	NO	0
15	NO	0
22	NO	0
19	NO	0
20	YES	1
20	NO	0
20	NO	0
18	NO	0
19	NO	0
20	YES	1
20	NO	0
19	YES	1
19	NO	0
18	NO	0
15	NO	0

Pada tabel 6 diatas terdapat penambahan atribut yaitu atribut X, atribut ini merupakan pengembangan dari atribut Health issue during lockdown menjadi atribut X, dengan menggantikan nilai kriteria YES menjadi 1 dan kriteria NO menjadi 0, maka diperoleh hasil X = 1 dan X = 0, adapun hasil klasifikasi dari atribut X adalah jumlah nilai 0 = 20 dan jumlah nilai 1 = 3. Dari hasil perubahan atribut yang terlihat pada tabel 6 diatas, maka dapat dihitung persamaan nya adalah sebagai berikut :

Diketahui :

B0 = 0,01 dan B1 = 0,02

Menghitung nilai Logit :

$g(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$

$g(x) = 0{,}01 + 0{,}02 * 15 = 0{,}31$

$g(x) = 0{,}01 + 0{,}02 * 19 = 0{,}39$

$g(x) = 0{,}01 + 0{,}02 * 20 = 0{,}41$

Menghitung exponensial Logit : Exp(Logit) :

Exp (0,31) = 1,363425114

Exp (0,39) = 1,476980794

Exp (0,41) = 1,506817785

Menghitung nilai P(x)

$$p(x) = \frac{1{,}363425114}{(1+1{,}363425114)} = 0{,}576885261$$

Menghitung Loglikelihood

Yi x LN(P(x)i+(1- Yi) x LN(1-P(x)i

$$L(\beta) = 0 * Ln(0{,}576885261) + (1-0) * Ln\,(1-\ 0{,}576885261)$$

$$= -0{,}860111886$$

Dari perhitungan persamaan diatas dapat dilihat hasil keseluruhan seperti terlihat pada tabel 3.8 sebagai berikut :

Tabel 3.8
Hasil perhitungan awal Loglikelihood

B0 = 0,01

B1 = 0,02

Logit	Exp Logit	p(x)= exp(logit)/(1+exp(logit)	Loglikelihood Yi x LN(P(x)i+(1- Yi) x LN(1-P(x)i
0,31	1,363425114	0,576885261	-0,860111886
0,39	1,476980794	0,596282699	-0,907040397
0,41	1,506817785	0,601087879	-0,919014134
0,35	1,419067549	0,586617579	-0,883382155
0,37	1,447734615	0,591458978	-0,89516295
0,41	1,506817785	0,601087879	-0,919014134
0,37	1,447734615	0,591458978	-0,89516295
0,45	1,568312185	0,610639234	-0,943248946
0,37	1,447734615	0,591458978	-0,89516295
0,31	1,363425114	0,576885261	-0,860111886
0,45	1,568312185	0,610639234	-0,943248946
0,39	1,476980794	0,596282699	-0,907040397
0,41	1,506817785	0,601087879	-0,509014134
0,41	1,506817785	0,601087879	-0,919014134
0,41	1,506817785	0,601087879	-0,919014134
0,37	1,447734615	0,591458978	-0,89516295
0,39	1,476980794	0,596282699	-0,907040397
0,41	1,506817785	0,601087879	-0,509014134
0,41	1,506817785	0,601087879	-0,919014134

0,39	1,476980794	0,596282699	-0,517040397
0,39	1,476980794	0,596282699	-0,907040397
0,37	1,447734615	0,591458978	-0,89516295
0,31	1,363425114	0,576885261	-0,860111886
Nilai Maksimal			-19,58433138

Berdasarkan perhitungan nilai awal Loglikelihood pada tabel diatas, maka di peroleh nilai akhir Loglikelihood seperti terlihat pada tabel 3.9 sebagai berikut :

Tabel 3.9
Hasil perhitungan akhir Loglikelihood

B0 = -8,864949558

B1 = 0,363570757

Logit	Exp Logit	p(x)= exp(logit)/(1+exp(logit)	Loglikelihood Y_i x LN$(P(x)_i$+(1- Y_i) x LN(1-$P(x)_i$
-3,41139	0,032995	0,031941444	-0,032462702
-1,95711	0,141267	0,123780675	-0,132138848
-1,59353	0,203206	0,168887209	-0,184989764
-2,68425	0,068273	0,063909347	-0,066042956
-2,32068	0,098207	0,089425003	-0,093679015
-1,59353	0,203206	0,168887209	-0,184989764
-2,32068	0,098207	0,089425003	-0,093679015
-0,86639	0,420465	0,296005416	-0,350984615
-2,32068	0,098207	0,089425003	-0,093679015
-3,41139	0,032995	0,031941444	-0,032462702
-0,86639	0,420465	0,296005416	-0,350984615
-1,95711	0,141267	0,123780675	-0,132138848
-1,59353	0,203206	0,168887209	-1,778524191
-1,59353	0,203206	0,168887209	-0,184989764
-1,59353	0,203206	0,168887209	-0,184989764
-2,32068	0,098207	0,089425003	-0,093679015
-1,95711	0,141267	0,123780675	-0,132138848
-1,59353	0,203206	0,168887209	-1,778524191
-1,59353	0,203206	0,168887209	-0,184989764
-1,95711	0,141267	0,123780675	-2,089244032
-1,95711	0,141267	0,123780675	-0,132138848
-2,32068	0,098207	0,089425003	-0,093679015
-3,41139	0,032995	0,031941444	-0,032462702
Nilai Maksimal			-8,433591991

Dari analisis kedua tabel diatas maka diperoleh hasil sebagai berikut :

Nilai awal B0 = 0,01 dan B1 = 0,02 menjadi Nilai Loglikelihood maksimum B0 = -8,864949558 dan B1 = 0,363570757

Sehingga memperoleh nilai persamaan linear Regression Logistic Y = -8,864 + 0,363 X1